



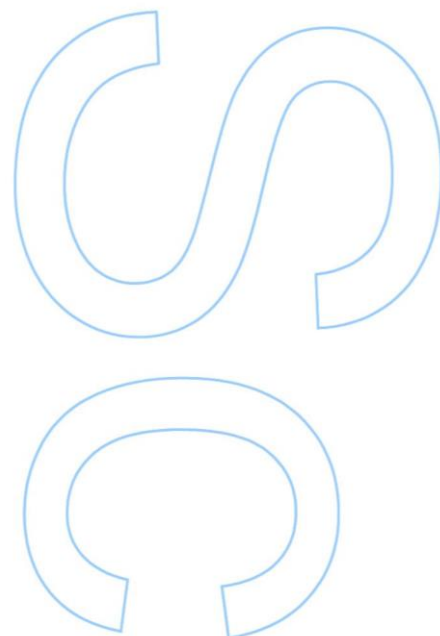
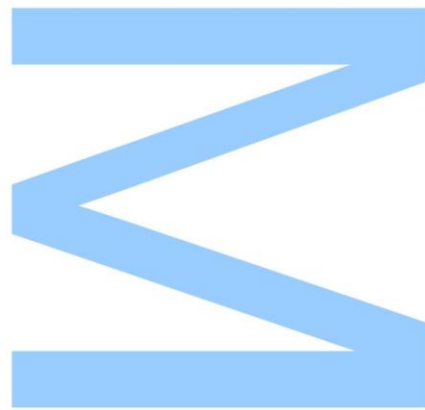
Previsão da compra de serviços em seguros de vida usando Data Mining

Catarina Inês Couto Correia

Engenharia Matemática
Departamento de Matemática
2019

Orientadores

Prof. Teresa Maria de Gouveia Torres Feio Mendonça, FCUP
Susana Maria Ferreira Pinto, Msg Life Iberia



Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____

N

S

O

Resumo

Este relatório relata todo o trabalho realizado ao longo do estágio curricular do Mestrado em Engenharia Matemática, na empresa Msg Life Ibéria. O presente trabalho teve como objetivo desenvolver um modelo que permita prever a compra de um serviço, usando técnicas de *Data Mining*. Para alcançar o objetivo foi necessário em primeiro lugar realizar um estudo sobre o ramo de negócio, a área seguradora, e posteriormente elaborar um estudo acerca dos dados a fim de perceber quais os melhores tratamentos a serem usados, um levantamento da informação relevante e os modelos mais adequados. Devido às características não equilibradas da compra e da não compra de serviços foi necessário aplicar técnicas de balanceamento de classes, tais como, *Oversampling*, *Downsampling*, *SMOTE* e *ROSE* a fim de se obter uma previsão menos enviesada. Compararam-se os modelos preditivos Árvore de Decisão, *Random Forest* e *Support Vector Machine*. Para este último modelo foi necessário aplicar Análise em Componentes Principais para reduzir a dimensão da base de dados. Depois da aplicação dos modelos preditivos compararam-se e escolheu-se o modelo com melhores resultados.

Palavras-chave: *Data Mining*, Seguros, Árvore de Decisão, *Random Forest*, SVM, PCA, *Imbalanced Domain Learning*, SMOTE, *Oversampling*, *Downsampling*, ROSE

Abstract

This report describes all the work realized throughout the Mathematical Engineering Master Degree's internship at Msg Life Iberia. The present study's aim was to develop a model that predicts the purchase of services, using techniques of Data Mining. Initially, to achieve this goal, it was necessary to do a study about the business, the insurance area, and then to perform a study about the data base in order to know what is the best techniques to apply, what are the most relevant information and the most applicable models. Due to characteristics of the purchase and not purchase of services it was necessary to apply imbalanced class techniques, such as Oversampling, Downsampling, SMOTE and ROSE in order to have a prediction with less bias. Decision Trees, Random Forest and Support Vector Machine models were compared. For the last model it was necessary to apply Principal Components Analysis to reduce the dimension of the data base. After the application of the predictive models, they were compared and the model with the best results was chosen.

Keywords: *Data Mining*, Insurance, Decision Tree, *Random Forest*, SVM, PCA, *Imbalanced Domain Learning*, SMOTE, *Oversampling*, *Downsampling*, ROSE

Agradecimentos

Em primeiro lugar quero agradecer à minha família, amigos e namorado pelo acompanhamento e por acreditarem comigo em todo este projeto.

Em segundo lugar quero dar meus sinceros agradecimentos a toda a equipa em que fui incorporada, mais especificamente, à Susana Pinto e ao Nuno Engana pelo apoio incansável em todos os percalços e pelo acompanhamento neste projeto. Também quero agradecer a todos os elementos da Msg Life Ibéria pelo acolhimento e bem estar.

Em terceiro lugar quero agradecer aos meus colegas de Mestrado e Licenciatura, mais especificamente à Ana Ribeiro, ao Alexandre Nogueira, à Catarina Pinto, à Cátia Costa, à Francisca Frias e ao Hugo Sousa por me ajudarem em todo este percurso académico e pela confiança depositada.

Em quarto lugar quero agradecer à minha orientadora da FCUP, Prof Teresa Mendonça e aos professores Joaquim Costa, André Marçal e Maria João Rodrigues pelo apoio académico.

Lista de abreviaturas, acrónimos, siglas e símbolos

Abreviatura	Significado
AD	Árvore de Decisão
AUC	<i>Area Under the Curve</i>
CRISP-DM	<i>CRoss Industry Standard Process of Data Mining</i>
DS	<i>Downsampling</i>
DM	<i>Data Mining</i>
ETL	<i>Extraction , Transformation and Loading</i>
ML	<i>Machine Learning</i>
OS	<i>Oversampling</i>
PCA	<i>Principal Component Analysis</i>
ROC	<i>Receiver Operating Characteristic</i>
ROSE	<i>Random Over-Sampling Examples</i>
RF	Floresta Aleatória (<i>Random Forest</i>)
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
SVM	Máquinas de Suporte Vectorial (<i>Support Vector Machine</i>)
TP	<i>True Positive</i>
TPR	<i>True Positive Rate</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FPR	<i>False Positive Rate</i>
FN	<i>False Negative</i>

Conteúdo

1	Introdução	15
1.1	Motivação	15
1.2	Objetivos	16
1.3	Organização do documento	16
2	Introdução empresarial e do ramo de negócio	17
2.1	Contextualização empresarial	17
2.2	Ramo de negócio	18
2.2.1	Visão geral sobre seguros	18
2.2.2	Seguros do ramo vida	19
2.2.3	Ciclo de vida de um seguro	20
3	Conceitos fundamentais e trabalho relacionado	25
3.1	<i>Machine Learning</i>	26
3.1.1	Árvores de decisão	27
3.1.2	Floresta Aleatória (<i>Random Forest</i>)	29
3.1.3	Máquinas de Suporte Vectorial (<i>Support Vector Machine</i> (SVM))	30
3.2	Métodos e técnicas aplicados à base de dados	33
3.2.1	Redução de dimensão	33
3.2.2	Balanceamento de classes	34
3.3	Metodologia de trabalho	37
4	Previsão da compra de coberturas de um seguro	41
4.1	Metodologia	41
4.2	Análise do problema em questão / do negócio	42

4.3	Extração e análise dos dados	43
4.3.1	Processo de extração de dados	43
4.3.2	Análise de dados - descrição numérica e gráfica	45
4.3.3	Análise da correlação das variáveis	47
4.4	Pré-processamento da base de dados	48
4.4.1	Pré-processamento inicial	50
4.4.2	Tratamento inicial de valores em falta	51
4.4.3	Tratamento inicial de valores atípicos (<i>Outliers</i>)	55
4.4.4	Redução do número de categorias	56
4.4.5	Pré-processamento SVM	57
4.4.6	Aplicação de técnicas de balanceamento de classes	59
4.5	Aplicação de modelos preditivos	61
4.6	Comparação de modelos	63
4.7	Previsão da compra de cobertura em novos exemplos	69
5	Conclusão e análise dos resultados	71
	Appendices	79
A	Glossário 1 : ramo de seguros	81
B	Glossário 2 : conceitos básicos	85

Lista de Figuras

3.1	Transformação SVM [14]	31
3.2	Hiperplano separador ótimo para classes linearmente separável, usando a abordagem SVM [14]	31
3.3	Hiperplano separador ótimo para classes linearmente não separável, usando a abordagem SVM [14]	32
3.4	Esquema <i>Random Downsampling</i>	35
3.5	Esquema <i>Random Oversampling</i>	36
3.6	Esquema SMOTE	37
3.7	Processo CRISP-DM [10]	39
4.1	Esquema global do processo de ETL	44
4.2	Etapa de "Staging Job" no processo de ETL	44
4.3	Descrição gráfica de algumas variáveis	45
4.4	Correlação de <i>Pearson</i> das variáveis numéricas	48
4.5	Exemplo da decodificação usada para obter as coberturas selecionadas.	51
4.6	Gráficos de dispersão dos prêmios anuais da cobertura B	53
4.7	Gráficos de dispersão dos prêmios anuais da cobertura C	53
4.8	Exemplo ilustrativo K-médias	54
4.9	Diagramas de dispersão das variáveis prêmio fracionado, prêmio anual, capital em dívida e duração inicial do crédito	55

4.10 Gráficos da proporção da variância explicada á esquerda e gráfico da proporção cumulativa explicada á direita, ambos relativos ao problema da previsão da cobertura A.	58
4.11 Distribuições das coberturas A e C	59
4.12 Técnicas de balanceamento aplicadas a cobertura A	60
4.13 Técnicas de balanceamento aplicadas a cobertura C	60
4.14 Validação Cruzada com k=5 [5]	62
4.15 Matriz de confusão	64
4.16 Exemplo Curvas ROC	66
5.1 Variáveis mais importantes para prever a cobertura A	72
5.2 Variáveis mais importantes para prever a cobertura C	72

Lista de Tabelas

4.1	Análise numérica das variáveis numéricas	47
4.2	Agrupamento usando K-médias das variáveis relativas aos prémios anuais das coberturas B e C.	55
4.3	Comparação de modelos e técnicas com as respetivas medidas de comparação, do problema de previsão da cobertura A	67
4.4	Comparação de modelos e técnicas com as respetivas medidas de comparação, do problema de previsão da cobertura C	68

Capítulo 1.

Introdução

"The price of success is hard work, dedication to the job at hand, and the determination that whether we win or lose, we have applied the best of ourselves to the task at hand."

Vince Lombardi

1.1 Motivação

Com a constante evolução tecnológica as empresas procuram, através desta, meios de otimizar e automatizar processos. Para tal procuram sistemas que idealmente se adaptem à mudança. Áreas como a da Inteligência Artificial, *Data Mining* e *Machine Learning*, são fulcrais no paradigma tecnológico atual para auxiliar sistemas informáticos na necessidade constante de adaptação.

A existência de processos de apoio à decisão pode ser fundamental para a compra de certos serviços, sendo essa compra consequentemente lucrativa para a empresa que a comercializa (como é o caso das companhias de seguro). Houve assim a necessidade de montar este processo de apoio à decisão usando as áreas acima. Para tal, é necessário entender o principal objetivo do problema, perceber quais as técnicas mais adequadas para o problema em questão e avaliar como se podem aplicar em casos reais.

1.2 Objetivos

O objetivo principal deste trabalho é desenvolver um modelo que permita prever a compra de um serviço, que neste caso se refere a uma cobertura, usando técnicas de *Data Mining*. Em paralelo é importante perceber quais as variáveis mais importantes/significativas para o problema em questão. Depois de desenvolvido este modelo de previsão é necessário perceber como se pode implementar/integrar este modelo nas ferramentas da empresa. Por fim, depois de integrado este modelo perceber se existe a necessidade e como se pode proceder à manutenção e reavaliação do modelo escolhido, a fim de se ter um modelo o mais atual possível. Este estudo foi elaborado usando o RStudio Server.

1.3 Organização do documento

O relatório de estágio está dividida em 5 capítulos principais. O presente Capítulo faz um pequeno enquadramento dos temas abordados. Aqui são descritos os conceitos fundamentais relacionados com a temática, motivação, importância e descrição do problema abordado no estágio. No Capítulo 2 é feita uma contextualização empresarial e são introduzidos alguns conceitos relacionada com o ramo de negócio em que se enquadra o tema . Seguidamente no Capítulo 3 são introduzidos conceitos fundamentais relacionados com *Data Mining*, *Machine Learning* e *Inteligência Artificial*, temas chave para a resolução do problema em questão. É feita uma breve explicação dos conceitos, a utilidade destas áreas, os algoritmos mais usados e uma breve explicação sobre modelos preditivos. No Capítulo 4 são apresentados os resultados e a metodologia de trabalho realizada. O último capítulo, o Capítulo 5, descreve as conclusões obtidas a partir do trabalho apresentado no capítulo anterior. Por fim em apêndice encontram-se duas listas de glossários sobre o ramo de seguros e sobre conceitos básicos usados. Estas listas servem para uma melhor compreensão dos temas.

Capítulo 2.

Introdução empresarial e do ramo de negócio

"Rethinking insurance."

Msg Life

2.1 Contextualização empresarial

A Msg Life é uma das principais especialistas em Tecnologias de Informação (TI) para seguradoras, com mais de 35 anos de experiência e especificação atuarial . Combinando conhecimentos atuariais especializados e profissionais de TI oferece aos clientes vantagens competitivas decisivas. A Msg Life disponibiliza soluções inovadoras, fundamentais para o sucesso dos clientes, ajudando-os a superar os desafios da digitalização de forma sustentável e com garantia de fiabilidade presente e futura. Com uma oferta que abrange desde os serviços de consultoria tecnológica até ao desenvolvimento de soluções de software para seguros, o seu portfólio cobre todo o ciclo de vida do negócio segurador [31].

A Msg Life conta com cerca de 900 colaboradores separados por 18 escritórios presentes em 11 países europeus e também nos EUA. Em relação ao trabalho realizado, existem mais de 100 implementações com sucesso em clientes de mais de 30 países na Europa, nos EUA e na Austrália.

A Msg Life Iberia pertence a um grande grupo internacional chamado Msg Life que tem sede na Alemanha. Em 1980 criou-se a empresa FJA, dez anos mais tarde criou-se outra em-

presa chamada COR e em 2009 houve a fusão destas duas empresas originando-se a COR&FJA AG. Em 2012 houve a criação da COR&FJA AG Portugal e passados dois anos esta empresa renomeou-se para a Msg Life Iberia, iniciando a sua atividade em Espanha.

Atualmente, desenvolve nos seus escritórios do Porto um produto para comercialização e distribuição de produtos de seguros, que pretende expandir para dar resposta às tendências de "massificação da personalização".

No âmbito do estágio curricular sobre previsão da compra de serviços em seguros de vida usando *Data Mining*, realizaram-se várias formações sobre o ramo dos Seguros e sobre as tecnologias e soluções usadas na empresa, tais como, *Product Machine* e msg.Sales, sendo a última uma solução pensada e desenvolvida para a criação e distribuição de produtos em diferentes dispositivos e canais. Este estágio tem como objetivo de criar ferramentas de auxílio à personalização ou recomendação da oferta, dentro do portfólio de produtos disponível, através da ferramenta de venda da Msg Life Iberia (recomendando produtos tipicamente comprados por grupos de clientes a identificar).

2.2 Ramo de negócio

O enquadramento empresarial da Msg Life Iberia é claramente a área dos Seguros. Esta área de negócio é muito vasta e completa, sendo que será descrita de um modo genérico nesta secção. No Capítulo 4 é complementada a informação e descrita com detalhe a aplicação proposta.

2.2.1 Visão geral sobre seguros

Como descrito no Glossário A , um contrato de Seguro é um contrato estabelecido entre duas partes, o Segurador/Companhia de Seguros e a Pessoa Segurada, através do qual o primeiro assume a cobertura de determinados riscos, comprometendo-se a satisfazer as indemnizações ou a pagar o capital seguro em caso de ocorrência de sinistro, nos termos acordados.

Deste modo, um contrato de seguro tem os seguintes intervenientes:

- Companhia de Seguros/Seguradora;
- Tomador de Seguro;
- Pessoa/Objeto Seguro;
- Subscritores (intervenientes em caso de operações de capitalização);
- Beneficiário.

Note-se que o Tomador de Seguro e a Pessoa Segura podem ser a mesma pessoa e todos os intervenientes de um seguro estão descritos na apólice. No entanto, vários intervenientes podem consistir na mesma pessoa, como acontece no caso dos Seguros de Vida, em que a Pessoa Segura, o Tomador de Seguro e o Beneficiário tipicamente são a mesma pessoa.

Os seguros podem ser de dois tipos: individuais ou de grupo [1]. Os primeiros são efetuados relativamente a uma pessoa, podendo o contrato incluir no âmbito de cobertura o agregado familiar ou um conjunto de pessoas que vivam em economia comum, ou o seguro efetuado conjuntamente sobre duas ou mais cabeças. Os segundos são seguros de um conjunto de pessoas ligadas entre si e ao Tomador do Seguro por um vínculo ou interesse comum, que não o de segurar.

Mais especificamente, os seguros de grupo podem ser contributivos ou não contributivos e tanto num como noutro, quem tem a obrigação de pagar o prémio ao Segurador/Companhia de Seguros é o Tomador do Seguro. Os primeiros correspondem aos seguros em que é a Pessoa Segura que paga o prémio ou parte do prémio, ao contrário dos seguros de grupo não contributivos em que a totalidade do prémio é paga pelo Tomador de Seguro.

