Understanding Customer Churn in Online Grocery Retail: an Inventory Management perspective

André Carneiro Cerejeira Campos

Master's Dissertation

Supervisor: Prof. Pedro Sanches Amorim

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia e Gestão Industrial

2018-07-02

Abstract

Even though online grocery retail is an industry that has been around for more than 15 years, only recently it is gaining significant exposure, with the rise of consumers' willingness to buy groceries online. This new channel for grocery sales allows companies to capture unfulfilled demand, opening new opportunities to understand customer behavior and more specifically to understand the impact that this unfulfilled demand can have on customer churn. Customer churn in non contractual settings has been already addressed in the literature, hence this work will build on previous research with a focus on the introduction of order fulfillment variables.

Therefore, using Logistic Regression and Gradient Boosting Machines several churn prediction models were applied to an online grocery retailer dataset. The first main objective was to evaluate the impact of product availability on the two main product categories (Food and Perishables) on the customer churn decision. The second main objective was to study the impact that substitutions of out of stock products could have on customer churn. Several iterations were produced to obtain the most reliable results possible. The methodology for this thesis was based on the work of Tamaddoni et al. (2016) which did a comparative study of churn prediction techniques in an online retail database.

The results indicate that when it comes to their churn decision, customers attribute more importance to the service level of the Food category rather than the Perishables category. This was confirmed in both techniques used to predict churn across several iterations. On the other hand, the impact of substitutions on out of stock products was not as consistent, although more importance was generally attributed to stockouts when compared to substitutions regarding customer churn.

These conclusions imply that managers should focus on the service level of the Food category since it has a larger impact on customer churn than Perishables. This should help to effectively reduce customer churn, since this is a variable that can be manipulated by managers, which does not happen on most of the other variables explaining customer churn. Additionally, the results also indicate that the product substitution policy is relevant to reduce the impact of out of stock products on customer churn.

Resumo

Apesar da mercearia *online* ser um sector que existe há mais de 15 anos, apenas recentemente deixou de ser um comércio de nicho, com o aumento da predisposição dos clientes para fazerem compras *online*. Este novo canal de vendas permite que as empresas capturem procura que não foi satisfeita por falta de inventário, o que não era possível no comércio tradicional. Deste modo, pretende-se entender o impacto que esta procura não satisfeita pode ter no comportamento do cliente, nomeadamente na sua desistência do serviço. Esta tese pretende complementar a literatura atual das desistências de clientes em situações extracontratuais, ao acrescentar estas novas realidades a metodologias em uso.

Assim, recorrendo à Regressão Logística e *Gradient Boosting Machines* vários modelos preditivos foram aplicados a uma base de dados de um retalhista. Em primeiro lugar, avaliou-se o impacto que o nível de serviço das duas maiores categorias de produtos (Alimentar e Frescos) teriam na desistência dos clientes. Seguidamente, estudou-se o impacto que as substituições de produtos que sofreram ruturas de inventário no momento da encomenda teriam na desistência de clientes. Várias iterações com diferentes variáveis foram feitas de forma a assegurar robustez nas conclusões. A metodologia usada teve como base o trabalho de Tamaddoni et al. (2016) que realizou um estudo comparativo de técnicas para previsão de desistências de clientes no retalho *online*.

Os resultados indicam que os clientes atribuem maior importância ao nível de serviço da categoria Alimentar do que da categoria Frescos, no que toca à decisão de abandonar o serviço. Isto foi corroborado em todas as iterações realizadas nas duas técnicas de previsão. Por outro lado, a diferença entre tipos de falha não foi tão consistente, mas com uma importância relativa maior atribuída às ruturas de inventário comparativamente às substituições de produtos.

Deste modo, conclui-se que as empresas conseguem reter mais clientes se se focarem no nível de serviço de produtos com um maior prazo de validade, já que os clientes parecem mais sensíveis a este tipo de produtos. Além disso, das variáveis estudadas estas eram que poderiam mais facilmente ser manipuladas pelas empresas, sendo as restantes características do comportamento do cliente. Os resultados relativos à política de substituição de produtos indicam que as medidas atuais aparentam efetivamente reduzir o impacto das ruturas de inventário nas desistências de clientes

Acknowledgements

This thesis would not be possible without the input from both my supervisor Prof. Pedro Amorim and Prof. Gonçalo Figueira, which guided me throughout these five months. Thanks for your constant availability and feedback to accomplish this milestone. Additionally, I would also like to thank Prof. José Luís Borges for the help regarding some of the technical aspects of this thesis.

This work ends my journey as a student, which was enabled by all the great opportunities provided by my parents Margarida and José, thank you for your guidance and education. I also need to express my gratitude to my grandparents and family, and despite being really annoying at times, to my sister Filipa for making this journey a bit more fun.

To my friends that accompanied since the early days, and to the ones that I met in Porto, this is just the end of a chapter from a great book. Thank you for all the great times, and for making my life so enjoyable both inside and outside school. To all the people in L203 and later in L204, thanks for creating such a great working environment and being always helpful throughout these five years.

"I have not failed. I've just found 10,000 ways that won't work."

Thomas Edison

Contents

Acronyms and Symbols

AIC	Akaike Information Criterion
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
FN	False Negative rate
FP	False Positive rate
GBM	Gradient Boosting Machine
LL	Log-Likelihood
LR	Logistic Regression
MAE	Mean Absolute Error
RFM	Recency, Frequency and Monetary
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SKU	Stock Keeping Unit
SV	Stock related variables
TN	True Negative rate
TP	True Positive rate
VIF	Variance Inflation Factor

List of Figures

List of Tables

Chapter 1

Introduction

Customer retention has been one of the biggest challenges faced by companies in an increasingly competitive environment (Buckinx and Van den Poel, 2005). Moreover, customer acquisition has been reported to be five to six times more expensive than retention, and long-term customers generate higher profits (Verbeke et al., 2011).

This concern with retention is especially important in the online setting where customers experience no switching costs, have a broad range of options and can easily compare offers and prices.

The birth of online grocery retail was troublesome, with sales initially not meeting expectations, it seemed that the online grocery market was poisoned for growth (Galante et al., 2013). However, in recent years, the transition from a purely online to a mixed online-physical approach, seems more appealing to consumers as they can choose to just pick the groceries from the store avoiding shipping costs and problems in delivery, which were pointed as some of their biggest obstacles (Galante et al., 2013). Models such as "Click and Collect" (Smith, 2013) that implement this online-physical trade-off are on the rise, with customers ordering products online and collecting them at the store. Thinking in terms of bricks versus clicks is outdated, bricks-and-clicks is the current and future retail reality (Contractors, 2018). Customers tend to be more disloyal as they learn to buy across multiple channels, with 64% switching preferred retailers when migrating from brick-and-mortar stores to the web (Galante et al., 2013).

This new trend is emphasized in a study by the Food Marketing Institute conducted by Nielsen, where online grocery sales are predicted to capture 20% of total grocery retail by 2025. This research also predicts that in as few as five to seven years, 70% of consumers will be doing grocery shopping online (Nielsen, 2018).

The growth of this new reality opens new opportunities for different management policies. Retailers are implementing innovative digital technologies that are transforming the shopping experience, in order to become more relevant to consumers' lifestyles and shopping occasions.

From an inventory management perspective, the proliferation of online grocery shopping introduces new challenging concepts. In the first place, customers view images of products rather than physical products in shelves. This absence of inventory display allows for unfulfilled demand

to be captured, since customers view and purchase items that could be unavailable at the time of picking of the order. This can provide new insights to customer relationship management, as now, in this online context, it is possible to associate service failures and purchase history. Besides, online retailers can often identify customers before transactions are completed, allowing targeted marketing efforts according to individual customer characteristics. Lastly, the online environment facilitates attempts to follow customers longitudinally, enabling companies to track the consequences of stockouts and substitutions.

There is already extensive research regarding customer churn, not only in the marketing but also in the data science literature, and some additional studies regarding the impact of order fulfilment in the online retail industry. However, online grocery retail is a growing sector in e-commerce, and it is characterized by perishable products that are known to yield though challenges in inventory management. A lot of studies have been done about online retail, especially regarding delays, but cases where there are no backorders and stock shortages are so frequent is still a less acknowledged area (Rao et al., 2011). Still no studies have been carried regarding the introduction of these perishable products and their impact on customer behaviour and future churn decision.

Using a case study with data regarding the purchase history and product availability from a one year period of a online grocery retailer, we seek to study these new realities and complement the current literature.

1.1 Objectives

Therefore, the goal of this thesis is to understand what the main drivers of grocery online retail customer churn are, based on data mining techniques. Hence, three main objectives were defined:

- Firstly, understand how to effectively predict customer churn;
- Secondly, study the impact of the service level of the two main product categories (Food and Perishables) in customer churn;
- Thirdly, study if substitutions and stockouts impact customer churn differently.

1.2 Approach

The basis of this thesis is the work of Tamaddoni et al. (2016), where several parametric and non-parametric churn prediction techniques are compared using empirical and simulated data from two on-line retailers. Since the on-line retail setting is common across both studies, the churn definition criterion used is the same. Initially, churn prediction is performed using Logistic Regression, as in Tamaddoni et al. (2016), mainly due to ease of use and interpretability. In a second phase, the same study is performed using Gradient Boosting Machines, in order to corroborate the previous results and check whether better performance could be achieved.

In each model created, variable importance is assessed and compared against the other models. Only when similar conclusions are corroborated by different models, we consider these insights to be robust.

1.3 Thesis outline

The remainder of this thesis is organized as follows.

Chapter 2 provides a theoretical background on some of the concepts addressed in this thesis, from partial churn to a review of the data mining methods used. Chapter 3 briefly describes our case-study and some of their practices regarding this thesis scope and provides an exploratory analysis regarding some of the most relevant variables. Chapter 4 presents the approach used to build the predictive models. This includes the explanation of all the criteria and thresholds used, especially in the selection of the data. Chapter 5 details the use of the predictive models and results obtained. Finally, Chapter 6 gives the insights gained from the results and proposes further research in this area.

Chapter 2

Literature review

In this section, an overview of the state of current literature on the topics addressed in this thesis is provided. From churn definition in non-contractual settings and some of the most relevant approaches used, to the prediction techniques and why they are the most appropriate for this application. After all, a summary and final perspective on the literature is presented.

2.1 Churn in non-contractual settings

"Churn refers to the tendency for customers to defect or cease business with company", (Kamakura et al., 2005).

As the case study for this thesis is online grocery shopping, the focus is going to be directed towards a non-contractual churn setting. This implies some ambiguity when defining churn, since there is no contract between the firm and its customers (Fader and Hardie, 2009). In non-contractual settings, and in particular in grocery retail, besides being difficult to identify the moment, customers usually change their grocery supplier gradually (Oliveira, 2012). So, in this case, churn can be considered to be partial and progressive.

