



Previsão de Churn no Retalho Alimentar

**Liliana Andreia de Sousa
Bernardino**

Orientação: João Portela da Gama

Setembro, 2011

Tese de Mestrado em Análise de Dados e
Sistemas de Apoio à Decisão

"A chave de todas as ciências é inegavelmente o ponto de interrogação."

Honoré de Balzac

Aos meus pais Alfredo e Helena, aos
meus avós Sousa e Carmelina e ao
meu avô Fernando.

Agradecimentos

Realizar um trabalho como este implica empenho, concentração, rigor, e o acompanhamento e estímulo das pessoas que estão mais próximas. Na conclusão desta etapa da minha formação académica não posso deixar de agradecer:

Aos meus pais e avós, que me ensinaram a vontade e a insatisfação pelo adquirido e me incentivaram sempre a procurar algo mais. A eles que fizeram de mim o que sou, me formaram e me valoraram como pessoa: é para vocês!

A essa fonte eterna e inigualável de inspiração que me olha e protege: avô Fernando.

À Matilde e ao Miguel, por serem meus desafiadores, e por eles eu sonhar com um mundo melhor todos os dias.

Ao meu orientador, o Professor João Gama, pela sua experiência, sugestões, pela sua visão do todo, paciência nos momentos de indefinição, a sua clareza de pensamento, e acima de tudo pela sua enorme disposição em ajudar.

Ao Dr. Nuno Rendeiro e Dr. Tomás Ribeiro por sempre me permitirem permanecer igual a mim própria e por me ensinarem a inquietação de espírito em querer conhecer todos os dias melhor o cliente Sonae.

À Sofia e à Patrícia, cuja disponibilidade e extraordinário apoio permitiram a possibilidade a este projecto.

À Sonae MC pela autorização da utilização dos dados que serviram de base à elaboração deste estudo, e respectiva publicação do mesmo em âmbito académico.

À Céu pelo incondicional apoio, pelo colo e carinho de todos os dias. À Helena, Mariela, Mónica, Ana, Cristina, Rui e Manuel pela insistência e crença no “és capaz”, e a todos os outros pelo entusiasmo e infinita amizade.

A todos, o meu MUITO OBRIGADA!

Resumo

A verdadeira riqueza para qualquer empresa da actualidade é claramente o seu cliente. Num mercado cada vez mais competitivo, como o sector do Retalho, as empresas foram a pouco e pouco centrando as suas estratégias de gestão na busca por uma vantagem diferenciadora face aos seus concorrentes. Neste âmbito passaram a ser utilizadas ferramentas de transformação de grandes quantidades de dados em retratos fies dos clientes, sendo o *Data Mining* uma fonte cada vez mais recorrente de chamada à tomada de decisões estratégicas na gestão do cliente e na sua satisfação, com a finalidade de promover e consolidar a fidelização deste com a organização.

Esta dissertação de Mestrado está inserida num projecto profissional na área de Retalho, que em tempos de crise como os de hoje, tem sido alvo de estudos cada vez mais assertivos e especializados. O projecto em causa baseia-se na construção e definição de sistemas que permitam classificar e explicar o fenómeno de *churn* dos clientes.

Baseada na premissa de que “manter é mais rentável que conquistar”, o objectivo principal desta dissertação é o de aprofundar o relacionamento entre a empresa e o cliente, utilizando o *Data Mining* para detecção e definição do conceito de cliente *churner* neste sector de mercado. Para o alcance deste objectivo, procurou-se extrair conhecimento através da exploração de um elevado volume de informação, que reside na base de dados transaccionais da empresa em estudo.

Uma das mais-valias da autora no trabalho desenvolvido, e a par da utilização das ferramentas de KDD (*Knowledge Discovery in Databases*), foi o ser portadora já de conhecimento do negócio. Factor de extrema importância na avaliação dos resultados obtidos. Este factor levou à escolha do sub-grupo de clientes Leais, considerados os mais rentáveis para a organização, e sobre eles recaiu todo o trabalho de modelação.

A formação do modelo final envolveu as etapas de análise do problema (*churn*), avaliação e entendimento dos dados, pré-processamento e classificação. Como algoritmos classificadores foram estudados e utilizados os métodos de: Redes Neurais Artificiais, Regressão Logística e Árvores de Decisão. Os resultados obtidos após scoring, validam o modelo desenvolvido, permitindo a sua utilização e aperfeiçoamento em ambiente real.

Palavras-chave: *Data Mining*; Classificação; Retalho; Cliente; *Churn*; Redes Neurais Artificiais; Árvores de Decisão; Regressão Logística.

Abstract

The real asset for any company nowadays is clearly the client. In a market increasingly competitive, as the retail sector, companies were gradually focusing their management strategies in pursue for a distinctive advantage over their competitors. Within this sphere were used tools for transformation of large amounts of data in order to give a trustworthy picture of the customers, being the Data Managing a source of recurring decision making in managing the customer and their satisfaction with the purpose of promoting and consolidate the loyalty of the organization.

This Master's thesis is part of a professional project in retail, which in times of crisis like the present, has been the aim of increasingly assertive and skilled studies. The project is based on the construction and definition of systems that allow the classification and explanation of the customer churn phenomenon.

Based on the premise that it “is more profitable to keep than conquer”, the main gold of this essay is to deepen the relationship between the company and the customer, using the Data Managing to detect and define the concept of customer churn in this market sector. To achieve this goal, was extracted knowledge by exploiting a high volume of information that lies in the database of the company submitted to study.

One of the author's avail in the work developed, together with the use of KDD (Knowledge Discovery in Databases) tools, was the knowledge of business, which was an extremely important factor in evaluating the results. This factor led to the choice of the sub-group of Loyal customers, considered to be the most profitable for the organization, and fell upon them all the work of modeling.

The formation of the final template involved problem analysis (churn), evaluation and understanding the data, processing and classification. As classifying algorithms were studied and used the following methods: Artificial neural networks, Logistic regression and Decision trees. The results obtained after scoring, validate the model developed, allowing its use and improvement in real environment.

Keywords: *Data Mining*, Classification, Retail, Customer, *Churn*, Artificial Neural Networks, Decision Trees, Logistic Regression

Índice

Agradecimentos	i
Resumo	ii
Abstract.....	iii
Índice	iv
Lista de figuras.....	viii
Lista de tabelas.....	x
I Introdução.....	13
1.1. Motivação	14
1.2. Objectivo.....	17
1.3. Estrutura da Dissertação	18
II Enquadramento	19
2. Contextualização.....	20
2.1. Comportamento do cliente.....	20
2.2. Criação de Valor como veículo de Fidelidade.....	22
2.3. Fidelidade/Lealdade.....	23
2.4. Fidelização	24
2.5. Retenção de Clientes.....	26
2.5.1. Quais os clientes que devem ser retidos	28
2.5.2. Abandono de Clientes.....	30
2.6. Gestão de <i>Churn</i>	31
2.6.1. Definição de <i>Churn</i>	31
2.6.2. Causas do <i>Churn</i>	33
2.6.3. Consequências e custos derivados.....	34
2.7. A filosofia de Marketing Cliente na Distribuição.....	35
2.7.1. A distribuição e logística em Portugal.....	37
2.8. A Sonae MC e a Fidelização.....	39
2.8.1. Fidelização dos Clientes na Sonae MC	41
3. Inteligência de processos	44

3.1. Introdução ao Data Mining	44
3.1.1. “Descendência” do Data Mining	45
3.2. Business Intelligence	46
3.2.1. Dados e DataWareHousing	47
3.3. O Processo KDD (Knowledge Discovery in Databases).....	48
3.3.1. Definição e análise do problema	50
3.3.2. Pré-processamento e compreensão dos dados	51
3.3.3. Transformação dos dados	57
3.3.4. Data Mining como parte do processo de KDD	59
3.3.5. Pós-processamento – Interpretação dos resultados	71
3.4. Customer Relationship Management (CRM)	72
3.4.1. Definição CRM	74
3.4.2. CRM na Distribuição.....	75
3.5. Conclusão.....	76
4. Objectivos e Abordagens em Data Mining.....	78
4.1. Introdução	78
4.2. Diferentes objectivos em Data Mining	79
4.2.1. Previsão	79
4.2.2. Descrição	81
4.3. O problema da classificação	83
4.3.2. Métodos e técnicas de classificação	85
4.4. Conclusão.....	90
III Implementação e Resultados	91
5. Trabalho realizado – Método	92
5.1. Descrição do Problema	92
5.1.1. Definição da Target	93
5.2. Software utilizado – <i>SAS Enterprise Miner</i>	95
5.3. Caracterização do <i>data set</i>	96
5.4. Metodologia utilizada	98
5.4.1. Preparação da Base de Dados de trabalho	98

5.4.2. Compreensão dos Dados	100
<i>Variáveis Socio-Demográficas</i>	101
<i>Variáveis Comportamentais</i>	102
<i>Variáveis de Envolvimento</i>	104
5.4.3. Etapas de Pré-processamento	106
5.4.3. Modelação do Churn	114
5.5. Conclusão.....	114
6. Análise de Resultados	116
6.1. Árvores de Decisão	116
6.2. Regressão Logística	120
6.3. Redes Neurais (RNA)	124
6.4. Discussão e comparação dos modelos escolhidos	128
6.5. <i>Scoring</i> do modelo escolhido.....	134
7. Conclusões e trabalho futuro	137
8. Bibliografia	140
ANEXO I	143
I. A Distribuição em Portugal e o papel da SONAE	144
I.1. Contexto e Evolução da Distribuição em Portugal.....	144
I.2. Planeamento/Organização da Grande Distribuição	146
I.3. A Distribuição no Futuro.....	147
I.4. A SONAE	148
I.4.1. O Continente	150
ANEXO II.....	151
II.1. Técnicas inerentes ao processo de KDD – Pré Processamento.....	152
II.1.2. Pré-processamento - Detecção de <i>Outliers</i>	152
II.1.3. Pré-processamento - Tratamento de Missing Values	154
II.1.4. Pré-processamento - Attribute Selection.....	156
II.2. Técnicas inerentes ao processo de KDD – Transformação	159
II.2.1. Transformação dos dados – Discretização de variáveis.....	160
II.2.2. Transformação dos dados – Redução de Dados.....	160

II.3. Técnicas inerentes ao processo de KDD – Pós-Processamento.....	162
II.3.1. Pós-processamento – Avaliação dos modelos.....	162
II.4. Data Mining como parte do processo KDD	167
II.4.1. Algoritmos e Técnicas de DM – Árvores de Decisão.....	167
II.4.1.1. Árvores de Decisão - Algoritmos de Implementação	169
II.4.1.2. Indução de Regras.....	170
II.4.2. Algoritmos e Técnicas de DM – Redes Neurais	171
II.4.3. Algoritmos e Técnicas de DM – Regressão Logística	179
ANEXO III	183
III.1. Caracterização do <i>data set</i>	184
III.1.1. As variáveis.....	184
III.1.1.1. Variáveis Socio-Demográficas	184
III.1.1.2. Variáveis Comportamentais	185
III.1.1.3. Variáveis de Envolvimento	186
III.1.2. Preparação dos Dados.....	187
III.1.2.1. Selecção dos dados e partição da base de dados	187
III.1.3. Estatísticas descritivas das variáveis – Base Loyal	189
III.1.3. Pré-Processamento.....	211
III.1.3.1. Transformações dos dados	211
III.2. Modelação.....	212
III.2.1. Árvores de decisão	212
III.2.2. Regressão Logística.....	213
III.3. Discussão dos Modelos.....	214
III.3. Diagrama SAS	215

Lista de figuras

Figura 1- Diagrama comportamento de compra do consumidor	22
Figura 2 - Processo de Business Intelligence.....	46
Figura 3 - Ciclo KDD (by FAYYAD et al. 1996)	50
Figura 4 - pilares para o sucesso do DM (Linoff 2002).....	61
Figura 5 - Estudo sobre a preferência de metodologia de DM	64
Figura 6 - Ciclo de vida da metodologia CRISP-DM.....	65
Figura 7 - Sequência comum das fases de CRISP-DM	67
Figura 8 - ciclo de vida da metodologia SEMMA (adaptado de SAS, 2005).....	68
Figura 9 - comparação das interacções entre as fases SEMMA e CRISP-DM	70
Figura 10 - processo de classificação de dados (by Han et al. 1996)	84
Figura 11 - Esquema de um classificador baseado em Árvores de Decisão.....	87
Figura 12 - Representação de uma Rede Neuronal Multicamada	88
Figura 13 - Variação da distribuição da variável Segm_age de Target=1 face a base .	102
Figura 14 - Variação da distribuição da variável Household de Target=1 face a base.	102
Figura 15 - histograma da variável Hygienicprd_ratio_cat_3_6_vl	104
Figura 16 - histograma da variável Hygienicprd_ratio_cat_6m_visitas	104
Figura 17- histograma da variável Tx_rsp_m.....	104
Figura 18 - Box-Plot das variáveis Visitas e Transacções no mês actual.....	105
Figura 19 - Esquema inicial SAS Miner	110
Figura 20 - Nó de transformação das variáveis	112
Figura 21 - Exemplo de transformação de variáveis	112
Figura 22 - Nó Attribute Selection by R-Square	113
Figura 23 - Selecção do melhor critério de divisão AD	117
Figura 24 - Árvore de Decisão Entropia	118
Figura 25 - Variação do Misclassification Rate X N° Folhas – Entropia	119
Figura 26 - Selecção do melhor método de Regressão Logística	121
Figura 27- Gráfico ROC Regressão Logística	122
Figura 28 - Selecção do melhor método de Redes Neurais	125
Figura 29 - Ganhos Cumulativos do modelo RN3	126
Figura 30 - Lift do modelo RN3	126
Figura 31 - Misclassification Rate X N° iterações RN3	127

Figura 32 - Diagrama SAS (comparação Global).....	129
Figura 33 - Matrizes de confusão modelos escolhidos	130
Figura 34 - Curvas ROC modelos seleccionados	131
Figura 35 - Score Distribution dos modelos seleccionados.....	131
Figura 36 - Ganhos Cumulativos dos modelos seleccionados.....	132
Figura 37 - Lift dos modelos seleccionados	132
Figura 38 - Scoring do modelo escolhido	135
Figura 39 - Matrizes Confusão VS Scoring.....	136
Figura 40 – Logo SONAE	148
Figura 41 - Logo Continente.....	150
Figura 42 - Procedimentos Selecção Atributos (Liu; Yu)	157
Figura 43 - Matriz de confusão de um classificador.....	162
Figura 44 - Espaço ROC	164
Figura 45 - Curvas ROC classificador B (gerador de vários pontos) e Classificador A (um só ponto no espaço ROC).....	165
Figura 46 – Estrutura de um neurónio natural típico	172
Figura 47 - Estrutura de um neurónio artificial típico	173
Figura 48 - Rede de uma só camada (<i>perceptron</i>)	175
Figura 49 – Topologia de uma rede <i>feedforward multicamada</i>	175
Figura 50 - Topologia típica de uma rede competitiva ou recorrente.....	176
Figura 51 -Forma da relação logística entre variáveis dependente e independente (by Hair et. al, 2005).	181
Figura 52 - Pré-Seleção e Partição inicial da BD em SAS Guide.....	188
Figura 53 – Histograma e Box-Plot da variável IPC original.....	189
Figura 54 - Histograma e Box-Plot da variável original CompetitionIndex.....	190
Figura 55 - Frequências da variável Gender_CD	190
Figura 56 - Frequências da variável Segm_Age	190
Figura 57 – Frequências da variável HouseHold.....	190
Figura 58 - Frequências da variável Municipality_cd	191
Figura 59 - Frequências da variável District_cd.....	191
Figura 60 - Frequências da variável Segm_LifeStyle.....	191
Figura 61- Frequências da variável Segm_Value	191

Figura 62 - Histogramas das variáveis comportamentais originais	200
Figura 63- Box-Plot das variáveis comportamentais originais	206
Figura 64 - Histogramas das variáveis de envolvimento	207
Figura 65 – Box-Plot das variáveis de envolvimento	207

Lista de tabelas

Tabela 1 - Caracterização dos diferentes segmentos “FV” de Clientes.....	93
Tabela 2 - Contadores dos <i>dataset</i> após a primeira filtragem e partição	100
Tabela 3 - Correlação com a <i>target</i> (variáveis de envolvimento).....	106
Tabela 4 - Quantidade e percentual de clientes por faixa de probabilidade de <i>Churn</i> (AD)	119
Tabela 5 – Estatísticas Seleccção dos modelos RL baseada no teste Kolmogorov-Smirnov.....	121
Tabela 6 - Teste de Significância RL Stepwise	123
Tabela 7 - Quantidade e percentual de clientes por faixa de probabilidade de Churn Stepwise	123
Tabela 8 - Estatísticas Modelos Redes Neurais	125
Tabela 9 - Matriz de confusão Modelo RN3	126
Tabela 10 - Quantidade e percentual de clientes por faixa de probabilidade de Churn RN3.....	127
Tabela 11 - Comparação dos Algoritmos de modelação	128
Tabela 12 - Estatísticas Modelos Seleccionados	130
Tabela 13 - Variáveis importantes ADEntropia (TOP 18)	133
Tabela 14 - Contador de registos BD de Classificação	135
Tabela 16 - Conjunto de variáveis Socio-Demograficas	184
Tabela 17 - Conjunto de variáveis Comportamentais (Parte 1).....	185
Tabela 18 - Conjunto de Variáveis Comportamentais (Parte 2).....	186
Tabela 19 - Conjunto de Variáveis de Envolvimento.....	187
Tabela 20 - Estatísticas Variáveis Comportamentais originais	193

Tabela 21 - Descrição das correlações de Pearson	208
Tabela 22 - Matriz de correlação das variáveis de Envolvimento	209
Tabela 23 - Variáveis para pré-processamento	210
Tabela 24 – Estatísticas das variáveis de entrada SAS Miner	211
Tabela 25- Estatísticas Árvores de Decisão.....	212
Tabela 26 - Legenda SAS para as árvores geradas	212
Tabela 27 - Estimativas dos parâmetros associadas ao modelo logístico Stepwise	213
Tabela 28 - Variáveis ADEntropia por ordem de importância	214

I Introdução

1.1. Motivação

Para qualquer empresa da actualidade a noção de onde reside a sua verdadeira riqueza, há muito que deixou de ser um segredo – os seus clientes. Estes são quem sustenta a operação da própria empresa, e em última análise, são a própria razão de existência da empresa.

Assim, e dada a constante mutação dos mercados, o foco de preocupação das empresas há muito que deixou de estar no produto que vende passando a ser direccionado para o mercado e para o cliente. Neste contexto, o marketing de massas foi-se tornando em marketing direccionado estando cada vez mais focalizado para uma vertente mais individualizada a cada cliente (*marketing one-to-one*¹).

Actualmente os clientes têm à sua disposição uma ampla gama de produtos e serviços, fazendo as suas escolhas sob percepções de qualidade, serviço e valor. Assim, e num mercado cada vez mais competitivo, as empresas foram a pouco e pouco centrando as suas estratégias de gestão na busca por uma vantagem diferenciadora face aos seus concorrentes.

É inequívoco que esta busca assenta principalmente na capacidade de inovação do domínio tecnológico e no recurso à informação que o próprio cliente dá de si próprio. Só com a relação bilateral destes dois factores é possível a qualquer empresa garantir um diferencial competitivo e vantajoso para ambas as partes (empresa e cliente).

Segundo Gensch *et al.*², a criação de um modelo inteligente sobre uma base de dados, permite e sob a óptica do cliente, a possibilidade de identificação de serviços mais adequados às suas necessidades, pagando menos e recebendo mais, enquanto para a empresa se percebe a possibilidade de vender a mais clientes com perfis de consumo

¹ Marketing One to One - O conceito (criado por Don Peppers) de marketing one to one (ou marketing individualizado) consiste na venda de tantos produtos e serviços quanto seja possível a cada cliente - o chamado conjunto alargado de necessidades - por uma vida inteira, em vez da venda de um produto ou serviço de cada vez para o maior número possível de consumidores. Este conceito serve, sobretudo, para obter a lealdade dos clientes.

² GENSCH, Denis; AVERSA, Nicola; MOORE, Steven.. **A Choice-Modeling Market Information System That Enabled ABB Eletric to Expand Its Market**. Interfaces Magazina, January-February, 1990.

potencialmente diferentes. Desta forma surgem as oportunidades de retenção da base de clientes criando laços de fidelidade por parte dos consumidores, bem como o aumento das receitas que advém dessa relação para com a empresa.

Ao longo dos últimos anos, a mutação dos mercados e das mentalidades gerou grandes alterações nas formas de “fazer negócio”. O investimento em clientes considerados “não lucrativos” passou a dar lugar à necessidade de garantir a permanência de um cliente cuja rentabilidade para a empresa seja segura e constante, justificando assim muitas vezes a renúncia ao lucro imediato por parte das empresas dando este lugar a uma rentabilidade a longo prazo.

Esta nova forma de operação fica explicada pela evolução dos custos que a aquisição de novos clientes implica face à receita média decrescente gerada pelos mesmos. Recorrendo à história recente das empresas, facilmente se prova que se torna mais dispendioso para uma empresa “recrutar” novos clientes provenientes de um concorrente, uma vez que a proposta terá que ser bastante mais apelativa do que aquela a apresentar a um cliente que já face parte do portfólio da empresa, com o qual a empresa já tem contacto e dados históricos para que seja possível uma retenção mais duradoura e rentável.

No âmbito da oferta de serviço e da recolha de dados, as ferramentas de fidelização, têm surgido com objectivos claros centrados na diferenciação em relação a empresas concorrentes, evitando a perda de clientes para essas mesmas empresas do mesmo sector, com uma série de medidas que contribuam com um valor acrescentado aos clientes. É desta forma que um qualquer acto de consumo identificado por parte de um cliente, se transforma numa oportunidade única para premiar a eleição do consumidor e a confiança na marca que escolhe para efectuar as suas compras. De uma forma geral, pode dizer-se que quaisquer acções de satisfação para com o cliente são os primordiais fios condutores para que este tenha menos motivação de optar por uma outra marca, tendo ao mesmo tempo um maior incentivo para a continuidade de compra de produtos da empresa.

Com ou sem ferramentas físicas de fidelização³, a recolha dos dados da actividade da empresa gera uma imensa quantidade de dados que representa uma valiosa fonte de informação, passível de ser transformada em conhecimento para o apoio nos processos de tomadas de decisão praticamente a todos os níveis do ramo empresarial.

No entanto, a transformação dessa massa de informação em conhecimento é uma tarefa fragosa e exigente de engenho científico e intuitivo. É exactamente neste contexto que surge Data Mining que não é nada mais do que a combinação de grande poder computacional, algoritmos inteligentes, conhecimento do negócio e intuição humana aplicada a grandes massas de dados, objectivando a descoberta do conhecimento útil para o negócio mas oculto em milhares de terabytes de dados (KELL, 1995).

É de extrema importância a percepção de que a eficácia da aplicação de modelos de Data Mining ou a solução ideal para os problemas de uma empresa, não se atinge através da simples aplicação de algoritmos sofisticados como Redes Neurais, árvores de decisão ou algoritmos genéticos, a um qualquer conjunto de dados. É essencial que à perícia na utilização deste tipo de modelos computacionais se una um profundo conhecimento do negócio e de todas as particularidades do problema que visa ser solucionado. Só a combinação destes factores permite gerar a inteligência efectiva sobre o negócio que pode vir a ser utilizada nos processos de tomada de decisão.

Neste trabalho procura-se a solução de um problema de gestão da relação com o cliente, através da criação, estudo e optimização do processo de Data Mining. A questão da perda de um cliente para um concorrente (Churn) numa determinada área do negócio de Retalho será avaliada em busca de uma solução satisfatória para que seja possível à empresa detectar e agir sobre o cliente que está prestes a perder. Um processo de Data Mining diligente, partindo do estudo do problema, passando pela recolha e tratamento adequado da informação, culminando na busca e optimização do modelo mais adequado para esta tarefa será o objectivo principal deste trabalho. Tem-se também como objectivo a demonstração da vantagem competitiva que pode ser gerada com o recurso a este tipo de técnicas e estratégias de retenção de clientes através da análise de rentabilidade do modelo desenvolvido.

³ Cartões magnéticos por exemplo

1.2. Objectivo

O objectivo geral deste trabalho é a criação de um modelo dinâmico de detecção de *Churners* no âmbito da base de dados de Clientes com cartão de Fidelização da Sonae MC. **Este modelo deverá ser desenvolvido com base no perfil de compra dos clientes** e deverá ser capaz de lançar alertas com uma fiabilidade elevada de quais os clientes que num espaço de tempo t irão deixar de comprar nas lojas da Marca.

O modelo deverá ser capaz de, com base no mesmo perfil de compra, identificar e retornar informação relativa ao comportamento indicativo de que o cliente irá deixar de fazer compras, de forma a que possa ser útil na prevenção de decrementos nas vendas para os vários segmentos de clientes. Sendo do conhecimento comum que os clientes se comportam de formas completamente distintas relativamente ao seu estilo de vida, existe uma pré-definição de *clusters* de clientes, sobre os quais está assumida e é esperada uma frequência de compra distinta. Assim, existe a necessidade de explorar vários tipos de *churn*. Não deve ser estudada a questão do *abandono* de clientes de forma cega e standard para todo o tipo de clientes. Não só porque a sua tipologia de compra, nomeadamente no que respeita à frequência é totalmente distinta, e um modelo global poderá estar provido de ruído, bem como a definição de Abandono de Cliente (*churn*) é visto como sendo algo difícil e muitas vezes enganador. Isto porque, a definição mais directa de *Churn* o é o facto de um cliente deixar de o ser (cancelamento de contratos) – algo que na indústria retalhista não é simples e directo de medir. Só porque alguém deixa de utilizar o seu cartão nos últimos 3 meses não significa, necessariamente, que efectivamente Abandonou.

Assim, propõe-se que o modelo criado possa ser adaptado de forma bilateral. Ou seja, a determinação de um cliente *churner* deverá contemplar a tipologia do seu consumo, e essa mesma tipologia deverá ser indicador da definição da variável Target. Ou seja, para um cliente que tenha um comportamento de compra bisemanal, será importante alertar para uma alteração de comportamento, a que neste trabalho vamos chamar de *churn*, se o cliente não compra no mês seguinte. No caso de clientes com um comportamento de compra mais esporádico, esta definição ficará alocada aos 3 meses sem compras.

Considera-se ainda importante e preponderante o factor de dinamismo no modelo construído. Dadas as constantes mutações no mercado do retalho, bem como a sazonalidade inerente às vendas no negócio retalhista, o modelo deverá ser capaz de reconhecer temporalmente a não preferência (ou abandono) do cliente.

1.3. Estrutura da Dissertação

A dissertação apresentada explora um modelo completo de *Data Mining*, com enfoque preditivo na classificação de clientes da Sonae MC. Apresenta-se uma revisão bibliográfica na qual se contemplam todas as etapas do processo de descoberta de conhecimento, aplicadas na construção de um modelo desenvolvido em ambiente de *case study*.

Os Capítulos 2, 3 e 4 são dedicados, respectivamente, à contextualização teórica do problema de *churn* e a uma exposição revista dos processos envolventes num projecto de Data Mining, nomeadamente ao problema específico de Classificação.

A descrição do trabalho desenvolvido, bem como contextualização prática do problema em análise são apresentados no Capítulo 5.

Por consequência, no Capítulo 6 são expostos e discutidos os resultados obtidos com a modelação de cada um dos algoritmos de classificação utilizados, com uma análise final comparativa de desempenho, e são também *scorizados* os clientes de uma nova base de dados, com base no modelo final.

II Enquadramento

2. Contextualização

Neste capítulo são apresentados os principais aspectos que compõem o cenário de actuação desta dissertação.

São focadas teorias de comportamento de compra do consumidor e as suas implicações nas estratégias de retenção para as empresas. São também discutidas alternativas estratégicas que permitem o crescimento e a manutenção do lugar num mercado cada vez mais competitivo e exigente.

Num contexto de gestão do *churn*, a importância da manutenção dos clientes actuais e a actuação sob medidas correctivas no processo de relação com o cliente, para para o benefício económico das empresas é também explorada.

2.1. Comportamento do cliente

Pode dizer-se que o conceito de ***marketing relacional*** propõe que uma organização deve satisfazer as necessidades e os desejos do consumidor para que possam ser gerados lucros. Para fazer valer esse mesmo conceito, as organizações precisam compreender os seus clientes e criar com eles uma ligação próxima, a fim de oferecer produtos e/ou serviços que por eles sejam valorizados. Esta ligação deverá ser estabelecida de forma transversal nas empresas, não sendo uma tarefa exclusiva dos responsáveis das áreas de marketing, já que muitas vezes, e por exemplo, a implementação de estratégias de geração de valor comportam valores e políticas alternativas ou complementares às sediadas nas organizações.

Uma vez que o propósito do *marketing* se centra em atender as necessidades dos consumidores para Kotler e Keller (2006), o conhecimento do comportamento de compra do consumidor torna-se fundamental para o delinear de estratégias comerciais e de negócio.

Para Engel *et al* (2000), entender o comportamento do consumidor é imprescindível para que as organizações qualifiquem os seus produtos e serviços, considerando os desejos e necessidades do consumidor para que desta forma possam orientar as suas ofertas no mercado e ainda atingir a compreensão de um dos maiores desafios da filosofia de marketing – a retenção de clientes.

Para que a comercialização de produtos ou serviços seja efectuada de forma a manter os clientes e a gerar lucro para a empresas, é necessário entender o **porquê** dos consumidores comprarem uns produtos/serviços ao invés de outros (Churchill; Peter, 2007).

Kotler e Armstrong (1999), defendem que o comportamento do consumidor tem um carácter dinâmico e que sofre mudanças causadas por factores sociais, culturais, pessoais e psicológicos. Qualquer alteração no quotidiano ou nas condições socio-economicas do individuo, acabam por afectar de alguma forma as suas decisões de consumo.

Autores como Churchill e Peter (2007) e Engel *et al* (2000) definem que o processo de compra e consumo parte do reconhecimento de uma necessidade até a avaliação pós-consumo.

O reconhecimento da necessidade pode acontecer em virtude de estímulos internos (fome ou cansaço, por exemplo) ou externos (como a o contacto com uma campanha publicitária), que levam o consumidor a querer satisfazer essa mesma necessidade ou desejo. Após esse reconhecimento, o consumidor procura informações sobre como satisfaz-la que posteriormente avalia para identificar a compra que lhe trará o melhor valor.

Após essa avaliação o cliente decide ou não pela compra do produto, que em caso afirmativo leva ao passo seguinte de decisão de quando e como pagar pelo produto ou serviço. Depois da compra, o consumidor tende sempre a avaliar formal ou informalmente, o resultado da sua experiência. Desta forma, a tomada de decisão não pára com o consumo, havendo portanto uma avaliação continua e dinâmica do produto ou serviço sobre a qual os consumidores se definem satisfeitos ou insatisfeitos com o seu contacto com a organização na qual adquiriram o bem comercial. O resultado desta

avaliação final afecta, e muitas vezes de forma definitiva, as intenções futuras do consumidor como é o caso da permanencia ou a recomendação a terceiros.

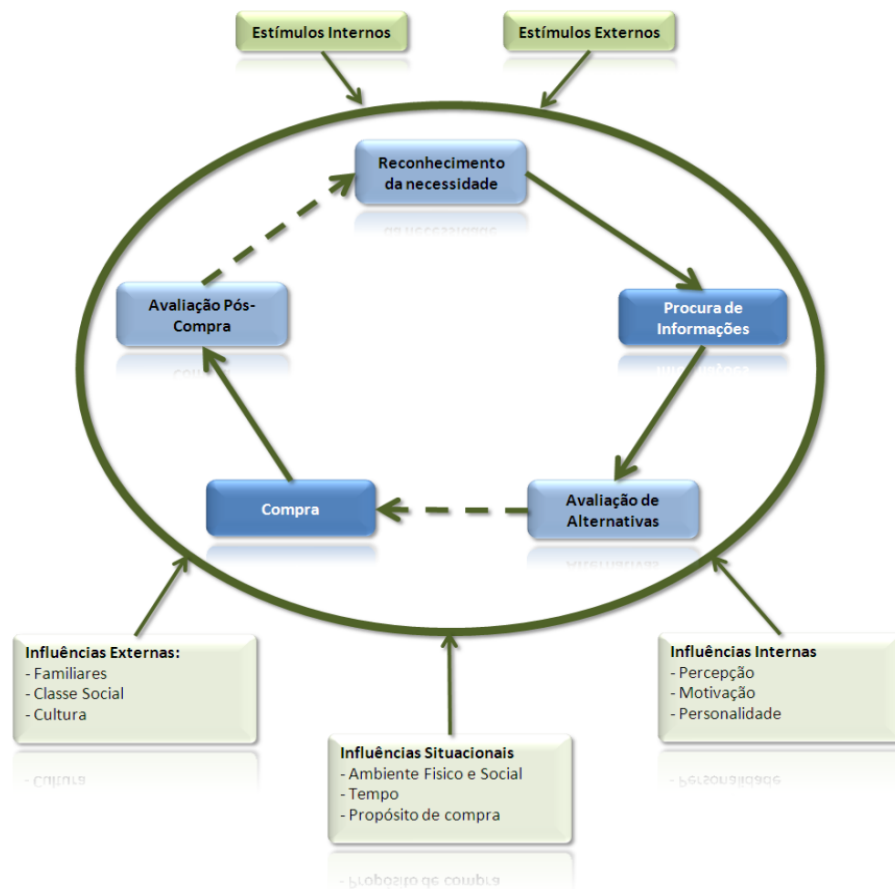


Figura 1- Diagrama comportamento de compra do consumidor

2.2. Criação de Valor como veículo de Fidelidade

Existem diversas estratégias para criar valor aos clientes de uma forma sustentável, ou seja gerando simultaneamente valor para a organização. De entre as alternativas estratégicas que permitem às empresas crescer e se desenvolverem num ambiente cada vez mais competitivo estão: a orientação para a produção, orientação para o produto,

orientação para vendas, orientação para marketing, estratégias genéricas, orientação para o valor e orientação para o mercado.

Sob diversas teorias o valor percebido pelo cliente não depende apenas de características do serviço como o produto, o preço ou mesmo a promoção de massas. Segundo Churchill e Peter (2007), os profissionais de marketing podem aumentar o valor para o consumidor aumentando os benefícios e/ou reduzindo os custos para os clientes.

Como foi já discutido na secção anterior, uma oferta de valor real para o cliente pode conduzir à fidelização dos clientes, e por consequência, levar a relações duradouras e lucrativas entre clientes e empresas. Um relacionamento organização-cliente que esteja bem sustentado, muda a perspectiva do cliente que passa a encarar a compra como uma necessidade de interacção continuada com a organização da venda, criando laços de comodidade e de confiança.

2.3. Fidelidade/Lealdade

Fidelidade corresponde ao grau em que os clientes estão predispostos a permanecer com a organização e a resistir a ofertas da concorrência.

Segundo Kotler (2003), e apesar da fidelidade à marca se poder ler aproximadamente na taxa de retenção de clientes, não pode ser medida de forma crua sobre estes valores. Muitas vezes, valores elevados nessas mesmas taxas podem simplesmente querer dizer que os clientes não abandonam a organização por indiferença ou inércia, ou mesmo por serem reféns de contratos de longo prazo.

Ainda que o objectivo da empresa seja desenvolver o seu portfólio de clientes fiéis, a fidelidade e a relação do consumidor não é forte o suficiente para resistir a, mesmo que momentaneas ou experimentais, investidas de um concorrente com propostas de valor mais convincentes.

Para Lovelock e Wright (2002), a **fidelidade** num contexto empresarial, corresponde apenas à vontade do cliente de prestigiar a longo prazo uma empresa, adquirindo os seus produtos de forma preferencialmente exclusiva e recomendando a marca da organização a outros indivíduos do seu círculo social.

Neste âmbito surge o conceito de **lealdade à marca**, que corresponde à probabilidade do consumidor mudar para outra marca, especialmente se a marca concorrente fizer uma mudança de preço ou nas características do produto, que o tornem mais apelativo para o cliente. O aumento da lealdade à marca, é inversamente proporcional à vulnerabilidade dos consumidores à acção da concorrência.

Mesmo sendo semelhantes, autores como Kotler (2000) defendem que podem ser interpretados de forma diferente. Enquanto **lealdade** é a medida percentual do volume de compras que os clientes fazem num dado fornecedor em relação ao total comprado em todas as organizações que vendem o mesmo produto, **fidelidade** pode ler-se aproximadamente como a taxa de retenção de clientes de um produto ou de uma marca.

2.4. Fidelização

O conceito de **Fidelização**, corresponde a um conjunto de acções realizadas pelas organizações no sentido de fidelizar ou reter os seus clientes mais valiosos, de forma a impedir a sua evasão para a concorrência.

Um cliente fiel é então um cliente envolvido com a marca, o que significa que não muda os seus hábitos de compra por pequenos detalhes ou desagradados, e que mantém o consumo frequente. Um cliente fiel é portanto um cliente que estabelece com a organização, a sua marca e os seus produtos uma relação de longo prazo.

Num mercado cada vez mais competitivo e difícil, a procura pelo envolvimento dos clientes fez com as empresas procurassem desenvolver ferramentas e metodologias que lhes permitissem estabelecer um processo de marketing relacional mais duradoura, simples e explorando da melhor forma o potencial dos clientes. Entre essas ferramentas, e uma das mais utilizadas é sem dúvida os **Programas de Fidelização**.

Este conceito envolve por si uma troca: a empresa oferece vantagens e benefícios aos consumidores que dão preferência aos seus produtos. Estas vantagens devem, no entanto, ser diferenciadoras. Num contexto de recompensa à fidelidade de um cliente, se os benefícios oferecidos pela empresa em nada se destacarem da oferta da concorrência, facilmente o cliente trocará o fornecedor, e desta forma não se está a promover a fidelização.

Um dos tipos de programas de fidelização mais utilizado por todo o mundo denomina-se de “programa de recompensa”. Estes programas são baseados na frequência de compra e são facilmente “copiados” pelos concorrentes fazendo com que o programa deixe de ser percebido como atractivo pelo cliente.

Para que as empresas disponham de uma oferta efectivamente valorizada e procurada pelo cliente, é necessário que as mesmas se dediquem a uma aprendizagem continua sobre as interacções com os seus clientes, por meio da integração dos diferentes canais de relacionamento e pela obtenção de informações prestadas pelos próprios clientes aquando da adesão aos programas, no momento em que efectuem as suas compras, ou mesmo quando contactam com a organização como sugestão ou reclamação de serviços.

Aprender com os clientes é conhecer, melhor do que outras empresas as suas características e necessidades, gerando um vínculo que torne inconveniente para o cliente, mudar para um concorrente com o qual teria que reiniciar todo o processo de relacionamento e confiança.

Para fidelizar o cliente é necessário oferecer mais do que brindes ou recompensas. É essencial uma constante aprendizagem com o cliente, para o consequente aprimoramento na oferta de soluções superiores às apresentadas pela concorrência para que a **retenção** seja possível.

Além disso, é importante que a empresa defina o foco da sua estratégia de retenção: devem ser concentrados esforços para evitar que os clientes abandonem a organização ou para recuperar clientes já evadidos? Se já é dispendioso conquistar, no lugar de manter, clientes, os custos envolvidos agravam-se quando o objectivo atenta na recuperação de clientes que já abandonaram a empresa, e neste caso as garantias de sucesso são também muito reduzidas.

2.5. Retenção de Clientes

Inúmeros autores como Engel *et al* (2000) e Kotler e Keller (2006), têm vindo a defender que a **retenção dos clientes** deve assumir um papel ainda mais importante do que a própria conquista de novos clientes. Esta afirmação justifica-se com inumeros estudos (Kotler e Asmstrong – 1997), que indiciam que é cerca de 5 vezes mais barato para uma organização manter os clientes actuais do que atrair novos clientes e que depois de reconquistado um cliente custa até doze vezes mais para que alcance o mesmo nível de benefício aquando do abandono. Acresce a estes estudos o conhecimento empirico de que num mercado de serviços cada vez mais maduro e competitivo, a perda de clientes pode tornar-se devastadora e decisória no fracasso de politicas comerciais.

Segundo Hoffman *et al* (1997) é ainda de realçar a actual escassez de novos consumidores. Tem vindo a observar-se nos ultimos anos, não só uma diminuição do crescimento populacional dos países desenvolvidos ou em desenvolvimento, bem como a redução do crescimento do produto interno bruto desses mesmos paises. Estes dois movimentos negativos conjugados, fazem com que as organizações se defrontem com uma realidade de relativa redução de quantidade de novos consumidores, proporcionalmente ao passado, e ainda com um decréscimo do nível de consumo dos consumidores existentes.

Uma outra razão acresce ainda à importância da retenção do consumidor. Esta é salientada por Hoffman e Bateson (1997), que consideram o aumento da concorrência como factor preponderante. Com o aumento da concorrência, aumentam as possibilidades de escolha e diminui a amplitude de uma organização se manter com uma proposta que *per-si* seja diferenciadora, tornando os clientes fiéis apenas se estiverem completamente satisfeitos.

Toda esta dinâmica de mercado fez com que as organizações e profissionais de marketing começassem a mudar o seu comportamento perante os novos e cada vez mais exigentes consumidores, fazendo com que o proprio conceito de marketing tenha evoluído.

É desta forma que o *velho* conceito de **marketing** deixa de ter o seu centróide na empresa e passa a tê-lo no mercado. A orientação passa então de vender com o objectivo de persuadir o consumidor e adaptar as necessidades deste às características do produto, para o foco no consumidor e nas suas necessidades, passando a ser a empresa a adaptar-se às necessidades dos seus consumidores.

Estas novas tendências de *fazer marketing* não são apenas orientadas à satisfação única do cliente. As empresas e organizações também ganham com a aplicabilidade destas novas políticas. Estudos como o de Reichheld e Sasser, comprovam que 5% de aumento na retenção de um cliente podem ser traduzidos em 85% de aumento de benefício (lucro) para a organização.

Validando este estudo, Kotler (2000) referenciou alguns aspectos acerca dos clientes **muito satisfeitos** que podem justificar estes valores:

- a) Os clientes muito satisfeitos permanecem **fiéis por mais tempo**;
- b) Esses clientes **compram mais produtos** da empresa à medida que ela lança produtos ou aperfeiçoa os produtos existentes;
- c) Estes tendem a tornar-se **advogados de defesa** da empresa, recomendando-a e aos seus produtos;
- d) Os clientes muito satisfeitos prestam **menos atenção às marcas e campanhas de outros concorrentes**;
- e) São menos sensíveis ao preço;
- f) **Sugerem melhorias** e dão idéias sobre produtos ou serviços;
- g) O atendimento é mais simples do que a novos clientes, uma vez que as transacções e contactos com a organização já se tornaram rotineiras;
- h) A taxa de **lucro por cliente tende a aumentar ao longo do tempo** de vida de um cliente retido.

Resumidamente pode concluir-se que à medida que a satisfação do consumidor aumenta, maior é o sentimento de afinidades e de ligação com os produtos e/ou serviços de uma empresa, que se manifesta pelo desejo do consumidor permanecer com a empresa retornando em lucros financeiros directos.

No entanto, pesquisas realizadas por alguns investigadores como Ganesh *et al* (2000), sugerem que os esforços de retenção de clientes e de construção de uma relação de fidelidade não devem ser aplicados a toda a base de clientes da empresa, uma vez que os consumidores diferem no seu valor/benefício para a empresa. Torna-se portanto primordial para as empresas que estas tenham um conhecimento detalhado da sua base de clientes por forma a conhecer quais são os de maior valor, que geram maior rentabilidade, e quais as suas necessidades de forma a mantê-los satisfeitos e leais.

2.5.1. Quais os clientes que devem ser retidos

Existem inúmeras definições de **cliente lucrativo**, e inúmeros estudos de opinião. No entanto, não obstante a definição transversal do conceito, a clarificação dependerá em grande parte da realidade da organização, do tipo de mercado e da medida de rentabilidade da própria empresa.

Kotler e Armstrong (1999) definem **cliente lucrativo** como “uma pessoa, família ou empresa que dá receitas ao longo do tempo, num montante considerável que seja superior aos custos da empresa para atrair, vender ou prestar o serviço a esse indivíduo”.

Já Day (2001), ressalva que os clientes ideais a serem retidos, podem variar de empresa para empresa, uma vez que os clientes ideais são aqueles para quem a empresa pode fazer mais pelo facto das suas aptidões se adaptarem melhor às necessidade e às condições do consumidor.

Reichheld (2001) incorpora à afirmação de Kotler e Armstrong uma visão de longo prazo e aspectos relativos a conquista, desenvolvimento e manutenção da lealdade dos clientes: “clientes que oferecem fluxos de retorno rentável constantes para o investimento da empresa nos próximos anos, e cuja lealdade possa ser conquistada e mantida. Este autor infere que alguns clientes fornecem à empresa maior recompensa por: serem inerentemente previsíveis e leais, preferirem relacionamentos duradouros e

estáveis, serem mais rentáveis que outros, e por acharem os produtos e serviços mais valiosos do que os da concorrência.

Estudos mais recentes de autores como Ryals (2005) e Reinartz (2005) indicam que para entender as diferenças entre os clientes que afectam positiva ou negativamente os fluxos de vendas da empresa é necessário fazer o cálculo aproximado do valor e da rentabilidade ao longo do ciclo de vida do cliente com a empresa (**Lifetime Value**). Ou seja, devem ser estimados os retornos futuros do cliente ou do segmento de clientes, e deduzir os custos estimados ao longo do tempo do relacionamento entre cliente e empresa. Desta forma, serão contabilizados os custos e benefícios das acções de marketing, de vendas e de todo o tipo de interacções com os consumidores. A maioria destes custos é de simples identificação uma vez que se tratam de custos de aquisição, atendimento, manutenção e retenção como a publicidade dirigida, comissões de venda, alterações de preços.

Para Reichheld (2001) e Ryals (2005), na estimativa dos custos futuros devem ser também contabilizados os custos dos prováveis produtos a serem consumidos pelo cliente e dos custos de serviço de todos os contactos e interacções da empresa com o cliente.

O desafio de reter clientes vai muito além da identificação dos clientes mais valiosos num longo prazo com a empresa, ou até mesmo da escolha das estratégias de retenção. Após a identificação dos clientes mais valiosos para a empresa, é necessário que a mesma se certifique das verdadeiras razões pelas quais o cliente se mantém fiel ou porque motivo abandonou o serviço prestado pela empresa. Não haverá com certeza muito a fazer, por exemplo com clientes que mudam de região ou que por motivos culturais deixam de adquirir o produto ou serviço (Vegetariano que deixa de adquirir carne)

2.5.2. Abandono de Clientes

Estatísticas do mercado americano revelam que as empresas perdem cerca de metade dos seus consumidores em cinco anos e que, nessa taxa de perda de clientes, as mesmas empresas vêm a sua performance financeira reduzida de 25% a 50% (Ganesh et al – 2000).

As elevadas taxas de abandono de clientes na maior parte dos mercados tornam de extrema importância que as empresas analisem a sua base de clientes, com vista a identificar quais os clientes propensos de abandono e quais as causas, de modo a que essa informação permita reduzir futuras *fugas* de clientes.

Dependendo do negócio, alguns clientes propensos a abandonar a empresa, são fáceis de identificar porque cancelam as suas contas ou os seus contratos de permanência, outros simplesmente vão reduzindo o volume das suas compras, passando a comprar muito menos dos produtos ou serviços.

Ryals (2005) aponta maioritariamente duas razões fulcrais para o abandono dos clientes:

1) Voluntárias

a. Preço

- i. Concorrente oferece preço mais competitivo
- ii. Promoções mais atrativas num concorrente com o mesmo serviço e/ou produto
- iii. Custos envolvidos no processo de realização da compra
- iv. (...)

b. Qualidade dos serviços

- i. Declínio na qualidade dos produtos
- ii. Concorrentes oferecem serviço pós-venda de maior qualidade
- iii. Produtos com baixo valor percebido pelo cliente
- iv. Má experiência ou experiências menos boas repetidas nos últimos contactos do cliente com a empresa, ou por indivíduos do círculo social do cliente
- v. (...)

2) Involuntárias

- a. Desregulação de mercados: fomenta a concorrência entre as empresas o que facilita a alguns consumidores a experimentação ou a ceder a apelos promocionais de empresas concorrentes
- b. Deixam de ter necessidade no produto ou serviço oferecido pela empresa (por exemplo um casal que deixou de comprar roupa de bebé)
- c. Motivos sociais
- d. Mudança de zona geográfica do consumidor para uma outra área em que a empresa não se faz representar.

Posto isto, e dada a diversidade de razões para um cliente abandonar a empresa, compete à empresa em todo o processo de retenção avaliar o motivo pelo qual o cliente abandonou o seu produto ou serviço, de forma a que para posteriores situações sejam possível evitar a causa de abandono do cliente. Não será de grande valia para a empresa saber o **como** se não conseguir diagnosticar o **porquê**.

2.6. Gestão de *Churn*

2.6.1. Definição de *Churn*

Segundo Ikeda (2006), *Churn* consiste no acto de um cliente abandonar uma organização em favor de uma concorrente, pondo termo a toda a relação com aquela e iniciando uma nova relação com esta.

Este fenómeno tem vindo a ser cada vez mais analisado na realidade *marketeer*, e principalmente na no mercado das telecomunicações. No entanto, e com os novos mercados e aplicações de práticas de fidelização em ascensão, torna-se uma problemática transversal para qualquer realidade comercial. Sendo visto pelo marketing como a perda de clientes sofrida em detrimento de um concorrente, é considerada como uma medida da **infidelidade** dos clientes.

Diz-se então que se a taxa de *churn* de uma organização varia entre 15% e 25% anualmente, então significa que a empresa está a perder anualmente entre 15% a 25% dos seus **clientes activos**⁴, que por algum motivo não ficaram satisfeitos e abandonaram a empresa (voluntária ou involuntariamente).

Em mercados competitivos, e principalmente aqueles em que são dirigidos grandes investimentos em publicidade, se a empresa não conquistar a lealdade dos seus clientes e os retém, corre o risco de que algum dos seus concorrentes consiga assedia-los de forma a que o cliente troque de *fornecedor*.

Para os consumidores, a vasta gama de opções que a maior parte dos mercados oferece dota a capacidade de escolha de uma liberdade e facilidade muito mais feroz. Sendo assim, a facilidade com que o consumidor muda de fornecedor é uma premissa de extrema relevância nos mercados actuais. Desta feita, caberá ao fornecedor o delineamento de estratégias específicas para atingir níveis de diferenciação maiores por forma a gerir de forma eficaz a sua base de clientes, o que sob a óptica do consumidor lhe permite a exigência de cada vez mais qualidade nos serviços que adquire.

O esforço das empresas no sentido de reduzir ou eliminar o *churn* centra-se na questão de evitar que o consumidor tome a decisão de abandonar a empresa, e não apenas no desenvolvimento de políticas de retenção paliativas.

Para evitar o abandono de um cliente, Engel, Blackwell e Miniard (2000), e em estudos aplicados às telecomunicações, mas que facilmente se aplicam a esta dissertação, defendem que é necessário um “sistema de aviso antecipado” que identifique os problemas em tempo suficiente para que sejam tomadas medidas correctivas, dado que uma vez que o cliente abandone a empresa, é já demasiado tarde para que tais medidas tenham sucesso.

Neste contexto, a **gestão de churn**⁵ aborda o desenvolvimento de técnicas capazes de medir e administrar a taxa de abandono dos clientes (taxa de *churn*), para que os clientes sejam retidos antes de o abandono acontecer, e uma vez que o abandono ocorra, seja possível determinar as razões pelas quais aconteceu.

Clientes Activos⁴ - Clientes que estão na base de dados da empresa por fazerem ou terem feito algum tipo de transacção com a mesma durante um determinado período P.

Gestão de Churn⁵ - *Churn Management*

Ikeda (2006), relembra ainda que a **gestão de churn** deve contemplar preocupações relacionadas com o tipo de cliente que tende a abandonar a empresa e ainda se vale ou não a pena investir esforços para manter alguns clientes. Assim a empresa tomará atitudes não só reactivas, mas também pró-activas no sentido de identificar os clientes mais lucrativos e gerir planos especiais que garantam a sua fidelidade, evitando assim a evasão dos clientes para a concorrência.

Para Berry e Linoff (2000), a **gestão de churn** pode ter vários objectivos:

- 1) Gerar valor para a empresa através da construção da lista de prováveis *Churners*⁶ para definição estratégica de acções de marketing dirigido⁷;
- 2) Construção de um modelo com aplicabilidade na mensuração da eficiência das acções de marketing dirigido.

Cobrindo estes objectivos, o estudo das taxas de *churn* de clientes, pode ser utilizado em acções de marketing que vão desde o desenvolvimento da base de clientes até programas de retenção dos mesmos. Permitindo ainda a monitorização, melhoria e correcção dos programas de fidelização das empresas.

2.6.2. Causas do Churn

Para autores como Cister (2005), existem três tipos de *churn*:

- 1) **Involuntário** – quando é resultado de uma acção da própria empresa, que por motivos de fraude, falta de pagamentos ou até mesmo falta de utilização do serviço, interrompe a sua relação com o cliente (cancelamento do cartão de fidelização por uso indevido, quando contemplado nas regras do programa, por exemplo);

*Churners*⁶ - Clientes propensos a abandonar a empresa

Marketing Dirigido⁷ - Provém do Marketing Relacional, e é definido por um conjunto de acções direccionadas a um cliente ou segmento de clientes.

- 2) **Voluntário** – ocorre quando o cliente pela sua própria vontade termina a sua relação com a empresa, seduzido por campanhas de marketing, promoções, qualidade dos produtos ou mesmo alargamento de gama de um concorrente;
- 3) **Inevitável (ou accidental)** – quando o *churn* ocorre por intermédio do cliente, mas por razões que fogem ao seu controlo, como por exemplo problemas financeiros, ou alteração de morada para localidades onde o fornecedor não tenha representação.

Para uma boa gestão do *churn* é essencial para as empresas reconhecer nos clientes, quais são os *Churners* voluntários, accidentais ou involuntários. O cliente *churner* involuntário, e pelas características explanadas acima, não faz parte das acções de retenção de clientes, uma vez que é a própria empresa que “expulsa” o cliente. Quanto ao *churn* accidental, dificilmente as empresas conseguem implementar algum tipo de acção capaz de evitar o abandono dos clientes, uma vez que as causas do *churn* têm características particulares e até inevitáveis. Posto isto, o foco de qualquer estratégia de retenção de clientes actuam sempre sobre consumidores do tipo *churner* voluntário. Este tipo de consumidor além causar os maiores prejuízos, é aquele que é passível de ser identificado através das suas características numa base de dados como sendo um possível *churner* ou não.

2.6.3. Consequências e custos derivados

Ao contrário do que acontecia nas primeiras aplicações deste tipo de *ferramentas de gestão*, em que a cultura empresarial adoptava a técnica de “conquistar o cliente a qualquer preço”, actualmente as empresas fomentam a estratégia de “reter o cliente” e “conquistar o bom cliente”. A conquista de um cliente com elevada rotatividade e com propensão compulsiva ao *churn* pode revelar-se custoso para a empresa, sendo nesses casos preferível permitir o abandono.

Não obstante o facto de o *churn* ser uma parte inevitável nos mercados de distribuição actuais, esta experiência podem advir consequências que venham a mostrar-se bastante caras. A perda de um grande número de clientes lucrativos para a empresa representará

obviamente uma perda na receita. No entanto, esta perda é apenas o factor mais impactante na demonstração de resultados. Outra característica que pode também ser desastrosa a nível de benefício económico, é a ocorrência de constantes baixas de preço, aquando de um esforço mal informado de tentar reduzir o abandono de clientes para a concorrência, demonstrando assim competitividade de preço.

É ainda um facto que por mais que a empresa invista os seus esforços num sentido oposto, é inevitável que mais cedo ou mais tarde venha a perder alguns dos consumidores para um seu concorrente. Nestas situações, e para que a base de clientes se mantenha numa ordem de grandeza constante ou crescente, as empresas investem muitas vezes em acções de marketing com elevados custos para reaver esses mesmos clientes.

Na verdade, o *churn* não é só um problema, mas sim o sintoma de um problema subjacente muito maior. Quando as empresas não têm habilidade de lidar e gerir o crítico processo de gestão de clientes, perdem a possibilidade de saber quem são os seus clientes, o que estes necessitam e querem e o porquê de os clientes ficarem ou saírem da empresa.

Vários estudos comprovam que o custo para manter um cliente é muito menor do que o custo gerado pela conquista de um novo. Assim, as empresas com vista a evitar o *churn*, utilizam então ferramentas de Data Mining e estatística multivariada que facultam uma análise aos bancos de dados com informação histórica das transacções do cliente com a empresa, permitindo assim a identificação de clientes leais, propensos ao *churn* e quais são efectivamente os clientes de elevado valor para a empresa (Cister, 2005).

2.7. A filosofia de Marketing Cliente na Distribuição

A incessante necessidade de criação de valor para o cliente nos mercados actuais, e a complexidade e elevados custos inerentes à distribuição, desafiam as organizações a encontrar uma forma de reduzir as distâncias entre o processo produtivo e o consumidor

final. Se por um lado as tecnologias de venda avançam a um ritmo acelerado, por outro lado a gestão dos canais de distribuição dificultam a redução dos custos operacionais. Esta questão constitui um problema de suporte financeiro complicado, o que por vezes leva a que as organizações optem por despende de parte do seu orçamento na raiz do processo de distribuição em vez de investirem na criação de valor para os seus clientes.

A falta de adaptação dos fornecedores à complexidade das necessidades actuais das cadeias de distribuição é na maioria das vezes o que dificulta o atendimento às necessidades dos clientes. Entender o processo de distribuição e logística requer um conhecimento aprofundado de suas fases evolutivas, além das tendências que se desenham num cenário de alta competitividade.

A distribuição e a logística devem ser compreendidas inicialmente pela identificação e análise de seus protagonistas: o fornecedor, o produtor, o distribuidor e o consumidor. A estrutura da cadeia é composta por esses elos, que permitem uma melhor administração desde que haja um alinhamento às estratégias e procura.

O *fornecedor*: elemento responsável pelo fornecimento da matéria-prima adequada ao processo de produção e às necessidades do cliente;

O *produtor*: elemento que se compromete a criar bens adequados ao uso e ao consumo e sempre disponíveis nos pontos de venda;

O *canal*: elemento que se dedica ao fluxo de mercadorias sem bloqueios, nem estratégias que dificultem a aquisição e satisfação dos clientes;

O *consumidor*: elemento chave no processo, pois é a partir dele que se concebe o sistema logístico. Enquanto os outros elementos desenvolvem suas estratégias, o consumidor desenvolve suas preferências baseado na forma como é tratado pela cadeia de valor.

Ao percorrer o caminho que leva o produto ao destino podem identificar-se três fases que irão determinar o seu grau de sucesso no mercado. A primeira tem o foco no cliente e nas suas necessidades quanto ao nível dos serviços e a qualidade do produto procurado. A segunda fase é compreendida entre a compra da matéria-prima e produto acabado, num processamento que prevê a adequação e os benefícios esperados pelo

consumidor. Na terceira fase tem-se o fluxo que permite uma distribuição ágil e eficaz, onde a transferência do produto ao consumidor deve acontecer naturalmente, sem que ele perceba ou questione qualquer falha.

2.7.1. A distribuição e logística em Portugal

Todas as mudanças que têm ocorrido no panorama Nacional e Internacional têm origem no processo de saturação dos mercados e mais na transformação incessante do comportamento de compra dos consumidores. Estes possuem hoje padrões muito mais sofisticados e mais exigentes, cobrando muitas vezes das empresas uma postura mais próxima e conhecedora das suas carências. Neste sentido, desenvolveu-se a gestão do marketing de distribuição e logística e passou a haver uma maior concentração nas políticas de gestão de todo o processo.

O marketing da cadeia de distribuição e logística tem a tarefa de traduzir a voz do mercado e encontrar mecanismos que possibilitem aumentar o valor do produto para o cliente, ampliando os benefícios percebidos. A ordem é maximizar a relação de forma que se possa garantir a satisfação dos consumidores. Também na relação que envolve duas empresas, onde uma é fornecedora e a outra cliente, pode se garantir maior valor para o cliente com um alto desempenho da logística, criando uma capacitação em cadeia, com stocks reduzidos e custos de pedidos mais baixos.

Em qualquer empresa produtora, onde se pretenda criar mecanismos de competição mais modernos e eficazes, é fundamental que seja elaborada uma proposta de cadeia de valor, que é seguramente uma declaração de como, onde e quando o valor será criado para clientes ou segmentos de mercado específicos. Esta proposta deve fazer parte dos princípios em que se baseiam todas as actividades da empresa, do desenvolvimento do produto à execução do pedido. Ao formalizá-la é preciso estar certo de que ela fornecerá uma base consistente para a obtenção da vantagem competitiva sem esquecer que os resultados positivos só terão possibilidade de aparecer se ela for definida pela óptica do cliente.

O consumidor realiza a compra quando sabe da importância que um determinado produto tem para ele. A organização deve estar sempre na busca de novas estratégias que possam traduzir a posse desse produto ou serviço em valor agregado, e que consigam mostrar que os benefícios da compra por um dos canais disponibilizados da empresa suplantam o preço pago. Isto desde que percebido pelo consumidor, cria o conceito de valor para o cliente.

Muitas vezes, os atritos gerados entre as áreas de distribuição e de logística acabam por prejudicar o consumidor final, quando não o faz migrar para a concorrência. Enquanto a distribuição produz uma cadeia cada vez mais extensa e custosa, engessando a flexibilidade e a rapidez nas respostas ao mercado, a logística preocupa-se em minimizar o impacto do preço final através de propostas de eliminação de etapas para reduzir o sistema de stock total, custos de armazenamento e tempos de operação.