2.2.2 Seguros do ramo vida

Nos Seguros do Ramo Vida são asseguradas indemnizações, o capital seguro, pelo Segurador/Companhia de Seguros em caso de sinistro sobre a Pessoa Segura [3].

Existem três grandes grupos ou modalidades de seguros do Ramo Vida:

- Seguro de Vida;

- Seguro Financeiro;
- Seguro de nupcialidade e natalidade.

Mais genericamente tem-se a terminologia "Seguros de Risco" que exploram a ocorrência de morte/vida e têm como objetivo proteger um interesse pré-contratualizado associado a um indivíduo, em que o Tomador de Seguro paga uma certa prestação e havendo sinistro ocorre a remuneração do valor acordado inicialmente. No caso dos seguros financeiros existe uma certa prestação a ser paga periodicamente e no final o cliente pode receber parte do investimento ou então não receber nada, dependendo do tipo de seguro financeiro contratado.

De seguida são abordados em mais detalhe os Seguros de Vida, que cobrem o risco de morte ou o risco de sobrevivência da(s) pessoa(s) segura(s). Nestes seguros existem três grandes vertentes:

- seguros de vida em caso de morte,
- seguros de vida em caso de vida;
- seguros de vida que englobam as duas situações anteriores.

No caso dos seguros de vida em caso de morte garante-se o pagamento ao beneficiário quando ocorre a morte da pessoa segura. No caso dos seguros de vida em caso de vida existe uma espécie de poupança até uma certa data e em caso de vida no final do prazo existe, geralmente, direito ao montante. No terceiro caso tem-se uma mistura dos dois, em que é garantido um capital em caso de vida e outro em caso de sobrevivência. Também existe a vertente dos Seguros Mistos em que é feita uma operação de capitalização referente a um investimento, ficando o montante a render a uma certa taxa.

2.2.3 Ciclo de vida de um seguro

O processo inicia-se com uma Proposta de Seguro entregue pelo mediador ou pela seguradora na qual o cliente ao assinar estará a confirmar que tomou conhecimento de toda a informação contida na Proposta. De seguida a seguradora avalia o risco garantindo que existem todas

as condições para não haver prejuízos, através do processo de Subscrição. Por fim é emitida uma Apólice de Seguro, que consta toda a informação detalhada, incluindo condições gerais, particulares e especiais.

Mais especificamente os diferentes estados no processo de venda são descritos abaixo:

1. **Oportunidade:** fase inicial da venda onde se pedem apenas dados básicos;
2. **Proposta:** fase da proposta;
 - (a) **Apólice:** fase da Apólice (caso seja aceite a proposta)
 - (b) **Subscrição:** quando é preciso uma autorização da proposta por parte da Seguradora.
 Por exemplo, caso esteja um adolescente a querer fazer um contrato com valores gigantescos. Tem-se dois casos:
 - i. Rejeitado por parte da Seguradora. Este contrato não segue para a Apólice e acaba neste ponto.
 - ii. Aceite por parte da seguradora. Pode haver uma alteração em algum ponto da proposta ou então estava tudo correto e prossegue-se;
 - A. Cliente aceita e vai diretamente para a fase 2(a)
 - B. Cliente não aceita e não se prossegue para a Apólice.

No processo de venda existem vários intervenientes tais como, o agente/vendedor e o gestor comercial (quem cria ofertas comerciais).

Mais detalhadamente existe o ciclo de vida de uma Apólice que passa pelos seguintes passos:

1. Criação/Emissão: o cliente recebe a proposta e tem oportunidade de aceitar ou rejeitar a oferta;
2. Submissão: o Segurador/Companhia Segura recebe a submissão da Apólice de seguro;
3. Subscrição da Apólice (*Underwriting*): onde ocorre a análise de risco, avaliação/revisão da informação, validação dos dados comunicados, etc;

4. Alteração da Apólice: pode envolver alterações simples de conteúdo (por exemplo, morada de alguma das entidades), alterações relativas ao que foi contratualizado com o Segurador / Agente (prémio ou fracionamento) ou alterações que comprometam o funcionamento da apólice;
5. Renovação: quando a Apólice está perto de expirar é revisto o histórico da mesma e sob validação de condições na renovação opta-se pela efetivação (ou não) do processo. No entanto o subscritor da apólice tem a última palavra sobre esta renovação;
6. Cancelamento da Apólice: pode ser cancelada pela Seguradora (por exemplo, se o cliente não efetuar o pagamento do seguro) ou pelo cliente (por exemplo, se não estiver satisfeito com o seguro).
7. Caducação: quando a Apólice não é renovada pelo cliente gera-se este processo automaticamente para que a Apólice não continue em vigor.
8. Termo/Vencimento da Apólice: a data do término de uma cobertura de uma Apólice é a do seu vencimento (no dia seguinte à duração previamente estipulada e contratualmente definida), sem sinistro, ou quando ocorre o pagamento do mesmo.

Por fim, as informações disponibilizadas no contrato são as seguintes:

- Informação pré-contratual;
- Informação contratual legal/estabelecida por lei;
- Condições Gerais e/ou Especiais: as condições gerais são aquelas que são previamente elaboradas e apresentadas pelo Segurador/Companhia de Seguros ; as condições especiais as cláusulas que complementam ou especificam o contrato;
- Incumprimentos (condições particulares): é o conjunto de cláusulas que adaptam o contrato à situação concreta de um específico Tomador de Seguro.

Tanto as informações pré-contratuais como as informações contratuais legais estão na Proposta de Seguro e as condições gerais e/ou especiais e os incumprimentos estão descritas na Apólice.

No Ramo de Vida a existência de uma Proposta de Seguro não é imprescindível, não havendo um contrato de seguro apenas só com a Apólice; contrariamente ao Ramo de Não Vida em que pode haver só a Apólice. Este último ramo não será analisado.

Capítulo 3.

Conceitos fundamentais e trabalho relacionado

"A learning machine is any device whose actions are influenced by past experience."

Nils John Nilsson

Hoje em dia as áreas de ***Data Mining***, ***Machine Learning*** e **Inteligência Artificial** são cada vez mais usadas para resolver inúmeros problemas da vida real e para tornar automáticos processos que até agora eram vistos apenas como processos manuais.

Data Mining usa técnicas de Estatística e *Machine Learning* (ML) para extrair conhecimento dos dados. Segundo *Arthur Samuel* ML é uma área de estudo que permite que os computadores aprendam sem serem explicitamente treinados para isso [29]. Assim, através de um conjunto de dados previamente conhecido, o computador pode ser treinado para construir regras de decisão a serem aplicadas no futuro sempre que se pretende prever algo. Dado que ML induz conhecimento a partir da experiência (dados), constitui uma grande área da Inteligência Artificial em que se estuda como criar agentes inteligentes. Em termos históricos sabe-se que entre 1940 e 1950 começou-se a discutir a possibilidade a criação de um cérebro artificial e em 1956 fundou-se uma disciplina académica sobre este assunto. Em meados de 1950 Arthur Samuel começou estudos sobre ML com algoritmos simples aplicado a jogos de damas [29]. Seguidamente em 1990 começou a aparecer o termo DM.

ML engloba principalmente dois grandes métodos de aprendizagem: **Aprendizagem Supervisionada** e **Aprendizagem Não Supervisionada**. O primeiro é o termo usado sempre

que o programa é treinado sobre um conjunto pré definido, em que o principal objetivo é a previsão de uma classe. O segundo é o termo usado quando um programa pode automaticamente encontrar padrões e relações num conjunto de dados, neste caso o programa não possui um conhecimento prévio sobre os dados, sendo o principal objetivo descobrir grupos ou estruturas num conjunto de dados. Existe uma grande diferença entre estes dois métodos, no primeiro conhece-se a variável a prever e esta é usada no processo de aprendizagem, contrariamente ao segundo método em que esta não se conhece [12].

Existem duas grandes categorias da aprendizagem supervisionada: **classificação** e **regressão** [12]. Para um problema de classificação, o objetivo dos algoritmos de ML é categorizar ou classificar *inputs* baseados no conjunto de treino. Muitos problemas de classificação são binários, ou seja, apenas estão envolvidas duas classes (por exemplo, verdadeiro ou falso), no entanto podemos ter problemas com várias classes. Assim, com esta metodologia pretende-se prever uma classe (em que a natureza da variável resposta é categórica). Para um problema de regressão o objetivo dos algoritmos de ML é desenvolver uma relação entre *outputs* e *inputs* usando uma função contínua para se entender como é que os *outputs* se alteram para um dado *input*. Assim, com esta metodologia pretende-se prever valores contínuos (em que a natureza da variável resposta é numérica).

O contexto deste projeto é prever a compra de um serviço, isto é, prever se um dado cliente compra ou não um dado serviço. Assim, o estudo centra-se na categoria classificação binária em aprendizagem supervisionada.

3.1 Machine Learning

O objetivo de *Machine Learning* é encontrar uma boa/útil aproximação $\hat{f}(x)$ de uma função $f(x)$ desconhecida. A função $f(x)$, cujo domínio é igual ao sub-espaco p-dimensional (em que p é o número de variáveis explicativas), é uma função que generaliza o fenómeno [17].

Pensando neste objetivo surgem várias questões tais como, como é que percebemos qual a

melhor aproximação, como é que medimos se a aproximação é boa e por fim como é que a encontramos. Para responder a estas questões existem algoritmos de *Machine Learning* que nos ajudam e que nos dão uma boa aproximação. Perceber se a aproximação é a melhor ou se é boa o suficiente depende do problema em questão e vai ser descrito mais a frente.

Existem vários algoritmos para previsão, tais como:

- **Árvores de decisão**
- **Floresta Aleatória (*Random Forest*)**
- **Máquinas de Suporte Vectorial (*Support Vector Machine* (SVM))**

Estes algoritmos são descritos nas sub-seções anteriores e ambos pertencem à aprendizagem supervisionada. Apenas estes três algoritmos serão abordados e a justificação do porquê de se ter usado estes três algoritmos será referida na secção 4.5.

A base de dados, neste tipo de projetos, é separada em dois conjuntos: **conjunto de treino** e **conjunto de teste**. Cerca de 70% da base de dados foi escolhida aleatoriamente para o conjunto de treino e os restantes 30% da base de dados ficam no conjunto de teste [23] . O primeiro conjunto é usado para aplicação de modelos e o segundo para perceber se os modelos aplicados tem uma boa *performance* ou não. Pode-se testar os modelos no conjunto de treino, no entanto, como os modelos aprenderam sob este conjunto então é muito provável que tenham muito bons resultados neste conjunto, a mais valia é perceber como se comporta o modelo testando noutro conjunto de dados onde ele não aprendeu.

3.1.1 Árvores de decisão

Em *Data Mining*, uma árvore de decisão é um modelo preditivo que pode ser usado em problemas de classificação e de regressão (normalmente chamadas árvores de regressão) que apareceu nos anos 60. Como o contexto é um problema de classificação vamos usar as primeiras. Assim, uma árvore de decisão particiona o espaço preditivo em conjuntos de regiões e em cada uma dessas regiões ajusta um modelo simples (como uma constante). Assim, o processo

de modelação é feito usando um conjunto de decisões hierárquicas sob as variáveis, arranjadas numa estrutura em árvore, produzindo regras em que dependendo das características da informação observada retorna uma certa classe [13] [15].

Uma árvore é constituída por nós e nós terminais, em que os primeiros sofrem uma divisão/partição em dois e os segundos são nós que não sofreram nenhuma divisão e estão no final da árvore, produzindo a classe prevista. Para fazer cada divisão existe um critério de divisão do nó em que o objetivo é maximizar a separação entre as diferentes classes nos nós abaixo [13].

Para cada nó da árvore é necessário escolher a melhor variável do conjunto de p variáveis para segmentar esse nó. Portanto, o objetivo aqui é que os descendentes do nó sejam mais "puros" (com menos mistura de classes) do que o nó que o originou. Para tal, é necessário medidas de impureza de um certo nó t [13]. As medidas de impureza mais comuns em classificação são as seguintes:

- Entropia ou quantidade de informação de Shannon:

$$i(t) = - \sum_{j=1}^K p_j \log(p_j) \quad (3.1)$$

- Índice de Gini:

$$i(t) = 1 - \sum_{j=1}^K p_j^2 \quad (3.2)$$

em que p_j representa a proporção de classes j observadas nesse nó. As classes j vão de 1 a K .

Depois de calculada a impureza de cada nó usa-se a divisão que maximiza a redução da impureza dada por:

$$\Delta I = I(t) - p(t_L)I(t_L) - p(t_R)I(t_R) \quad (3.3)$$

sendo que o nó t se divide em dois nós t_L e t_R sendo o primeiro o nó "filho" esquerdo e o segundo o direito [34].

As vantagens deste método são [15]:

- transparência/interpretabilidade nos resultados, isto é, consegue-se dar uma explicação dos resultados olhando por exemplo para a estrutura da árvore;
- pode ser aplicado a bases de dados com variáveis explicativas de qualquer natureza
- consegue-se perceber quais as variáveis mais importantes para o modelo

Como referido acima, uma das vantagens deste algoritmo é capacidade de se extrair as variáveis mais importantes na previsão. As variáveis mais importantes são as que aparecem no topo da árvore e à medida que se desce da árvore a sua importância vai diminuindo. Posto isto, aplicando este algoritmo consegue-se não só visualizar as "regras" bem como perceber quais as variáveis mais importantes usadas para a previsão [34].

Para a aplicação deste algoritmo foi usado a função `rpart` (*Recursive Partitioning And Regression Trees*) do R, que implementa muitas ideias encontradas no livro CART [7] e nos programas produzidos pelos autores. Esta função tem vários parâmetros e que serão descritos na secção 4.5.

3.1.2 Floresta Aleatória (Random Forest)

A ideia fundamental da Floresta Aleatória (*Random Forest*) é a combinação de várias árvores de decisão num único modelo, fazendo parte de um **Ensemble Method** que correspondem à combinação de vários modelos a fim de obter valores previstos mais acertados.

Para classificar uma nova observação, cada árvore de decisão fornece a classificação, sendo colecionadas cada uma dessas classificações e escolhidas as previsões que são mais votadas como o resultado do modelo. As variáveis da base de dados que entram para cada árvore são aleatoriamente escolhidas da base de dados original.

Com este algoritmo apesar de conseguir melhorar os resultados perde-se a interpretabilidade nos resultados que se tinha usando apenas uma árvore de decisão. No entanto, consegue-se perceber quais as variáveis mais importantes para o modelo, a partir de duas medidas de

importância de cada variável: *Mean Decrease Accuracy* e *Mean Decrease Gini* [20].

Cada árvore tem uma amostra que não é usada na construção e esta amostra é usada para calcular a importância de cada árvore. Primeiro mede-se a *accuracy* da previsão nessa amostra, depois os valores da variável nesta amostra são aleatoriamente misturados mantendo as outras variáveis intactas e é medido o decréscimo na *accuracy* nos dados misturados. Por fim, calcula-se a média do decréscimo na *accuracy* para todas as árvores, obtendo assim a medida *Mean Decrease Accuracy*. Assim, esta medida mede quanto é que diminui a *accuracy* retirando essa variável ou quanto é que aumenta a *accuracy* incluindo essa variável [20].

Quando uma árvore é construída calcula-se o índice de Gini de cada nó. Assim, para cada variável calcula-se a soma do decréscimo no índice de Gini relativamente a todas as árvores. Por fim divide-se cada valor de decréscimo pelo número de árvores na floresta, dando uma média e obtendo-se assim *Mean Decrease Gini* [20].

Para a aplicação deste algoritmo foi usado a função `randomForest` do R que foi implementado baseando-se nas árvores CART [7]. Esta função tem vários parâmetros e serão descritos na secção 4.5.

3.1.3 Máquinas de Suporte Vectorial (Support Vector Machine (SVM))

As SVM's foram introduzidas em 1992 na *the COLT-92 conference* [6] que deram origem a uma nova classe de algoritmos chamados *kernel machines*. A ideia básica das SVM's é mapear a base de dados original num novo sistema (com dimensão elevada) de coordenadas onde as classes são linearmente separáveis, transformando cada ponto x num ponto $z = \{z_j = \phi_j(x)\}_1^M$ num espaço M-dimensional em que $M \gg \dim(x) = p$. [14]

A transformação num novo espaço pode ser observada na figura 3.1 .

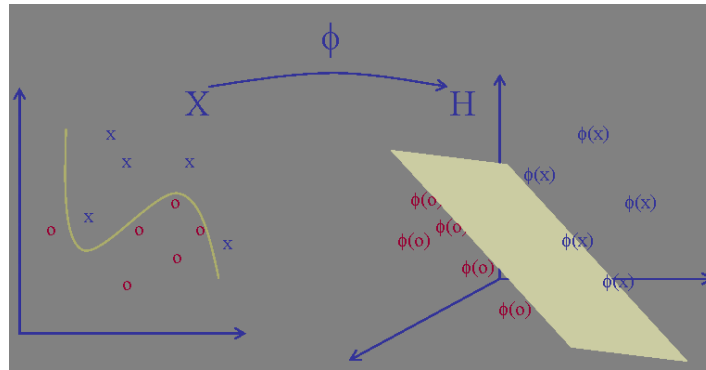


Figura 3.1: Transformação SVM [14]

Como é de esperar, existe um número infinito de hiperplanos que separam as duas classes e assim procura-se o hiperplano que resulta numa classificação mais acertada nos dados não observados. A abordagem SVM resolve este problema fazendo a procura pela margem máxima do hiperplano, assumindo implicitamente que quanto maior for a margem melhor será o erro no teste [14]. Apenas será abordado o caso em que se tem apenas 2 classes.