Buckinx and Van den Poel (2005) focused on the study of partial defection to identify which customers were at risk of leaving, in order to apply preventive measures. Oliveira (2012) also studied partial churn by adding the succession of first products' categories purchased as a proxy of the state of trust and demand maturity of a customer towards a company in grocery retailing. In Figure 2.1, a visualization of what is partial churn can be seen.

Finally, in Tamaddoni et al. (2016), a comparison of churn prediction models in online retail settings was performed. They characterize churn as a temporary phenomenon being studied for a finite time period. The goal was to predict customer's partial churn in the next period.

2.1.1 Identification of churners

As it was mentioned above, there is a need to define a threshold in order to group customers as churners or non-churners. In Buckinx and Van den Poel (2005), where a similar problem was faced, the time frame was split in two sub-periods of five months: a "Calibration period"

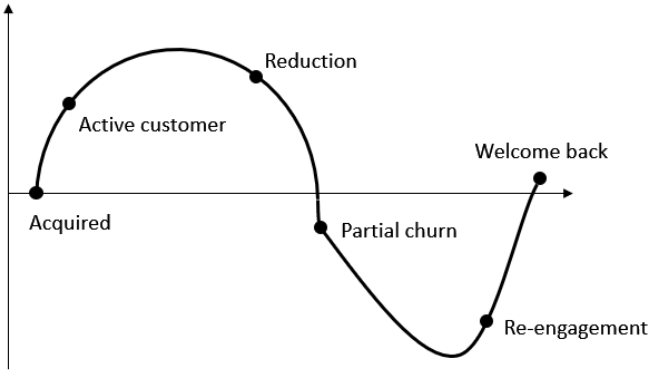


Figure 2.1: Customer journey in a company adapted from Goran et al. (2015)

and a "Prediction period". In the calibration period the most valuable customers were selected according to their purchase frequency and regularity. Then, if these conditions were not fulfilled in the prediction period customers were classified as churners. With these two sub-periods it was possible to make an individual level prediction of the probability to churn in a near future. All the models built were binary classification models where the explanatory variables classified a customer as a partial churner or not.

Following this paper, Tamaddoni et al. (2016) used an evolution of this criterion, applied to two online retailers. Here, instead of the predefined five months used in Buckinx and Van den Poel (2005), in Tamaddoni et al. (2016) the biggest individual interpurchase time was used as the prediction period, and the remaining data frame as the calibration period. With this strategy they expect to avoid problems with the prediction period being too long (taking too long to detect churn) or too short (misclassifying customers with long interpurchase times).

In order to define a customer as a churner or non-churner in the prediction period, a criterion must be defined. Tamaddoni et al. (2014), Chen et al. (2015) and Tamaddoni et al. (2016) flagged a non-active customer when the customer had at least one purchase in the calibration period and no purchases in the prediction period. On the other hand, Buckinx and Van den Poel (2005) focused their analysis on the customers who had a frequency of purchases and a standard deviation of the interpurchase time above average in the calibration period. If these selected customers fell below these averages in the prediction period, they were identified as a churner. Oliveira (2012), when defining churners in grocery retail, focused on spending and classified a customer as a partial churner if he spent less 40% in the next periods than in the reference period.

2.1.2 Independent variables for churn modeling

Independent variables are the set of features that are used to predict the dependent variable in a machine learning model. Previous studies of churn in grocery retail gave special importance to behavioural predictors known as recency, frequency and monetary (RFM) variables. Buckinx

and Van den Poel (2005) used as predictors only individual level behavioural data, customer demographics and perceptions as these are the most used group of variables in previous research. They concluded that behavioural predictors (more specifically RFM) were more relevant than demographics. Oliveira (2012) confirmed the importance of RFM variables while introducing the succession of first products' categories purchased to predict partial churn.

In Tamaddoni et al. (2016) apart from recency and frequency-based variables, some other variables such as number of failures (not inventory related) and loyalty card membership significantly contributed to building the models. Additionally, decreases in customers spending on some categories could also help improve accuracy.

2.2 Churn prediction

Churn prediction problems are often formulated as binary classification problems in which the aim is to assign customers, to one of the two classes Y = {Churner; Non Churner} on the basis of their observed features (Stripling et al., 2017).

As mentioned previously, in Tamaddoni et al. (2016) the prediction power of several data mining techniques was compared in an online retail setting. The conclusion was that Boosting and Logistic Regression were the best performing techniques, so these were the selected models for the current study.

2.2.1 Logistic regression

Logistic Regression (LR) is a very popular technique in customer churn prediction, known for good predictive performance (Coussement et al., 2017; Neslin et al., 2006). Verbeke et al. (2012) confirmed the good predictive performance of LR in a benchmark study. Other known characteristics of LR are the ease of use and the robustness of the results (Buckinx and Van den Poel, 2005). Besides, it does not rely on assumptions of normality for the predictor variables or the errors and may handle non-linear effects (Fahrmeir and Kaufmann, 1985). LR handles linear relations between variables very well, but it does not take into consideration the interaction effects between variables (De Caigny et al., 2018). In a benchmark study (Tamaddoni et al., 2016) LR was benchmarked against several data mining and probability models and had a consistently robust performance. Moreover, while Boosting had a better predictive performance, it can be argued that LR has better comprehensibility, so it should be a technique to consider. Finally, LR is computationally fast, enabling the use of feature selection methods that require the training of many different models.

2.2.2 Gradient boosting machine

Gradient boosting is a machine learning technique that combines two powerful tools: gradient-based optimization and boosting. Gradient-based optimization uses gradient computations to minimize a model's loss function in terms of the training data. Boosting additively collects an

ensemble of weak models to create a robust learning system for predictive tasks (Malohlava and Candel, 2017).

It is an ensemble of classification or regression trees, that has shown considerable success in multiple practical applications. This is due to the fact that they are highly customizable to the particular needs of the application, such as being learned with respect to different loss functions (Natekin and Knoll, 2013).

A Gradient Boosting Machine (GBM) sequentially builds models based on the errors previously made. It is an iterative procedure that helps to improve the accuracy of trees, starting with simpler trees (weaker learners) and stacking them.

Friedman (2002) concluded that, the approximation accuracy and execution speed of gradient boosting can be improved by incorporating randomization into the procedure. Specifically, at each iteration, a subsample of the training data is drawn at random (without replacement) from the full training data set. This randomly selected subsample is then used in place of the full sample to fit the base learner and compute the model update for the current iteration. This randomized approach also increases robustness against overcapacity of the base learner.

One critique made to this algorithm is that it will capture noise and may be prone to overfit if too many models are stacked. To overcome this, it is common to use a validation set to stop the model when it tries to learn the noise of the data (section 4.7). Moreover, it can also be sensitive to extreme values in the data.

2.3 Churn prediction vs comprehensibility

Literature on churn analysis is generally divided into descriptive and predictive analysis (Tamaddoni et al., 2010). The former is focused on explaining churn and is more popular throughout the marketing literature (Ahn et al., 2006). On the other hand, the latter is focused on identifying potential churners before actual churn becomes apparent and it appears more in data mining literature (Mohammad et al., 2016).

In customer retention management it is not only important to know if customers are leaving the company but also why they are leaving. Thus, the focus should not only be in prediction but also in comprehensibility of the models to make better informed decision when combating churn (Gustafsson et al., 2005). The comprehensibility of a model however also depends on the number of variables included. Clearly a model including ten variables is easier to interpret than a model containing fifty variables or more (Verbeke et al., 2012).

These were key factors when choosing the predictive techniques for this study, with LR being a more intuitive and simple method, stronger for comprehensibility. On the other hand, GBM is a more complex method but better on prediction.

Taking this into consideration, this thesis will focus on comprehensibility, but always trying to maintain a high prediction accuracy in order to keep the models relevant and provide valuable insights into the business.

As LR is a more simple and intuitive method it will be used as the primary method for understanding churn, with GBMs as a support.

2.4 Customer churn and operations management

Certifying that on-hand inventory meets customer demand is a basic operations function that affects both short-term revenues and long-term customer loyalty (Jing and Lewis, 2011). But meeting demand with appropriate supply can be tough due to demand uncertainty or supply inflexibility. This results in inventory shortages that might lead to lost sales and lower revenue (Consulting, 1996).

In an alternative approach to evaluate the impact of stockouts in online retailing, Jing and Lewis (2011) proposed the adaptation of Zhang and Krishnamurthi (2004) modelling approach to capture the consumer's joint decision of purchase incidence and amount. Using data from the first 14 months of operations of an online non-perishable grocery and drugstore company, Jing and Lewis (2011) concluded that stockouts can have a dramatic but nonlinear impact on the seller's profitability and that prioritizing inventory for targeted types of customers can be helpful.

2.5 Final perspective on current literature

To sum up, customer churn analysis in non-contractual settings is an area that is still incipient. The fact that there is no clear measure to tell if a customer ceased relations with a company, leads to several studies in different areas to be performed.

In the literature referring to grocery retail and customer churn, the impact that unfulfilled demand can have on customer churn has never been fully explored. This may be due to the fact that only with the introduction of the online channels, this unfulfilled demand can be properly captured.

Product category information had already been explored in this context, but only at a quantity and spending levels. This thesis introduces this category differentiation oriented to an inventory perspective.

In the selection of the best method to compute churn and assess the performance of the predictors, several methods were considered, but LR and Boosting seemed to have the best balance between performance and comprehensibility in comparative studies (Tamaddoni et al., 2016).

Chapter 3

Problem description

In this chapter, a brief introduction to the case-study is performed, followed by a description of the available data. Afterwards, the focus turns to how the company manages stockouts, followed by a description of the process of substitutions. Lastly, an exploratory analysis of the data is made in order to get better understand the business.

3.1 Case study

The database used for this analysis belongs to a Portuguese leader in food retail, which has been operating in the on-line grocery retail since 2001. The website serves as an extension of the well-established brand, with almost 300 stores spread across the country. Consequently, the brand uses their developed infrastructure to secure this evolving market.

In 2013, this on-line store had more than 20 thousand different products, both food and non-food, with more than 12 million annual website visits (Ferreirinha, 2014).

When a customer places an order, the day of delivery and time slot to receive the order can be chosen. Delivery prices vary for different slots. Orders could be received at home and in certain selected stores collected by car or inside the store. The company also sells a subscription service for home deliveries in which, for a fixed value, all home deliveries for 100 days are free of delivery charges. If a customer wants to return a product, it can be done at any store of the company up to 15 days after the order arrives.

3.1.1 Available data

The raw data consists of customers transactional information at an individual level, from 01/10/2016 until 29/09/2017. This included information about:

- **Order details:** when the order was placed and by which customer as well as all the SKUs in each order, with quantity, price and discount discriminated;
- **Product specifications:** product characteristics such as brand and category;

- **Delivery status:** delivery time, delivery cost and description of problems in delivery such as delays, damaged goods, missing items and items returned;
- **Picking information:** information relative to the stock level of a SKU. It could be in stock, out of stock, substituted or partially in stock. Products are only substituted when they are out of stock and if an appropriate substitute is available;
- **Geographic distribution:** shipping postal code and information about the store where picking for the order was made.