A logística hoje é, sobretudo, serviço e tornou-se um factor fundamental para a conquista e retenção do cliente. Esse cliente exige cada vez mais níveis de desempenho dos fornecedores elevados. É necessário que a entrega seja ágil, rápida, dentro das novas expectativas do tempo de resposta concebível. A tolerância do consumidor em esperar respostas eficientes é cada vez mais reduzida.

Com isso, a tarefa da cadeia de valor é identificar as exigências do segmento em que actua e fazer uma constante reestruturação do processo de logística e distribuição. Desta forma, quando uma organização resolve desenvolver um sistema integrado de logística e de distribuição, que seja flexível e dinâmico, objectivando sempre a criação de valor para o cliente, estará sintonizada com o que há de mais moderno em termos de competição mercadológica.

Um caminho ideal é o da parceria entre fornecedores e distribuidores, utilizando a tecnologia da informação e a previsão de vendas. A indústria deve estar interligada com os seus fornecedores, distribuidores e pontos de venda. Deve estar atenta à procura, aos novos desejos e necessidades de quem revende e de quem consome seus produtos. A ordem é substituir stocks por informações, chegar mais rápido ao consumidor, eliminar ao máximo os intermediários; reagir positivamente aos movimentos e demandas do

mercado; agregar valor a cada etapa da cadeia, sem aumentar preços; seleccionar bem os canais; ou seja **satisfazer o cliente**.

2.8. A Sonae MC e a Fidelização

Para qualquer empresa da actualidade a noção de onde reside a sua verdadeira riqueza, há muito que deixou de ser um segredo – os seus clientes. Estes são quem sustenta a operação da própria empresa, e em última análise, são a própria razão de existência da empresa.

Assim, e dada a constante mutação dos mercados, o foco de preocupação das empresas há muito que deixou de estar no produto que vende passando a ser direccionado para o mercado e para o cliente. Neste contexto, o marketing de massas foi-se tornando em marketing direccionado estando cada vez mais focalizado para uma vertente mais individualizada a cada cliente (marketing one-to-one).

Num mercado cada vez mais competitivo, as empresas foram a pouco e pouco centrando as suas estratégias de gestão na busca por uma vantagem diferenciadora face aos seus concorrentes. Neste âmbito passaram a ser utilizadas ferramentas de transformação de grandes quantidades de dados em retratos fies dos clientes → Data Mining.

Este tipo de analogias é transversal à maior parte dos sectores de actividade das empresas. No entanto, e mais especificamente sobre o Retalho têm vindo a ser desenvolvidos inúmeros estudos e investigações de forma a transformar o que outrora era apenas uma grande quantidade de informação em conhecimento para o apoio a tomadas de decisão.

As ferramentas de fidelização têm objectivos claros e que se centram na diferenciação em relação a empresas concorrentes, evitando a perda de clientes para essas mesmas empresas do mesmo sector, com uma série de medidas que contribuam com um valor

acrescentado aos clientes. É desta forma que um qualquer acto de consumo identificado por parte de um cliente, se transforma numa oportunidade única para premiar a eleição do consumidor e a confiança na marca que escolhe para efectuar as suas compras. De uma forma geral, pode dizer-se que quaisquer acções de satisfação para com o cliente são os primordiais fios condutores para que este tenha menos motivação de optar por uma outra marca, tendo ao mesmo tempo um maior incentivo para a continuidade de compra de produtos da empresa.

O mecanismo de desenvolvimento de um plano de fidelização requer o entendimento de passos importantes, para que os objectivos sejam alcançados de forma completa. Os passos principais a ter em conta neste mecanismo são:

Identificação dos Clientes: É necessário que a empresa conheça o cliente para que se possa estabelecer uma relação com ele. Os dados do cliente são meios essenciais para que a comunicação possa ser estabelecida entre as duas entidades.

Armazenamento da informação sobre o cliente: Para que a informação do cliente esteja disponível para consulta por parte da organização de forma a sustentar a análise e a tomada de decisão, é necessário que sejam desenvolvidos processos capazes de recolher a informação.

Definição de métricas de avaliação da fidelidade dos clientes: É importante que a empresa seja capaz de avaliar o valor e grau de fidelidade de cada cliente.

Qualificar clientes: Um dos principais objectivos de um plano de fidelização é a diferenciação dos clientes. Cada cliente tem as suas necessidades e comportamento e estas características reflectem-se na importância do cliente para a empresa. É por isso necessário que os clientes sejam classificados quanto ao seu valor, com a finalidade de desenvolver estratégias eficazes de fidelização. Desta forma, pode deduzir-se quais os mais interessantes e fidelizáveis.

Criação de Valor para o cliente: Esta meta é atingida através da criação de ofertas com valor para o cliente. Estas ofertas devem ser planeadas e analisadas através da informação que é recolhida em cada interacção da empresa com o cliente.

Uma das formas mais adoptadas pelas empresas para fidelizar os clientes é o cartão de fidelização. Este oferece por norma bonificações e descontos aos clientes e em troca, o cliente fornece às empresas o conhecimento dos seus hábitos de consumo permitindo assim que estas definam de forma sustentada a sua política comercial e de Marketing.

2.8.1. Fidelização dos Clientes na Sonae MC

Neste contexto, a Sonae MC⁸, que desde a sua implementação se tentou destacar dos seus concorrentes, lançou a 23 de Janeiro de 2007 o seu Cartão de Fidelização com o principal objectivo de gerar um valor diferenciado quer para a companhia quer para o próprio cliente. Com esta nova ferramenta de gestão a Modelo Continente pôde de forma mais específica premiar os seus melhores clientes, e gerir a sua carteira de clientes de modo a garantir a manutenção da mesma e a angariação de novos clientes.

Apesar de um histórico de mais de 54 meses de informação de cliente, e tendo em grande parte ultrapassado todos os passos principais no desenvolvimento do seu plano de Fidelização, existem ainda algumas preocupações no que diz respeito a análise da base de clientes. Sendo uma empresa ainda muito jovem no que diz respeito à exploração de dados deste nível, existem ainda lacunas nos sistemas até agora explorados e que são merecedoras de atenção.

Uma área que em termos de negócio é justificável de algum cuidado é, e por todos os motivos já descritos anteriormente é a manutenção da base de clientes. O desenvolvimento de mecanismos de detecção atempada de abandono de clientes é neste momento um factor crítico para a companhia.

Com o crescimento exponencial da concorrência e com a facilidade de acesso que cada vez mais os clientes têm quer sobre informação concorrencial, quer de acessibilidade longitudinal, é cada vez mais fulcral a análise sobre a base de clientes de forma a promover a retenção de clientes à Marca.

⁸ Sonae MC – Sonae Modelo Continente. Enquadramento da Empresa e breve abordagem à Distribuição em Portugal em Anexo I.

O abandono dos clientes – churn - é um dos principais problemas com que se depara qualquer empresa e a Sonae MC não é excepção. No entanto este processo não é repentino. Salvo algumas excepções que se prendem com experiências desagradáveis num determinado contacto com as lojas ou por exemplo alteração da residência para locais onde a empresa não se encontra representada, o churn acontece progressivamente e é por norma provocado pelo descontentamento face à própria empresa ou em detrimento de uma outra, originando a descredibilidade e abandono. Até a esta fase final cabe à empresa agir sobre o cliente de forma a prevenir a evidente prejudicial perda do cliente.

O estabelecimento de planos e estratégias de retenção dos clientes não se baseiam apenas no conhecimento de quando em que o cliente vai abandonar, mas também, e o mais importante, saber o porquê. Se os motivos que despoletam o abandono não forem conhecidos, não se saberá como reter o cliente. É neste contexto que a “causa” é um elemento de extrema importância para os responsáveis de Marketing e Comerciais, pois só desta forma lhes é possível actuar adequadamente.

Apesar dos inúmeros estudos elaborados nesta área específica, e não obstante a oportunidade inquestionável da exploração deste problema para a Sonae MC, sob este mesmo problema de Churning optou-se por uma abordagem um pouco mais abrangente.

Dadas as condicionantes do mercado actual como: proximidade dos concorrentes; comunicação exaustiva e comparativa de preços por parte dos Retalhistas; diversidade na oferta de produtos diferenciada de concorrente para concorrente, entre outros, sabe-se que grande parte dos consumidores não tem por norma apenas uma loja ou marca de destino para a compra da totalidade do seu cabaz de compras. É cada vez mais comum um cliente efectuar as suas compras de Mercearia numa determinada loja e as compras relativas a bens de consumo Perecível numa outra loja ou Marca concorrente. Além desta questão, surge ainda a experiência de compra dos consumidores. O perfil de compra, nomeadamente da tipologia de artigos que adquire ou deixa de adquirir em dada loja pode ser um factor indicador do que o cliente vai fazer a seguir.

Como se sabe, tipologias de produtos diferentes sustentam rentabilidade diferente. Desta forma torna-se um factor importante a determinação da captação (ou abandono)

dos clientes ao maior número de áreas de negócio possíveis, de forma a fidelizar o cliente mais assertivamente.

A identificação de áreas fracas para um cliente ou conjunto de clientes mostra-se de extrema importância, uma vez que como se sabe quanto maior for o domínio da marca sobre as compras do cliente, maior será o valor que esse mesmo cliente gera para a empresa. Existe ainda inerente a este ponto questões como a compra cruzada (ou cross-selling) que indicam que o facto de um cliente comprar determinado tipo de produtos é também propenso a comprar um outro. Por exemplo, um cliente com compras regulares em artigos de carisma perecível, pode revelar-se um cliente que tem grande confiança na marca e que dificilmente, e a menos que recaia sobre uma das excepções citadas, abandona a empresa como marca de eleição.

Posto isto, o domínio do conhecimento do comportamento do cliente em termos de abandono não só global como a nível comercial, poderá constituir uma ferramenta poderosa de gestão e retenção da base de clientes.

Derivado do elevado ambiente concorrencial, e mesmo adoptando uma política de mercado generalista, torna-se cada vez mais difícil fidelizar a 100% um cliente. Factores como a distância geográfica ou mesmo como a concorrência urbana, podem levar a que um cliente por uma questão de comodidade se desloque de forma menos ou mais recorrente a determinada loja.

Posto isto, e não obstante a importância da definição de um modelo global de *Churn*, a diferente caracterização dos clientes em termos de frequência de compra, despoleta o interesse na criação de modelos distintos para os diferentes tipos de clientes. Actuar sobre um cliente Leal, quando este não efectua compras nas lojas há mais de um mês pode ser prioritário, enquanto que por outro lado se trata de um cliente Ocasional, o mesmo período de tempo pode não ser relevante.

3. Inteligência de processos

Neste capítulo serão apresentados os conceitos básicos que constituem a descoberta de conhecimento em bases de dados (Data Mining) dando especial ênfase às suas aplicações em CRM.

Serão ainda abordadas as principais e mais comuns técnicas e modelos de Data Mining aplicadas em CRM, que serão também utilizadas na construção dos modelos preditivos deste trabalho. Apesar de não ser intuito a utilização de todos os métodos e técnicas, optou-se neste capítulo por realizar um levantamento mais exaustivo e exploratório no que diz respeito às opções de trabalho experimental posterior.

Numa secção final será ainda explorado sucintamente o tema do Churn no que respeita a estudos existentes em diferentes sectores do mercado Internacional.

3.1. Introdução ao Data Mining

Data Mining é a terminologia que ficou popularizada como denominação do processo de descoberta de conhecimento em bases de dados. Esta “ciência” utiliza ferramentas computacionais com o fim de identificar informação que seja potencialmente útil, descrita de forma padronizada, partindo dos grandes volumes de dados que são armazenados actualmente pelas empresas. A identificação desse conhecimento implícito nos padrões descobertos tem sido cada vez mais útil no conhecimento dos consumidores e na tomada de decisões por parte das organizações.

Segundo FAYYAD et al. (1996), este conceito pode ser resumido como “o processo não-trivial de identificar padrões novos, válidos, potencialmente úteis e, principalmente, compreensíveis por meio da observação dos dados contidos numa base de dados”. Ou seja, o objectivo primordial do Data Mining não passa apenas por encontrar padrões e relações da informação, mas sim extrair conhecimento inteligível e utilizável de forma mais ou menos imediata no processo de tomada de decisões.

Para o mesmo autor, o crescente interesse neste tema deve-se à necessidade das organizações que armazenam de forma exponencial grandes quantidades de dados, não havendo por parte dos métodos de análise tradicionais capacidade de leitura de informação e conteúdos relevantes.

Desde a chamada de atenção por parte de FAYYAD et al. (1996) para a falta de ferramentas de Data Mining bem desenvolvidas, estas têm apresentado uma evolução quase que vertiginosa quanto à robustez e versatilidade. No entanto, o conhecimento do especialista é ainda um factor de elevada relevância quando as empresas procuram este tipo de soluções. Factores como a compreensão do negócio, a origem dos dados e a própria diversidade dos mesmos são críticos e criam um esforço inicial premente na consolidação da informação que servirá de base para o processo de Data Mining, bem como serão cruciais para a leitura dos resultados.

3.1.1. “Descendência” do Data Mining

O Data Mining descende fundamentalmente de três linhagens:

- 1) Estatística Clássica: que envolve conceitos que permitem estudar os dados e os relacionamentos entre os mesmos. É nos elementos fundamentais como as distribuições probabilísticas, medidas de dispersão e outros, que as análises estatísticas mais avançadas se apoiam sendo portanto a base da maioria das tecnologias a partir das quais o DM é construído.
- 2) Inteligência Artificial (AI): esta segunda linhagem do DM é construída a partir de fundamentos heurísticos, que, ao contrário da Estatística Clássica, tenta “imitar” o pensamento humano na resolução dos problemas estatísticos.
- 3) Machine Learning (ML): pode ser descrita como a junção de melhores práticas de Estatística Clássica e Inteligência Artificial. Baseia-se essencialmente no pressuposto de fazer com que as ferramentas computacionais “aprendam” com os dados que “estudam”, de forma a que essas mesmas ferramentas tomem diferentes decisões baseadas nas características dos dados estudados, utilizando a Estatística Clássica para os conceitos fundamentais, adicionando a componente

heurística avançada da Inteligência Artificial e algoritmos para alcançar os objectivos.

Desta forma, podemos definir o Data Mining como a união da evolução histórica e recente da Estatística Clássica em AI e ML, técnicas que são utilizadas conjuntamente para avaliar os dados e encontrar tendências e padrões nos mesmos.

3.2. Business Intelligence

O termo *business intelligence* é utilizado como referência às técnicas e ferramentas que fornecem suporte às decisões de negócio baseadas em tecnologia de Informação, no qual o Data Mining assume, actualmente um papel preponderante (Figura 2).



Figura 2 - Processo de Business Intelligence

Por norma, o valor que a informação tem no suporte à tomada de decisão, cresce da base para o topo da pirâmide (Figura 2). Ao contrário das decisões tomadas com base em dados sumarizados e tratados (camadas superiores), uma decisão que seja baseada em dados de camadas inferiores, onde tipicamente existem milhões de registos, ainda

poluídos (nas camadas mais básicas), poderão estar providos de ruído levando muitas vezes, e por esse motivo a decisões precipitadas que poderão englobar por exemplo apenas uma transacção não identificativa do portfolio de clientes.

3.2.1. Dados e DataWareHousing

Para que qualquer conhecimento seja gerado a partir de informação, o primeiro requisito é que essa informação exista sob a forma de dados e que os mesmos estejam disponíveis. Dada a constante necessidade de recorrer a informação organizada, actualmente as empresas recorrem a “grandes” infra-estruturas de tecnologia para garantir o correcto armazenamento e a disponibilidade dos dados para servir ao apoio na tomada de decisões.

A forma mais comum de organização, consolidação e disponibilização de dados são denominados de Data Warehouses, que são grandes repositórios de bases de dados alimentados por sistemas operacionais. Permitem às organizações uma visão de dados e operações num único ponto, garantindo a integridade e consistência de todos os dados armazenados. Por norma, aquando do encaminhamento dos dados para um Data Warehouse (DW), os mesmos são limpos e transformados de forma a garantir que dados de diferentes fontes e formatos passem a possuir as mesmas definições.

Os Data Warehouses dispõem da possibilidade de extracção (SQL⁹), tratamento e agregação de múltiplos sistemas operacionais em *data marts* ou múltiplos DW. Armazenam cada vez mais frequentemente dados em formato multidimensional (cubo OLAP¹⁰).

A utilização da visão multidimensional dos dados por parte da metodologia OLAP, permite que sejam efectuadas consultas analíticas complexas em pequenos intervalos de tempo de execução.

⁹ SQL – Structured Query Language: método utilizado par a execução de consultas sobre um DW.

¹⁰ OLAP – Online Analytical Processing

A formação e configuração de um Data Warehouse não é uma tarefa de todo trivial. Desta forma, as organizações optam muitas vezes pela formação de *data marts*, que são conjuntos de dados significativamente mais pequenos que um DW, e por isso mais facilmente acessíveis. Estes *data marts* funcionam muitas vezes orientados para um problema ou departamento específico dentro de uma organização ou empresa, potenciando assim uma maior flexibilidade de análise e um processamento mais especializado.

3.3. O Processo KDD (Knowledge Discovery in Databases)

Para que o objective primordial do Data Mining seja atingível, isto é, para que a extracção de conhecimento inteligível e utilizável de forma a apoiar o processo de decisão, é necessário ter em consideração todo um conjunto de pressupostos não triviais, no que respeita à organização e tratamento dos dados que servirão de origem a esse conhecimento.

É então crucial, e dependendo da diversidade e heterogeneidade dos dados recolhidos, concentrarem-se esforços de pré-processamento e limpeza da informação original, para a geração de dados passíveis de serem utilizados pelas técnicas e modelos de DM. Em primeira instância é essencial que seja realizada a investigação de inconsistências devidas, por exemplo, a problemas de escalas, ao tratamento de valores *anormais* (outliers) e à limpeza de observações erróneas. Só após esta primeira fase de homogeneização dos dados, o processo de DM pode ser iniciado. Por fim, a interpretação, compreensão e aplicação dos resultados encontrados é o passo que torna o conhecimento adquirido num real input para o apoio à tomada de decisões.

Para CIOS (2000), existem diferentes tipos de Data Mining. Sendo que estes dependem da generalização do problema e das expectativas relacionadas com o tratamento do mesmo. Estes níveis são:

- 1) **DM não direccionado ou Puro:** aplica-se a um cenário mais genérico, em que o analista por norma procura simplesmente “informação relevante” para o negócio. Neste tipo de DM, não existem restrições de sistema, no entanto não são também indicadas ao pesquisador directrizes sobre o que esperar. Deste tipo de abordagem incorre por exemplo o risco de que os padrões encontrados possam ser já conhecidos, não sendo por isso traduzíveis em valor acrescentado;
- 2) **DM direccionado:** esta é uma abordagem mais específica, em que é indicado ao analista que tipos de perfis serão interessantes para a organização;
- 3) **Testes de Hipóteses e refinamento:** neste caso, a pesquisa torna-se ainda mais específica, sendo que o analista é previamente provido de algumas hipóteses, sendo expectável a validação das mesmas, e no caso de insucesso neste objectivo deverão ser formuladas e testadas novas hipóteses no mesmo contexto.

Como foi referido anteriormente, o processo de descoberta de conhecimento numa base de dados, é um processo tudo menos trivial, sendo composto por um conjunto de fases actualmente já bem providas de técnicas e métodos.

Ocorrendo apenas algumas diferenças em terminologias ou descrições dessas etapas, autores como FAYYAD et al. (1996), CIOU (2000), e KLOSGEN et al. (2002), definem de forma semelhante o ciclo de Descoberta de Conhecimento em Bases de Dados (ou ciclo KDD¹¹).

A estrutura e formação do ciclo de KDD proposta por FAYYAD e tal. (1996) está ilustrada na Figura 3.

¹¹ KDD – Knowledge Discovery in Databases

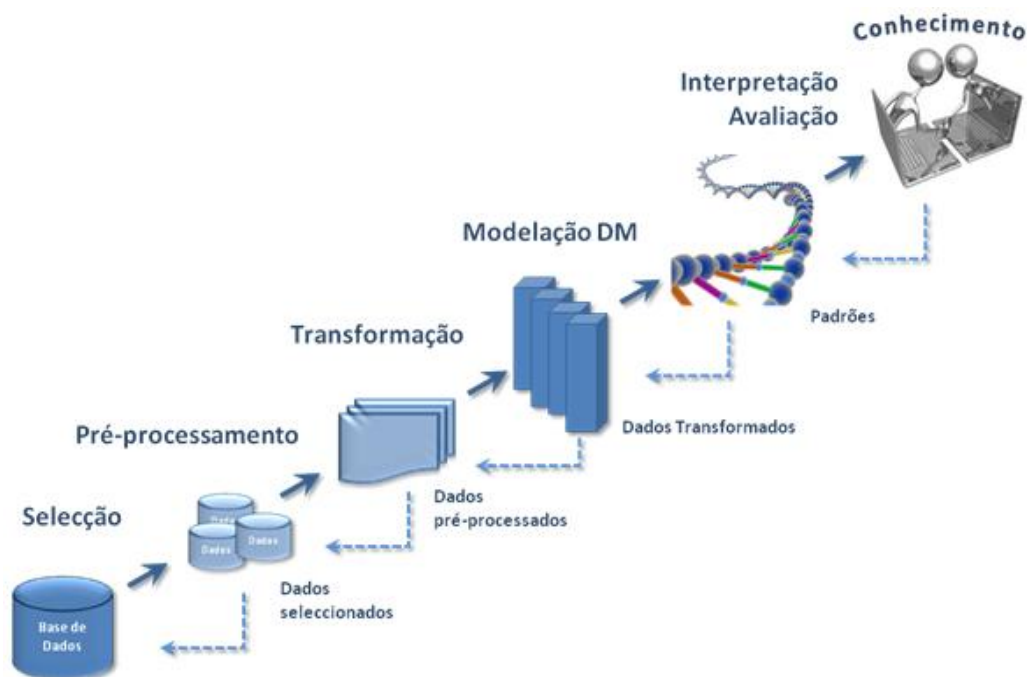


Figura 3 - Ciclo KDD (by FAYYAD et al. 1996)

Na estrutura deste trabalho será utilizada uma composição mista do ciclo de KDD, sendo referenciado a proposta de FAYYAD et. al (1996), agregando itens do modelo de KLOSGEN et. al (2002), sendo que este ultimo se diferencia do modelo proposto pelo primeiro autor mais no que respeita à apresentação e organização do que propriamente no conteúdo e método.

3.3.1. Definição e análise do problema

Os problemas relacionados com Marketing encontram-se nos lugares de topo da lista das diversas aplicações do processo de KDD. Enigmas relacionados com os perfis de clientes e produtos, a predição da lealdade e retenção de clientes e a avaliação de campanhas promocionais, ou até mesmo a predição de clientes alvo a campanhas direccionadas, são alguns dos interesses típicos por parte das organizações.

A saturação dos mercados actuais, o forte ambiente concorrencial e a maturidade dos produtos e serviços levaram a que as organizações transpusessem as suas preocupações

no que respeita apenas à qualidade, para a competição por informação detalhada do comportamento dos consumidores e utilização da mesma como factor diferenciador.

3.3.2. Pré-processamento e compreensão dos dados

As metodologias aplicadas no KDD podem prover as organizações de conhecimento utilizável, só e apenas se os dados que gerem esse mesmo conhecimento estiverem disponíveis.

Segundo KLOSGEN et. al (2002), é necessário que existam dados relevantes para a resolução do problema, e ainda que a generalização dos mesmos possa ser aplicada em questões objectivas. Será então necessário para cada tabela fonte de dados decidir quais os atributos mais relevantes ao problema. Se, por exemplo, for requisito a predição de um determinado atributo A (Cliente abandona ou não abandona a empresa?), esse mesmo atributo deverá constar do banco de dados a trabalhar, como também serão necessários atributos que em primeira instância possam conduzir à predição do atributo A. Para o mesmo autor, os dados mesmo que provenientes de diferentes fontes, deverão ser agrupados num banco único de dados (*dataset*). O modelo deve ainda garantir que o volume de dados tenha alcance suficiente sobre a variedade de resultados das variáveis relevantes, isto é deve ser um conjunto de dados homogéneo em termos de qualidade, e heterogéneo em termos de quantidade.

Após a definição do problema, os dados relevantes para o estudo do mesmo, precisam ser extraídos. Na maioria dos casos, este conjunto de informação é extraído de bases de dados operacionais ou de um Data Warehouse.

Em alguns casos, a melhor estratégia pode passar pela formação de uma base de dados, ou de um Data Mart específico. Por exemplo, no problema de Detecção de Churn (estudado neste trabalho), dados como os contactos de Call Center com um cliente, podem ser preponderantes para o modelo, e o feedback desses contactos poderá não estar armazenado na base de dados transaccional, onde constam os contactos *front-office* do cliente com a organização. Em casos como estes, em que variáveis que possam

ser relevantes ao problema não constem na mesma fonte de dados, poderá, e dependendo dos custos associados a esta operação, ser necessário a construção de um novo esquema relacional de forma a que o modelo aplicado seja provido de toda a informação necessária.

A etapa do processo de KDD de pré-processamento dos dados é frequentemente interpretada como uma fase que envolve um elevado conhecimento e domínio. Por norma, os dados extraídos das bases de dados operacionais possuem informações imprecisas, além de ser providos de uma grande quantidade de *outliers* oriundos de erros de armazenamento ou erros de operação. Se bem que muitos algoritmos de DM foram desenvolvidos para manipular os dados nestas condições, é de esperar que os mesmos gerem resultados mais precisos se a informação for de qualidade, ou seja, se estes problemas nos dados forem previamente tratados e removidos. Em alguns casos específicos, e dependendo da natureza da informação, é muitas vezes necessário recorrer a um especialista no negócio nesta fase, uma vez que a correcção deste tipo de anomalias requerem um conhecimento mais especialista.

É então principal objectivo desta fase validar e verificar a consistência da informação, correcção de possíveis erros e preenchimento ou eliminação de valores nulos ou redundantes. Nesta etapa são identificados e removidos os dados duplicados e corrompidos.

De uma forma geral, o trabalho de pré-processamento de dados é um processo semi-automático. Desta forma, depende da capacidade de identificação dos problemas nos dados por parte do analista, e similar capacidade de utilização dos métodos mais apropriados a resolução dos mesmos.

As tarefas associadas a esta fase não são, no entanto, todas igualmente dependentes do conhecimento de domínio. Isto é, existem tarefas que podem ser desempenhadas automaticamente por métodos específicos que extraem dos dados as informações necessárias para o tratar o problema (Métodos de tratamento de *Missing Values*¹², e Detecção de *Outliers*). E existem outras que dependem de que conhecimento específico seja fornecido como as tarefas de verificação de Integridade ou de selecção de atributos.

¹² *Missing Values* – valores desconhecidos ou ausentes de um determinado campo (ex. #NA; null).

3.3.2.1. Detecção de Outliers

Um exemplo já enumerado, e que é bastante comum na limpeza dos dados (que ocorre no pré-processamento) é a procura por valores absurdos (*detecção de outliers*), valores esses que à partida não deveriam existir na base de dados por serem impossíveis na prática. Muitas vezes, nesta etapa é necessário recorrer a um especialista de negócio, pois esses que podem a uma primeira análise parecer valores “aberrantes”, podem de facto ser valores importantes no processo, dependendo claramente do objectivo do processo de análise. Um exemplo de valores que devem ser limpos da base de dados, acontece recorrentemente na base que será explorada neste trabalho. Quando, por exemplo está em causa o processo de digitalização de um cupão de adesão, podem ocorrer frequentemente erros de digitalização ou mesmo de reconhecimento de caligrafia, produzindo erros como por exemplo a idade de um cliente. Quando não tratado este campo, pode muitas vezes ser provido de valores como idades negativas ou idades superiores a 200 anos, valores esses que serão obviamente erróneos. No caso de valores díspares que devem ser considerados, pode assumir-se por exemplo o caso da banca no número de utilizações de um determinado cartão de crédito. Se o objectivo do modelo é por exemplo prever o “abandono” de um cliente, observações neste campo, por exemplo com o valor de 1 numa base de dados anual em que o número médio de utilizações é de 60, não deverá ser considerado como um *outlier*, mesmo que o método de tratamento escolhido assim o indique.

Nesta tarefa de limpeza e consistência dos dados, pode optar-se por eliminar o valor que é considerado como equivoco, ou então pode ser substituído por outros valores possíveis, utilizando neste caso medidas de localização como a média ou mediana da variável em questão.

Os processos de DM dispõem de inúmeras técnicas de detecção e eliminação deste tipo de valores. No entanto, e como foi referido anteriormente, em determinado tipo de análises estes valores *raros* ou *estranhos* podem ser eventualmente interessantes e provir a análise de inputs importantes para a compreensão de determinado tipo de comportamentos.

Os *outliers* podem ser detectados com testes estatísticos que assumem um modelo de distribuição ou probabilístico para os dados, ou então através de medidas distância

(valores mais distantes de um determinado conjunto de dados são considerados como *outliers*). Existem ainda métodos que identificam estes valores segundo uma análise de desvios.

Uma breve abordagem a alguns dos métodos de detecção de *outliers* está descrita no Anexo II.

3.3.2.2. Tratamento de Missing Values

O tratamento de valores desconhecidos ou ausentes é um problema comum na etapa de pré-processamento. Podem ser aplicadas inúmeras técnicas, algumas delas bastante simples, como a simples exclusão da instância em que ocorre o valor omissivo, a substituição desses valores pela média, moda ou mediana do atributo em questão. No entanto podem ser aplicadas técnicas mais *inteligentes* baseadas por exemplo em algoritmos preditivos que induzem o valor pelo qual o desconhecido deve ser substituído.

Esta tarefa deve no entanto ser tratada com especial atenção, na medida em que a utilização de técnicas simples, como a substituição de todos os valores desconhecidos por uma única variável global, pode desvirtuar o processo de predição, trazendo resultados indesejados. Por outro lado a utilização de algoritmos de predição para a substituição de valores desconhecidos, traz uma indiscutível complexidade adicional a todo o processo.

A ocorrência de *missing values* deve-se geralmente a erros humanos, à não disponibilidade da informação aquando da entrada dos dados em sistema pelo facto do atributo não ser de preenchimento obrigatório.

Para tratar este tipos de atributos pode recorrer-se à simples exclusão das instâncias onde um *missing value* ocorre, ou então à substituição desses valores por outros que sejam estatisticamente aceitáveis¹³.

¹³ Alguns métodos de tratamento de Missing Values podem ser consultados no Anexo II.

3.3.2.3. Identificação de Inconsistências, ruído, redundância e valores default

As inconsistências ocorrem quando dados diferentes são representados pelo mesmo valor de atributo, ou mesmo quando valores de um atributo diferentes representam o mesmo dado. Um exemplo deste último acontece por exemplo na utilização de siglas e o extenso de Marcas ou nomes de empresas para um mesmo campo. Nestes casos a sigla e o nome por extenso refere a mesma identidade, no entanto está representada nos dados de forma diferente.

Já o ruído nos dados acontece quando valores que supostamente não respeitam a tipologia do atributo são encontrados na base de dados. Um exemplo de ruído nos dados pode por exemplo ser o facto de encontrar um valor diferente de “F” ou “M” no campo *género* de uma base de dados de clientes.

No que diz respeito à redundância e informação, pode citar-se por exemplo o caso de numa base de dados existir informação idêntica armazenada em diferentes atributos. Um exemplo típico de redundância nos dados é a existência do PVP, número de quantidades compradas e o valor da compra numa mesma tabela, quando o que acontece é que estes atributos recorrem uns dos outros, podendo assim ser calculados em tempo real. É necessário ter em atenção este problema na fase de pré-processamento, uma vez que a redundância provoca o aumento no tempo de processamento, particularmente em métodos que sejam mais sensíveis ao número de atributos utilizados.

A informação duplicada pode muitas vezes ser um problema importante a ser trabalhado, nomeadamente em bases de dados de clientes. O não tratamento deste problema pode fazer por exemplo com um determinado que está registado em diferentes tabelas com nomes diferentes, seja tratado pelos algoritmos como clientes diferentes, o que faz com que informação comportamental importante seja perdida no processo.

É ainda de notar a importância da identificação de valores padrão ou *default*. Na maioria das bases de dados, e muitas vezes tendo sido implementado para evitar a existência de *missing values*, existem campos que assumem valores por defeito. É importante que o analista tenha conhecimento das variáveis que irá considerar nestas condições para que possa ter essa consideração na etapa de interpretação de resultados.

3.3.2.4. Selecção de Atributos (*attribute selection*)

Este é um problema importante no processo de *KDD* e consiste na selecção de um subconjunto de atributos da base de dados, no qual o algoritmo de DM aplicado possa ser implementado de forma a acrescentar valor.

Esta selecção deve no entanto ser a mais criteriosa possível, e depende fortemente do problema que está a ser analisado. É um estágio importante não só porque muitos dos algoritmos de DM não têm grande desempenho com uma grande quantidade de atributos, mas também porque o conhecimento que é induzido por esses mesmos algoritmos é mais compreensível na fase de interpretação quando envolve um menor conjunto de variáveis. Para alguns algoritmos como as redes neuronais, a indução do algoritmo pode ser prejudicada aquando da existência de muitos atributos irrelevantes.

Quando a natureza dos dados e do problema o permite, os algoritmos de DM utilizam por norma *datasets* obtidos por amostragem, com um conjunto restrito de exemplos (registos) e numero igualmente restrito de atributos. Os algoritmos de selecção de atributos são frequentemente utilizados para a identificação, através das possíveis combinações, os atributos que melhor representam os dados a analisar.

O objectivo destas técnicas é portanto, optimizar a relação de informação entre as variáveis de entrada e saída de um modelo, reconhecendo quais as variáveis que desempenham um papel relevante nas saídas e eliminando as menos relevantes. Além da óbvia redução na dimensão da base de dados, melhorando o desempenho dos algoritmos, é de realçar também que a identificação de quais as variáveis mais relevantes ou que exercem maior influência num determinado modelo, provem o analista de informação valiosa sobre o problema em estudo.

A literatura possui um grande número de métodos de *attribute selection*. Algumas dessas abordagens técnicas podem ser encontradas no Anexo II deste trabalho.

3.3.2.5. Construção de atributos

Quando os atributos utilizados para a descrição do conjunto de dados são inadequados, o que acontece recorrentemente é que os algoritmos de DM criam classificadores

imprecisos ou com uma complexidade excessiva. Alguns atributos podem ser individualmente inadequados, no entanto, quando combinados gerando novos, podem revelar-se representativos para a descrição de um determinado conceito.

Desta forma, a construção de atributos é então o processo de composição de atributos ditos *primários*, produzindo novos atributos que possam ser relevantes para a descrição de uma realidade. De uma forma ampla, pode dizer-se que este processo pode ser dividido em duas abordagens, a automática na qual o processo é guiado automaticamente pelo método de construção em que os atributos construídos vão sendo avaliados relativamente aos dados, e uma outra dependente do conhecimento empírico do analista.

3.3.3. Transformação dos dados

O principal objectivo desta tarefa é a de transformar a representação dos dados a fim de superar quaisquer limitações existentes por parte dos algoritmos de DM que serão utilizados, e a decisão de quais as transformações que são necessárias recai essencialmente na escolha do algoritmo que será utilizado. De uma forma geral, a transformação nos dados envolve a aplicação de técnicas sobre as variáveis de forma a maximizar a informação contida nas mesmas, tornando-a assim mais apropriada para a modelação. Contudo, o objectivo da transformação dos dados não se fica pela adequação das variáveis, mas visa também responder às limitações impostas pelos algoritmos de DM. Esta tarefa é comumente utilizada por exemplo para transformar variáveis nominais em variáveis numéricas.

3.3.3.1. Normalização dos dados

Esta técnica de transformação consiste em transformar os valores dos atributos para um intervalo específico e é especialmente valiosa para os métodos que calculam distâncias entre atributos. É realizada principalmente com o objectivo de homogeneizar a variabilidade dos atributos de uma base de dados, criando um intervalo de amplitude

similar para todas as variáveis. Esta técnica é então necessária sempre que existem nos *datasets*, variáveis com unidades diferentes ou dispersões heterogêneas. Apesar de se saber que métodos como as Redes Neurais têm um melhor desempenho quando aplicados sobre conjuntos de dados com uma amplitude de valores dos atributos pequena, a outros como as Árvores de Decisão (e outros que induzem representações simbólicas) a normalização tende a diminuir a compreensibilidade do modelo gerado.

Duas das formas mais comuns e simples de normalização são:

1) Normalização pelo desvio padrão (z-score): —

2) Normalização pela faixa de variação: —

3.3.3.2. Discretização de atributos quantitativos em qualitativos e vice-versa

Muitos algoritmos de DM trabalham sob atributos qualitativos. Essa limitação exige que nos casos dos bancos de dados serem de natureza quantitativa, exista uma transformação de informação. O exemplo mais comum consiste em transformar os valores de um determinado atributo em valores intervalares. Aliada ainda à necessidade desta transformação para que alguns algoritmos possam ser aplicados, tem-se ainda o facto de que a compreensibilidade dos resultados é na maioria dos casos mais simples com atributos discretos.

Por outro lado, e para a aplicação de algoritmos que manipulem apenas variáveis de natureza quantitativa, e existam atributos de ordem qualitativa (ou por exemplo nominal), é necessário proceder à transformação inversa. Esta manipulação de dados, deve no entanto ser criteriosa no sentido em que deve ser tido em conta se o atributo mesmo que qualitativo, segue ou não uma ordem, e nesse caso aquando da transformação, a mesma ordem numérica deve ser levada em conta, mesmo para que os resultados do modelo sejam providos de maior compreensibilidade.

Outro tipo de transformação que pode ser alocada no mesmo grupo destas técnicas tem a ver com a manipulação de dados complexos, como é o caso de datas ou horas, presentes na maioria dos conjuntos de dados, e que são muitas vezes de extrema relevância para o modelo (p.ex: data de nascimento do cliente). Neste caso a

transformação efectuada sobre o atributo transforma estes dados em números inteiros (p.ex: data de nascimento → idade).

3.3.3.3. Redução de dados

Apesar das claras vantagens da utilização de grandes volumes de dados para treino e teste, o conjunto de dados pode no entanto tornar-se excessivamente grande na prática. Além da dimensão elevada dos dados poder exercer um peso negativo no desempenho dos algoritmos, muitas ferramentas (como por exemplo o WEKA), têm uma limitação quanto ao tamanho do *dataset* sob o qual aplicam os algoritmos de modelação.

Esta técnica é de elevada importância neste trabalho, na medida em que a base de dados transaccional dos clientes da SonaeMC atinge uma dimensão que impossibilita a aplicabilidade da maioria senão de todos os algoritmos de DM, quer por questão de fiabilidade de resultados, quer por motivos de desempenho.

A redução de dados consiste então na construção de um subconjunto de dados C' com origem no *dataset* original C sem descorar obviamente a qualidade dos dados. Uma vez reduzido para o *dataset* C' , o conjunto de dados encontra-se em condições de ser então dividido em **conjunto de treino** (sob o qual vai ocorrer a modelação) e **conjunto de teste** (no qual vai ser testado o modelo). Uma das técnicas mais mencionadas na literatura foi proposta por Klosgen no ano de 2002, e pode ser consultada no Anexo II.

3.3.4. Data Mining como parte do processo de KDD

Sob o ponto de vista de alguns autores, o Data Mining é o núcleo do processo de KDD, sendo a fase deste processo que tem recebido mais atenção por parte dos investigadores. Para autores como Shepard (1998) ou Drokdenko et al. (2002), a escolha das abordagens a aplicar sobre os dados depende directamente dos objectivos de marketing definidos e deverá começar pela expressão das tarefas de alto nível.

Normalmente o Data Mining, distingue-se de outras técnicas de Análise de Dados pela forma como explora as relações entre os dados. Se por um lado, no que respeita às diversas ferramentas de análise de dados disponíveis, o utilizador constrói hipóteses sobre relações específicas que corrobora ou refuta através da ferramenta utilizada, o processo de Data Mining é responsável pela geração dessas hipóteses, o que potencia maior rapidez, qualidade, autonomia e fiabilidade aos resultados.

Todo este processo sugere obviamente uma hierarquia que começa em instâncias elementares (embora muitas vezes de grande volume) e termina num ponto relativamente concentrado, mas de elevada importância – o conhecimento. A procura de padrões requer que os dados em bruto sejam sistematicamente simplificados de modo a desvalorizar aquilo que é específico e a privilegiar o que é genérico, este é um dos conceitos mais importantes do processo de KDD.

Segundo Ramachadran (2001), e numa perspectiva orientada para os processos, as actividades de DM podem agrupar-se em três classes:

- 1) **Descoberta de padrões:** este é o processo que pesquisa padrões “escondidos” na base de dados sem hipóteses pré-concebidas acerca daquilo que os padrões poderão ser. Nesta perspectiva os algoritmos têm a iniciativa de realizar a pesquisa sem qualquer orientação especificada pelo utilizador (e.g. tendências e variações ou associações entre determinados elementos);
- 2) **Modelação preditiva:** esta classe de actividades de DM, aplica na prática o resultado da descoberta de padrões e consiste na sua utilização para previsão do futuro. A modelação preditiva permite então ao utilizador submeter registos com valores desconhecidos em alguns atributos ao sistema, e este tendo em conta unicamente os padrões previamente descobertos, irá prever os valores desconhecidos;
- 3) **Análise Exploratória:** consiste na aplicação dos padrões extraídos para detectar dados anómalos ou pouco comuns. Neste tipo de análise, é primeiramente definido o padrão habitual e, tendo em conta uma determinada margem de desvio, detectam-se os elementos que se afastam desse padrão. Ao contrário da Descoberta de Padrões que procura “acontecimentos comuns”, a Análise

Exploratória ajuda a identificar elementos pouco comuns e casos específicos dentro da Base de Dados. Uma aplicação exemplo pode ser a detecção de fraude.

A temática do DM tem vindo a ser alvo de grande discussão e investigação nos últimos anos. Essa investigação desenvolvida viabilizou que o surgimento de metodologias de implementação, quer o surgimento de algoritmos de aprendizagem que se distinguem na forma como traduzem a informação descoberta e na forma como é realizada essa mesma descoberta, o que faz com que para determinado problema, haja algoritmos mais apropriados que outros. Para Linoff et al. (2002), o DM como parte do processo de KDD é sustentado em três pilares fundamentais, e dos quais depende o sucesso do projecto:

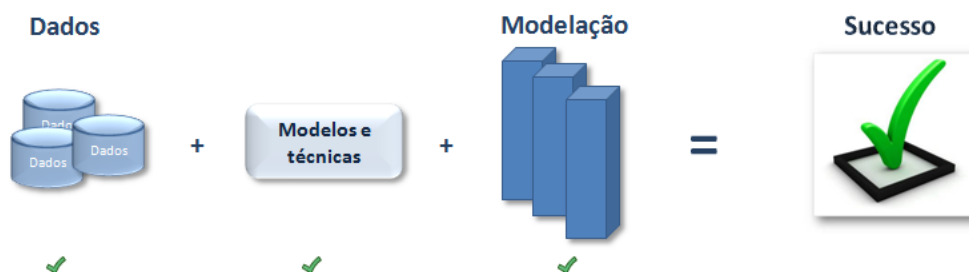


Figura 4 - pilares para o sucesso do DM (Linoff 2002)

3.3.4.1. Desafios no processo de DM

Existem diversos requisitos importantes para a obtenção de sucesso na aplicação de um processo de DM, essa mesma relevância incute a todo o processo e aos sistemas sustentados pela descoberta de conhecimento, um conjunto impar de desafios que devem ser tomados em conta.

Segundo Han et. al (2001), alguns desses desafios primordiais dos sistemas de DM são:

1) Metodologia de DM e questões de interacção do utilizador:

- a. **Procura de diferentes tipos de conhecimento em bases de dados:**
deve ser coberto um largo espectro de tarefas de descoberta de conhecimento e análise de dados, incluindo caracterização, discriminação, associação, classificação, agrupamento, análise de tendências e desvios e análise de similaridade dos dados;

- b. Procura interactiva de conhecimento a níveis múltiplos de abstracção e incorporação de conhecimento básico:** permite ao utilizador focalizar a sua pesquisa em padrões, provendo e refinando os requisitos do DM em função dos resultados obtidos. A informação referente ao domínio do problema estudado deve ser utilizada como fio condutor de todo o processo de forma a permitir que os padrões descobertos sejam expressos em termos concisos e em níveis diferentes de abstracção;
- c. Linguagens de consulta de DM e DM *ad hoc*:** as linguagens de consulta relacionais, como o SQL, permitem a construção de consultas específicas para um determinado fim (*ad hoc*). De forma similar, devem ser desenvolvidas linguagens de consulta para DM e o ideal é que estas venham a ser integradas em linguagens de DW e optimizadas para um DM eficiente e flexível;
- d. Apresentação e visualização dos resultados de DM:** o sistema deve ser provido de requisitos de expressão do conhecimento descoberto em linguagens de alto nível, representações visuais ou outras formas expressivas que tornem esse mesmo conhecimento de fácil compreensão e de usabilidade coerente para os utilizadores não especialistas;
- e. Suporte a ruídos e a dados incompletos:** o sistema deve possuir métodos de limpeza e análise para o tratamento de ruído e de dados incompletos, os quais comprometem a eficiência dos modelos;
- f. Avaliação de padrões:** este item diz respeito à utilização de medidas de interesse para guiar todo o processo de descoberta e reduzir o espaço de busca.

2) Itens relacionados ao desempenho:

- a. Eficácia e escalabilidade dos algoritmos de DM:** para a extracção de informação de grandes volumes de dados, os algoritmos de DM devem ser eficientes e escaláveis, ou seja, não só deve ser possível prever o tempo de execução, como também este deve ser aceitável;
- b. Algoritmos de DM paralelos, distribuídos e incrementais:** o sistema deve permitir a divisão de dados em partições, para que estas sejam

processadas em paralelo para posterior junção e avaliação de desempenho.

3) Diversidade dos tipos de bases de dados:

a. Suporte de tipos de dados relacionais e complexos: os sistemas de DM devem suportar não só os tipos de dados relacionais como os presentes em *Data Warehouses*, como também dados de natureza mais complexa, como dados espaciais ou temporais;

b. Bases de dados heterogêneas e sistemas de informação globais: referente à conexão de múltiplas fontes de dados.

4) Protecção da privacidade e segurança da informação: a utilização cada vez mais popular e crescente de ferramentas de DM e redes de telecomunicações, incute a estes sistemas a necessidade da existência de métodos que assegurem a protecção, privacidade e segurança da informação.

3.3.4.2. Metodologias Usadas no Data Mining

O Data Mining é um processo que tem inerente uma grande complexidade não só por envolver um grande número de tarefas, mas também pela existência do elevado número de algoritmos e ainda pela necessidade em alguns problemas do envolvimento de mais do que um indivíduo.

As entidades precursoras nestas áreas seguiram as suas próprias estratégias e métodos. Por conseguinte, foram sendo desenvolvidas diferentes estratégias para o mesmo problema, com a obtenção de diferentes resultados. Por este motivo, ao fim de algum tempo surgiu a necessidade de definir uma metodologia que servisse de referência para o desenvolvimento de projectos que envolvessem DM. Variadas metodologias foram então sendo propostas, no entanto actualmente podem distinguir-se apenas duas dessas metodologias que estão disseminadas e bem definidas: a metodologia CRISP-DM (*C***R***oss-Industry Standard Process for Data Mining*) e a metodologia SEMMA (*S***a***mple, E***x***plore, M***odify, M***odel, A***c***cessement*).

Apesar da disparidade de opiniões, encontra-se disponível um estudo na internet (Figura 5), onde pode verificar-se um domínio da metodologia CRISP-DM face a outras e

nomeadamente à metodologia SEMMA. No entanto, vale a pena salientar que o SEMMA, é cada vez mais utilizado por profissionais.

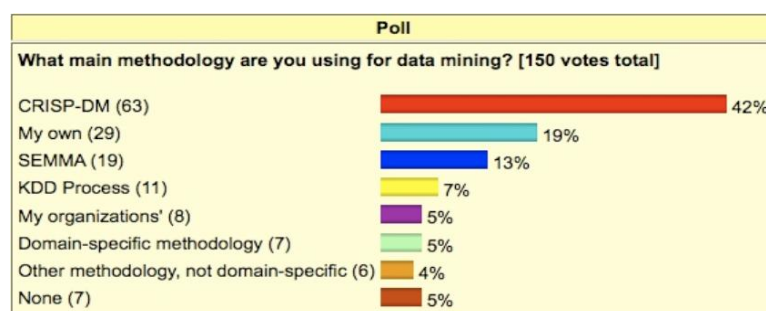


Figura 5 - Estudo sobre a preferência de metodologia de DM

(Fonte: estudo realizado por www.kdnuggets.com, 2007)

3.3.4.2.1. CRISP DM

Em finais de 1996, esta metodologia foi concebida por um consórcio formado por quatro empresas de consultadoria, oriundos de diferentes países como: Alemanha¹⁴, Estados Unidos¹⁵, Dinamarca¹⁶ e Holanda¹⁷. Este desenvolvimento foi motivado pelo crescente e generalizado interesse para o mercado de Data Mining. Esta metodologia foi sofrendo melhoramentos ao longo dos anos e em 2000, surgiu o CRISP-DM 1.0, que tem vindo desde então a ser utilizado para tornar os projectos de DM mais fáceis de gerir e mais fiáveis (Chapman, 2000). Esta ultima versão da metodologia é orientada para aspectos práticos, e tem como base os princípios académicos e também a experiência prática na aplicação de DM.

O CRISP-DM é descrito em termos de um processo hierárquico, que contém seis fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação (Figura 6). As seis fases foram pensadas para que pudessem ser aplicadas em qualquer área de negócio.

As fases que fazem parte do ciclo de vida de um projecto de exploração de dados segundo a metodologia CRISP-DM, relacionam-se de forma interactiva e a sua sucessão não deve ser encarada de forma rígida, embora exista uma ordem natural para a

¹⁴ DaimlerChrysler AG (<http://www.daimlerchrysler.com/dcom>)

¹⁵ NRC – proporciona soluções de Data Warehousing (<http://www.mrc.com>) e SPSS – que disponibiliza soluções baseadas em DM (<http://www.spss.com>)

¹⁶ NRC (<http://www.mrc.com>)

¹⁷ OHRA – grupo bancário (<http://www.ohra.nl>)

execução de cada uma das fases. As setas indicam as dependências mais importantes e mais frequentes entre as fases. O círculo externo do diagrama simboliza a natureza cíclica intrínseca ao próprio processo de Data Mining uma vez que o processo continua mesmo após a descoberta de uma solução que servirá de base para experiências posteriores.

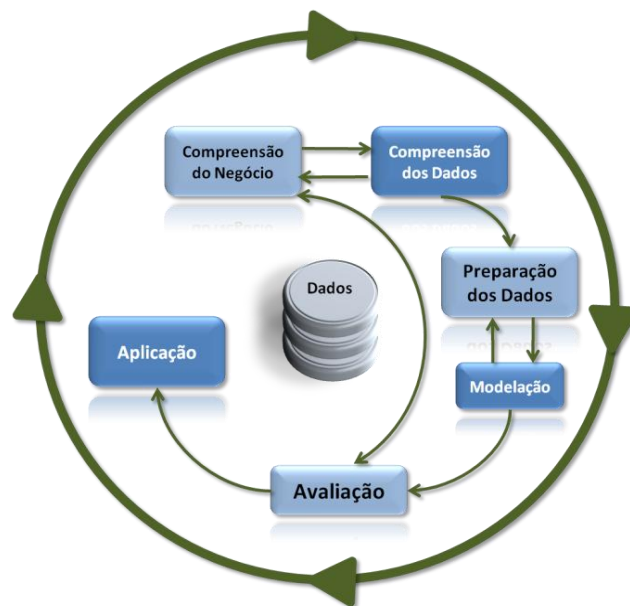


Figura 6 - Ciclo de vida da metodologia CRISP-DM

Os conceitos para cada uma das fases são bastante semelhantes e de acordo com os modelos propostos na literatura, nomeadamente por Fayyad et al. (1996), Cabena et al. (1997), Cios (2000), Han et al. (2001) e Klosgen et al. (2002), e podem ser resumidos como:

Compreensão do Negócio (*Business Understanding*): Esta fase inicial procura identificar as metas, os objectivos e as necessidades numa perspectiva de negócio e converter esse conhecimento, na definição de um problema de Data Mining e num plano inicial de ataque ao problema;

Compreensão dos Dados (*Data Understanding*): Nesta fase identifica-se a informação que possa vir a ser relevante para o estudo, procura-se um primeiro contacto com a informação identificando problemas na qualidade dos dados ou detectando subconjuntos interessantes para formar hipóteses sobre a informação;

Preparação dos Dados (*Data Preparation*): Esta fase cobre toda a construção do *data set* final. As tarefas intrínsecas a esta fase são, muitas vezes, realizadas várias vezes incluindo: selecção de tabelas, registos e atributos, bem como transformação e limpeza dos dados para a posterior importação para as ferramentas computacionais indicadas à aplicação de algoritmos de Data Mining;

Modelação (*Modeling*): Cabe nesta fase seleccionar e aplicar com ajuste dos respectivos parâmetros, várias técnicas de modelação. Por norma, existe várias técnicas que podem ser utilizadas sobre o mesmo problema. Algumas delas possuem requisitos específicos na formatação da informação, sendo por isso muitas vezes necessário regressar à fase de preparação de dados;

Avaliação (*Evaluation*): O modelo contruido na fase de modelação é avaliado tendo em conta os requisitos do negócio e é definida nesta fase a escolha do modelo óptimo;

Implementação (*Deployment*): Depois de construido e avaliado, o modelo construido está pronto a ser aplicado. Esta fase pode colidir com uma simples elaboração de relatórios para o apoio à tomada de decisão, ou a processos mais complexos como a incorporação do modelo criado em CRM¹⁸.

Esta metodologia não centra os princípios unicamente na tecnologia, mas também nas necessidades, mas também nas necessidades dos utilizadores para a resolução de problemas de negócio (Han e Kamber, 2006).

Convém ainda referir que o CRISP-DM proporciona documentos distintos como ferramenta da ajuda no desenvolvimento dos projectos de DM. Assim, pode dizer-se que esta metodologia é extremamente completa e documentada, sendo que cada uma das suas fases está devidamente organizada e definida, permitindo assim que o projecto possa ser compreendido e revisto com grande facilidade (Figura 7).

¹⁸ **Customer Relationship Management**: base de dados de um sistema de relacionamento com clientes. Ver 4.4.

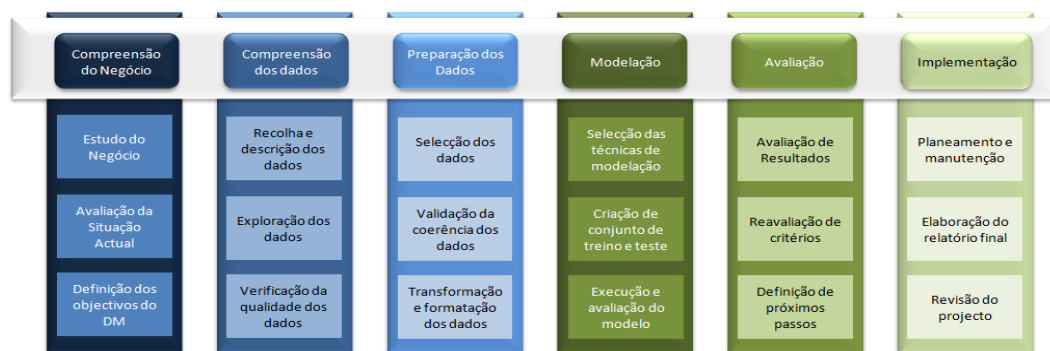


Figura 7 - Sequência comum das fases de CRISP-DM

3.3.4.2.2. SEMMA

A metodologia SEMMA foi proposta pelo Instituto SAS¹⁹, que se dedica ao desenvolvimento de soluções para estatística, análise de dados, Business Intelligence, Data Mining e Sistemas de Apoio à Decisão. Esta metodologia surgiu como resposta à necessidade de definição, padronização e integração de processos de DM nos ciclos de produção, para que a solução fosse aceite mais facilmente no ambiente do negócio (Groth, 2000). Mais do que uma metodologia de DM, é também considerada como uma ferramenta auxiliar na condução de um projecto por todas as suas etapas, desde a especificação do problema até à implementação dos resultados.

O Data Mining é definido pelo Instituto SAS como o “processo de extrair informação valiosa e relações complexas de um grande volume de dados” e foi exactamente neste sentido, que dividiram o processo de DM em 5 etapas (Figura 8), dando origem ao acrónimo SEMMA (SAS, 2005; Groth, 2000). Estas etapas, distintas, correspondem a um ciclo, em que as tarefas internas podem ser executadas de forma repetida sempre que se verifique necessário.

¹⁹ SAS Institute Inc. (<http://www.sas.com>)

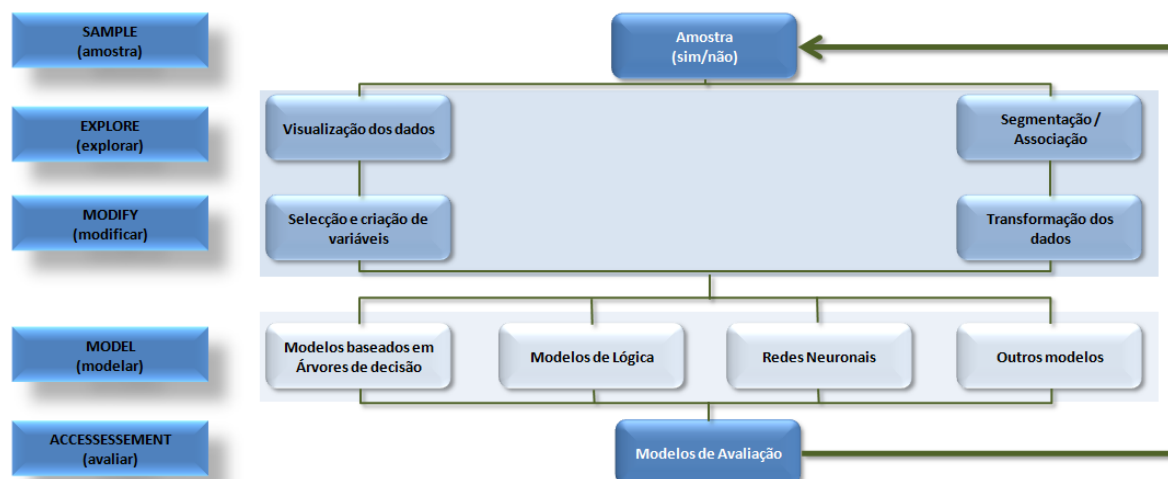


Figura 8 - ciclo de vida da metodologia SEMMA (adaptado de SAS, 2005)

As fases que completam a metodologia SEMMA podem ser resumidas como:

Amostragem (*Sample*): é realizada uma amostra, extraindo uma parte dos dados que se considere possuir informação significativa e efectua-se ainda à partição dos dados.

Exploração (*Explore*): é efectuada uma exploração exaustiva dos dados, uma pesquisa de relações antecipadas (através de métodos de associação e/ou segmentação, e são ainda identificadas as tendências e as anomalias de forma a garantir uma maior compreensão sobre os dados.

Modificação (*Modify*): nesta etapa são criados, seleccionados e transformados os dados com o objectivo fulcral de poder decidir-se qual o modelo a ser seleccionado.

Modelação (*Model*): é concretizada pela aplicação das ferramentas de análise disponíveis como árvores de decisão, redes neuronais, entre outras. Esta etapa é a responsável pela descoberta de conhecimento propriamente dito.

Avaliação (*Accessment*): nesta última etapa é realizada a avaliação dos resultados dos modelos, de forma a aferir acerca do desempenho dos modelos. Por norma, esta avaliação corresponde à apresentação de um conjunto de dados de teste ao modelo, avaliando a resposta do mesmo perante instâncias que não foram utilizadas na fase de

treino. Esta apresentação é efectuada através do método *Holdout*²⁰ com o objectivo exacto de determinar a eficácia dos modelos de previsão.

3.3.4.2.3. Comparação das metodologias (CRISP-DM VS SEMMA)

As metodologias CRISP-DM e SEMMA compartilham da mesma essência, estruturando o processo de DM em fases que se encontram inter-relacionadas entre si, convertendo o processo de descoberta de conhecimento num processo interactivo e iterativo.

A metodologia SEMMA centra-se mais nas características do desenvolvimento das técnicas e do processo, enquanto a metodologia CRISP-DM contém uma perspectiva mais ampla em relação aos objectivos empresariais do projecto. A diferença estabelece-se imediatamente desde a primeira fase, onde no SEMMA existe uma maior preocupação com a recolha da amostra de dados, enquanto o CRISP-DM realiza uma análise do problema de negócio com vista a transforma-lo num problema técnico. Pode assim afirmar-se que a metodologia CRISP-DM está mais virada para uma concepção real do projecto, podendo ser integrada numa metodologia de gestão de projectos.

No entanto, e apesar das diferenças evidentes, analisando estas duas metodologias, pode verificar-se que muitas das etapas do processo SEMMA estão directamente correlacionadas com as etapas do CRISP-DM (Figura 9).