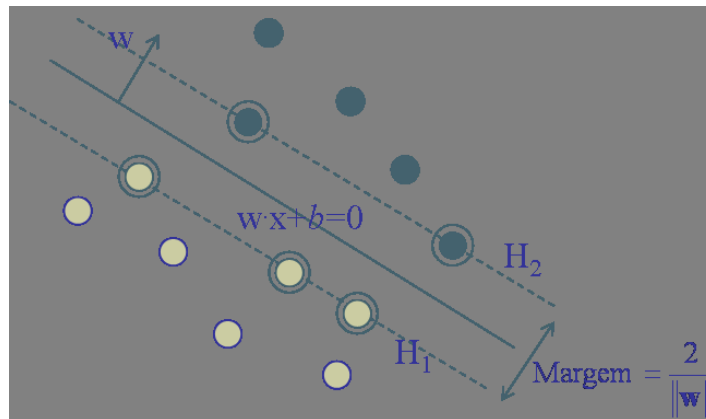


Figura 3.2: Hiperplano separador ótimo para classes linearmente separável, usando a abordagem SVM [14]

Na figura 3.2 pode ser observado como funcionam as SVM's no caso em que existe uma separação linear entre as classes. Nesta figura estão representados os pontos da base de dados a amarelo e a azul separados por cores representando as respetivas classes. O hiperplano satisfaz a equação $wx + b = 0$, onde x é um ponto, w é o vetor normal ao hiperplano (perpendicular ao hiperplano), $\frac{|b|}{\|w\|}$ é a distância perpendicular desde o hiperplano até a origem e $\|w\|$ é a norma Euclidiana de w . Também está representado os vetores de suporte (neste caso são os pontos que tocam na margem), a margem dada por $\frac{2}{\|w\|}$ e os hiperplanos H_1 e H_2 representados a tracejado [9].

O caso da separação linear é o caso mais simples, que raramente acontece. Assim, existe o caso linearmente não separável que pode ser observado na figura 3.2.

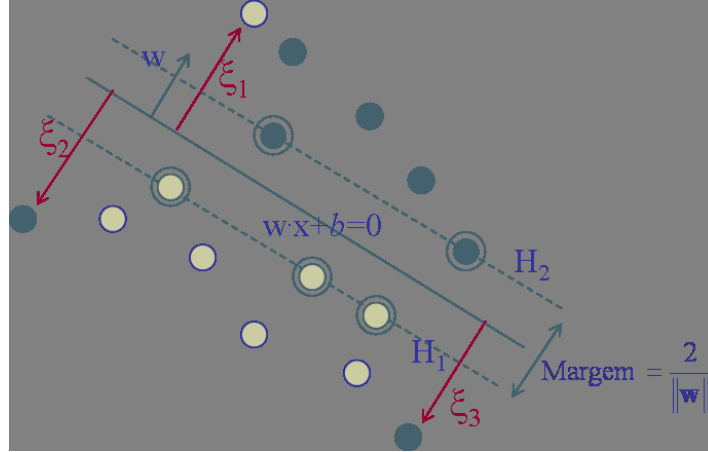


Figura 3.3: Hiperplano separador ótimo para classes linearmente não separável, usando a abordagem SVM [14]

No caso linearmente não separável admite-se que existam pontos que ficam do lado errado da margem, como se pode observar na figura 3.3. A quantidade total, representada pela soma dos ζ_i , que é permitida é um parâmetro de regularização do método. Neste caso os vetores de suporte são os pontos que estão em cima da margem e no lado errado da fronteira.

Um ingrediente importante nesta abordagem é a definição de produto interno em H dado por $z_i^T z_j = K(x_i, x_j)$, em que $K(x_i, x_j)$ representa a função núcleo [9]. As funções núcleo mais usadas são:

- **núcleo Linear:**

$$K(x_i, x_j) = x_i^T x_j \quad (3.4)$$

- **núcleo Radial:**

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (3.5)$$

, onde γ é um parâmetro a determinar

- **núcleo Polinomial:**

$$K(x_i, x_j) = (x_i^T x_j + 1)^d \quad (3.6)$$

, sendo d o grau do polinómio nos dados.

Estes núcleos vão ser testados na secção 4.5.

3.2 Métodos e técnicas aplicados à base de dados

3.2.1 Redução de dimensão

Hoje em dia, estamos perante uma enorme produção de dados devido à evolução da tecnologia. Assim, por vezes é necessário e útil aplicar-se técnicas de redução da dimensão da base de dados, pois grandes quantidades de dados envolvem mais tempo computacional, mais tempo envolvido no tratamento dos dados e mais memória de armazenamento.

Um dos métodos de redução de dimensão mais usados é **Análise em Componentes Principais (PCA)** proposta por *Karl Pearson* em 1901 [27], mais tarde desenvolvido independentemente por *Harold Hotelling* em 1933 [21] e expandido por *Karhunen–Loève* em 1969 [22]. O objetivo desta técnica corresponde a se houver um conjunto de n observações em $\mathbb{R}^p$ (sendo p o número de variáveis explicativas), encontrar uma forma de representar esses dados usando um menor número de variáveis. Assim, é necessário projetar os dados em sub-espacos interessantes, permitindo não só ver os dados mas também reduzir a dimensão. A projeção para um novo sistema de eixos para PCA corresponde a que os novos eixos apresentem grande variância, conseguindo assim ser analisado a contribuição de cada variável original para esse espaço.

Esta técnica é usada para extrair informação importante a partir de um conjunto multivariado e exprimi-la como um conjunto de novas variáveis chamadas Componentes Principais. A informação contida num conjunto de dados corresponde à variação total que este contém. Fica-se assim com Componentes Principais, em vez das variáveis originais, em que estas com-

ponentes são combinação linear das variáveis originais. Desta forma, o objetivo desta técnica é identificar componentes principais tal que a variação do conjunto de dados seja máxima [28].

3.2.2 Balanceamento de classes

Diz-se que se está perante um problema de classes desequilibradas quando não se tem a mesma percentagem/frequência nas duas classes. Isto revela ser um problema pois os modelos preditivos assumem alguns pressupostos, como por exemplo, a distribuição da variável a prever é equilibrada e os utilizadores (clientes) tem preferências uniformes [16]. Assim, entra-se numa área de aprendizagem diferente, chamada *Imbalanced Domain Learning* que tem os seguintes pressupostos [16]:

1. Os utilizadores (clientes) têm preferências não uniformes, isto é , os utilizadores indicam uma preferência enviesada para um subconjunto da variável a prever;
2. Os casos mais importantes/relevantes são aqueles que são raros ou extremos.

É de notar que [16]:

- apenas com as duas condições combinadas é que estamos perante um problema de *Imbalanced Domain*;
- o facto do utilizador ter uma preferência por um subconjunto de valores não é um problema se esse subconjunto for o comum;
- o facto de alguns valores serem raros não é um problema se esses não forem os valores mais importantes para o utilizador.

Existem duas estratégias habitualmente usadas para este tipo de problemas: técnicas de balanceamento das classes e alterar/adaptar algoritmos a este problema [16]. Optou-se pelo uso de técnicas de balanceamento de classes antes da aplicação dos modelos (na fase de pré-processamento), ao invés de usar algoritmos adaptados pois um dos objetivos é aplicar os modelos introduzidos neste capítulo, que por sua vez assumem que as classes a prever são

aproximadamente equilibradas, e isto só é possível usando técnicas de balanceamento de classes como pré-processamento antes da aplicação dos modelos.

Assim, uma das grandes vantagens de se usar esta estratégia é que se pode aplicar qualquer algoritmo de aprendizagem depois de a usar. No entanto, uma das desvantagens é que existe a dificuldade de decidir qual a distribuição ótima e estas técnicas podem reduzir ou aumentar o número de exemplos.

As técnicas de balanceamento de classes mais populares são as seguintes:

- ***Random Downsampling*** (ou ***Random Undersampling***)
- ***Random Oversampling***
- **SMOTE**
- **ROSE**

Na primeira técnica, *Random Downsampling*, reduz-se aleatoriamente o número de observações da classe maioritária (classe com mais elementos) a fim de haver o mesmo número de elementos nas duas classes [19]. Assim, a base de dados modificada obtida tem menos observações do que a original. A grande vantagem da aplicação desta técnica é que com bases de dados grandes reduzindo o número de observações melhora o tempo de execução e os problemas de armazenamento. No entanto, a grande desvantagem é que se pode perder informação importante pertencente à classe maioritária. Está ilustrado na Figura 3.4 um esquema desta técnica.

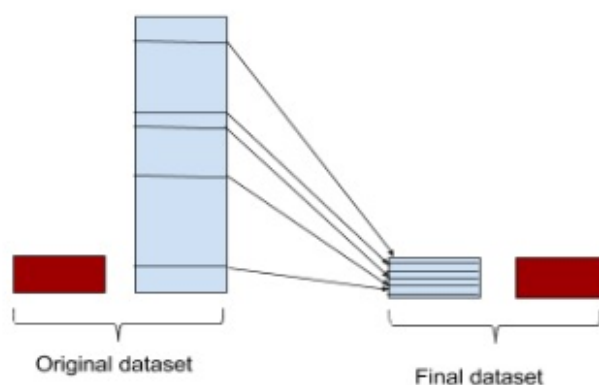


Figura 3.4: Esquema *Random Downsampling*

Na segunda técnica, *Random Oversampling* (ou *up-sampling*), replica-se aleatoriamente as observações da classe minoritária (classe com menos elementos) a fim de haver o mesmo número de elementos nas duas classes [19]. A grande vantagem desta técnica é que não se perde informação (contrariamente ao que acontece na técnica anterior). No entanto, pode haver *overfitting* (sobre-ajustamento) dos dados pois está-se a multiplicar/replicar observações de vários tipos da base de dados. Está ilustrado na Figura 3.5 um esquema desta técnica.

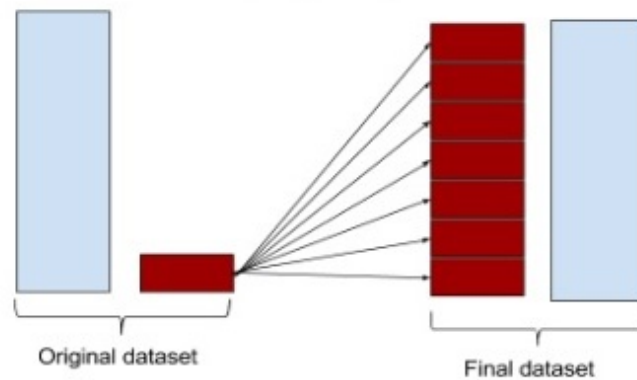


Figura 3.5: Esquema *Random Oversampling*

Na terceira técnica, SMOTE, em vez de se replicar ou reduzir observações este método gera um conjunto aleatório de dados artificiais/sintéticos semelhantes às observações da classe minoritária [11]. A geração de casos sintéticos usa interpolação de dois casos da classe minoritária. Um novo exemplo da classe minoritária é obtido usando um exemplo da classe e um dos seus vizinhos que é aleatoriamente selecionado usando K- vizinhos mais próximos. Assim, cria-se exemplos artificiais, não mexendo nos exemplos da classe majoritária. Usando esta técnica não há perda de informação e atenua o problema de *overfitting* causado por *Random Oversampling*. No entanto, pode não ser muito eficaz com bases de dados com altas dimensões e pode introduzir um adicional ruído, pois a informação artificial adicionada pode não ser real e interpretável. É ilustrado um esquema desta técnica na figura 3.6.

Por último, a mais recente técnica ROSE usa amostras *bootstrap* suavizadas para desenhar amostras artificiais a partir da vizinhança da classe minoritária [24]. Esta técnica combina *Random Oversampling* e *Random Downsampling* gerando uma amostra aumentada da base de dados, seguindo os seguintes passos [32]:

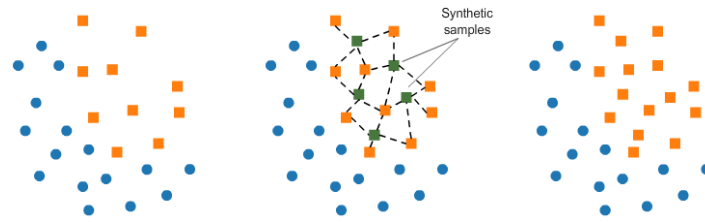


Figura 3.6: Esquema SMOTE

1. Reamostragem dos dados da classe maioritária usando *bootstrap* para remover o excesso da classe maioritária (*downsampling*)
2. Reamostragem dos dados da classe minoritária usando *bootstrap* para repetir as observações da classe minoritária (*oversampling*)
3. Geração de novas observações sintéticas na sua vizinhança

Estes três passos são repetidos para se produzir uma nova amostra sintética com aproximadamente o mesmo tamanho da base de dados original mas onde as classes estão agora igualmente representadas. No entanto podem-se produzir amostras irreais (fora dos valores possíveis).

3.3 Metodologia de trabalho

Existem várias metodologias de trabalho usadas e testadas com diferentes objetivos e adequadas para diferentes contextos, pois dependendo do contexto e do objetivo do problema pode ser necessário uma metodologia mais geral ou então uma metodologia que seja mais adequada ao contexto de negócio. A metodologia escolhida foi CRISP-DM e será introduzida de seguida.

O processo CRISP-DM, feito em 1996 através de um consórcio entre *Daimler-Chrysler*, NCR1 e SPSS2 [10] . Este processo centra-se nas seis fases seguintes e que podem ser analisadas na Figura 3.7:

- **Entender o Negócio:** foca-se no entendimento do objetivo do projeto e requer uma perspetiva de negócio;

- **Entender a Base de Dados:** começa-se com uma recolha inicial de dados, análise dos dados, identificando problemas de qualidade, descobrindo primeiro *insights* ou detetando subconjuntos de informação de interesse para formar hipóteses;
- **Preparação da Base de Dados:** preparação da base de dados para a fase de Modelação, começando com uma seleção de dados (por exemplo, remoção de *Outliers*), limpeza da base de dados (por exemplo, datas com formato incorretos, números interpretados como *strings*, etc), criação de novos dados e integração de bases de dados (caso seja necessário juntar duas ou mais fontes e dados diferentes);
- **Modelação:** selecionar várias técnicas e aplicados vários modelos com parâmetros diferentes a fim de obter os melhores parâmetros e modelos;
- **Avaliação:** os modelos são avaliados e existe a verificação se o objetivo é cumprido;
- **Deployment:** quando já se tem o modelo final e este é colocado em produção para que possa ser usado.

Uma das grandes vantagens desta metodologia é que pode ser aplicada a qualquer tipo de negócio pelo seu processo ser muito geral e não tem dependência da ferramenta para ser executada.

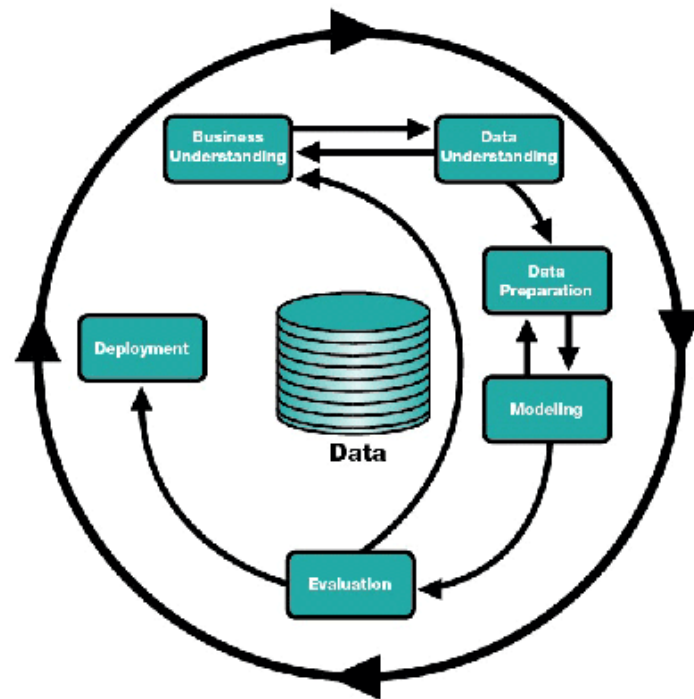


Figura 3.7: Processo CRISP-DM [10]

Percebe-se assim que o processo CRISP-DM é adequado devido à sua grande generalidade e por dar atenção ao contexto de negócio, que é fundamental para este projeto pois está integrado num contexto empresarial.

Capítulo 4.

Previsão da compra de coberturas de um seguro

"The best way to predict the future is to study the past, or prognosticate."

Robert Kiyosaki

4.1 Metodologia

Como citado no capítulo anterior, usou-se uma metodologia de trabalho mais adequada ao contexto do ramo segurador e adaptou-se a metodologia CRISP-DM, mais especificamente descrita pelos seguintes passos:

Passo 1: Entender o negócio

Passo 2: Extração dos dados e entender a base de dados

Passo 3: Pré-Processamento Geral da base de dados

Passo 4: Aplicação de Modelos Preditivos

Passo 5: Comparação e Avaliação de Modelos

Passo 6: Determinação do Modelo Final

Passo 7: Previsão com base em novos exemplos

Mais detalhes sobre cada passo serão feitos nas próximas secções.