3.1.2 Stockouts

Orders are fulfilled from the current physical grocery stores of the company. Therefore, a trade-off between reserving stock for the online customers and serving the customers visiting the stores arises. To tackle this, picking for the online orders is distributed by different shifts, according to the time of the order delivery. Since some shifts are at the end of the day, there could be less stock meaning that stockouts are more prone to happen. Picking for perishables is always made on the delivery day to assure that the items are fresh. The company in question has an online dedicated store, that does not sell to customers visiting the store. Consequently, this store experiences on average 17% less stockouts for online orders than stores that combine online and physical shopping.

3.1.3 Substitutions

When a stockout happens on an ordered item the picker tries to choose a similar product to substitute it. An e-mail is then sent to the customer informing about the substitutions made on the order. The picker tries to follow some rules in order to maintain consistency and make the best possible choice for the customer. Here are some examples for some products:

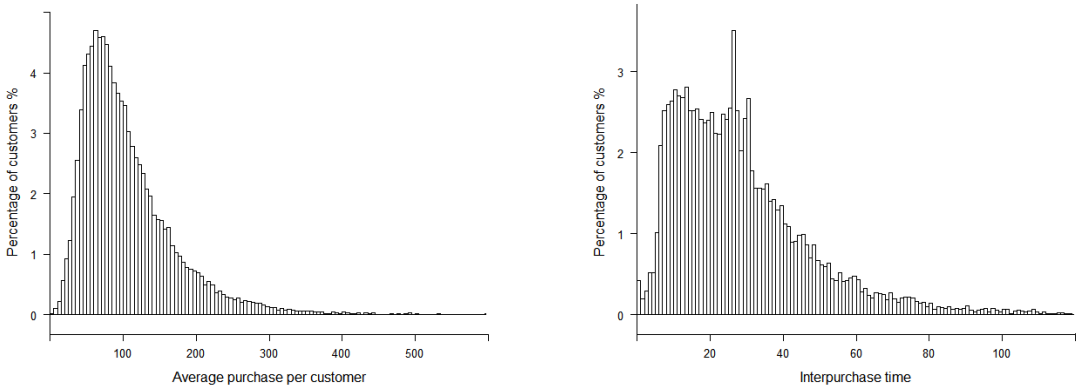
- for products with a certain flavour like yogurts, the item is replaced by one with a similar flavour;
- for basic ingredients like rice and pasta, the items are replaced by white label products;
- for hygiene and cleanliness products, the items are replaced with items from the same brand;
- when products have promotions, they are replaced by equivalent items in promotion.

There was a time in which the retailer used to call their customers asking if they accepted specific substitutions or not. However, during the period of the analysis, the customer had only the option to select whether they want to receive product substitutions or not. They could additionally right a note on each product when ordering, stating if they did not want to receive substitutions for that product, or what preferences they may had. If the price of the product substituted is lower than the original product that was out of stock, the price difference is reimbursed through the loyalty card for future purchases.

3.2 Exploratory analysis on the calibration data

This section focuses on the calibration period, where the independent variables will be calculated, and algorithm will be trained. This analysis tries to characterize the more relevant variables with some preliminarily insights. The focus is not only in variables such as recency, frequency and monetary (highly studied in the marketing literature), but also on product related and service failure variables.

Analysing Figure 3.1 (a) it is possible to see the distribution of the average purchase value per online customer, which has a mean of around 105€. It is interesting to note that this is considerably higher than the average transaction at conventional grocery retail stores, where the mean is around 60€ (Oliveira, 2012). The larger average purchase could be influenced by the delivery costs that encourage larger orders.



(a) Average purchase value per customer (€)

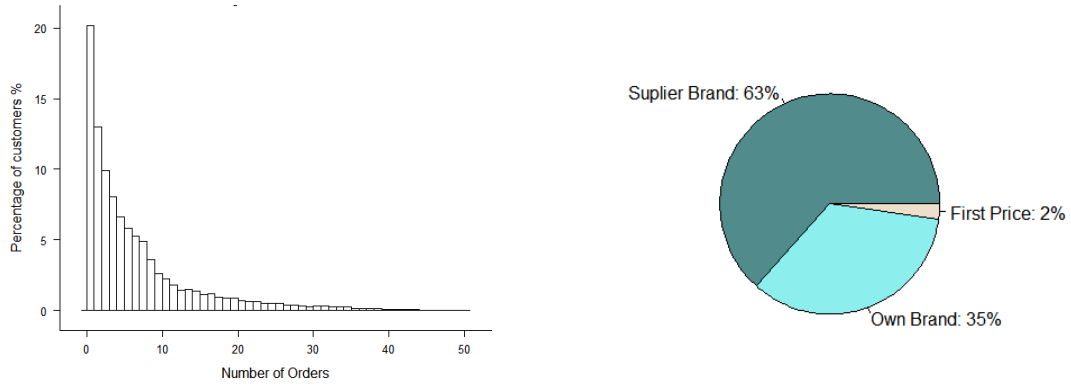
(b) Average interpurchase time per customer (days)

Figure 3.1: Histograms of purchase amount and frequency

Thus, as these online customers do larger purchases it makes sense that interpurchase times are longer when compared to conventional grocery shopping. In Figure 3.1 (b) this can be confirmed, with a mean value of 28 days and standard deviation of 18 days, the distribution of interpurchase times seems to vary substantially from customer to customer. In fact, it is interesting to note that there is sizable proportion of customers with interpurchase times from around seven days until slightly over a month. This is quite high when compared to the mean of 14 days for conventional grocery shopping (Oliveira, 2012).

With respect to the number of orders, around 20% of customers have only a single order, as it can be seen in Figure 3.2 (a). This could be due to some aggressive marketing strategies, dissatisfaction with the service, or incompatibility with the offering. Being a growing sector, not all customers are acquainted with buying groceries online, leading to some wanting to try this concept but not converting to repeat customers. As these customers are such a big part of the customer base, they will be further discussed in the next chapter.

The company in question is known for having two types of private labels. One is focused on the essentials, known to be the first price in the categories it is present. The other is a more common private label, which intends to substitute most supplier brand products. As it is expected, not all supplier brand products have equivalent private label products, so the presence of the former is superior (Figure 3.2 (b)). It should also be noted that private label products are substitutes of the supplier brand products, which makes the percentages of these two types of brands highly correlated.



(a) Histogram of the number of orders per customer

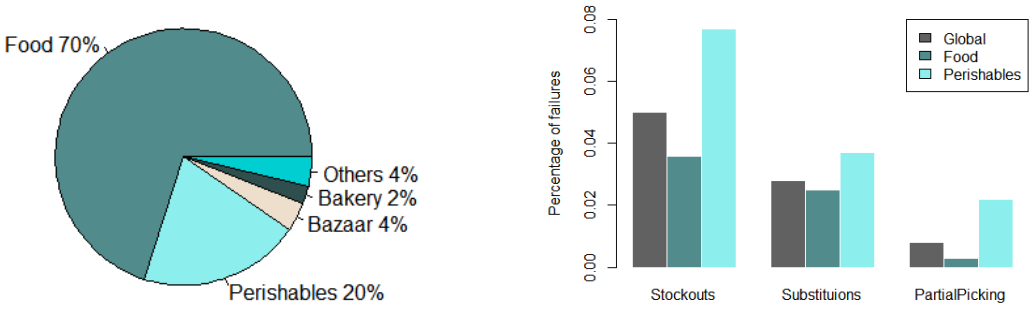
(b) Distribution of brand types

Figure 3.2: Distributions of the number of orders and brand types

Regarding the product categories, as this is grocery retail, most items sold by the company can be segmented into two big categories: Food and Perishables. Food refers to all the packaged goods with higher expiration dates, accounting to around 19 thousand different SKUs. On the other hand, Perishables refer to the fresh goods known for the lower expiration date and has around 4 thousand different SKUs. Packaged goods are known for having a big diversity of characteristics such as flavors and sizes that can create variations around one product, resulting in a large number of SKUs. Consequently, the Food category represents around 70% of the total SKUs that are bought (Figure 3.3(a)).

In Figure 3.3 (b), the comparison between the percentage of Stockouts, Substitutions and Partial Picking for each of the two biggest categories (Food and Perishables) is shown. From the analysis of this chart, a couple of facts stand out. First, as it is expected the Perishables category dominates with higher percentage of failures throughout the different types of picking. With a lower expiring date, increased inventory costs result in lower stock levels. Furthermore, it can be seen that the percentage of stockouts is the largest, followed more closely by substitutions but with a big difference to Partial Picking. Partial picking happens when there isn't enough stock to match the order, but some items are sent anyway. Due to the lack of information in the database regarding this situation, i.e. number of items sent, allied to the low frequency of observations this type of picking was omitted from the methodology.

The loyalty card is a big part of the company's efforts regarding customer retention and is



(a) Distribution of product categories

(b) Types of picking for Perishables and Food categories

Figure 3.3: Distribution of product categories and types of picking

mainly used to accumulate discounts for future purchases. Moreover, the collection of the customer's purchase behaviour allowed the company to make customized promotions. As there was no information in the database relative the ownership of the card, a proxy had to be made. If a customer ever used their loyalty card for discounts, then it is because they have one. Otherwise, it was assumed that the customer didn't have a loyalty card. The fraction of customers that used the loyalty card in the database is 67.54%.

Chapter 4

Methodology

The methodology for this thesis can be divided in two segments. Initially, a more global methodology was idealized, to be used by all the specific models. This serves as a base for the problem analysis and will be described in this chapter. The, specific details for the construction of each model will be described in the next chapter.

4.1 Approach

In order to tackle the goals defined previously, the approach could be divided in three main steps:

- Firstly, the data that was spread across different tables in the dataset was aggregated into one table, where each row represented a different customer. For this, several variables had to be created to translate the past purchase behaviour. The criteria for churn definition (division into sub periods and assignment of churners) was also performed in this stage. This first step was performed using SQL Server;
- Secondly, the data was prepared for the prediction models. This meant doing customer selection and performing standard data preparation techniques;
- Thirdly, training and testing sets were defined, followed by the execution of the models, evaluation of performance and comparison of results. For this multiple evaluation metrics were considered according to the technique used.

To perform these last two steps the software used was R. It is an environment embedded with implemented statistical techniques. Through additional packages, R can easily be extended. A recent feature of the R package H2O that enables automatic machine learning was used. Through the iteration of several different methods and hyperparameters, this package outputs several ranked optimized models.

Regarding model construction, two different techniques will be used. First, LR will be used for predicting churn, with posterior variable importance assessment. Then, GBMs will be used with the same goal, in an effort to increase robustness.