²⁰ Método Houldout – segundo Hill (2002) é baseado no seguinte princípio: “a amostra total é dividida aleatoriamente em duas sub-amostras *A* e *B* (o número de casos em cada uma delas deve ser aproximadamente igual). O método *holdout* assume que as duas sub-amostras são representativas da amostra total e do Universo, e portanto, os resultados de uma técnica estatística aplicada à sub-amostra *A* devem ser muito semelhantes aos resultados obtidos por aplicação da mesma técnica à sub-amostra *B*”

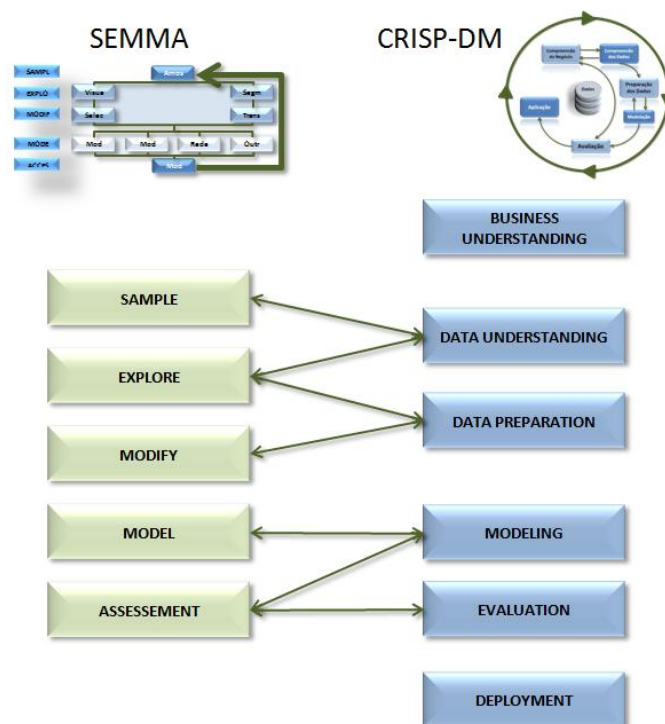


Figura 9 - comparação das interações entre as fases SEMMA e CRISP-DM

Uma outra diferença significativa entre as duas metodologias é a sua relação com as ferramentas comerciais. A metodologia SEMMA está muito ligada a produtos SAS e perfeitamente integrado com o *SAS Enterprise Miner* (Giudici, 2003), enquanto a metodologia CRISP-DM foi desenhada como uma metodologia neutra em relação à ferramenta que se utiliza para o desenvolvimento do projecto de DM, sendo que a sua distribuição é livre e gratuita.

Não obstante as evidentes diferenças entre os dois processos, e assente nas considerações de que este projecto será desenvolvido em *SAS Enterprise Miner* e *SAS Guide* (explicado em secção específica), o que incute uma diminuição de complexidade ao processo, e ainda a evidente importância no que respeita à compreensão do problema e dos dados e à implementação do modelo resultante, neste trabalho optou-se pela procura de conjugação das duas metodologias.

3.3.5. Pós-processamento – Interpretação dos resultados

Tal como a etapa de pré-processamento, esta etapa do processo de KDD, constitui uma espécie de engenharia dos dados, desta vez processando os dados de saída em vez dos de entrada. O objectivo é, portanto, construir um modelo de saída para tornar os resultados dos esquemas de DM mais eficientes, ou seja, para tornar os resultados mais compreensíveis para os seus utilizadores finais.

Existem dois aspectos que merecem especial atenção nesta fase: a **tradução da informação encontrada para uma forma perceptível ao utilizador do sistema**, e a **verificação da qualidade dessa informação**. Esta verificação pode ser encarada como um passo da etapa de DM, mais concretamente a fase de Avaliação que é parte integrante das duas metodologias de Data Mining apresentadas (CRISP-DM e SEMMA). Inclui-se esta fase na etapa de pós-processamento uma vez que a determinação dos valores de confiança, suporte e o grau de generalização do modelo resultante são dados extremamente importantes para a validação das regras encontradas e portanto devem ser fornecidos ao utilizador final.

A importância inerente a esta fase do processo de KDD prende-se com o facto de que, caso a análise dos resultados não seja efectuada, não pode descartar-se o risco premente de que sejam apresentados resultados errados ou simplesmente desinteressantes. Existem vários métodos de amostragem para estimar a capacidade de generalização de um modelo: *estatística simples*, *validação Percentage Split*, *validação cruzada* ou *bootstrapping*.

A par do método de **Validação Cruzada**, um dos métodos mais populares, pela sua simplicidade de implementação, para a estimação do erro de generalização é o método de **Percentage Split**. Este baseia-se numa divisão do *data set* em dois subconjuntos, um deles formado por instâncias que serão utilizadas para treino do modelo e as restantes instâncias formam o conjunto de teste, que será utilizado para estimar o erro de validação. Este método tem como pontos fortes a sua simplicidade e rapidez, embora possa produzir uma redução efectiva nos dados disponíveis para treino do modelo de aprendizagem.

Já o método de **Validação Cruzada** é considerado por muitos autores como um melhoramento ao método **Percentage Split**, permitindo a utilização de todos os casos disponíveis para treino do modelo. Na validação cruzada, o conjunto de dados original C , é subdividido em k subconjuntos mutuamente exclusivos de tamanho aproximadamente igual ($C'_1, C'_2, \dots, C'_k$). Os modelos são treinados e testados k vezes. O erro final de generalização do modelo é dado pela média dos erros de validação obtidos nos k testes. O valor de k pode variar entre 2 e n , embora o valor $k=10$, seja o mais frequentemente utilizado. Este método destaca-se positivamente, mas à custa de um maior esforço computacional, face ao método *Percentage Split* para *data sets* de pequena dimensão.

Existem várias técnicas de avaliação²¹, que podem ser seleccionadas em função do tipo de problema ou em função do algoritmo de DM aplicado. Por exemplo, em problemas de classificação, os mais utilizados são: a *Matriz de Confusão*, as *Curvas ROC*, e nos casos de regressão as medidas mais utilizadas são: o *Mean absolute deviation*, *Sum Squared Error*, *Mean Squared Error* e *Root Mean Squared Error*. Face a outros problemas, podem ainda ser utilizadas outras medidas de avaliação das quais se falará atempadamente neste trabalho.

3.4. Customer Relationship Management (CRM)

Como já foi várias vezes referido, uma etapa essencial e o objectivo primordial da procura de conhecimento em bases de dados (Data Mining) é o da interpretação dos resultados e a transformação destes numa base real e trivial de decisões. Depois da colecta, preparação e estudo dos dados, os resultados obtidos por qualquer que seja o modelo ou técnica utilizados requerem uma avaliação cuidada de forma a encontrar a aplicabilidade dos resultados ao problema proposto.

²¹ Breve abordagem às técnicas de avaliação, pode ser consultada no Anexo II deste trabalho.

Crescendo a uma velocidade vertiginosa, os problemas propostos pelas organizações a esta “recente ciência” – o Data Mining, têm como objectivo comum, e independentemente da vertente ou sector da organização, que assenta na procura pela compreensão cada vez mais individualizada dos seus clientes. Para isso, as empresas contam actualmente com as técnicas e modelos de DM, para que com base no conhecimento gerado, possam garantir a satisfação e a fidelidade dos seus clientes. Este tipo de mecanismos de gestão são relativamente recentes, e aparte algum tipo de resistência pelos sistemas de gestão mais instalados, têm cada vez mais sido um foco de preocupação e de mudança das organizações.

Em pequenas empresas existe uma estreita relação entre os clientes e a organização, o que permite que a empresa aprenda mais e com maior rapidez sobre o seu cliente, e possa utilizar esse mesmo conhecimento para melhor o satisfazer e fidelizar. Em grandes empresas, este tipo de relação é impraticável. É então necessário que sejam desenvolvidos meios para que a relação com o cliente seja explorada. Neste contexto nasce o CRM (*Customer Relationship Management*).

A inteligência de *marketing* para a manutenção de clientes está intimamente ligada à filosofia do *Customer Relationship Management* (CRM), um sistema que ajuda a empresa a reunir todo o conhecimento colectivo sobre os seus clientes, concentrando os dados num lugar onde possam ser partilhados por todos os funcionários, por meio do uso de um conjunto variado de ferramentas. Com este sistema, a empresa consegue entender melhor os clientes, oferecer-lhes mais apoio, servi-los melhor e, em última análise, ganhar novos clientes (Gaset, 2001).

Antes de ser um software, a filosofia CRM é o relacionamento da empresa com seus clientes. Por meio da tecnologia, a organização procura o resgate qualitativo das relações, estabelecendo um diálogo interactivo com o cliente. Para isso, é essencial retornar aos modelos de relacionamento históricos, e procurar compreender os métodos, os números e as restrições de tempo inerentes às organizações.

3.4.1. Definição CRM

Segundo CISTER, 2005 CRM pode ser definido como uma estratégia de negócio virada para o entendimento e antecipação das necessidades dos clientes actuais e potenciais de uma empresa ou organização. Pode também dizer-se que CRM é a infra-estrutura que permite que seja implementada a filosofia one-to-one (um para um) no relacionamento de uma empresa com os seus consumidores.

De uma visão mais técnica, o CRM envolve todo o processo desde a colecta, consolidação e análise dos dados do cliente, possibilitando no final da análise uma passagem de conhecimento dos resultados aos vários pontos de contacto com o cliente para que essa mesma informação possa ser utilizada de forma a que a interacção com o cliente seja o mais “lucrativa” por parte da empresa e mais “agradável” por parte do cliente possível.

Actualmente as estratégias corporativas deixaram de assentar numa visão de marketing focada no produto, para assentar um contexto mais personalizado, centrado em estratégias e processos com o foco específico no consumidor. De uma forma cada vez mais evidente, as empresas organizam-se em função dos vários tipos de clientes, permitindo assim a diferenciação dos clientes, em primeiro lugar pelas necessidades dos vários tipos de clientes, e posteriormente pelas suas necessidades individuais. Neste contexto, a utilização de um modelo de gestão baseado em CRM tornou-se cada vez mais premente, sendo actualmente uma vantagem competitiva.

Segundo CISTER, 2005 existem 3 padrões diferenciados para um sistema de CRM:

- 1) **CRM Analítico:** cuja função é determinar quais são os clientes que devem ser tratados de forma personalizada e quais os que devem ser transladados para patamares de prioridade inferior;
- 2) **CRM Operacional:** cujo objectivo é racionalizar e optimizar os processos da empresa numa vertente mais operacional (exemplo Call Centre);
- 3) **CRM Colaborativo:** que é a aplicação das tecnologias de informação que permitem a automatização e a integração de todos os pontos de contacto do cliente com a organização. Esses pontos de contacto deverão estar providos de

mecanismos e ferramentas que permitam a interacção com o cliente, e ao mesmo tempo disseminem as informações recolhidas do cliente ao longo do contacto para os sistemas de CRM Operacional.

A título de resumo pode então dizer-se que o CRM está directamente relacionado com o armazenamento, processamento, análise e distribuição dos dados, sendo este processo provido de uma total preocupação com o cliente, para que as organizações consigam tornar a relação com o cliente o mais eficiente possível. O CRM pode também ser visto como um elo de integração entre diferentes áreas de negócio como o Marketing, Serviços Financeiros, Tecnologia de Informação e Comitês de Decisão, áreas que até há poucos anos, trabalhavam de forma quase que autónoma e singular.

3.4.2. CRM na Distribuição

Nas empresas da grande distribuição a quantidade de informação foi crescendo exponencialmente, e cada vez mais foi necessário explorar a utilização de variáveis, e essencialmente na área da identificação e compreensão das necessidades dos clientes.

A implementação de ferramentas de Data Mining e de CRM permitem a estas organizações definir perfis e direccionar e aplicar estratégias de acordo com esses perfis, uma vez, que estes sistemas são um recurso para apoiar a tomada de decisão. Os sistemas de suporte à decisão, permitem retirar benefícios da sua utilização, essencialmente ao nível dos gestores de loja, sendo sistemas interactivos, permitem a definição de modelos de decisão, com base em séries hipotéticas que poderão ser criadas. Estas metodologias poderão por exemplo ser utilizadas para a compreensão de hábitos de consumo, conseguindo determinar os artigos que são frequentemente comprados em conjunto e, com isso, identificar uma oportunidade para fazer promoções sobre esses mesmos produtos e aumentar as vendas.

Os avanços tecnológicos estão a produzir enorme impacto em todas as organizações, independentemente da sua dimensão ou sector de actividade. Alguns desses avanços permitiram que diversas funções desenvolvidas pelas organizações fossem descentralizadas

para parceiros de negócio, possibilitando que com menos recursos humanos se desenvolvesse maior volume de trabalho, que condicionalmente conduziu ao declínio do tamanho da organização.

No âmbito do Marketing relacional, este tipo de tecnologia possibilitou automatizar tarefas e processos repetitivos, identificar os clientes através do seu potencial e gerir a bolsa de clientes de forma proactiva. Há que salientar, ainda a possibilidade e a capacidade de interagir com os clientes, e de certo modo, reduzir os custos.

Não obstante a importância da implementação destas ferramentas, as estratégias de CRM e Marketing Relacional, devem ser suportadas por soluções tecnológicas que permitam adequar as campanhas de Marketing face aos clientes que a organização possui, através de campanhas direccionadas, de acordo com o perfil e necessidades e desejos dos clientes. Esse é actualmente, e como já foi anteriormente referido um factor diferenciador e torna-se cada vez mais numa vantagem competitiva.

3.5. Conclusão

Neste capítulo foram apresentados os principais conceitos referentes à inteligência de processos relacionados com a descoberta de conhecimento em bases de dados. Foram descritas as principais definições a respeito do termo *Data Mining*, procurando a diversificação e abrangência através da citação de várias fontes bibliográficas. Descreveu-se a fundamentação básica de todo o processo de KDD, citando ainda as raízes teóricas da etapa de Data Mining. E foram ainda apresentadas também as definições de sistemas de CRM e *Data Warehouse*.

A título de resumo do trabalho exposto neste capítulo pode ainda citar-se que o elevado crescimento das Bases de Dados empresariais, governamentais e institucionais ultrapassa a capacidade humana de interpretar e assimilar toda esta informação. Tal facto deu origem à necessidade de geração de novas ferramentas e técnicas inteligentes,

capazes de realizar um tratamento, análise e extracção de conhecimento de forma automática (ou semi-automática) (Fayyad et al, 1996).

A área de KDD e de DM encontra-se em progressiva expansão e é uma das áreas na qual a comunidade científica tem investido. A primeira pode ser considerada como todo o processo necessário para atingir o conhecimento, enquanto a segunda pode ser encarada apenas como parte integrante da primeira, embora seja uma das etapas que tem merecido maior atenção por parte dos investigadores.

O KDD tem como objectivo extrair padrões e modelos a partir de dados em bruto. Os padrões podem ser obtidos com a realização das fases de selecção, pré-processamento e transformação dos dados. Após o processamento dos dados, são aplicados os algoritmos de DM. Por fim, realiza-se a interpretação dos resultados obtidos por todo o processo.

O processo de KDD é interactivo, uma vez que as fases estão muito relacionadas e qualquer alteração numa delas poderá afectar as fases seguintes e até o sucesso de todo o processo. Importa também referir que a elaboração de um projecto de KDD é um processo complexo, por isso é necessário recorrer a metodologias padrão para o desenvolvimento de descoberta de conhecimento, no qual se enquadram o CRISP-DM e o SEMMA.

4. Objectivos e Abordagens em Data Mining

O presente capítulo tem o objectivo de apresentar os principais problemas e abordagens utilizados no processo de Data Mining, evidenciando as propriedades associadas, com particular ênfase naqueles que foram usados na aquisição de conhecimento no âmbito do caso de experimentação *Churn*.

4.1. Introdução

Como tem vindo a ser dito ao longo deste trabalho, o Data Mining procura a detecção automática de padrões, previamente desconhecidos e potencialmente úteis, que não se encontram explicitamente descritos numa Base de Dados (Hongxia e tal., 2009).

Com o fim de atingir esse seu primordial objectivo, o DM oferece actualmente uma panóplia cada vez mais alargada de algoritmos. Esses algoritmos são utilizados para resolver problemas como: descoberta de afinidades ou associações, *clustering*, predição, estimação de propriedades, classificação entre outros.

Na maior parte da investigação sobre *Churn* procura-se identificar e caracterizar os clientes que podem vir a abandonar a organização, i.e. potenciais *Churners*. Pretende-se, deste modo, poder estabelecer medidas de forma proactiva, no sentido de evitar que no futuro clientes com características semelhantes possam vir a ter um comportamento de *Churn* (Hongxia et al., 2009).

O problema do *Churn* é tradicionalmente conceptualizado como um problema de classificação, e é resolvido na maioria dos casos com recurso a técnicas de Data Mining. A análise da literatura sobre técnicas de Data Mining mais usadas para classificação de *Churn* demonstra que no *top list* se encontram as Árvores de Decisão e as Redes Neurais. No entanto, algoritmos como classificadores de Bayes (*Naive Bayes* e

Redes Bayesianas) são algoritmos, que respondendo de forma bastante positiva a problemas de classificação, também são bastante utilizados.

De seguida, e uma vez que a temática deste trabalho assenta no problema de *churn*, serão explanadas com maior detalhe estas técnicas.

4.2. Diferentes objectivos em Data Mining

Como foi atrás referido, a escolha do algoritmo de DM a aplicar depende do tipo de objectivo do mesmo. Genericamente e segundo Fayyad et al., 1996, podem distinguir-se dois objectivos principais no Data mining:

- 1) **Verificação** - que consiste apenas na verificação da hipótese do utilizador;
- 2) **Descoberta** – que consiste na procura de novos padrões, que se subdivide em:
 - a. **Previsão** – procura de padrões que permitam prever o futuro;
 - b. **Descrição** – procura de padrões que apresentem o conhecimento de forma compreensível.

4.2.1. Previsão

A previsão envolve a utilização de algumas variáveis ou campos de uma base de dados, no sentido de prever valores desconhecidos ou valores para outras variáveis de interesse. O principal objectivo do **Data Mining Preditivo** consiste em automatizar o processo de tomada de decisão, criando um modelo capaz de estimar ou prever um determinado valor. Por norma, os resultados do modelo são utilizados directamente nos dados, tornando a *accuracy* do modelo como a medida de desempenho mais importante para a sua avaliação.

Para estudos com o objectivo de previsão, os problemas são tratados como pertencentes a uma das seguintes classes:

- 1) **Classificação** – encontrar uma função que faça o mapeamento dos dados em classes pré-definidas (e.g. diagnóstico de uma doença a partir de um conjunto de sintomas; ou detecção de abandono de um cliente perante uma série de variáveis descritivas do seu comportamento;
- 2) **Regressão** – encontrar uma função desconhecida cuja saída tem um domínio de valores reais (e.g. previsão do valor das acções da bolsa).

4.2.1.1. Classificação

Trata-se de um dos tipos de problema mais frequente em DM e consiste basicamente na análise das características de um objecto através de um *classificador*, que classifica o objecto num número de possíveis categorias, definidas *a priori* pelo analista. Uma vez que as classes ou categorias estão previamente definidas, o objectivo do processo é a aprendizagem de uma função que faça o mapeamento de um novo elemento numa das várias classes, tendo em conta o seu “comportamento” ou seja, os valores que esse mesmo elemento tem nas variáveis explicativas.

O processo de classificação é supervisionado, o que implica ser necessária que a especificação de características com os respectivos valores para cada classe ou categoria, quer proporcionar ao sistema um conjunto de dados preparado (em que existam instâncias pertencentes a diferentes categorias), permitindo assim ao sistema aprender, generalizar e classificar novos padrões.

Algumas das técnicas de aplicação comum em problemas de classificação são as **árvores de decisão**, a **indução de regras**, e as **redes neuronais artificiais**.

4.2.1.2. Regressão

A regressão, que é frequentemente intitulada de previsão, tem como objectivo encontrar uma função de mapeamento, entre um conjunto de variáveis independentes (de entrada), e uma variável dependente (de saída) que é do tipo numérico. Normalmente, a regressão

é utilizada quando se pretende prever o comportamento de uma determinada variável. Os algoritmos que mais se aplicam nesta abordagem são a **regressão múltipla**, as **árvores de decisão**, as **redes neuronais artificiais**, e as **máquinas de vectores de suporte**.

4.2.2. Descrição

Pode dizer-se sumariamente que o objectivo primário do **Data Mining Descritivo** consiste em aumentar o conhecimento e a compreensão sobre os dados, na medida em que foca a procura de padrões descritivos reconhecidos ou interpretáveis pelo ser humano. Ao contrário do DM Preditivo, os resultados obtidos nem sempre se traduzem em acções de aplicação directa. As novas perspectivas e o conhecimento adquirido sobre os dados durante a construção do modelo, são o aspecto mais importante do processo.

Para este objectivo de descrição há vários métodos, tais como:

- 1) **Segmentação (*clustering*)** - procura de um número finito de conjuntos (*clusters*) que descrevam os dados;
- 2) **Sumariação** – procura de uma descrição compacta de um conjunto ou subconjunto de dados;
- 3) **Associação ou Análise de Dependências** – procura de um modelo que descreva as relações entre as variáveis;
- 4) **Detecção de desvios** – descoberta de alterações significativas nos dados.
- 5) **Visualização** – possibilitar a representação gráfica dos resultados finais ou intermédios.

4.2.2.1. Segmentação (Clustering)

A segmentação consiste na identificação de um conjunto finito de categorias ou *clusters* usados para descrição dos dados de acordo com uma métrica, segundo a qual se procura maximizar as semelhanças e minimizar as diferenças dos *clusters* entre si.

Neste tipo de modelos, a aprendizagem é não supervisionada, ao que corresponde na prática, à inexistência de qual a informação inicial no sistema, assumindo-se portanto que este aprende por si mesmo. Nesta técnica, a pesquisa é realizada sobre os dados tendo em conta todas as características dos mesmos.

Ao contrário do que acontece na classificação, não existe um conjunto predefinido de categorias nos dados. Estas são descobertas em função da análise dos dados e da sua natureza. Esta técnica permite o agrupamento de dados em diferentes clusters, sobre os quais se poderão realizar estudos mediante o recurso a técnicas estatísticas, a árvores de decisão ou a redes neuronais. Como exemplos de técnicas que podem ser aplicadas tem-se o algoritmo **K-means** e as **redes neuronais do tipo Kohonen**.

4.2.2.2. Sumariação

O objectivo da sumariação consiste em determinar uma descrição compacta para um subconjunto de dados através da aplicação de métodos próprios. Os métodos de sumariação mais sofisticados derivam de regras de resumo e descobertas de relações funcionais entre variáveis, sendo sempre aplicadas à análise exploratória de dados e à geração automática de relatórios.

4.2.2.3. Associação ou Análise de Dependências

A Análise de Dependências é a tarefa responsável por encontrar um modelo que descreva dependências significativas entre as variáveis. Normalmente esta abordagem é utilizada em dados transaccionais. Identificam-se atributos que estão directa ou indirectamente relacionados, ou seja, são utilizadas regras que permitam identificar essas relações. Esta técnica é muito na análise de compras (*basket analysis*), e um dos algoritmos mais utilizado é o **apriori**.

4.2.2.4. Visualização

A visualização possibilita a representação gráfica dos resultados (finais ou intermédios) de DM recorrendo a formas visuais facilmente perceptíveis. O objectivo da visualização

consiste em descrever as informações complexas através de diagramas, permitindo uma melhor representação de padrões e tendências. Quanto melhor for a descrição de um conjunto de dados, maior é a possibilidade de o entender e de compreender o domínio em que está inserido.

4.3. O problema da classificação

Foi já mencionado que a classificação é o processo de encontrar um conjunto de modelos que descrevam e distingam classes de dados, que são posteriormente utilizados para a predição de objectos cujas classes são desconhecidas, baseada na análise de um conjunto de dados de treino onde as classes são conhecidas.

É então objectivo da tarefa de classificação, encontrar algum tipo de relação entre os atributos preditivos e o atributo objectivo ou *target*, de forma à obtenção de um conhecimento que possa vir a ser utilizado para prever a classe de uma determinada instancia não classificada.

As tarefas de classificação requerem muitas vezes a análise de relevância para identificar atributos que não contribuem para esses processos, de forma a exclui-los do conjunto de variáveis explicativas de forma não só a acelerar o processo de descoberta de conhecimento, mas também para aumentar o grau de qualidade e legibilidade do modelo gerado. O desenvolvimento deste tipo de modelos caracteriza-se pela definição do grupo de classes, das variáveis relevantes e de um conjunto de treino que contem instâncias pré-classificadas.

Grande parte dos problemas de classificação de interesse real têm duas fases: uma etapa *in sample* ou de treino (de aprendizagem), que é executada a partir do *data set* existente, e uma etapa *out of sample* ou de generalização, onde são apresentados dados que não foram utilizados para o treino do modelo. O objectivo primordial é extrair do *data set*, as características que definem o mapeamento entre a entrada e a saída, para que possam

ser posteriormente utilizadas na tomada de decisão sobre novos dados, que não tenham sido ainda avaliados pelo sistema.

Segundo Han et al. (2001), o primeiro passo (**Figura 10 a**) na tarefa de classificação é a construção de um conjunto pré-determinado de classes e conceitos. O modelo é construído através da análise das instâncias do *data set* que são descritas por um conjunto de variáveis, e cada instância é assumida como pertencente a uma determinada classe (atributo que contem a informação da classe: *class label attribute*).

Estas instâncias ou exemplos formam o conjunto de treino, que são seleccionadas do *data set* original de forma aleatória, através de métodos de redução de dados e amostragem. A Figura 10, ilustra este processo definido por Han et al.

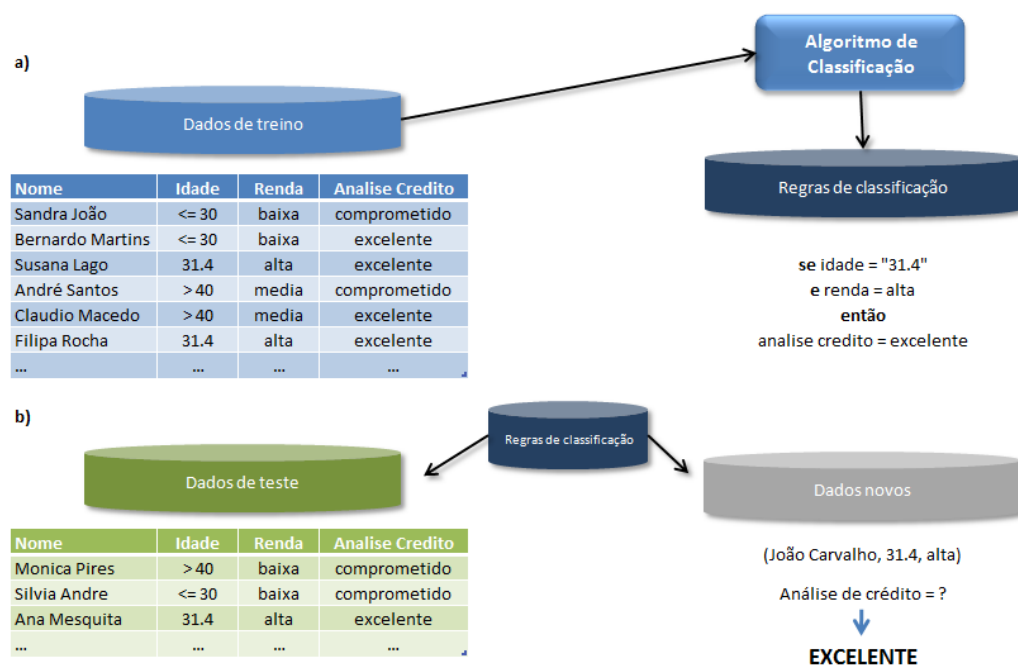


Figura 10 - processo de classificação de dados (by Han et al. 1996)

Uma vez que a descrição de todas as variáveis são fornecidas ao modelo, esta etapa de classificação é também denominada de **aprendizagem supervisionada** (a aprendizagem do modelo é dita de supervisionada uma vez que se conhece a classe a que pertence cada instância do conjunto de treino). Tomando como exemplo o *data set* utilizado na formação do modelo proposto para este trabalho, para cada instância (registo) do conjunto de treino, formado por diversos atributos relacionados com cada

cliente, são conhecidos os registos pertencentes à classe alvo da predição, ou seja a classe *churner*.

A segunda fase ou segundo passo do processo proposto por Han et al. (2000), (ilustrado na Figura 10b), o modelo obtido é utilizado para classificar novas instâncias. Neste passo, estima-se então a precisão do modelo (ou classificador). Segundo o mesmo autor, os principais critérios para avaliar a precisão do modelo são:

- 1) **Precisão da predição:** mede a habilidade do modelo em predizer correctamente a classe dos novos dados;
- 2) **Velocidade:** mede os custos computacionais envolvidos na utilização do modelo;
- 3) **Robustez:** mede a capacidade do modelo efectuar predições correctas mesmo sob um conjunto de dados com ruído ou *missing values*;
- 4) **Escalabilidade:** mede a eficiência da aplicação do modelo numa grande quantidade de dados;
- 5) **Interpretabilidade:** refere-se ao nível de informação útil e conhecimento descoberto pelo modelo.

A precisão de um modelo num determinado *data set* é medida como a percentagem de instâncias de teste correctamente classificadas pelo classificador. Para cada registo de teste, o valor que é previamente conhecido da classe, é comparado com o valor previsto do algoritmo. Se a precisão do modelo for considerada aceitável, o modelo pode ser utilizado para classificar dados futuros.

4.3.2. Métodos e técnicas de classificação

Num projecto de DM com foco definido em Classificação e Predição, a utilização de diferentes abordagens e algoritmos é fundamental para um processo de avaliação. Segundo Han et al. (2001), não é apenas a precisão do modelo que deve ser avaliada mas também a velocidade, robustez, escalabilidade e interpretabilidade. E neste aspecto,

a comparação entre modelos gerados segundo várias abordagens, torna-se numa mais-valia no que respeita à avaliação de resultados.

Dependendo do modelo e da estrutura de predição num cenário de aprendizagem supervisionada, alguns algoritmos e ferramentas podem ser mais adequados ao problema em questão. Alguns dos algoritmos mais utilizados em tarefas de classificação são Árvores de Decisão, Classificadores Bayesianos, Algoritmos Genéticos²² e Redes Neurais. A escolha do algoritmo a utilizar tal como a ferramenta na qual o modelo será construído dependem essencialmente de três factores: disponibilidade de uso, performance e custo; factores esses altamente ultrapassados pela ferramenta utilizada neste trabalho (SAS Enterprise Miner).

4.3.2.1. Árvores de Decisão

As árvores de decisão²³ foram inicialmente desenvolvidas pela Universidade de Michigan, com origem na área de *Machine Learning*.

Este algoritmo baseia-se na estratégia de ***dividir para conquistar***. De uma forma geral, esta estratégia segue a directriz de uma sucessiva divisão do problema em vários *sub-problemas* de menor dimensão, até que uma solução para cada um dos problemas mais simples seja encontrada. Os algoritmos de indução de Árvores de Decisão, vão então construindo a árvore de forma recursiva a partir dos dados de treino, dividindo os dados em subconjuntos até que esses representem uma só classe/valor.

Tendo como fundamento este princípio, os classificadores baseados em árvores de decisão procuram encontrar formas de dividir sucessivamente o problema em vários subconjuntos, criando nós contendo os respectivos testes, até que cada um destes contemple apenas uma classe, ou pelo menos até que uma das classes demonstre uma clara maioria que não justifique posteriores divisões (problema trivial).

Mais resumidamente, a classificação consiste em seguir o caminho ditado pelos sucessivos testes colocados ao longo da subdivisão até que seja encontrada uma folha que contenha uma só classe a atribuir à nova instância (Figura 11).

²² Algoritmos Genéticos – técnica não utilizada neste trabalho. Referência explicativa em Anexo II.

²³ Para uma abordagem mais descritiva deste Algoritmo, consultar Anexo II.

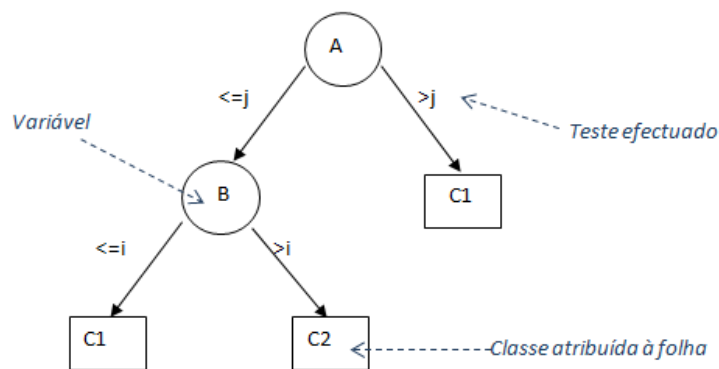


Figura 11 - Esquema de um classificador baseado em Árvores de Decisão

As árvores de decisão são representações gráficas de regras de classificação e a sua estrutura é constituída por:

- **Nó raiz:** nó com o primeiro teste (**A** na Figura 11);
- **Nós internos:** cada um possui um teste a um atributo dos dados e têm duas ou mais sub-árvores que correspondem às respostas possíveis (**B** na Figura 11);
- **Ramos:** contendo valores dos atributos;
- **Folhas:** que representam as classes (**C1** e **C2** na Figura 11).

No âmbito dos objectivos de DM, as árvores são mais adequadas a problemas de classificação, produzindo resultados de forma mais rápida que outras técnicas, além de que, e pela sua representação gráfica, os resultados obtidos podem também ser providos de uma maior compreensibilidade por parte do usuário. No entanto pode apontar-se, como desvantagem desta técnica, o desempenho obtido em problemas com uma elevada não linearidade.

4.3.2.2. Redes Neurais

Nas redes neurais artificiais, a ideia é realizar o processamento de informação tendo como princípio a organização de neurónios do cérebro. Como o cérebro humano é capaz de aprender e tomar decisões baseadas na aprendizagem, as redes neurais artificiais devem fazer o mesmo. Assim, uma rede neuronal pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado em

aprendizagem (experiência) e disponibilizar este conhecimento para a aplicação em questão.

As Redes Neurais artificiais, também conhecidas como modelos conexionistas, são redes densamente interligadas de unidades computacionais simples, designadas neurónios. As entradas da rede consistem num conjunto de N valores numéricos, assim como as suas saídas. Um neurónio calcula a diferença da soma ponderada das suas entradas com um *threshold*, e determina o valor de uma função não linear (por exemplo a *função sigmóide*) para esse valor. As redes neuronais são criadas a partir da ligação da saída de um neurónio à entrada de um ou mais outros. Uma das possíveis topologias das redes neuronais é o perceptrão multi-camada (Figura 12), em que a saída de cada neurónio apenas é usada como entrada de unidades que se encontrem numa camada posterior. Este tipo de redes são especialmente úteis para tarefas de classificação, atribuindo um valor alto à saída que representa a classe referente aos valores de entrada, e um valor mais baixo às saídas referentes às restantes classes.

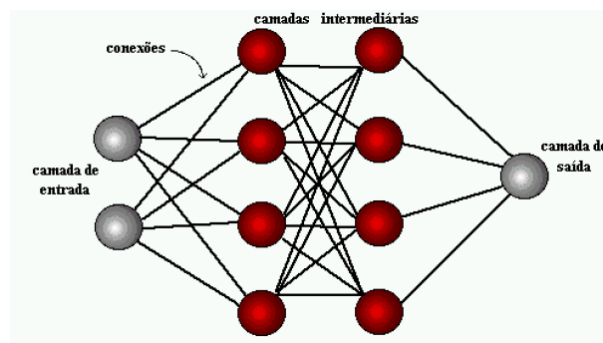


Figura 12 - Representação de uma Rede Neuronal Multicamada

O objectivo é determinar os pesos associados às entradas de cada unidade computacional, e o treino é efectuado fornecendo um exemplo de cada vez, e alterando os valores dos pesos de acordo com a classificação do exemplo, pelo que se trata de um mecanismo de aprendizagem supervisionada.

Existem algumas desvantagens na utilização de redes neuronais em DM, a primeira das quais diz respeito à baixa velocidade do processo de aprendizagem quando comparado com os mecanismos de aprendizagem simbólica. Uma segunda desvantagem é que a informação gerada pelas redes é representada sob a forma de uma rede, traduzindo-se na sua topologia e nos valores de cada um dos seus pesos, em vez de explicitamente

representada sob a forma de regras ou padrões conceptuais. Por último, é difícil, embora possível, incorporar conhecimento do domínio ou interacção do utilizador durante o processo de aprendizagem (Holsheimer, 1994).

Uma abordagem mais exploratória sobre esta técnica pode ser consultada no Anexo II.

4.3.2.3. Regressão Logística

O modelo de Regressão Logística (RL) é uma técnica de análise multivariada cuja variável dependente (ou de resposta) é uma variável binária. Esta técnica permite ainda estimar a **probabilidade de ocorrência** de um evento (no presente caso, o evento de um cliente abandonar) e identificar as variáveis independentes que contribuem efectivamente para a predição da variável de resposta. Mesmo quando a resposta de Interesse (variável dependente) não é originalmente do tipo binário, alguns investigadores tendem a torná-la dicotómica de modo que a probabilidade de sucesso possa ser modelada através da regressão logística.

No caso de um modelo de previsão de *Churn* – abandono = *sim* ou *não* – é modelada a proporção de resposta de uma das duas classes em função das variáveis independentes. Logo, a variável resposta de interesse Y pode ser classificada como igual a 1, geralmente denotado para a resposta de maior interesse, também chamado de “sucesso”, que pode ser utilizado no caso de um cliente ser *churner* o serviço da empresa, ou igual a 0 caso contrário.

Suponha-se então que existe um conjunto de i variáveis independentes que caracterizam um cliente, e seja Y a variável de resposta que se pretende prever com $1 = \text{churner}$, o modelo de Regressão Logística pode ser escrito da seguinte forma:

—, onde

Os coeficientes são estimados a partir do conjunto de dados, pelo método da máxima verosimilhança. Esse método encontra uma combinação de coeficientes, a qual maximiza a probabilidade de a amostra ter sido observada (Hosmer e Lemeshow, 1989).

Uma explicação mais abrangente e detalhada desta técnica pode ser consultada no Anexo II deste trabalho.

4.4. Conclusão

Neste capítulo foi apresentado um breve estudo sobre os diferentes objectivos de um projecto de Data Mining, nomeadamente a Previsão e Descrição. No que respeita ao objectivo da Descrição foram sucintamente explanados os métodos de Segmentação, Sumariação, Associação e Visualização.

No que respeita ao Data Mining Preditivo, e focando a atenção no *case study* deste trabalho, foi dada especial atenção aos métodos de classificação, nomeadamente Árvores de Decisão, Classificadores Bayesianos, Redes Neurais e Regressão Logística.

A escolha destas técnicas deu-se em função da sua disponibilidade nas ferramentas utilizadas, da sua adaptabilidade ao problema em estudo, da performance dos algoritmos (ponto essencial de avaliação, considerando-se a natureza e o tamanho do *data set* em estudo) e do custo (principalmente sob a óptica dos recursos necessários para a execução das ferramentas).

A revisão sobre a fundamentação teórica de cada uma das técnicas foi de extrema importância para a compreensão dos principais parâmetros e variáveis de cada ferramenta, ponto essencial na procura do melhor conjunto de dados de configuração de utilização de cada algoritmo.

A utilização dos três tipos de classificadores tem como objectivo responder a um parâmetro importante da etapa de pós-processamento do processo de KDD, e que diz respeito à comparação de vários métodos, procurando o que se verifica com melhor comportamento face a casos reais futuros derivados do *case study* proposto.

III Implementação e **Resultados**

5. Trabalho realizado – Método

O presente capítulo tem como objectivo, apresentar sucintamente o problema, descrever os principais conceitos utilizados, bem como o de caracterizar o *data set* que será utilizado, e a própria ferramenta. Serão ainda evidenciados os princípios recolhidos junto da empresa, principais preocupações e assumpções que servirão de base ao trabalho desenvolvido.

5.1. Descrição do Problema

Como foi já sucintamente descrito na secção de objectivos, a finalidade deste trabalho baseia-se na construção de modelos distintos de *churn* consoante o perfil de consumo dos clientes. Esta caracterização deverá ser desenvolvida com base no perfil de compras e no perfil demográfico dos consumidores.

Em estudos anteriores realizados na companhia, foram determinados diferentes segmentos de clientes com base na sua frequência de compra e valor de compras por cada visita efectuada pelo cliente. Esta segmentação é semelhante à segmentação RFV (*Recency Frequency Value*), no entanto não é considerada a variável de *Recency* que mede a distância temporal à última compra do cliente. Estes mesmos segmentos são recalculados numa base periódica de 8 semanas, e são denominados de Loyal, Frequent e Occasional. Na Tabela 1, estão descritas as características principais destes segmentos.

Nº de Visitas em 8 semanas		Mensal <= 3	Quinzenal 4 a 6	Semanal 7 a 9	+1x semana 10 a 19	Diário >= 20
Compra média		OCCASIONAL		FREQUENT		LOYAL
V Small Shop < €10						
Small Shop €10 a €19						
Medium Shop €20 a €30						
Big Shop €31 a €48						
Large Shop €49 a €79						
Very Large Shop >= €80						

Tabela 1 - Caracterização dos diferentes segmentos “FV” de Clientes

Como é facilmente percebido, um dos principais objectivos de qualquer empresa que disponha deste tipo de segmentação deve ser o de movimentar os clientes no sentido Loyal, ou seja “forçar” os clientes a visitarem a loja mais vezes e comprar em maior valor em cada visita. Para este efeito, devem ser realizadas acções específicas a cada segmento de clientes de forma a “cativar” e fidelizar os clientes nos segmentos mais à esquerda. No entanto, e não sendo esse o tema em discussão deste trabalho, vão assumir-se estes segmentos como aceites e estáticos (não obstante o facto do cliente poder movimentar-se bimensalmente entre eles).

5.1.1. Definição da Target

Dado o objectivo proposto para este trabalho, e uma vez que não foi ainda desenvolvido pela Sonae MC nenhum modelo de predição de Churners, não existe a partida nas bases de dados da empresa, uma variável associada aos clientes que indique que um cliente é *churner* ou não. A variável target essencial, que dá ao sistema capacidade de diferenciar um cliente activo de um cliente que abandona a empresa, deve então ser criada baseada nas definições e regras do negócio.

Não obstante a não existência de modelos já previamente desenvolvidos, e derivado também à dificuldade de definir o que é o “Abandono” no mercado retalhista, tornou-se

então necessário usufruir do conhecimento de negócio, extremamente importante neste tipo de trabalhos para a definição da Target.

Dados os diferentes perfis definidos pelos segmentos “FV”²⁴, e uma vez que se sustentam na recorrência com que um cliente efectua as suas compras, e acreditando que essa mesma frequência se revela um factor importante na definição da variável Target, enunciou-se a relevância da criação de modelos distintos para os diferentes segmentos, dado que também a variável Target será definida de formas distintas.

Ficou então definido que a Target (“Churn” ou “não Churn”), ficaria definida por:

1. **Clientes Loyal** (cliente semanal a diário): Prever $Y = 1$ (cliente é *churner*) se não efectua compras no mês seguinte;
2. **Clientes Occasional + Clientes Frequent** (cliente quinzenal a mensal): Prever $Y = 1$ (cliente é *churner*) se não efectua compras nos 3 meses seguintes.

Pretende ainda comparar-se estes dois modelos com um modelo de *churn global*, onde serão considerados todos os clientes do universo, e o segmento “FV” será uma variável (prevê-se que se demonstre explicativa). Para este modelo, e dada a representatividade dos clientes Occasional e Frequent (~90%), a variável ficará definida a 3 meses.

Dado que a segmentação “FV” é refrescada a cada 8 semanas, e assumindo a possível movimentação dos clientes entre segmentos, e uma vez que se torna necessário categorizar os clientes apenas com um segmento, procedeu-se a um pequeno trabalho de pré-processamento que consistiu muito resumidamente no seguinte:

- 1) Para a base de clientes foi recolhida a amostra dos últimos 12 meses de histórico de segmentação “FV”;
- 2) Ao segmento Loyal foi atribuído o valor **3**, ao segmento Frequent o valor **2**, e ao segmento Occasional o valor **1**;
- 3) Cada cliente recebe no final o valor da moda da série recolhida para 3 meses;
 - a. No caso de ser “bimodal”:
 - i. Se o sentido do movimento for positivo, aplica-se o segmento de maior valor;

²⁴ FV = Frequency & Value Segmentation

- ii. Se o sentido do movimento for negativo, aplica-se o segmento de menor valor;
 - iii. Se o sentido do movimento for nulo, aplica-se o valor do último segmento observado.
- 4) A cada cliente ficou então atribuído um único segmento “FV” de valor 3 (Loyal), 2 (Frequent) ou 1 (Occasional).

5.2. Software utilizado – *SAS Enterprise Miner*

Este projecto foi desenvolvido utilizando a ferramenta SAS Enterprise Guide e Enterprise MinerTM. Este software é utilizado em todos os sectores de actividade e para todas as necessidades: redução de custos (optimização de campanhas de marketing), domínio do risco, fidelização (antecipação das partidas, escolha de clientes a fidelizar, diferenciação das ofertas), ou prospecção (determinação de pessoas que interessem, ou potenciais).

O *Enterprise Miner* é uma solução integrada, que, através de um interface totalmente gráfico, dá ao utilizador o acesso a diferentes etapas da metodologia SAS Institute para o *Data Mining*: a metodologia *SEMMA* (oportunamente explanada na secção 3 deste trabalho). Esta metodologia, baseada na experiência adquirida pelos consultores de SAS, decompõe-se em 5 etapas: Sample; Explore; Manipulate; Model; Access.

Apoiando-se sobre o Data Warehouse, o Enterprise Miner é capaz de tratar quaisquer dados, qualquer que seja a origem. Respeitando a metodologia Data Mining da SAS, garante ao utilizador a coerência do processo de análise.

Enterprise Miner integra, na mesma interface, todas as funcionalidades necessárias à preparação de dados, etapa indispensável antes da modelização. Ao propor pôr em paralelo e comparar várias técnicas de modelização, garante o resultado óptimo. Ao dar a possibilidade ao especialista de visualizar os resultados sobre forma financeira e de fazer simulações, permite tomar a decisão mais correcta.

Por fim, ao dar as ferramentas necessárias à industrialização desta decisão, com a ajuda de scripts completos podendo ser utilizados em *Data Warehouses*, o *Enterprise Miner*TM permite pôr em prática as soluções encontradas pelos projectos.

Relativamente ao SAS Enterprise Guide, pode resumidamente dizer-te que é “uma ferramenta OLAP para Windows, orientado por projectos, e que possibilita acesso rápido a uma grande parte da potencialidade analítica do SAS para estatísticos, analistas de negócios e programadores SAS”²⁶.

Nesta ferramenta efectuaram-se os primeiros trabalhos de pré-processamento e análise descritiva dos dados.

5.3. Caracterização do *data set*

Nessa etapa, são apresentados os dados envolvidos na elaboração das variáveis referentes à previsão de *churn*, são também enfocadas as actividades realizadas para conhecimento, identificação de problemas e detecção do subconjunto de dados de interesse para este estudo.

O estudo começou com o levantamento, nos sistemas da empresa, de todas as variáveis disponíveis, passíveis de relacionamento com o *churn* e/ou com o comportamento do cliente quanto à sua tendência ou não de continuar a comprar nas lojas da empresa.

Este levantamento foi levado a cabo envolvendo vários factores como: avaliação dos dados disponíveis, entendimento do negócio, e posterior selecção dos atributos que se considerou como sendo os mais explicativos do comportamento de compra dos clientes.

Assim, e no processo de Data Mining proposto para os modelos que serão desenvolvidos, serão utilizados dados compostos de diferentes tabelas existentes no DataWarehouse e CRM da SonaeMC, para formação do *data set* final utilizado para treino dos modelos.

Os dados extraídos e consolidados dos sistemas da empresa podem ser divididos em:

- 1) **Informação Socio-Demográfica:** contém dados referentes ao cadastro do cliente na empresa, como a idade, o género, o número de elementos do agregado familiar, a morada, o índice de poder de compra e o índice concorrencial associado à sua loja de preferência;
- 2) **Informação Comportamental:** que agrupa todo o tipo de informação relacionada com o consumo do cliente, no que respeita a bens ou serviços (como promoções direccionadas), bem como as *flags* associadas a segmentos de clientes previamente definidos na empresa.
- 3) **Informação de Envolvimento:** contém os dados relativos aos dados agregados de compra do cliente, nomeadamente informação de carácter temporal e que permitirá a definição da própria variável Target.

A extracção e consolidação dos dados e posterior construção da ABT necessária para a utilização do Software SAS, deu-se basicamente pela solicitação através de linguagem SQL aos sistemas da empresa.

A estratégia de escolha dos registos foi a de escolher 13 meses completos sendo tirada a fotografia de status, activo ou inactivo (comprou ou não comprou), no 13º mês (no caso dos clientes Loyal) ou nos 3 meses imediatamente seguintes nos casos dos clientes Occasional e Frequent.

Dessa forma, a massa de dados foi escolhida entre os meses de Abril de 2010 a Março de 2011, tendo sido definida a Target, e pela lógica descrita na secção 5.1.1, para o mês de Abril de 2011, ou no caso do *data set* a treinar no modelo Occasional + Frequent, para os meses de Abril + Maio + Junho de 2011. Uma vez que o estudo será reflectido sobre o histórico de 12 meses a Março de 2011, foram considerados apenas os clientes com pelo menos uma compra realizada nesse mês.

Neste estágio inicial, e derivado da experiência de negócio, sentiu-se a necessidade de expurgar do *data set* alguns registos de que se tem conhecimento poderiam provir os dados de muito ruído. Assim, foram desconsiderados os registos relativos a Colaboradores e clientes do Arquipélago dos Açores.

As variáveis extraídas dos sistemas e que serão submetidas aos trabalhos de pré-processamento e consequentes de todo o processo de Data Mining estão descritas na secção 1 do Anexo III.

5.4. Metodologia utilizada

As etapas abrangidas na elaboração deste projecto são descritas em função da metodologia proposta em CRISP (2000) e da metodologia SEMMA directamente relacionada com a ferramenta utilizada (SAS Enterprise Miner). Os principais objectivos da metodologia aplicada são:

- 1) Transformação necessidades de negócio em tarefas de Data Mining;
- 2) Sugestão de transformações nos dados e técnicas adequadas ao estudo do problema;
- 3) Utilização de métricas para a avaliação da qualidade dos resultados obtidos;
- 4) Documentação de todo o projecto.

A evidente importância no que respeita à compreensão do problema e dos dados e à implementação do modelo resultante, e uma vez que a metodologia SEMMA (embebida no software utilizado) possui essa deficiência, fez com que neste trabalho se tenha optado pela procura de conjugação das duas metodologias apresentadas.

5.4.1. Preparação da Base de Dados de trabalho

Uma vez que a tabela de dados preparada (ABT²⁵) para integrar o software SAS contém todos os clientes, não estando em primeira instância limpa de registos que podem ser considerados como ruído para a análise, foi necessário proceder a uma primeira fase de limpeza e pré-selecção dos dados a analisar. Esta fase foi executada em SAS Guide.

²⁵ ABT – Analytical Base Table

Seleccionaram-se então, através de cláusulas “*where*” em SQL aos seguintes filtros para criar o grupo de clientes elegíveis:

1. **Clientes com compras em Março 2011** – exigido pela definição da Target nos dois grupos de clientes;
2. **Clientes com comportamentos atípicos face à base global de clientes** – tendo como base o conhecimento do negócio, sabe-se que existem alguns conjuntos de clientes, que pelas suas características têm um comportamento muito distinto da base geral de clientes, podendo incutir ruído na amostra dos dados. Assim, e sendo que estes grupos representam uma pequena percentagem do total da base, optou-se por retirá-los em primeira instância:
 - a. **Colaboradores** – clientes que são colaboradores da empresa (dispõem inclusive de condições especiais de utilização do cartão);
 - b. **Grossistas** – clientes com comportamento de compra *grossista*;
 - c. **Clientes Açores** – clientes que compram nas lojas dos Açores;
 - d. **Duplicados** – existe um conjunto de clientes que dispõe de mais do que uma conta cartão, optou-se então por eliminar a conta mais antiga destes, uma vez que uma agregação de dados nesta fase iria incutir ao projecto tempo de processamento adicional.

Posteriormente a esta limpeza inicial de dados em primeira instância ter-se-á que subdividir os registos do *data set* entre Loyal e Occasional + Frequent. De hora em diante estes dois grupos de clientes serão denominados de L (Loyal) e OF (Occasional & Frequent).

Após esta partição, procedeu-se então à criação da variável Target (*Churn* ou *não Churn*). Esta criação efectuou-se através do cruzamento da ABT de Março já filtrada pelas condições referidas acima, com a ABT de Abril, em que:

Clientes Loyal (L):

- Cliente com valor de compras > 0 em Abril → Target = 0 (*não churn*)
- Cliente com valor de compras = 0 em Abril → Target = 1 (*churn*)

Clientes Occasional & Frequent (OF):

- Cliente com valor de compras > 0 em Abril ou Maio ou Junho → Target = 0 (*não churn*)
- Cliente com valor de compras = 0 em Abril e Maio e Junho → Target = 1 (*churn*)

Os conjuntos de dados resultantes destas primeiras operações de filtragem ficaram então com as seguintes dimensões:

Target	Descr Target	OF data		L data	
0	<i>Não Churn</i>	1 677 437	94.86%	157 353	99.19%
1	<i>Churn</i>	90 898	5.14%	1 285	0.81%
Total		1 768 335		158 638	

Tabela 2 - Contadores dos *dataset* após a primeira filtragem e partição

Através desta breve análise, identificou-se de imediato que se trata de uma base de dados não balanceada, onde apenas 5% dos registos relativos aos clientes OF e apenas 1% dos clientes L, correspondem à classe *Churn*. Este desbalanceamento terá que ser considerado quer no momento da modelação, quer com maior rigor nos trabalhos de pré-processamento.

Estes trabalhos de pré-selecção e partição, podem ser consultados na secção 1.2 do Anexo III.

Dada a partição tipológica efectuada sobre os dados globais disponíveis, o estudo recairá apenas sob o conjunto de clientes Loyal.

5.4.2. Compreensão dos Dados

Essa fase compreende a identificação da informação que possa ser relevante para o estudo e uma primeira familiarização com o conteúdo, descrição, qualidade e utilidade dos dados.

O primeiro passo na etapa de reconhecimento de uma base de dados (que será efectuado em SAS Guide) é analisar as medidas descritivas da mesma, tanto da base como um

todo, como dos grupos de clientes separados. As métricas utilizadas nesta análise foram a média, desvio padrão, mínimo, máximo, quantidade de registos e percentual dos mesmos, abrindo esta informação por classe (*churn* ou *não churn*).

Uma vez que a base de dados inicial foi particionada em Loyal e Occasional & Frequent, a análise descritiva dos dados e posteriores trabalhos de pré-processamento será realizada individualmente.

Variáveis Socio-Demográficas

Analizando os histogramas e medidas de dispersão das variáveis Socio-Demográficas (Anexo III), verifica-se que à excepção da variável COMPETITIONINDEX, todas as variáveis provêm os dados de *missing values*. Para variáveis como SEGM_AGE ou HOUSEHOLD, o tratamento inicial aquando da extracção dos dados determinou que os valores *omissos* assumiriam um segmento específico, S*_0.

Pode ainda notar-se por observação do histograma e box-plot da variável IPC, a existência de *outliers*. Este dado deverá ser tomado em conta aquando das etapas de pré-processamento.

Ainda e avaliando na diagonal a distribuição das frequências das variáveis Socio-Demográficas categóricas, pode brevemente concluir-se que:

1. Clientes sem dados de morada (MUNICIPALITY_CD ou DISTRICT_CD) que assumem nos dados os valores de *missing values* para estas variáveis, são os clientes cuja representatividade no conjunto *Churn* (Target = 1) é superior. Isto pode levar a inferir que:
 - a. Como os clientes não são contactos via mailing (por não terem morada na base da dados) são mais propensos a abandonar a relação com a companhia;
 - b. Clientes novos ou cujo cupão de adesão não foi ainda processado ou não entregue, são clientes com uma relação mais “fraca” com a organização, pelo que mais facilmente se tornam *churners*;

Estas questões poderão ser avaliadas de forma mais assertiva aquando da exploração dos modelos. Assim, e uma vez que esta variável parece ser, em primeira instância

relevante, não deverão ser desconsiderados os *missing values*, mas sim tratados para que possam ser *lidos* pelos algoritmos dos modelos.

Um facto que se encontrou de algum interesse nesta descrição de dados foi que aparentemente, existem diferenças significativas entre os clientes *churn* e *não churn*, no que respeita às variáveis SEGM_AGE e HOUSEOLD. Como pode verificar-se na Figura 13 e Figura 14, a distribuição dos clientes *churn* distinguem-se da base (cálculo através do Index face à base²⁶) como sendo clientes mais novos e com um agregado familiar mais reduzido.

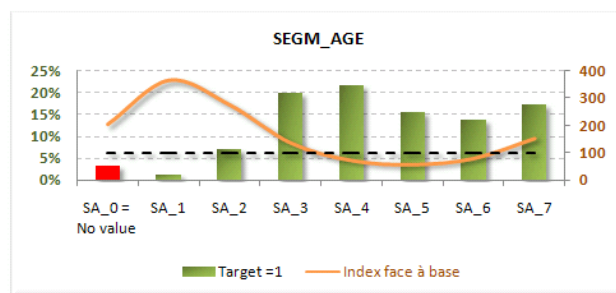


Figura 13 - Variação da distribuição da variável Segm_age de Target=1 face a base

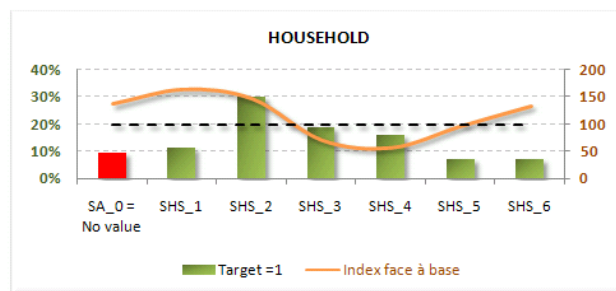


Figura 14 - Variação da distribuição da variável Household de Target=1 face a base

Variáveis Comportamentais

Por análise dos gráficos de distribuição das frequências das variáveis categóricas comportamentais como SEGM_VALUE e SEGM_LIFESTYLE, denotam-se algumas diferenças significativas entre os clientes *churn* e *não churn*. Através da mesma análise, e pela identificação de um elevado número de outliers, optou-se por expurgar da modelação as variáveis: TOP1CAT_12; TOP2CAT_12; TOP3CAT_12; TOP1CAT_24; TOP2CAT_24; TOP3CAT_24. A variável SEGM_RECENCY será também rejeitada, uma vez que não tem significado para a modelação.

²⁶ Index face à base: diferença entre a distribuição percentual por grupo face à distribuição existente no dataset

Através da análise descritiva das variáveis comportamentais numéricas (ANEXO III.1.3.1.1.), foram tomadas algumas decisões relativamente à rejeição de algumas variáveis. Esta decisão foi tomada pelo elevado número de *missing values* e inconsistências nos dados (Tabela 19).

De notar que para as variáveis comportamentais relativas ao gasto nas categorias de produto, não existem *missing values*, isto porque a ausência de valor de compras assume-se como o valor 0. Desta forma, foram também rejeitadas as variáveis em que a proporção de *zeros* fosse superior a 25% (valor do primeiro quartil = 0).

Pelo entre os Coeficientes de Variação das duas classes (*churn* e *não churn*), pode verificar-se uma maior variabilidade nos dados no conjunto *churn*. Esta variabilidade evidencia-se nas variáveis: GROCERIES_FLG_VISIT_LAST12M, DRINKS_FLG_VISIT_LAST12M, HYGIENICPRD_FLG_VISIT_LAST12M, FRTSVGTBLS_FLG_VISIT_LAST12M, e DAIRY_FLG_VISIT_LAST12M.

Da observação das diferenças entre a *amplitude semi-interquartil*, pôde também verificar-se que nas variáveis GROCERIES* e DRINKS*, existem mais observações próximas da média nos clientes *não churn*. Este pode ser um factor indicativo de que nestas categorias de produtos, existe uma maior variabilidade nos clientes *churn*.

Identificou-se ainda a presença de outliers em algumas das variáveis, não só pela presença de valores negativos em variáveis referentes a compras, como também por observação dos correspondentes Box-Plot (ANEXO III.1.3.1.1.).

Pode ainda verificar-se nos histogramas das variáveis (e.g.) HEALTHDP_INDEX_3M_VISITAS, DIETETICS_CAT_6M_VL e FISH_CAT_6M_VL que os eixos apresentam grande amplitude existindo uma grande concentração de observações à direita ou à esquerda desses mesmos eixos. Esta análise aponta também valores máximos bastante discrepantes, podendo significar a presença de *outliers*. Para dar-se continuidade às comparações entre as variáveis faz-se necessário normalizá-las, visto que apresentam ordens de grandeza bastante diferentes.

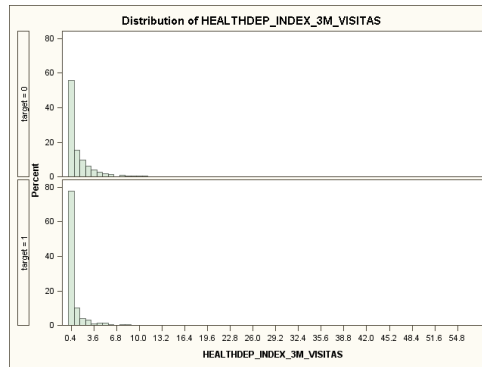


Figura 15 - histograma da variável Hygienicprd_ratio_cat_3_6_vl

Através desta primeira análise descritiva observa-se que para algumas variáveis existe uma clara separação entre os clientes *churn* e *não churn* (como por exemplo nas variáveis TX_RSP_M - Figura 17- e HYGIENICPRD_CAT_6M_VISITAS - Figura 16).