4.2 Análise do problema em questão / do negócio

Foi estudado um produto de **Seguro de Vida indexado a um crédito de habitação**, isto é, é um Seguro de Vida com coberturas complementares para quem tem ou tencionam contrair um empréstimo habitação, em que o beneficiário é uma entidade credora , a pessoa segura. Associadas a este produto podem ser oferecidas coberturas opcionais e obrigatórias. Existem portanto duas coberturas obrigatórias, a cobertura M, cobertura de morte, e duas coberturas complementares mutuamente exclusivas de invalidez, as coberturas A e B. Adicionalmente é dada a oportunidade de escolha uma cobertura adicional C, sendo esta uma cobertura de morte por acidente de circulação.

Deste modo, existem vários cenários possíveis de compra:

- M+A
- M+B
- M+A+C
- M+B+C

A fim de o projeto ser o mais completo possível foi realizado a previsão da compra das coberturas A e C separadamente, pois a não compra de A implica a compra de B e vice versa. Assim, existem dois problemas de previsão feitos em paralelo, previsão da compra de A e previsão da compra de C.

4.3 Extração e análise dos dados

4.3.1 Processo de extração de dados

Hoje em dia com a constante evolução na tecnologia o volume de informação que circula nas empresas é cada vez maior, sem que toda seja devidamente captada, tratada e extraído o devido valor. Esta informação é armazenada sob diversos formatos, tais como bases de dados relacionais, folhas de cálculo, formulários digitais ou em papel, imagens, emails, etc. O passo fundamental para o desenvolvimento de um processo de *Data Mining* é a aquisição dos dados, pois sem eles não existe a base do trabalho para este processo. Assim, existe a necessidade de se montar um processo de extração de dados que se revelou complexo e dispendioso tanto temporalmente como tecnicamente, pois houve a necessidade de se perceber que informação relevante para o problema, qual a melhor forma de ser apresentada/retirada para análise e depois perceber como se iria extrair esta informação.

Anteriormente à extração de dados houve um processo de anonimização, zelando pela privacidade de cada cliente e da seguradora, uma vez que qualquer entidade que lide com dados pessoais de clientes está sujeita ao cumprimento das regras definidas no GDPR desde 25 de Maio do ano passado.

A extração de dados foi feita usando um processo de ETL, consistindo em 3 fases:

- **Passo 1: Extração** - a extração dos dados pode ser feita de várias origens e a partir de vários formatos
- **Passo 2: Transformação** - este passo pode envolver operações tais como filtragem, ordenação, agregação, limpeza, eliminação de duplicados e validação de dados. Este Processo ocorre de acordo com as regras de negócio.
- **Passo 3: Carregamento** - correspondendo ao carregamento dos dados para um arquivo de destino.

Em paralelo com a elaboração deste processo analisou-se as variáveis que fariam mais sentido serem incluídas na análise e como seria a melhor forma de as incluir.

Uma visão global do processo de ETL pode ser feita através da Figura 4.1, onde se pode observar as várias fases que se efetuou, iniciando pela etapa de "Staging Job" que está apresentada na Figura 4.2 e representa a extração dos contratos, da informação das coberturas, pessoas intervenientes no contrato (onde, por exemplo, se efetua uma ligação com um *webservice* da empresa CTT - Correios de Portugal S.A. com o intuito de enriquecer a informação disponível a partir do código postal.

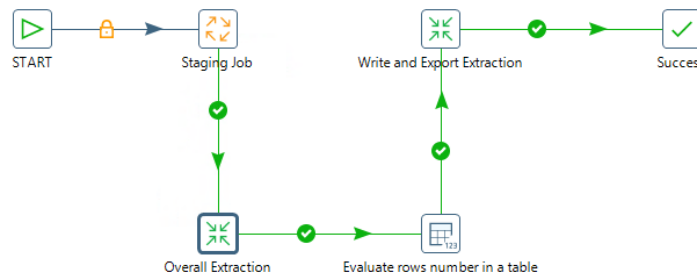


Figura 4.1: Esquema global do processo de ETL

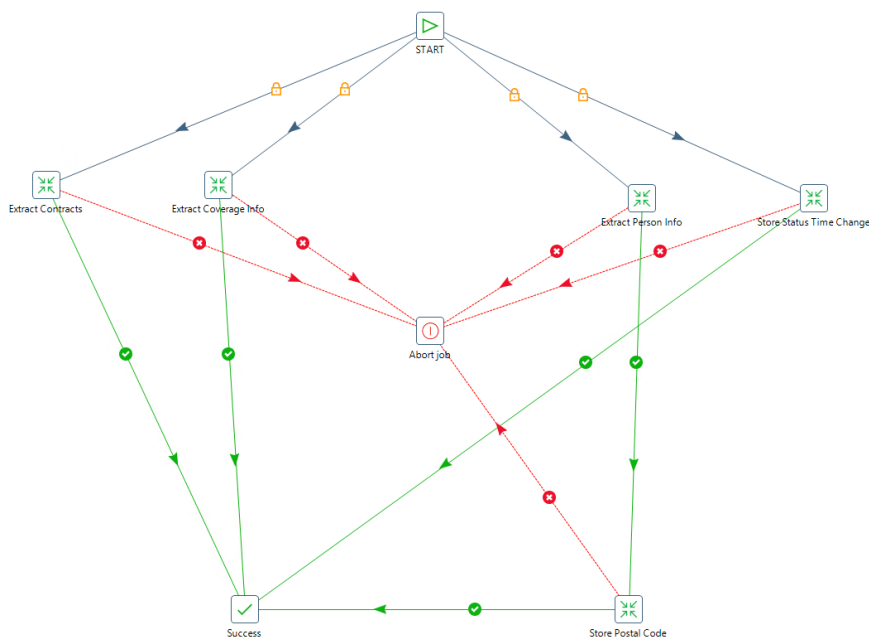


Figura 4.2: Etapa de "Staging Job" no processo de ETL

Finalizando o processo de extração de dados seguiu-se o processo de análise de dados.

4.3.2 Análise de dados - descrição numérica e gráfica

Do processo de ETL descrito na secção anterior resulta uma base de dados composta por 16750 clientes que aprovaram a simulação de uma proposta de seguro (cerca de aproximadamente 18% dos casos) ou que contraíram uma apólice de seguro. Além disso, existem 43 variáveis representando informações sobre o contrato da pessoa em questão.

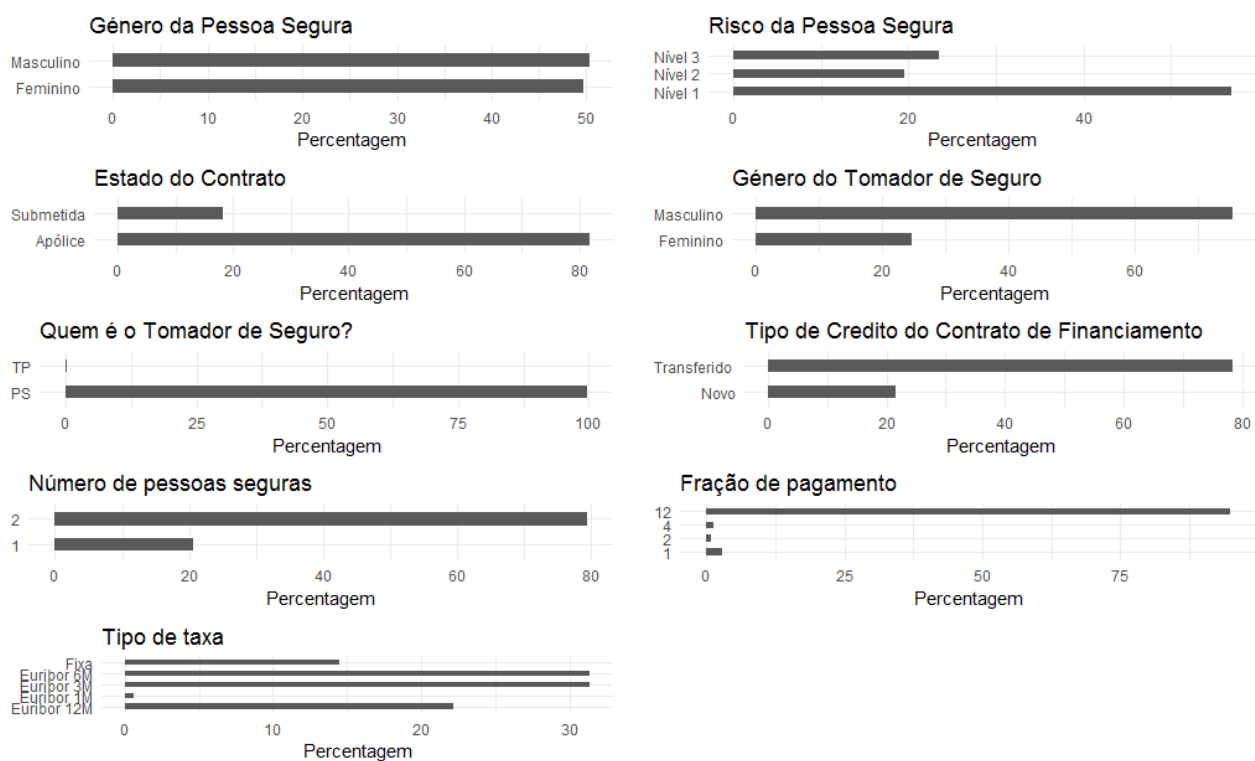


Figura 4.3: Descrição gráfica de algumas variáveis

Na Figura 4.3 estão representadas algumas das variáveis presentes na análise, as restantes variáveis são correspondentes tanto a valores de capitais seguros como valores de prémios anuais, percentagem coberta pela pessoa segura, data de nascimento tanto da primeira pessoa como do tomador de seguro, profissão e respetivo setor profissional da pessoa segura, localidade, concelho e distrito tanto da pessoa segura como do tomador de seguro, coberturas seleccionadas e respetivos prémios anuais por cobertura, capital em dívida do crédito, duração do crédito, entre muitas outras variáveis que vão sendo analisadas posteriormente. Existe uma percentagem semelhante dos diferentes géneros na amostra das pessoas seguras. No entanto, o tomador de

seguro é maioritariamente do género masculino, o que acontece porque a pessoa segura do contrato não tem de ser necessariamente o tomador. Além disso, só existe um tomador de seguro por apólice, ao passo que podem existir até 2 pessoas seguras por contrato. Também se percebe que geralmente o tomador de seguro é a mesma pessoa que a pessoa segura.

É de notar que a pessoa segura pode ser: a mesma pessoa que o tomador de seguro ou uma terceira pessoa. Deste modo, no gráfico do género da pessoa segura contém vários tipos de pessoas, o que pode ser verificado no gráfico do número de pessoas seguras (que podem ser 2 ou 1). Neste gráfico repara-se que existem mais contratos com 2 pessoas seguras (cerca de aproximadamente 80% dos casos).

Analisando o gráfico relativo ao tipo de crédito do contrato de financiamento observa-se que cerca de aproximadamente 78% são casos de contratos transferidos e os restantes são contratos de financiamento novos. Também olhando para o tipo de taxa contraída há uma preferência para taxas Euribor a 3 meses e a 6 meses. Por fim olhando para a fração de pagamento verifica-se que mais de 89% dos casos pagam o contrato mensalmente e os restantes pagam anualmente, de 3 em 3 meses ou 2 vezes por ano.

Juntamente com estas variáveis também se tem a variável relativa ao tipo de plano de amortização, isto é, possíveis maneiras de fazer o abatimento do capital para quando se chega ao final do crédito bancário se ter pago tudo o que se devia. Existem quatro possibilidades:

1. **Amortização Clássica:** corresponde à amortização mais comum (99% dos casos) onde ao longo dos anos se paga o capital mais os juros;
2. **Capital Diferido:** ao longo de um período contratualmente pré-definido vai-se pagando uma parte do capital e quando chega o final do contrato de financiamento/bancário tem de ser feito um depósito do capital em falta. Neste contexto existem duas variáveis adicionais relativas ao período de carência (representando o tempo de espera) e o valor residual em percentagem (durante x tempo vai se pagando uma percentagem do capital e no final do contrato tem de se pagar o valor restante, assim a percentagem do resíduo é o valor que se deixa para o final do contrato);

3. **Carência de Amortização de Capital:** durante meses só se paga juros e não existe abatimento de capital pois continua-se a dever o mesmo do que no início do contrato.

4. **Carência de Amortização de Capital + Capital Diferido:** junção dos dois últimos.

Relativamente às variáveis numéricas podem ser analisados valores da média, desvio padrão, mínimo e máximo na Tabela 4.1. Nota-se que existem valores muito dispares relativamente às variáveis que tem em conta prémios e capitais os diferentes fracionamentos (semestrais, trimestrais, ...) dos contratos, anualmente renováveis, originam valores de prémio muito díspares. Também o caso do capital em dívida pode tomar valores muito diversificados pois o contrato pode ser transferido e estar no final do contrato e assim haver um valor mais baixo do que , por exemplo, o caso o contrato ser novo e haver uma grande dívida a ser paga.

Variáveis Numéricas	Média	Desvio Padrão	Mínimo	Máximo
PS - Percentagem Coberta	98.94	7.38	10.00	100.00
Contrato - Fração do Pagamento	11.49	2.22	1.00	12.00
Contrato - Prémio fracionado	39.93	70.57	0.19	3145.06
Contrato - Prémio anual	398.74	572.65	2.28	37740.72
Credito - Capital em divida	77913.00	90492.77	457.00	5807266.00
Credito - Duração inicial em meses	378.60	112.48	12.0	4444.0
Credito - Duração remanescente em meses	289.80	122.01	12.00	4286.00
Credito - Período de Carência	24.05	23.22	1.00	90.00
Credito - Valor Residual em Percentagem	30.81	10.91	0.00	86.00

Tabela 4.1: Análise numérica das variáveis numéricas

4.3.3 Análise da correlação das variáveis

Antes da análise das correlações obtidas introduz-se teoricamente a que se refere.

Definição 1 *O coeficiente de correlação de Pearson é uma medida estatística que mede o grau de relação linear entre duas variáveis quantitativas. Esta medida toma valores entre -1 e 1 , sendo que valores extremos revelam alta correlação, tanto negativa como positiva.*

É de notar que se duas variáveis estão muito relacionadas será redundante a introdução das duas na análise, daqui percebe-se a importância de uma análise deste tipo antes do Pré-

Processamento. Analisando a Figura 4.4 em que está representada a respetiva correlação de *Pearson* entre as variáveis numéricas, nota-se que existe uma correlação muito elevada positivamente entre o prémio anual e o capital em dívida, o que é natural pois, em geral, quanto maior for o capital em dívida maior será o prémio anual a ser pago. Também se nota uma correlação elevada entre as variáveis relativas ao prémio fracionado e ao prémio anual, o que também é natural pois quanto maior for o valor do prémio fracionado maior será o valor do prémio anual. Apesar destas correlações serem elevadas não existe, nesta fase, evidência suficiente para as eliminar da análise.

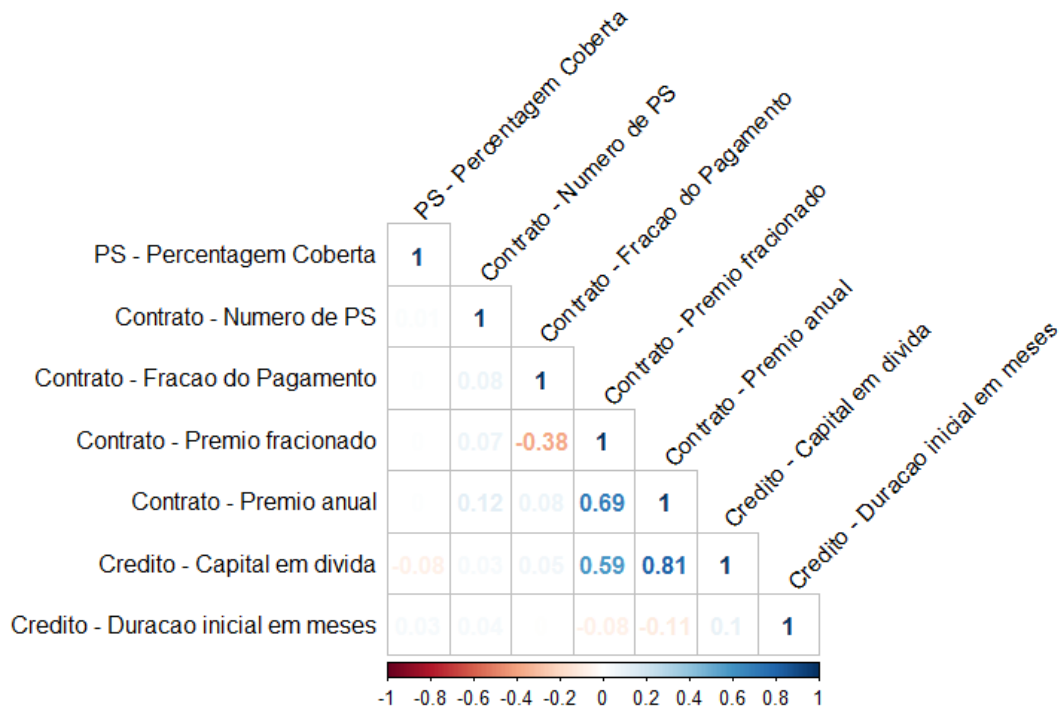


Figura 4.4: Correlação de *Pearson* das variáveis numéricas

4.4 Pré-processamento da base de dados

Definição 2 *O pré-processamento da base de dados é um conjunto de passos que envolvem a transformação da base de dados num formato compreensível.*

Existem vários motivos para se fazer este processo, que é diferente de base de dados para base de dados. Do processo de extração há informação referente a campos de texto livre, o que aumenta a possibilidade de erro, como por exemplo, "Portugal" e "PORTUGAL" que correspondem ao mesmo. Levantando assim a necessidade de se efetuar uma limpeza na base de dados, transformando-a num formato coerente. Também pode haver variáveis com valores em falta, ou seja, informação que deveria estar preenchida e não está. Pode também ser preciso criar variáveis novas fazendo a junção de duas, que pode ser mais informativo para o problema em questão. Um dos problemas mais comuns enfrentados nesta fase, é o tipo/natureza das variáveis, pois existem métodos de *Machine Learning* que apenas podem ser aplicados com determinados tipos/naturezas de bases de dados, o que acrescenta uma maior complexidade ao processo. Por fim, pode também existir um problema relacionados com a dimensão, isto é, a existência de demasiadas variáveis em estudo. Como é natural, quantas mais variáveis existirem mais complexo fica o problema e mais difícil de resolver e assim levanta-se a necessidade de aplicar técnicas para diminuir a dimensão do problema.