At the end, the relative importance of features among the different models is compared, and the rank of the different variables is assessed.

4.2 Churn definition

As it was addressed in Chapter 2, there are several approaches to define what is a churner in non-contractual settings, and all are very context dependent. In this thesis the method used was the same as Tamaddoni et al. (2016), since it was performed on a similar setting (online grocery retail) and seemed the most suitable for this business.

In order to define what is a churner, the time frame was split into calibration and prediction periods. Based on the criteria used by Tamaddoni et al. (2016), the duration of the prediction period was calculated with 2 steps. Firstly, the customers were sorted in ascending order based on their average interpurchase times. Then the prediction period was set to be approximately equal to the average interpurchase time of the last customer in the 99-quantile of the sorted customer base. This is made to capture the activity not only of customers with a long interpurchase time, but also those with a short interpurchase time as soon as possible.

This resulted in the division presented in Figure 4.1, with a calibration period of 250 days and a prediction period of 114 days. This was later confirmed by a domain expert from the company in question, to be a realistic value for this business.

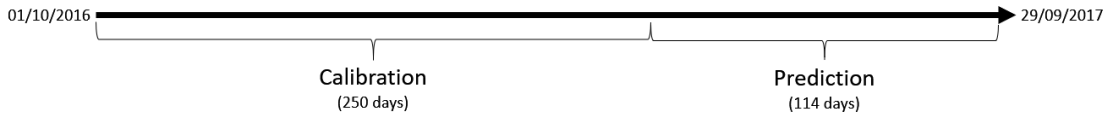


Figure 4.1: Timeline division into calibration and prediction periods

After the division into subperiods a rule had to be set to define what would separate the churners from the non-churners in the prediction period. Due to the type of purchases made in this setting (larger quantities and with longer interpurchase times - Chapter 3) and considering the literature on this subject, the following criterion was set: to be defined as a churner a customer needed to have at least one purchase in the calibration period and no purchases in the prediction period. If a customer made at least one purchase in each period, then it would be classified as a non-churner.

4.3 Customer selection

This work deals with a concept that is difficult to quantify churn in non-contractual settings. So, the selection and computation of the explanatory variables is crucial. To properly evaluate certain variables, the focus should be on the relevant segment of customers. With this selection, the purpose is to create a homogeneous set of customers, so that more insightful conclusions can be

achieved. In the following subsections it is detailed why four sets of customers will be excluded from the analysis.

4.3.1 Time of first purchase

In order to evaluate the purchase behaviour of a customer we need to consider his evolution throughout the time frame. In this case the time frame where the explanatory variables are predicting churn is the calibration period. So, similarly to Tamaddoni et al. (2016), only the customers that made their first purchase in the first half of the calibration period are being considered for this analysis. This allows the engineered variables that reproduce the changing behaviour from the first to the second period to be properly interpreted.

4.3.2 One-time buyers

Trying to understand the behaviour of a customer that never repeated a purchase in an extended period does not seem reasonable, since this is a business characterized by repeated purchases.

Moreover, most customers that only made one purchase probably never intended to continue to purchase online and were just compelled by aggressive marketing strategies. As mentioned in Section 3.2, these one-time buyers represent around 20% of the customer base, which can heavily influence churn analysis. So, as churn analysis is based on established customers, these one-time buyers were removed from the analysis.

4.3.3 Bulk buyers

On the other hand, some customers are extremely loyal and make only high-volume purchases. Actually, some of these extreme spenders could be small companies or other grocery retail businesses. Additionally, at the time of this study, this company had recently introduced a business-oriented platform, so some of these extreme buyers could be in the process of migrating. To select these extreme buyers, customers with a total purchased amount over 7000€ were excluded from the dataset (average of around 200€ per week). This results in a selection of around 110 removed customers.

4.3.4 Product categories

In order to evaluate how both product categories and product substitutions/stockouts influence customer churn these had to be distinguished in the analysis. As it was mentioned in Section 3.2, there are two categories that clearly stand out in terms of sales: Food and Perishables.

To properly assess how stockouts and substitutions in these categories influence churn behaviour, only the customers that had purchased products in both categories were considered for analysis. This meant eliminating about 8% of customers.

When a customer has never made a purchase in one of these categories, the variables for service level in this case appear as a missing value. At least for the LR, missing values can heavily influence results, reducing model performance.

The point could also be made that these customers were less relevant to the company, since they did not purchase their main product categories. This was also confirmed by a domain expert in the company, pointing out that some customers only use the service to buy books or other more niche products, having a very distinct purchasing behaviour from the rest of the customer base.

4.4 Engineered variables

In this subsection, it is explained how the variables were created, in order to gather information about the available customer purchase history.

The explanatory variables can be divided in three main categories: purchasing behaviour, service failure and customer demographics. These are all computed solely on the calibration period. On the other hand, the response variables are computed in the prediction period. There are two response variables, that are going to be used to reproduce different models.

In the following subsections the list of variables based on existing literature and available data is presented.

4.4.1 Purchasing behaviour

According to the literature, the most relevant variables of purchasing behaviour are recency, frequency and monetary (RFM) (De Caigny et al., 2018). These are used in multiple occasions and are the most popular for predicting churn among authors (Oliveira, 2012; Tamaddoni et al., 2014; Chen et al., 2015). Buckinx and Van den Poel (2005) found some other relevant predictors such as length of the relationship, number of categories purchased, usage of loyalty points and brand preferences.

The purchasing behaviour variables considered were:

- *NumberOrders:* frequency variable, with extended use in the literature. Counts the number of orders made by the customer in the calibration period;
- *TimeSinceFirstPurchase:* as in Tamaddoni et al. (2016), to account for recency two variables were used this first one relates to the time since the customer's first purchase;
- *TimeBetweenFirstLastPurchase:* this is the second variable regarding recency. It expresses the effective relationship duration in the calibration period;
- *ChangeTotalSpending:* difference in spending from the first to the second half of the calibration period, to determine if the customer is increasing or decreasing their spending (Figure 4.2). If positive the customer is increasing spending;

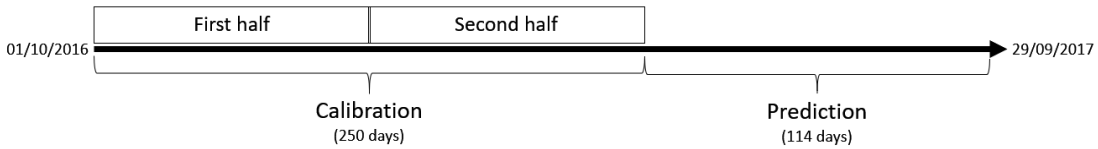


Figure 4.2: Division of the calibration period into two sub periods

- *SupplierBrand:* ratio of supplier brand products over the total number of products. Customers, usually either buy supplier brand or private label products, as it can be seen in Chapter 3. Since there is a big correlation between private label and supplier brand variables, only supplier brand was kept in the study;

- *LoyaltyCardOnwership:* binary variable that describes if a customer has used the loyalty card for discounts on any purchase. No information was available regarding loyalty card ownership, so this proxy had to be used;

- *DeliveryCost:* average delivery cost regarding home deliveries. It only depends on the time slot chosen by the customer to receive the order at home;

- *DeliverySubscription:* change in spending on the delivery subscription service from the first to the second half of calibration period. This delivery subscription allows the customer to have free delivery fees for 100 days. Additionally, the total amount spent in this subscription service was also considered;

- *NumberSubcategories:* number of different subcategories purchased;

- *Promotions:* value saved in promotions over the total purchased amount.

4.4.2 Service failures

Service failure is an area that has been less acknowledged by the literature. In Tamaddoni et al. (2016) some product delivery failures, like damaged items and refunds were included but not with the extent of this study. On the other hand, Jing and Lewis (2011) studied stockouts more in depth but not applied to churn, and instead to a proxy of customer value, separating categories by product penetration depth.

In this thesis, product stockouts and substitutions will be the focus of the analysis, both separated by the two largest product categories (Food and Perishables). Furthermore, incidents in delivery (delays and damaged items) will also be considered.

A Stock keeping unit (SKU) is a product and service identification code that is unique due to some characteristic (brand, size, colour, model). Since this is online grocery retail, there is a vast diversity of products, which implies a large number of SKUs. This multiplicity of products makes it difficult to compare quantities and prices paid between products.

Thus, for the calculation of the service failure variables, it was only assessed if a certain SKU in an order was in stock or if it was substituted. In the following subsections, when it is mentioned the "number of SKUs", we are referring to the number of different products in each order. This can better understood in Figure 4.3.

Order	SKU	Picking
1	A	Substituted
1	B	Full
1	C	Stockout
2	D	Substituted
2	A	Stockout
2	C	Stockout
2	E	Full

Customer	Substitutions	Stockouts
ID	$^{2}/_{7}$	$^{3}/_{7}$

Figure 4.3: Calculation of service failure variables by customer

Following the example of Figure 4.3, if a customer purchased on both orders "SKU A" and "SKU C", the total amount of SKUs considered is seven. Although the customer only bought five different products in total, two of them were purchased in two different orders which leads to different inventory levels and possible stockouts.

The service failure variables considered were:

- *Stockouts:* consists in the number of SKUs that were out of stock at the time of the picking and that were not substituted, over the total amount of ordered SKUs;
- *Substitutions:* ratio regarding the amount of substitutions over the total amount of received SKUs;
- *FoodFailures:* sum of the ratio of stockouts and substitutions for the Food category. It can be seen as the probability that a customer does not receive the desired SKU, when ordering the Food category;
- *PerishablesFailures:* sum of the ratio of stockouts and substitutions for the Perishables category. It can be seen as the probability that a customer does not receive the desired SKU, when ordering the Perishables category;
- *DeliveryIncidents:* ratio of SKUs that had delays or damaged items during delivery.

Both *Stockouts* and *Substitutions* are also separated by category, referring to the percentage of that type of failure inside each category.

4.4.3 Demographics

In line with Tamaddoni et al. (2016), the relationship between satisfaction and loyalty can be distinct for people with different socioeconomic status. Therefore, the inclusion of the variable *SocioeconomicStatus* in the study seemed relevant, to capture differences in the purchasing power of different customers.

Socioeconomic status is an index regarding the purchasing power of the customer. Considering that this type of data was not available in the database a proxy had to be produced. So, using the postal code to where the orders were delivered and the country's database with purchasing power by municipality (PORDATA, 2015), this index was extracted.

4.4.4 Response variable

As it was mentioned previously, the response variable is *churn*, which is a binary variable that indicates if a customer has churned in the prediction period according to the criteria above mentioned. If it is one represents a customer that churned in the prediction period while a zero represents a non-churner;

4.5 Data preparation

After the selection of the customer base it is now important to prepare the data in a way that it could be used for modelling purposes. "Raw data" contains all sorts of "pollution", such as discrepancies, missing values, outliers and input errors, which will lead to a poor model performance. Data preparation enhances the chance of building models that are relevant for practice.