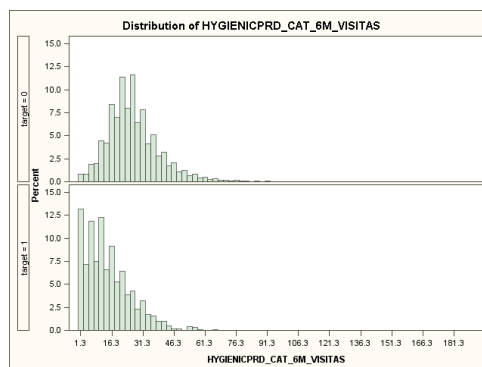


Figura 16 - histograma da variável Hygienicprd_ratio_cat_6m_visitas

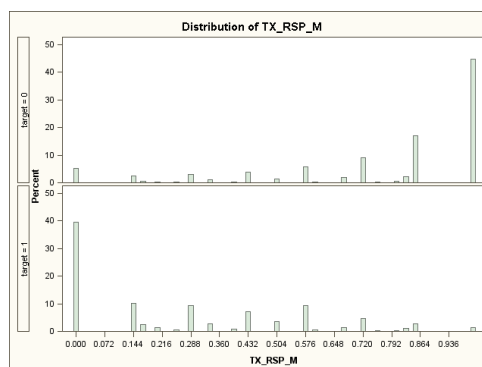


Figura 17- histograma da variável Tx_rsp_m

Variáveis de Envolvimento

Pela análise das medidas de dispersão, histogramas e Box-Plot das variáveis de envolvimento (ANEXO III.1.3.1.3.), pôde decidir-se quanto à rejeição de variáveis cujo conjunto de valores não se mostra relevante para a construção dos modelos. Nestas

condições encontram-se as variáveis referentes ao montante de descontos nos períodos considerados (DEVOLUC_*) e as variáveis que dizem respeito ao número de lojas visitadas (LOJAS_ACTUAL_NR).

Uma vez que as variáveis relativas ao consumo em outras áreas de negócio que não o alimentar foram rejeitadas no estudo das variáveis comportamentais, optou-se pela rejeição das variáveis referentes à proporção do montante comprado nas insígnias MCH (MCH_*).

Por uma questão conceitual, as variáveis VISITAS*²⁷ e TRANSAC*²⁸ são muito similares. Desta forma, e uma vez que as variáveis TRANSAC* são providas de mais *outliers* (Figura 18) optou-se por rejeitar esta.

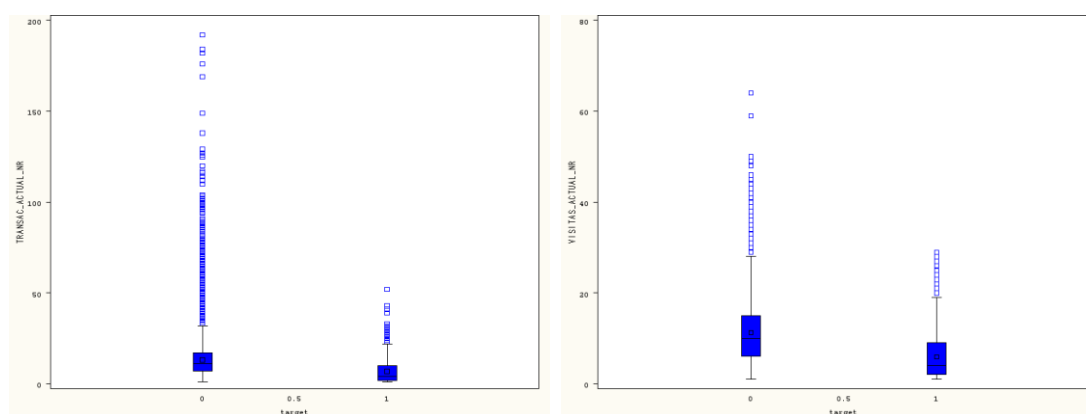


Figura 18 - Box-Plot das variáveis Visitas e Transacções no mês actual

Através da observação das distribuições das variáveis relativas às proporções de descontos, e dado que por definição a utilização das nove variáveis, pouco acrescentará ao modelo, dever-se-á estudar a correlação das variáveis entre si e destas com a *target*. Ainda, e como conceitualmente a compra média é calculada com base no gasto global e nas visitas efectuadas no período, sabe-se de antemão que estas três variáveis poderão estar correlacionadas.

De forma a sustentar a hipótese levantada acima, foi calculado o Coeficiente de Correlação de Pearson (r). Os resultados obtidos podem ser consultados na Tabela 21 (ANEXO III.1.3.1.3.1).

²⁷ VISITAS* - Número de visitas num determinado período

²⁸ TRANSAC* - Número de transacções num determinado período

Sobre as variáveis com um valor de $|r| \geq 0.8$ (correlação forte) foi analisada a correlação com a *target*. No entanto, e uma vez que nenhuma das variáveis analisadas possui *per-si* uma correlação forte com a *target* (Tabela 3), optou-se por não rejeitar nenhuma nesta fase.

	GASTO_ACTUAL_VL	TICKET_MED_ACTUAL	VISITAS_ACTUAL_NR	GASTO_ULT_3M_VL	TICKET_MED_ULT_3M	VISITAS_ULT_3M_NR	TRANS_DESCONTOS_3M_VL_PR	GASTO_ULT_6M_VL	TICKET_MED_ULT_6M	VISITAS_ULT_6M_NR	DESCONTOS_ULT_6M_VL	PROD_DESCONTOS_6M_VL_PR	PROD_PROMOCAO_6M_VL_PR	TRANS_DESCONTOS_6M_VL_PR	GASTO_ULT_12M_VL	TICKET_MED_ULT_12M	VISITAS_ULT_12M_NR	DESCONTOS_ULT_12M_VL	PROD_DESCONTOS_12M_VL_PR	PROD_PROMOCAO_12M_VL_PR	TRANS_DESC_12M_VL_PR	CATEGORIAS_LAST6M_NRS	CATEGORIAS_LAST12M_NRS
r (target)	-0.09	-0.03	-0.07	-0.08	-0.03	-0.06	-0.07	-0.08	-0.02	-0.06	-0.06	-0.02	-0.03	-0.06	-0.08	-0.02	-0.05	-0.05	-0.01	-0.02	-0.07	-0.11	-0.10

Tabela 3 - Correlação com a *target* (variáveis de envolvimento)

Pode ainda comprovar-se pela observação dos Box-Plot das restantes variáveis, que a amostra está provida de *outliers* praticamente em todas as variáveis. Este é um factor que terá que se ter em conta na etapa de pré-processamento.

5.4.3. Etapas de Pré-processamento

Após a análise descritiva da *data set*, a fase de Pré-processamento é das mais importantes em todo o processo de modelação, para que os resultados obtidos sejam providos do mínimo ruído possível.

Na fase anterior de compreensão dos dados foram rejeitadas 59 variáveis das consideradas inicialmente. Na Tabela 22 encontram-se as variáveis pré-seleccionadas.

Os trabalhos de pré-processamento seguintes foram desenvolvidos em SAS Guide, e SAS Miner, conforme as especificidades do problema.

5.4.3.1. Tratamento de Missing Values (valores omissos)

O tratamento dos valores omissos, é de fundamental importância quando se faz uso da regressão logística e de redes neuronais. As árvores de decisão tratam os valores *missing* como válidos e, portanto, não necessitam desse tratamento. A técnica utilizada para as substituições foi a de alocar o valor ‘-1’ que equivale a “Sem Informação” para as variáveis

qualitativas, e **a média** para as variáveis quantitativas, à excepção das variáveis TX_RSP_M e PROD_DESC_ACTUAL_NR_PR nas quais foi assumido o valor 0 nos valores omissos. Esta operação foi efectuada em SAS Guide.

5.4.3.2. Tratamento ruído e inconsistências nos dados

Por observação da Tabela 19, e restantes análises descritivas das variáveis em estudo, verifica-se a existência ruído e inconsistências na informação. Por definição sabe-se que as variáveis relativas a proporção deverão ter uma distribuição entre [0;1], e que todas as variáveis respectivas a gasto ou número de compras não poderão assumir valores negativos.

Este tratamento foi efectuado em SAS Guide a par do tratamento de valores omissos.

5.4.3.3. Detecção de Outliers

A análise de *outliers* é uma das fases mais importantes do processo de pré-processamento, não só por dela depender o tratamento de valores extremos nos dados, mas também porque é necessário que seja acompanhada de uma análise qualitativa, para que não sejam eliminados ou substituídos valores que podem vir a mostrar-se relevantes nos modelos.

Neste âmbito, e tendo em conta a definição das variáveis utilizadas para o modelo, optou-se por não considerar como *outliers* todas as observações de valor 0 ou 1 nas variáveis relativas a proporções, e as observações de valor 0 para as variáveis representativas de gasto em valor ou em número de visitas. Posto isto, e dado que se procedeu ao tratamento de inconsistências, onde foram tratados os valores negativos, foram apenas analisados os outliers *à direita da média*.

Para uma breve avaliação e tomada de decisão acerca do tipo de tratamento a efectuar sobre os *outliers*, calcularam-se os *z-scores* das variáveis nas quais residia a suspeita de valores extremos. Adoptando a técnica dos *z-scores*, conclui-se que a eliminação dos registos com valores extremos ($z > +3.3$), reduziria a dimensão da amostra em 75%. O que, no caso dos clientes *churn*, poderia revelar-se problemático. Desta forma, optou-se

então por proceder à substituição dos valores extremos pelo valor $Q3+1,5L$ ²⁹ ($Q3$ =terceiro quartil; L =amplitude inter-quartil).

5.4.3.4. Construção de variáveis derivadas

Apesar do elevado número de variáveis existentes no banco de dados, torna-se nesta fase crucial a criação de novas variáveis a partir destas, de forma a tentar enfatizar certos aspectos dos clientes que podem revelar-se uma grande ajuda na definição do *churn*. Esse incremento na riqueza da informação pode realizar-se de inúmeras formas.

A forma mais comum, é a da criação de rácios entre variáveis correlacionadas de forma a acrescentar valor ao *data set*. No entanto, e uma vez que o conjunto de dados em estudo foi produzido exclusivamente com o propósito deste estudo, e ainda dada a experiência de negócio, a maior partes das variáveis foram criadas aquando da extracção dos dados do *Data Warehouse*.

Assim, e aparte a variável **TARGET**, já definida em secção anterior, optou-se nesta fase por criar apenas uma variável dependente do numero de respostas a promoções direccionadas (**N_RESPOSTAS_MC**) e do numero de promoções a que o cliente foi alvo nos últimos 12 meses (**N_CONTACTOS_MC**), criando assim uma variável que mede a **Taxa de Resposta (TX_RSP_M)** do cliente a este tipo de promoções.

5.3.3.5. Oversampling

Por norma, e mais especificamente em problemas relacionados com o *churn*, torna-se essencial um procedimento de *oversampling* para que o *data set* possa ser tratado por modelos de classificação de padrões.

Neste tipo de problemas, o número de instâncias classificadas como *churn* é muito menor (classe rara) que a classe *não churn*. Este facto provoca a impossibilidade de gerar um modelo com uma margem de erro baixa. Ainda, como o interesse em identificar as características dos clientes *churn* é maior, torna-se crucial minimizar a probabilidade de classificação incorrecta.

²⁹ $Q3+1,5L$: Método Box-Plot para detecção de outliers, combinado com o método **binning** para substituição de valores extremos. Anexo II.1.2.

No *data set* em questão as proporções entre as classes *churn* e *não churn* encontram-se totalmente desbalanceadas, onde apenas 5% dos registos relativos aos clientes OF e 1% dos clientes L, correspondem à classe *Churn*.

Esta proporção deve ser corrigida através de *oversampling* para um valor entre 10% e 50%, de modo a que os modelos a serem aplicados consigam distinguir as duas classes da variável *Target*.

A aplicação de tamanhos diferentes para as amostras, implica uma interpretação mais cuidada dos resultados dos modelos. Desta forma, optou-se por balancear e dada a reduzida dimensão do *sub-set* de registos da classe *churn* para o conjunto L, optou-se neste caso por aplicar uma percentagem de 80% *não churn* e 20% *churn*. Uma vez que o desbalanceamento entre as classes é muito grande, uma distribuição mais equitativa, poderia enviesar os resultados dos modelos.

Esta operação foi efectuada em SAS Miner através do nó “Sampling”.

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	target	TARGET	0	5196	80
TRAIN	target	TARGET	1	1299	20

Após avaliar os resultados dos modelos obtidos numa primeira fase da experiência, constatou-se que a consistência dos modelos ficava aquém das expectativas. Verificou-se nomeadamente que a escolha por qualquer dos modelos desenvolvidos ao longo das experiências, não apresentava ganhos significativos face à aleatoriedade (medido através do *lift* dos modelos). Após alguma análise, verificou-se que este facto se ficava a dever ao baixo número de instâncias traduzidas como *churn*. Uma vez que diferentes percentagens aplicadas em *oversampling*, não resolveram o problema detectado, dado que na população recolhida não existiam mais exemplos que pudessem ser classificados como *churners* Loyal, e por forma a evitar uma reformulação da variável *target*, optou-se por proceder a uma duplicação dos registos *churn* (*Bootsing*) de forma a “dar aos modelos” mais informação acerca do evento que se pretendia prever. Após esta

transformação os dados de treino (sujeitos a um *oversampling* de 80% - 20%) ficaram divididos com **2598** exemplos na *target* = 1 e **10392** exemplos na *target* = 0.

5.4.3.6. Amostragem

A operação de amostragem, de forma a separar os conjuntos de treino, teste e validação foi efectuada em SAS Miner através do nó “Data Partition” (Figura 19)

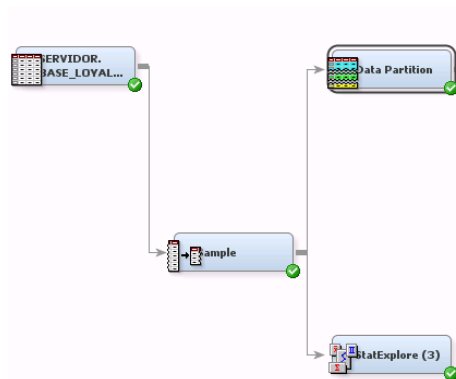


Figura 19 - Esquema inicial SAS Miner

Dado que se trata de um *data set* relativamente pequeno, optou-se por não reservar dados para a fase de teste. Parametrizou-se em 70% para conjunto de Treino e 30% para Validação. A partição foi efectuada pelo método “Simple Random” através do qual todas as instâncias do *data set* têm a mesma probabilidade de ser seleccionadas.

5.4.3.7. Transformação dos dados

O passo da transformação de dados é de grande importância para incrementar o desempenho dos modelos que serão executados. Assim como o processo de criação de variáveis derivadas, também esta fase depende directamente dos dados disponíveis.

Em problemas de *churn* como o caso em estudo, é muito importante proceder à análise da necessidade de transformações sobre os dados originais, uma vez que a variabilidade e os valores máximos e mínimos de cada variável podem ser muito distintos.

Analizando as estatísticas básicas das variáveis do *data set*, verifica-se que existem variáveis com distribuições altamente assimétricas. Esta conclusão pode ser extraída dos elevados valores de *skewness*³⁰, uma vez que pelo valor do coeficiente *Kurtosis*, se pode

³⁰ *Skewness*: Medida que caracteriza uma distribuição não normal. Próximo de zero indica simetria.

inferir que nenhuma das variáveis tem distribuição normal (*kurtosis* $\neq 3$). Estes valores podem ser revistos na Tabela 23.

Nestas distribuições uma pequena percentagem de observações pode ter grande influência sobre os dados, e nestes casos, as transformações podem melhorar em muito a resposta dos modelos, especialmente os de regressão logística, que não respondem tão bem para comportamentos não lineares que, no entanto, são muito bem tratados pelas redes neurais, por exemplo.

O SAS Enterprise Miner, disponibiliza ao utilizador uma panóplia de opções de transformação de variáveis em ordem à melhor relação com a variável de resposta (ou *target*). Uma vez que se pretendia proceder futuramente a uma selecção de variáveis, optou-se nesta fase por aplicar alguns dos métodos a todas as variáveis do *data set*, para que depois pudessem ser aceites ou rejeitadas pelo método r^2 ou **Qui-Quadrado**.

Das experiências realizadas, optou-se por escolher as transformações:

- “*optimal binning for relationship to target*” – faz automaticamente os agrupamentos testando os resultados em relação à variável resposta;
- “*Bucket*” - são criados intervalos dividindo os valores em intervalos igualmente espaçados com base na diferença entre o máximo e o mínimo;
- “*Maximize Normality*” - escolhe a transformação que produz quantis na amostra que estão mais próximos dos quantis teóricos da distribuição normal;
- “*Equalize*” – procede a várias transformações e escolhe a que minimiza a variância com a *target*.
- “*Standardize*” – standardização normal das variáveis;
- “*multiple*” – aplica múltiplas transformações que serão analisadas posteriormente em Selecção de Variáveis.

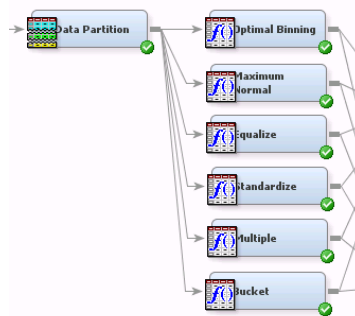


Figura 20 - Nó de transformação das variáveis

As transformações realizaram-se sobre todas as variáveis de *input*. O software aplica um prefixo com a sigla da transformação aplicada à variável ao *bias* do atributo. Ex: BIN_ALCDRNKS_IDX3MVIS, é o resultado da transformação por “optimal binning” a oito intervalos da variável ALCDRNKS_IDX3MVIS.

Na Figura 21 podem observar-se dois exemplos de transformações realizadas sobre as variáveis originais, em que o *sombreado* é a distribuição da classe *não churn*.

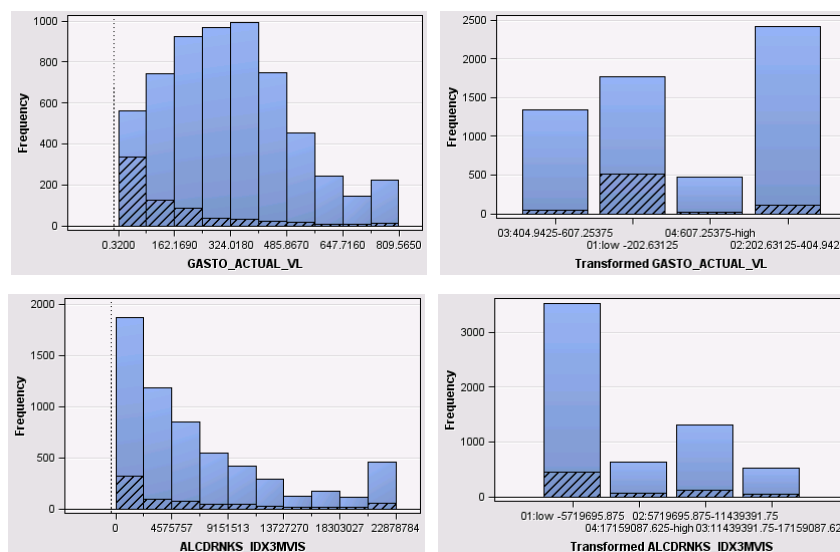


Figura 21 - Exemplo de transformação de variáveis

Outra transformação que se denota de elevado interesse é o mapeamento das variáveis categóricas com n classes em n variáveis *dummy*³¹ com um único valor não nulo que indica a que classe pertence o registo. Esta transformação é particularmente importante para os modelos de Regressão Logística, e para a avaliação da correlação das variáveis

³¹ Variável *dummy* – variável auxiliar. (exemplo: Se género do cliente = F então 1 ; Se género do cliente = M então 2; Se *null* então 0)

catóricas. No entanto, e uma vez que o software utilizado procede a esta criação automaticamente, optou-se por não realizar este passo.

5.4.3.8. Selecção de Variáveis

Esta fase é responsável por encontrar e reconhecer que variáveis presentes no *data set* contribuem mais significativamente para a definição de *churn*, sem que se perca informação importante para a modelação. Além de reduzir a dimensão da base de dados a analisar através da eliminação de variáveis com pouca ou nenhuma relevância para a caracterização da variável *Target*, ou mesmo variáveis redundantes, este procedimento também fornece ideias claras sobre quais os atributos que levam um cliente a deixar de comprar.

Sob o conjunto de variáveis resultantes do passo de transformação, foram testados 2 métodos de selecção de variáveis (*feature selection*): medida r^2 e o método Qui-Quadrado.

O SAS Enterprise Miner, disponibiliza um nó de Variable Selection, que permite definir os parâmetros de escolha, sobre os quais o algoritmo de selecção irá operar.



Figura 22 - Nó Attribute Selection by R-Square

O critério R^2 utiliza um método *stepwise* de selecção de variáveis. Neste procedimento é calculado o valor de R^2 (que mede a capacidade de discriminação de cada variável relativamente à *target*). O método foi parametrizado com o critério de paragem = 0.0005 (o que significa que o processo de selecção pára quando o valor global do R^2 não aumenta em mais de 0.0005 com a selecção da variável que está a ser testada. Foi ainda considerado o valor 0.05 como o “Minimum R-Square”³².

No critério **Qui-Quadrado**, a selecção de variáveis é efectuada utilizando *splits* binários entre as variáveis disponíveis, de forma a maximizar o valor da estatística Qui-Quadrado na tabela de contingência 2X2 gerada. As variáveis catóricas são

³² Minimum R-Square: valor mínimo a partir da qual a variável é elegível para o processo de selecção.

transformadas em variáveis *dummy*, e as variáveis numéricas são previamente divididas em categorias para serem posteriormente testadas. Foi neste critério considerado o valor de 3.84 para a estatística Qui-Quadrado ($p\text{-value} = 0.05$) como valor mínimo para a variável ser considerada no processo de selecção.

Uma vez que a execução deste método levou à rejeição de 155 variáveis, optou-se pelo método de selecção R^2 . No entanto, as variáveis SEG* que dizem respeito a segmentações de clientes previamente conhecidas do negócio, foram mantidas, mesmo com a recomendação de serem rejeitadas.

5.4.3. Modelação do Churn

Com o *data set* L pronto para a modelação e dividido em conjuntos de treino (70%) e validação (30%) realiza-se então a modelação e avaliação dos diversos modelos de classificação de padrões, na busca pelo modelo que realize a classificação mais adequada dos clientes e possua a melhor capacidade de generalização.

Como já foi referido, os modelos utilizados e testados no sistema desenvolvido foram:

- 1) Árvores de Decisão;
- 2) Redes Neurais;
- 3) Regressão Logística.

Os resultados obtidos por cada um dos modelos e antecedentes experiências de ajustamento dos parâmetros serão discutidos na secção 6 deste trabalho.

5.5. Conclusão

Neste capítulo foi apresentado e estudado o conjunto de dados adoptado para a modelação. Foram ainda analisadas descritivamente as variáveis que serão utilizadas

nos modelos criados, tendo sido dado especial ênfase a todo o trabalho de pré-processamento realizado sobre o conjunto de dados.

De notar a explicação da definição da variável Target, e ainda a diferenciação do *data set* original em dois subconjuntos de dados (L e OF). Foi ainda nesta fase decidido recair o trabalho de investigação apenas sobre o grupo de clientes Loyal.

6. Análise de Resultados

O desenvolvimento dos modelos é um processo altamente iterativo, e leva a inúmeros processamentos para que se chegue a algo que demonstre robustez e um alto poder de explicação, este número de iterações verificou-se mais exponencial no desenvolvimento de redes neurais e árvores de decisão, que são realmente processos baseados em iterações. Pequenos detalhes como os ajustes aos parâmetros dos modelos ou no que concerne à escolha das variáveis, podem conduzir a modelos robustos ou pelo contrário a modelos completamente instáveis.

Não obstante as iterações que foram levadas a cabo neste trabalho, serão apenas explanadas com detalhe, aquelas que conduziram a resultados satisfatórios.

6.1. Árvores de Decisão

Uma vez que as árvores de decisão categorizam as variáveis de entrada em intervalos, e dado ainda que este é um método que pode ser utilizado para a selecção de variáveis. Dado que do nó transformação de variáveis, resultou um grande número de variáveis categorizadas em intervalos, e uma vez que esta operação pode reduzir as divisões que a árvore poderá vir a considerar e assim resultar num modelo com uma performance inferior, optou-se para a aplicação deste modelo, utilizar como input o modelo com as variáveis transformadas e as variáveis originais, e sobre ele não efectuar qualquer operação prévia de selecção.

Para a aplicação dos algoritmos de Árvores de decisão em SAS Miner devem definir-se os seguintes parâmetros:

- Critério de divisão:
 - **Entropia:** A escolha do atributo de partição é concretizada tendo em conta o Ganho de Informação, que mede a eficácia de um atributo

classificar o conjunto de treino (qual o atributo que mais reduz a Entropia);

- **Gini**: Mede a impureza de um determinado nó, quando utilizado um atributo para divisão. Este índice contabiliza a proporção de observações em cada classe da variável dependente num nó relativamente ao total, isto é, ao nó raiz.
- **Qui-Quadrado ou Teste de Fisher** (para variáveis *target* intervalares): Tem por base os testes de Qui-Quadrado de Pearson numa tabela de contingência entre as categorias da variável dependente e as categorias das variáveis independentes.
- Profundidade máxima da árvore: dada a variabilidade presente nos dados, restringiu-se a profundidade máxima a **20** níveis, de forma a permitir alguma liberdade no crescimento da árvore.
- Número de *branches* (ramos): após alguns testes de comportamento optou-se fixar a **2 ramos** (árvore binária).
- Métrica de avaliação da árvore: uma vez que a *target* é uma variável categórica, fixou-se a medida de avaliação e selecção da melhor árvore como sendo o “Misclassification Rate”³³.

Utilizou-se o nó de comparação de modelos “Model Comparison”, disponível na ferramenta (Figura 23), para seleccionar o melhor modelo gerado pelas Árvores de Decisão.

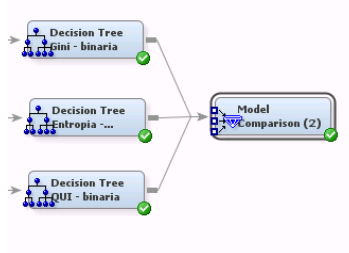


Figura 23 - Selecção do melhor critério de divisão AD

³³ Misclassification Rate: Mede a eficácia do modelo no que respeita à proporção de instâncias mal classificadas

O modelo seleccionado pelo algoritmo foi a Árvore de Decisão com critério de divisão Entropia que é o modelo no qual é atingido um menor valor de “Misclassification Rate” no conjunto de validação = 7.5%

Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Tree2	Decision Tree Entropia - binaria	0.075404
	Tree	Decision Tree Gini - binaria	0.079251
	Tree3	Decision Tree QUI - binaria	0.093357

A Figura 24 representa a Árvore com melhor performance. Não é possível visualizar os resultados de forma imediata, no entanto é possível verificar que a árvore atingiu uma profundidade de 13 níveis.

A primeira variável a ser testada foi a TX_RSP_M que diz respeito à taxa de resposta a promoções direccionadas, seguida da variável VST_ACTUAL que diz respeito à frequência de compra no mês actual.

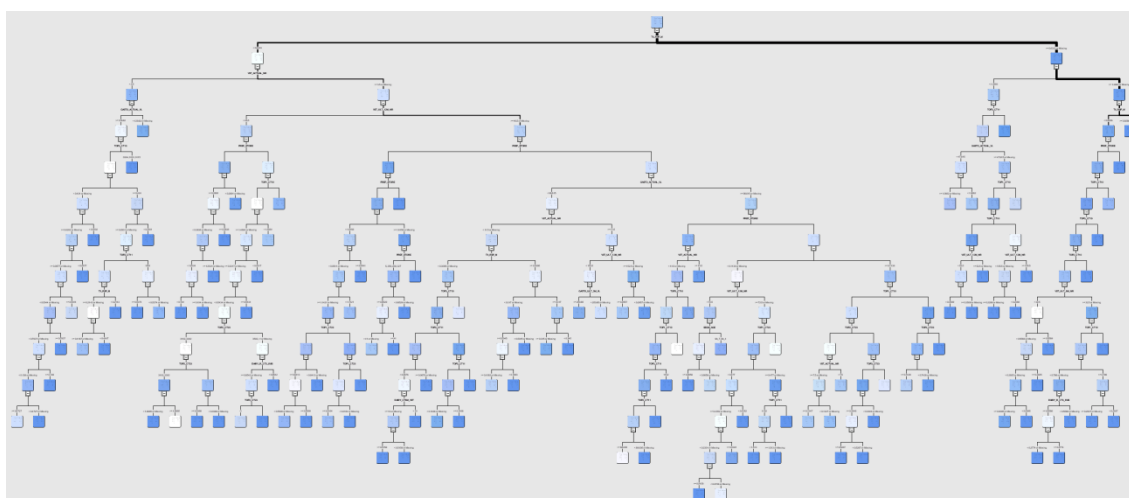


Figura 24 - Árvore de Decisão Entropia

Uma vez que a métrica de avaliação escolhida foi o “MisclassificationRate”, e dado que o menor erro (como se pode visualizar na Figura 25) no conjunto de validação é obtido para árvores com 117 folhas.

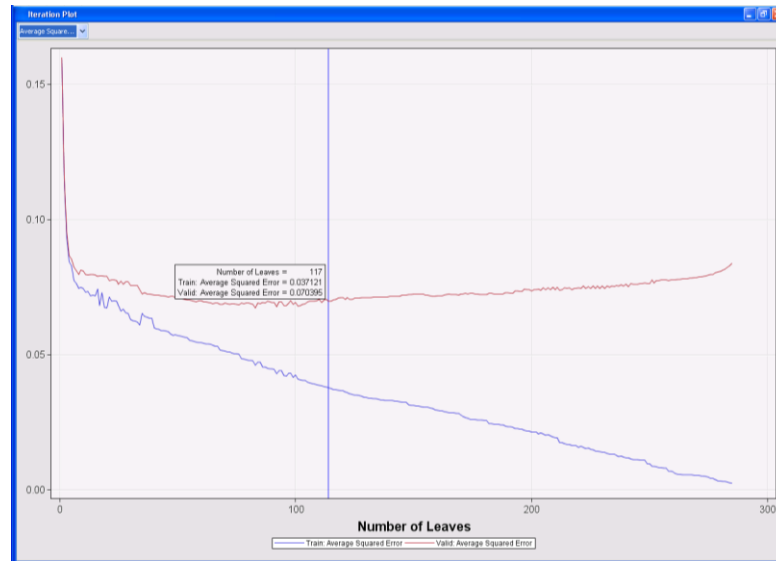


Figura 25 - Variação do Misclassification Rate X Nº Folhas – Entropia

A Tabela 4 apresenta a quantidade de clientes *churn*, obtida pelo modelo, por faixa de probabilidade P, dentro dos 3899 clientes da amostra de validação, o que representa o risco de evasão por faixa de score. Verifica-se que, quanto menor o score, menor o risco de *churning*, o que mostra o grande potencial do modelo na demonstração do risco dos clientes abandonarem. Pela tabela observa-se a probabilidade média de um cliente abandonar por faixa de probabilidade (P). Por exemplo, os clientes classificados na faixa de $90 < P < 100$, a probabilidade de abandono é de 99%.

Dados Conjunto Validação			
Probabilidade (P) de <i>churn</i>	% Clientes	Nº clientes	Probabilidade média de Churn
0.00-0.05	67%	757	0%
0.05-0.10	2%	26	8%
0.10-0.15	0%	2	13%
0.15-0.20	6%	70	16%
0.20-0.25	0%	0	.
0.25-0.30	1%	16	30%
0.30-0.35	0%	4	33%
0.35-0.40	0%	5	38%
0.40-0.45	0%	1	42%
0.45-0.50	3%	30	48%
0.50-0.55	0%	2	55%
0.55-0.60	0%	0	.
0.60-0.65	1%	10	62%
0.65-0.70	1%	6	69%
0.70-0.75	0%	3	75%
0.75-0.80	1%	11	79%
0.80-0.85	1%	10	81%
0.85-0.90	1%	16	87%
0.90-0.95	9%	104	92%
0.95-1.00	5%	60	99%

Tabela 4 - Quantidade e percentual de clientes por faixa de probabilidade de *Churn* (AD)

6.2. Regressão Logística

No caso do emprego dos algoritmos de Regressão Logística, e dadas as particularidades e susceptibilidades destes algoritmos, utilizou-se o *data set* com as variáveis transformadas, uma vez que a escolha do *data set* original implicaria que o modelo, de forma a compensar a desnormalização das variáveis, elegeria um maior número de variáveis, como mecanismo de compensação, inculindo complexidade desnecessária ao modelo.

No nó “Regression”, a ferramenta utilizada permite manipular os seguintes parâmetros:

- Método de selecção das variáveis:
 - **None (Main Effects):** por defeito todas as variáveis de input são candidatas a serem incluídas no modelo final
 - **Backward:** começa por considerar todas as variáveis (“efeitos”) e vai eliminando aquelas que estão menos correlacionadas com a *target*. Este critério não é recomendado para *targets* binárias, e por esse motivo não foi testado;
 - **Forward:** vai acrescentando variáveis correlacionadas com a *target*;
 - **Stepwise:** é similar ao método *Forward*, no entanto variáveis que tenham sido consideradas, podem ser rejeitadas.
- Critério de selecção (de eliminação ou adição): valor de significância de entrada fixo a **0.2** e de saída fixo a **0.05**. Após a realização de inúmeras iterações, estes foram os valores que conduziram a modelos com melhor desempenho.
- Critério de avaliação: foi utilizado o “Misclassification Rate”.

Foram então parametrizados 3 modelos distintos (Stepwise / Forward / Main Effects), e utilizou-se posteriormente o nó de comparação de modelos (Figura 26), de forma a seleccionar o melhor modelo gerado pelas Regressões. A medida de avaliação escolhida para que o software sugerisse o melhor modelo foi estatística o “misclassification rate” no conjunto de Validação.

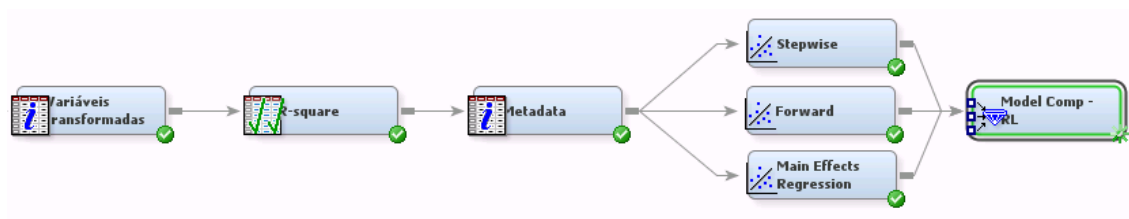


Figura 26 - Selecção do melhor método de Regressão Logística

Pelo nó de comparação de modelos, o modelo escolhido foi o modelo gerado por procura *Forward*, no entanto e por observação da Tabela 5, verifica-se que o modelo gerado por procura *Stepwise* apresenta melhores resultados no que respeita ao Índice ROC e à estatística de Kolmogorov-Smirnov, que aliás é uma das métricas comumente utilizada para a avaliação de modelos de Regressão Logística.

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Roc Index	Valid: Kolmogorov-Smirnov Statistic
	Reg2	Forward	0.0949	0.0730	0.9360	0.7320
Y	Reg	Stepwise	0.0954	0.0722	0.9370	0.7340
	Reg3	Main Effects Regression	0.0985	0.0738	0.9370	0.7480

Tabela 5 – Estatísticas Selecção dos modelos RL baseada no teste Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov (KS) é uma estatística não paramétrica para testar se as funções de distribuição de probabilidades de dois grupos são iguais. O valor do KS do modelo é a maior diferença entre as distribuições acumuladas das probabilidades dos grupos de clientes *churn* e *não churn*. O valor da estatística pode variar entre 0 a 1, sendo que quanto mais próximo de 1, maior o poder discriminatório do modelo ajustado.

Não obstante o facto de ser o modelo que apresenta uma maior taxa de não acertos (*misclassification rate*), por observação da Figura 27, verifica-se que a Regressão Stepwise é aquela que apresenta melhor desempenho no conjunto de validação (maior AUC). É de notar ainda que no que respeita ao valor do erro médio quadrático, este modelo é também o que apresenta valores mais satisfatórios. Desta forma, resolveu aceitar-se a sugestão de considerar este como sendo o modelo que melhor se adapta às características dos dados.

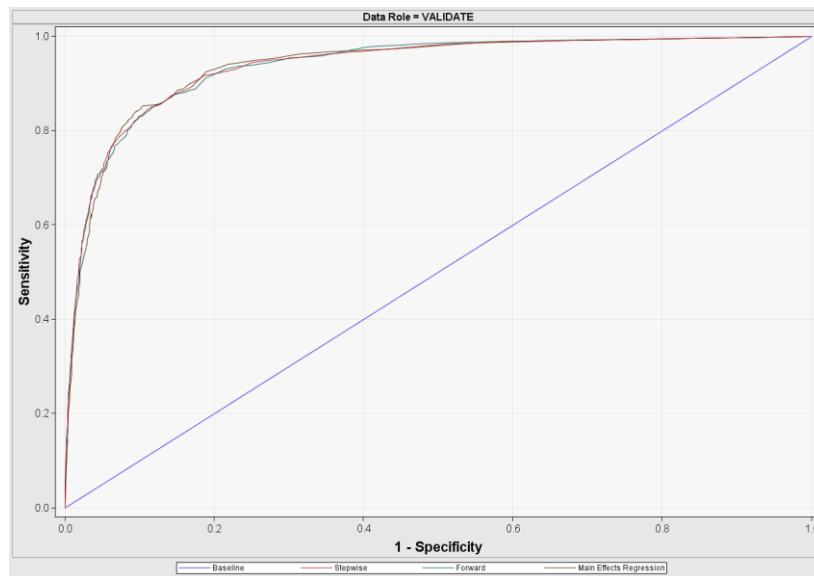


Figura 27- Gráfico ROC Regressão Logística

O modelo final fornece scores que quanto maior o valor obtido para os clientes, maior o risco do cliente ser *churner*. Como o score é uma probabilidade, o valor pode variar de 0% a 100%. Para obtenção do score de cada cliente é necessário atribuir o valor para cada uma das variáveis do modelo final de acordo com a equação de acordo com os coeficientes estimados resultantes do modelo Stepwise.

Depois de calcular as estimativas (coeficientes), o passo seguinte consiste em determinar se as variáveis independentes são associadas significativamente à variável resposta, através de um teste de hipótese – Teste Wald, e foram rejeitados todos os *effects*³⁴ com um valor de p-value > 0.05 (95% de significância).

Para testar a significância de cada variável em relação às demais, utilizou-se também o teste de Wald, cujas estatísticas relativas às variáveis escolhidas pelo modelo, podem ser consultadas na Tabela 6.

³⁴ Effects: pares variável, valor. Por exemplo no caso de uma variável suportar dois valores (1 ou 2), à variável estão associados dois *effects* (efeitos).

Variável	Wald Chi-Square	P-value
CULTURE_FLG_12M	8.5732	0.0034
EXP_CATEGORIAS_NACTUAL	59.7789	<.0001
FRTSVGTBLS_PRDL0J6M	28.8982	<.0001
HOUSEHOLD	28.1346	<.0001
PWR_TRANS_DESC_ACTUAL_VL	16.7982	<.0001
SEGM_AGE	23.5957	0.0013
SEGM_SPENDVARIATION	66.3096	<.0001
SEGM_VALUE	95.0877	<.0001
SQRT_BUTCHER_IDX3MVIS	19.7619	<.0001
SQRT_FRTSVGTBLS_CT6M_VL	25.4852	<.0001
SQRT_TRANS_DESC_12M_VL	64.1924	<.0001
TOP1_CT10	75.6326	0.0015
TOP1_CT23	59.4776	0.0003
TOP2_CT11	64.5798	<.0001
TOP3_CT10	57.5379	0.0828
TX_RSP_M	409.2839	<.0001
VST_ACTUAL_NR	161.4757	<.0001
VST_ULT_3M_NR	7.2667	0.007

Tabela 6 - Teste de Significância RL Stepwise

A Tabela 7 apresenta a quantidade de clientes *churn*, obtida pelo modelo, por faixa de probabilidade P, dentro dos 3899 clientes da amostra de validação, o que representa o risco de evasão por faixa de score. Verifica-se que, quanto menor o score, menor o risco de *churning*, o que mostra o grande potencial do modelo na demonstração do risco dos clientes abandonarem. Pela tabela observa-se a probabilidade média de um cliente abandonar por faixa de probabilidade (P). Por exemplo, os clientes classificados na faixa de $90 < P < 100$, a probabilidade de abandono é de 98%.

Dados Conjunto Validação			
Probabilidade (P) de churn	% Clientes	Nº clientes	Probabilidade média de Churn
0.00-0.05	67%	2604	0%
0.05-0.10	2%	88	8%
0.10-0.15	0%	7	13%
0.15-0.20	6%	242	16%
0.20-0.25	0%	0	.
0.25-0.30	1%	55	30%
0.30-0.35	0%	13	33%
0.35-0.40	0%	18	38%
0.40-0.45	0%	4	42%
0.45-0.50	3%	102	48%
0.50-0.55	0%	8	55%
0.55-0.60	0%	0	.
0.60-0.65	1%	33	62%
0.65-0.70	1%	21	69%
0.70-0.75	0%	11	75%
0.75-0.80	1%	38	79%
0.80-0.85	1%	36	81%
0.85-0.90	1%	55	87%
0.90-0.95	9%	358	92%
0.95-1.00	5%	206	99%

Tabela 7 - Quantidade e percentual de clientes por faixa de probabilidade de Churn Stepwise

6.3. Redes Neurais (RNA)

O processo de construção de um modelo por Redes Neurais, engloba etapas específicas, como a escolha da topologia (estrutura) da rede e do método de aprendizagem.

Faz parte da primeira etapa a escolha do número de nós (neurónios) de entrada, nós de saída, número de camadas escondidas, e número de nós escondidos. Para a segunda etapa é necessário seleccionar um algoritmo de aprendizagem que identifique um conjunto de pesos de conexão que melhor abranja os dados de treino e tenha a melhor precisão preditiva. Estes pesos, por norma são inicializados sob um conjunto aleatório de valores, que é aprimorado e modificado consoante o *feedback* acerca do desempenho na aprendizagem no modelo. Para uma topologia MLP³⁵ (que foi a utilizada neste projecto), o algoritmo de aprendizagem supervisionado utilizado é o algoritmo *Backpropagation*.

Adequando estas fases ao problema em estudo, foram testados modelos com os seguintes parâmetros e métodos:

- **Número de nós de entrada:** variáveis transformadas e pré-seleccionadas
- **Número de nós de saída:** apenas um (*target*) que assume o valor 0 ou 1
- **Número de camadas escondidas:** 3 ; 5 ; 7
- **Arquitectura da rede:** MLP – Multilayer Perceptron
- **Algoritmo de aprendizagem supervisionado:** *Backpropagation*
- **Critério de selecção do modelo:** “Misclassification Rate”
- **Outras características topográficas da rede:** rede com um grafo sem ciclos (rede directa). Ou seja, não existem conexões directas entre os nós de entrada e os nós de saída. Essa ligação é estabelecida apenas pelos nós das camadas escondidas.

Após a parametrização dos 3 modelos distintos (com 3, 5 e 7 camadas escondidas) utilizou-se o nó de comparação de modelos (Figura 26), de forma a seleccionar o

³⁵ MLP (Multilayer Perceptron) – Rede Neuronal do tipo feedforward multicamada.

melhor modelo gerado. A medida de avaliação escolhida para que o software sugerisse o melhor modelo foi estatística o “Misclassification rate” no conjunto de Validação.

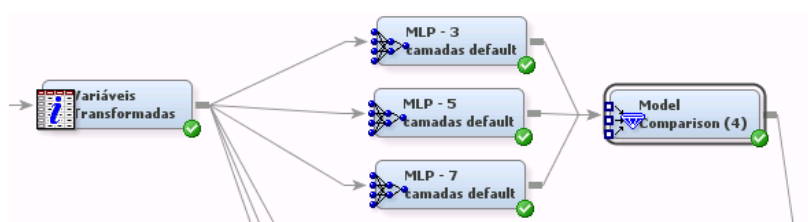


Figura 28 - Selecção do melhor método de Redes Neurais

O modelo escolhido foi o modelo gerado com **3 camadas escondidas**. Esta escolha pelo Taxa de acertos do modelo, fica validada quando consultada a Tabela 8. Nela podem ser observadas as principais estatísticas de avaliação de modelos. Como pode facilmente verificar-se pelos *high-lights* de formatação da tabela, o modelo RN3³⁶ escondidas é claramente mais adequado do que o modelo RN5³⁷ em todas as estatísticas à excepção da KS³⁸.

Estatística	Treino			Validação		
	3 camadas	5 camadas	7 camadas	3 camadas	5 camadas	7 camadas
Misclassification Rate	0.08	0.09	0.08	0.09	0.09	0.09
Lift	4.41	4.45	4.52	4.49	4.31	4.36
Kolmogorov-Smirnov Statistic	0.76	0.78	0.77	0.75	0.76	0.75
Average Squared Error	0.06	0.06	0.06	0.07	0.07	0.07
Roc Index	0.95	0.95	0.96	0.94	0.94	0.94
Cumulative Percent Captured Response	45.98	45.65	46.59	44.87	43.72	44.74
Percent Captured Response	22.06	22.28	22.61	22.44	21.54	21.79
Gain	359.4	356.1	365.4	348.60	337.07	347.32
Cumulative Lift	4.59	4.56	4.65	4.49	4.37	4.47
Mean Squared Error	0.07	0.09	0.08	0.07	0.07	0.07
Cumulative Percent Response	91.87	91.21	93.08	89.74	87.44	89.49
Percent Response	88.13	89.01	90.33	89.74	86.15	87.18
Number of Wrong Classifications	760	783	733	358	361	362

Tabela 8 - Estatísticas Modelos Redes Neurais

Comparando o modelo RN5 com o modelo RN7³⁹, verifica-se que existe uma proximidade acentuada entre o valor das estatísticas, dada esta proximidade a escolha recai naturalmente sobre o modelo com menor complexidade.

³⁶ RN3: Modelo gerado MLP com 3 camadas escondidas

³⁷ RN5: Modelo gerado MLP com 5 camadas escondidas

³⁸ KS: Kolmogorov-Smirnov Statistic

³⁹ RN7: Modelo gerado MLP com 7 camadas escondidas

		Previsto	
		0	1
Target	0	2 962	157
	1	201	579

Tabela 9 - Matriz de confusão Modelo RN3

A análise da matriz de confusão do modelo RN3 (Tabela 9) revela que este modelo tem alguma tendência para obter falsos negativos, i.e. para classificar como *não churners* indivíduos que de facto o são. Os *não churners* são, no geral, identificados correctamente pelo modelo, contudo no que respeita à classificação dos *churners*, o modelo não parece muito ajustado.

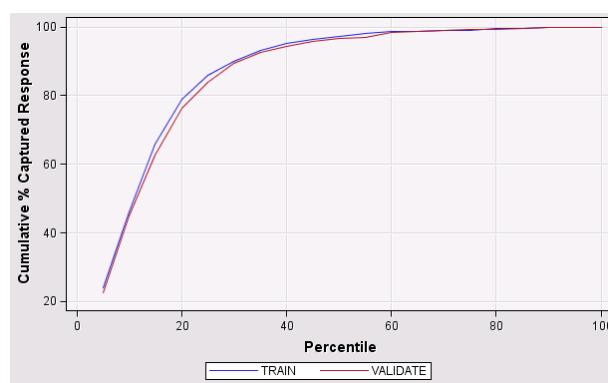


Figura 29 - Ganhos Cumulativos do modelo RN3

No gráfico de ganhos cumulativos (Figura 29) pode verificar-se que entre os 20% de clientes que o modelo RN3 identifica como tendo maior probabilidade de abandonarem estão cerca de 76% dos *churners* efectivos. Observando o gráfico da Figura 30 que representa o valor do *lift* para o modelo, por percentil 10, pode dizer-se que o valor anteriormente analisado se traduz num ganho de 2,76 vezes face a uma decisão aleatória (i.e. sem modelo).

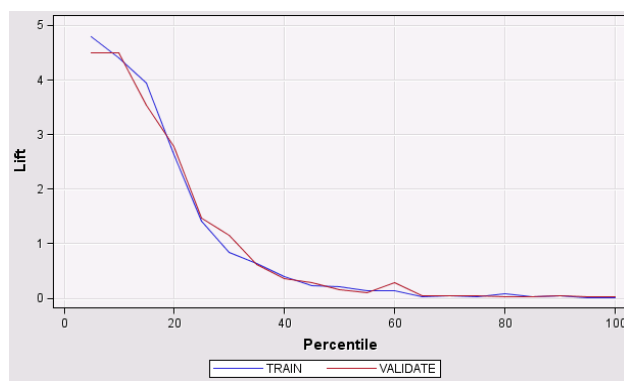


Figura 30 - Lift do modelo RN3

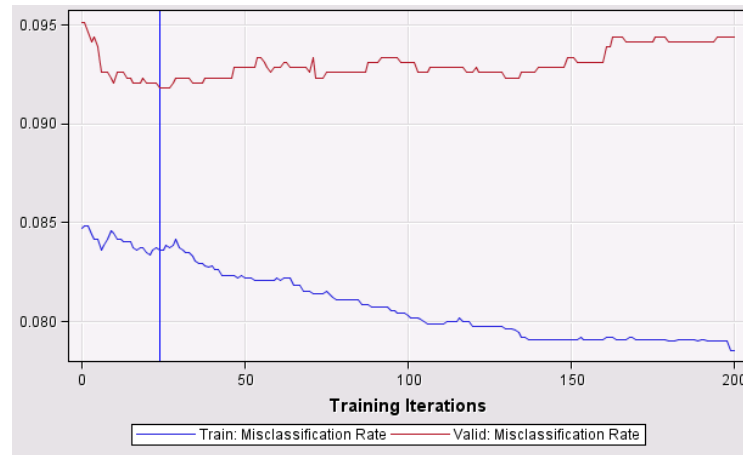


Figura 31 - Misclassification Rate X N° iterações RN3

O gráfico da Figura 31, apresenta a evolução do *misclassification rate* em relação ao número de iterações (total = 26), quando se chega a um patamar que novos pesos já não afectam mais o erro cometido.

A Tabela 10 apresenta a quantidade de clientes *churn*, obtida pelo modelo, por faixa de probabilidade P, nos clientes da amostra de validação, o que representa o risco de *churn* por faixa de probabilidade. Pela tabela observa-se que no conjunto de validação 5% dos clientes têm uma probabilidade entre 95% a 100% de se tornar *churners*.

Dados Conjunto Validação			
Probabilidade (P) de <i>churn</i>	% Clientes	Nº clientes	Probabilidade média de Churn
0.00-0.05	59%	2304	2%
0.05-0.10	9%	348	7%
0.10-0.15	4%	144	12%
0.15-0.20	2%	85	17%
0.20-0.25	1%	57	22%
0.25-0.30	1%	58	28%
0.30-0.35	1%	48	32%
0.35-0.40	1%	40	37%
0.40-0.45	1%	39	43%
0.45-0.50	1%	40	47%
0.50-0.55	1%	50	52%
0.55-0.60	1%	47	58%
0.60-0.65	1%	36	62%
0.65-0.70	1%	47	67%
0.70-0.75	2%	60	72%
0.75-0.80	1%	43	78%
0.80-0.85	2%	71	82%
0.85-0.90	2%	79	87%
0.90-0.95	3%	126	93%
0.95-1.00	5%	177	97%

Tabela 10 - Quantidade e percentual de clientes por faixa de probabilidade de Churn RN3

6.4. Discussão e comparação dos modelos escolhidos

Após a selecção de um modelo por técnica de descoberta de padrões (Árvores de Decisão; Regressão Logística e Redes Neurais), torna-se necessário e imperativo a escolha de apenas um modelo para proceder posteriormente ao *Scoring* do modelo e operacionalização do modelo.

Esta escolha deve recair sobre o modelo que apresente melhor performance quer a nível técnico (através das medidas de avaliação), quer a nível mais subjectivo sobre o qual recairá a decisão da escolha do modelo com maior interpretabilidade para o negócio. A Tabela 11 mostra uma visão geral dos pontos fortes e fracos de cada tipo de ferramenta de modelação, bem como as Estatísticas frequentemente utilizadas para comparação de resultados.

Critério de escolha	Algoritmo de DM		
	Árvores de Decisão	Regressão Logística	Redes Neurais
Facilidade da explicação do resultado	↑	↑	↓
Velocidade de processamento	↑	↑	↓
Flexibilidade no desenho	→	→	↑
Flexibilidade na aplicação a problemas diversos (Escalabilidade)	→	→	↑
Robustez Estatística	↑	↑	→
KS	↑	↑	↑
Curva ROC	↑	↑	↑
Lift	↑	↑	↑
Matriz de Confusão		↑	
Sensibilidade a mudanças nas informações de entrada	→	→	↑
Risco de Overfitting	→	↑	↓
Necessidade de trabalhar valores omissos	↑	↓	↓

Tabela 11 - Comparação dos Algoritmos de modelação

No que respeita a uma comparação directa entre a Regressão Logística e as Redes Neurais pode destacar-se o facto de que as redes neurais não fornecem informações directas sobre o valor de um único parâmetro na previsão (como é o caso da RL que tem como output os coeficientes da equação logística). Ou seja, entre estas duas metodologias, caso os valores das estatísticas avaliadoras dos modelos não forem suficientemente apelativas, a escolha não deve recair sob um modelo RNA ao invés de um RL. Se por um lado não

existem estratégias simples e claras de interpretabilidade dos modelos gerados por técnicas RNA, por outro lado, o risco de *overfitting* ou mesmo a velocidade de processamento, são pontos desfavoráveis que não conferem aos modelos RNA em grande parte dos problemas, ganhos mensuráveis que permitam que a escolha recaia sobre eles.

No que respeita às Árvores de Decisão, a sua simplicidade de emprego, e ainda a fácil interpretabilidade dos resultados, faz dos modelos resultantes fortes candidatos à escolha final. Nomeadamente em negócios como o retalho, em que cada cliente conta, e a percepção de quais as variáveis que estão a contribuir para determinado comportamento, e de que forma estão “encadeadas” são de extrema importância. Todos estes factores acrescem positivamente na avaliação dos modelos gerados por Árvores de Decisão.

Estando a par das nuances acima discutidas, procedeu-se então à comparação estatística dos modelos gerados. Foi incorporado, à semelhança do efectuado na avaliação das técnicas, um nó de comparação de modelos no diagrama SAS (Figura 32).

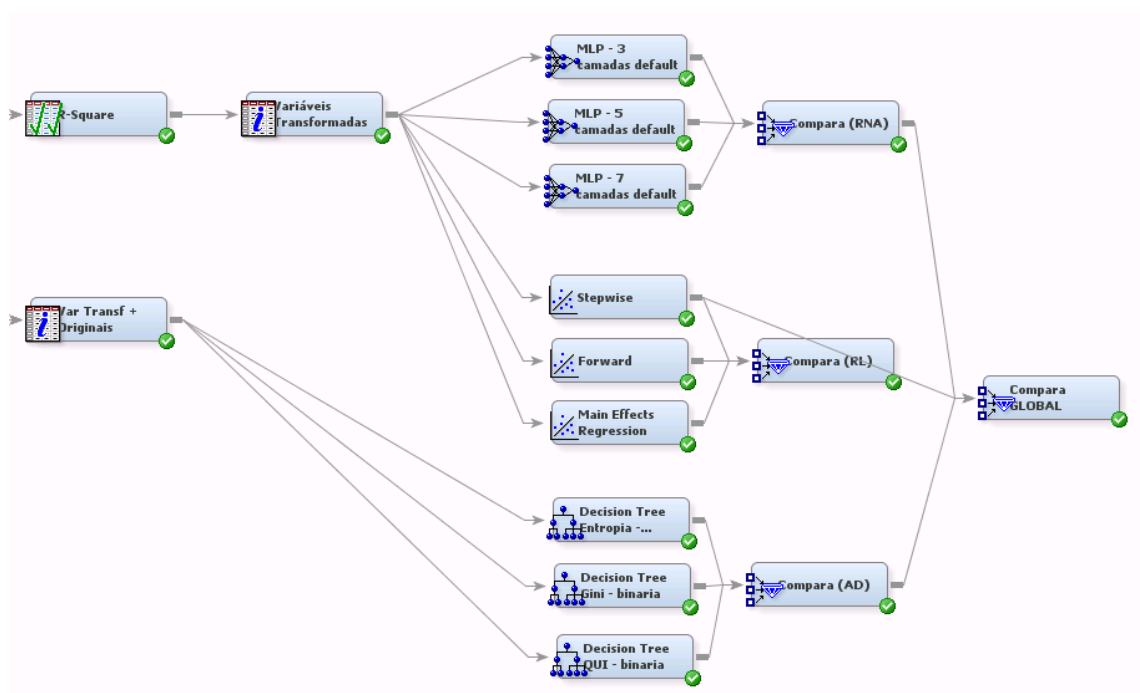


Figura 32 - Diagrama SAS (comparação Global)

Foi seleccionado o método “Misclassification Rate in Validation”, como critério de escolha dos modelos, apenas por uma questão de simplicidade. No entanto todas as estatísticas de avaliação serão discutidas de seguida.

Estatística	RN3	RLStepwise	ADEntropia
Valid: Kolmogorov-Smirnov Statistic	0.749	0.734	0.764
Valid: Roc Index	0.939	0.937	0.916
Valid: Lift	4.486	4.230	4.508
Valid: Misclassification Rate	0.092	0.095	0.075

Tabela 12 - Estatísticas Modelos Seleccionados

Avaliando os valores das estatísticas presentes na Tabela 12, verifica-se que o modelo que parece ser o mais ajustado em termos de desempenho face às medidas de avaliação é o modelo gerado através da técnica Árvores de Decisão (Entropia).

Este modelo apresenta um valor de *misclassification rate* claramente inferior aos restantes (7,5%) e pelo valor do KS pode dizer-se que é também o modelo que melhor separa os eventos *churn* dos não eventos *não churn*. Pelo valor de *lift* obtido por este modelo pode dizer-se a aplicação deste modelo para a predição de clientes *churners* é 4,5 vezes melhor do que a aleatoriedade.

Uma vez que para o caso em estudo é importante a clara separação entre clientes *churn* e *não churn*, dado que desta classificação poderá resultar, por exemplo medidas de comunicação diferentes, ou mesmo uma comunicação única para potenciais *churners*, é importante avaliar as taxas de acerto nos eventos e não eventos de cada um dos modelos.

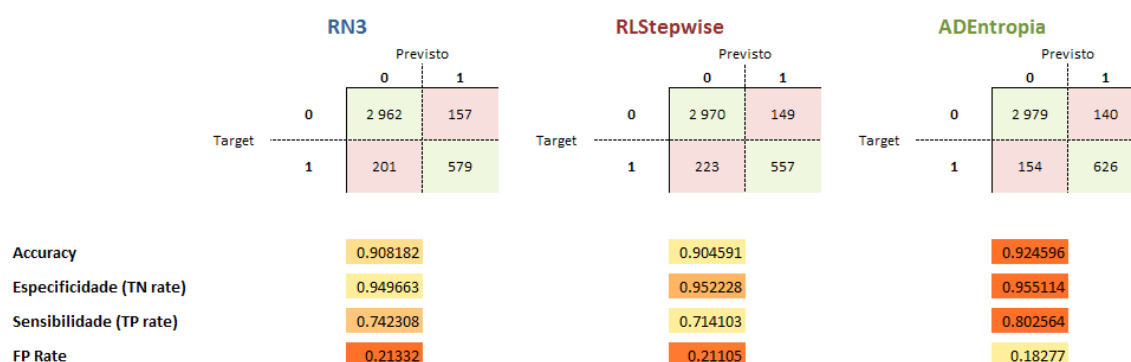


Figura 33 - Matrizes de confusão modelos escolhidos

No que respeita às métricas de comparabilidade dos modelos seleccionados, o modelo ADEntropia apresenta apenas um valor de Roc Index ligeiramente inferior aos restantes, no entanto e avaliando, pela Figura 33, as medidas TP Rate (verdadeiros positivos), e TN (verdadeiros negativos), esta perda face aos restantes é claramente compensada

principalmente pelo valor da Taxa de acerto de verdadeiros positivos (acerto em clientes *churners*), que difere dos restantes modelos em 6pp e 8.5pp.

A medida de *Accuracy* não é mais do que o inverso do *Misclassification Rate*, ou seja, enquanto a segunda mensura a taxa de acerto da previsão, a segunda mede a taxa de erro. Também nesta estatística (Figura 33), o modelo ADEntropia é aquele que apresenta melhor desempenho.

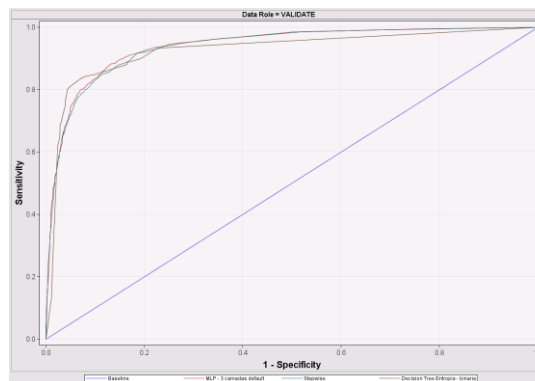


Figura 34 - Curvas ROC modelos seleccionados

O objectivo primário da Curva ROC (Figura 34) é a mensuração dos falsos positivos e dos. Quanto melhor um modelo for capaz de reduzir os dois tipos de erro (Falsos Positivos & Falsos Negativos), mais a curva tenderá a concentrar-se no canto superior esquerdo (ou seja, uma pequena taxa de falsos positivos é acompanhada de um taxa também pequena de falsos negativos). A análise das Curvas ROC da Figura 34, só vem então comprovar o analisado através das matrizes de confusão dos modelos.