Deste modo, percebe-se a necessidade de fazer um tratamento prévio da base de dados e devido à sua complexidade nota-se que é a fase mais demorada de todo o estudo. Esta fase divide-se em várias sub-fases:

1. **Pré-processamento inicial** : onde se faz uma limpeza inicial a base de dados, analisa-se a introdução de variáveis novas e elimina-se informação redundante. Nesta fase também se estruturou a base de dados a fim de se ter a estrutura mais intuitiva e fácil de trabalhar;
2. **Breve tratamento de valores em falta** : onde são tratadas as variáveis incompletas;
3. **Breve análise da existência de *outliers*** : onde se faz um pequeno estudo da existência de valores demasiados extremos que podem prejudicar o estudo.

Estas 3 fases descritas anteriormente são comuns a todos os modelos aplicados, no entanto houve a necessidade de se acrescentar outra fase que vai ser explicada mais à frente. Cada

sub-fase será analisada seguidamente em secções diferentes.

4.4.1 Pré-processamento inicial

Além das variáveis categóricas mostradas na Secção 4.3.2 e na Figura 4.3, existem duas variáveis referentes à agência com a qual se contratualizou o crédito bancário, informações tanto da localização do balcão (por exemplo "Porto") bem como do banco que é associado o crédito contraído (por exemplo, "Banco BPI, S.A."). Houve a necessidade de se criar uma nova variável chamada "Agência- Banco e Balcão" que junta as variáveis relativas ao banco e ao balcão, pois juntas são mais informativas. Depois da criação desta nova variável eliminaram-se as duas variáveis usadas para a construção.

De seguida efetuou-se o tratamento da variável que contém as coberturas seleccionadas. Esta variável está codificada como um número decimal e o primeiro passo foi descodificar esta variável de decimal para binário. De seguida acrescenta-se um ou mais zeros à direita até este número ser igual ao número total de coberturas que o produto tem e inverte-se o número resultante dos passos anteriores por ser um número binário e se ler de trás para a frente. Por fim o número resultante deste último passo é dividido em n variáveis diferentes sendo n o número total de coberturas. Estes passos estão exemplificados na Figura 4.5, acabando assim com n variáveis binárias novas em que em cada uma delas, se tomar valor 1 representa que o cliente seleccionou a cobertura em questão, 0 caso contrário. Depois do tratamento da variável relativa às coberturas seleccionadas foi feito a mesma desmultiplicação para n variáveis que contem o prémio anual por cada cobertura.

Posteriormente iniciou-se um tratamento das variáveis relacionadas com datas, pois como não se pretende fazer um estudo temporal é mais informativo uma variável numérica do que com datas. A partir das variáveis relativas às datas de nascimento da pessoa segura, datas de nascimento do tomador de seguro e data de inicio do contrato criou-se as duas variáveis: idade no inicio do contrato da pessoa segura e do tomador de seguro, pois estes valores são mais

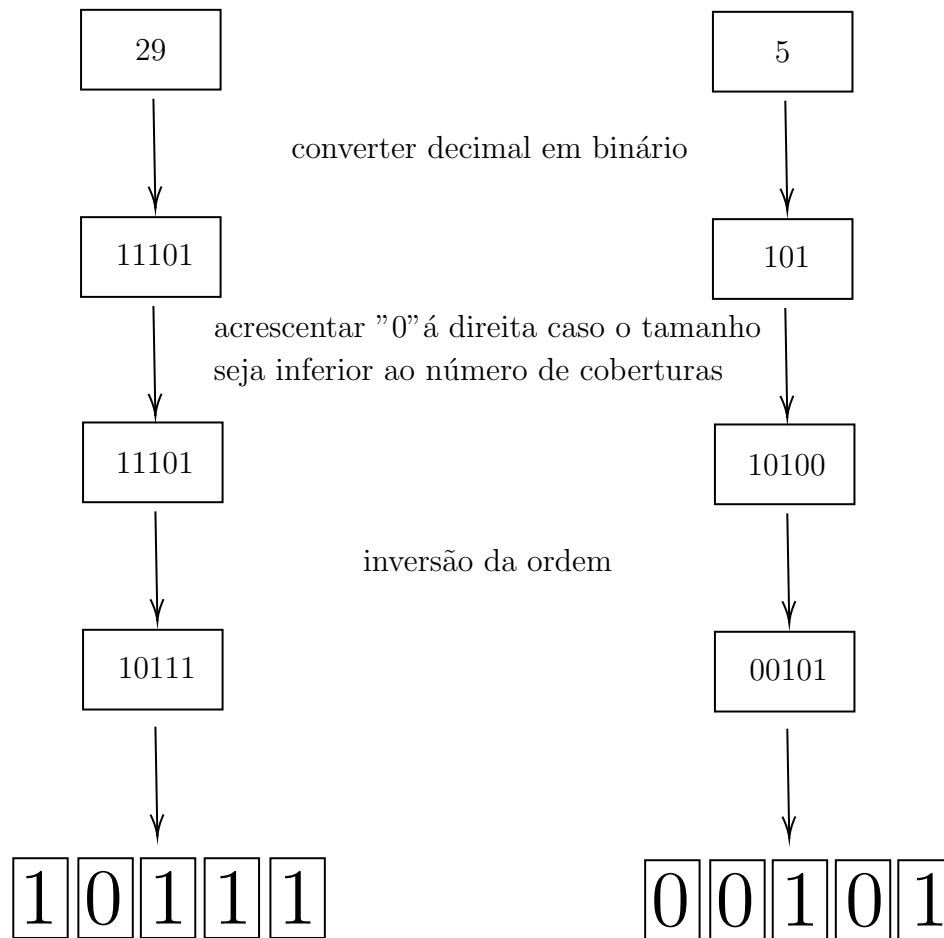


Figura 4.5: Exemplo da decodificação usada para obter as coberturas selecionadas.

informativos que as datas, para este tipo de problema. Também a partir das variáveis relativas à data de início do contrato e do crédito criou-se as variáveis com apenas o mês de início, tanto do crédito como do contrato. Optou-se por não usar os dias nem o ano por ser irrelevante o ano em que se criou o contrato e o dia.

4.4.2 Tratamento inicial de valores em falta

Variáveis com valores em falta são variáveis que não contém valores em todas as suas observações. Tipicamente no lugar destes valores aparece "NA" ou apenas um espaço em branco.

Em geral, os modelos não conseguem lidar com valores em falta e assim é necessário haver um tratamento quando tal situação acontece. Existem essencialmente dois tratamentos possíveis: eliminação do registo ou variável e substituição/imposição do registo usando valores "aproximados" (por exemplo média, mediana, moda). Os valores em falta estão em variáveis que não

são comuns a todas as coberturas, ou seja, são específicas de cada cobertura e por isso não faz sentido incluir estas variáveis para o estudo. Não se optou por substituir os valores em falta porque não fazem sentido para outros tipos de coberturas.

Depois da desmultiplicação dos valores dos prémios pelas coberturas seleccionadas, referido na Secção 4.4.1, verificou-se que as coberturas que não estão seleccionadas (isto é, que o cliente não compra) não vão ter valor nenhum no prémio anual dessa cobertura. A opção de eliminar estas observações não foi posta em causa, pois as observações seriam todas eliminadas devido ao facto de existirem 2 coberturas obrigatórias e mutuamente independentes, isto é, o cliente só pode ter uma das duas havendo sempre o valor da outra cobertura em falta.

Assim, optou-se por fazer a imposição do valor em falta para as variáveis dos prémios anuais. Para tal, pensou-se em substituí-los por "0" pois não se está a pagar por essa cobertura, mas isto faria com que a pessoa em questão tivesse todas as coberturas seleccionadas com pelo menos um prémio grátis, o que iria estragar a previsão das coberturas porque os modelos irão certamente preferir uma cobertura em que não se paga do que outra que se paga 10 euros. Com esta lógica não faz sentido impor um valor numérico nos valores em falta, pelo que se optou por categorizar estas variáveis da seguinte forma:

- valor em falta são transformados na categoria : "não seleccionada"
- valores iguais a 0 são transformados na categoria : "gratuita"
- valores diferentes de zero vão ser divididos em diferentes intervalos da forma : "(valor mínimo) até (valor máximo)"

Enfrenta-se agora outro problema que é a divisão dos valores em intervalos diferentes. Note-se que não se optou por criar uma única classe com os valores numéricos todos lá contidos pois existem valores muito distintos como se pode observar nas Figuras 4.6 e 4.7.

Esta divisão de intervalos não pode ser feita sem critério pois convém que os grupos escolhidos sejam de alguma forma equilibrados. Analisando os valores do prémio anual das coberturas verificou-se que os prémios da cobertura M, que é obrigatória não vai ter valores na categoria

”não selecionada” mas terá valores nas outras categorias, os prémios da cobertura A apenas tem as categorias ”não selecionada” e ”gratuita” e as restantes coberturas B e C terão de ser analisados.

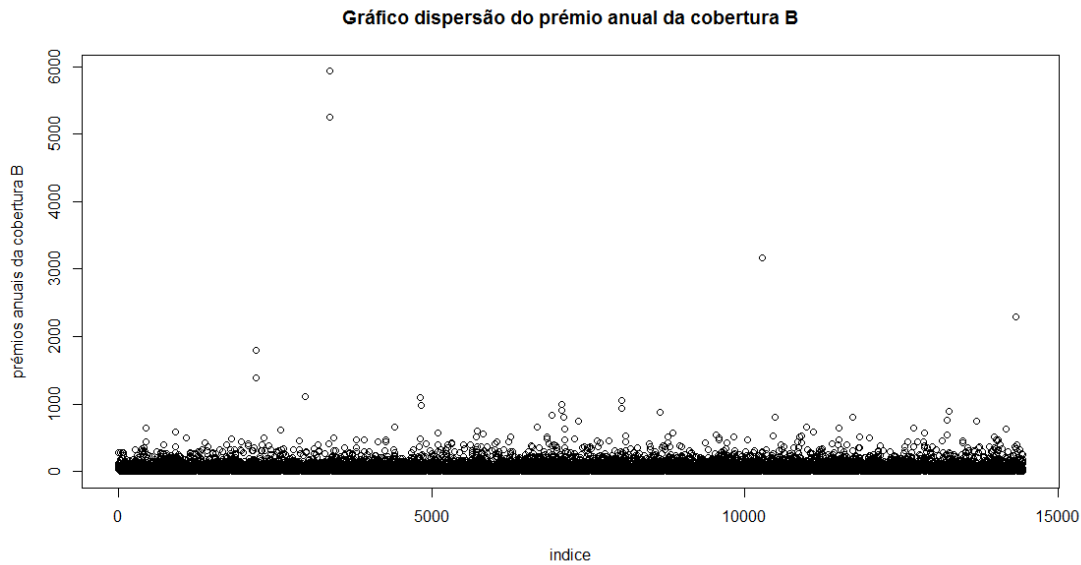


Figura 4.6: Gráficos de dispersão dos prémios anuais da cobertura B

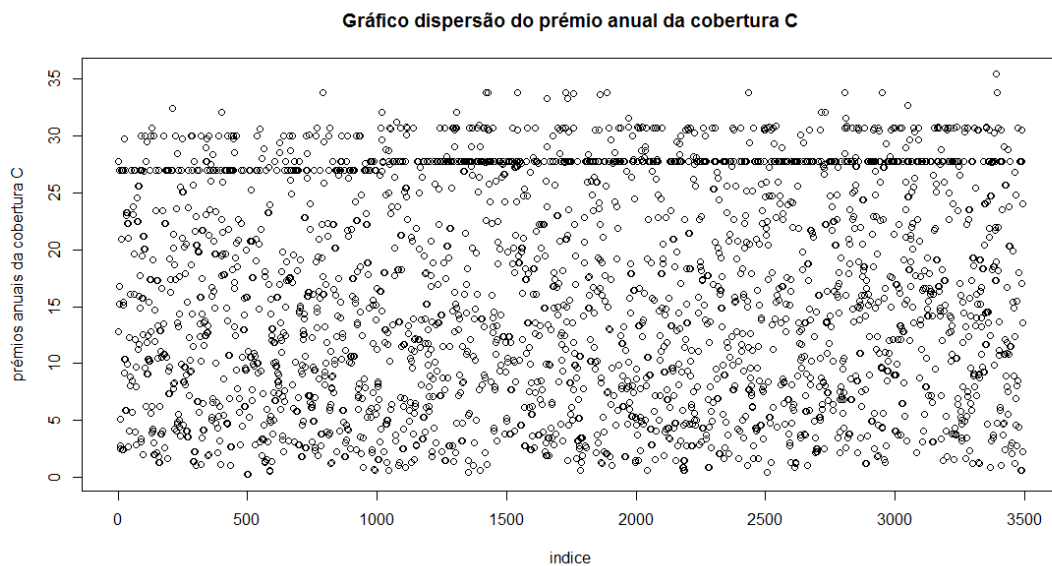


Figura 4.7: Gráficos de dispersão dos prémios anuais da cobertura C

Na figura 4.6 e 4.7 são analisados os valores não nulos dos prémios das coberturas B e C. A ideia é fazer uma divisão que se adapte á distribuição dos prémios e que seja generalizada a todos os prémios. Como se pode ver nos dois gráficos das figuras acima referidas, não parece existir uma maneira geral para fazer esta divisão, pois tem-se valores demasiado diferentes por-

que são coberturas com características diferentes.

Assim, optou-se por usar uma abordagem diferente usando a técnica de agrupamento chamada ***K-médias*** que usa o algoritmo descrito a seguir e descrito na figura 4.8:

1. Escolhe-se aleatoriamente K pontos iniciais, chamado centros iniciais dos K grupos;
2. As observações são atribuídas ao grupo cujo centro é mais próximo;
3. Os centros dos respetivos grupos são recalculados para os centros dos grupos com as novas observações;
4. Repete-se os passos 2 e 3 até a iteração em que nenhuma observação é alterada.

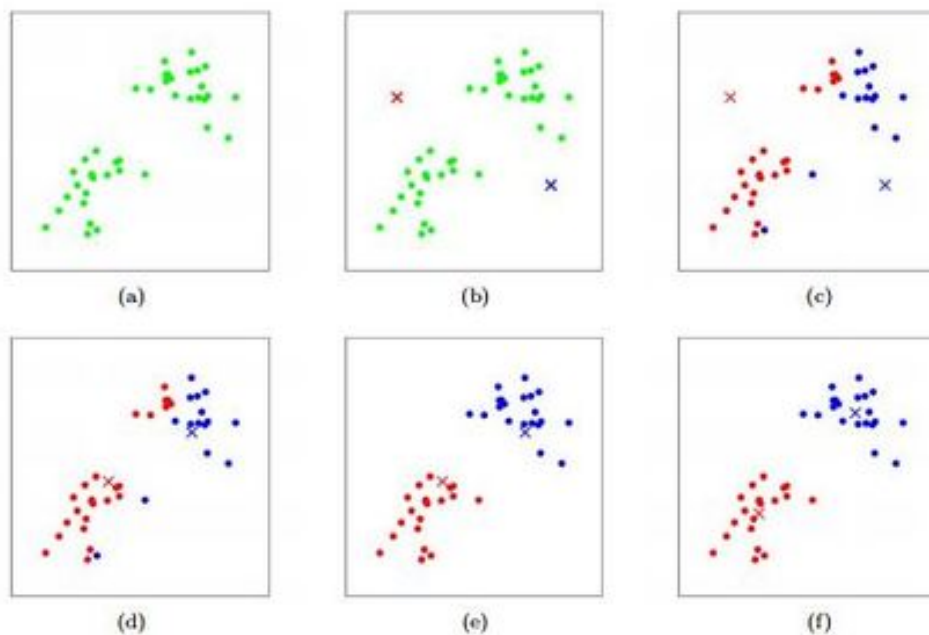


Figura 4.8: Exemplo ilustrativo K-médias

Na figura 4.8 estão descritos os passos acima, em que a) representa as observações a serem agrupadas, em b) estão descritos os $K=2$ centros iniciais (o vermelho e o azul), em c) representa o passo 2, em d) representa o passo 3 e as imagens e) e f) representam os passos 2 e 3 repetidos.

Com esta abordagem surge outra questão que é perceber que valor de K se deve usar, para tal usou-se a função *NbClust* que combina vários métodos que determinam o melhor valor para

K e assim tem-se uma noção mais próxima do melhor valor, apesar de que este valor pode não ser valor perfeito.

Os resultados deste agrupamento estão apresentados na Tabela 4.2, sendo por exemplo uma das categorias do prémio anual da cobertura B é "0.28 até 70.52" correspondente á primeira linha da tabela.