The "raw data" that was available had already suffered some data cleaning. The products without any orders were removed, as well as customers with no orders and other kinds of mismatches. Still, a lot had to be done since this is one of the most important phases in predictive analysis.

4.5.1 Standardization

Standardization is the process of putting different variables on the same scale, by subtracting each variable by its mean and dividing it by its standard deviation.

When performing LR, the numeric predictors were standardized so that scale is the same, and the magnitude of the LR coefficients can be directly compared. Taking the absolute value of the standardized coefficients enables them to be ranked from highest to lowest, in order of strength of association with the outcome. A downside of this method is that standardized scale may not be as intuitive as the original scales, since we need to think in terms of standard deviations instead of one unit increments (Thompson, 2009).

4.5.2 Multicollinearity

As it was already mentioned, LR coefficients are affected by correlations between explanatory variables. To avoid this, firstly the pairwise Pearson correlations were checked and the threshold of 0.7 was defined as in Dormann et al. (2013). If two explanatory variables were above this threshold then they would be further investigated to find the reasoning behind it. When two variables portrayed the same information then only one was used for the modelling purposes.

When performing regression analysis special attention should be given to the effects of multicollinearity and the risks that may exist in the detection of linear dependencies. The Variance Inflation Factor (VIF) provides the user with a measure of how many times larger the variance is for multicollinear data than for orthogonal data (where each VIF is one). Neter et al. (1996) considered that if the value of VIF is greater than ten, then multicollinearity may be causing problems with the estimations. VIF for the independent variables considered in this study was below three.

4.5.3 Missing values and outliers

The data was already pre-processed for other uses, so cases where, for example, there were products with no orders, or just empty orders were already excluded. Still, preventive measures were taken.

Due to the nature of the engineered variables (aggregations and ratios) a lot of missing values appeared after their creation. So, in this case the most common case of missing values is when, for example, one customer did not buy a certain category of products. This results in one of the relevant variables, for example, *FoodSubtitutions* having a missing value. Here, a zero would mean that the customer never had a substitution in this category, but that is misleading because he never bought it. To mitigate this, as it was mentioned in customer selection (Section 4.3), customers with no purchases in the two biggest categories were excluded, since it would create misleading conclusions regarding category importance.

On the other hand, outliers should also be evaluated carefully since they can be part of the relevant range of analysis. In this case, as it was mentioned in customer selection (Section 4.3), some outlier customers had extreme values of spending and could be input mistakes or small companies, so these were excluded from the analysis. The remaining variables in the dataset were also checked for these discrepancies (comparing ranges and distributions), but no other relevant cases were found, apart from input errors.

4.5.4 Training and testing division

As it is good practice in data mining, the data set is split in two: a training and a testing set. The former is used for the algorithm to learn how to classify orders. While the latter is kept isolated, to be used just at the end of the modelling procedure to estimate the performance of the chosen model. According to the literature, the division between the number of observations in the testing set usually lays between 20-30% of the total number of data entries. In this work 25% of the instances were used for testing, hence 75% of the data set was used for training.

The purpose of a testing set is to replicate as close as possible the behaviour of the model with new data. As the dataset is rather imbalanced, by randomly selecting samples an unrealistic representation of new data could be created. So, it was certified that the rate of churners in the test set was similar to the overall rate, by performing stratification.

The end result of this data preparation process culminated in a training set with 23.583 customers and a test set with 7.859 customers.

4.6 Feature importance

Having prepared the data for analysis it is now crucial to understand the importance that the different predictors have in predicting customer churn.

The absolute value of the LR coefficients can be used to access feature importance (Nathans et al., 2012). However, as the coefficients depend on the scale of the feature, it is common practice use standardized features/coefficients (Thompson, 2009). As it was already addressed though, LR relies on the assumption that there should be little or no multicollinearity among the independent variables, so interactions between variables should be analysed.

In order to strengthen the conclusions of the LR analysis, GBM was also used to check for variable importance. This technique uses very different assumptions from LR which improves robustness (no standardization needed and not sensitive to multicollinearity). One of the reasons GBMs were used is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. In Friedman and Meulman (2003) variable importance in GBM is defined as the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees.

4.7 Tuning of GBMs

As it was previously mentioned, a recent function of the H2O package was used to create the GBMs. This package searches the hyperparameter space through an iterative process, in order to find the best possible model according to a previously set metric.

Parameter tuning in this function can be done in three different ways:

- **Random grid search:** randomized version of grid search were instead of running all possible combinations of parameters, only a few are selected randomly;

- **Tuning of individual models:** for GBM models where trees are stacked, an early stopping criteria should be defined. In this case, the model is tested on a validation set to check if it is overfitting. This happens when the model tries to learn the noise of the training data and does not improve performance;

- **Bayesian Hyperparameter Optimization:** on each iteration, saves a surrogate model with the hyperparameters used and the corresponding performance. Taking all the past iterations into consideration, return a new set of hyperparameters to be tested.

The GBM hyperparameters supported and tested in this process are:

- **Score tree retrieval:** score the model after every so many trees;
- **Histogram type:** option to specify the type of histogram to use for finding optimal split points (UniformAdaptive, Random, QuantilesGlobal and RoundRobin);
- **Number of trees:** increasing N reduces the error on training set, but setting it too high may lead to overfitting;
- **Maximum depth:** deeper trees can seem to provide better accuracy on a training set because deeper trees can overfit. Also, the deeper the algorithm goes, the more computing time is required;
- **Minimum rows:** specifies the minimum number of observations for a leaf in order to split;
- **Learning rate (shrinkage):** used for reducing, or shrinking, the impact of each additional fitted base-learner (tree). It reduces the size of incremental steps and thus penalizes the importance of each consecutive iteration. The intuition behind this technique is that it is better to improve a model by taking many small steps than by taking fewer large steps. If one of the boosting iterations turns out to be erroneous, its negative impact can be easily corrected in subsequent steps (Bhalla, 2016);
- **Row sample rate and column sample rate:** can improve generalization and lead to lower validation and test set errors. Higher values can improve training accuracy;
- **Column sample rate per tree:** it is multiplicative with Column sample rate. Setting both parameters to 0.8, for example, results in 64% of columns being considered at any given node to split;
- **Minimum split improvement:** search a few minimum required relative error improvement thresholds for a split to happen. When properly tuned, this option can help reduce overfitting because the algorithm will stop splitting when all the possible splits lead to worse error measures.

4.8 Evaluation

After creating various models, some criteria are needed to compare and evaluate them. Thus, each model will be evaluated with the most relevant metrics according to technique used, based on the literature.

4.8.1 Cross-validation

The type of cross validation used was k-fold cross-validation, which splits the dataset into k mutually exclusive subsets of an approximately equal size. K models are produced, and each is tested on the fold that was excluded on the respective training set. This method avoids biased estimates since all the data is considered both in the training and testing set. By averaging the accuracy of the various models, we can obtain a more reliable estimate. The k used for this work was 5 since it was suggested by the function used in H2O, so to standardize across all models the same k was used.

4.8.2 Classification

Evaluation criteria that are commonly used for testing the validity of a classification model are: accuracy, recall and the area under the ROC curve (AUC) (He and Garcia, 2009).

Accuracy and recall are calculated with the use of the confusion matrix. As can be seen from the confusion matrix (Table 4.1) a predicted value can be: a True Positive (TP), a False Positive (FP), a False Negative (FN) or a True Negative (TN). In case of a TP or a TN, the case is predicted correctly and in case of a FP or a FN the case is predicted incorrectly.

False positives can also be called *type I errors*, while false negatives can be called *type II errors*. In this case a *type I error* would mean that the model incorrectly assigned a customer as a churner, while a *type II error* would mean that the model incorrectly assigned a customer as a non-churner.

Table 4.1: Example of confusion matrix

		Real class	
		Non-churner	Churner
Predicted	Non-churner	True Positive	False Positive
	Churner	False Negative	True Negative

Accuracy is the ratio of the accurately classified instances over the total number of cases. When the data is unbalanced, accuracy is not a very reliable performance measure. Common values for imbalanced data are generally in the order of 100:1 or higher (He and Garcia, 2009). This dataset has a 10:3, which should not be problematic but it still should be taken into consideration when interpreting some metrics, particularly accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4.1}$$

As we can see from (Equation 4.1), accuracy uses both columns' information. Therefore, as class distribution varies, measures of the performance will change even though the underlying fundamental performance of the classifier does not. To tackle this, other performance measures

can be derived from the confusion matrix, such as:

$$Recall = \frac{TP}{TP+FN} \tag{4.2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4.3}$$

ROC curve

The Receiver Operating Characteristic graph is formed by plotting the TP rate as a function of the FP rate, and any point in ROC space corresponds to the performance of a single classifier on a given distribution. An example of a ROC curve can be seen in figure Figure 4.4. In the case of soft-type classifiers, i.e. classifiers that output a continuous numeric value to represent the confidence of an instance belonging to the predicted class, a threshold can be used to produce a series of points in ROC space (He and Garcia, 2009). A ROC curve plots the TP rate and FP rate for each threshold between $-\infty$ and $+\infty$ (Fawcett, 2006).

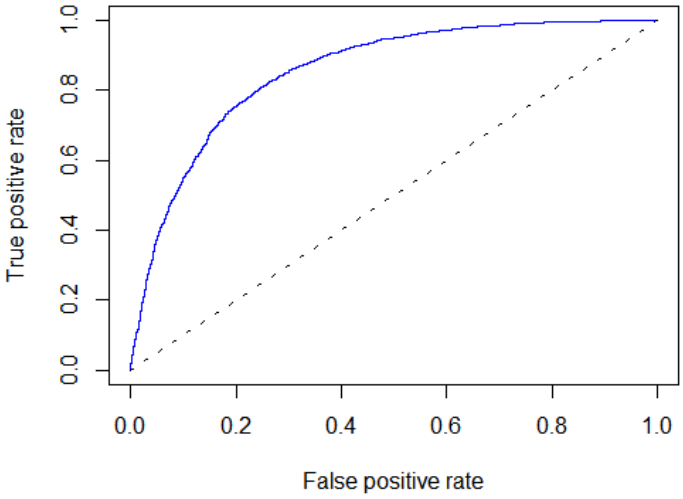


Figure 4.4: Example of a ROC curve

A way of interpreting the ROC curve is to use the area under the curve (AUC) as a measure of performance. AUC is equivalent to the probability that the classier will give a higher score to a randomly chosen observation of class 1 than to a randomly chosen observation of class 0 (Hanley and McNeil, 1982). Fawcett (2006) notes that the AUC performs very well in practice as a general measure of classifier performance.

4.8.3 Alternative metrics

Demler et al. (2012), based on numerical simulations and supported by a theoretical explanation, concluded that the test proposed by DeLong for comparison of two AUCs should not be extended

to compare two nested AUCs for models developed and validated on the same data. This misuse of the DeLong test leads to a conservative test with low power for a few practical scenarios. Vickers et al. (2011), Pepe et al. (2013) and Cook (2007) arrived at a similar conclusion.