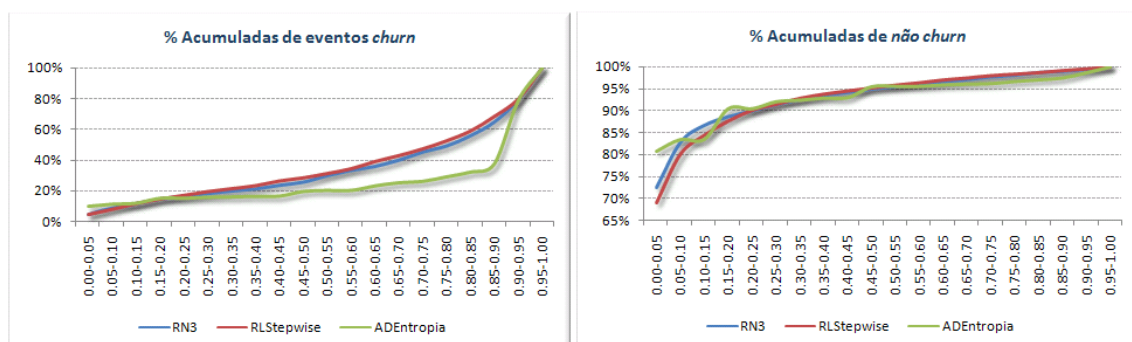


Figura 35 - Score Distribution dos modelos seleccionados

A Figura 35 representa a percentagem de eventos (*churn*), e não eventos (*não churn*), por intervalo de score atribuído por cada um dos modelos. Um modelo útil deve então

mostrar uma maior percentagem de eventos em níveis superiores de *score* (probabilidade de *churn*), e uma maior percentagem de não eventos em níveis inferiores de *score*. Por análise da Figura 35, verifica-se efectivamente que o modelo que scoriza com menor probabilidade de *churn* uma menor percentagem de eventos (*churners* efectivos) é o modelo ADEntropia, cuja percentagem acumulada de eventos cresce acentuadamente apenas a partir do intervalo de *score* (probabilidade) $\geq 85\%$.

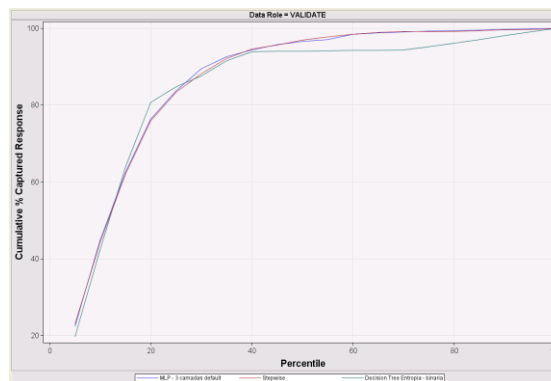


Figura 36 - Ganhos Cumulativos dos modelos seleccionados

No gráfico de ganhos cumulativos (Figura 29) pode verificar-se que entre os 20% de clientes que o modelo ADEntropia identifica como tendo maior probabilidade de abandonarem estão cerca de 80% dos *churners* efectivos. Este valor nos modelos RLStepwise e RN3 atinge apenas os 76%.

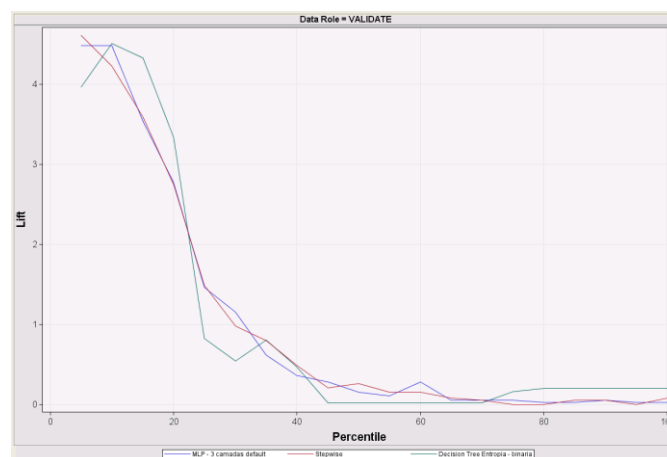


Figura 37 - Lift dos modelos seleccionados

Também, por observação do gráfico da Figura 37, Figura 30 que representa o valor do *lift* para os modelos, por percentil 10, pode dizer-se que o valor anteriormente analisado se traduz num ganho de 3.33 vezes face a uma decisão aleatória (i.e. sem modelo),

enquanto que no que respeita aos restantes dois modelos, o valor do lift para os primeiros 20% de clientes classificados como *churners* é de apenas 2.74.

Após analisar exaustivamente as estatísticas resultantes dos modelos, e incorporando também algumas medidas de carácter subjectivo como a interpretabilidade do modelo e a usabilidade dos outputs do mesmo em termos de negócio (como é o caso da importância das variáveis e encadeamento das mesmas), justifica-se assim a escolha do modelo final: Árvore de Decisão com critério de divisão pela Entropia.

Este modelo utiliza 101 variáveis para previsão e as primeiras 18 em termos de importância podem ser consultadas na Tabela 13. As restantes variáveis estão listadas em Anexo III na Tabela 27.

Variável	Descrição
TX_RSP_M	Taxa de Resposta a Mailings
TRANS_DESC_12M_VL	Proporção do montante de descontos à transacção ultimos 12 meses
VST_ACTUAL_NR	Número total de visitas no mês actual
VST_ULT_12M_NR	Número total de visitas nos últimos 12 meses
GROCERIES_R_CT3_6VL	Rácio do Montante Total de Produtos comprados em Mercearia nos últimos 3 meses / 6 meses
GASTO_ACTUAL_VL	Montante total gasto no mês actual
PREF_STORE	Loja Preferencial do Cliente
CATEGORIAS_NACTUAL	Número de categorias compradas no mês actual
GASTO_ULT_3M_VL	Montante total gasto nos últimos 3 meses
TOP1_CT10	Top1 Categoria em Transacções no Departamento Alimentar
TRANS_DESC_6M_VL	Proporção do montante de descontos à transacção ultimos 6 meses
TOP3_CT10	Top3 Categoria em Transacções no Departamento Alimentar
TICKET_MED_ULT_12M	Ticket Médio por visita nos últimos 12 meses
SQR_CATEGORIAS_NLAST3	Número de categorias compradas nos ultimos 3 meses
(...)	(...)

Tabela 13 - Variáveis importantes ADEntropia (TOP 18)

Analisando as primeiras 18 variáveis importantes para o modelo, pode dizer-se que fazem sentido em termos de negócio. Para um cliente Leal (Fiel), pode de facto ser um factor indicativo de propensão a abandono uma taxa de resposta a mailings reduzida. Sendo o negócio bastante focado em Mercearia, e sendo que esta é uma das categorias de produtos consideradas pela organização como uma categoria de destino, não é de todo surpreendente que quer as categorias TOP neste tipo de produtos, quer o rácio do total deste tipo de artigos dos últimos 3 meses face aos últimos 6, apareçam no top.

Ainda e no que respeita ao número de categorias compradas nos últimos 3 meses e no mês actual, são também variáveis que fazem sentido em termos de negócio e mercado. O conhecimento comum poderia já levar a dizer como razão para abandono, por exemplo, “se um cliente reduz a variedade do seu cabaz de um período para outro, pode ser indicativo de que está a abandonar o relacionamento com a organização”.

Outro factor interessante encontrado no conjunto de variáveis utilizadas e rankeadas em termos de importância do modelo, foi o facto de o modelo utilizar variáveis que mesmo a título inicial do estudo foram ponderadas como sendo variáveis correlacionadas. São elas as variáveis relativas ao total do Gasto do cliente, frequência de visita e mesmo o valor médio por compra do cliente, nos períodos actual, últimos 3 meses, últimos 6 meses e últimos 12 meses. Este facto confere ao modelo seleccionado uma certa conotação temporal, que pode revelar-se extremamente interessante e importante em termos de negócio. Uma vez que o modelo não se baseia a avaliar o comportamento estático do cliente (no mês actual), mas depende também de comportamentos anteriores do cliente. Por exemplo, um decréscimo do valor de compras de um período para outro pode ser indicativo da propensão ao abandono.

6.5. *Scoring* do modelo escolhido

Por último, e dado que o objectivo deste trabalho não era meramente académico, surgiu o interesse em observar o comportamento do modelo seleccionado em situação de Classificação (*Scoring*).

Para este procedimento utilizou-se o SAS Guide com o nó “Model Scoring”. Esta aplicação procede ao cálculo da *Scorização* de todos os clientes de uma base que contenha os mesmos atributos sobre os quais o modelo resultante foi construído com uma probabilidade de *churn*.

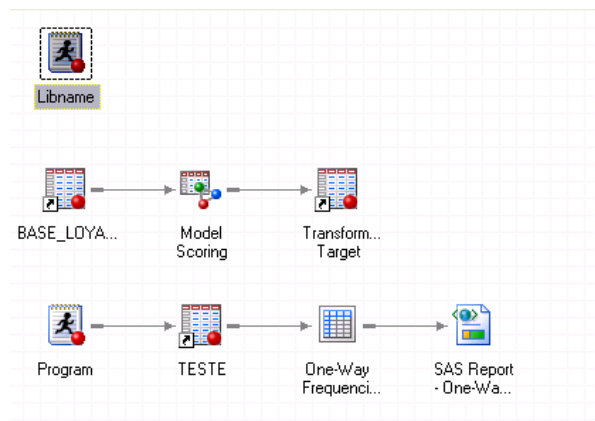


Figura 38 - Scoring do modelo escolhido

Foi então utilizada uma nova ABT (a Abril de 2011), sobre a qual foi efectuado o mesmo tratamento de dados que na ABT utilizada para modelação.

Esta nova tabela (a que se chamará de hora em diante de BD Classificação) está provida de novos registos, não existentes na base de dados de treino do modelo. É constituída por um total de 172.841 instâncias Loyal, e após a definição da *target* a Maio (se compra → *target* =0 ; se não compra → *target* =1), e como pode visualizar-se na Tabela 14, verifica-se uma taxa de *churn* bastante similar à existente na base de dados de treino (0.81% ~ 0.87%), o que parece ser de já um indicio da estabilidade dos dados a utilizar.

Target	Descr Target	L data	
0	Não Churn	171 337	99.13%
1	Churn	1 504	0.87%
Total		172 841	

Tabela 14 - Contador de registos BD de Classificação

A simulação de *Scoring* foi efectuada em três passos cruciais:

- 1) Procedeu-se à classificação da Base de Dados de Classificação (ABT de Abril), e posteriormente;
- 2) A base de dados foi submetida ao modelo de Árvores de Decisão com critério de Divisão Entropia, gerado em SAS Miner;

3) Considerando como resultado *churn* um valor de probabilidade (score) > 50% foi calculada a matriz de confusão.

```
data teste;
set SERVIDOR.SCORE_BASE_LOYAL_ABT_201104Tese;
if P_target1>0.50 and target=1 then Result='VP';
if P_target1<0.50 and target=0 then Result='VN';
if P_target1>0.50 and target=0 then Result='FP';
if P_target1<0.50 and target=1 then Result='FN';
run;
```

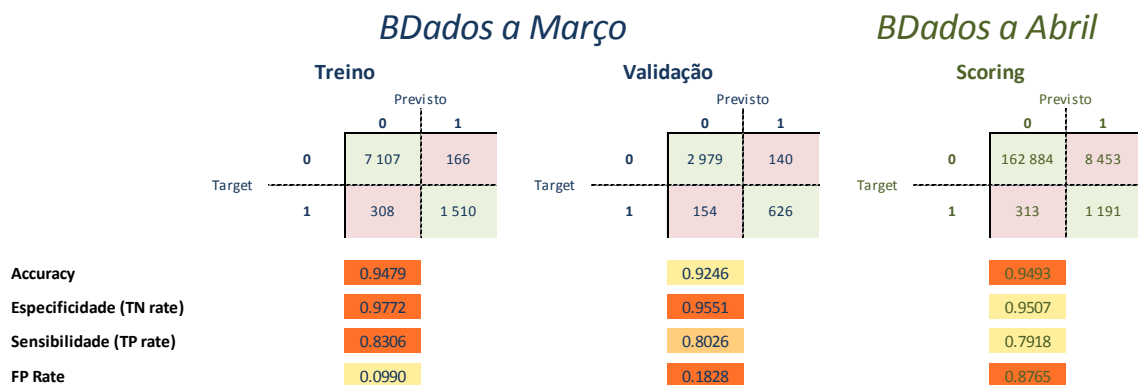


Figura 39 - Matrizes Confusão VS Scoring

Por análise das matrizes de confusão da Figura 39, pode verificar-se que a Árvore de Decisão não se comporta de forma muito coerente no que respeita à classificação de Falsos Positivos. Ou seja, no conjunto testado em *Scoring*, a Árvore classifica como potenciais *churners* cerca de 8400 clientes que efectivamente não o são. Aliado à baixa taxa de *churn* neste segmento de clientes, este fenómeno não se considera muito crítico. Em termos de negócio, e dado que a aplicabilidade destes métodos é a de exercer uma qualquer força comunicativa sobre os clientes propensos a abandonar, seria mais crítico que a taxa de Verdadeiros Positivos fosse muito baixa, o que significaria que o modelo não estava a classificar como *churners* clientes que efectivamente o são, e o assim perder da oportunidade de os *fidelizar*.

É ainda de ressaltar o facto de que à excepção da FP Rate, as taxas de erro obtidas na BD de Scoring e nos conjuntos de treino e teste são bastante semelhantes, o que parece indicar a não existência de overfitting.

7. Conclusões e trabalho futuro

A execução deste projecto permitiu à autora uma exaustiva exploração das bases de dados comportamentais dos clientes da Sonae MC.

O objectivo primordial proposto foi alcançado com sucesso, tendo sido desenvolvido um modelo de classificação de probabilidade de *Churn* com base na informação de dados transaccionais dos clientes portadores de cartão de Fidelização da companhia. O modelo final, foi seleccionado de uma base de três modelos desenvolvidos sob as técnicas de Árvores de Decisão, Redes Neurais Artificiais e Regressão Logística. A eleição do modelo gerado por Árvores de Decisão foi devidamente justificada através dos indicadores de performance e interpretabilidade dos resultados. E foi ainda possível, numa fase final proceder à *scorização* de novos registos, não sujeitos a treino, de forma a avaliar em novas instâncias a performance dos modelos.

Do modelo final seleccionado pode verificar-se que os valores da taxa de resposta a acções promocionais direccionadas são um factor preponderante indicativo da propensão ao abandono. Outro dos factores interessantes no *rankeamento* das variáveis seleccionadas pelo modelo, foi o aparecimento de variáveis que respeitam ao tipo de cabaz pelo cliente. Alterações na gama de produtos Alimentares básicos, ou mesmo o volume do cabaz abrangido pelo acto de compra do cliente parecem então ser um agente no abandono.

É ainda de realçar, e ainda no âmbito das variáveis utilizadas pelo modelo, o carácter temporal que as mesmas conferem ao modelo, uma vez que as “mesmas” variáveis, são utilizadas pelo modelo em períodos de medição diferentes (como o gasto médio ou a frequência de compra).

Os resultados obtidos quer em termos de modelação, quer em termos de *scorização*, foram satisfatórios, levando então ao cumprimento do objectivo subjacente a este trabalho, e que se prende com a criação de ferramentas de alerta de clientes propensos a abandonar o relacionamento com a empresa, para que os mesmos possam ser contactados por via de canais de Marketing.

Contribuições e Limitações:

Após inúmeras experiências com os dados pré-processados, verificou-se que a performance dos modelos desenvolvidos ficava muito aquém do esperado, não apresentando inclusive qualquer ganho face à decisão aleatória (*lift*). De modo a ultrapassar esta dificuldade, e uma vez que a reparametrização da variável *target*, ou mesmo a escolha de um outro conjunto de análise, seria um processo muito moroso, optou-se por recorrer a uma técnica de *Bootsing*, sobre a qual foram duplicados os registos a que correspondiam os *eventos Churn*.

Sobre esta base de dados alterada, foram aplicados os algoritmos e os resultados obtidos, quer na modelação, quer na comparação do modelo final com os dados da *scorização*, levam a acreditar na não existência de *overfitting*, e na aplicabilidade do modelo em situações de tomada de decisão.

Uma das principais inquietações iniciais deste estudo foi sem dúvida a limpeza de dados. Muito embora na Sonae MC exista já um modelo de dados muito bem focalizado para o cliente, e como é transversal em todas as grandes Bases de Dados, problemas de outliers e missing values foram uma das principais preocupações mesmo aquando da modelação das ABT para o desenvolvimento do estudo.

O modelo resultante deste estudo tem no entanto algumas limitações. Apesar de a Árvore de Decisão ser o modelo com a melhor matriz de confusão, sendo portanto aquele que tem potencialmente mais valor para aplicação de uma campanha de retenção de clientes, aquando da *scorização* verificou-se que, apesar do modelo permitir agir de forma proactiva em grande parte dos clientes *Churners*, ele também identifica como potenciais clientes ao abandono clientes que efectivamente não o são. No entanto, e baseados na premissa de que “é melhor reter que conquistar”, os resultados levam a crer que a comunicação de retenção para com um cliente que efectivamente não vá abandonar, terá com certeza um retorno muito mais atractivo para a Companhia, ao invés de não proceder a comunicação e consequente retenção dos verdadeiros *churners*. Campanhas de retenção bem escalonadas, poderão eventualmente fazer aumentar o *share-of-wallet* desses mesmos clientes.

Propostas para investigação e trabalho futuro:

Apesar da satisfação global com os resultados obtidos, uma das melhorias sugeridas em trabalhos futuros, diz respeito à tentativa de diminuição da classificação de Falsos Positivos (clientes que não são *churners* e que ficam classificados como tal). Esta melhoria, poderia eventualmente impedir gastos “desnecessários” com o envio de cartas, por exemplo.

Uma vez que os clientes foram separados em sug-grupos (Loyal e Occasional+Frequent), um dos trabalhos a propor será então, a modelação de um sistema de previsão de *Churn* para o segundo grupo de clientes. Dado que o comportamento natural dos dois grupos será diferente em termos de consumo, os modelos gerados terão com certeza outro tipo de nuances, mesmo no que respeita às variáveis ou mensuração das mesmas, utilizadas nos modelos. Será interessante e revelador, avaliar as diferenças entre os dois tipos de modelos.

Outra das propostas passaria por um estudo de canais de comunicação. Uma vez bem identificados e classificados os clientes, a análise de qual o melhor canal para comunicar com o cliente, poderá eventualmente mostrar-se uma mais-valia no sector.

Ainda e no que respeita a trabalho futuro, sugere-se ainda abordar o tema do *Churn* numa óptica mais parcial de negócio. Dado que o negócio do Retalho possui especificidades que não permite dizer por exemplo “o cliente nunca mais voltará a comprar nas lojas”, dado que esse “nunca” será difícil de definir, a ideia, poderia passar pela construção de um modelo Global de *Churning* através de modelos locais. Estes últimos seriam construídos com base em comportamentos de compra em áreas específicas de negócio como o Talho ou os Frescos, sendo então o objectivo a previsão de *churn* nessas mesmas áreas. Um modelo global de *churn* poderia eventualmente ser construído com base nos resultados dos vários modelos globais obtidos (Ex: se cliente tem probabilidade $> 0,7$ de ser *churner* em Talho e cliente tem probabilidade $> 0,4$ de ser *churner* em Bazar => Cliente é *churner* na companhia).

8. Bibliografia

- CABENA, P, Hadjinian, P, Stadler, R, Verhees, J, Zanasi, A, (1998) “Discovering Data Mining – From Concept to Implementation”, Prentice- Hall, 1998.
- CHAPMAN, Pete (NCR), CLITON, Julian (SPSS), KHABAZA, Thomas (SPSS), REINARTZ, Thomas (DaimlerChrysler), WIRTH, Rüdiger (DaimlerChrysler), (1999) “The CRISP-DM Process Model, CRISP-DM Discussion Paper”
- CHURCHILL, A. G.; Ford, N. M.; Walker, O. C.M (2007); “Sales Force Management”, Boston: Irwin Mcgraw-Hill, 7a Edição
- CHURCHILL, Jr., Gilbert A.; PETER, J. Paul. “**Marketing**: criando valor para o cliente”. 2. ed. São Paulo: Saraiva, 2000.
- CISTER, Angelo M.; EBECKEN, Nelson F. F. (2002)- “CRM through DM: a case study” – Third International Conference on Data Mining- DATA MINING III, Bologna, Italy
- COELHO, P., & Vilares, M. (2006). “Satisfação e Lealdade do Cliente: Metodologias de Avaliação, Gestão e Análise”. Lisboa: Escolar Editora.
- DROZDENKO, Ronald; Drak Perry D. (2002): “Optimal Database Marketing”, SAGE Publications, Thousand Oaks, USA, 2002
- ENGEL, James. F. *et al.* “Comportamento do consumidor”. 8. ed. Rio de Janeiro: LTC, 2000.
- FAYYAD, Usama M., Gregory. Piatetsky-Shapiro, Padhraic. Smyth e Ramasamy. Uthurusamy. (1996) “Advances in knowledge discovery and data mining”. American Association for Artificial Intelligence Menlo Park, CA, USA
- FREEMAN, James A. e David M. Skapura. (1991), “Neural networks: algorithms, applications, and programming techniques”, Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA
- GANESH, J., Arnold, M., & Reynolds, K. (2000). “Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayer”. *Journal of Marketing*, 64(3), 65-87.

- Getting Started with SAS® Enterprise Miner™ 5.2. Cary, NC: SAS institute Inc.: SAS Publishing, 2006
- GRAHAM, J. (2009). “Missing Data Analysis: Making It Work in the Real World”. Annual Review of Psychology, 60, 549-576.
- GROTH, Robert (2000); “Data Mining: Building Competitive Advantage”, Prentice Hall PTR, USA
- GUARINO A. J; Meyers L. S; GAMST G., (2002), “Data Analysis using SAS Guide”, Cambridge.
- HAGAN, M. T., Demuth, H. B; Beale, M. (1996): “Neural Network design”, PWS Publishing Company
- HAIR, Jr., Joseph. F., Barry Babin, Arthur H. Money e Phillip Samouel (2005). “Fundamentos de métodos de pesquisa em administração”, Porto Alegre, Bookman.
- HAN J., Huang, Y., Cercone, N., Fu, Y., (1996), “Intelligente query answering by knowledge discovery techniques”. IEEE Transactions on Knowledge and Data Engineering, v.8, n.3, <http://academic.csuohio.edu/fuy/Pub/tkde96.pdf>
- HAN, J; KAMBER, M. (2001), “Data Mining: Concepts and Techniques”. San Diego: Academic Press, p. 94-165.
- HOFFMAN, P.; GRISTEIN, G.; MARX, K., GROSSE, I.; STANLEY, E. (2003), “DNA Visual Analytic Data Mining”. Lowell, p. 199-231. Department of Computer Science, University of Massachusetts Lowell.
- KLOSGEN, Willi e Jan M. Zytkow (2002), “Handbook of data mining and knowledge discovery”, Willi Klösgen and Jan M. Zytkow: Oxford University Press, Inc. New York, NY, USA.
- KOTLER P.; ARMSTRONG G., (2005) “Princípios do Marketing” (4. Edição). Prentice Hall, p.77.
- LINOFF, Gordon.; Berry, Michael, (2000), “Mastering Data Mining: The Art and Science of Customer Relationships Management”, John Wiley & Sons Inc, USA
- LIU, H; ONG, C. (2006), “Variable selection in clustering for marketing segmentation using genetic algorithms”, Expert Systems with Applications, Vol 34, pp. 502-510

- MATTISON, Rob (2005), “The Telco Churn Management Handbook”, USA
- NEVILLE, P. G (1999), “Decision Trees for predictive Modeling”, SAS Institute Inc.
- PETERMAN, R. (2006). “Modelo de mineração de dados para classificação de clientes em telecomunicações”. Universidade Católica do Rio Grande do Sul, Porto Alegre.
- RAMACHANDRAN, Pushpa (2001) : “White Paper – Mining for Gold”, Wipeo Technologies, Santa Clara CA, USA
- SAS Institute Inc. Data Mining Using SAS® Enterprise Miner™: “A Case Study Approach,” Second Edition. 2003
- SAS® Enterprise Miner™ 5.2. Cary, NC: SAS institute Inc.: SAS Publishing. (2008) “Predictive Modeling Using Enterprise Miner: Course Notes”
- SHEARER C, (2000) “The CRISP-DM Model: The New Blueprint for DataMining”. JOURNAL of Data Warehousing, Volume 5, Number 4, p.13-22.
- SHEPARD, David (2000); “The New Direct Marketing: How to Implement a profit-driven Database Marketing Strategy”, ed David SHEPARD Associates, Mcgraw-Hill
- TORGO, L (1999), “Inductive Learning of Tree-based Regression Models”, PhD Thesis, Univerdade do Porto, Portugal
- YANG, Harry e David Carlin (2000), “ROC SURFACE: A GENERALIZATION OF ROC CURVE ANALYSIS”. Journal of Biopharmaceutical Statistics, v.10, n.2, 04, p.183.

Sites:

- CRISP-DM, www.crisp-dm.org - acedido em 14 de Julho de 2011.
- Site de suporte do SAS: <http://support.sas.com/documentation/> - acedido em 20 de Julho de 2011.
- Wikipedia: <http://pt.wikipedia.org>

ANEXO I

I. A Distribuição em Portugal e o papel da SONAE

O presente capítulo pretende descrever uma visão global da evolução do sector da distribuição em Portugal, tendo em linha de conta, as diferentes etapas de crescimento e os diferentes formatos existentes. Serão também referenciados alguns aspectos relacionados com o futuro das cadeias de distribuição. E será por fim efectuado um pequeno *briefing* acerca do papel da Sonae na Distribuição em Portugal, bem como, uma pequena caracterização da empresa.

I.1. Contexto e Evolução da Distribuição em Portugal

Segundo Rousseau (2002), a distribuição é um conjunto de actividades que colocam os produtos e serviços à disposição das pessoas para que estas possam adquiri-los ou utilizá-los de acordo com as suas exigências e à medida das suas necessidades (Rousseau, 2002).

Segundo Carvalho (2001), a Evolução da Distribuição em Portugal divide-se em 6 grandes etapas. A primeira etapa, iniciou com a década de 60, focada na abertura de pequenas cadeias de supermercado, localizadas nas grandes cidades. A segunda etapa, inicia na década de 70, com surgimento da cadeia Pão de Açúcar, tornando-se pouco tempo depois líder de mercado. A terceira etapa, surgiu com alteração do regime político, no 25 de Abril, onde os grandes grupos foram nacionalizados. A quarta etapa, teve início em 1979, com a cadeia de supermercados Pingo Doce, tendo posteriormente entrado no mercado as insígnias Inô e o MiniPreço, pertencentes ao grupo Pão de Açúcar. O impulsionador da quarta etapa, foi o projecto Continente, em 1985. Nos anos posteriores, ficaram marcados pela abertura, reestruturação e venda de superfícies comerciais. Sendo de referenciar, que as aberturas sucessivas de novos Continentes e hipermercados Jumbo, originaram o desenvolvimento do conceito de hipermercados, constituindo esta a quinta etapa. Por fim a sexta etapa, é marcada pela entrada da cadeia *discount* Dia em Portugal, em 1955, reforçada mais tarde com a entrada do Lidl.

O Sector da Distribuição, em Portugal, assume nos dias de hoje, uma posição de referência na Economia Portuguesa, oferecendo um vasto conjunto de bens e serviços, não só na quantidade mas também na diversificação e sofisticação da oferta.

Nos últimos anos, surgiram novas cadeias internacionais no nosso país, quer no retalho alimentar (Carrefour, adquirida pelo Grupo SONAE em 2008), quer no retalho especializado (ex: Leroy Merlin, IKEA), no entanto, face à rentabilidade e a factores de crescimento, bem como ao estado da economia portuguesa, a instalação de grandes grupos internacionais foi estagnando gradualmente.

Os efeitos não só da crise actual, mas também da crise que se vive actualmente à escala global, também se têm vindo a fazer sentir na economia Portuguesa, com o crescimento do Produto Interno Bruto em termos reais registando cada vez mais recorrentemente valores trimestrais negativos. Efectivamente, a diminuição da procura externa aliada às dificuldades de financiamento colocaram novos entraves às empresas portuguesas, com impacto imediato na retracção dos níveis de investimento.

Como acontece transversalmente em todos os sectores, o sector da Distribuição ressentiu-se efectivamente com este panorama de crise económica, tanto que é de notar uma certa contenção em estratégias de expansão.

No entanto, e apesar do panorama geral, o sector da Distribuição, em Portugal, tem sido um dos sectores onde se tem denotado um crescimento significativo e um aumento da sua contribuição para PIB, resultando da sua importância para a dinamização da actividade económica e aumento da competitividade sectorial.

Não obstante o crescimento acentuado do número de grandes superfícies comerciais, e face ao problema do encerramento do comércio tradicional devido à concorrência por parte dos supermercados e hipermercados, a Autoridade da Concorrência concluiu que das aquisições não resultavam aspectos negativos para o consumidor final, muito pelo contrário, poderia vir a beneficiá-lo, no resultado dos preços. Sendo um facto que a área de venda implantada pelos grandes Retalhistas, Portugal é ainda um dos países europeus com maior número de pontos de venda de comércio tradicional.

I.2. Planeamento/Organização da Grande Distribuição

No mercado da Distribuição, existem diferentes formatos de lojas, tendo em atenção os diferentes tipos de clientes existentes. Todos os formatos são de livre serviço, ou seja, cada cliente tem ao seu dispor produtos na superfície de exposição e venda da loja de modo a permitir a livre circulação dos clientes dentro do espaço comercial e a livre escolha de produtos de acordo com o seu nível de preferência (Rousseau, 2002).

Cada cliente, é atraído por qualquer um dos formatos, dependendo das circunstâncias, tais como geográficas, financeiras, gama de produtos, marcas entre outros aspectos.

Cada formato define um tipo de cliente fiel, ou seja, um supermercado, por exemplo, atrai pessoas que necessitam de produtos (na maioria alimentos) para o seu dia-a-dia, ao passo que lojas de conveniência atraem consumidores que precisam de poucos produtos (na maioria de impulso), sem ter que enfrentar longas filas ou percorrer grandes distâncias para chegar às grandes superfícies. Quando se analisa o seu comportamento, pode-se perceber que os clientes transitam por diversos formatos de loja, dependendo da sua necessidade.

Logo, uma loja de conveniência não irá competir directamente com um supermercado, mesmo que o público-alvo seja idêntico, uma vez que a ocasião de consumo é diferente. As pequenas superfícies apenas poderão competir com as grandes quando identificam e compreendem os hábitos e as exigências dos seus clientes.

Numa grande superfície, como um hipermercado, o consumidor procura variedade de produtos a um baixo preço. O que significa que contrário das lojas de conveniência, os clientes acabam por utilizar estas lojas pelo seu carácter prático, económico, rápido e próximo.

Apesar da grande variedade de formatos, consoantes com a oferta e com o tipo de clientes, pode dizer-se que a Grande Distribuição em Portugal pode ser catalogada pelos seguintes tipos:

1. **Supermercados** – caracterizam-se essencialmente com a dimensão e pequena gama de produtos. Possuem no máximo até 2500 m² de área de exposição e venda.

Através essencialmente da área alimentar. São lojas, que por norma privilegiam a proximidade ao consumidor, geralmente no interior das localidades e acessíveis, proporcionando uma compra cómoda e económica. (Exemplos: Pingo Doce, Continente Modelo).

2. **Hipermercado** – distinguem-se pela dimensão e pela larga e profunda gama de produtos. Possuem um mínimo de 2500 m² de área de exposição e venda. Através por vezes da combinação da área alimentar e não alimentar. Encontram-se localizados, na periferia das cidades de implantação das lojas. (Rousseau, 2002). Ou seja, os hipermercados são caracterizados pelas suas grande áreas horizontais, com elevado nível de cobertura de todos os tipos de produto. (Exemplos: Jumbo, Continente).
3. **Hard Discount** – São lojas de ofertas que operam com estrutura de despesas muito baixa. O seu layout é simples, o sortido limitado, contendo poucos serviços agregados. O tamanho de loja é pequeno, essencialmente encontram-se junto de aglomerados populacionais. O tipo de cliente deste tipo de loja não é homogéneo, nem tipificado. (Exemplos: Lidl, Plus).

I.3. A Distribuição no Futuro

No futuro, a internacionalização e a cooperação entre as empresas da distribuição será um facto, uma vez, que no mercado nacional, o número de licenças aprovadas começa a diminuir. O crescimento actual permitiu às empresas uma forte estrutura competitiva quer ao nível de produtos e serviços, quer ao nível de *know-how* adquirido. Outro dos factores prende-se, com a liberalização dos mercados e normalização dos hábitos e comportamentos de consumo.

Estamos a assistir em Portugal, a um novo formato, associado a um conceito de pequena dimensão, mas de especialização e/ou *gourmet*, consistindo deste modo numa nova oportunidade de negócio, já aproveitado pelas grandes superfícies que implantam pequenas áreas de venda destinadas a este tipo de segmento de mercado. Outros dos factores, predominante, é o forte crescimento da rede de supermercados urbanos,

aproveitando a proximidade do consumidor. O mesmo não se verifica, na rede de hipermercados, uma vez que, o seu crescimento está associado a grandes superfícies comerciais, como lojas âncora. A cooperação, entre produtores e distribuidores, deve ser maximizada, de forma a conseguirem rentabilizar o seu investimento e otimizar processos de compra e venda. Devido à desaceleração do crescimento deverá brevemente assistir-se a associações entre empresas (parcerias), para potenciar o poder de compra e representação das empresas de distribuição em Portugal, trazendo vantagens para os clientes de ambas as empresas.

Este contexto de parceria não é já uma novidade, na medida em que, actualmente se assiste, por exemplo a cooperações entre empresas de diferentes sectores (exemplo: hipermercados com postos de abastecimento).

I.4. A SONAE



Figura 40 – Logo SONAE

A SONAE⁴⁰ iniciou a sua actividade em Portugal em 18 de Agosto 1959, no sector dos produtos derivados de madeira, mais especificamente como produtor de painéis decorativos de alta pressão. Os negócios da SONAE cresceram de forma contínua até à década de 80, ocasião em que se iniciou o processo de diversificação, através da aquisição de uma cadeia de supermercados, seguida da abertura do primeiro hipermercado em Portugal, na cidade de Matosinhos.

Presentemente, a SONAE é uma empresa composta por Holdings nos sectores das telecomunicações (SONAE COM), Retalho Especializado (SONAE SR), Retalho Não Especializado (SONAE MC), Imobiliária (SONAE SIERRA). As várias empresas são geridas de forma independente, por equipas de gestão em dedicação total, que

⁴⁰ SONAE: Sociedade Nacional de Estratificados

desenvolvem estratégias baseadas nos seus próprios factores de geração de valor acrescentado.

A área do retalho encontra-se agora dividida em:

1. Retalho Alimentar (SONAE MC) - A área de retalho alimentar da SONAE é hoje uma referência no mercado, após ter iniciado uma verdadeira revolução nos hábitos de consumo e no panorama comercial português, com a implementação do primeiro hipermercado em Portugal, em 1985 (Continente de Matosinhos). A SONAE MC é líder de mercado nacional, no retalho alimentar, com um conjunto de formatos distintos.

- a. Continente (hipermercados);
- b. Continente Modelo (mini-hipermercados);
- c. Continente Bom dia (supermercados);
- d. Bom Bocado (cafeteiras);
- e. Well's (parafarmácias e Ópticas);
- f. Book.it (livraria/papelaria);
- g. Pets & Plants (Pet Care & Jardim).

2. Retalho não Alimentar (SONAE SR) - é a sub-holding na área do retalho não-alimentar e detém um universo de insígnias que detêm posições de referência nos respectivos segmentos de mercado. A oferta é bastante diferenciada:

- a. SportZone (equipamento e vestuário desportivo);
- b. Worten (electrodomésticos e electrónica de consumo);
- c. Vobis (equipamento informático);
- d. Modalfa (vestuário);
- e. Zippy (vestuário de bebé e criança);
- f. Worten Mobile (telecomunicações móveis);
- g. Worten Gamer (entretenimento);
- h. Loop (calçado).

Tendo em conta o âmbito desta dissertação, e dado que o foco é o comportamento dos consumidores no sector alimentar, irão apenas ser focadas as Insígnias Continente.

I.4.1. O Continente



Figura 41 - Logo Continente

A insígnia Continente, foi a primeira cadeia de hipermercados em Portugal, e desde então permanece como referência no sector de retalho alimentar do país.

As suas lojas, encontram-se localizadas em grandes superfícies das principais cidades portuguesas. A marca Continente assume hoje elevado grau de notoriedade, junto da população portuguesa, possuindo um posicionamento de preços muito competitivo, aliado à variedade de produtos e serviços de que dispõem.

O formato Continente Modelo, possui hipermercados, com aproximadamente 2.000 m² localizados em focos populacionais de média dimensão, bem como supermercados de conveniência e proximidade com cerca de 800m² de área de venda e vocacionados para as compras mais frequentes do dia-a-dia. Com um núcleo muito forte na área alimentar e de perecíveis, as lojas aliam de forma positiva a proximidade às características de variedade dos produtos

As lojas Continente Modelo inserem-se predominantemente em pequenos centros comerciais *stand-alone* com parque de estacionamento, que apresentam uma oferta complementar na área do vestuário, dos electrodomésticos e da electrónica de consumo. Contam igualmente com parafarmácias e pequenas unidades de serviços.

ANEXO II

II.1. Técnicas inerentes ao processo de KDD – Pré Processamento

II.1.2. Pré-processamento - Detecção de *Outliers*

Os processos de DM dispõem de inúmeras técnicas de detecção e eliminação de valores erróneos, denominados na estatística de *Outliers*. No entanto, e como foi referido anteriormente, em determinado tipo de análises estes valores *raros* ou *estranhos* podem ser eventualmente interessantes e provir a análise de inputs importantes para a compreensão de determinado tipo de comportamentos.

Os *outliers* podem ser detectados com testes estatísticos que assumem um modelo de distribuição ou probabilístico para os dados, ou então através de medidas distância (valores mais distantes de um determinado conjunto de dados são considerados como *outliers*). Existem ainda métodos que identificam estes valores segundo uma análise de desvios.

A identificação destes valores pode ou não ser trivial, consoante a origem e a natureza dos dados disponíveis para análise. Algumas das técnicas mais comuns de identificação são:

- 1) **Clustering:** os valores similares são organizados dentro de grupos ou *clusters*, e os valores fora desses conjuntos são identificados como *outliers* e desconsiderados;
- 2) **Inspecção humana combinada com a máquina:** os valores excedentes de uma determinada banda de valores são colocados numa lista e o analista ou especialista examina-os para identificar aqueles que devem realmente ser excluídos;
- 1) **Graficamente (Box Plot)** – é construído um gráfico de Box da seguinte forma:
 - a. Calcula-se a mediana, o quartil inferior (Q1) e o quartil superior (Q3);
 - b. Subtrai-se o quartil superior do quartil inferior = (L)

Os valores que estiverem no intervalo de $Q3+1,5L$ e $Q3+3L$ e no intervalo $Q1-1,5L$ e $Q1-3L$, serão considerados *outliers* podendo, portanto ser aceites na população com alguma suspeita; Os valores que forem maiores que $Q3+3L$ e menores que $Q1-3L$ devem ser considerados suspeitos de pertencer à população, devendo ser investigada a origem da dispersão → *outliers*.

- 2) **Testes estatísticos:** (teste de Grubbs – para distribuições normais, ou teste de Dixon) a identificação de *outliers* vai ser sustentada com base em que se assume que os dados têm uma distribuição normal.
- 3) **Z-Scores:** a aplicação desta técnica implica o cálculo dos z-scores (z-standardizados) dos dados.
 - a. Se o conjunto dos dados é pequeno (inferior a 50), valores que tenham z-scores inferiores a -2.5 ou superiores a 2.5 devem ser considerados *outliers*.
 - b. Se o conjunto dos dados é grande, valores que tenham z-scores inferiores a -3.3 ou superiores a 3.3 são tipicamente considerados *outliers*.
 - c. Se o conjunto dos dados é muito grande (1000 ou mais), também valores mais extremos do que ± 3.3 podem ser considerados dados normais e não *outliers*.

No que respeita à eliminação destes valores extremos, algumas técnicas que podem ser adoptadas para a são [Han; Kamber (2001)]:

- 1) **Eliminação do valor:** ao invés da substituição do valor, este pode simplesmente ser eliminado. No entanto optar por esta estratégia pode originar lacunas em variáveis explicativas importantes.
- 2) **Binning:** os dados relativos a um determinado atributo que contenha *outliers* são previamente classificados, e distribuídos em intervalos (*bins*), de seguida os valores são substituídos pela média do intervalo (*smoothing by Bin medians*) ou então são considerados os valores máximo e mínimo de cada intervalo como os limites do mesmo e os valores do intervalo são substituídos pelo limite mais próximo (*smoothing by Bin boundaries*);
- 3) **Regressão:** os valores extremos podem ser substituídos pelo resultado de uma função, através por exemplo de regressão. Existem dois tipos de regressão que

podem ser utilizadas: a regressão linear e regressão linear múltipla. Na regressão linear o objectivo é encontrar a melhor linha para ajustar dois atributos de modo a que um atributo possa ser utilizado para prever o outro (). Já a regressão linear múltipla é uma extensão da regressão linear, em que são envolvidos mais do que dois atributos, e os dados são ajustados a uma superfície multidimensional, isto é, um atributo de resposta Y é modelado como uma função linear de um vector de atributos multidimensional (p.ex: ().

II.1.3. Pré-processamento - Tratamento de Missing Values

Os valores omissos, ou ausentes, são um problema bastante comum nos *datasets* utilizados para a resolução de problemas estatísticos. Desta forma, este tipo de valores são o alvo de uma das tarefas de Pré-processamento. Apesar de serem de fácil identificação, na maioria dos casos, esta tarefa deve no entanto ser tratada com especial atenção, na medida em que a utilização de técnicas simples, como a substituição de todos os valores desconhecidos por uma única variável global, pode desvirtuar o processo de predição, trazendo resultados indesejados. Por outro lado a utilização de algoritmos de predição para a substituição de valores desconhecidos, traz uma indiscutível complexidade adicional a todo o processo.

A ocorrência de *missing values* deve-se geralmente a erros humanos, à não disponibilidade da informação aquando da entrada dos dados em sistema pelo facto do atributo não ser de preenchimento obrigatório.

Para tratar este tipos de atributos têm-se então as seguintes alternativas:

- 1) **Excluir Instâncias:** Esta técnica pode eventualmente ser adoptada diante de *data sets* muito grandes, uma vez que a exclusão das instâncias em que ocorrem os valores ausentes, não produz uma perda de informação muito significativa;

2) **Substituir os *missing values*:** Segundo Han; Kamber (2001), algumas das formas mais eficazes de substituir valores ausentes são:

- a. **Substituição Manual:** esta alternativa torna-se bastante inviável no caso de se tratar de um conjunto de dados muito grande com muitos *missing values*;
- b. **Substituição com uma constante global:** a utilização desta técnica baseia-se na substituição do valor omissivo por um valor neutro como por exemplo *null*;
- c. **Substituição com o valor mais provável:** esta técnica pode ser aplicada utilizando regressão, ferramentas baseadas em inferências usando formalismo bayesiano ou por indução de árvores de decisão;
- d. **Método de Kononenko; Roscar:** neste método é considerado o valor mais frequente como um bom candidato a substituir o valor ausente. No entanto este método foi aprimorado no sentido de auxiliar em problemas de classificação em que o valor da classe numa dada instância não existe;
- e. **Substituição com o valor médio do atributo:** a complementação ocorre considerando que o conjunto de dados é inteiro, ou então que todas as instâncias pertencem a uma mesma classe que aquela que apresenta o valor ausente;
- f. **Método MVC (*Missing Values Completion*):** este método consiste em dois passos:
 - i. Geração de todas as regras de associação ($X \rightarrow Y$) através do algoritmo *Robust Association Rules*, que permite manipular o *data set* com múltiplos valores ausentes. O objectivo da utilização deste algoritmo é o de descobrir regras em conjuntos de dados válidos, i.é, em subconjuntos de dados que não apresentem *missing values*
 - ii. A substituição dos atributos com valores ausentes é então efectuada utilizando a conclusão da regra (Y)

NOTA: No método MVC é necessário ter em conta que: se todas as regras obtidas indicarem a mesma conclusão, então esta deve ser a utilizada para substituir o *missing value*, caso contrário deve ser utilizada a medida de confiança e o número de regras que

concluem um mesmo valor para ajudar na escolha do valor a ser adoptado para a substituição.

II.1.4. Pré-processamento - Attribute Selection

Este é um problema importante no processo de *KDD* e consiste na selecção de um subconjunto de atributos da base de dados, no qual o algoritmo de DM aplicado possa ser implementado de forma a acrescentar valor. Apesar de simples à primeira vista, essa selecção de variáveis ou atributos deve ser efectuada de forma criteriosa, uma vez que serão as variáveis seleccionadas para realizar o estudo que irão ditar a adaptação do modelo gerado.

Pode dizer-se de forma resumida que os efeitos imediatos da selecção de atributos são: a execução mais rápida do algoritmo de DM e a melhoria da qualidade dos dados, o que conduz a um melhor desempenho do modelo obtido e um aumento da compreensibilidade dos resultados de DM [LIU; YU, 2002].

A literatura possui um grande número de métodos de *attribute selection*. Essas técnicas dividem-se entre aquelas que dependem da modelação completa para que seja verificada a importância das variáveis (*model based*), e aquelas que são totalmente independentes dos modelos e são baseadas em testes estatísticos entre os vários subconjuntos de variáveis de entrada e as saídas desejadas do modelo (*model free*).

A selecção de atributos pode ser efectuada de forma **manual** ou **através de algoritmos**.

- 1) **Seleccção Manual:** os atributos são escolhidos com base na experiência do analista e no entendimento do domínio do problema, bem como com o auxílio de especialistas de negócio;
- 2) **Seleccção automática (por algoritmos):** os atributos são seleccionados através de algoritmos específicos por uma abordagem supervisionada ou não supervisionada. Como por exemplo os algoritmos que determinam medidas de importância das variáveis:
 - a. **R^2 :** É medida a capacidade de discriminação dos atributos com relação às classes e ainda o grau de correlação entre eles. Quanto maior for a

correlação dos atributos com o atributo classe, e menor a correlação entre eles, maior é o valor de importância atribuído a esse atributo;

- b. **Qui-Quadrado:** é calculada para cada variável o valor da estatística Qui-Quadrado (por norma com um alpha de 0.025) com respeito a cada classe;
- c. **Árvores de Decisão:** Cada variável é testada, calculando o Ganho da Informação (que é a redução na Entropia da classe devido à informação adicional fornecida pela variável em teste). As variáveis seleccionadas são aquelas a que ficam associados maiores valores de IG.

Liu; Yu (2002) sugeriu que um método típico de selecção de atributos compreende quatro passos básicos (Figura 42).

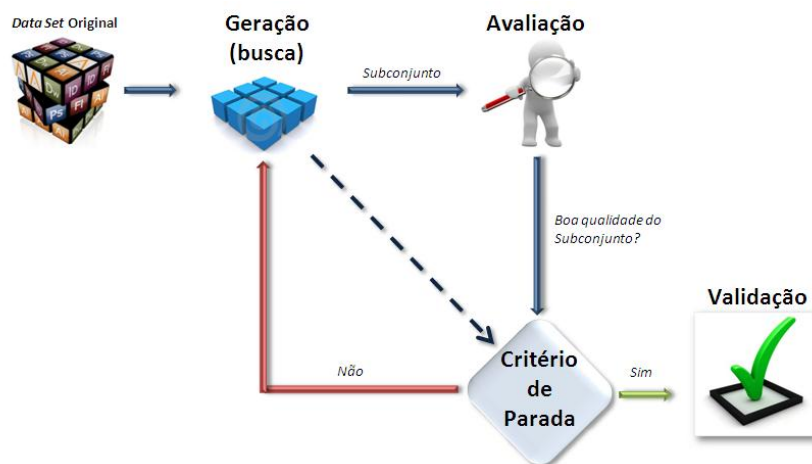


Figura 42 - Procedimentos Selecção Atributos (Liu; Yu)

Esta metodologia foi inicialmente projectada para selecção supervisionada, no entanto, e com o desenvolvimento dos trabalhos realizados sobre a selecção não-supervisionada, foi mantida com o acréscimo de considerar um novo critério de avaliação a partir de modelos filtro e *wrapper* dentro do passo de avaliação do subconjunto.

a) Fase de geração ou busca

Este procedimento consiste na geração de subconjuntos de atributos para avaliação. Esta geração pode assumir diferentes direcções: *forward*, *backward* ou *aleatória*.

a.1.1) Direcção *forward*: um subconjunto de atributos é expandido a partir de um conjunto vazio, sendo que os atributos são adicionados um a um em cada iteração, o

melhor atributo entre os não seleccionados é escolhido baseado no critério de avaliação. O subconjunto de atributos vai crescendo até ao conjunto original ou até ao **critério de parada** ser satisfeito;

a.1.2) Direcção *backward*: a busca é iniciada a partir do conjunto completo de atributos. Em cada iteração é efectuada a eliminação do atributo menos relevante através do critério de avaliação. O subconjunto de atributos vai reduzindo até ser composto apenas por um atributo ou até ao **critério de parada** ser satisfeito;

a.1.3) Direcção aleatória: esta técnica refere-se a que a busca é iniciada num subconjunto aleatório de atributos, sendo que os atributos restantes vão sendo adicionados e removidos de forma aleatória.

Nesta fase é ainda importante a questão da estratégia de busca que Liu classificou como:

a.2.1) Busca completa: através desta estratégia, todos os subconjuntos óptimos de atributos são encontrados, no entanto não significa que todos os atributos são avaliados.

a.2.2) Busca Heurística: esta estratégia de busca tende a não encontrar todos os subconjuntos óptimos de atributos. O tipo de busca é em profundidade (*depth-first*). Este tipo de algoritmos são considerados de implementação muito simples e de produção de resultados rápida.

a.2.3) Busca Não-Determinística: adoptando-se esta estratégia, a busca do conjunto seguinte de atributos é efectuada de forma aleatória.

b) Função de avaliação

A função de avaliação é responsável por medir a boa qualidade de um subconjunto produzido pelo procedimento de geração, sendo este valor comparado com o do subconjunto que foi anteriormente considerado melhor. Caso o que está no momento a ser avaliado seja melhor que o anterior, substitui o mesmo.

O subconjunto óptimo está sempre relacionado a um determinado critério de avaliação, i.é, um subconjunto de atributos que tenha sido escolhido segundo um critério de

avaliação, pode não coincidir, no caso de ser utilizado um outro critério aquando de uma segunda experiência.

Os critérios de avaliação podem ser divididos em dois grandes grupos:

b.1) Critérios Independentes: são utilizados pelos métodos de selecção filtro⁴¹ (*filter*), e avaliam a qualidade de um atributo ou subconjunto de atributos sem o envolvimento de algoritmos de aprendizagem. Alguns desses critérios são: medidas de distância, medidas de informação (medem o ganho de informação de um atributo, ou seja a diferença entre a incerteza anterior e posterior à utilização de um atributo X), medidas de dependência (ou correlação que qualificam a habilidade de prever o valor de um atributo a partir do valor de outro) e medidas de consistência (que encontram o subconjunto de tamanho mínimo que satisfaça uma taxa de inconsistência aceitável previamente definida pelo analista).

b.2) Critérios Dependentes: são utilizados comumente pelos métodos de selecção *wrapper*⁴², e tentam avaliar a boa qualidade de um subconjunto de atributos pela avaliação do desempenho do algoritmo de aprendizagem aplicado sobre o mesmo, ou seja acaba por ser a mesma medida de desempenho do algoritmo. São exemplos destes critérios as medidas para classificação (que têm o objectivo de maximizar a *accuracy*⁴³ preditiva) ou as medidas para *Clustering* (que estimam a qualidade dos resultados de *clustering*).

II.2. Técnicas inerentes ao processo de KDD – Transformação

⁴¹**Métodos de selecção *filter*:** Tentam encontrar o conjunto de atributos relevantes antes do processo de aprendizagem indutiva através da selecção de alguns atributos e exclusão de outros irrelevantes e/ou redundantes.

⁴²**Métodos de selecção *wrapper*:** O classificador é a própria função de avaliação, ou seja, o classificador selecciona os atributos e posteriormente utiliza-os na predição de rótulos de classes de instâncias não inspeccionadas, o que torna o nível de *accuracy* muito alto.

⁴³Accuracy – exactidão

II.2.1. Transformação dos dados – Discretização de variáveis

A discretização consiste então em transformar valores de atributos conjuntos em valores simbólicos (discretos), no entanto pode também equivaler à formação de categorias de valores que já sejam originalmente discretos mas que se apresentem em grande número de diversidade. O processo de discretização⁴⁴ não é de todo trivial.

Segundo Hussain et al. (1999), para a realização de um processo de discretização é necessário levar a cabo quatro etapas importantes:

- 1) **Classificação do atributo:** os valores contínuos do atributo a ser tratado são classificados por ordem crescente ou decrescente.
- 2) **Seleção de *cut-points*/intervalos adjacentes:** entende-se por *cut-point* o valor real dentro de uma faixa de valores contínuos que divide a faixa dentro de dois intervalos, um intervalo é menor ou igual ao *cut-point* e o outro intervalo é maior ao *cut-point*;
- 3) **Divisão / Combinação:** para a divisão é preciso avaliar *cut-points* e escolher o melhor deles para dividir a faixa de valores contínuos entre duas partições. A discretização continua em cada parte até o critério de parada ser satisfeito. Similarmente para combinação, os intervalos adjacentes são avaliados para encontrar o melhor par de intervalos para combinar em cada iteração;
- 4) **Parada do processo:** um critério de parada específica quando o processo de discretização deve parar. Para isso pode utilizar-se um critério simples como a definição inicial do número de intervalos, ou então pode recorrer-se a uma função de avaliação.

II.2.2. Transformação dos dados – Redução de Dados

Apesar das claras vantagens da utilização de grandes volumes de dados para treino e teste, o conjunto de dados pode no entanto tornar-se excessivamente grande na prática.

⁴⁴ Processo de discretização de variáveis segundo Hussain et al. (1998), pode ser consultado no Anexo II.

Além da dimensão elevada dos dados poder exercer um peso negativo no desempenho dos algoritmos, muitas ferramentas (como por exemplo o WEKA), têm uma limitação quanto ao tamanho do *dataset* sob o qual aplicam os algoritmos de modelação. Surge com estas questões, a necessidade de **redução dos dados**.

Existem diversas técnicas de redução que podem ser aplicadas ao conjunto inicial de dados. KLOSGEN (2002), sugere uma abordagem baseada em *sampling*⁴⁵, que segundo o autor são rotinas que permitem estimar características das populações de forma eficiente e com distorção mínima, descrevendo os principais tipos de técnicas de amostragem como:

- 1) **Amostragem aleatória simples:** neste tipo de abordagem (*random*), todos os registos do conjunto de dados inicial C, têm a mesma probabilidade de serem recolhidos para C'. Probabilisticamente pode dizer-se que quanto maior for o tamanho de C', este será provido de maior representatividade do conjunto inicial;
- 2) **Amostragem aleatória estratificada:** a população é subdividida em grupos e é aplicada a amostragem aleatória simples a cada um dos grupos, que no final serão conjugados para a formação do conjunto C';
- 3) **Amostragem por Cluster:** esta técnica é similar à amostragem aleatória estratificada, no entanto parte-se do princípio que os subgrupos da população já existem naturalmente (género ou região);
- 4) **Amostragem sistemática:** pode ser útil em casos em que os indivíduos de uma dada população são de alguma forma numerados. De forma sucinta, a técnica envolve a escolha aleatória de um membro da população entre 1 e N, incluindo-o sistematicamente no subconjunto C'. Não sendo totalmente aleatória, esta pode não ser ideal para amostras que se pretendam o mais representativas possível da população;
- 5) **Amostragem em duas fases:** deve ser utilizada quando se pretende organizar os dados segundo de algumas das variáveis e se desconhece a distribuição dessas mesmas variáveis.

⁴⁵ *Sampling* - Amostragem

II.3. Técnicas inerentes ao processo de KDD – Pós-Processamento

II.3.1. Pós-processamento – Avaliação dos modelos

II.3.1.1. Avaliação de Modelos de Classificação

Quando se trabalha um problema de classificação, uma das técnicas usadas é a *Matriz de confusão* (Kohavi & Provost, 1998). Esta é utilizada para avaliar o resultado de uma classificação, mapeando os valores previstos por um modelo com os valores desejados.

II.3.1.1.1. Matriz de Confusão

A matriz de confusão (Figura 43) reflecte a quantidade de classificações correctas e erradas sobre um conjunto de exemplos X . O número de acertos para cada classe é indicado na diagonal principal da matriz $M(C_i, C_j)$, esses acertos podem ser *verdadeiros positivos* (*TP* (*true positive*)) – correspondem ao número de exemplos positivos e classificados como tal) e *verdadeiros negativos* (*TN* (*true negative*)) – correspondem ao número de exemplos negativos classificados como tal. Os restantes elementos da matriz $M(C_i, C_j)$ para $i \neq j$, representam erros na classificação que podem ser *falsos positivos* (*FP* – *false positive*) ou *falsos negativos* (*FN* – *false negative*).

		Classe Verdadeira	
		p	n
Classificada Como	p	True Positive	False Positive
	n	False Negative	True Negative
Totais		P	N

Figura 43 - Matriz de confusão de um classificador

Da leitura da matriz e da quantificação das entradas da mesma é possível derivar as seguintes medidas:

- $\frac{TP}{TP + FN}$, que determina a percentagem de respostas dadas pelo modelo face aos casos equacionados;
- $\frac{TN}{TN + FP}$, que determina a percentagem de acertos negativos correctos relativamente a todos os casos classificados como negativos;
- $\frac{TP}{TP + FN + FP}$, também denominada de *recall* ou *TP Rate* em alguns softwares, calcula a probabilidade de acerto ou taxa de verdadeiros positivos;
- $\frac{FP}{FP + FN}$, utilizada essencialmente nas curvas ROC, mede a probabilidade de *falsos alarmes* ou taxa de falsos positivos.

Estas medidas de desempenho são independentes do custo e das probabilidades das classes.

II.3.1.1.2. Curvas ROC

A análise ROC (Receiver Operating Characteristic) é uma ferramenta muito útil na avaliação de modelos de classificação. Surgiu no âmbito da “*Teoria da Detecção de Sinal*”, desenvolvida durante a II Guerra Mundial para a análise de imagens de radar. A tarefa consistia em descobrir se um dado sinal era indicador de um barco inimigo, aliado ou se o dito sinal era simplesmente ruído. A utilidade deste método foi reconhecida mais tarde e a sua utilização foi estendida para a comparação de exames clínicos para diagnóstico médico. Esta curva permite a avaliação do desempenho de um classificador, sendo apropriado quando existem dois estados possíveis, ou seja, duas classes.

Os gráficos ROC são muitas vezes utilizados como alternativa à avaliação utilizando medidas, uma vez que permitem uma melhor visualização da multidimensionalidade do problema de avaliação como por exemplo o limiar entre taxas de erro e taxas de acerto dos classificadores.

Estes gráficos são baseados na relação entre a *TP Rate (sensitivity)* e na *FP Rate*, variando num determinado *threshold*. Para se construir o gráfico ROC, coloca-se a FP rate no eixo das ordenadas (eixo x) e a TP rate no eixo das abcissas (eixo y).

No espaço ROC, um classificador é um ponto, e para este ser obtido são calculadas as taxas FP e TP com base na matriz de confusão gerada pelo modelo (Figura 43).

Pontos de destaque no espaço ROC

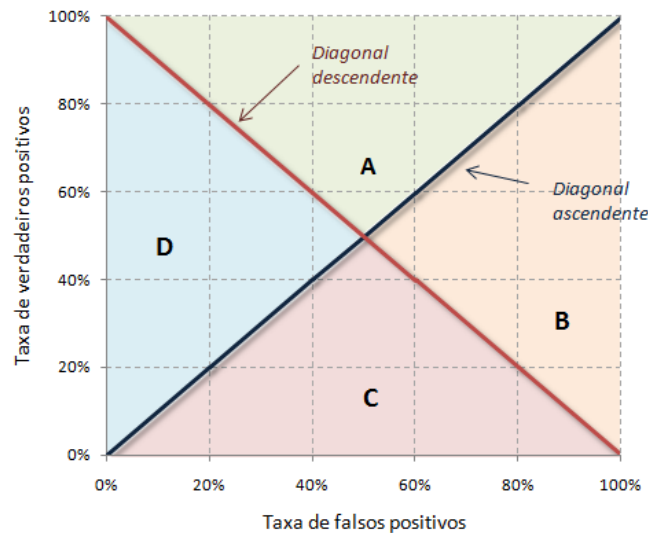


Figura 44 - Espaço ROC

No espaço ROC é relativamente perceber se um modelo é melhor que outro. Isto acontece se e somente se o respectivo ponto está acima e à esquerda do ponto do modelo *concorrente*, isto é se tem uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos.