Variável	Grupo	Mínimo	Máximo
Prémio anual da cobertura B	1	0.28	70.52
	2	70.56	184.20
	3	184.68	467.86
	4	470.88	2291.64
	5	3169.92	5927.54
Prémio anual da cobertura C	1	0.24	7.21
	2	7.34	19.94
	3	20.04	35.40

Tabela 4.2: Agrupamento usando K-médias das variáveis relativas aos prémios anuais das coberturas B e C.

4.4.3 Tratamento inicial de valores atípicos (Outliers)

Definição 3 *Um valor atípico (Outlier) é uma observação com um valor demasiado extremo, fora do normal, olhando para o resto das observações.*

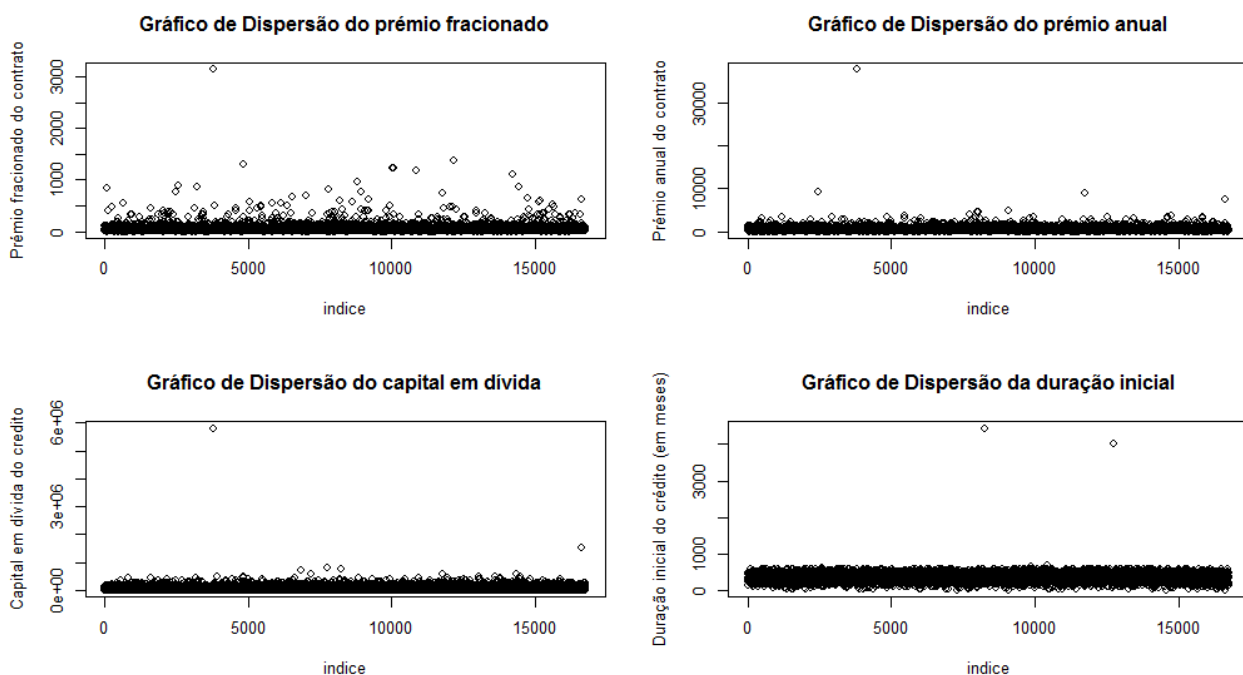


Figura 4.9: Diagramas de dispersão das variáveis prémio fracionado, prémio anual, capital em dívida e duração inicial do crédito

Olhando para a Figura 4.9, onde se reuniu os casos onde parece haver valores extremos, podemos verificar no gráfico de dispersão da duração inicial em meses dois valores absurdos : um contrato com 4008 meses (correspondente a 334 anos) outro com 4444 meses (correspondente a 370,33 anos). Estes dois valores são claramente erros de produção e o seu registo foi eliminado. Também se verifica alguns valores demasiado elevados no prémio fracionado, anual e capital em dívida, o que é natural como já foi explicado anteriormente. No entanto não serão eliminados da análise pois existem modelos que são capazes de aprender com este tipo de observações, assim se for necessário voltar a esta fase para melhor o modelo será feito.

4.4.4 Redução do número de categorias

De seguida verificou-se que existiam variáveis com categorias pouco representativas, isto é, categorias com poucos elementos contidos. Assim, optou-se por se juntar as categorias pouco significativas a fim de se ter todas as categorias representativas. Daqui surge a questão do que são classes representativas e qual será a percentagem adequada a aplicar. De [33] sugere-se juntar classes com menos de 5% de representação, e começou-se por seguir esta sugestão. No entanto , por exemplo, no caso das profissões onde existe um vasto número de diferentes categorias em que a maior parte delas com percentagens muito baixas (a classe mais representativa tem 5.48% de representação) optou-se por alterar o limiar de 5%, apenas para estes casos. Assim, juntaram-se as categorias com representação inferior a 1% para uma nova categoria chamada "Outras".

Relativamente às variáveis referentes a prémios (variáveis numéricas), o critério foi o seguinte: usou-se o valor limite de 5% e em vez de se criar a categoria "Outras" alargou-se o intervalo definido em 4.4.2.

Por fim eliminaram-se as variáveis que continham apenas uma categoria, uma vez que não eram informativas, como por exemplo a cobertura M, como é obrigatória o valor será sempre 1 (cobertura seleccionada).

4.4.5 Pré-processamento SVM

Como modelos diferentes exigem pré-processamentos diferentes, para a aplicação do modelo SVM a base de dados tem de ter todas as variáveis numéricas, pelo simples facto de se pretender encontrar o hiperplano que maximiza a distância da margem entre as duas classes. Esta transformação apenas é necessária para a metodologia SVM, sendo que os restantes modelos são aplicados à base de dados resultante da Secção 4.4.4.

O primeiro passo será então a passagem das variáveis categóricas para numéricas. Começou-se por distinguir variáveis categóricas ordinais e nominais, em que as primeiras foram transformadas segundo a sua ordem, isto é, na variável responsável pelo estado do contrato a transformação fica "Submetida"=1 e "Apólice"=2 (pois, primeiro o contrato passa pelo estado de proposta submetida e depois é que passa para o estado de apólice) e para as segundas foi aplicado a técnica **One-Hot-Encode** em que cada categoria é desmultiplicada em k variáveis binárias, sendo k o número de categorias dessa variável, por exemplo na variável relativa ao género da pessoa se for feminino então na variável relativa á classe feminino terá valor 1 e na variável relativa á classe masculina tem valor 0.

De seguida normalizou-se as variáveis, a fim de haver uma melhor comparação entre as variáveis escrevendo-as todas na mesma escala de 0 a 1. O método de **normalização** usado foi do **Min-Max** dado por, sendo x cada ponto da base de dados e $\tilde{x}$ o vetor onde o ponto x está:

$$x = \frac{x - \min(\tilde{x})}{\max(\tilde{x}) - \min(\tilde{x})} \quad (4.1)$$

Depois destas transformações a base de dados cresceu bastante tendo passado de 30 variáveis para 246 variáveis, sofrendo agora do problema da "maldição da dimensão" o que faz com que piore o tempo computacional e o desenvolvimento do modelo. É de notar que esta transformação na base de dados é feita depois da fase de redução de categorias das variáveis categóricas. Caso esta fase não acontecesse a base de dados teria crescido ainda mais necessariamente.

Desta maneira, houve a necessidade de se aplicar técnicas de redução de dimensão para lidar

com este problema. A técnica usada foi **Análise de Componentes Principais (PCA)** e começou por se eliminar a variável a prever pois esta técnica é de análise não supervisionada. Seguidamente dividiu-se a base de dados em 30% para teste e 70% para treino e aplicou-se a técnica à base de dados de treino. Posteriormente à aplicação da técnica é necessário perceber quais as componentes que explicam uma variância considerável, não perdendo muita informação. Analisaram-se assim os gráficos de proporção e proporção cumulativa da variância explicada mostrada na Figura 4.10.

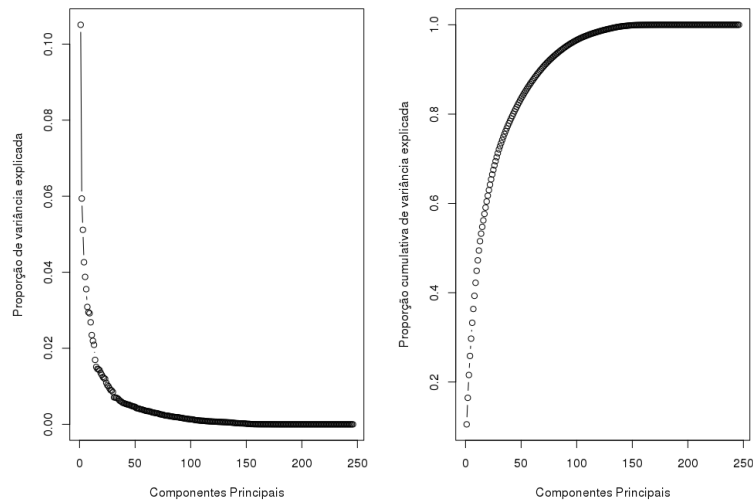


Figura 4.10: Gráficos da proporção da variância explicada á esquerda e gráfico da proporção cumulativa explicada á direita, ambos relativos ao problema da previsão da cobertura A.

Visualizando os gráficos da Figura 4.10 pode concluir-se que a partir de 50 componentes principais as curvas começam a estabilizar, isto é, os valores de variância explicada começam a ficar aproximadamente constantes podendo eliminar componentes porque não estão a explicar mais informação. Assim, escolhendo exatamente 69 componentes estas vão explicar 90.33% da variância, perdendo-se 10% da informação mas assim consegue-se passar de 246 variáveis para 69 componentes principais eliminando-se 177 variáveis.

Note-se que o processo acima foi feito separadamente para o problema da previsão da cobertura A e para o problema da previsão da cobertura C, de uma forma análoga.

4.4.6 Aplicação de técnicas de balanceamento de classes

Depois da fase de pré-processamento verificou-se que não existe o mesmo número de pessoas a comprar e a não comprar as coberturas A e C (pode ser verificado na Figura 4.11), o que faz sentido em termos de negócio. Assim no caso da previsão da compra de A, observa-se que cerca de 86% das pessoas não compram a cobertura A e apenas 14% compram e no caso da previsão da compra de C existem 68% de clientes que optam por não comprar e os restantes 32% optam por comprar. Analisando estes factos percebe-se que as classes são desequilibradas, não tendo a mesma percentagem nas duas classes, o que é um problema pois os modelos assumem que se tem a mesma probabilidade de comprar e não comprar, o que também pode ser observado na Figura 4.11, onde estão representadas as duas classes (a seleção e a não seleção da cobertura correspondendo à compra e não compra desta).

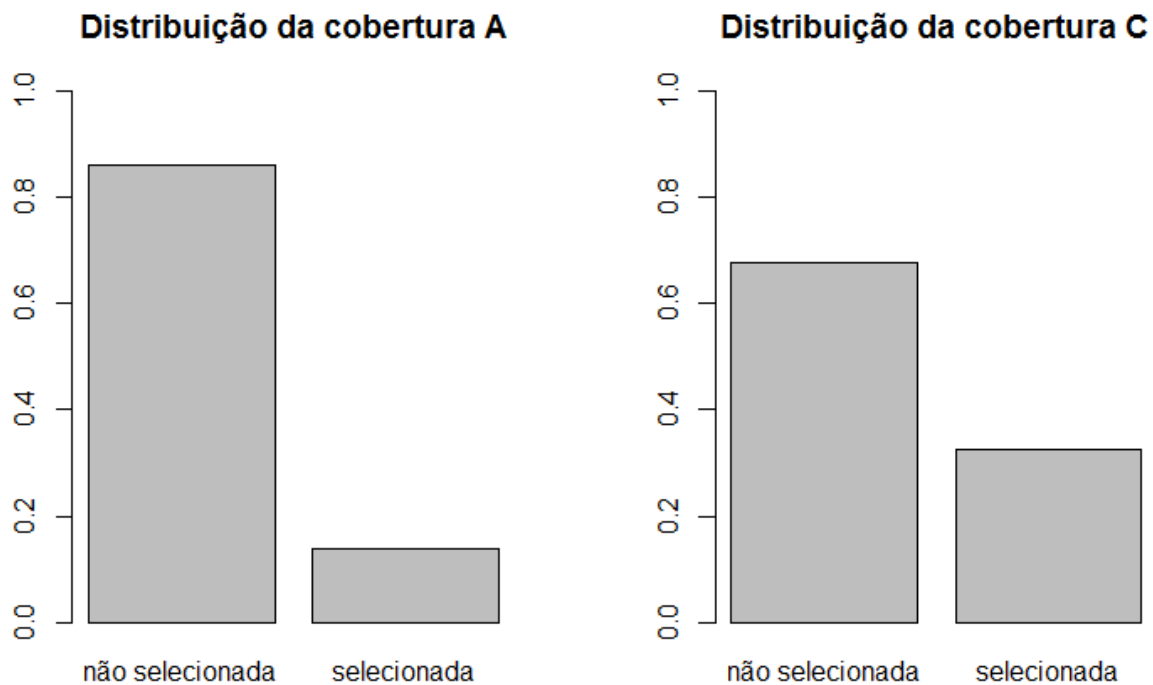


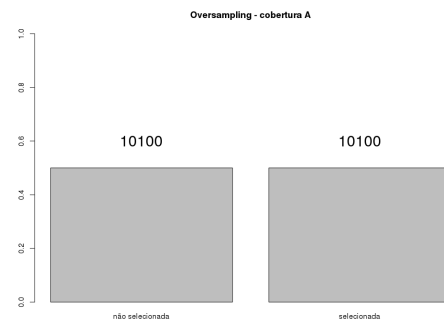
Figura 4.11: Distribuições das coberturas A e C

Com o objetivo de se resolver o problema das classes desequilibradas foram aplicadas as técnicas de balanceamento de classes abordadas na Secção 3.2.2. Nas Figuras 4.12 e 4.13 são

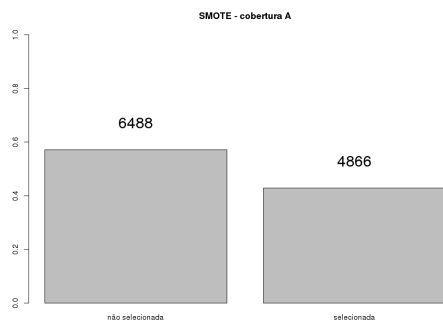
demonstradas as dimensões resultantes da aplicação destas técnicas, tanto na cobertura A como na C. Note-se que a técnica que envolve maior quantidade de dados é a técnica de *Oversampling* ficando com na totalidade 20200 observações, sugerindo que seja a técnica que leve maior tempo de execução.



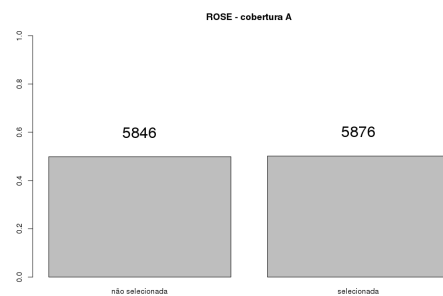
(a) Oversampling aplicado a cobertura A



(b) Downsampling aplicado a cobertura A

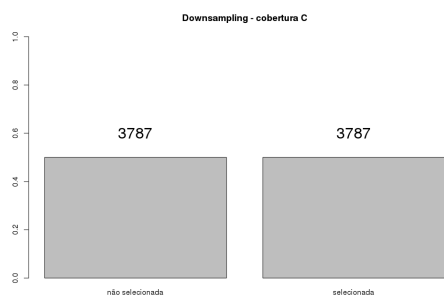


(c) SMOTE aplicado a cobertura A

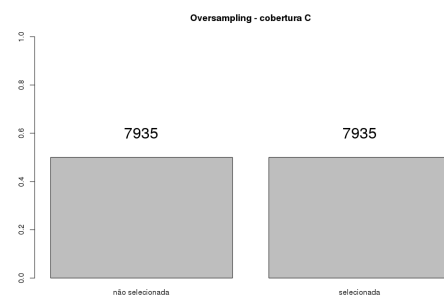


(d) ROSE aplicado a cobertura A

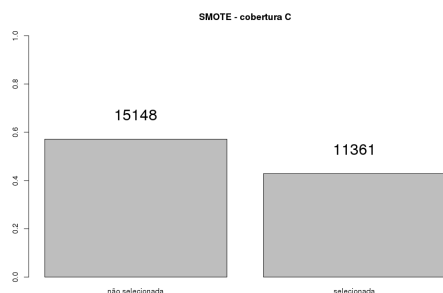
Figura 4.12: Técnicas de balanceamento aplicadas a cobertura A



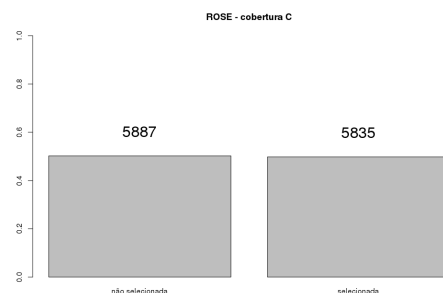
(a) Oversampling aplicado a cobertura C



(b) Downsampling aplicado a cobertura C



(c) SMOTE aplicado a cobertura C



(d) ROSE aplicado a cobertura C

Figura 4.13: Técnicas de balanceamento aplicadas a cobertura C

Depois da aplicação destas técnicas passou-se para a aplicação dos modelos abordados na Secção 3.1.