Two models are nested when one of them, the null model, is a special case of the other, the alternative model. So, in this case, the models where variables are added step by step should be evaluated with caution.

To avoid redundant testing and use of potentially problematic methods for inference, the authors recommend that hypothesis testing for no improvement to be limited to evaluate the added variables as a risk factor, for which methods are well developed and widely available. Analyses of measures of prediction performance should focus on estimation rather than on testing for no improvement in performance (Pepe et al., 2013).

In Cook (2007) a similar point is made regarding the ROC curve. As an alternative, the Bayes information criterion is advised as a good measure to evaluate model fit to complement AUC.

Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) can be seen in Equation 4.4 where LL is the log-likelihood of the model, k is the number of independent parameters, and n is the sample size.

$$BIC = -2LL + k\log(n) \tag{4.4}$$

BIC is a criteria for model selection that measures the trade-off between model fit and complexity. A lower BIC means that a model is considered to be more likely to be the true model. Due to the penalty term depending on the number of independent parameters, BIC tends to favour parsimonious models. According to Raftery (1995), a difference of the BIC value lower than two between two models is barely worth mentioning, a difference between two and five is positive, a difference between five and ten is strong, and a difference larger than ten is very strong.

BIC penalizes model complexity very heavily, so an alternative measure will be considered, the AIC.

Akaike information criterion (AIC)

AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model (see Equation 4.5 where $\hat{L}$ is the maximum value of the likelihood function for the model). A lower AIC means a model is considered to be closer to the truth.

$$AIC = 2k - 2\ln(\hat{L}) \tag{4.5}$$

The only way they should disagree is when AIC chooses a larger model than BIC. In general, it might be best to use AIC and BIC together in model selection.

Chapter 5

Results and discussion

The focus of this thesis is to evaluate the impact of service failure in customer churn, so here the results of the several models are going to be presented, followed by a detailed analysis of the outcomes.

One the main objectives was to study how the different product categories affect customer churn. To accomplish this, two variables were created representing all the service failures of the two main product categories: *FoodSF* and *PerishablesSF*.

On the other hand, another goal was to study the influence that the type of failures could have in customer churn. When it comes to types of failures the main objective is to detect differences between *Substitutions* and *Stockouts*, so only these two variables are being considered.

Finally, to increase the specificity of the analysis, the combination of the above variables is considered in one last model. So, in this last stage, four different variables are being considered: *FoodSubstitutions*, *FoodStockouts*, *PerishSubstitutions* and *PerishStockouts*. With this analysis the objective is to detect differences in types of failures across different categories.

To properly evaluate the performance of these three sets of variables, and provide robust conclusions they are going to be reproduced in three different main models:

- **1.0 Tamaddoni et al (2016):** firstly, the reproduction of Tamaddoni et al. (2016) will be done for this case study to check if we can separate churners from non-churners effectively. Then step by step the service level variables will be added to assess what is the best approach, according to the regression coefficients and evaluation metrics;

- **2.0 Base Model:** a posterior analysis of the regression coefficients, together with some insights from the business, led to some changes in the independent variables of Model 1.0. In addition, some new independent variables were also introduced, in an attempt to increase model performance and fit, this lead to the creation of a Base Model inherent to this case study;

- **3.0 RFM:** a similar procedure will then be performed on a RFM model, where, in order to corroborate the results, we will focus on the main variables responsible for predicting churn so that the stock related variables are not influenced by possible correlations.

This division can be better visualized in Figure 5.1, all these models were performed in LR. It should be noted that in Figure 5.1 SV stands for these four stock related variables.

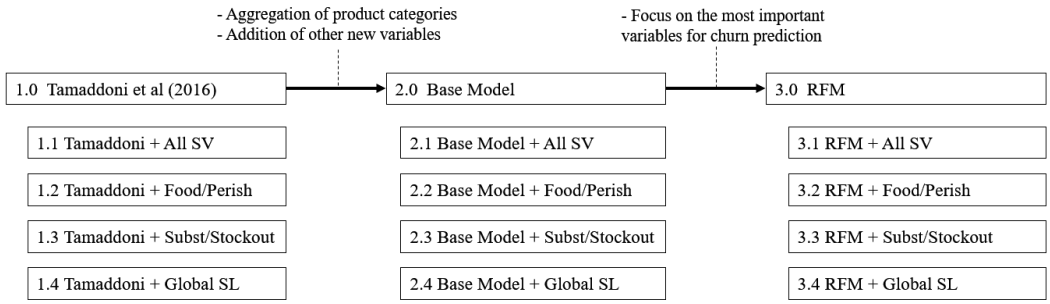


Figure 5.1: Diagram of the different models created with LR

Lastly a GBM Model was created with the same variables as Model 2.1, to corroborate previous conclusions.

The evaluation metrics presented in this chapter are computed on the testing set. These values were checked against the cross-validation results which were similar, indicating that performance of the model is consistent.

5.1 Model construction

In this first section, the focus will be on comparing the performance metrics from each of the three main models, and the contribution of the stock related variables for each one.

5.1.1 Reproduction of Tamaddoni et al. (2016) (Model 1)

The introduction of the stock variables and the comparison to the standard model (Tamaddoni et al., 2016) in this database is represented in Table 5.1. It can be seen that stock variables only lower the BIC value when all the variables that enter the model are statistically significant (refer to Table 5.2). When some of the added variables are not significant, as it happens for the perishables in both Models 1.1 and 1.2, the BIC increases substantially (check LR coefficients in Table 5.2).

Table 5.1: Performance measures on reproduction and improvement of predictive model

	AIC	BIC	AUC
1.0 Tamaddoni	20 486	20 655	0.8588
1.1 Tamaddoni + All Stock variables	20 480	20 681	0.8590
1.2 Tamaddoni + Food/Perishables	20 476	20 661	0.8590
1.3 Tamaddoni + Stockouts/Substitutions	20 459	20 644	0.8543
1.4 Tamaddoni + Global Service Level	20 463	20 640	0.8541

This is due to the high penalization that this metric gives to an increase in complexity of the model. Still, it should be noted that when the relevant stock variables were introduced, the BIC

value reduced substantially. Regarding the AIC, where the increase in number of variables is not as penalized, the introduction of the stock related variables always reduces its value. The AUC values only suffered minor changes, as it was expected considering these are nested models.

Table 5.2: Coefficients regarding Model 1: Tamaddoni et al. (2016)

	Model 1.0	Model 1.1	Model 1.2	Model 1.3
Intercept	-1.1898***	- 1.1901***	- 1.1899***	-1.1912***
TimeBetweenFirstLastPurchase	- 0.9307***	- 0.9245***	- 0.9245***	- 0.9189***
NumberOrders	- 0.6998***	- 0.7052***	- 0.7050***	- 0.7148***
FoodSKUs	- 0.3083***	- 0.3088***	- 0.3091***	- 0.3092***
LoyaltyCard	- 0.1770***	- 0.1773***	- 0.1776***	- 0.1776***
TimeSinceFirstPurchase	+ 0.1142***	+ 0.1127***	+ 0.1127***	+ 0.1129***
FoodSpending	- 0.0937*	- 0.0894.	- 0.892.	- 0.3092***
SocioEconomicStatus	- 0.0959***	- 0.0916***	- 0.0916***	- 0.0941***
PerishablesSKUs	- 0.0711	- 0.0709*	- 0.0710	- 0.0709
HouseSKUs	+ 0.0402	+ 0.0410	0.0410	+ 0.0458
HygeneSKUs	- 0.0244	- 0.0240	- 0.0239	- 0.0252
ClothesSpending	+ 0.0183	+ 0.0164	+ 0.0164	+ 0.0111
DeliveryIncidences	+ 0.0155	+ 0.0138	0.0139	+ 0.0138
BakerySpending	- 0.0097	- 0.0104	- 0.0012	- 0.0102
HouseSpending	+ 0.0089	+ 0.0086	0.0085	+ 0.0096
BakerySKUs	- 0.0061	- 0.0055	-0.0054	- 0.0063
HygeneSpending	- 0.0024	- 0.0028	-0.0026	- 0.0032
PerishablesSpending	- 0.0010	- 0.0709	- 0.0012	- 0.0014
ClothesSKUs	+ 0.0100	+ 0.0114	0.0116	+ 0.0111
BazaarSKUs	- 0.0194	- 0.0187	- 0.0189	- 0.0188
FoodStockouts		+ 0.0473**		
FoodSubstitutions		+ 0.0359*		
PerishablesSubstitutions		+ 0.0234		
PerishablesStockouts		+ 0.0219		
FoodFailures			+ 0.0527**	
PerishablesFailures			+ 0.0297.	
Stockouts				+ 0.0851***
Substitutions				+ 0.0619***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

Performing an analysis on the standardized coefficients from Model 1.0, some conclusions were in line with Tamaddoni et al. (2016) (see variable importance in Table 5.2). As expected, RFM variables are responsible for most of the predictive power, but variables such as *LoyaltyCardMembership* and some product categories were also considered important when predicting churn. Yet, in contrast to what was found in Tamaddoni et al. (2016) *SocioEconomicStatus* in this dataset was assigned as statistically significant. This may be due to different proxies used to assess the economic status. Additionally, *DeliveryIncidences* was not found to be statistically significant in this dataset, in opposition to Tamaddoni et al. (2016).

It should also be noted that, it is clear that the separation of quantity and spending by categories is not relevant to the model. In fact, only some of the Food and Perishables variables have statistically significant coefficients, meaning that the other variables increase the complexity of the model unnecessarily (refer to Table 5.2). Besides, this separation between quantity and spending leads to highly correlated variables which are not advised in LR. This led to some changes to the current model, along with the addition of some other new variables.

5.1.2 Introduction of new variables specific to this case study (Model 2)

Since in this case study the separation of spending for all categories did not seem to provide good predictors to the model a different approach was idealized. By aggregating all the categories into one variable, the complexity of the model was reduced, while not compromising much on predictive performance. This also allowed to introduce new variables that could be derived from this case study like brand purchasing behaviour, delivery subscription, promotional campaigns and delivery cost.

As it can be seen in Table 5.3, this new approach lowers AIC and BIC by more than 250 units as well as slightly increase AUC. Following what was previously mentioned, comparisons for nested models solely based on the AUC can be very misleading, but here a clear pattern can be seen. The ROC curve for Model 2.1 is presented in Figure 5.2.