- (1) **Ponto (0,0)** – representa a estratégia de **nunca classificar um exemplo como positivo**. Ou seja, modelos que correspondem a esse ponto, não apresentam nenhum falso positivo.
- (2) **Ponto (100%,100%)** – representa a estratégia de **classificar sempre um novo exemplo como positivo**.
- (3) **Ponto (0,100%)** – representa um **modelo perfeito**. Ou seja, todos os exemplos positivos e negativos são correctamente classificados.
- (4) **Ponto (100%,0)** – indica que todas as predições do modelo são erradas.
- (5) **Modelos próximos ao canto inferior esquerdo** – indicam que o modelo classifica como positivos apenas os exemplos em cuja segurança é muito elevada. Como consequência classificam poucos falsos positivos, mas frequentemente têm baixas taxas TP.
- (6) **Modelos próximos ao canto superior direito** – A maioria dos exemplos são classificados como positivos, o que muitas vezes cria taxas de FP bastante elevadas.

- (7) **A linha diagonal ascendente** – representa um modelo de comportamento estocástico: cada ponto (j, j) pode ser obtido pela previsão da classe positiva com probabilidade j e da classe negativa com probabilidade $1-j$.
- (8) **Triângulo A e B** – Modelos *liberais*. Isto é, fazem classificações positivas com pouca evidência
- (9) **Triângulo C e D** – Modelos *conservadores*. Isto é, classificam um novo exemplo como positivo se e só se existir uma forte evidência.
- (10) **A linha diagonal descendente** – modelos com bom desempenho igualmente em ambas as classes.

Área Abaixo da Curva (AUC)

Uma curva ROC é uma demonstração bidimensional da performance de um classificador. Para comparar classificadores é necessário reduzir a curva ROC a um valor escalar, ou seja uma métrica de comparação entre os classificadores.

Um método comum para realizar esta redução é calcular a **área abaixo da curva ROC** gerada pelo classificador (no caso do classificador produzir apenas um ponto é comum unir esse ponto aos pontos $(0,0)$ e $(100,100)$ do espaço ROC).

Como a AUC é uma fracção de um quadrado de lado 1 (espaço ROC), o seu valor está sempre entre 0 e 1. Esta área pode ser interpretada como a probabilidade de um exemplo positivo, escolhido aleatoriamente, ser classificado como tal.

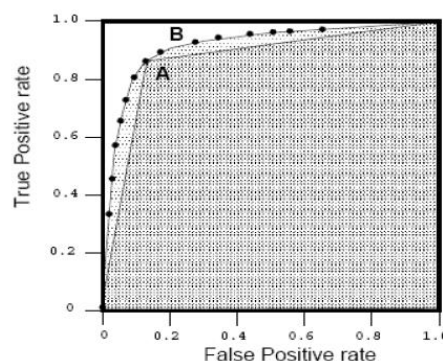


Figura 45 - Curvas ROC classificador B (gerador de vários pontos) e Classificador A (um só ponto no espaço ROC)

Na Figura 45, o classificador B possui uma área maior, logo pode inferir-se que tem uma performance média melhor que o classificador A.

Análise ROCCH (ROC Convex Hull)

Esta análise permite declarar um subconjunto de classificadores como potencialmente óptimos. Incluídos todos os pontos que constituem as curvas ROC dos diferentes classificadores, sendo formada a *Convex Hull* correspondente, é realizada uma análise dos pontos que estão sobre a linha de *Convex Hull*. Se um ponto está sobre a *Convex Hull*, existe então uma linha tangente a esse ponto que tem uma *TP rate* superior, sendo o classificador representado por esse ponto considerado como o classificador óptimo.

Existem ainda outras formas gráficas de avaliação como o *lift chart* que proporciona um sumário visual da usabilidade da informação que provem de um ou mais modelos preditivos. O Lift é uma medida calculada com base no rácio entre os resultados obtidos com e sem a previsão. A interpretação do gráfico é bastante simples, e semelhante às curvas ROC: quanto maior for a area entre o valor do lift e a *baseline*, melhor será o modelo.

II.3.1.1. Avaliação de Modelos de Regressão

Um modelo de regressão é um estimador, que tenta estimar o valor associado a cada um dos exemplos. O principal objectivo de um método de regressão é conceber o “melhor” modelo segundo uma medida da estimativa do erro. Quando se trata de problemas de regressão, o erro (ou resíduo) e é medido por $e = d - d'$ onde d representa o valor desejado e d' o valor ajustado pelo modelo.

Para um conjunto de dados , pode ser calculado o erro global:

- $\frac{1}{n} \sum_{i=1}^n (d_i - d'_i)^2$;
- $\frac{1}{n} \sum_{i=1}^n |d_i - d'_i|$;
- $\frac{1}{n} \sum_{i=1}^n (d_i - d'_i)$;
- $\frac{1}{n} \sum_{i=1}^n (d_i - d'_i)^2$.

As medidas *MAD* e *MSE* são duas das medidas de avaliação mais usadas no âmbito da regressão. Ambas possuem origens na estatística, tendo como intuito inferir quão melhor é um estimador para um dado conjunto de valores. Se ele for uma boa medida de

centralidade, representa a distribuição dos valores, melhor que qualquer outro valor. A melhor medida de centralidade é aquela que minimiza as medidas *MAD* e *MSE*.

Segundo Torgo (1999), essas medidas quando aplicadas à avaliação de modelos de regressão, servem a diferentes propósitos. Se na previsão de valores for aceitável cometer alguns erros extremos, desde que aproximada na maior parte das vezes ao valor real, a medida *MAD* é a mais adequada. Quando se está perante situações em que é crucial não cometer erros extremos, mesmo que possam ocorrer erros pequenos mas frequentes, então a medida *MSE* deve ser a utilizada, uma vez que esta utiliza o quadrado das distâncias das previsões aos valores reais, o que faz com que grandes distâncias sejam amplificadas relativamente a pequenas distâncias.

II.4. Data Mining como parte do processo KDD

II.4.1. Algoritmos e Técnicas de DM – Árvores de Decisão

As Árvores de Decisão merecem actualmente uma grande atenção da comunidade científica de várias áreas do conhecimento (Geurts, Irrthum, & Wehenkel, 2009).

As Árvores de Decisão podem ser caracterizadas como representações do conhecimento e como robustas ferramentas de classificação. A simplicidade de leitura dos resultados, e da sua implementação está na origem da sua popularidade.

As Árvores de Decisão são constituídas por:

- **Nós** (internos ou raiz) que representam os atributos (i.e. variáveis independentes), e sobre os quais são efectuados os testes;
- **Ramos** provenientes dos nós, que recebem os valores possíveis para as variáveis consideradas;

- **Folhas**, que representam as diferentes classes da variável dependente de um conjunto de dados de treino, assim cada folha fica associada a uma determinada classe.

Ao contrário das Árvores na natureza – que crescem da raiz para as folhas - as Árvores de Decisão são construídas numa metodologia *topdown*. Representa-se no topo o nó de raiz da Árvore, deste crescem ramos que dão origem a novos nós, e assim sucessivamente até se chegar a um nó folha.

Nos ramos superiores da análise ficam, desta forma, representadas as variáveis que nos trazem maiores ganhos de informação e que contribuem para uma convergência mais rápida para um nó folha e, conseqüentemente, para uma conclusão. A Classificação é representada por este mesmo percurso.

Analizando uma Árvore de Decisão é intuitivo que cada percurso na Árvore - da raiz a uma dada folha - corresponde a uma regra de classificação. Estas regras são facilmente traduzíveis em acção, em linguagem de programação ou de Base de Dados (e.g. linguagem SQL).

O objectivo de uma Árvore de Decisão é operar partições sucessivas num conjunto de dados de treino, até que cada um dos subconjuntos obtidos contenha casos de uma única classe (Petermann, 2006). O modelo obtido servirá para futuras classificações.

Esta definição implica que para induzir Árvores de Decisão, é necessário um conjunto de dados de treino previamente classificados (i.e. registo dos quais se conhece a classe da variável dependente, sendo na maior parte dos casos dados históricos). O conjunto de treino deve ter exemplos negativos e positivos de uma classe.

Existem alguns aspectos importantes a ter em conta na construção de uma árvore de decisão. Destes pontos podem depender a performance do classificador. Entre eles encontram-se, por exemplo os processos de *prunning*⁴⁶. Estes algoritmos produzem muitas vezes estruturas com mais ramificações do que o necessário, pelo que se torna imperativo “podar” a estrutura obtida. A “poda” pode ser realizada durante a aprendizagem, no entanto isso torna o processo mais complexo. Além disso,

⁴⁶ *Prunning* – processo de poda da árvore de decisão

determinadas “podas” só podem e devem ser decididas após a construção da estrutura, pelo que a maior parte dos algoritmos faz o *pruning* apenas nessa fase.

De forma a obter um melhor classificador, é ainda importante ter em conta:

- O critério para a escolha da variável a utilizar em cada nó;
- Qual a melhor forma de calcular a partição do conjunto de treino;
- De que forma se deve decidir que um nó é uma folha → Critério de paragem
- Qual o melhor critério para seleccionar a classe a atribuir em cada folha.

No que respeita ao teste do modelo gerado, se houver dados em elevado número, pode-se reservar parte deles para teste após a construção da árvore (técnica de *reduced-error pruning*). Caso os dados sejam escassos, pode-se utilizar um esquema de *Cross-Validation*.

As vantagens das árvores de decisão são:

- 1) As árvores de decisão podem ser aplicadas a qualquer tipo de dados.
- 2) A estrutura final do classificador é bastante simples podendo ser guardada e manipulada de uma forma bastante eficiente.
- 3) Manipula de uma forma muito eficiente informação condicional subdividindo o espaço em sub-espacos que são manipulados individualmente.
- 4) Os modelos resultantes destes classificadores revelam-se normalmente robustos e insensíveis a erros de classificação no conjunto de treino.
- 5) As árvores resultantes são normalmente bastante compreensíveis podendo ser facilmente utilizadas para se obter uma melhor compreensão do fenómeno em causa.

Como desvantagem desta técnica, pode apontar-se o desempenho obtido em problemas com uma elevada não linearidade. Nestas situações, as Árvores de Decisão tendem a ser ultrapassadas por técnicas com uma maior complexidade de aprendizagem, tais como as Redes Neurais ou SVM's.

II.4.1.1. Árvores de Decisão - Algoritmos de Implementação

Existem diferentes algoritmos de classificação que elaboram árvores de decisão. Não existe uma regra que determine parâmetros de performance para a definição apriori de

qual o melhor algoritmo a aplicar, sendo que as diferentes abordagens podem traduzir resultados igualmente satisfatórios em função do problema em análise.

- 1) **Automatic Interaction Detection (AID)**: proposto por Morgan e Sonquist em 1963, é um método de criação de árvores através da separação dos dados em grupos sucessivos com base nos seus valores;
- 2) **CHAID**: resultou da introdução, por Hartigan 1975, do teste do chi-quadrado para identificar variáveis independentes ao método **AID**. Este novo método permitiu diminuir o crescimento inicial das árvores facilitando a poda;
- 3) **Classification and Regression Trees (CART)**: desenvolvido por Breiman et al., este algoritmo cria árvores binárias pelo que, se o teste de uma variável apontar para mais que dois valores, os ramos do nó de teste terão que incluir sub-testes o que pode levar a árvores de grande extensão;
- 4) **Iterative Dichotomizer (ID)**: desenvolvido por Ross Quinlan teve várias versões e serviu de base ao **C4.5** que, por sua vez, também tem tido várias versões, entre as quais algoritmos para regressão. Tal como o **CHAID**, estes algoritmos identificam as variáveis independentes, mas recorrendo aos conceitos de “entropia” e “ganho de informação”.

A entropia está relacionada com a distribuição dos valores de uma variável, ou seja, entropia elevada quer dizer que há uma distribuição mais uniforme, ao invés uma entropia pequena quer dizer que há um valor predominante, e o ganho de informação indica a capacidade de uma variável em separar os casos de treino.

-
-

II.4.1.2. Indução de Regras

Trata-se de uma técnica muito conhecida, geralmente utilizada para representar o conhecimento associado às Árvores de Decisão. A representação é escrita utilizando uma sintaxe do tipo: **SE condição ENTÃO ação**. São reconhecidas vantagens

importantes na **Indução de Regras**, das quais se destaca a facilidade de explicação e compreensão das regras.

II.4.2. Algoritmos e Técnicas de DM – Redes Neurais

As **Redes Neurais Artificiais** (RNA) têm ganho nos últimos anos bastante popularidade no seio das soluções de suporte à tomada de decisão. Estas oferecem uma arquitectura bastante potente, com elevadas capacidades de aprendizagem e de representação de padrões dos dados (Petermann, 2006). No entanto, e apesar de serem um instrumento poderoso, são também uma das ferramentas de Data Mining mais difíceis de interpretar para os analistas.

As características mais relevantes das RNA's (segundo Hagan et al., 1996), são:

- 1) **Aprendizagem e generalização:** conseguem descrever o todo a partir de algumas partes, constituindo-se como formas eficientes de aprendizagem e armazenamento de conhecimento;
- 2) **Processamento paralelo:** permitem que tarefas complexas sejam realizadas num curto espaço de tempo;
- 3) **Não linearidade:** atendendo a que a maioria dos problemas reais são de natureza não linear;
- 4) **Adaptabilidade:** podem adaptar a sua topologia de acordo com mudanças do ambiente;
- 5) **Robustez e degradação suave:** permitem processar o ruído ou informação incompleta de forma eficiente, sendo capazes de manter o seu desempenho quando acontece a desactivação de algumas conexões e/ou neurónios;
- 6) **Flexibilidade:** possuem um grande domínio de aplicabilidade.

As RNA's têm um funcionamento análogo ao do sistema nervoso central nos humanos, sendo constituídas, tal como nos humanos, por um conjunto de unidades básicas de processamento fortemente interligadas, denominadas de neurónios artificiais. Cada

neurónio recebe uma série de valores e, em função deles, determina um valor a apresentar como saída. Os valores de saída de alguns neurónios, são em alguns casos, valores de entrada de outros, dependendo da interligação existente entre eles. A forma como esta interligação se encontra estabelecida é também importante para o tipo de resultados obtidos.

A anatomia e o funcionamento do neurónio artificial são muito semelhantes à de um neurónio natural (Figura 46). O neurónio natural recebe informação dos seus vizinhos através das suas Dendrites. A informação é processada no Núcleo e o *output* é transportado através do Axónio para os Terminais do Neurónio, de onde é passada aos neurónios seguintes caso o valor da estimulação exceda um determinado valor limiar.

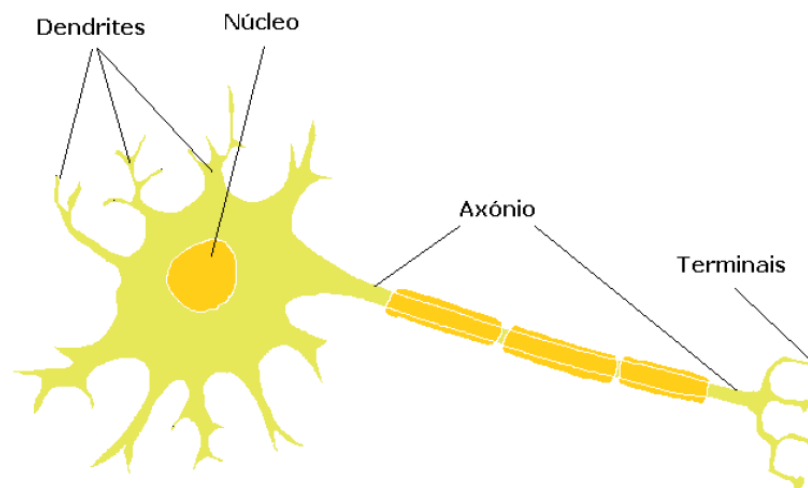


Figura 46 – Estrutura de um neurónio natural típico

(Fonte: www.schizophrenia.com/sznews/archives/005490)

O neurónio artificial, designado também por *nó* (Figura 47), é a unidade de processamento chave para a operação de uma RNA. Embora existam diversos tipos de neurónios artificiais, a norma base desta unidade é a de produzir uma saída, quando o efeito cumulativo das entradas excede um dado valor limite pré-definido. Segundo Hagan et al., um neurónio artificial é composto por três elementos fundamentais:

- 1) **Conjunto de conexões:** que representam as sinapses ou conexões entre os neurónios. Cada conexão tem associado um peso (numero real binário w_{ij}), que tem um efeito *activante* ou *inibidor*, caso os valores sejam respectivamente positivos ou negativos. Assim, o sinal ou estímulo x_j , dado como entrada da

conexão é multiplicado pelo correspondente peso w_{ij} , onde i representa o neurónio objecto de estudo e j o neurónio emissor do sinal. Em algumas situações, pode ainda existir uma conexão extra, denominada de *bias*, cuja entrada é fixada no valor **1**, cuja função é estabelecer as correctas condições operacionais para o neurónio;

- 2) **Integrador (g)**: reduz os n argumentos de entrada (estímulos) a um único valor. Frequentemente é utilizada a função de adição (Σ), pesando todas as entradas numa combinação linear;
- 3) **Função de activação (f)**: que pode condicionar o sinal de saída, introduzindo uma componente de não linearidade no processo computacional. Esta função determina com que grau o resultado obtido pela combinação dos estímulos, causa a excitação ou inibição dos neurónios aos quais este está conectado.

À semelhança do neurónio, o neurónio artificial recebe informação externa à rede através dos estímulos (*inputs*). A informação recebida é processada por uma função de transferência. O valor obtido é então a entrada para a função de activação, cujo resultado é o *output* do neurónio que quando activado envia informação, através de conexões para os neurónios aos quais está ligado.

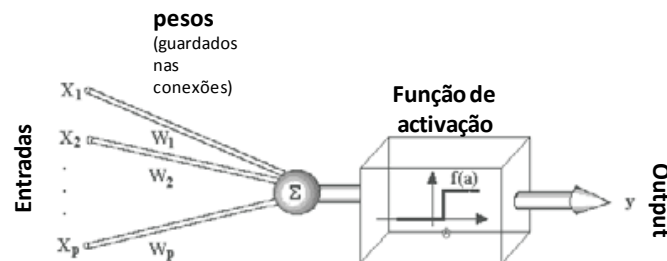
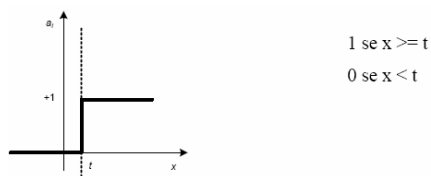


Figura 47 - Estrutura de um neurónio artificial típico

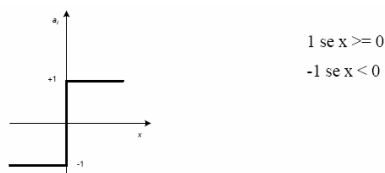
As três funções de activação mais utilizadas em RNA's são:

- 1) **Função degrau ou step function:**



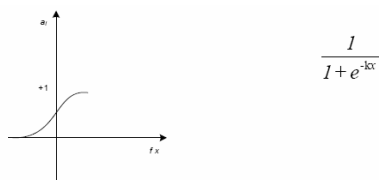
É normalmente utilizada quando se pretende que os neurónios adoptem valores de saída = **+1** apenas se o ganho for não-negativo, de acordo com a filosofia do tudo ou nada;

2) *Função sinal:*



Esta função tem um modo de funcionamento similar a um interruptor (ligado/desligado). Adota valores **+1** ou **-1** consoante a entrada;

3) *Função logística ou sigmóide:*



Esta função é das mais utilizadas em RNA. Trata-se de uma função crescente que exhibe um balanceamento muito bom entre um comportamento linear e não linear. Quando se varia a inclinação **k**, obtêm-se funções com diferentes declives.

Os neurónios interligam-se numa estrutura de rede denominada por **topologia**. Existem vários tipos de topologias de RNA's, que podem ser organizadas nas seguintes categorias:

- 1) **Redes Feedforward de uma só camada - Perceptron** (Figura 48): As RNA *feedforward* podem ser organizadas por camadas, uma vez que não existem ciclos, e dado que as conexões são unidireccionais (sejam convergentes ou divergentes). A topologia mais simples é composta apenas por uma camada de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída. As camadas de entrada não são contabilizadas devido ao facto de nestas não serem efectuados cálculos. Cada neurónio calcula a soma pesada das suas entradas, sendo o valor de saída do tipo binário (0 ou 1) de acordo com um

determinado limite. A função de *Step* é a função de activação comummente utilizada neste tipo de redes.

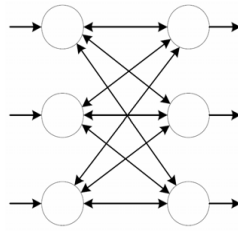


Figura 48 - Rede de uma só camada (*perceptron*)

2) **Redes Feedforward Multicamada** (Figura 49): Esta classe de redes *feedforward* distingue-se por possuir uma ou mais camadas de neurónios intermédias, sendo a sua função intervir de forma útil entre a entrada e a saída da rede. O aumento de camadas intermédias, eleva a capacidade da rede em modelar funções de maior complexidade. No entanto este acréscimo implica o aumento exponencial do tempo necessário para a aprendizagem.

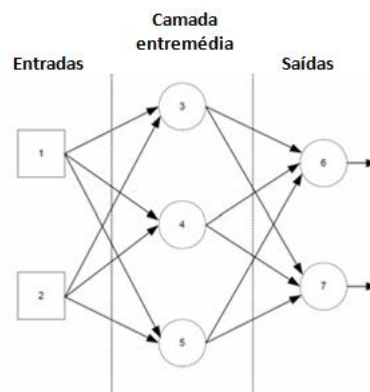


Figura 49 – Topologia de uma rede *feedforward multicamada*

Estas redes constituem uma das mais importantes e populares classes de RNA, sendo utilizadas em múltiplos domínios de aplicação, em problemas de memória associativa, classificação, optimização e regressão, sendo portanto a classe mais utilizada em problemas de *Churn*. As Redes Feedforward Multicamada (RFMC) são caracterizadas por possuírem:

- a) Um conjunto de neurónios de entrada onde surgem os estímulos do ambiente;

- b) Um conjunto de unidades internas de processamento que extraem progressivamente mais características do processo (neurónios intermédios);
- c) Um conjunto de conexões pesadas unidireccionais;
- d) Um conjunto de funções de activação, normalmente do tipo não linear e diferenciável (função logística é a mais utilizada)

O sinal de entrada propaga-se para a frente através da rede, camada por camada. O erro é calculado como a diferença entre a resposta desejada e a resposta produzida pela RNA (aprendizagem supervisionada), ajustando o peso de forma a que se o erro se torne zero.

3) Redes competitivas ou Recorrentes (Figura 50): A recorrência existe em sistemas dinâmicos quando uma saída de um elemento influencia de algum modo a entrada para esse mesmo elemento, criando-se assim um ou mais circuitos fechados. Ao serem incluídas uma ou mais conexões cíclicas numa rede, esta passa a ter um comportamento não linear, de natureza espacial e/ou temporal.

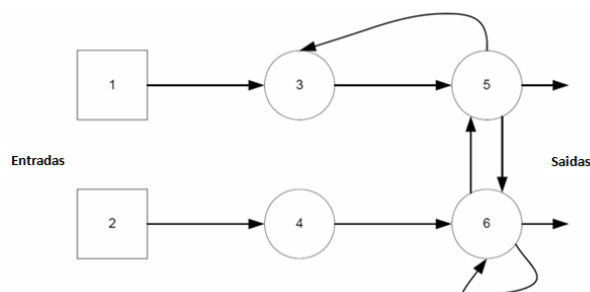


Figura 50 - Topologia típica de uma rede competitiva ou recorrente

A aprendizagem da rede é executada a partir de um algoritmo de aprendizagem que se relaciona com o modelo do ambiente em que opera. Existem três paradigmas fundamentais de aprendizagem em Redes Neurais Artificiais:

- 1) Aprendizagem supervisionada:** perante uma configuração apresentada, a RNA produz uma resposta, que é comparada com a resposta correcta. A rede aprende a partir do conjunto de treino composto por um vector de entrada e um vector de saída. Durante o processo de aprendizagem é efectuada uma comparação entre o

valor desejado e o valor real de saída da rede originando um erro que é utilizado para ajustar os pesos das conexões, reajuste esse que ocorre com o objectivo de minimizar o erro;

2) Aprendizagem de Reforço: Neste caso a resposta correcta não é apresentada à rede. Apenas se fornece a informação se a resposta da rede é correcta ou incorrecta. A partir dessa informação, a rede reajusta-se de forma a melhorar a sua eficácia;

3) Aprendizagem não supervisionada: Segue uma abordagem diferente das duas anteriores, uma vez que não é fornecida ao sistema qualquer indicação externa acerca da resposta correcta. A aprendizagem é realizada através da identificação de características nos dados de entrada, adaptando-se a regularidades estatísticas ou agrupamentos de padrões das instâncias do conjunto de treino.

Nas RNA's, o paradigma de aprendizagem mais comum é o de aprendizagem supervisionada. Assim, o processo de aprendizagem ou treino da rede consiste em ajustar os valores dos pesos das conexões (também denominadas de sinapses) de modo que as entradas na rede produzam as saídas correctas. Para tal, são utilizados algoritmos de treino sendo que o mais conhecido é o ***Backpropagation***.

No algoritmo *Backpropagation*, em cada ciclo de aprendizagem, os erros obtidos pelas diferenças entre as saídas da rede e os valores de treino são propagados para trás, desde os neurónios de saída até aos de entrada, ocorrendo depois um reajuste dos pesos das conexões. O treino termina usualmente quando se obtém o mínimo erro à saída, no caso de regressão, ou o mínimo de classificações erradas.

À primeira vista pode ser criada a ilusão de que quantas mais iterações do algoritmo de aprendizagem forem executadas melhor. Isto pode eventualmente não ser verdade, na medida em que poder-se-á cair numa situação em que a rede se adapta muito bem aos casos de aprendizagem, mas responde mal a outros casos. Neste caso estar-se-á perante uma situação de sobre-ajustamento ou *overfitting*.

É portanto importante ter em consideração os critérios de paragem na aprendizagem de forma a conseguir um equilíbrio entre a precisão e a generalização. Na busca desse difícil equilíbrio, vários critérios de paragem podem ser utilizados, entre eles: erro

máximo, gradiente do erro, número de iterações, ou por validação cruzada, sendo este último método considerado merecedor de maior confiança, mas computacionalmente mais pesado. Quando se opta pelo critério do número de iterações corre-se o risco de sobre-ajustamento, especialmente se esse número for muito elevado. Quanto aos dois outros critérios, dependem muito da escolha do método de cálculo de erro.

O algoritmo *Backpropagation* é computacionalmente pesado e de convergência lenta pelo que surgiram melhorias a este, tais como o *QuickPropagation* desenvolvido por Fahlman em 1998, ou o RPPROP (*Resilient Backpropagation*) proposto por Riedmiller. No entanto, e embora não sendo perfeito, este algoritmo permitiu uma forma automática de ajustamento dos pesos.

Uma das vantagens “competitivas” das RNA’s face a outros algoritmos de DM, é a capacidade de aprender a partir do seu próprio ambiente. O processo de aprendizagem de uma rede neuronal artificial, envolve a seguinte consequência de eventos:

- 1) A RNA é estimulada por um dado ambiente;
- 2) Alguns parâmetros livres (pesos das conexões) são alterados em resultado do estímulo recebido;
- 3) A RNA responde de uma nova forma ao ambiente em virtude das alterações na sua estrutura interna.

As Redes Neurais Artificiais tem-se apresentado cada vez mais como uma solução válida para a resolução de problemas de regressão não linear, uma vez que os resultados obtidos produzem normalmente taxas de erro reduzidas, embora o seu resultado seja uma função matemática que possa ser facilmente utilizada. Na resolução deste tipo de problemas são normalmente utilizadas redes com uma única saída (correspondente à variável cujo valor se pretende prever).

Já no que respeita a problemas de classificação, a utilização de RNA’s tem-se mostrado cada vez mais apelativa, sendo que os softwares desenvolvidos mais recentemente já assentam grande parte das suas soluções nesta metodologia. Neste tipo de problemas a topologia da rede comporta tantas saídas quantas as classes existentes. A saída com valor mais elevado determina a classificação feita pela rede.

As principais desvantagens encontradas na utilização das RNA decorrem do facto de exigirem dados numéricos, de preferência na gama de valores $[0,1]$ ou $[-1,1]$ dependendo da função de activação. Aquando da decisão da adopção deste tipo de técnica, este problema deverá ser tomado em conta e tratado na fase de pré-processamento dos dados, contudo o processo de treino da rede pode revelar-se um problema na medida em que se revela por norma bastante moroso, quando são utilizadas quantidades consideráveis de dados, dando origem à sugestão de várias alterações do método de treino.

II.4.3. Algoritmos e Técnicas de DM – Regressão Logística

De todas as técnicas utilizadas pelos investigadores para separação de instâncias ou exemplos em duas classes, a Análise Discriminante e a Regressão Logística são as mais encontradas na literatura. Estas técnicas enquadram-se na classe de métodos estatísticos multivariados de dependência, uma vez que relacionam um conjunto de variáveis independentes com uma variável dependente categórica (Sharma, 1996; Hair et al., 2005).

Hair et al. (2005) aponta alguns motivos para se optar pela regressão logística face à Análise Discriminante:

- Não é necessário supor normalidade multivariada das variáveis discriminantes;
- Não é necessário que as matrizes de variância/co-variância sejam iguais;
- É uma técnica mais genérica e mais robusta, sendo a sua aplicação apropriada no caso de grande variedade de situações.

Segundo o mesmo autor, a regressão logística, mesmo que tenha uma única variável estatística composta por coeficientes estimados para cada variável independente – como na regressão múltipla – é estimada de uma maneira inteiramente diferente. A regressão múltipla emprega o método de mínimos quadrados, o qual minimiza a soma das diferenças quadradas entre os valores reais e os previstos para a variável dependente. A natureza não linear da transformação logística adopta o método da máxima

verossimilhança de forma iterativa para encontrar as estimativas “mais prováveis” para os coeficientes.

Uma das vantagens mais evidentes da Regressão Logística é que é preciso saber apenas se um evento ocorreu para então se usar um valor binário como variável dependente. A partir desse valor dicotômico (binário), o procedimento prevê a sua estimativa de probabilidade de que o evento ocorrerá ou não. Assim cada elemento é classificado de acordo com a maior probabilidade prevista de pertencer a um grupo.

O procedimento que calcula o coeficiente logístico, compara a probabilidade de um evento ocorrer com a probabilidade de ele não ocorrer. Essa razão de desigualdade pode ser expressa como:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i)$$

Onde os coeficientes estimados $\beta_0, \beta_1, \beta_2, \dots, \beta_i$, são na verdade, medidas das variações na proporção das probabilidades, chamada de razão de desigualdade. Coeficiente positivo indica aumento da probabilidade, o contrário representa a diminuição da probabilidade prevista.

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada directamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1: se o elemento pertence ao grupo ou 0: caso contrário) e haver um conjunto de i variáveis independentes $x_1, x_2, \dots, x_i$, o modelo de regressão logística pode ser registado da seguinte forma:

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i)}$$

Considerando certa combinação de coeficientes $\beta_0, \beta_1, \beta_2, \dots, \beta_i$ (que são estimados a partir do conjunto de dados pelo método da máxima verossimilhança) e variando os valores de $x_1, x_2, \dots, x_i$, observa-se que a curva logística tem comportamento probabilístico no formato da letra S, o que é característica da regressão logística (Figura 51).

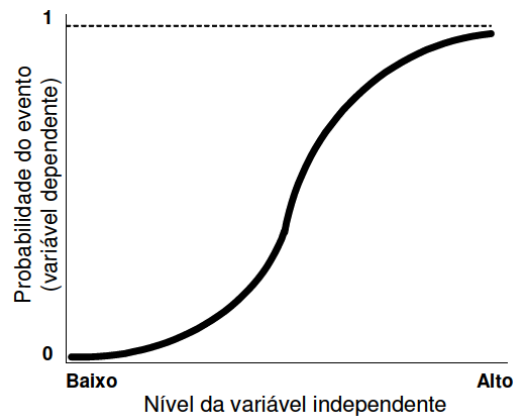


Figura 51 -Forma da relação logística entre variáveis dependente e independente (by Hair et. al, 2005).

Esse formato dá à Regressão Logística um elevado grau de generalidade, aliada a aspectos muito desejáveis:

- Quando ;
- Quando .

Assim, e uma vez que através deste modelo é possível estimar directamente a probabilidade de ocorrência de um evento, pode também estimar-se a probabilidade de não ocorrência, através da diferença: . A principal suposição da utilização de modelos de RL, é que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear, onde a expressão de probabilidade é denominada de *Odds*⁴⁷:

A variável *Odds*, resolve os problemas de estimativa de probabilidades entre 0 e 1, abaixo desse limite a solução é o logaritmo de *Odds*, denominada de função *logit*:

⁴⁷ *Odds* – Segundo Hair et. al, 1995, é a razão entre a probabilidade de um evento ocorrer sobre a probabilidade do mesmo não ocorrer, esta razão é utilizada para mensurar a variável dependente em regressão logística.

Posto isto, ao interpretar os coeficientes da regressão logística, opta-se pela interpretação de β e não directamente de σ . Para utilizar o modelo de RL para discriminação de dois grupos, a regra de classificação é a seguinte:

- Se $P(Y=1) > 0.5$, então classifica-se $Y=1$ (*churner*);
- Caso contrário, classifica-se $Y=0$ (*não churner*).

Para comprovar a adequação do modelo, deve verificar-se a precisão dos parâmetros estimados β , construir intervalos de confiança, testar hipóteses e por ultimo realizar uma análise de diagnóstico e de resíduos.

No caso específico deste estudo, o modelo ajustado tem como objectivo principal discriminar os indivíduos que deixarão de comprar e os que permanecerão clientes. Existem várias medidas que permitem medir e comparar o desempenho de modelos na realização desse propósito. Três das medidas de desempenho mais populares são: curva ROC, estatística de Kolmogorov-Smirnov e distância de Mahalanobis.

ANEXO III

III.1. Caracterização do *data set*

III.1.1. As variáveis

III.1.1.1. Variáveis Socio-Demográficas

Foram recolhidas as variáveis disponíveis no CRM da SonaeMC e compilados com variáveis calculadas e transformadas para o efeito: Índice de Poder de compra P/C e Índice de Concorrência. Esta última variável foi criada com o intuito de mensurar até que ponto o factor concorrencial dita o *churn* dos clientes.

Nome	Descrição	Tipo dados	Observações	Valores Possíveis
CUSTOMER_ACCOUNT_CD	Identificação do Cliente	NOMINAL		
GENDER	Código do Género do cliente	NOMINAL		F, M, ND
SEGM_AGE	Segmento de Idade do Cliente	NOMINAL		SA_0 - No Value; SA_1 - [0;18]; SA_2 -]18;25]; SA_3 -]25;35]; SA_4 -]35;45]; SA_5 -]45;55]; SA_6 -]55;65]; SA_7 - >65
HOUSEHOLD	Agregado Familiar do Cliente	NOMINAL		SHS_0 - No Value; SHS_1 - 1; SHS_2 - 2; SHS_3 - 3; SHS_4 - 4; SHS_5 - 5; SHS_6 - >5;
PREF_STORE	Loja Preferencial do Cliente	NOMINAL	Loja onde o cliente efectua o maior valor em compras.	
MUNICIPALITY_CD	Concelho residência do cliente	NOMINAL		
DISTRICT_CD	Distrito residência do cliente	NOMINAL		
IPC	Índice de Poder de Compra P/C	INTERVAL	Adoptado o Índice de Poder de compra do Concelho de residência do Cliente (Fonte: INE)	
COMPETITIONINDEX	Índice de Concorrência	INTERVAL	Calculado com base nos metros quadrados de concorrência existente nas freguesias até 20 minutos da loja preferencial.	

Tabela 15 - Conjunto de variáveis Socio-Demográficas

III.1.1.2. Variáveis Comportamentais

Este tipo de informação agrupa todo o tipo de registos relacionados com o consumo do cliente, no que respeita a bens ou serviços, bem como o valor de cada cliente nas diferentes segmentações adoptadas pela empresa.

Tipo	Nome	Descrição	Tipo de dados	Observações	Valores Possíveis
Comportamental	SEGM_LIFESTYLE	Segmento Vida do Cliente	NOMINAL	Segmentação Orientada para a tipologia de consumo do cliente (com base num histórico de 12 meses)	KC, KB, TS, BF, PP, WH, AA, NV
Comportamental	SEGM_VALUE	Segmento Valor do Cliente	NOMINAL	Segmentação com base na frequência de compra e valor médio de compra por visita	SV_0 - No Value; SV_1 - Loyal Large; SV_2 - Loyal Medium; SV_3 - Loyal Small; SV_4 - Frequent Medium; SV_5 - Frequent Large; SV_6 - Frequent Small; SV_7 - Occasional Medium; SV_8 - Occasional Small
Comportamental	SEGM_SPENDVARIATION	Segmento de Variação de Gasto	NOMINAL	Inactive Active: New, High, Low, Stable	SCC_2 - Inactive; SCC_3 - New; SCC_5 - High; SCC_6 - Low; SCC_7 - Stable
Comportamental	SEGM_RECENCY	Segmento Abandono	NOMINAL	Inactive: N/A Active: número de meses desde a última compra	SRQ_1 - Last Buy [0;3] M; SRQ_2 - Last Buy [3;6] M; SRQ_3 - Last Buy [6;9] M; SRQ_4 - Last Buy [9;12] M; SRQ_7 - Inactive
Comportamental	N_RECLAMACOES	Nº reclamações efectuadas nos últimos 12 meses	INTERVAL		
Comportamental	N_CONTACTOS_MC	Nº contactos promoções direccionadas	INTERVAL	Nº de cartas com ofertas recebidas	
Comportamental	N_RESPOSTAS_MC	Nº respostas positivas a promoções direccionadas	INTERVAL	Nº de acções a que o cliente respondeu	
Comportamental	TOP1_CAT_10	Top1 Categoria mais vezes comprada	NOMINAL	em Alimentar	
Comportamental	TOP2_CAT_10	Top2 Categoria mais vezes comprada	NOMINAL	em Alimentar	
Comportamental	TOP3_CAT_10	Top3 Categoria mais vezes comprada	NOMINAL	em Alimentar	
Comportamental	TOP1_CAT_11	Top1 Categoria mais vezes comprada	NOMINAL	em Perecíveis	
Comportamental	TOP2_CAT_11	Top2 Categoria mais vezes comprada	NOMINAL	em Perecíveis	
Comportamental	TOP3_CAT_11	Top3 Categoria mais vezes comprada	NOMINAL	em Perecíveis	
Comportamental	TOP1_CAT_12	Top1 Categoria mais vezes comprada	NOMINAL	em Take Away	
Comportamental	TOP2_CAT_12	Top2 Categoria mais vezes comprada	NOMINAL	em Take Away	
Comportamental	TOP3_CAT_12	Top3 Categoria mais vezes comprada	NOMINAL	em Take Away	
Comportamental	TOP3_CAT_23	Top1 Categoria mais vezes comprada	NOMINAL	em Bazar	
Comportamental	TOP3_CAT_23	Top2 Categoria mais vezes comprada	NOMINAL	em Bazar	
Comportamental	TOP3_CAT_23	Top3 Categoria mais vezes comprada	NOMINAL	em Bazar	
Comportamental	TOP3_CAT_24	Top1 Categoria mais vezes comprada	NOMINAL	em Têxtil	
Comportamental	TOP3_CAT_24	Top2 Categoria mais vezes comprada	NOMINAL	em Têxtil	
Comportamental	TOP3_CAT_24	Top3 Categoria mais vezes comprada	NOMINAL	em Têxtil	
Comportamental	PROD_PROM_ACTUAL_NR_PR	Proporção do número de artigos distintos com desconto no mês actual	INTERVAL		
Comportamental	TRANSAC_FDS_ACTUAL_PR	Proporção do número de compras com promoção no mês actual	INTERVAL		
Comportamental	PROD_DESC_ACTUAL_NR_PR	Proporção do montante comprado em promoção no mês actual	INTERVAL		

Tabela 16 - Conjunto de variáveis Comportamentais (Parte 1)

Foi definido neste conjunto de variáveis um grupo de atributos relativos ao gasto específico em áreas de negócio que foram definidas como condutoras do comportamento do cliente. De forma a facilitar a visualização das variáveis, entenda-se por “*” o nome da área de negócio correspondente (Tabela 17).

Nome	Descrição	Tipo dados
*_CAT_6M_VL	Montante total comprado na Área de negócio * nos últimos 6 meses	INTERVAL
*_RACIO_CAT_3_6_VL	Rácio do Montante total comprado na Área de Negócio * nos últimos 3 meses / 6 meses	INTERVAL
*_CAT_6M_VISITAS	Número de visitas à Área de Negócio * nos últimos 6 meses	INTERVAL
*_RACIO_CAT_3_6_VIS	Rácio do Nº de visitas na Área de Negócio * nos últimos 3 meses / 6 meses	INTERVAL
*_FLAG_VISIT_LAST12M	Flag de Compras na Área de Negócio * nos últimos 12 meses	INTERVAL
*_INDEX_3M_VISITAS	Número de visitas na Área de Negócio nos últimos 3 meses sobre a média das visitas em do total de clientes na Área de Negócio *	INTERVAL
*_MAXPRD_MLOJ_6M_NR	Número máximo de produtos distintos comprados em cada loja na Área de Negócio * nos últimos 6 meses	INTERVAL

Descrição possível do sufixo *	Área de Negócio
GROCERIES	Mercearia
DIETETICS	Dietéticos
DRINKS	Bebidas ã Alcoolicas
HYGIENICPRD	DPH
BUTCHER	Talho
BIRDS	Aves
FISH	Peixaria
CODFISH	Bacalhau
DELICATESSEN	Charcutaria
FRTSVGTBLS	Frutas e Legumes
BAKERY	Padaria
PETCARE	Pet Care
CULTURE	Cultura
TOYS	Brinquedos
TAKEAWAY	Take Away
TEXTIL	Textil
LIGHTBAZAAR	Bazar Ligeiro
ALCHLCDRNKS	Bebidas Alcoolicas
PERSONALHYGIENE	Higiene Pessoal
DAIRY	Lactínicos
FROZEN	Congelados

Tabela 17 - Conjunto de Variáveis Comportamentais (Parte 2)

III.1.1.3. Variáveis de Envolvimento

Este conjunto de variáveis contém os dados relativos aos dados agregados de compra do cliente, nomeadamente informação de carácter temporal e que permitirá a definição da própria variável Target.

Nome	Descrição	Tipo dados
LOJAS_ACTUAL_NR	Número de lojas totais distintas frequentadas no mês actual	INTERVAL
TRANSAC_ACTUAL_NR	Número de transacções no mês actual	INTERVAL
GASTO_ACTUAL_VL	Montante total gasto no mês actual	INTERVAL
TICKET_MED_ACTUAL	Ticket Médio por visita no mês actual	INTERVAL
VISITAS_ACTUAL_NR	Número total de visitas no mês actual	INTERVAL
DESCONTOS_ACTUAL_VL	Total de descontos obtidos no mês actual	INTERVAL
PROD_DESC_ACTUAL_VL_PR	Proporção do montante de descontos ao artigo no mês actual	INTERVAL
PROD_PROM_ACTUAL_VL_PR	Proporção do montante de artigos com promoção no mês actual	INTERVAL
TRANS_DESC_ACTUAL_VL_PR	Proporção do montante de descontos à transacção no mês actual	INTERVAL
DEVOLUC_ACTUAL_PR	Proporção do montante de artigos devolvidos no mês actual	INTERVAL
MCH_ACTUAL_PR	Proporção do montante comprado nas insígnias MCH no mês actual	INTERVAL
CATEGORIAS_ACTUAL_NRS	Número de categorias compradas no mês actual	INTERVAL
TRANSAC_ULT_xM_NR	Número de transacções nos últimos x meses	INTERVAL
GASTO_ULT_xM_VL	Montante total gasto nos últimos x meses	INTERVAL
TICKET_MED_ULT_xM	Ticket Médio por visita nos últimos x meses	INTERVAL
VISITAS_ULT_xM_NR	Número total de visitas nos últimos x meses	INTERVAL
DESCONTOS_ULT_xM_VL	Total de descontos obtidos nos últimos x meses	INTERVAL
PROD_DESCONTOS_ULT_xM_VL_PR	Proporção do montante de descontos ao artigo nos últimos x meses	INTERVAL
PROD_PROMOCAO_ULT_xM_VL_PR	Proporção do montante de artigos com promoção nos últimos x meses	INTERVAL
TRANS_DESCONTOS_ULT_xM_VL_PR	Proporção do montante de descontos à transacção nos últimos x meses	INTERVAL
DEVOLUC_ULT_xM_PR	Proporção do montante de artigos devolvidos nos últimos x meses	INTERVAL
MCH_ULT_xM_PR	Proporção do montante comprado nas insígnias MCH nos últimos x meses	INTERVAL

Observacoes
x meses assume 3, 6 ou 12

Tabela 18 - Conjunto de Variáveis de Envolvimento

III.1.2. Preparação dos Dados

III.1.2.1. Selecção dos dados e partição da base de dados

Foi necessário proceder a uma pré-selecção dos dados e posterior partição da base de dados inicial. Seguidamente foi criada a variável Target (*churn* =1 ; *não churn* =0), tendo como auxiliar o cruzamento entre a ABT de Março e a ABT de Abril.

Esta operação foi realizada em SAS Guide e está documentada na Figura 52.

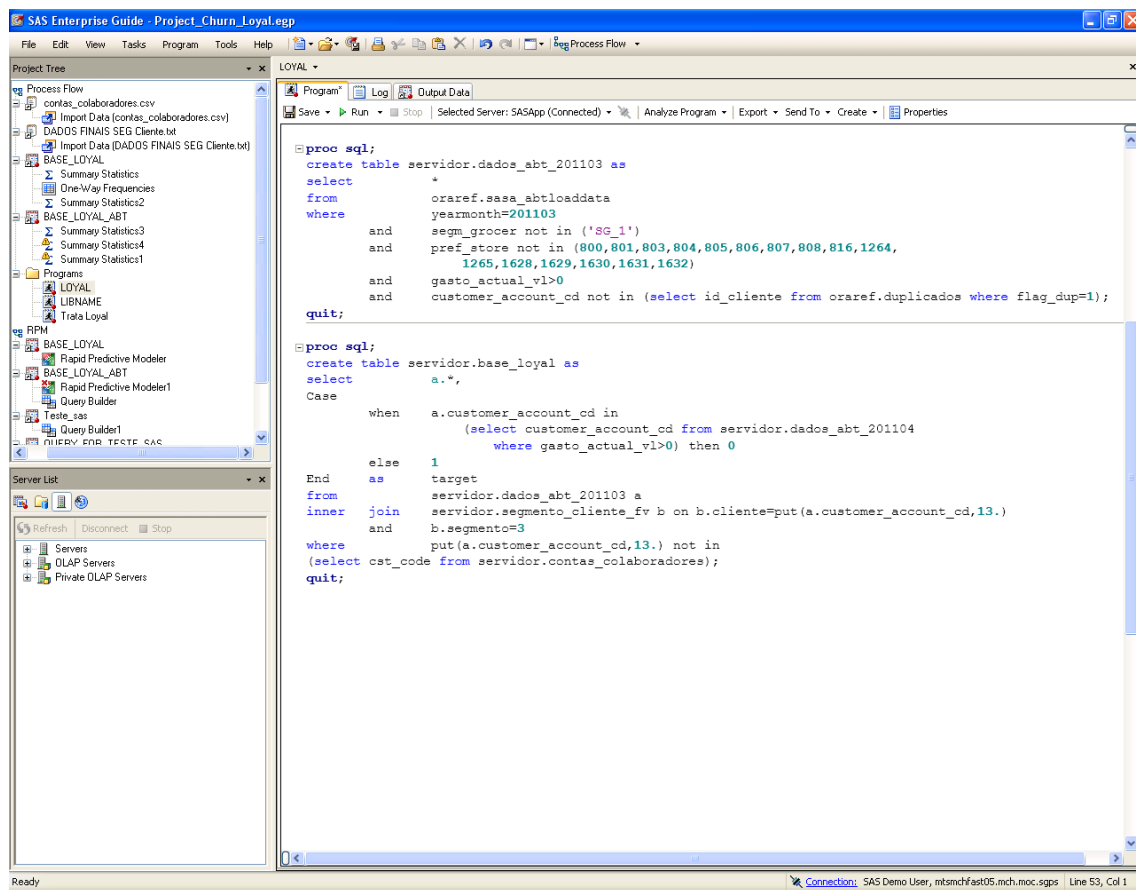


Figura 52 - Pré-Seleção e Partição inicial da BD em SAS Guide

III.1.3. Estatísticas descritivas das variáveis – Base Loyal

III.1.3.1. Variáveis Socio-Demográficas

IPC

target = 0

Basic Statistical Measures			
Location		Variability	
Mean	113.003	Std Deviation	40.96551
Median	100.4	Variance	1678
Mode	100.4	Range	189.86
		Interquartile Range	42.2

target = 1

Basic Statistical Measures			
Location		Variability	
Mean	116.9848	Std Deviation	43.22697
Median	107.46	Variance	1869
Mode	235.74	Range	189.86
		Interquartile Range	36.02

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	45.88	11	275	134.02	413
2	46.34	3	276	135.44	2503
3	46.95	1	277	137.64	1898
4	47.55	48	278	139.13	4370
5	47.58	8	279	141.55	645
6	47.64	2	280	144.81	395
7	47.71	1	281	155.74	3726
8	47.75	4	282	170.5	5208
9	47.95	8	283	172.95	4418
10	48.29	6	284	235.74	8292

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	45.88	1	126	134.02	2
2	47.58	1	127	135.44	22
3	49.26	1	128	137.64	13
4	51.26	3	129	139.13	17
5	51.83	1	130	141.55	7
6	51.84	1	131	144.81	4
7	51.98	6	132	155.74	26
8	52.41	1	133	170.5	36
9	53.39	3	134	172.95	38
10	54.03	1	135	235.74	74

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	14826	9.43	100

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	284	21.86	100

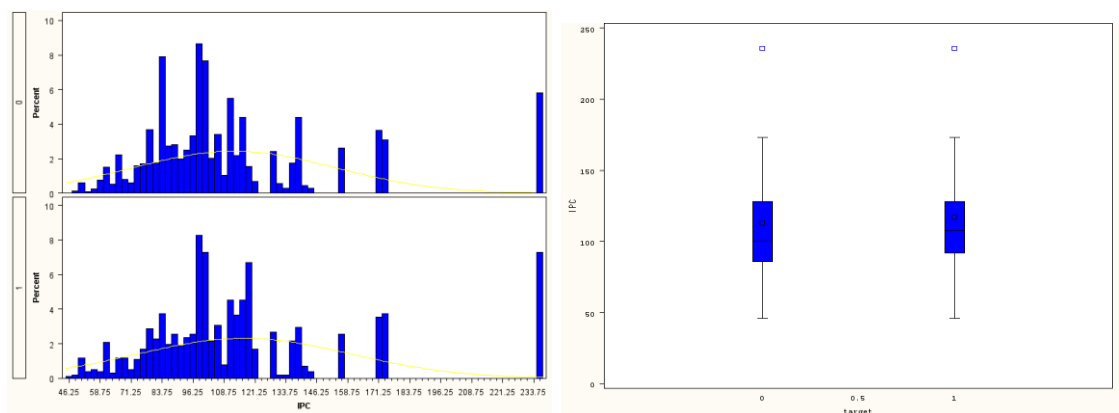


Figura 53 – Histograma e Box-Plot da variável IPC original

COMPETITIONINDEX

target = 0

Basic Statistical Measures			
Location		Variability	
Mean	1.322263	Std Deviation	1.29035
Median	0.75	Variance	1.66501
Mode	0.29	Range	4.39
		Interquartile Range	1.75

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	916	0.58	100

target = 1

Basic Statistical Measures			
Location		Variability	
Mean	1.252045	Std Deviation	1.2572
Median	0.64	Variance	1.58055
Mode	0.27	Range	4.39
		Interquartile Range	1.77

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	13	1	100

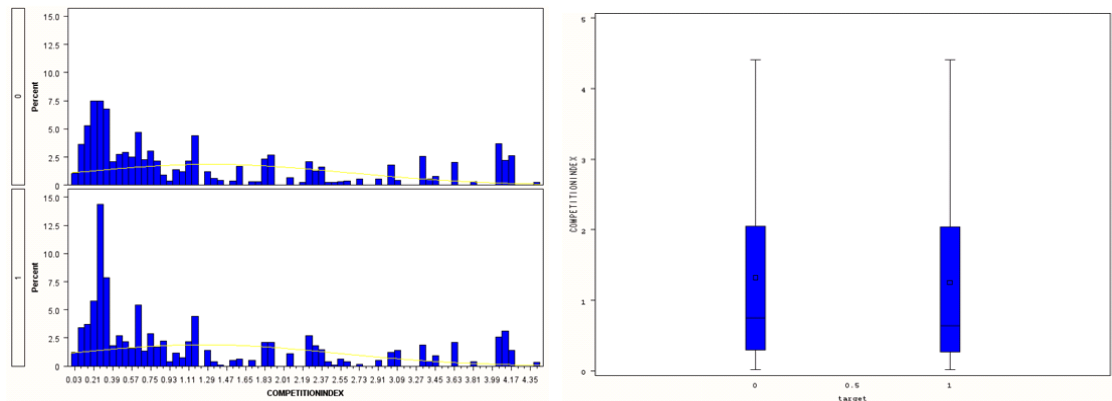


Figura 54 - Histograma e Box-Plot da variável original CompetitionIndex

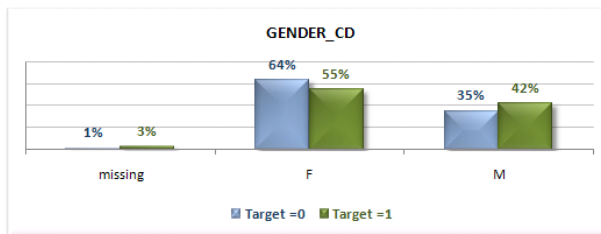


Figura 55 - Frequências da variável Gender_CD

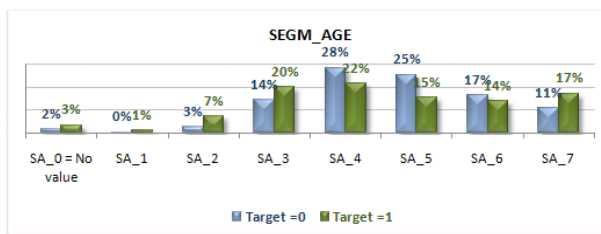


Figura 56 - Frequências da variável Segm_Age

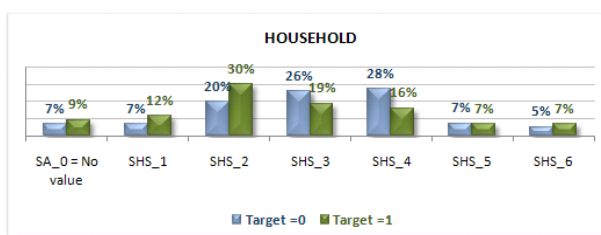


Figura 57 - Frequências da variável HouseHold

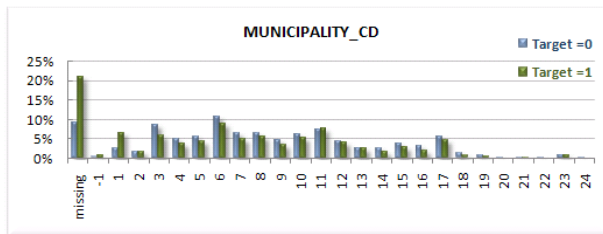


Figura 58 - Frequências da variável Municipality_cd

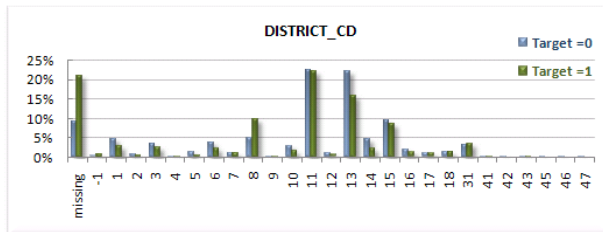


Figura 59 - Frequências da variável District_cd

III.1.3.2. Variáveis Comportamentais

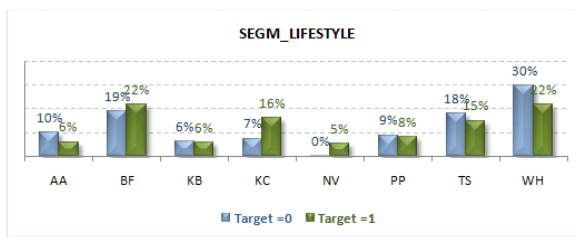


Figura 60 - Frequências da variável Segm_LifeStyle

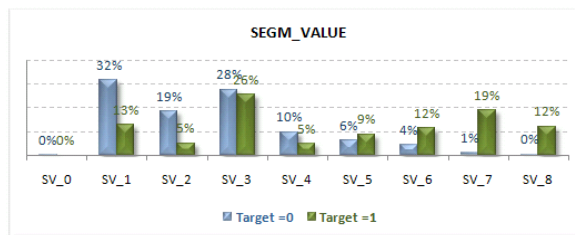


Figura 61- Frequências da variável Segm_Value

As variáveis TOP_CAT* foram brevemente analisadas, no entanto, e uma vez que se referem a categorias em que os clientes efectuaram compras, optou-se por não fazer uma análise exploratória dos dados, deixando que os modelos criados decidissem quanto à sua relevância.

Na TABELA estão resumidas algumas das estatísticas observadas para analisar descritivamente as restantes variáveis comportamentais.