4.5 Aplicação de modelos preditivos

Existem dois problemas distintos: a previsão da cobertura A e a previsão da cobertura C. Estes dois problemas são tratados em separado mas aos dois problemas aplicou-se os seguintes passos:

1. Eliminação das variáveis relativamente a coberturas que não sejam a que se quer prever, ou seja, é mantida em análise apenas a informação da cobertura que se quer prever;
2. Divisão da base de dados em 70% para os dados de treino e 30% para os dados de teste;
3. Determinação dos melhores parâmetros utilizados nos modelos usando **Validação Cruzada**;
4. Aplicação do modelo com os parâmetros otimizados

No passo 3 usa-se a abordagem Validação Cruzada (*k-fold cross validation*), em que se começa por dividir a base de dados em k blocos. Depois usa-se um bloco para testar o modelo e os restantes $k-1$ blocos para treinar o modelo. Este processo repete-se k vezes usando diferentes blocos para fazer o teste e obtém-se por fim k modelos. Por fim calcula-se a média da *performance* obtida pelos modelos nos blocos de teste e esta é considerada a verdadeira *performance* do algoritmo [5]. Um esquema exemplificativo desta abordagem usando $k=5$ blocos é ilustrado na Figura 4.14.

Deste modo, usa-se a mesma metodologia para estimar o melhor parâmetro do modelo, em que para cada valor possível para o parâmetro calcula-se o erro usando esse parâmetro para cada k . Para cada parâmetro calcula-se a média do erro. Por fim, o parâmetro escolhido é o que produz o menor erro.



Figura 4.14: Validação Cruzada com $k=5$ [5]

Ainda sobre os passos descritos no início da secção, é de notar que os passos 3. e 4. são aplicados a bases de dados diferentes, ou seja, aplica-se o modelo às 4 técnicas de balanceamento diferentes e à base de dados original tendo no total 5 resultados diferentes para cada modelo. Também é de notar que a base de dados de treino serve para aplicação/treinamento dos modelos e a base de dados de teste serve para se medir se os resultados são bons ou não.

Inicialmente a aplicação dos modelos no passo da determinação do melhor parâmetro (passo 4.) revelou ser muito demorada computacionalmente (dias de processamento), o que levantou a necessidade de se usar **processamento paralelo**, isto é, paralisam-se processos independentes em que, por exemplo, cada processo é processado num core diferente (*core* representa núcleo de processamento). Usando este processo, na validação cruzada cada processamento na sub-amostragem pode ser feita usando cores diferentes, otimizando assim o processo dividindo o "trabalho" por *cores* diferentes. Para o uso do processamento paralelo foi usada a ferramenta *RStudio Server* que usa um sistema operativo em Linux. É de notar que os pacotes existentes no R para processamento paralelo apenas existem em Linux levantando assim a necessidade de se usar uma ferramenta diferente.

Entrou-se assim na fase de aplicação de modelos usando processamento paralelo. Começou-se por fazer um levantamento de quais os algoritmos mais indicados para a previsão. Como o contexto é um contexto empresarial existe uma forte necessidade de os resultados terem

de ser explicados e interpretados. Deste modo percebeu-se que pela sua alta capacidade de interpretação as Árvore de Decisão (descrito teoricamente na Sub-secção 3.1.1) seriam um bom ponto de partida [25]. É de notar que não foi escolhido um modelo linear devido às características da base de dados, nomeadamente do número de variáveis explicativas ser elevado, e devido à complexidade do problema em questão claramente não ser resolvido com um modelo simples, linear.

Depois de escolhido as Árvore de Decisão como um ponto de partida aplicou-se um modelo mais sofisticado que usa esta metodologia, designado *Random Forest* (descrito teoricamente na Sub-secção 3.1.2). É de notar que querendo aumentar a sofisticação dos resultados e a sua *accuracy* perde-se a interpretabilidade referida acima.

Testou-se ainda o algoritmo SVM (descrito teoricamente na Sub-secção 3.1.3) com vários núcleos diferentes, devido a ser provavelmente o algoritmo mais popular em ML para análise supervisionada e conhecido pela sua robusteza, grande capacidade de generalização e por dar soluções ótimas globais e únicas. [4]

Por fim, depois de aplicados os três algoritmos referidos acima, comparou-se os resultados na Secção 4.6.

4.6 Comparação de modelos

Em problemas de classificação são muito usadas as **matrizes de confusão** (*Confusion Matrix*), que são tabelas que permitem visualizar a *performance* de um modelo. Cada linha representa os instantes nas classes previstas e cada coluna representa os instantes na classe atual (exemplos no conjunto de teste), como se pode visualizar na Figura 4.15, no caso do problema em que se tem 2 classes (uma denominada positiva e outra negativa). Nesta são representados 4 valores:

- TP : número de casos previstos corretamente como positivos ;

- TN : número de casos previstos corretamente como negativos;
- FP : número de casos previstos como positivos mas na verdade são negativos (são previstos incorretamente, normalmente chamado Erro do Tipo I);
- FN : número de casos previstos como negativos mas na verdade são positivos (são previstos incorretamente, normalmente chamado Erro do Tipo II);

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Figura 4.15: Matriz de confusão

Define-se também dois tipos de classes: **classe maioritária** e **classe minoritária**. A primeira corresponde à classe com mais elementos, que neste caso é a classe de não compra, e a segunda corresponde à classe com menos elementos que corresponde à classe de compra, respetivamente. O objetivo desta fase é determinar qual o melhor modelo, comparando a *performance* dos vários modelos no conjunto de teste. Para esta comparação é necessário definir uma métrica, ou seja, uma medida de comparação dos diferentes modelos.

Tradicionalmente, as métricas mais populares e frequentes são as seguintes:

- **Accuracy**: mede a percentagem na qual todas as observações são corretamente classificadas, definida pela equação (4.2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

- **Taxa de Erro (*Error Rate*)**: mede a percentagem na qual todas as observações são

incorretamente classificadas, definida pela equação (4.3)

$$TaxadeErro = 1 - Accuracy \quad (4.3)$$

No entanto, estas métricas não são muito úteis para o problema das classes desequilibradas, pois, por exemplo, suponhamos que existe 5% de exemplos na classe positiva e os restantes 95% encontram-se na classe negativa, assim neste pequeno exemplo teríamos cerca de 95% de *accuracy*, o que era excelente, no entanto, pouco realista porque a classe maioritária está a enviesar os resultados e a não conseguir identificar a classe minoritária. Este fenómeno designa-se *Accuracy Paradox* [30].

Assim, houve a necessidade de se procurar métricas mais adequadas para este tipo de problema e que consigam identificar a classe minoritária. As métricas mais conhecidas são [19]:

- **Precision**: proporção de identificações positivas que foram atualmente corretas, normalmente designada de medida de exatidão, dada pela equação (4.4).

$$Precision = \frac{TP}{TP + FP} \quad (4.4)$$

- **Recall**: proporção de atual positivos que foram identificados corretamente, normalmente designada medida de completude, dada pela equação (4.5).

$$Recall = \frac{TP}{TP + FN} \quad (4.5)$$

- **F1**: média harmónica entre Precision e Recall, dada pela equação (4.6).

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (4.6)$$

- **Curvas ROC/AUC**: a curva ROC corresponde a representação gráfica de TPR contra FPR, dados pelas equações (4.7) e (4.8). Usando esta métrica, pode-se calcular a área debaixo da curva (AUC) e comparando as diferentes áreas dos diferentes modelos dizer

qual o melhor modelo.

$$TPR = recall \quad (4.7)$$

$$FPR = \frac{FP}{FP + TP} \quad (4.8)$$

Como identificado na Figura 4.16, apenas pela visualização das curvas consegue-se perceber qual o melhor modelo. Assim, o melhor modelo é o modelo cuja curva ROC esteja mais próxima do canto superior esquerdo, em que consequentemente a área debaixo desta curva terá maior valor. Neste gráfico existe uma reta a tracejado que representa a linha de referência, isto é, cada ponto que se encontra na diagonal representa que o modelo forneceu uma previsão aleatória da classe, geralmente falando, corresponde a um classificador aleatório. [19]

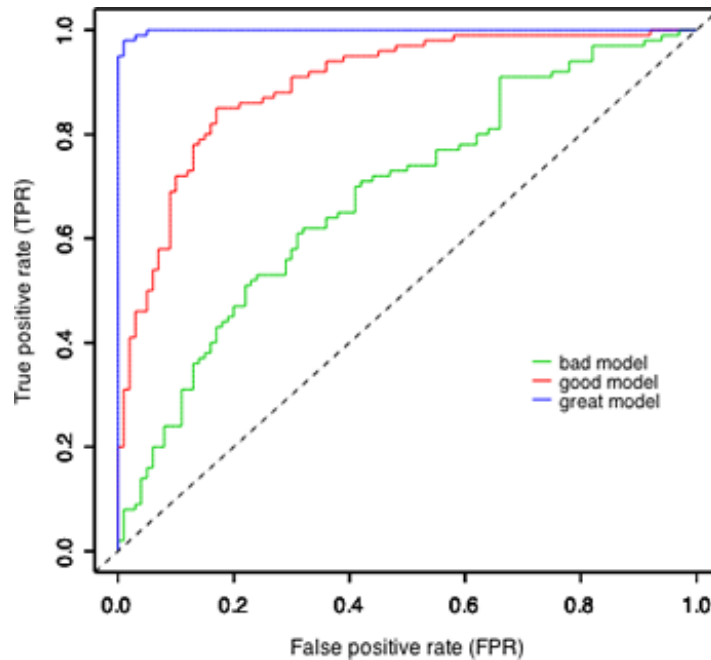


Figura 4.16: Exemplo Curvas ROC

No caso em questão, em que se verifica a existência de classes desequilibradas, não existe uma métrica que seja mais adequada a usar que outra. No entanto, como introduzido existem 4 métricas muito usadas para este contexto e cada métrica consegue identificar vantagens diferentes, consoante o objetivo do problema.

As métricas *precision* e *recall* são normalmente usadas em conjunto, pois apenas diferem numa parcela do denominador. Um bom modelo será o modelo que apresenta valores elevados nestas duas métricas. No entanto, o que normalmente acontece é que quando uma métrica é alta a outra é baixa, havendo a necessidade de balançar entre estas duas métricas, quando escolhidas.

Nas Tabelas 4.3 e 4.4 analisa-se os diferentes resultados com as diferentes medidas apresentadas anteriormente. Para cada modelo e técnica calculou-se as medidas de avaliação descritas anteriormente e adicionou-se também os tempos de execução, em que o Tempo 1 corresponde ao tempo computacional que levou a encontrar os melhores parâmetros e o Tempo 2 corresponde ao tempo computacional da aplicação do modelo com os parâmetros otimizados. Estes tempos estão em segundos e se não houver valor, isto é aparece '—' na tabela, representa que o tempo de execução foi instantâneo.

Modelos e técnicas	<i>precision</i>	<i>recall</i>	<i>F1</i>	<i>AUC</i>	Tempo 1	Tempo 2
AD	0.782	0.141	0.239	0.567	7.784	-
AD + DS	0.173	0.943	0.293	0.614	2.393	-
AD + OS	0.230	0.631	0.337	0.648	10.629	-
AD + SMOTE	0.215	0.599	0.317	0.626	3.333	-
AD + ROSE	0.335	0.440	0.381	0.651	4.236	-
RF	0.854	0.638	0.730	0.811	6088.975	52.040
RF + DS	0.279	0.917	0.428	0.771	413.192	4.592
RF + OS	0.880	0.714	0.788	0.849	6911.523	680.793
RF + SMOTE	0.513	0.711	0.596	0.803	200.885	228.964
RF + ROSE	0.580	0.747	0.653	0.831	315.714	-
SVM Linear	0.725	0.145	0.242	0.568	569.506	433.404
SVM Linear + DS	0.236	0.751	0.359	0.683	13.197	13.197
SVM Linear + OS	0.244	0.754	0.369	0.692	2257.172	861.613
SVM Linear + SMOTE	0.281	0.573	0.377	0.670	461.024	153.528
SVM Linear + ROSE	0.670	0.249	0.363	0.674	1178.747	-
SVM Radial	0.805	0.217	0.341	0.604	329.811	49.827
SVM Radial + DS	0.807	0.310	0.448	0.761	24.224	5.069
SVM Radial + OS	0.721	0.523	0.606	0.808	831.079	230.225
SVM Radial + SMOTE	0.495	0.631	0.555	0.764	289.293	52.269
SVM Radial + ROSE	0.353	0.534	0.425	0.652	408.306	-
SVM Polinomial	0.174	0.783	0.285	0.583	3123.107	45.840
SVM Polinomial + DS	0.452	0.573	0.505	0.731	147.084	4.729
SVM Polinomial + OS	0.529	0.651	0.584	0.780	11498.677	192.059
SVM Polinomial + SMOTE	0.571	0.503	0.535	0.721	2568.489	60.318
SVM Polinomial + ROSE	0.557	0.517	0.536	0.737	6873.564	-

Tabela 4.3: Comparação de modelos e técnicas com as respetivas medidas de comparação, do problema de previsão da cobertura A

Analisando a Tabela 4.3 verifica-se em todos os modelos o que foi constatado acima relativamente a quando se tem valores altos de *precision* tem-se valores baixos de *recall* e vice versa. Relativamente a estas duas métricas os valores máximos de *precision* é 0.880 para RF com

oversampling e de *recall* é 0.943 para AD com *downsampling*. No entanto olhando para a sua média harmónica (F1) o valor máximo é 0.788 para RF com *oversampling*. Analisando também os valores de AUC verifica-se o máximo de 0.849 para RF com *oversampling*.

Deste modo, olhando para estas 4 métricas o modelo ganhador é RF com *oversampling* para a previsão da compra da cobertura A. No entanto, olhando para os tempos computacionais demora cerca de 2 horas a processar.

Modelos e técnicas	<i>precision</i>	<i>recall</i>	<i>F1</i>	<i>AUC</i>	Tempo 1	Tempo 2
AD	0.626	0.328	0.430	0.616	4.782	1.022
AD + DS	0.539	0.495	0.516	0.644	2.576	0.555
AD + OS	0.574	0.419	0.485	0.634	4.100	1.219
AD + SMOTE	0.511	0.438	0.472	0.617	6.325	1.405
AD + ROSE	0.605	0.387	0.472	0.632	3.601	-
RF	0.869	0.734	0.796	0.840	9533.211	59.056
RF + DS	0.639	0.862	0.734	0.814	1983.590	21.398
RF + OS	0.864	0.785	0.823	0.862	6667.725	411.431
RF + SMOTE	0.798	0.796	0.797	0.849	1061.740	1352.074
RF + ROSE	0.692	0.748	0.719	0.793	304.279	-
SVM Linear	NA	0	NA	0.500	475.570	225.676
SVM Linear + DS	0.434	0.617	0.509	0.613	186.248	75.170
SVM Linear + OS	0.434	0.627	0.513	0.615	1388.968	581.283
SVM Linear + SMOTE	0.488	0.515	0.501	0.626	3963.403	1573.097
SVM Linear + ROSE	0.619	0.432	0.509	0.611	1355.576	-
SVM Radial	0.755	0.330	0.459	0.639	374.347	101.109
SVM Radial + DS	0.681	0.526	0.594	0.691	160.742	30.264
SVM Radial + OS	0.665	0.592	0.627	0.721	655.985	183.439
SVM Radial + SMOTE	0.649	0.608	0.628	0.724	1863.866	431.383
SVM Radial + ROSE	0.509	0.491	0.500	0.626	373.069	-
SVM Polinomial	0.117	0.801	0.205	0.552	7646.858	56.906
SVM Polinomial + DS	0.514	0.554	0.533	0.649	1634.121	23.372
SVM Polinomial + OS	0.565	0.603	0.583	0.689	12185.60	191.600
SVM Polinomial + SMOTE	0.651	0.726	0.687	0.769	19520.847	426.180
SVM Polinomial + ROSE	0.565	0.510	0.536	0.651	8054.855	-

Tabela 4.4: Comparação de modelos e técnicas com as respetivas medidas de comparação, do problema de previsão da cobertura C

Analisando os resultados da previsão da compra da cobertura C, mostrados na Tabela 4.4, nota-se que os valores máximos para cada medida são 0.864 (RF com *oversampling*), 0.862 (RF com *downsampling*), 0.823 (RF com *oversampling*) e 0.862 (RF com *oversampling*) para as medidas *precision*, *recall*, F1 e AUC respetivamente. Deste modo o modelo e técnica que se obteve melhores resultados foi *Random Forest* com a técnica de balanceamento *Random Oversampling*. É de notar que SVM Linear preveu a classe "não selecionada" para todas as observações, estando claramente uma previsão enviesada. Também se observa que o modelo RF com qualquer uma das técnicas consegue dar muito melhores resultados em praticamente todas as métricas.

4.7 Previsão da compra de cobertura em novos exemplos

No final da Secção 4.6 tem-se dois modelos finais, um para prever a compra da cobertura A e outro para C. A próxima fase é a previsão da compra das duas coberturas em novos exemplos, ou seja, usando nova informação.