Table 5.3: Changes to the Tamaddoni model and improvement in performance

	AIC	BIC	AUC
1.1 Tamaddoni + All Stock variables	20 480	20 681	0.8590
2.0 Base Model	20 204	20 317	0.8604
2.1 Base Model + All Stock variables	20 201	20 345	0.8606
2.2 Base Model + Food/Perishables	20 198	20 326	0.8606
2.3 Base Model + Stockouts/Substitutions	20 185	21 314	0.8608
2.4 Base Model + Global Service Level	20 190	20 311	0.8607

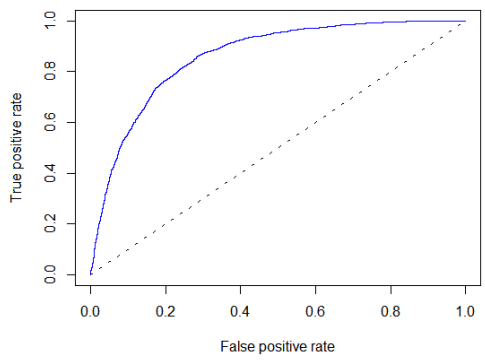


Figure 5.2: Roc curve for Model 2.1

Some other relevant evaluation metrics for these models are accuracy, recall and specificity which can be calculated from the confusion matrix in Table 5.4. We will focus on Model 2.1, to have a better idea on how the Base Model is performing (where all the stock variables are discriminated). While accuracy had a value of 0.8026, it should be considered with caution since the dependent variable is slightly imbalanced. Recall measures how good a test is at detecting the non-churners and had a corresponding value of 0.9050. On the other hand, Specificity evaluates how good the test is at detecting churners and had a value of 0.5585. From this analysis we can conclude that the model performs better at predicting non-churners than churners, but for our work this does not have a significant impact because it focuses on separating the classes. The metric commonly used for evaluating class separation is the AUC, which can be seen in Table 5.3 for Model 2.1. An approach with balanced classes was also performed but it did not improve significantly the previous results, probably due to the mild imbalance experienced.

Table 5.4: Confusion matrix for model 2.1

		Real class	
		Non Churner	Churner
Predicted	Non Churner	5011	1025
	Churner	526	1297

Looking at the coefficients regarding the new variables added to the model (see Table A.1) some remarks should be mentioned. By descending order of importance: *DeliverySubscription*, *DeliveryCost* and *SupplierBrand* all significantly contribute to explaining customer churn. The negative sign associated with these variables indicates that an increase in one of these parameters decreases the churn probability. It is interesting to note that an increase in delivery cost decreases the churn probability, but this should be related to the fact that customers that are willing to pay more for delivery (larger or more convenient delivery slots) are less likely to churn. Also, the positive sign of *TimeSinceFirstPurchase* indicates that older customers are more likely to churn. As it was expected the delivery subscription variables had a strong and negative impact on churn, since customers that are willing to pay for this service are considered to be loyal. *SupplierBrand* also negatively impacted churn since customers that tend to buy these more expensive products are more likely to use an online service such as this one. The stock related coefficients from these new models will be discussed in the following section.

5.1.3 Introduction of stock related variables on a simpler model (Model 3)

Lastly, a simpler model based only on the RFM variables will be assessed, with the goal of validating previous conclusions. RFM variables, as it can be seen in Table A.1 have the highest standardized predictors and are highly regarded in the literature as the best churn predictors. In line to what was found in Table 5.1, when introducing the stock variables to the standard RFM model, the same conclusions were attained (Table 5.5). Stock variables only lower the BIC value when all the variables that enter the model are statistically significant (check variable importance in Appendix B). This indicates that there can be significant differences on how different stock

variables affect customer churn. With this in mind, the following sections do a more detailed analysis regarding the different types of stock variables available.

Table 5.5: Introduction of stock variables into standard RFM model

	AIC	BIC	AUC
3.0 RFM	20 546	20 586	0.8587
3.1 RFM + All Stock variables	20 535	20 607	0.8590
3.2 RFM + Food/Perishables	20 532	20 588	0.8589
3.3 RFM + Stockouts/Substitutions	20 516	20 572	0.8593
3.4 RFM + Global Service Level	20 522	20 570	0.8592

5.2 The impact of stock related variables

Now that the relevance of the models is set, first we will focus on studying the impact of the service level of the two main product categories (Food and Perishables) in customer churn. Then, we will check if substitutions and stockouts impact customer churn differently. Beyond the models from the previous section, here the GBM model will also be analysed.

5.2.1 Impact of the main product categories' service level on customer churn

The focus here will be on the models represented in Figure 5.3, since here the picking performance is discriminated by category. Additionally a GBM model will be presented in the end to corroborate the results. By looking at the ranked standardized coefficients of these two categories (Food and Perishables), their importance impacting customer churn can be depicted.

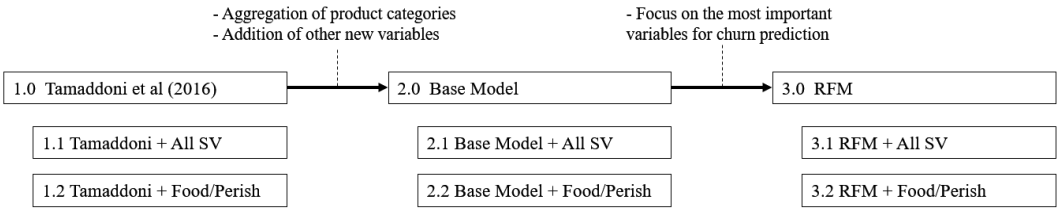


Figure 5.3: Diagram of the different models created regarding the product category division

In Model 2.1, where both Food and Perishables are also discriminated by type of failure (Substitutions and Stockouts), a clear distinction between the rank of the variables is present. In Table A.1 (Appendix), it can be seen that both Food related variables have higher coefficients than the Perishables' variables. Also, *FoodSubstitutions* is statistically significant with a p-value lower than 0.05, and for *FoodStockouts* the p-value was only lower than 0.1. On the other hand, both Perishables variables were not significant to the model.

In Model 2.2, the relative importance of Food over Perishables was confirmed, with the aggregated variables achieving similar results. Once again, the Food category had a higher coefficient

and was statistically significant with a p-value below 0.01, while the Perishables variable had only a p-value lower than 0.1 (refer to Table A.1 in Appendix).

Since inferring variable importance by looking just at these two models could be ambiguous, several other models were built to increase robustness. Concentrating on Models 3, the RFM (Recency, Frequency and Monetary) variables were isolated along with the stock related variables. Firstly, as it can be seen in Table B.1 (Appendix), Model 3.1 achieved an even larger distinction between these two categories than Model 2.1. It reported the two Food related variables as significant (p-value$<$0.01), while the Perishables variables were almost not significant to the model. So, it is expected that when considering the aggregated variables, similar conclusions are achieved. In Model 3.3, *FoodFailures* clearly stands outs, having a larger coefficient and being statistically significant with a p-value lower than 0.001 while *PerishablesFailures* was almost non significant to the model. Hence, it can be argued that stock shortages regarding the Food category have more impact in the customer decision to churn than the Perishables category.

In Models 1.1 and 1.2, used for the reproduction of Tamaddoni et al. (2016) and represented in Table 5.2, these previous conclusions were corroborated. The Food related variables seen in this table always have larger coefficients and are more significant to the models than the Perishables variables.

Following the methodology stated, a different approach using GBM is employed to the same variables as in Model 2.1. A total of 123 GBM models were created, due to the limited processing capability (the AUC for the top 5 scoring models can be seen in Table C.2 in Appendix). The models were ranked per AUC and the highest scoring GBM achieved a value of 0.8665, which is higher than the previous LR models (see detailed hyperparameters in Table C.1 in Appendix). This is in line with the observations of Tamaddoni et al. (2016), where boosting also scored slightly higher. Looking at the ranked variable importance function provided by the package, it is possible to assess predictors relative importance. As it can be seen in Figure 5.4 in all the top five ranked GBM models, the findings of LR were validated. The Food category variables (seen in dark blue in Figure 5.4) were always ranked above the Perishables (represented in light blue), confirming previous results and increasing the robustness of this analysis. Here it is also visible how the RFM variables are consistently best predictors. The performance metrics for the top five GBM models sorted by AUC can be seen in Table C.2.

Since as it can be seen in Figure 3.3, the distribution of sales in the two biggest product categories is unbalanced, with the Food category having more sales than Perishables, the concern that this could be influencing the results came up. Therefore, a new variable was added to the analysis, representing the ratio of spending in Perishables over Food for each customer. The interaction of this new variable with all the stock related variables was assessed, to check if this imbalance could influence previous conclusions. All the interaction coefficients in Models 2.1 and 2.2 where the category distinction is present in the stock variables proved to be non significant. This implies that the difference in spending between categories does not significantly affect previous conclusions.

In summary, in this section it was found that the Food category consistently has more impact in the customer decision to defect than the Perishables category. This was corroborated by the

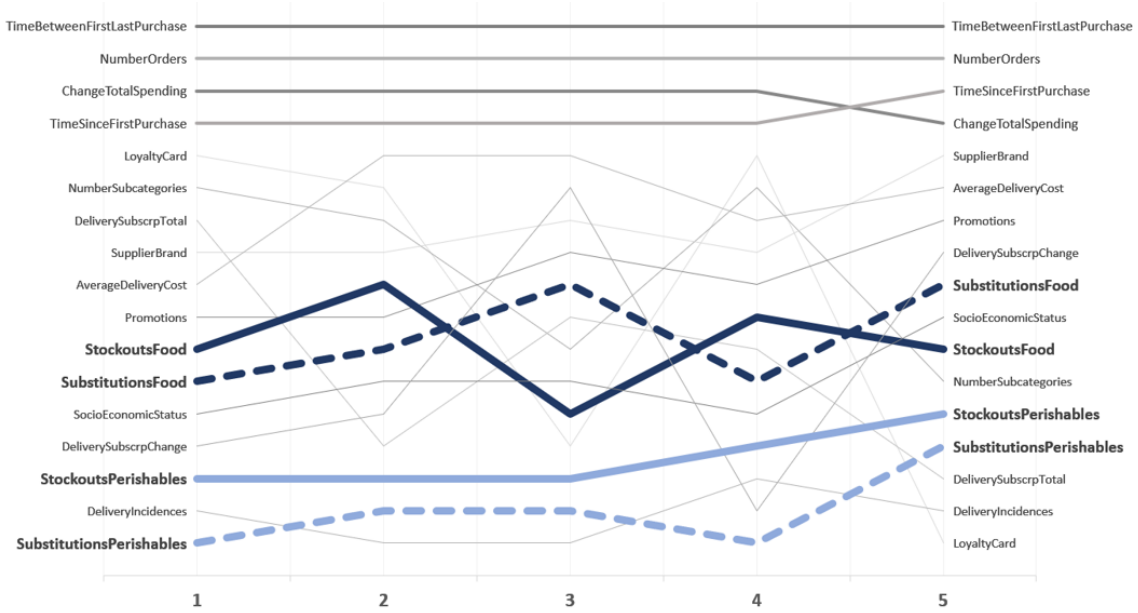


Figure 5.4: Ranked variable importance (descending order) across the top five scoring GBM models

several models executed with different predictors and methods. From a business perspective, it is understandable that customers are more sensitive to shortages in categories were they are not expecting this to happen (Food category). On the other hand, as Perishables are known for being more difficult to handle, customers seem to have a much higher tolerance for failures in this category. Thus, a focus from management in the fulfilment of the Food category should have a larger impact on the customer churn decision, than the Perishables category.