Leia-se a legenda da tabela:

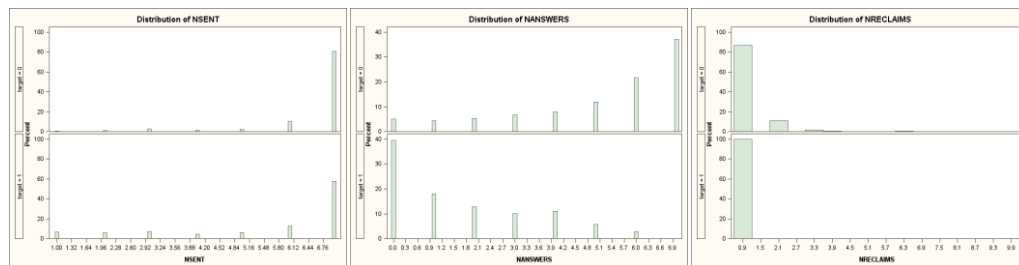
Decisão Dq* - classe com mais observações proximas da média
Observações

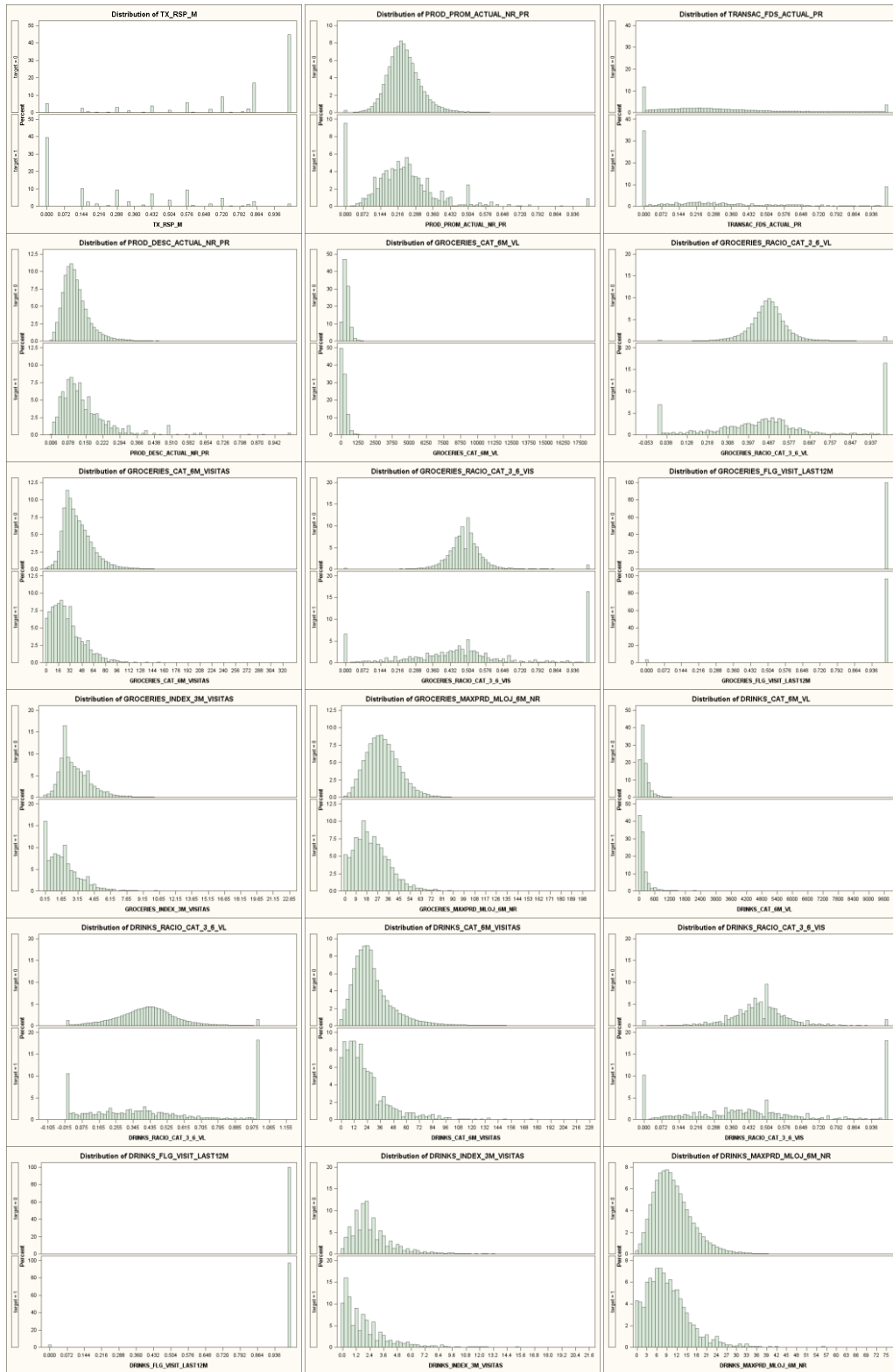
- 1) Variável rejeitada - elevado numero de *missing values*
- 2) Tratamento de *missing values* necessario
- 3) Valores negativos relativos a vendas indicam **presença de outliers**
- 4) Variável rejeitada - >= 25% de obs = 0

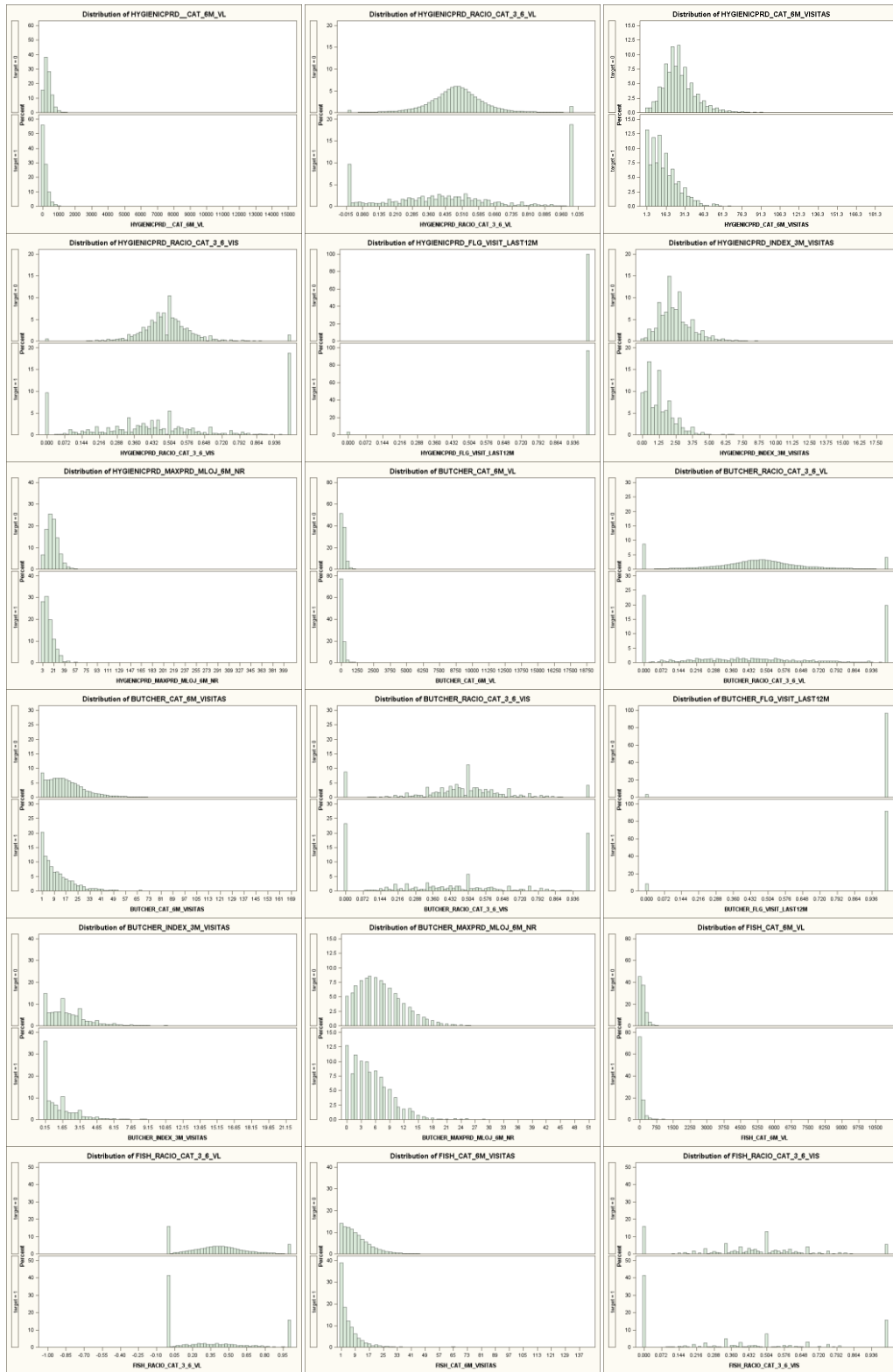
target = 0										target = 1					
Variable	Obs.	Decisão Dq*	CV 1 / CV0	Mean	Coef Var*	Minimum	% Miss	amplitude semi-interquartil (Dq)	Mean	Coef Var*	Minimum	Mode	% Miss	amplitude semi-interquartil (Dq)	
NSENT	2)	não churn	2.08	6.58	17%	1.00	3%	0	5.66	35%	1.00	7	14%	1	
NANSWERS	2)		2.65	5.16	41%	0.00	3%	2	1.63	108%	0.00	0	14%	2	
NRECLAIMS	1)		0.00	1.18	49%	1.00	99%	0	1.00	0%	1.00	1	100%	0	
TX_RSP_M	2)	não churn	2.80	0.77	38%	0.00	3%	0	0.26	106%	0.00	0	14%	0	
PROD_PROM_ACTUAL_NR_PR	2)	não churn	2.11	0.24	30%	0.00	0%	0	0.24	63%	0.00	0	3%	0	
TRANSAC_FDS_ACTUAL_PR	2)	não churn	1.35	0.34	84%	0.00	0%	0	0.29	113%	0.00	0	3%	0	
PROD_DESC_ACTUAL_NR_PR	2)	não churn	1.40	0.11	55%	0.00	1%	0	0.14	77%	0.01	0	22%	0	
GROCERIES_CAT_6M_VL	3)	churn	1.63	366.54	67%	-14.03	0%	135	194.09	109%	0.00	0	0%	115	
GROCERIES_RATIO_CAT_3_6_VL	3)	não churn	2.67	0.49	21%	-0.06	0%	0	0.51	57%	0.00	1	0%	0	
GROCERIES_CAT_6M_VISITAS		não churn	1.59	41.94	49%	0.00	0%	13	26.88	78%	0.00	0	0%	13	
GROCERIES_RATIO_CAT_3_6_VIS		não churn	2.95	0.50	19%	0.00	0%	0	0.51	55%	0.00	1	0%	0	
GROCERIES_FLG_VISIT_LAST12M			6.04	1.00	3%	0.00	0%	0	0.97	18%	0.00	1	0%	0	
GROCERIES_INDEX_3M_VISITAS		não churn	1.59	2.89	51%	0.00	0%	1	1.83	81%	0.00	0	0%	1	
GROCERIES_MAXPRD_MLOJ_6M_NR			1.49	31.15	44%	0.00	0%	10	21.78	66%	0.00	0	0%	10	
DIETETICS_CAT_6M_VL	4)	churn	1.35	9.96	291%	0.00	0%	4	4.71	394%	0.00	0	0%	1	
DIETETICS_RATIO_CAT_3_6_VL	4)	churn	1.51	0.27	140%	-0.74	0%	0	0.16	212%	0.00	0	0%	0	
DIETETICS_CAT_6M_VISITAS	4)	churn	1.40	2.49	204%	0.00	0%	2	1.12	286%	0.00	0	0%	1	
DIETETICS_RATIO_CAT_3_6_VIS	4)	churn	1.51	0.27	139%	0.00	0%	0	0.16	210%	0.00	0	0%	0	
DIETETICS_FLG_VISIT_LAST12M	4)		1.51	0.64	76%	0.00	0%	1	0.43	114%	0.00	0	0%	1	
DIETETICS_INDEX_3M_VISITAS	4)	churn	1.50	0.59	210%	0.00	0%	0	0.26	315%	0.00	0	0%	0	
DIETETICS_MAXPRD_MLOJ_6M_NR	4)	churn	1.52	1.15	158%	0.00	0%	1	0.63	240%	0.00	0	0%	1	
DRINKS_CAT_6M_VL		churn	1.36	194.32	125%	0.00	0%	87	144.01	170%	0.00	0	0%	70	
DRINKS_RATIO_CAT_3_6_VL	3)	não churn	1.84	0.42	41%	-0.11	0%	0	0.45	76%	0.00	1	0%	0	
DRINKS_CAT_6M_VISITAS		não churn	1.38	29.39	70%	0.00	0%	10	21.99	96%	0.00	0	0%	11	
DRINKS_RATIO_CAT_3_6_VIS		não churn	2.32	0.48	29%	0.00	0%	0	0.48	67%	0.00	1	0%	0	
DRINKS_FLG_VISIT_LAST12M			4.79	1.00	4%	0.00	0%	0	0.97	17%	0.00	1	0%	0	
DRINKS_INDEX_3M_VISITAS		não churn	1.43	2.66	76%	0.00	0%	1	1.93	108%	0.00	0	0%	1	
DRINKS_MAXPRD_MLOJ_6M_NR		não churn	1.30	10.88	54%	0.00	0%	4	9.25	71%	0.00	6	0%	4	
HYGIENICPRD_CAT_6M_VL		churn	1.66	318.38	73%	0.00	0%	142	147.03	121%	0.00	0	0%	84	
HYGIENICPRD_RATIO_CAT_3_6_VL	3)	não churn	2.29	0.49	29%	-0.02	0%	0	0.49	66%	0.00	1	0%	0	
HYGIENICPRD_CAT_6M_VISITAS			1.58	26.55	49%	0.00	0%	8	14.17	77%	0.00	1	0%	8	
HYGIENICPRD_RATIO_CAT_3_6_VIS		não churn	2.62	0.49	24%	0.00	0%	0	0.49	63%	0.00	1	0%	0	
HYGIENICPRD_FLG_VISIT_LAST12M			5.39	1.00	3%	0.00	0%	0	0.97	18%	0.00	1	0%	0	
HYGIENICPRD_INDEX_3M_VISITAS		churn	1.60	2.49	52%	0.00	0%	1	1.24	83%	0.00	0	0%	1	
HYGIENICPRD_MAXPRD_MLOJ_6M_NR		churn	1.54	18.35	52%	0.00	0%	7	11.65	80%	0.00	7	0%	6	
BUTCHER_CAT_6M_VL		churn	1.31	168.55	127%	0.00	0%	93	88.41	165%	0.00	0	0%	49	
BUTCHER_RATIO_CAT_3_6_VL		não churn	1.63	0.46	50%	0.00	0%	0	0.44	82%	0.00	0	0%	0	
BUTCHER_CAT_6M_VISITAS		churn	1.28	17.36	85%	0.00	0%	9	10.78	109%	0.00	0	0%	7	
BUTCHER_RATIO_CAT_3_6_VIS		não churn	1.72	0.47	46%	0.00	0%	0	0.45	79%	0.00	0	0%	0	
BUTCHER_FLG_VISIT_LAST12M			1.70	0.97	18%	0.00	0%	0	0.92	30%	0.00	1	0%	0	
BUTCHER_INDEX_3M_VISITAS		churn	1.31	2.15	89%	0.00	0%	1	1.30	117%	0.00	0	0%	1	
BUTCHER_MAXPRD_MLOJ_6M_NR		churn	1.20	7.09	68%	0.00	0%	4	5.05	82%	0.00	0	0%	3	
FISH_CAT_6M_VL	3)	churn	1.44	131.76	141%	-9.95	0%	75	65.80	204%	0.00	0	0%	37	
FISH_RATIO_CAT_3_6_VL	3)	não churn	1.62	0.39	69%	-1.00	0%	0	0.33	113%	0.00	0	0%	0	
FISH_CAT_6M_VISITAS		churn	1.59	9.56	95%	0.00	0%	5	4.90	151%	0.00	0	0%	3	
FISH_RATIO_CAT_3_6_VIS		não churn	1.74	0.42	60%	0.00	0%	0	0.35	105%	0.00	0	0%	0	
FISH_FLG_VISIT_LAST12M			2.18	0.96	21%	0.00	0%	0	0.83	46%	0.00	1	0%	0	
FISH_INDEX_3M_VISITAS		churn	1.60	1.60	104%	0.00	0%	1	0.77	166%	0.00	0	0%	1	
FISH_MAXPRD_MLOJ_6M_NR		não churn	1.42	3.83	74%	0.00	0%	2	2.42	105%	0.00	0	0%	2	
DELICATESSEN_CAT_6M_VL	3)	churn	1.44	131.76	141%	-9.95	0%	75	65.80	204%	0.00	0	0%	37	
DELICATESSEN_RATIO_CAT_3_6_VL	3)	não churn	1.62	0.39	69%	-1.00	0%	0	0.33	113%	0.00	0	0%	0	
DELICATESSEN_CAT_6M_VISITAS		churn	1.59	9.56	95%	0.00	0%	5	4.90	151%	0.00	0	0%	3	
DELICATESSEN_RATIO_CAT_3_6_VIS		não churn	1.74	0.42	60%	0.00	0%	0	0.35	105%	0.00	0	0%	0	
DELICATESSEN_FLG_VISIT_LAST12M			2.18	0.96	21%	0.00	0%	0	0.83	46%	0.00	1	0%	0	
DELICATESSEN_INDEX_3M_VISITAS		churn	1.60	1.60	104%	0.00	0%	1	0.77	166%	0.00	0	0%	1	
DELICATESSEN_MAXPRD_MLOJ_6M_NR		não churn	1.42	3.83	74%	0.00	0%	2	2.42	105%	0.00	0	0%	2	
FRTSVGTBLS_CAT_6M_VL	3)	churn	1.43	190.12	88%	-3689.02	0%	86	99.88	126%	0.00	0	0%	54	
FRTSVGTBLS_RATIO_CAT_3_6_VL	3)	não churn	2.19	0.48	32%	-0.23	0%	0	0.47	70%	0.00	1	0%	0	
FRTSVGTBLS_CAT_6M_VISITAS		churn	1.41	34.02	60%	0.00	0%	11	20.17	84%	0.00	1	0%	10	
FRTSVGTBLS_RATIO_CAT_3_6_VIS		não churn	2.52	0.50	26%	0.00	0%	0	0.49	65%	0.00	1	0%	0	
FRTSVGTBLS_FLG_VISIT_LAST12M			4.11	1.00	4%	0.00	0%	0	0.98	16%	0.00	1	0%	0	
FRTSVGTBLS_INDEX_3M_VISITAS		não churn	1.53	2.80	63%	0.00	0%	1	1.65	96%	0.00	0	0%	1	
FRTSVGTBLS_MAXPRD_MLOJ_6M_NR		não churn	1.36	17.61	52%	0.00	0%	6	13.61	71%	0.00	10	0%	7	
BAKERY_CAT_6M_VL	3)	churn	1.40	92.24	82%	-1.02	0%	41	50.50	114%	0.00	0	0%	28	
BAKERY_RATIO_CAT_3_6_VL		não churn	1.90	0.46	38%	0.00	0%	0	0.47	72%	0.00	1	0%	0	
BAKERY_CAT_6M_VISITAS		churn	1.31	34.17	74%	0.00	0%	13	21.09	97%	0.00	0	0%	11	
BAKERY_RATIO_CAT_3_6_VIS		não churn	2.20	0.49	30%	0.00	0%	0	0.49	67%	0.00	1	0%	0	
BAKERY_FLG_VISIT_LAST12M			3.91	1.00	5%	0.00	0%	0	0.97	19%	0.00	1	0%	0	
BAKERY_INDEX_3M_VISITAS		churn	1.38	2.84	79%	0.00	0%	1	1.73	109%	0.00	0	0%	1	
BAKERY_MAXPRD_MLOJ_6M_NR			1.28	8.09	52%	0.00	0%	3	6.17	66%	0.00	6	0%	3	
CULTURE_CAT_6M_VL	3)	churn	2.52	54.20	223%	-58.41	0%	29	30.96	562%	0.00	0	0%	12	
CULTURE_RATIO_CAT_3_6_VL	3)	não churn	1.58	0.40	78%	-5.60	0%	0	0.31	123%	0.00	0	0%	0	
CULTURE_CAT_6M_VISITAS		churn	1.30	10.27	131%	0.00	0%	6	4.47	169%	0.00	0	0%	3	
CULTURE_RATIO_CAT_3_6_VIS		não churn	1.73	0.42	68%	0.00	0%	0	0.31	118%	0.00	0	0%	0	
CULTURE_FLG_VISIT_LAST12M			2.29	0.94	25%	0.00	0%	0	0.76	57%	0.00	1	0%	0	
CULTURE_INDEX_3M_VISITAS		churn	1.29	1.76	139%	0.00	0%	1	0.69	179%	0.00	0	0%	0	
CULTURE_MAXPRD_MLOJ_6M_NR		churn	1.45	4.38	93%	0.00	0%	2	2.40	135%	0.00	0	0%	2	
TAKEAWAY_CAT_6M_VL	3)	churn	1.40	92.24	82%	-1.02	0%	41	50.50	114%	0.00	0	0%	28	
TAKEAWAY_RATIO_CAT_3_6_VL		não churn	1.90	0.46	38%	0.00	0%	0	0.47	72%	0.00	1	0%	0	
TAKEAWAY_CAT_6M_VISITAS		churn	1.31	34.17	74%	0.00	0%	13	21.09	97%	0.00	0	0%	11	
TAKEAWAY_RATIO_CAT_3_6_VIS		não churn	2.20	0.49	30%	0.00	0%	0	0.49	67%	0.00	1	0%	0	
TAKEAWAY_FLG_VISIT_LAST12M			3.91	1.00	5%	0.00	0%	0	0.97	19%	0.00	1	0%	0	
TAKEAWAY_INDEX_3M_VISITAS		churn	1.38	2.84	79%	0.00	0%	1	1.73	109%	0.00	0	0%	1	
TAKEAWAY_MAXPRD_MLOJ_6M_NR			1.28	8.09	52%	0.00	0%	3	6.17	66%	0.00	6	0%	3	
BIRDS_CAT_6M_VL		churn	1.30	57.67	126%	0.00	0%	35	28.87	165%	0.00	0	0%	19	
BIRDS_RATIO_CAT_3_6_VL		não churn	1.58	0.44	66%	0.00	0%	0	0.37	105%	0.00	0	0%	0	
BIRDS_CAT_6M_VISITAS		churn	1.39	8.56	97%	0.00	0%	5	4.67	135%	0.00	0	0%	3	
BIRDS_RATIO_CAT_3_6_VIS		não churn	1.61	0.44	64%	0.00	0%	0	0.37	103%	0.00	0	0%	0	
BIRDS_FLG_VISIT_LAST12M			1.61	0.93	28%	0.00	0%	0	0.83	46%	0.00	1	0%	0	
BIRDS_INDEX_3M_VISITAS		churn	1.49	1.57	102%	0.00	0%	1	0.80	152%	0.00	0	0%	1	
BIRDS_MAXPRD_MLOJ_6M_NR			1.32	2.79	74%	0.00	0%	2	1.84	98%	0.00	1	0%	2	

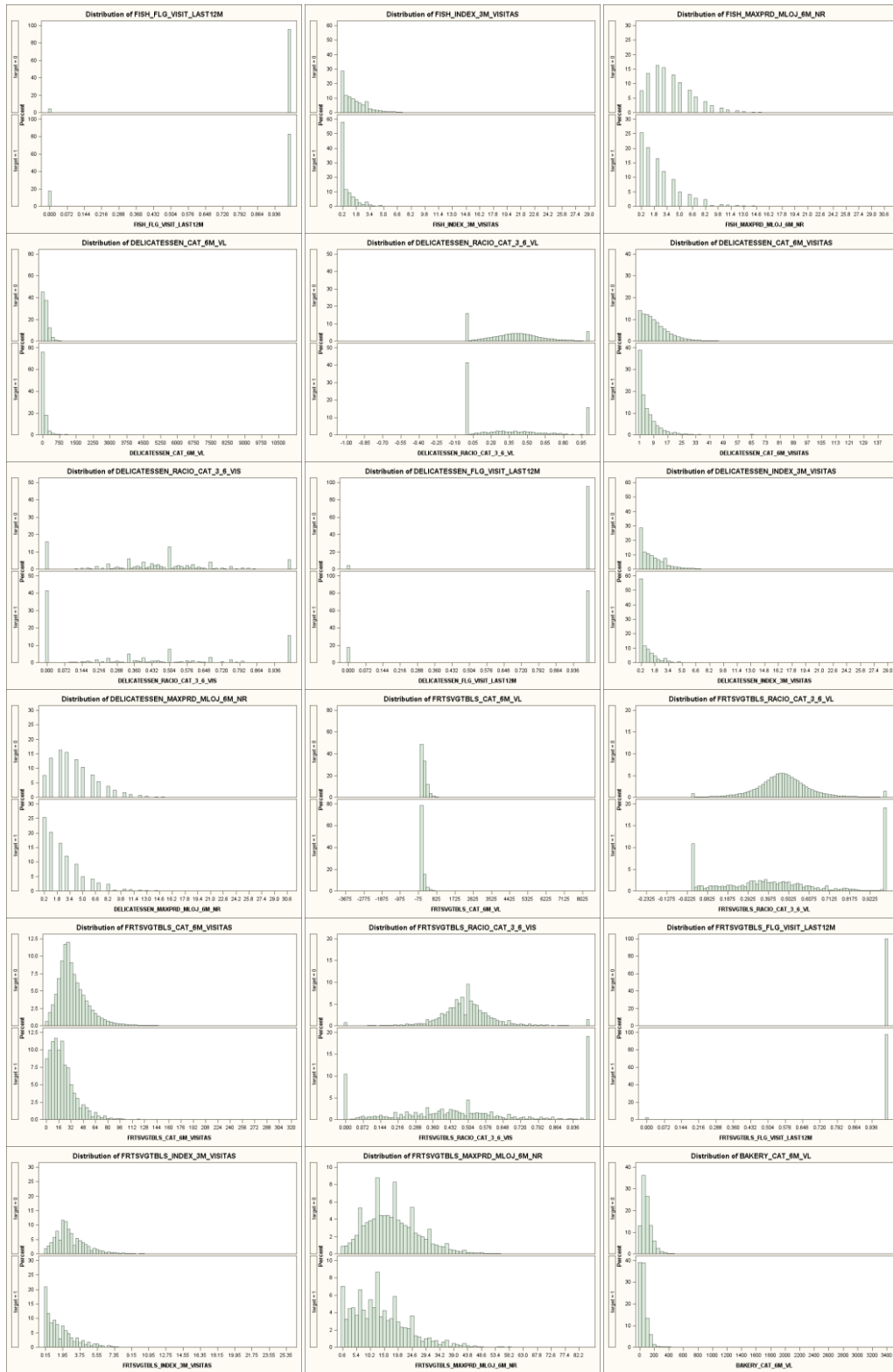
target = 0														target = 1					
Variable	Obs.	Decisão Dq*	CV 1 / CV0	Mean	Coef Var*	Minimum	% Miss	amplitude semi-interquartil (Dq)	Mean	Coef Var*	Minimum	Mode	% Miss	amplitude semi-interquartil (Dq)					
CODFISH_CAT_6M_VL	3)	churn	1.10	30.19	158%	-11.05	0%	17	18.31	175%	0.00	0	0%	11					
CODFISH_RATIO_CAT_3_6_VL		não churn	1.34	0.40	82%	-14.03	0%	0	0.36	110%	0.00	0	0%	0					
CODFISH_CAT_6M_VISITAS		churn	1.35	7.36	124%	0.00	0%	5	4.81	167%	0.00	0	0%	3					
CODFISH_RATIO_CAT_3_6_VIS		não churn	1.38	0.41	77%	0.00	0%	0	0.37	107%	0.00	0	0%	0					
CODFISH_FLG_VISIT_LAST12M			1.50	0.91	31%	0.00	0%	0	0.82	47%	0.00	1	0%	0					
CODFISH_INDEX_3M_VISITAS		churn	1.38	1.31	135%	0.00	0%	1	0.82	186%	0.00	0	0%	1					
CODFISH_MAXPRD_MLOJ_6M_NR			1.20	2.77	99%	0.00	0%	2	2.10	119%	0.00	1	0%	2					
PETCARE_CAT_6M_VL	4)	churn	1.21	44.00	257%	-8.98	0%	22	17.69	312%	0.00	0	0%	4					
PETCARE_RATIO_CAT_3_6_VL	4)	churn	1.50	0.26	123%	-16.74	0%	0	0.17	185%	0.00	0	0%	0					
PETCARE_CAT_6M_VISITAS	4)	churn	1.44	6.03	186%	0.00	0%	4	2.91	269%	0.00	0	0%	1					
PETCARE_RATIO_CAT_3_6_VIS	4)	churn	1.53	0.27	119%	0.00	0%	0	0.17	182%	0.00	0	0%	0					
PETCARE_FLG_VISIT_LAST12M	4)		1.38	0.63	77%	0.00	0%	1	0.47	106%	0.00	0	0%	1					
PETCARE_INDEX_3M_VISITAS	4)	churn	1.49	0.90	193%	0.00	0%	1	0.38	287%	0.00	0	0%	0					
PETCARE_MAXPRD_MLOJ_6M_NR	4)	churn	1.20	2.35	169%	0.00	0%	2	1.32	202%	0.00	0	0%	1					
TOYS_CAT_6M_VL	4)	churn	1.46	46.19	206%	-94.20	0%	27	21.86	301%	0.00	0	0%	6					
TOYS_RATIO_CAT_3_6_VL	4)	churn	0.92	0.12	324%	-59.99	0%	0	0.08	297%	-0.03	0	0%	0					
TOYS_CAT_6M_VISITAS	4)	churn	1.26	2.02	163%	0.00	0%	2	0.93	206%	0.00	0	0%	1					
TOYS_RATIO_CAT_3_6_VIS	4)	churn	1.46	0.16	178%	0.00	0%	0	0.10	259%	0.00	0	0%	0					
TOYS_FLG_VISIT_LAST12M	4)		1.53	0.62	79%	0.00	0%	1	0.41	120%	0.00	0	0%	1					
TOYS_INDEX_3M_VISITAS	4)	churn	1.20	0.41	231%	0.00	0%	0	0.16	276%	0.00	0	0%	0					
TOYS_MAXPRD_MLOJ_6M_NR	4)	churn	1.49	1.76	165%	0.00	0%	1	0.87	245%	0.00	0	0%	1					
TEXTIL_CAT_6M_VL	4)	churn	1.32	52.35	207%	-56.94	0%	29	26.81	273%	0.00	0	0%	8					
TEXTIL_RATIO_CAT_3_6_VL	4)	churn	1.27	0.20	173%	-24.90	0%	0	0.14	220%	0.00	0	0%	0					
TEXTIL_CAT_6M_VISITAS	4)	churn	1.28	2.90	168%	0.00	0%	2	1.35	215%	0.00	0	0%	1					
TEXTIL_RATIO_CAT_3_6_VIS	4)	churn	1.41	0.21	148%	0.00	0%	0	0.15	209%	0.00	0	0%	0					
TEXTIL_FLG_VISIT_LAST12M	4)		1.42	0.66	72%	0.00	0%	1	0.49	102%	0.00	0	0%	1					
TEXTIL_INDEX_3M_VISITAS	4)	churn	1.36	0.59	192%	0.00	0%	0	0.28	261%	0.00	0	0%	0					
TEXTIL_MAXPRD_MLOJ_6M_NR	4)	churn	1.39	2.73	166%	0.00	0%	2	1.67	231%	0.00	0	0%	1					
LIGHTBAZAAR_CAT_6M_VL	3)	churn	1.78	253.52	103%	-86.35	0%	126	133.03	183%	0.00	0	0%	75					
LIGHTBAZAAR_RATIO_CAT_3_6_VL	3)	não churn	1.54	0.41	52%	-21.31	0%	0	0.44	81%	-0.01	1	0%	0					
LIGHTBAZAAR_CAT_6M_VISITAS		churn	1.42	24.53	73%	0.00	0%	9	12.28	104%	0.00	2	0%	7					
LIGHTBAZAAR_RATIO_CAT_3_6_VIS		não churn	2.08	0.46	34%	0.00	0%	0	0.47	71%	0.00	1	0%	0					
LIGHTBAZAAR_FLG_VISIT_LAST12M			3.23	1.00	6%	0.00	0%	0	0.96	20%	0.00	1	0%	0					
LIGHTBAZAAR_INDEX_3M_VISITAS		churn	1.40	2.57	79%	0.00	0%	1	1.20	111%	0.00	0	0%	1					
LIGHTBAZAAR_MAXPRD_MLOJ_6M_NR		churn	1.39	11.56	72%	0.00	0%	5	7.39	100%	0.00	2	0%	4					
ALCHLCDRNKS_CAT_6M_VL	3)	churn	1.20	122.76	173%	-13.33	0%	65	101.50	207%	0.00	0	0%	52					
ALCHLCDRNKS_RATIO_CAT_3_6_VL	3)	não churn	1.36	0.36	75%	-1.06	0%	0	0.37	101%	0.00	0	0%	0					
ALCHLCDRNKS_CAT_6M_VISITAS		churn	1.24	13.54	123%	0.00	0%	7	11.17	152%	0.00	0	0%	6					
ALCHLCDRNKS_RATIO_CAT_3_6_VIS		não churn	1.53	0.40	61%	0.00	0%	0	0.39	94%	0.00	0	0%	0					
ALCHLCDRNKS_FLG_VISIT_LAST12M			2.01	0.98	16%	0.00	0%	0	0.91	32%	0.00	1	0%	0					
ALCHLCDRNKS_INDEX_3M_VISITAS			1.22	#####	138%	0.00	0%	4 328 419	6587117.04	168%	0.00	0	0%	4 328 419					
ALCHLCDRNKS_MAXPRD_MLOJ_6M_NR	1)						100%						100%						
PRSNLHYGIENE_CAT_6M_VL	3)	churn	1.63	171.28	85%	-4.58	0%	84	83.63	139%	0.00	0	0%	49					
PRSNLHYGIENE_RATIO_CAT_3_6_VL	3)	não churn	2.00	0.47	39%	-4.12	0%	0	0.44	78%	0.00	0	0%	0					
PRSNLHYGIENE_CAT_6M_VISITAS		churn	1.56	15.94	54%	0.00	0%	6	7.84	84%	0.00	0	0%	5					
PRSNLHYGIENE_RATIO_CAT_3_6_VIS		não churn	2.31	0.49	32%	0.00	0%	0	0.45	74%	0.00	0	0%	0					
PRSNLHYGIENE_FLG_VISIT_LAST12M			3.85	0.99	7%	0.00	0%	0	0.93	27%	0.00	1	0%	0					
PRSNLHYGIENE_INDEX_3M_VISITAS			1.60	#####	58%	0.00	0%	4 714 480	6701166.95	92%	0.00	0	0%	4 714 480					
PRSNLHYGIENE_MAXPRD_MLOJ_6M_NR	1)						100%						100%						
DAIRY_CAT_6M_VL	3)	churn	1.50	218.32	69%	-1.12	0%	90	95.33	103%	0.00	0	0%	56					
DAIRY_RATIO_CAT_3_6_VL		não churn	2.67	0.49	26%	0.00	0%	0	0.47	69%	0.00	1	0%	0					
DAIRY_CAT_6M_VISITAS		churn	1.52	33.30	52%	0.00	0%	10	20.10	80%	0.00	0	0%	9					
DAIRY_RATIO_CAT_3_6_VIS		não churn	2.83	0.49	23%	0.00	0%	0	0.47	66%	0.00	1	0%	0					
DAIRY_FLG_VISIT_LAST12M			5.23	1.00	4%	0.00	0%	0	0.96	20%	0.00	1	0%	0					
DAIRY_INDEX_3M_VISITAS		não churn	1.65	#####	56%	0.00	0%	9 478 611	20020940.90	92%	0.00	0	0%	11 584 969					
DAIRY_MAXPRD_MLOJ_6M_NR	1)						100%						100%						
FROZEN_CAT_6M_VL	3)	churn	1.46	91.42	106%	-3.56	0%	48	42.58	155%	0.00	0	0%	25					
FROZEN_RATIO_CAT_3_6_VL	3)	não churn	1.88	0.47	50%	-0.14	0%	0	0.39	94%	0.00	0	0%	0					
FROZEN_CAT_6M_VISITAS		churn	1.50	13.31	75%	0.00	0%	6	7.35	112%	0.00	0	0%	5					
FROZEN_RATIO_CAT_3_6_VIS		não churn	2.08	0.48	43%	0.00	0%	0	0.40	90%	0.00	0	0%	0					
FROZEN_FLG_VISIT_LAST12M			3.18	0.99	11%	0.00	0%	0	0.89	36%	0.00	1	0%	0					
FROZEN_INDEX_3M_VISITAS		churn	1.64	#####	80%	0.00	0%	4 369 368	4818974.51	131%	0.00	0	0%	3 641 140					
FROZEN_MAXPRD_MLOJ_6M_NR	1)						100%						100%						

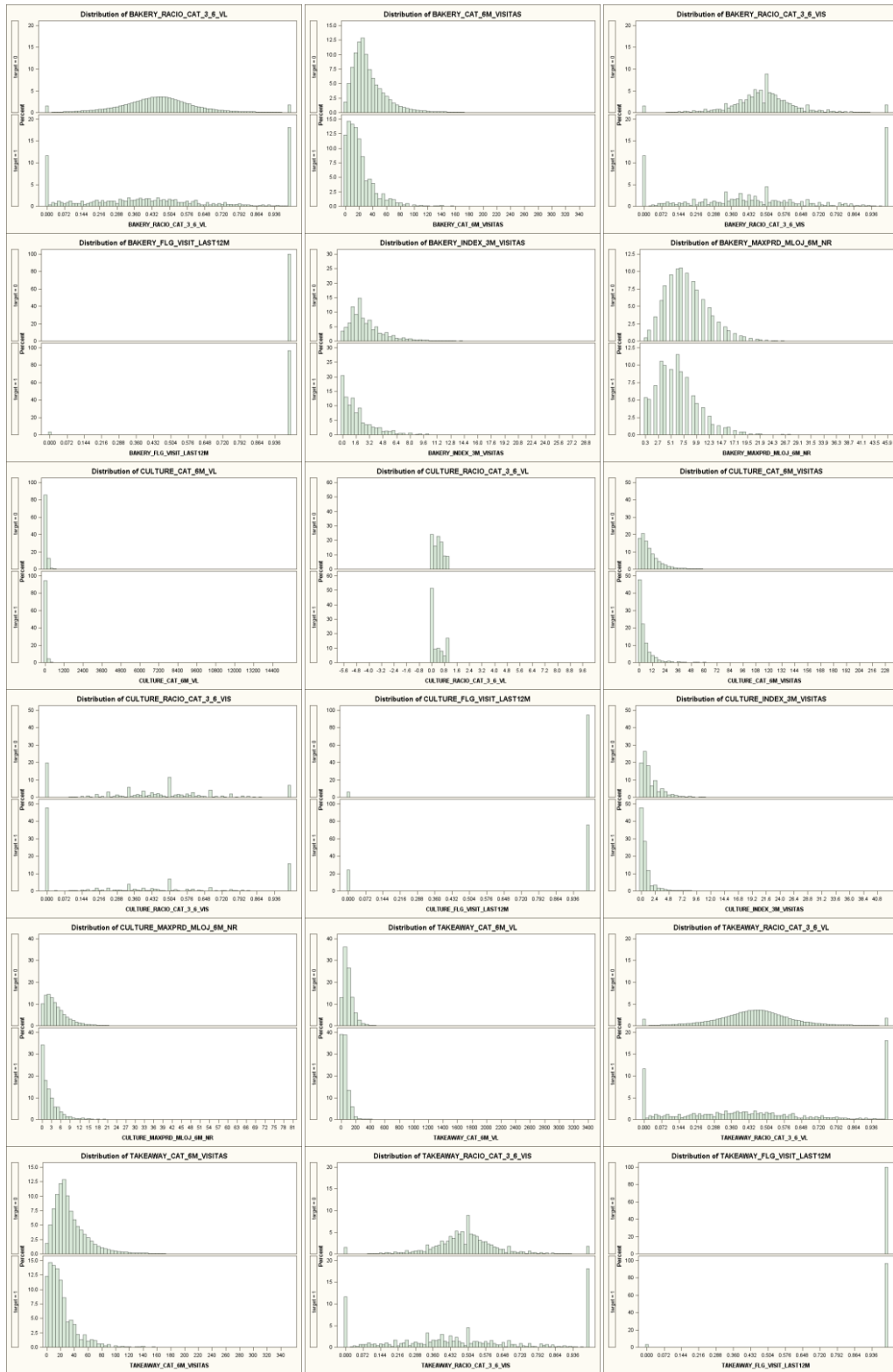
Tabela 19 - Estatísticas Variáveis Comportamentais originais



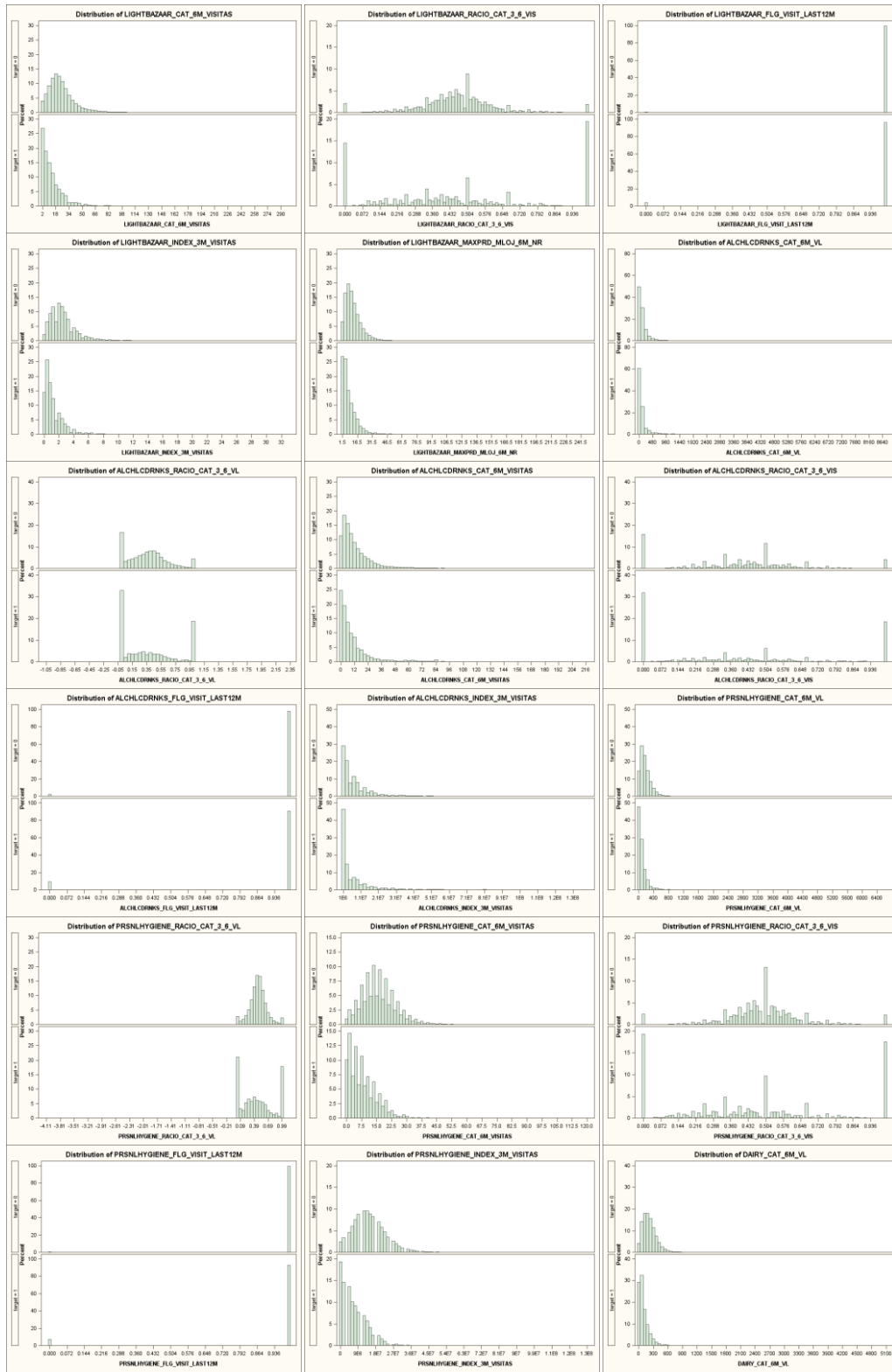


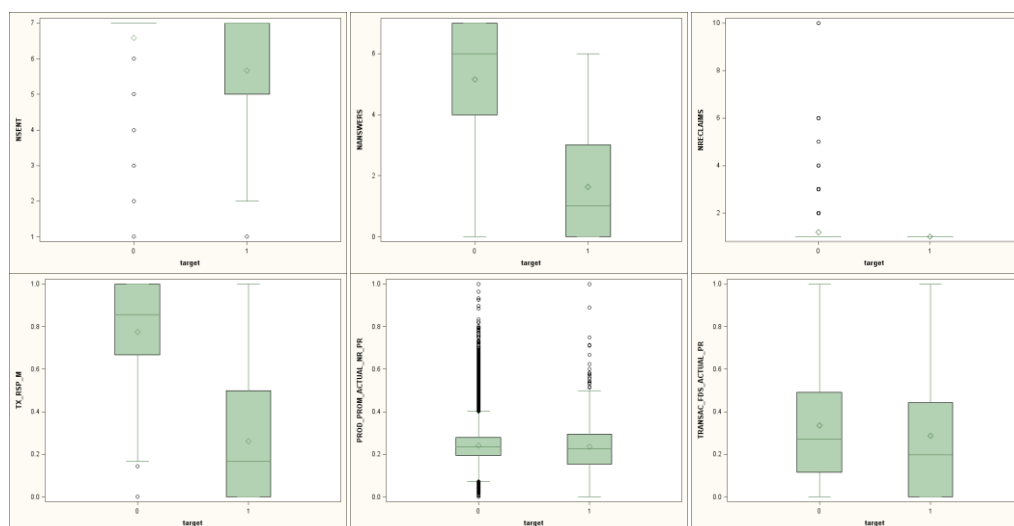
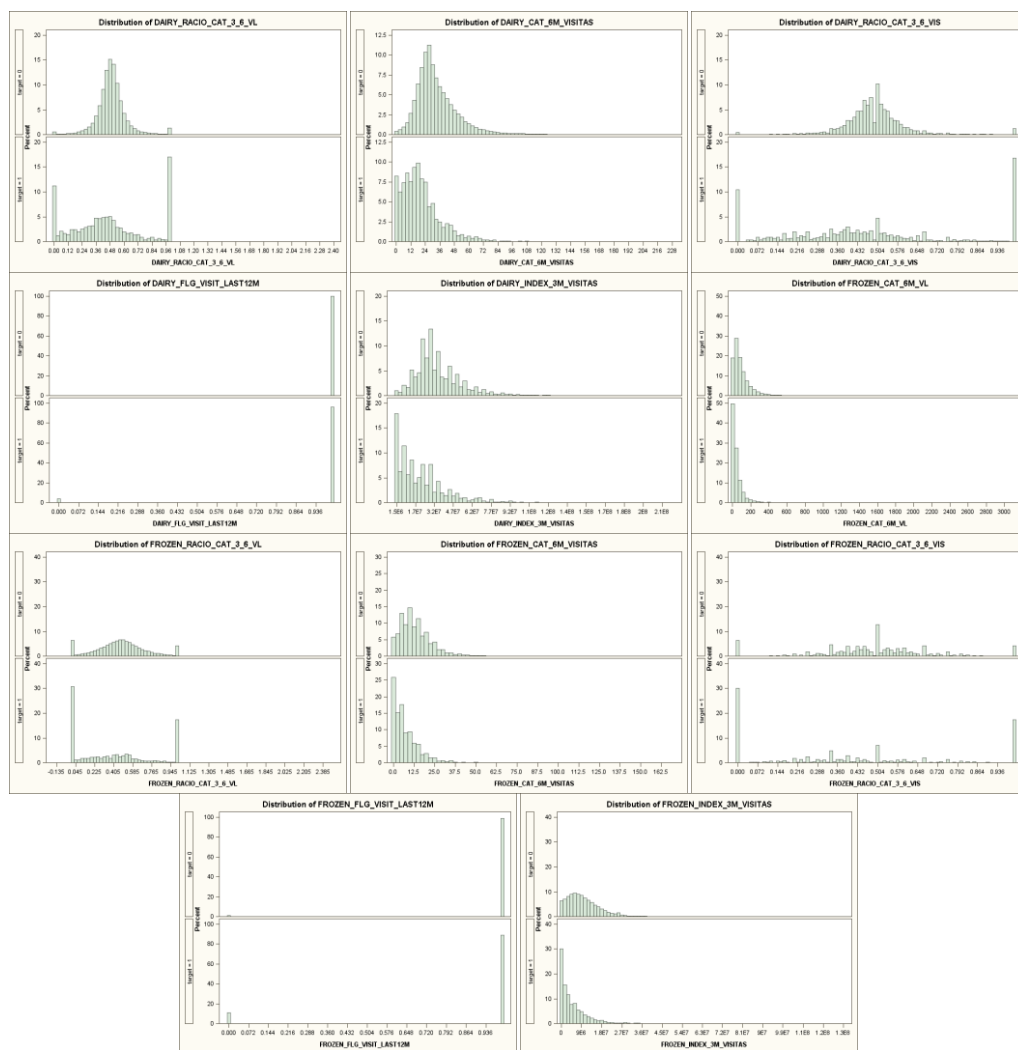


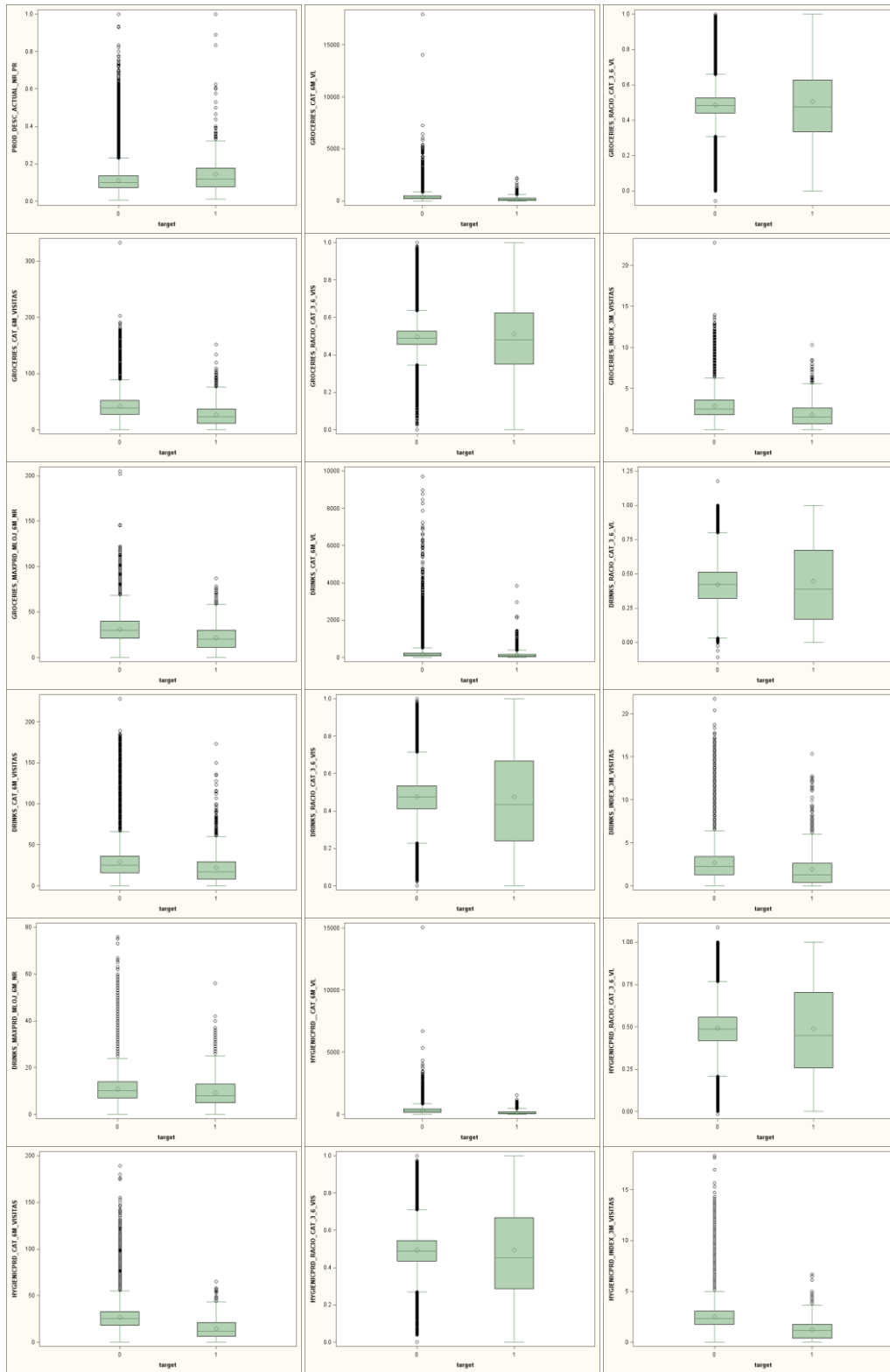


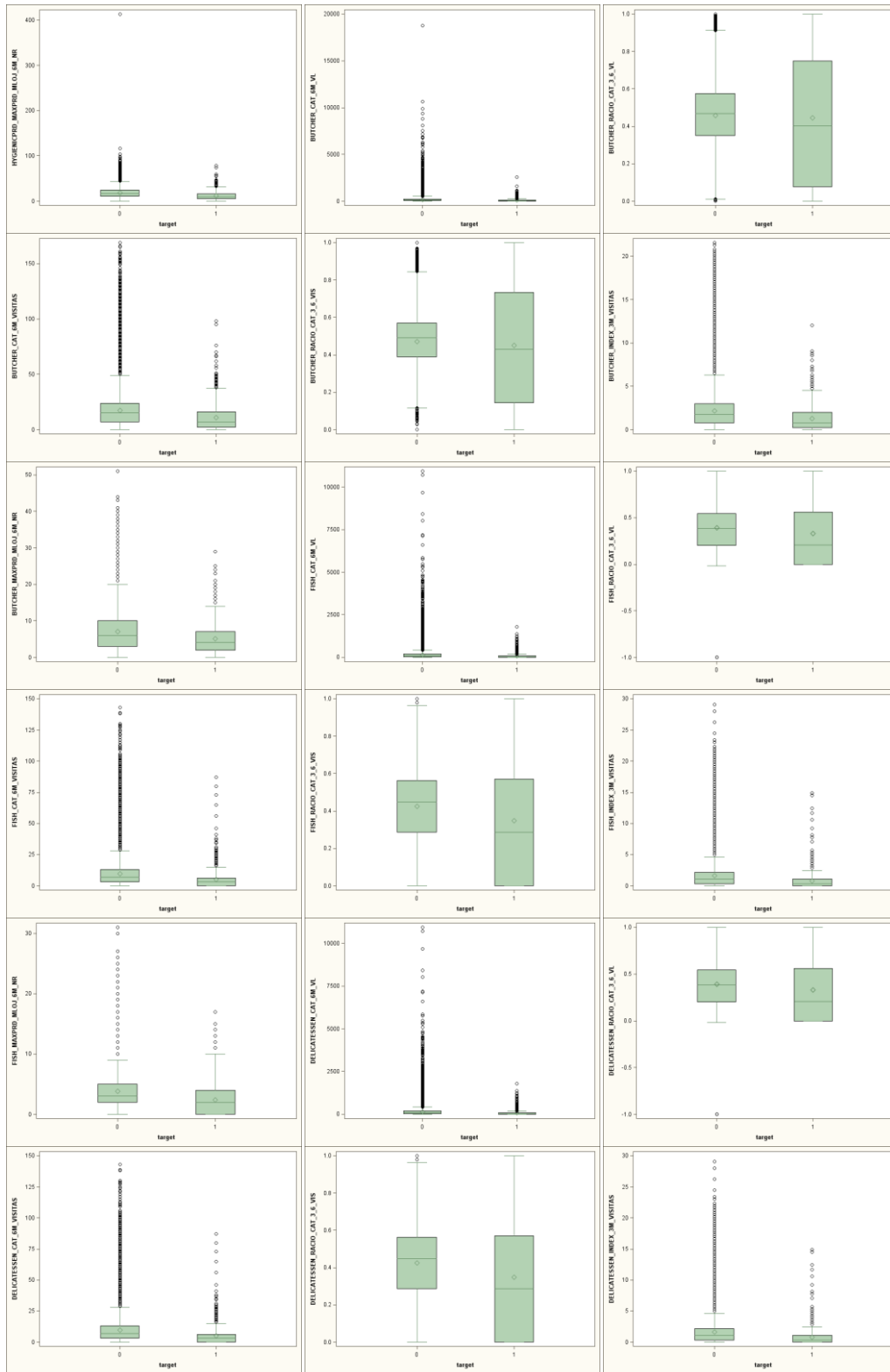


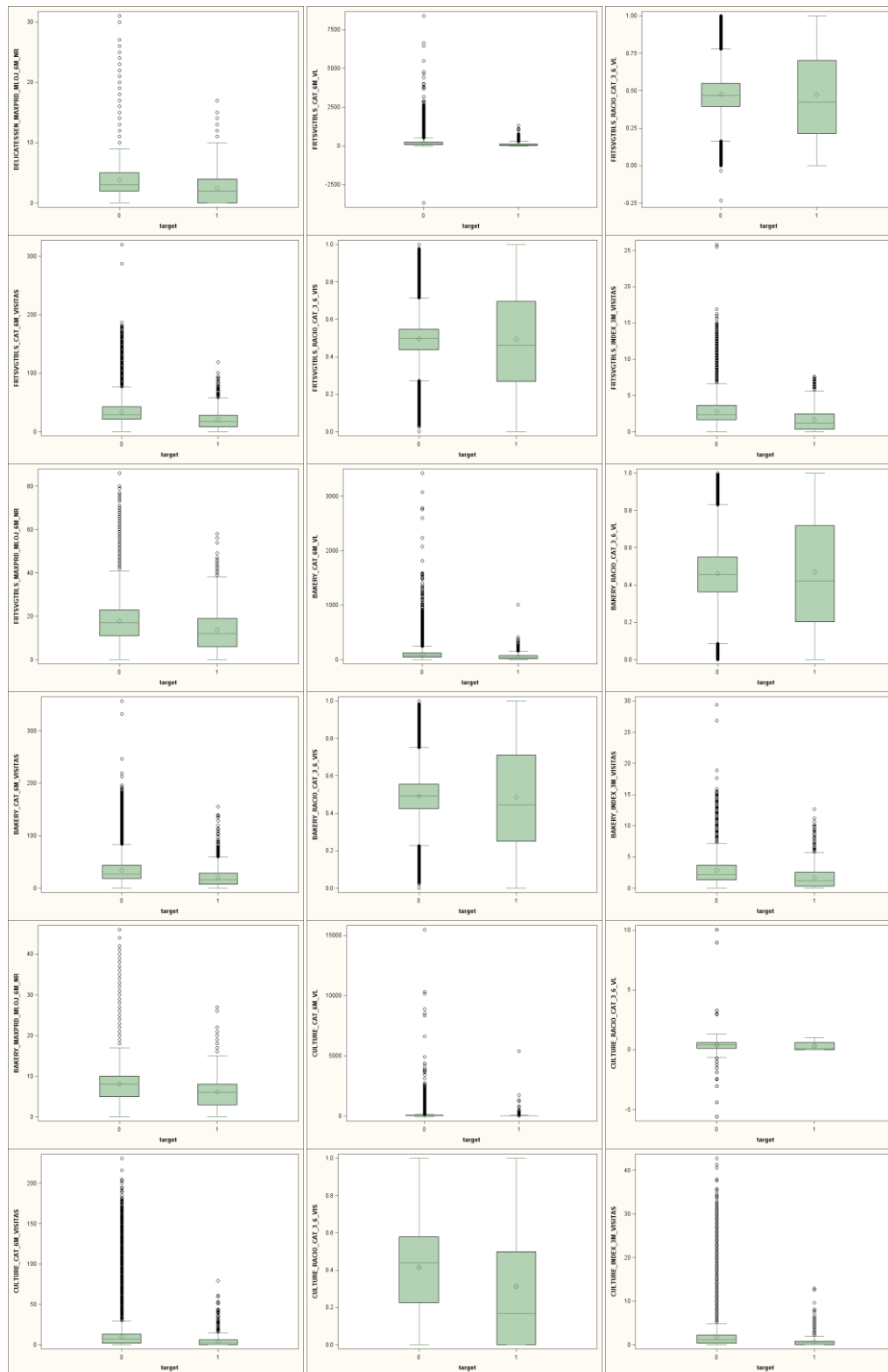


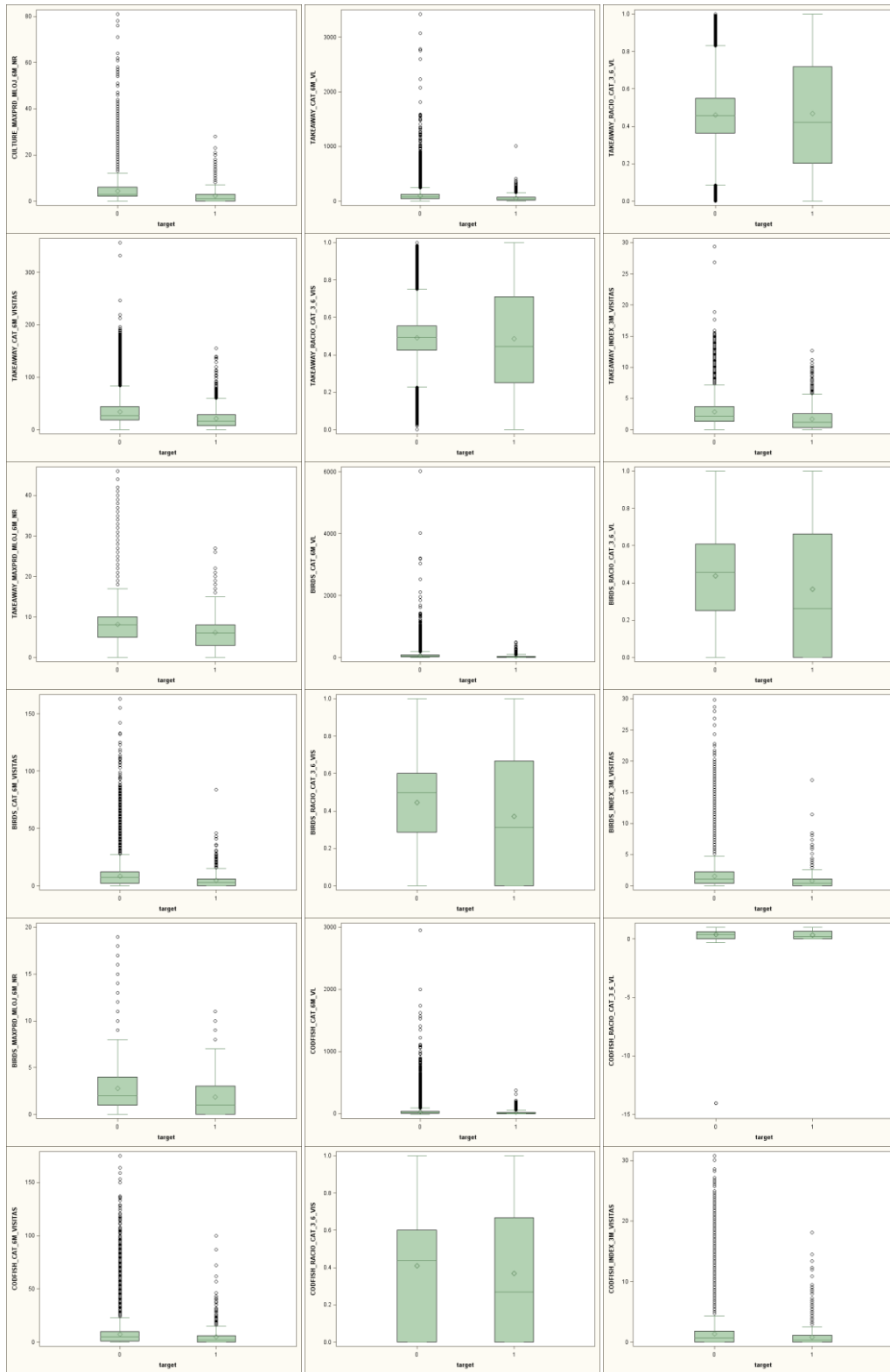


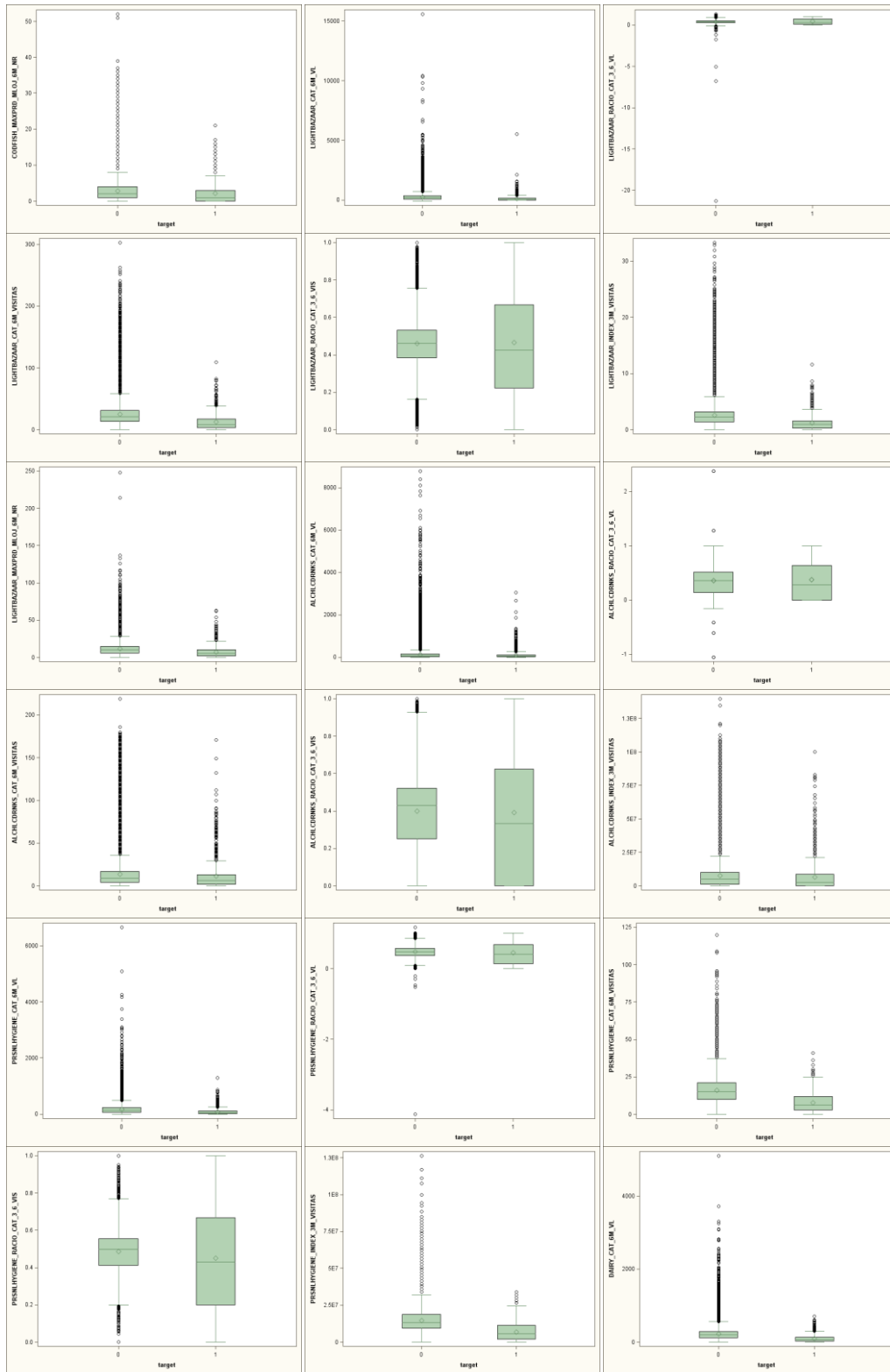












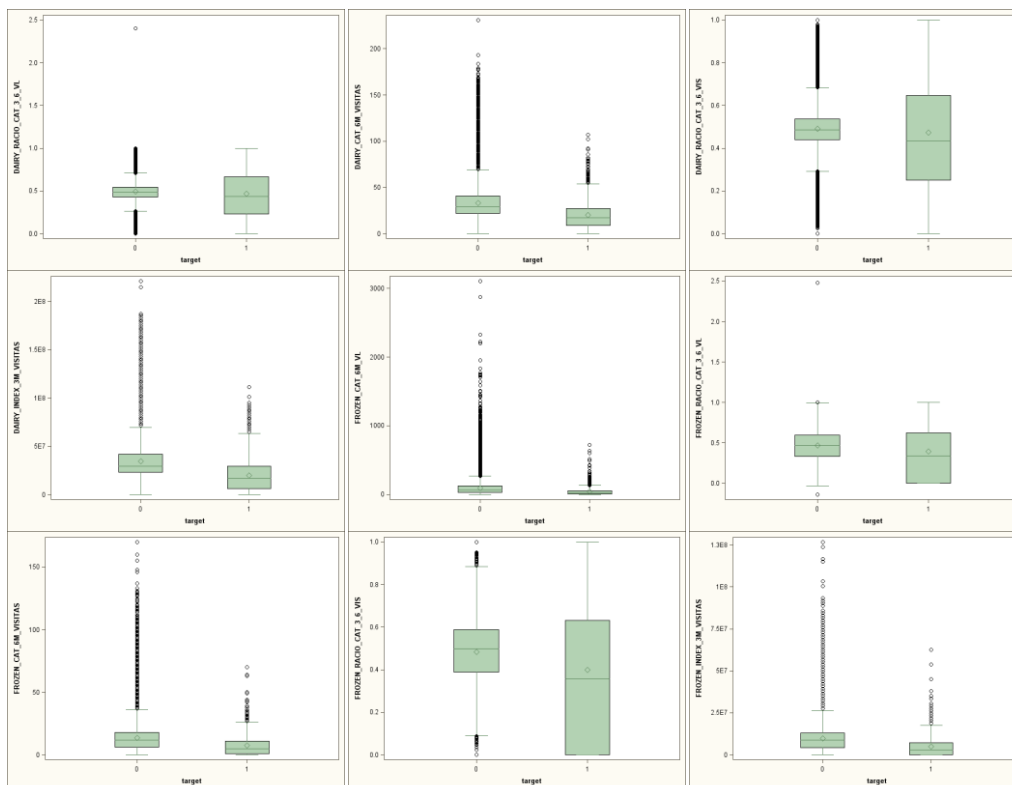
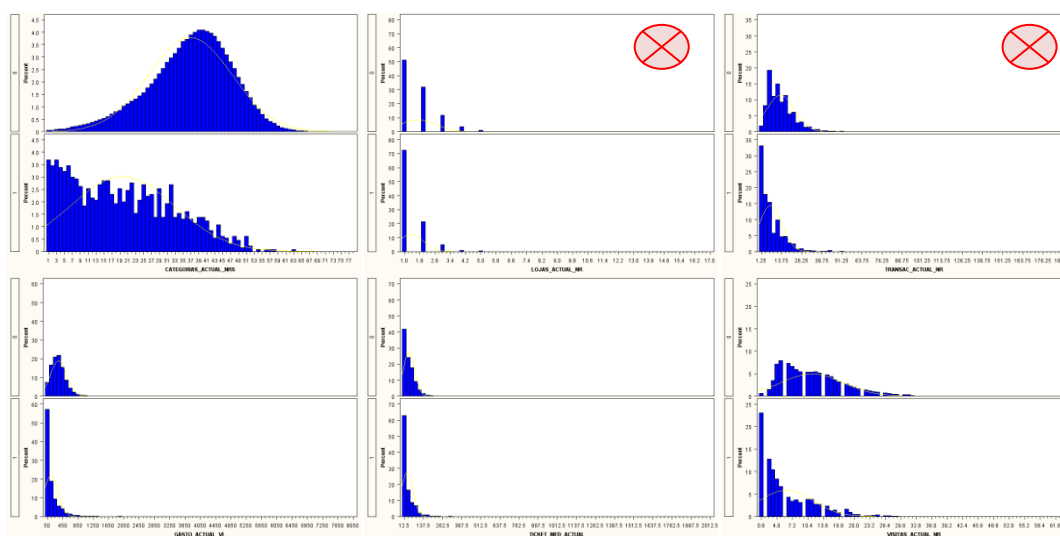


Figura 63- Box-Plot das variáveis comportamentais originais

III.1.3. 3. Variáveis de Envolvimento

Uma vez que estas variáveis de multiplicam pelos períodos analisados (actual, últimos 3 meses, últimos 6 meses e últimos 12 meses). E dado que por observação dos resultados das estatísticas de todas as variáveis não se encontraram diferenças significativas nas distribuições pelos períodos, optou-se por fazer referência, e a título de exemplo de um dos conjuntos de variáveis.