Na prática a ideia é carregar-se o modelo final que foi anteriormente treinado e guardado, receber-se a informação em tempo real e efetuar a previsão acerca da comprar da cobertura em questão [8]. No entanto, esta fase envolve mais trabalho do que apenas os passos descritos anteriormente devido ao facto da informação recebida em tempo real não estar no mesmo formato da informação com que o modelo aprendeu. Assim, a nova informação tem de passar pelos passos todos referentes ao pré-processamento descrito na Secção 4.4. Deste modo foi criada uma tabela de conversão com as variáveis resultantes da Secção 4.4 e as categorias que foram transformadas em "Outras" (da fase descrita na Secção 4.4.4). Usando esta tabela transforma-se a informação do novo exemplo para o formato adequado.

Por fim, tendo o novo exemplo transformado efetua-se a previsão da compra da coberturas coberturas, através da função `predict()` do R.

Capítulo 5.

Conclusão e análise dos resultados

"Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do."

Pele

Como referido na Secção 4.6, tendo um modelo final e sendo este Floresta Aleatória (*Random Forest*) pode-se analisar as variáveis mais importantes para esse modelo.

Começando por analisar as variáveis mais importantes para a previsão da compra da cobertura A, analisa-se a Figura 5.1. Observa-se que usando as duas medidas, a variável mais importante é "Agencia - Banco", pois é a variável com valores mais elevados, evidenciando que para um cliente é bastante importante a agência bancária e o seu banco para contrair um empréstimo bancário com a cobertura A. Pode-se assim concluir que a escolha da agência e do banco, para o cliente, é mais importante do que, por exemplo, as características da cobertura em questão. Também se verifica que a profissão da pessoa segura é um fator importante, usando as duas medidas.

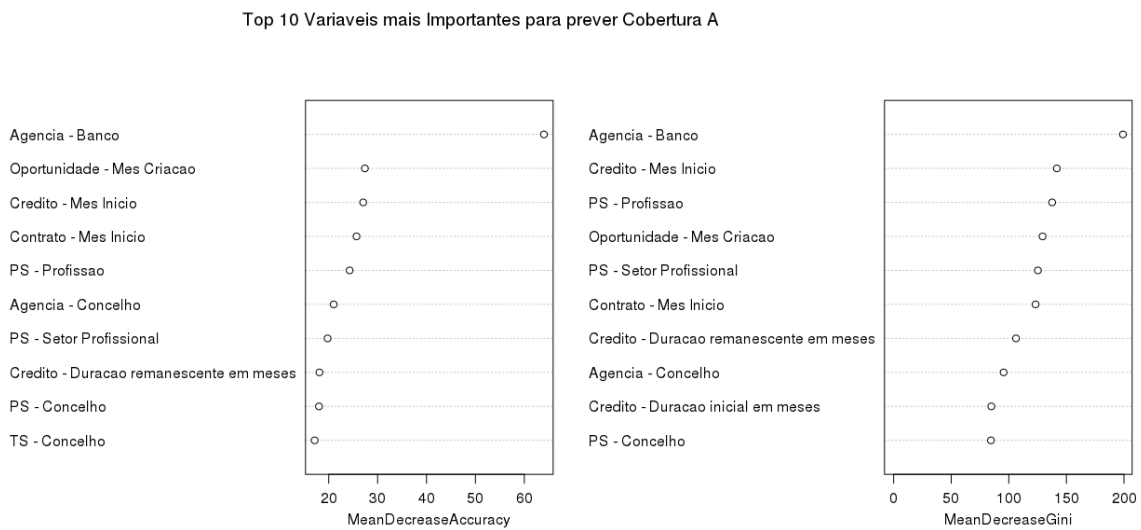


Figura 5.1: Variáveis mais importantes para prever a cobertura A

Olhando para a Figura 5.2 analisa-se as variáveis mais importantes para prever a compra da cobertura C. Nota-se que o mês de criação da oportunidade, usando as duas medidas, é a variável mais importante para a compra da cobertura em questão, contrariamente ao caso da compra da cobertura A em que a variável mais importante foi a agência e banco. Também se verifica que tanto o mês de início do contrato como o mês de início do crédito são bastante importante para a decisão de compra da cobertura C. Ressalta-se também a variável profissão da pessoa segura que se revela importante na compra da cobertura C, como também verificado para a compra da cobertura A.

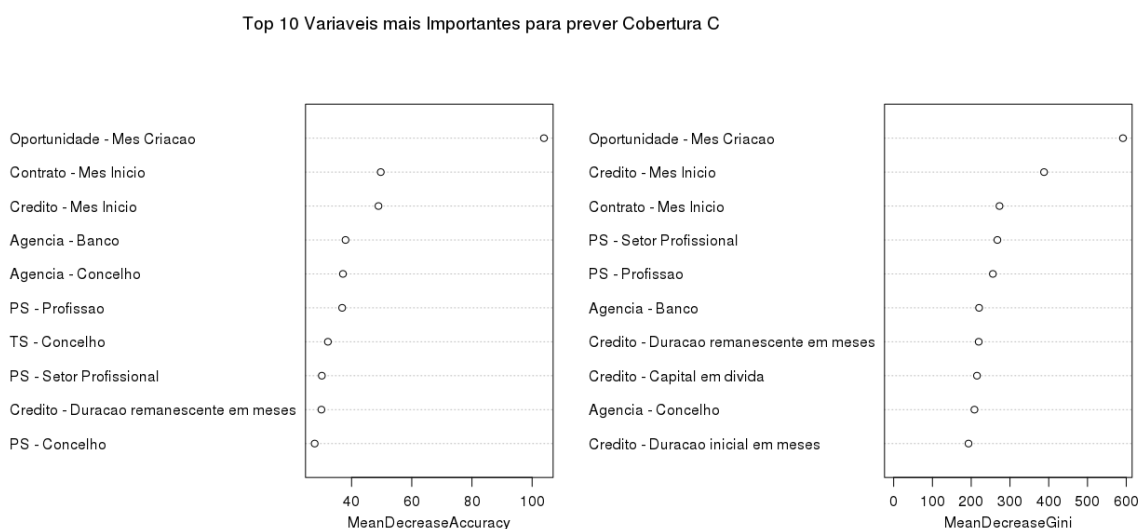


Figura 5.2: Variáveis mais importantes para prever a cobertura C

Em trabalhos futuros, é necessário voltar a treinar os modelos com mais dados e perceber

se continuam a ser estes dois modelos os que dão melhores resultados. Também é necessário averiguar de quanto em quanto tempo ou com quantos dados a mais é necessário voltar a treinar os modelos. Adicionalmente, implementação de novos modelos como por exemplo XGBoost a fim de se perceber se se consegue melhorar os resultados obtidos.

Bibliografia

- [1] Ageas - Glossário de Seguros. URL: <https://www.ageas.pt/glossario-de-seguros#>.
- [2] ASF - Glossário. URL: <https://www.asf.com.pt/NR/exeres/BF408252-BF68-4555-8AD1-34E31950AE11.htm>.
- [3] ASF - SEGUROS DO RAMO VIDA. URL: https://www.asf.com.pt/NR/rdonlyres/AF85F6E6-10B4-4368-A713-399F4929A009/0/ASF_Brochura_Vida_2015.pdf.
- [4] Mariette Awad e Rahul Khanna. «Support vector machines for classification». Em: *Efficient Learning Machines*. Springer, 2015, pp. 39–66.
- [5] Parth Bhavsar et al. «Machine learning in transportation data analytics». Em: *Data Analytics for Intelligent Transportation Systems*. Elsevier, 2017, pp. 283–307.
- [6] Bernhard E Boser, Isabelle M Guyon e Vladimir N Vapnik. «A training algorithm for optimal margin classifiers». Em: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152.
- [7] Leo Breiman. *Classification and regression trees*. Routledge, 2017.
- [8] Jason Brownlee. *Save And Finalize Your Machine Learning Model in R*. February 19, 2016. URL: <https://machinelearningmastery.com/finalize-machine-learning-models-in-r/>.
- [9] Christopher JC Burges. «A tutorial on support vector machines for pattern recognition». Em: *Data mining and knowledge discovery 2.2* (1998), pp. 121–167.
- [10] Pete Chapman et al. «The CRISP-DM user guide». Em: *4th CRISP-DM SIG Workshop in Brussels in March*. Vol. 1999. 1999.

- [11] Nitesh V Chawla et al. «SMOTE: synthetic minority over-sampling technique». Em: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.
- [12] Mashrur Chowdhury, Amy Apon e Kakan Dey. *Data analytics for intelligent transportation systems*. Elsevier, 2017.
- [13] Joaquim Costa DMA-FCUP. *Arvores de Decisão e de Regressão*. Abril 24, 2015.
- [14] Joaquim Costa DMA-FCUP. *MAQUINAS DE SUPORTE VECTORIAL*. 2018.
- [15] Alipio Jorge DCC FCUP. *Trees as learning models*. 16 March 2019.
- [16] Nuno Moniz DCC FCUP. *Embalanced Domain Learning*. 2019.
- [17] Jerome Friedman, Trevor Hastie e Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.
- [18] *Genial Seguros - Como funciona a franquia dos seguros*. URL: <https://genialseguros.com.br/artigo/como-funciona-a-franquia-dos-seguros>.
- [19] Haibo He e Eduardo A Garcia. «Learning from imbalanced data». Em: *IEEE Transactions on Knowledge & Data Engineering* 9 (2008), pp. 1263–1284.
- [20] Jake Hoare. *How is Variable Importance Calculated for a Random Forest?* 30 July, 2018. URL: <https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/>.
- [21] Harold Hotelling. «Analysis of a complex of statistical variables into principal components.» Em: *Journal of educational psychology* 24.6 (1933), p. 417.
- [22] Harold Hotelling. «Relations between two sets of variates». Em: *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.
- [23] Han Liu e Mihaela Cocea. «Semi-random partitioning of data into training and test sets in granular computing context». Em: *Granular Computing* 2.4 (2017), pp. 357–386.
- [24] Nicola Lunardon, Giovanna Menardi e Nicola Torelli. «ROSE: A Package for Binary Imbalanced Learning.» Em: *R journal* 6.1 (2014).

- [25] Christoph Molnar. *Interpretable Machine Learning*. Vol. 1st edition. 2018, pag 28 - 29.
- [26] NOSSA VIDA - Glossário. URL: <http://www.nossaseguros.ao/pt/glossario>.
- [27] Karl Pearson. «LIIL. On lines and planes of closest fit to systems of points in space». Em: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.
- [28] *Principal Component Methods in R: Practical Guide*. URL: <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>.
- [29] Arthur L Samuel. «Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress». Em: *Computer Games I*. Springer, 1988, pp. 366–400.
- [30] Tejumade Afonja. Towards Data Science. *Accuracy Paradox*. URL: <https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b>.
- [31] *Sobre a msg life : Visão, Missão e História*. URL: <https://www.msg-life.pt/sobre-msg-life/msg-life/#/>.
- [32] Chakkrit Tantithamthavorn, Ahmed E Hassan e Kenichi Matsumoto. «The impact of class rebalancing techniques on the performance and interpretation of defect prediction models». Em: *IEEE Transactions on Software Engineering* (2018).
- [33] ANALYTICS VIDHYA CONTENT TEAM. *This Machine Learning Project on Imbalanced Data Can Add Value to Your Resume*. URL: <https://www.analyticsvidhya.com/blog/2016/09/this-machine-learning-project-on-imbalanced-data-can-add-value-to-your-resume/>.
- [34] Mayo Foundation Terry M. Therneau Elizabeth J. Atkinson. *An Introduction to Recursive Partitioning - Using the RPART Routines*. April 11, 2019. URL: <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>.

Appendices

Capítulo A.

Glossário 1 : ramo de seguros

Algumas definições relacionadas com a área dos Seguros [2] [26]:

- **Apólice de Seguro:** Documento que titula o contrato celebrado entre o Tomador do Seguro e o Segurador provando a existência de um contrato de seguro. Consiste em Condições Gerais, Especiais, Particulares e Atas Adicionais.
- **Beneficiário:** Pessoa singular ou coletiva definida nas Condições Particulares a favor de quem reverte a prestação do Segurador/Companhia de Seguros decorrente de um contrato de seguro ou de uma operação de capitalização.
- **Capital Seguro:** Montante máximo de responsabilidade do Segurador/Companhia de Seguros em caso de ocorrência de sinistro. Descrito nas condições particulares do contrato.
- **Carteira:** Portefólio de clientes.
- **Cobertura:** Proteção que a seguradora oferece para uma dada situação acordada.
- **Comissão:** Remuneração pela angariação ou gestão de um contrato de Seguro, de Resseguro ou de retrocessão. Dado que é uma remuneração atribuída pela Seguradora a parceiros normalmente externos, é incluída nos prémios a pagar pelo Tomador. É de notar que normalmente é paga em percentagem, mas pode não ser.
- **Contrato de Seguro:** Contrato estabelecido entre duas partes, o Segurador/Companhia de Seguros e a Pessoa Segurada, através do qual o primeiro assume a cobertura de determinados riscos, comprometendo-se a satisfazer as indemnizações ou a pagar o capital

seguro em caso de ocorrência de sinistro, nos termos acordados.

- **Encargo:** Importância acrescentada ao prêmio simples de um seguro que serve para cobrir as despesas de aquisição, gestão e cobrança do Segurador/Companhia de Seguros, cujos limites são previamente fixados e do conhecimento do instituto de seguros. Por exemplo, encargo de gestão, aquisição e cobrança de Apólices.
- **Franquia:** Importância que, caso haja sinistro, fica a cargo da Pessoa Segura e/ou do Tomador de Seguro, definida na Apólice. Caso o custo de reparo dos danos seja inferior ao valor da franquia a Pessoa Segura tem de pagar na totalidade o reparo, não recebendo qualquer indemnização do Segurador/Companhia de Seguros. Caso o custo do reparo seja superior ao valor da franquia a Seguradora tem de pagar uma indemnização, pagando a Pessoa Segura o valor da franquia e o restante paga o Segurador/Companhia de Seguros[18].
- **Operação de Capitalização:** Contratos pelos quais o Segurador/Companhia de Seguros se compromete a pagar um valor previamente fixado, decorrido um certo número de anos, em troca do pagamento de um prêmio único ou periódico. Este valor pode ser determinado em função de um “valor de referência” e é pago ao subscritor ou ao portador do título da operação de capitalização na data do seu vencimento. Esta operação irá ser explorada no Ramo Vida.
- **Pessoa Segura/Objeto Seguro:** Pessoa/Objeto sob o qual recai o interesse da celebração do contrato.
- **Prémio:** Importância paga, incluindo taxas e impostos, que o Tomador de Seguro deve pagar ao Segurador/Companhia de Seguros pelo seguro.
- **Proposta de Seguro:** documento pelo qual uma pessoa singular ou coletiva declara que pretende subscrever um contrato de seguro. Através da sua assinatura, o Tomador de Seguro declara que tomou conhecimento de toda a informação contida na Proposta.

- **Ramo de Seguro:** Conjunto de operações ou atividades referentes a contratos de seguro da mesma natureza.
- **Reembolso:** Devolução, ao Tomador do Seguro, de uma parte do prémio do seguro já pago (ou pago em excesso devido a alterações na apólice).
- **Ressegurador:** Segurador/Companhia de Seguros que cobre parte dos riscos de uma empresa de seguros através de tratados de resseguro.
- **Resseguro:** Seguro realizado pelo Segurador/Companhia de Seguros através do qual se responsabiliza, total ou parcialmente, pelo risco de uma operação já coberta por outro Segurador/Companhia de Seguros.
- **Retrocessão:** Operação pela qual uma empresa de resseguros faz, por sua vez, segurar parte dos riscos que aceitou em resseguro;
- **Segurador/Companhia de Seguros:** Entidade legalmente autorizada a exercer a atividade seguradora e que é parte no contrato de seguro.
- **Sinistro:** Acontecimento ou conjunto de acontecimentos suscetível de acionar as garantias do contrato.
- **Subscritor:** Pessoa que contrata uma operação de capitalização com uma empresa de seguros, sendo responsável pelo pagamento da respetiva prestação.
- **Subscrição:** Termo que representa todo o processo que envolve a tomada de decisão seletiva, de riscos aceitáveis, determinação do prémio a ser cobrado, os termos do contrato e o monitorizamento das decisões tomadas.
- **Tomador do Seguro:** Pessoa singular ou coletiva que, por sua conta ou por conta de uma ou várias pessoas, celebra o contrato de seguro com o Segurador/Companhia de Seguros, sendo responsável pelo pagamento do prémio.

Capítulo B.

Glossário 2 : conceitos básicos

- **Amostras *bootstrap***: são amostras retiradas de uma amostra maior, com o mesmo tamanho, que são repetidamente desenhadas, com repetição, de uma única amostra original.
- **Estatística** : é a ciência que permite tirar conclusões e de uma maneira geral tirar conclusões sobre a população a partir de uma amostra, usando conjuntos de dados.
- **Variável** : é uma característica dos elementos da amostra que é de interesse averiguar estatisticamente. Pode ser de dois tipos:
 - variável qualitativa, representando uma variável que pode ser medida qualitativamente. Estas variáveis são ainda ordinais ou nominais, sendo as primeiras aquelas que têm ordenação entre as suas categorias e as segundas não tem esta propriedade.
 - variável quantitativa que é representada por categorias. Estas variáveis são ainda contínuas, podendo assumir todos os valores numéricos possíveis, ou discretas podendo assumir apenas valores inteiros.
- **Variável resposta/dependente** : representa a variável de interesse estudar. No caso da previsão pode ser chamada de variável prevista;
- **Variáveis explicativas/independentes** : representam as variáveis que explicam a variável resposta. No caso da previsão podem ser chamadas variáveis preditoras. Ao contrário da variável resposta que é só uma as variáveis explicativas podem ser várias consoante os recursos que se tem.