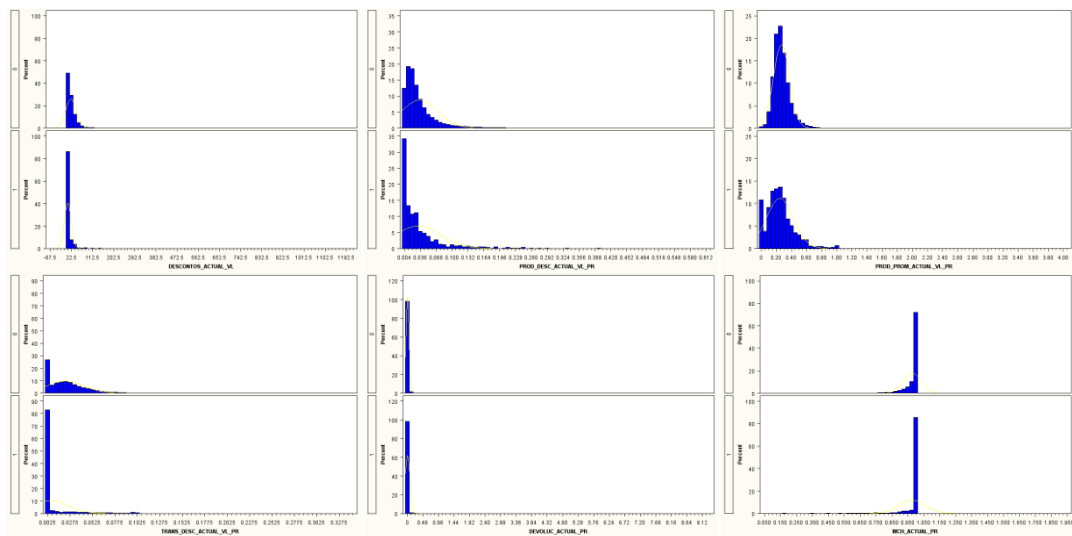


Figura 64 - Histogramas das variáveis de envolvimento

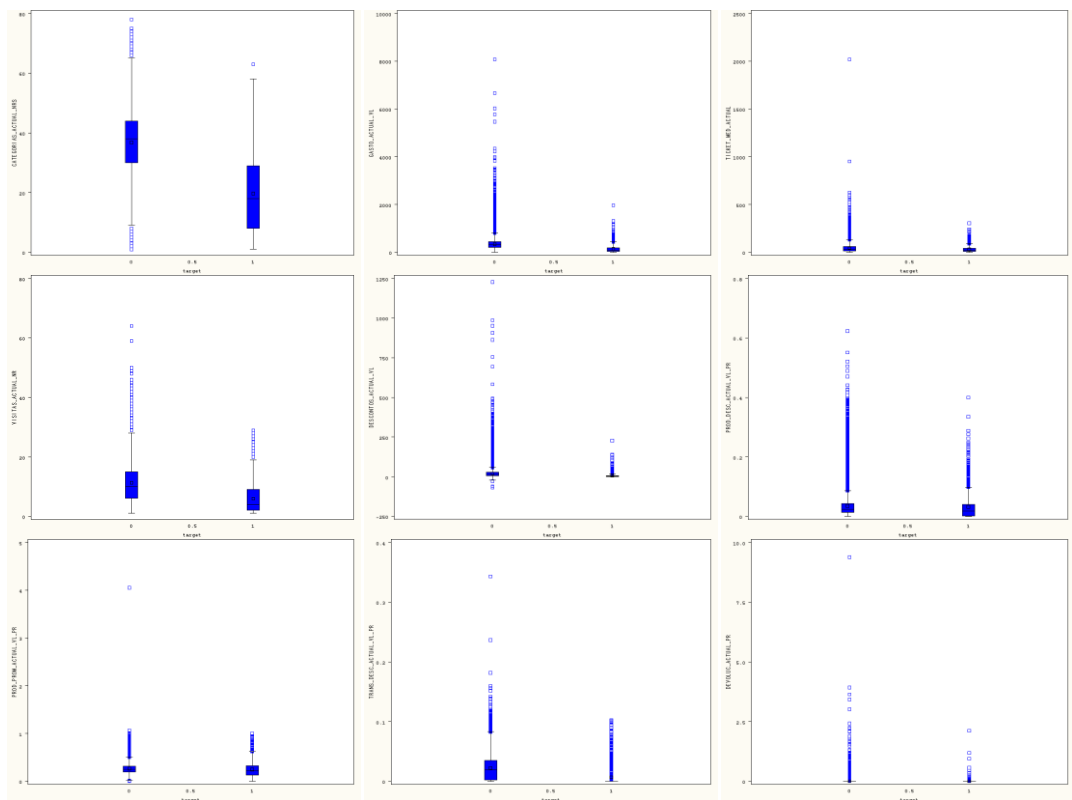


Figura 65 – Box-Plot das variáveis de envolvimento

III.1.3.3.1. Análise de correlações

Coeficiente de correlação	Correlação
$r = 1$	Perfeita positiva
$0,8 \leq r < 1$	Forte positiva
$0,5 \leq r < 0,8$	Moderada positiva
$0,1 \leq r < 0,5$	Fraca positiva
$0 < r < 0,1$	Ínfima positiva
0	Nula
$-0,1 < r < 0$	Ínfima negativa
$-0,5 < r \leq -0,1$	Fraca negativa
$-0,8 < r \leq -0,5$	Moderada negativa
$-1 < r \leq -0,8$	Forte negativa
$r = -1$	Perfeita negativa

Tabela 20 - Descrição das correlações de Pearson

De forma a sustentar a hipótese levantada acerca de correlações entre as variáveis de envolvimento, foi calculado o Coeficiente de Correlação de Pearson (r). Os resultados obtidos podem ser consultados na Tabela 21.

r = Coeficiente de correlação de Pearson																																				
	GASTO_ACTUAL_VL	TICKET_MED_ACTUAL	VISITAS_ACTUAL_NR	DESCONTOS_ACTUAL_VL	PROD_DESC_ACTUAL_VL_PR	PROD_PROM_ACTUAL_VL_PR	TRANS_DESC_ACTUAL_VL_PR	GASTO_ULT_3M_VL	TICKET_MED_ULT_3M	VISITAS_ULT_3M_NR	DESCONTOS_ULT_3M_VL	PROD_DESCONTOS_3M_VL_PR	PROD_PROMOCAO_3M_VL_PR	TRANS_DESCONTOS_3M_VL_PR	GASTO_ULT_6M_VL	TICKET_MED_ULT_6M	VISITAS_ULT_6M_NR	DESCONTOS_ULT_6M_VL	PROD_DESCONTOS_6M_VL_PR	PROD_PROMOCAO_6M_VL_PR	TRANS_DESCONTOS_6M_VL_PR	GASTO_ULT_12M_VL	TICKET_MED_ULT_12M	VISITAS_ULT_12M_NR	DESCONTOS_ULT_12M_VL	PROD_DESCONTOS_12M_VL_PR	PROD_PROMOCAO_12M_VL_PR	TRANS_DESC_12M_VL_PR	CATEGORIAS_ACTUAL_NRS	CATEGORIAS_LAST12M_NRS	CATEGORIAS_LAST6M_NRS	PROD_PROM_ACTUAL_NR_PR				
GASTO_ACTUAL_VL		0.57	0.05	0.58	0.08	0.03	0.12	0.91	0.55	-0.01	0.48	-0.01	-0.05	0.13	0.84	0.53	-0.04	0.39	-0.03	0.00	0.15	0.78	0.53	-0.07	0.33	-0.02	0.01	0.11	0.64	0.51	0.43	0.37	0.02			
TICKET_MED_ACTUAL	0.57		-0.58	0.43	0.05	0.02	0.34	0.58	0.93	-0.59	0.42	0.01	-0.03	0.37	0.57	0.89	-0.59	0.35	-0.01	0.04	0.36	0.55	0.87	-0.58	0.30	0.00	0.06	0.30	0.31	0.30	0.29	0.27	0.03			
VISITAS_ACTUAL_NR	0.05	-0.58		-0.08	0.01	0.00	-0.31	-0.04	-0.61	0.94	-0.15	-0.02	-0.03	-0.36	-0.11	-0.61	0.88	-0.15	-0.03	-0.08	-0.33	-0.14	-0.61	0.81	-0.14	-0.03	-0.10	-0.29	0.15	0.02	-0.07	-0.11	-0.03			
DESCONTOS_ACTUAL_VL	0.58	0.43	-0.08		0.65	0.47	0.38	0.51	0.40	-0.12	0.78	0.47	0.35	0.38	0.49	0.39	-0.12	0.68	0.40	0.38	0.37	0.46	0.38	-0.13	0.63	0.40	0.39	0.34	0.34	0.31	0.30	0.28	0.33			
PROD_DESC_ACTUAL_VL_PR	0.08	0.05	0.01	0.65		0.75	0.05	0.05	0.04	-0.01	0.52	0.72	0.59	0.13	0.06	0.03	0.00	0.50	0.63	0.55	0.14	0.05	0.03	0.00	0.50	0.61	0.56	0.15	0.00	0.05	0.08	0.10	0.54			
PROD_PROM_ACTUAL_VL_PR	0.03	0.02	0.00	0.47	0.75		0.08	0.02	0.01	-0.01	0.41	0.58	0.69	0.14	0.03	0.01	0.00	0.41	0.54	0.60	0.14	0.02	0.01	0.00	0.41	0.53	0.58	0.16	0.00	0.05	0.10	0.10	0.79			
TRANS_DESC_ACTUAL_VL_PR	0.12	0.34	-0.31	0.38	0.05	0.08		0.16	0.34	-0.30	0.39	0.14	0.10	0.81	0.18	0.33	-0.28	0.35	-0.16	0.19	0.74	0.18	0.32	-0.26	0.30	0.16	0.19	0.69	0.11	0.14	0.16	0.17	0.08			
GASTO_ULT_3M_VL	0.91	0.58	-0.04	0.51	0.05	0.02	0.16		0.62	-0.04	0.57	0.03	-0.01	0.19	0.94	0.61	-0.07	0.47	0.01	0.03	0.20	0.87	0.61	-0.10	0.40	0.00	0.05	0.16	0.57	0.59	0.51	0.44	0.01			
TICKET_MED_ULT_3M	0.55	0.93	-0.61	0.40	0.04	0.01	0.34	0.62		-0.65	0.47	0.02	-0.01	0.41	0.62	0.97	-0.65	0.40	0.00	0.06	0.41	0.60	0.95	-0.63	0.34	0.01	0.08	0.34	0.29	0.33	0.32	0.30	0.03			
VISITAS_ULT_3M_NR	-0.01	-0.59	0.94	-0.12	-0.01	-0.01	-0.30	-0.04	-0.65		-0.16	-0.02	0.00	-0.38	-0.10	-0.65	0.96	-0.16	-0.02	-0.07	-0.35	-0.14	-0.65	0.88	-0.14	-0.02	-0.09	-0.31	0.08	0.02	-0.05	-0.10	-0.04			
DESCONTOS_ULT_3M_VL	0.48	0.42	-0.15	0.78	0.52	0.41	0.39	0.57	0.47	-0.16		0.67	0.51	0.48	0.56	0.46	-0.16	0.88	0.58	0.53	0.48	0.52	0.45	-0.16	0.83	0.57	0.54	0.45	0.28	0.36	0.36	0.35	0.32			
PROD_DESCONTOS_3M_VL_PR	-0.01	0.01	-0.02	0.47	0.72	0.58	0.14	0.03	0.02	-0.02	0.67		0.80	0.18	0.04	0.02	0.00	0.64	0.86	0.74	0.19	0.03	0.01	0.01	0.64	0.84	0.75	0.21	-0.06	0.03	0.08	0.11	0.46			
PROD_PROMOCAO_3M_VL_PR	-0.05	-0.03	-0.03	0.35	0.59	0.69	0.10	-0.01	-0.01	0.00	0.51	0.80		0.11	0.01	-0.01	0.02	0.50	0.71	0.80	0.12	0.01	-0.02	0.03	0.51	0.69	0.76	0.15	-0.05	0.06	0.11	0.15	0.57			
TRANS_DESCONTOS_3M_VL_PR	0.13	0.37	-0.36	0.38	0.13	0.14	0.81	0.19	0.41	-0.38	0.48	0.18	0.11	0.23	0.41	-0.35	0.44	0.22	0.24	0.93	0.23	0.39	-0.32	0.39	0.21	0.24	0.86	0.12	0.17	0.20	0.22	0.14				
GASTO_ULT_6M_VL	0.84	0.57	-0.11	0.49	0.06	0.03	0.18	0.94	0.62	-0.10	0.56	0.04	0.01	0.23	0.65	-0.08	0.54	0.04	0.09	0.22	0.95	0.65	-0.10	0.47	0.04	0.09	0.18	0.54	0.57	0.57	0.51	0.03				
TICKET_MED_ULT_6M	0.53	0.89	-0.61	0.39	0.03	0.01	0.33	0.61	0.97	-0.65	0.46	0.02	-0.01	0.41	0.65	-0.66	0.42	0.02	0.08	0.40	0.63	0.98	-0.65	0.36	0.02	0.10	0.34	0.28	0.33	0.34	0.32	0.03				
VISITAS_ULT_6M_NR	-0.04	-0.59	0.88	-0.12	0.00	0.00	-0.28	-0.07	-0.65	0.96	-0.16	0.00	0.02	-0.35	-0.08	-0.66	-0.13	0.01	-0.05	-0.35	-0.10	-0.66	0.95	-0.11	0.00	-0.08	-0.30	0.06	0.01	-0.01	-0.04	-0.03				
DESCONTOS_ULT_6M_VL	0.39	0.35	-0.15	0.68	0.50	0.41	0.35	0.47	0.40	-0.16	0.88	0.64	0.50	0.44	0.54	0.42	-0.13	0.74	0.66	0.44	0.50	0.41	-0.13	0.93	0.70	0.66	0.42	0.23	0.33	0.38	0.39	0.34				
PROD_DESCONTOS_6M_VL_PR	-0.03	-0.01	-0.03	0.40	0.63	0.54	0.16	-0.01	0.00	-0.02	0.58	0.86	0.71	0.22	0.04	0.02	0.01	0.74	0.86	0.20	0.03	0.01	0.01	0.71	0.94	0.84	0.23	-0.07	0.04	0.11	0.16	0.45				
PROD_PROMOCAO_6M_VL_PR	0.00	0.04	-0.08	0.38	0.55	0.60	0.19	0.03	0.06	-0.07	0.53	0.74	0.80	0.24	0.09	0.08	-0.05	0.66	0.86	0.23	0.07	0.07	-0.05	0.64	0.80	0.93	0.26	-0.02	0.11	0.18	0.22	0.52				
TRANS_DESCONTOS_6M_VL_PR	0.15	0.36	-0.33	0.37	0.14	0.14	0.74	0.20	0.41	-0.35	0.48	0.19	0.12	0.93	0.22	0.40	-0.35	0.44	0.20	0.23	0.22	0.39	-0.32	0.39	0.20	0.23	0.94	0.13	0.18	0.18	0.20	0.15				
GASTO_ULT_12M_VL	0.78	0.55	-0.14	0.46	0.05	0.02	0.18	0.87	0.60	-0.14	0.52	0.03	0.01	0.23	0.95	0.63	-0.10	0.50	0.03	0.07	0.22	0.65	-0.06	0.48	0.04	0.06	0.17	0.49	0.52	0.54	0.53	0.02				
TICKET_MED_ULT_12M	0.53	0.87	-0.61	0.38	0.03	0.01	0.32	0.61	0.95	-0.65	0.45	0.01	-0.02	0.39	0.65	0.98	-0.66	0.41	0.01	0.07	0.39	0.65	-0.66	0.36	0.02	0.09	0.32	0.28	0.32	0.34	0.33	0.03				
VISITAS_ULT_12M_NR	-0.07	-0.58	0.81	-0.13	0.00	0.00	-0.26	-0.10	-0.63	0.88	-0.16	0.01	0.03	-0.32	-0.10	-0.65	0.95	-0.13	0.01	-0.05	-0.32	-0.06	-0.66	-0.09	0.00	-0.10	-0.28	0.02	-0.03	0.00	0.00	-0.03				
DESCONTOS_ULT_12M_VL	0.33	0.30	-0.14	0.63	0.50	0.41	0.30	0.40	0.34	-0.14	0.83	0.64	0.51	0.39	0.47	0.36	-0.11	0.93	0.71	0.64	0.39	0.48	0.36	-0.09	0.78	0.70	0.37	0.18	0.27	0.33	0.36	0.35				
PROD_DESCONTOS_12M_VL_PR	-0.02	0.00	-0.03	0.40	0.61	0.53	0.16	0.00	0.01	-0.02	0.57	0.84	0.69	0.21	0.04	0.02	0.00	0.70	0.94	0.80	0.20	0.04	0.02	0.00	0.78	0.88	0.22	-0.07	0.03	0.10	0.14	0.45				
PROD_PROMOCAO_12M_VL_PR	0.01	0.06	-0.10	0.39	0.56	0.58	0.19	0.05	0.08	-0.09	0.54	0.75	0.76	0.24	0.09	0.10	-0.08	0.66	0.84	0.93	0.23	0.06	0.09	-0.10	0.70	0.88	0.25	-0.01	0.12	0.19	0.22	0.51				
TRANS_DESC_12M_VL_PR	0.11	0.30	-0.29	0.34	0.15	0.16	0.69	0.16	0.34	-0.31	0.45	0.21	0.15	0.86	0.18	0.34	-0.30	0.42	0.23	0.26	0.94	0.17	0.32	-0.28	0.37	0.22	0.25		0.12	0.16	0.17	0.18				
CATEGORIAS_ACTUAL_NRS	0.64	0.31	0.15	0.34	0.00	0.00	0.11	0.57	0.29	0.08	0.28	-0.06	-0.05	0.12	0.54	0.28	0.06	0.23	-0.07	-0.02	0.13	0.49	0.28	0.02	0.18	-0.07	-0.01	0.12		0.83	0.71	0.60	0.00			
CATEGORIAS_LAST12M_NRS	0.51	0.30	0.02	0.31	0.05	0.05	0.14	0.59	0.33	0.02	0.36	0.03	0.06	0.17	0.57	0.33	0.01	0.33	0.04	0.11	0.18	0.52	0.32	-0.03	0.27	0.03	0.12	0.16	0.83	0.90	0.80	0.06				
CATEGORIAS_LAST6M_NRS	0.43	0.29	-0.07	0.30	0.08	0.07	0.16	0.51	0.32	-0.05	0.36	0.08	0.11	0.20	0.57	0.34	-0.01	0.38	0.11	0.18	0.18	0.54	0.34	-0.03	0.33	0.10	0.19	0.17	0.71	0.90	0.93	0.09				
CATEGORIAS_LAST12M_NRS	0.37	0.27	-0.11	0.28	0.10	0.10	0.17	0.44	0.30	-0.10	0.35	0.11	0.15	0.22	0.51	0.32	-0.04	0.39	0.16	0.22	0.20	0.53	0.33	0.00	0.36	0.14	0.22	0.18	0.60	0.80	0.93					
PROD_PROM_ACTUAL_NR_PR	0.02	0.03	-0.03	0.33	0.54	0.79	0.08	0.01	0.03	-0.04	0.32	0.46	0.57	0.14	0.03	0.03	-0.03	0.34	0.45	0.52	0.15	0.02	0.03	-0.03	0.35	0.45	0.51	0.17	0.00	0.06	0.09	0.11				
TRANSAC_FDS_ACTUAL_PR	0.02	0.17	-0.19	0.05	0.09	0.07	-0.01	0.06	0.19	-0.19	0.07	0.04	0.05	0.06	0.08	0.19	-0.18	0.07	0.03	0.05	0.06	0.08	0.18	-0.16	0.06	0.02	0.05	0.06	0.07	0.09	0.10	0.09	0.09			
PROD_DESC_ACTUAL_NR_PR	-0.05	-0.06	0.03	0.36	0.69	0.63	0.02	-0.06	-0.05	0.03	0.33	0.56	0.51	0.09	-0.04	-0.05	0.04	0.35	0.51	0.48	0.10	-0.04	-0.06	0.05	0.37	0.51	0.47	0.12	-0.16	-0.11	-0.07	-0.04	0.73			

Tabela 21 - Matriz de correlação das variáveis de Envolvimento

Para uma melhor interpretabilidade foram salientados os valores de *r*, acima de 0.80, que são já indicativos de uma correlação positiva forte (Tabela 20).

Sobre estas variáveis foi analisada a correlação com a *target*, de modo a poderem ser rejeitadas à partida as variáveis com menos relação com a *target*, e assim reduzir o conjunto de variáveis de input ao modelo. No entanto, e uma vez que nenhuma das variáveis analisadas possui *per-si* uma correlação forte com a *target*, optou-se por não rejeitar nenhuma nesta fase.

Após a análise descritiva dos dados, foi possível rejeitar algumas das variáveis consideradas inicialmente. Após esta rejeição o conjunto de variáveis ficou reduzido ao conteúdo da Tabela 22.

Tipo	Variável	Missing values	Outliers
ID	CUSTOMER_ACCOUNT_CD		
t	target		
Socio-Demográfica	COMPETITIONINDEX	✓	
Socio-Demográfica	DISTRICT_CD		
Socio-Demográfica	GENDER	✓	
Socio-Demográfica	HOUSEHOLD		
Socio-Demográfica	IPC	✓	✓
Socio-Demográfica	MUNICIPALITY_CD		
Socio-Demográfica	PREF_STORE		
Socio-Demográfica	SEGM_AGE		
Comportamental	ALCHLCDRNKS_CAT_6M_VISITAS		✓
Comportamental	ALCHLCDRNKS_CAT_6M_VL		✓
Comportamental	ALCHLCDRNKS_FLG_VISIT_LAST12M		✓
Comportamental	ALCHLCDRNKS_INDEX_3M_VISITAS		✓
Comportamental	ALCHLCDRNKS_RATIO_CAT_3_6_VIS		✓
Comportamental	ALCHLCDRNKS_RATIO_CAT_3_6_VL		✓
Comportamental	BAKERY_CAT_6M_VISITAS		✓
Comportamental	BAKERY_CAT_6M_VL		✓
Comportamental	BAKERY_FLG_VISIT_LAST12M		✓
Comportamental	BAKERY_INDEX_3M_VISITAS		✓
Comportamental	BAKERY_MAXPRD_MLOI_6M_NR		✓
Comportamental	BAKERY_RATIO_CAT_3_6_VIS		✓
Comportamental	BAKERY_RATIO_CAT_3_6_VL		✓
Comportamental	BIRDS_CAT_6M_VISITAS		✓
Comportamental	BIRDS_CAT_6M_VL		✓
Comportamental	BIRDS_FLG_VISIT_LAST12M		✓
Comportamental	BIRDS_INDEX_3M_VISITAS		✓
Comportamental	BIRDS_MAXPRD_MLOI_6M_NR		✓
Comportamental	BIRDS_RATIO_CAT_3_6_VIS		✓
Comportamental	BIRDS_RATIO_CAT_3_6_VL		✓
Comportamental	BUTCHER_CAT_6M_VISITAS		✓
Comportamental	BUTCHER_CAT_6M_VL		✓
Comportamental	BUTCHER_FLG_VISIT_LAST12M		✓
Comportamental	BUTCHER_INDEX_3M_VISITAS		✓
Comportamental	BUTCHER_MAXPRD_MLOI_6M_NR		✓
Comportamental	BUTCHER_RATIO_CAT_3_6_VIS		✓
Comportamental	BUTCHER_RATIO_CAT_3_6_VL		✓
Comportamental	CODFISH_CAT_6M_VISITAS		✓
Comportamental	CODFISH_CAT_6M_VL		✓
Comportamental	CODFISH_FLG_VISIT_LAST12M		✓
Comportamental	CODFISH_INDEX_3M_VISITAS		✓
Comportamental	CODFISH_MAXPRD_MLOI_6M_NR		✓
Comportamental	CODFISH_RATIO_CAT_3_6_VIS		✓
Comportamental	CODFISH_RATIO_CAT_3_6_VL		✓
Comportamental	CULTURE_CAT_6M_VISITAS		✓
Comportamental	CULTURE_CAT_6M_VL		✓
Comportamental	CULTURE_FLG_VISIT_LAST12M		✓
Comportamental	CULTURE_INDEX_3M_VISITAS		✓
Comportamental	CULTURE_MAXPRD_MLOI_6M_NR		✓
Comportamental	CULTURE_RATIO_CAT_3_6_VIS		✓
Comportamental	CULTURE_RATIO_CAT_3_6_VL		✓
Comportamental	DAIRY_CAT_6M_VISITAS		✓
Comportamental	DAIRY_CAT_6M_VL		✓
Comportamental	DAIRY_FLG_VISIT_LAST12M		✓
Comportamental	DAIRY_INDEX_3M_VISITAS		✓
Comportamental	DAIRY_RATIO_CAT_3_6_VIS		✓
Comportamental	DAIRY_RATIO_CAT_3_6_VL		✓
Comportamental	DELICATESSEN_CAT_6M_VISITAS		✓
Comportamental	DELICATESSEN_CAT_6M_VL		✓
Comportamental	DELICATESSEN_FLG_VISIT_LAST12M		✓
Comportamental	DELICATESSEN_INDEX_3M_VISITAS		✓
Comportamental	DELICATESSEN_MAXPRD_MLOI_6M_NR		✓
Comportamental	DELICATESSEN_RATIO_CAT_3_6_VIS		✓
Comportamental	DELICATESSEN_RATIO_CAT_3_6_VL		✓
Comportamental	DRINKS_CAT_6M_VISITAS		✓
Comportamental	DRINKS_CAT_6M_VL		✓
Comportamental	DRINKS_FLG_VISIT_LAST12M		✓
Comportamental	DRINKS_INDEX_3M_VISITAS		✓
Comportamental	DRINKS_MAXPRD_MLOI_6M_NR		✓
Comportamental	DRINKS_RATIO_CAT_3_6_VIS		✓
Comportamental	DRINKS_RATIO_CAT_3_6_VL		✓
Comportamental	FISH_CAT_6M_VISITAS		✓
Comportamental	FISH_CAT_6M_VL		✓
Comportamental	FISH_FLG_VISIT_LAST12M		✓
Comportamental	FISH_INDEX_3M_VISITAS		✓
Comportamental	FISH_MAXPRD_MLOI_6M_NR		✓
Comportamental	FISH_RATIO_CAT_3_6_VIS		✓
Comportamental	FISH_RATIO_CAT_3_6_VL		✓
Comportamental	FROZEN_CAT_6M_VISITAS		✓
Comportamental	FROZEN_CAT_6M_VL		✓
Comportamental	FROZEN_FLG_VISIT_LAST12M		✓
Comportamental	FROZEN_INDEX_3M_VISITAS		✓
Comportamental	FROZEN_RATIO_CAT_3_6_VIS		✓
Comportamental	FROZEN_RATIO_CAT_3_6_VL		✓
Comportamental	FRTSVGTBLS_CAT_6M_VISITAS		✓
Comportamental	FRTSVGTBLS_CAT_6M_VL		✓
Comportamental	FRTSVGTBLS_FLG_VISIT_LAST12M		✓
Comportamental	FRTSVGTBLS_INDEX_3M_VISITAS		✓
Comportamental	FRTSVGTBLS_MAXPRD_MLOI_6M_NR		✓
Comportamental	FRTSVGTBLS_RATIO_CAT_3_6_VIS		✓
Comportamental	FRTSVGTBLS_RATIO_CAT_3_6_VL		✓
Comportamental	GROCERIES_CAT_6M_VISITAS		✓
Comportamental	GROCERIES_CAT_6M_VL		✓
Comportamental	GROCERIES_FLG_VISIT_LAST12M		✓
Comportamental	GROCERIES_INDEX_3M_VISITAS		✓
Comportamental	GROCERIES_MAXPRD_MLOI_6M_NR		✓
Comportamental	GROCERIES_RATIO_CAT_3_6_VIS		✓
Comportamental	GROCERIES_RATIO_CAT_3_6_VL		✓
Comportamental	HYGIENICPRD_CAT_6M_VISITAS		✓
Comportamental	HYGIENICPRD_CAT_6M_VL		✓
Comportamental	HYGIENICPRD_FLG_VISIT_LAST12M		✓
Comportamental	HYGIENICPRD_INDEX_3M_VISITAS		✓
Comportamental	HYGIENICPRD_MAXPRD_MLOI_6M_NR		✓
Comportamental	HYGIENICPRD_RATIO_CAT_3_6_VIS		✓
Comportamental	HYGIENICPRD_RATIO_CAT_3_6_VL		✓
Comportamental	LIGHTBAZAAR_CAT_6M_VISITAS		✓
Comportamental	LIGHTBAZAAR_CAT_6M_VL		✓
Comportamental	LIGHTBAZAAR_FLG_VISIT_LAST12M		✓
Comportamental	LIGHTBAZAAR_INDEX_3M_VISITAS		✓
Comportamental	LIGHTBAZAAR_MAXPRD_MLOI_6M_NR		✓
Comportamental	LIGHTBAZAAR_RATIO_CAT_3_6_VIS		✓
Comportamental	LIGHTBAZAAR_RATIO_CAT_3_6_VL		✓
Comportamental	PERSONALHYGIENE_CAT_6M_VISITAS		✓
Comportamental	PERSONALHYGIENE_CAT_6M_VL		✓
Comportamental	PERSONALHYGIENE_INDEX_3M_VISITAS		✓
Comportamental	PERSONALHYGIENE_MAXPRD_MLOI_6M_NR		✓
Comportamental	PERSONALHYGIENE_RATIO_CAT_3_6_VIS		✓
Comportamental	PERSONALHYGIENE_RATIO_CAT_3_6_VL		✓
Comportamental	PROD_DESC_ACTUAL_NR_PR	✓	✓
Comportamental	PROD_PROM_ACTUAL_NR_PR		✓
Comportamental	PRSNLHYGIENE_FLG_VISIT_LAST12M		✓
Comportamental	SEGM_LIFESTYLE		
Comportamental	SEGM_SPENDVARIATION		
Comportamental	SEGM_VALUE		
Comportamental	TAKEAWAY_CAT_6M_VISITAS		✓
Comportamental	TAKEAWAY_CAT_6M_VL		✓
Comportamental	TAKEAWAY_FLG_VISIT_LAST12M		✓
Comportamental	TAKEAWAY_INDEX_3M_VISITAS		✓
Comportamental	TAKEAWAY_MAXPRD_MLOI_6M_NR		✓
Comportamental	TAKEAWAY_RATIO_CAT_3_6_VIS		✓
Comportamental	TAKEAWAY_RATIO_CAT_3_6_VL		✓
Comportamental	TOP1_CAT_10	✓	
Comportamental	TOP1_CAT_11	✓	
Comportamental	TOP1_CAT_23	✓	
Comportamental	TOP2_CAT_10	✓	
Comportamental	TOP2_CAT_11	✓	
Comportamental	TOP2_CAT_23	✓	
Comportamental	TOP3_CAT_10	✓	
Comportamental	TOP3_CAT_11	✓	
Comportamental	TOP3_CAT_23	✓	
Comportamental	TRANSAC_FDS_ACTUAL_PR		✓
Envolvimento	CATEGORIAS_ACTUAL_NRS		✓
Envolvimento	CATEGORIAS_LAST12M_NRS		✓
Envolvimento	CATEGORIAS_LAST3M_NRS		✓
Envolvimento	CATEGORIAS_LAST6M_NRS		✓
Envolvimento	DESCONTOS_ACTUAL_VL		✓
Envolvimento	DESCONTOS_ULT_12M_VL		✓
Envolvimento	DESCONTOS_ULT_3M_VL		✓
Envolvimento	DESCONTOS_ULT_6M_VL		✓
Envolvimento	GASTO_ACTUAL_VL		✓
Envolvimento	GASTO_ULT_12M_VL		✓
Envolvimento	GASTO_ULT_3M_VL		✓
Envolvimento	GASTO_ULT_6M_VL		✓
Envolvimento	PROD_DESC_ACTUAL_VL_PR		✓
Envolvimento	PROD_DESCONTOS_ULT_12M_VL_PR		✓
Envolvimento	PROD_DESCONTOS_ULT_3M_VL_PR		✓
Envolvimento	PROD_DESCONTOS_ULT_6M_VL_PR		✓
Envolvimento	PROD_PROM_ACTUAL_VL_PR		✓
Envolvimento	PROD_PROMOCAO_ULT_12M_VL_PR		✓
Envolvimento	PROD_PROMOCAO_ULT_3M_VL_PR		✓
Envolvimento	PROD_PROMOCAO_ULT_6M_VL_PR		✓
Envolvimento	TICKET_MED_ACTUAL		✓
Envolvimento	TICKET_MED_ULT_12M		✓
Envolvimento	TICKET_MED_ULT_3M		✓
Envolvimento	TICKET_MED_ULT_6M		✓
Envolvimento	TRANS_DESC_ACTUAL_VL_PR		✓
Envolvimento	TRANS_DESCONTOS_ULT_12M_VL_PR		✓
Envolvimento	TRANS_DESCONTOS_ULT_3M_VL_PR		✓
Envolvimento	TRANS_DESCONTOS_ULT_6M_VL_PR		✓
Envolvimento	TX_RSP_M	✓	
Envolvimento	VISITAS_ACTUAL_NR		✓
Envolvimento	VISITAS_ULT_12M_NR		✓
Envolvimento	VISITAS_ULT_3M_NR		✓
Envolvimento	VISITAS_ULT_6M_NR		✓

Tabela 22 - Variáveis para pré-processamento

III.1.3. Pré-Processamento

III.1.3.1. Transformações dos dados

Variables - Trans2

(none) ☐ not Equal to ☐ ...

Columns: ☐ Label ☐ Mining ☐ Basic ☒ Statistics

Name	Is	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
PROD_DESC_ACTUAL_VL	.	0	0	0.623325	0.034822	0.036657	3.005578	13.90225
PROD_DESCONTOS_12M_VL	.	0	0	0.342991	0.041149	0.041126	2.677186	8.916643
PROD_DESCONTOS_3M_VL	.	0	0	0.451264	0.040587	0.036975	2.536608	9.042206
PROD_DESCONTOS_6M_VL	.	0	0	0.401241	0.045447	0.040754	2.263832	6.760085
PROD_DESC_ACTUAL_NR_PR	.	0	0	1	0.109285	0.062346	2.210607	12.39469
GROCERIES_R_CT3_6VIS	.	0	0	1	0.495622	0.095232	1.514343	11.07382
PROD_PROMOCAO_12M_VL	.	0	0	1	0.277117	0.085794	1.408123	3.538321
CULTURE_CT6M_VL	.	0	0	150.845	43.08245	45.94388	1.162641	0.225672
CODFISH_CT6M_VL	.	0	0	90.15	25.5728	27.40271	1.156655	0.222595
PROD_PROM_ACTUAL_VL	.	0	0	1	0.267245	0.107708	1.156278	2.993277
ALCDRNKS_CT6M_VL	.	0	0	342.135	100.5157	104.3435	1.155617	0.17975
ALCDRNKS_IDX3MVIS	.	0	0	22878784	6618481	6671034	1.147007	0.333609
CODFISH_CT6M_VST	.	0	0	23.5	6.734383	6.73052	1.121697	0.319591
BIRDS_CT6M_VL	.	0	0	183.7	53.31031	52.01282	1.079905	0.255725
PROD_PROMOCAO_6M_VL	.	0	0	0.991462	0.309157	0.093166	1.079616	2.227875
CULTURE_CT6M_VST	.	0	0	29.5	9.044404	8.426157	1.073001	0.221543
DELICATESSEN_CT6M_VL	.	0	0	404.55	120.0463	111.8308	1.053794	0.260207
FISH_CT6M_VL	.	0	0	404.55	120.0463	111.8308	1.053794	0.260207
FROZEN_CT6M_VL	.	0	0	267.28	85.22143	74.35274	1.043451	0.193518
COMPETITIONINDEX	.	0	0.02	4.41	1.321487	1.286315	1.041497	-0.25412
ALCDRNKS_CT6M_VST	.	0	0	36.5	11.85949	10.54187	1.041112	0.082402
GROCERIES_R_CT3_6VL	.	0	0	1	0.487652	0.1069	1.040869	7.448812
DRINKS_CT6M_VL	.	0	0	501.63	173.1917	137.6019	1.025683	0.160472
PROD_PROMOCAO_3M_VL	.	0	0	1	0.312542	0.091093	1.025059	2.590388
BUTCHER_CT6M_VL	.	0	0	510.8	156.28	139.0914	1.016091	0.222451
PROD_PROM_ACTUAL_NR_PR	.	0	0	1	0.240567	0.073453	1.007298	5.518201
DESCONTOS_ACTUAL_VL	.	0	0	58.66	19.2854	15.88074	1.002765	0.213732
TICKET_MED_ACTUAL	.	0	0.31	129.4886	41.8477	32.36193	0.99842	0.207659
LIGHTBAZAAR_CT6M_VL	.	0	0	715.16	237.442	192.8071	0.993979	0.177179
DESCONTOS_ULT_12M_VL	.	0	0	619.845	220.9476	168.9759	0.981839	0.123389
DESCONTOS_ULT_6M_VL	.	0	0	381.575	132.092	102.1053	0.975863	0.158084
TRANS_DESC_ACTUAL_VL	.	0	0	0.343251	0.022374	0.020728	0.973657	1.10307
DESCONTOS_ULT_3M_VL	.	0	0	182.07	62.21418	47.59879	0.95024	0.186066
TRANS_DESC_6M_VL	.	0	0	0.120739	0.015491	0.012367	0.939975	1.243353
BIRDS_CT6M_VST	.	0	0	27	8.243947	7.198351	0.925701	0.112457
DRINKS_IDX3MVIS	.	0	0	6.467712	2.53152	1.608031	0.911354	0.234615
DELICATESSEN_CT6M_VST	.	0	0	28	9.117974	7.453166	0.907245	0.088591
FISH_CT6M_VST	.	0	0	28	9.117974	7.453166	0.907245	0.088591
BAKERY_IDX3MVIS	.	0	0	7.173957	2.709949	1.870993	0.906456	0.063004
TAKEAWAY_IDX3MVIS	.	0	0	7.173957	2.709949	1.870993	0.906456	0.063004
FRTSVGTBLB_CT6M_VL	.	0	0	508.88	181.5435	129.6034	0.903733	0.130868
BAKERY_CT6M_VST	.	0	0	84.5	32.77095	21.31485	0.90272	0.109966
TAKEAWAY_CT6M_VST	.	0	0	84.5	32.77095	21.31485	0.90272	0.109966
TICKET_MED_ULT_3M	.	0	0.32	132.7445	42.60871	31.60536	0.900123	0.066985
BAKERY_PRDLQJ6M	.	0	0	46	8.074929	4.177671	0.898713	1.522097

Tabela 23 – Estatísticas das variáveis de entrada SAS Miner

III.2. Modelação

III.2.1. Árvores de decisão

Data Role=Valid

Statistics	Tree	Tree3	Tree6
Valid: Kolmogorov-Smirnov Statistic	0.66	0.65	0.66
Valid: Average Squared Error	0.08	0.09	0.07
Valid: Roc Index	0.88	0.86	0.91
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.01	0.03	0.01
Valid: Cumulative Percent Captured Response	76.42	75.29	74.04
Valid: Percent Captured Response	9.96	9.65	6.39
Valid: Divisor for VASE	3896.00	3896.00	3896.00
Valid: Gain	663.07	649.12	638.74
Valid: Gini Coefficient	0.76	0.72	0.83
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.67	0.66	0.66
Valid: Kolmogorov-Smirnov Probability Cutoff	0.01	0.01	0.01
Valid: Cumulative Lift	7.65	7.51	7.40
Valid: Lift	2.01	1.92	1.27
Valid: Maximum Absolute Error	1.00	1.00	1.00
Valid: Misclassification Rate	0.09	0.10	0.14
Valid: Sum of Frequencies	1948.00	1948.00	1948.00
Valid: Average Squared Error with Priors	0.06	0.06	0.08
Valid: Misclassification Rate with Priors	0.02	0.03	0.01
Valid: Root Average Squared Error	0.28	0.30	0.27
Valid: Cumulative Percent Response	6.27	6.15	6.07
Valid: Percent Response	1.65	1.57	1.04
Valid: Sum of Squared Errors	308.22	340.41	289.21
Valid: Sum of Case Weights Times Freq	3896.00	3896.00	3896.00

Tabela 24- Estatísticas Árvores de Decisão

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected	Model		Valid:
Model	Node	Model Description	Misclassification Rate
Y	Tree	Decision Tree Gini - binaria	0.09189
	Tree3	Decision Tree Entropia - binaria	0.09805
	Tree6	Decision Tree QUI - binaria	0.13758

Tabela 25 - Legenda SAS para as árvores geradas

III.2.2. Regressão Logística

Variável	Nível	DF	Coefficiente	Erro Padrão	Wald Chi-Square	P-value
G_OPT_DESCONTOS_ULT_3M_VL	0	1	-0.7366	0.2215	11.06	0.0009
G_OPT_DESCONTOS_ULT_3M_VL	1	1	-0.2103	0.1453	2.09	0.148
G_OPT_DESCONTOS_ULT_3M_VL	2	1	0.2188	0.1688	1.68	0.1949
G_OPT_DESCONTOS_ULT_3M_VL	3	1	0.1961	0.1312	2.23	0.1352
G_OPT_DRINKS_R_CT3_6VL	0	1	0.7005	0.1907	13.49	0.0002
G_OPT_DRINKS_R_CT3_6VL	1	1	-0.0685	0.1483	0.21	0.6442
G_OPT_DRINKS_R_CT3_6VL	2	1	-0.2387	0.1159	4.24	0.0394
G_OPT_FROZEN_R_CT3_6VIS	0	1	0.186	0.1103	2.84	0.0919
G_OPT_FROZEN_R_CT3_6VIS	1	1	0.1222	0.0906	1.82	0.1771
G_OPT_HYGIENICPRD_CT6M_VL	0	1	0.5538	0.3957	1.96	0.1616
G_OPT_HYGIENICPRD_CT6M_VL	1	1	-0.0489	0.146	0.11	0.7376
G_OPT_HYGIENICPRD_CT6M_VL	2	1	-0.00904	0.1567	0	0.954
G_OPT_HYGIENICPRD_CT6M_VL	3	1	0.00426	0.1762	0	0.9807
G_OPT_TAKEAWAY_R_CT3_6VIS	0	1	0.2041	0.1877	1.18	0.277
G_OPT_TAKEAWAY_R_CT3_6VIS	1	1	-0.1865	0.2466	0.57	0.4495
G_OPT_TAKEAWAY_R_CT3_6VIS	2	1	0.3112	0.1328	5.49	0.0191
G_OPT_TAKEAWAY_R_CT3_6VIS	3	1	0.00637	0.1326	0	0.9617
G_OPT_TRANS_DESC_12M_VL	0	1	-0.5558	0.1583	12.33	0.0004
G_OPT_TRANS_DESC_12M_VL	1	1	-0.3285	0.1238	7.04	0.0079
G_OPT_TRANS_DESC_12M_VL	2	1	0.2676	0.1156	5.36	0.0206
G_OPT_TRANS_DESC_ACTUAL_VL	0	1	0.2897	0.0807	12.9	0.0003
G_OPT_TX_RSP_M	0	1	1.6626	0.1553	114.6	<0.001
G_OPT_TX_RSP_M	1	1	0.7715	0.1107	48.55	<0.001
G_OPT_TX_RSP_M	2	1	-0.311	0.1523	4.17	0.0412
G_OPT_VST_ACTUAL_NR	0	1	0.9495	0.2803	11.47	0.0007
G_OPT_VST_ACTUAL_NR	1	1	0.557	0.2201	6.4	0.0114
G_OPT_VST_ACTUAL_NR	2	1	0.0596	0.1916	0.1	0.7557
G_OPT_VST_ACTUAL_NR	3	1	-0.4165	0.1385	9.05	0.0026
G_OPT_VST_ULT_12M_NR	0	1	0.4628	0.1981	5.46	0.0195
G_OPT_VST_ULT_12M_NR	1	1	0.0573	0.2135	0.07	0.7884
G_OPT_VST_ULT_12M_NR	2	1	-0.0416	0.1285	0.1	0.7464
G_TOP1_CT10	0	1	3.5492	.	.	.
G_TOP1_CT10	1	1	-0.2048	0.1775	1.33	0.2487
G_TOP1_CT10	2	1	-0.1349	0.133	1.03	0.3106
G_TOP1_CT10	3	1	-0.4224	0.1415	8.92	0.0028
G_TOP1_CT10	4	1	-0.358	0.1738	4.25	0.0394
G_TOP1_CT10	5	1	-0.6502	0.1281	25.78	<0.001
G_TOP1_CT10	6	1	-0.4988	0.1421	12.33	0.0004
G_TOP1_CT11	0	1	4.6881	21.3898	0.05	0.8265
G_TOP1_CT11	1	1	0.1813	3.5905	0	0.9597
G_TOP1_CT11	2	1	-0.9579	3.5719	0.07	0.7886

Variável	Nível	DF	Coefficiente	Erro Padrão	Wald Chi-Square	P-value
G_TOP1_CT11	3	1	-0.4754	3.5678	0.02	0.894
G_TOP1_CT11	4	1	-0.7978	3.5671	0.05	0.823
G_TOP1_CT11	5	1	-0.7832	3.5701	0.05	0.8264
G_TOP2_CT11	0	1	-1.5656	0.7387	4.49	0.0341
G_TOP2_CT11	1	1	0.5472	0.3496	2.45	0.1176
G_TOP2_CT11	2	1	0.5089	0.1902	7.16	0.0075
G_TOP2_CT11	3	1	0.3786	0.1713	4.88	0.0271
G_TOP2_CT11	4	1	0.2494	0.1926	1.68	0.1953
G_TOP2_CT11	5	1	0.0426	0.1735	0.06	0.8059
G_TOP2_CT23	0	1	-0.3933	0.2802	1.97	0.1604
G_TOP2_CT23	1	1	0.7327	0.3424	4.58	0.0324
G_TOP2_CT23	2	1	0.3295	0.1606	4.21	0.0402
G_TOP2_CT23	3	1	-0.1029	0.1274	0.65	0.419
G_TOP3_CT10	0	1	6.0272	44.7783	0.02	0.8929
G_TOP3_CT10	1	1	-0.1146	7.4699	0	0.9878
G_TOP3_CT10	2	1	-0.8377	7.4657	0.01	0.9107
G_TOP3_CT10	3	1	-1.1077	7.464	0.02	0.882
G_TOP3_CT10	4	1	-1.1797	7.464	0.02	0.8744
G_TOP3_CT10	5	1	-1.3087	7.4641	0.03	0.8608
G_TOP3_CT23	0	1	0.4508	0.2084	4.68	0.0305
G_TOP3_CT23	1	1	0.0182	0.2033	0.01	0.9286
G_TOP3_CT23	2	1	0.1221	0.1109	1.21	0.2709
G_TOP3_CT23	3	1	0.0361	0.1222	0.09	0.7677
LOG_BUTCHER_CT6M_VST		1	0.2503	0.0692	13.09	0.0003
OPT_COMPETITIONINDEX	1	1	-0.1586	0.1189	1.78	0.182
OPT_COMPETITIONINDEX	2	1	0.3547	0.1494	5.64	0.0176
OPT_GASTO_ACTUAL_VL	1	1	1.2547	0.3018	17.28	<0.001
OPT_GASTO_ACTUAL_VL	2	1	0.4287	0.165	6.75	0.0094
OPT_GASTO_ACTUAL_VL	3	1	-0.0582	0.1411	0.17	0.6797
OPT_GASTO_ACTUAL_VL	4	1	-0.5897	0.1451	16.53	<0.001
OPT_IPC	1	1	-0.224	0.0812	7.61	0.0058
OPT_IPC	2	1	0.1685	0.0939	3.22	0.0727
OPT_VST_ULT_3M_NR	1	1	0.301	0.2295	1.72	0.1896
OPT_VST_ULT_3M_NR	2	1	0.4007	0.1984	4.08	0.0434
OPT_VST_ULT_3M_NR	3	1	-0.0711	0.1265	0.32	0.574
Intercept		1	1.9945	8.262	0.06	0.8092

Tabela 26 - Estimativas dos parâmetros associadas ao modelo logístico Stepwise

III.3. Discussão dos Modelos

Obs	NAME	NRULES	NSURROGATES	IMPORTANCE	VIMPORTANCE	RATIO
1	TX_RSP_M	4	1	1	1	1
2	SQRT_TRANS_DESC_12M_VL	1	2	0.90661	0.91827	1.01286
3	VST_ACTUAL_NR	4	0	0.69781	0.69375	0.99418
5	VST_ULT_12M_NR	5	1	0.50346	0.4891	0.97147
6	GROCERIES_R_CT3_6VL	1	2	0.43337	0.43432	1.00218
7	GASTO_ACTUAL_VL	3	2	0.38243	0.39316	1.02805
8	PREF_STORE	6	0	0.36708	0.20479	0.55788
9	EXP_CATEGORIAS_NACTUAL	5	5	0.33864	0.29924	0.88366
11	GASTO_ULT_3M_VL	1	1	0.28416	0.3185	1.12083
12	TOP1_CT10	7	0	0.26068	0.14041	0.53864
13	SQRT_TRANS_DESC_6M_VL	1	3	0.20534	0.12928	0.62959
15	TOP3_CT10	4	1	0.1862	0.08436	0.45306
16	PWR_TICKET_MED_ULT_12M	1	3	0.18592	0.15086	0.81144
17	SQR_CATEGORIAS_NLAST3	1	1	0.17962	0.12816	0.71354
18	LOG_CULTURE_CT6M_VL	1	2	0.17905	0.07342	0.41009
19	TOP3_CT11	5	0	0.16316	0.07978	0.48897
21	TOP2_CT23	3	1	0.15461	0.0148	0.09574
22	LOG_PROD_DESC_ACTUAL_NR_PR	4	0	0.15427	0.15342	0.99445
23	VST_ULT_6M_NR	1	1	0.15007	0.11465	0.76398
24	PROD_DESC_ACTUAL_VL	1	2	0.14251	0.09723	0.68229
25	SQRT_HYGIENICPRD_CT6M_VL	1	1	0.14221	0	0
26	SQRT_LIGHTBAZAAR_CT6M_VST	2	1	0.13958	0.13748	0.98494
27	ALCDRNKS_CT6M_VL	2	1	0.13612	0.09774	0.7181
28	TOP1_CT23	3	0	0.13379	0.11859	0.88638
29	TOP2_CT10	1	2	0.13172	0.10921	0.82911
30	LOG_LIGHTBAZAAR_R_CT3_6VL	1	1	0.12766	0.03387	0.26531
31	DRINKS_R_CT3_6VIS	2	1	0.12708	0.08634	0.6794
32	EXP_TICKET_MED_ULT_6M	1	2	0.1251	0.11029	0.88163
34	PWR_CATEGORIAS_NLAST6	1	2	0.12404	0.04757	0.38348
36	PROD_DESCONTOS_3M_VL	3	0	0.12124	0.13597	1.12152
37	LOG_BUTCHER_CT6M_VL	2	1	0.12086	0.10556	0.87346
39	BAKERY_R_CT3_6VL	2	0	0.12004	0.14566	1.21341
40	DAIRY_R_CT3_6VIS	2	1	0.11686	0.09106	0.77917
44	TOP3_CT23	3	0	0.11072	0.07456	0.67339
49	SQRT_DELICATESSEN_CT6M_VL	1	0	0.10012	0	0
50	PWR_TRANS_DESC_ACTUAL_VL	1	0	0.09986	0.09184	0.91974
51	SQRT_FRTSVGTBLS_CT6M_VL	1	1	0.09908	0.09211	0.92967
52	PROD_DESCONTOS_12M_VL	1	1	0.09795	0.08085	0.82547
53	HYGIENICPRD_R_CT3_6VL	1	1	0.09641	0.07503	0.77825
56	PRSNHYGIENE_R_CT3_6VL	1	1	0.09364	0.07237	0.77281
57	PWR_BUTCHER_R_CT3_6VIS	1	0	0.09254	0.05828	0.62976
63	BAKERY_R_CT3_6VIS	2	0	0.08923	0.10644	1.19283
66	LOG_DELICATESSEN_IDX3MVIS	1	0	0.08613	0.0332	0.38549
67	TOP2_CT11	1	1	0.08476	0.10221	1.20585
68	HYGIENICPRD_IDX3MVIS	1	0	0.08183	0.1209	1.47738
69	SQRT_BAKERY_CT6M_VL	1	0	0.07751	0.07104	0.91646
73	PWR_BIRDS_R_CT3_6VL	2	0	0.07556	0.11201	1.48252
76	DAIRY_CT6M_VST	1	0	0.07455	0	0
78	LOG_CULTURE_R_CT3_6VIS	1	0	0.07394	0.11104	1.50167
82	SQRT_DESCONTOS_ULT_12M_VL	1	0	0.0648	0.06567	1.01346
85	LOG_DELICATESSEN_R_CT3_6VIS	1	0	0.06324	0.07628	1.20608
86	SQRT_BUTCHER_IDX3MVIS	1	0	0.06245	0	0
88	FRTSVGTBLS_CT6M_VST	1	0	0.0611	0.04272	0.69911
89	GROCERIES_R_CT3_6VIS	1	0	0.06025	0.04195	0.69633
90	SEGM_AGE	1	0	0.05928	0	0
91	LOG_DELICATESSEN_R_CT3_6VL	1	0	0.05916	0.05339	0.90254
94	LOG_BIRDS_CT6M_VL	1	0	0.05508	0.02079	0.37756
96	PWR_CATEGORIAS_NLAST12	1	0	0.04944	0.06572	1.32917
97	PWR_CODFISH_R_CT3_6VL	1	0	0.04848	0.06624	1.36625
101	TOP1_CT11	1	0	0.03416	0.07048	2.06346

Tabela 27 - Variáveis ADEntropia por ordem de importância

III.3. Diagrama SAS