5.2.2 Impact of stockouts and substitutions on customer churn

A similar methodology was implemented for the distinction between *Stockouts* and *Substitutions*. This time focusing on the models in Figure 5.5 where these types of failures are discriminated. As in the previous subsection, the same GBM model will also be used to corroborate the conclusion here.

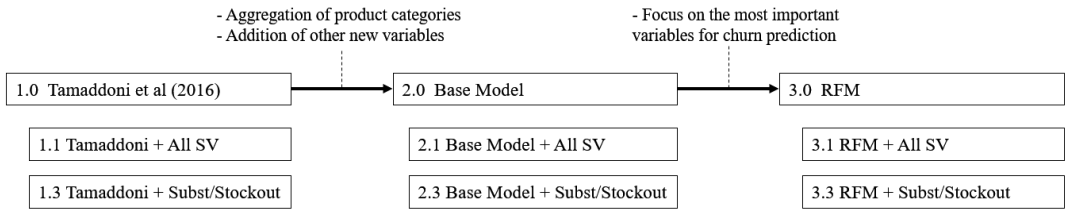


Figure 5.5: Diagram of the different models created regarding the type of failure

In this case, the distinction between stockouts and substitutions in Model 2.1 is not clear (Table A.1). In both product categories, substitutions are ranked ahead of stockouts, but for the

Food category the difference was marginal. In Model 2.3, all the newly engineered variables were included along the variables *Substitutions* and *Stockouts*. Both these variables were found to be statistically significant with a p-value below 0.001. The coefficient for *Stockouts* was slightly higher, but the difference being very small.

From the RFM Model 3 (Table B.1) very similar conclusions as in Model 2 were achieved, but with Model 3.3 giving a slight edge to substitutions. When analysing the Models 1, used for the reproduction of Tamaddoni et al. (2016) and represented in Table 5.2, it can be seen that stockouts are usually ranked slighly higher, with the exception for the perishables category.

Hence, for substitutions and stockouts the distinction is not as clear as for the categories in the previous section, with very marginal differences. Nonetheless, stockouts seems to be usually higher ranked than substitutions.

To improve the confidence of the conclusions, the exact same GBM model as in the previous subsection is used, but with the focus now on failure types. In Figure 5.4 stockouts are highlighted as the continuous lines and substitutions are the dashed lines. The failure types do not have consistency throughout the models. In the Perishables category stockouts are always ranked above, but in the Food category this is not as clear with two cases where that does not happen.

Due to some discrepancies in the results, most likely influenced by the category, no firm conclusions can be drawn regarding failure types. On the one hand, in both the LRs and GBMs stockouts seem to have slightly more impact on churn than substitutions. However, there are some exceptions in both cases, resulting in no global robust conclusion regarding failure types.

From a management perspective, this seems to indicate that the current policies of substituting products are reducing the impact of stockouts on churn for this case study. These conclusions should be taken with caution though, due to some inconsistencies in the results.

Chapter 6

Conclusions

This work addressed the problem of customer defection applied to the evolving online grocery retail market. This is a growing sector with new characteristics that can be exploited to improve operations management and decrease customer churn. More specifically, the fact that customers do not see actual inventory when buying a product and the delay between purchase and delivery, allows companies to capture unfulfilled demand.

This new reality in grocery retail was the main motivation behind this thesis. Here, we intend to capture how the availability of different product categories can influence customer churn and understand if the current policy of substitutions is indeed valuable to mitigate the effect of stockouts.

After predicting customer churn with LR and GBMs several insights could be taken. As it was expected from the literature available, boosting performed slightly better with an AUC of 0.8665 (compared to 0.8606 for the LR). The introduction of new variables specific to this dataset proved to be helpful in modelling churn, besides the common Recency, Frequency and Monetary variables. From the new variables introduced, delivery subscription service and the brand purchase behaviour were among the most important. The average delivery cost also had a big relevance, indicating that a higher delivery cost reduces churn probability, which can be explained by the way that the delivery cost is differentiated for deliveries. Customers that are willing to pay more, according to the model are more loyal and less likely to churn.

From the variables employed in this case-study, stock related variables are the ones that can be more easily targeted by the company to affect customer experience and consequently churn behaviour. The two techniques employed confirmed that the Food product category was steadily ranked above the perishables category, reflecting that the former has a higher importance regarding the customer churn decision. The Perishables category, known for having reduced shelf life, consequently has a lower average service level than the Food category (0.86 and 0.93 respectively). This may be the reason why customers appear to be more sensitive to Food rather than the Perishables category. According to the results achieved, a focus from management in the service level of the Food category should have a larger impact on customer experience and churn behaviour.

Regarding stockouts and substitutions the difference between the two was not has consistent

across methods as the previously mentioned categories. Even though, a general trend could be seen were stockouts are more important than substitutions (with a few exceptions) regarding the customer decision to churn. This implies that the current policies regarding product substitutions are being effective, with the substitutions made lowering the general impact of stockouts in customer churn.

For future work, this analysis can be further investigated employing different metrics and even using different approaches. The first obstacle was to define customer churn in online grocery retail. For this, based on the literature several criteria were set, but always using a global approach. Maybe applying a criterion that considers individual purchase history and for each customer makes a customized decision, could be an interesting approach. Also, instead of focusing on the big spectrum of clients, maybe doing segmentation and clustering of customer groups could provide different insights, having a more targeted approach. Additionally, an approach using purchase behaviour should be considered, since it can provide a more in depth look at customer behaviour than churn, which is a binary outcome. Several other data mining techniques can also be employed to this problem in order to further investigate the relation between churn and the predictor variables. On the other hand, instead of focusing on churn, the impact that the unfulfilled demand can have on customer experience should be further studied with customer lifetime value. This highly studied metric in the literature should give another insight on how these variables impact future earnings, and even calculate what is the cost of having different service levels. With this customer lifetime value approach, the option to reserve stock for the more profitable or high-risk customers could also be evaluated, as picking is currently done in physical stores.

Bibliography

Ahn, J.-H., Han, S.-P., and Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry. *Telecommunications Policy*, 30(10):552–568.

Bhalla, D. (2016). Gbm (boosted models) tuning parameters.

Buckinx, W. and Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164(1):252–268.

Chen, K., Hu, Y.-H., and Hsieh, Y.-C. (2015). Predicting customer churn from valuable b2b customers in the logistics industry: a case study. *Information Systems and e-Business Management*, 13(3):475–494.

Consulting, A. (1996). Where to look for incremental sales gains: the retail problem of out-of-stock merchandise. *The Coca-Cola Retailing Research Council, Atlanta, GA*.

Contractors, I. (2018). Will online grocery shopping be the end of brick and mortar stores?

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935.

Coussement, K., Lessmann, S., and Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction. *Decision Support Systems*, 95(C):27–36.

De Caigny, A., Coussement, K., and De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2):760–772.

Demler, O. V., Pencina, M. J., and D'Agostino, R. B. (2012). Misuse of delong test to compare aucs for nested models. *Statistics in Medicine*, 31(23):2577–2587.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., and Leitão, P. J. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Fader, P. S. and Hardie, B. G. S. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing*, 23(1):61–69.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Ferreirinha, C. A. G. (2014). *Modelos logísticos de alargamento geográfico de entregas no formato Click and Pick: o caso do E-commerce da Sonae MC*. Thesis.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378.

Friedman, J. H. and Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, 22(9):1365–1381.

Galante, N., López, E. G., and Monroe, S. (2013). The future of online grocery in europe. *McKinsey & Company*, pages 22–31.

Goran, K., Robert, K., and Leo, M. (2015). *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. IGI Global, Hershey, PA, USA.

Gustafsson, A., Johnson, M. D., and Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing*, 69(4):210–218.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Jing, X. and Lewis, M. (2011). Stockouts in online retailing. *Journal of Marketing Research*, 48(2):342–354.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P. C., Wedel, M., and Wilcox, R. (2005). Choice models and customer relationship management. *Marketing Letters*, 16(3):279–291.

Malohlava, M. and Candel, A. (2017). Gradient boosting machine with h2o.

Mohammad, F., Yaser, H., and Behrouz, M.-B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. *Kybernetes*, 45(5):732–743.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(21).

Nathans, L. L., Oswald, F. L., and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research Evaluation*, 17(9).

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2):204–211.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

Nielsen (2018). Developing your omnichannel collaboration model.

Oliveira, V. L. M. (2012). *Analytical customer relationship management in retailing supported by data mining techniques*. Thesis.

Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in Medicine*, 32(9):1467–1482.

PORDATA (2015). Poder de compra per capita.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, pages 111–163.

Rao, S., Griffis, S. E., and Goldsby, T. J. (2011). Failure to deliver? linking online order fulfillment glitches with future purchase behavior. *Journal of Operations Management*, 29(7-8):692–703.

Smith, C. (2013). The rise of "click and collect" purchases.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., and Snoeck, M. (2017). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*.

Tamaddoni, A., Sepehri, M. M., Teimourpour, B., and Choobdar, S. (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *Journal of Strategic Marketing*, 18(7):587–598.

Tamaddoni, A., Stakhovych, S., and Ewing, M. (2014). Managing b2b customer churn, retention and profitability. *Industrial Marketing Management*, 43(7):1258–1268.

Tamaddoni, A., Stakhovych, S., and Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: A contingent perspective. *Journal of Service Research*, 19(2):123–141.

Thompson, D. (2009). Ranking predictors in logistic regression. *Paper D10-2009. Online available at http://www. mwsug. org/proceedings/2009/stats/MWSUG-2009-D10.*

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364.

Vickers, A. J., Cronin, A. M., and Begg, C. B. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC medical research methodology*, 11(1):13.

Zhang, J. and Krishnamurthi, L. (2004). Customizing promotions in online stores. *Marketing Science*, 23(4):561–578.

Appendix A

Variables coefficients' from Model 2

Table A.1: Coefficients regarding Model 2: Base Model

	Model 2.0	Model 2.1	Model 2.2	Model 2.3
Intercept	- 1.2238***	- 1.2270***	- 1.2267***	- 1.2294***
TimeBetweenFirstLastPurchase	- 0.8961***	- 0.8906***	- 0.8909***	- 0.8862***
NumberOrders	- 0.8756***	- 0.8757***	- 0.8760***	- 0.8791***
ChangeTotalSpending	- 0.3743***	- 0.3709***	- 0.3707***	- 0.3688***
DeliverySubscriptionChange	- 0.3134***	- 0.3139***	- 0.3138***	- 0.3145***
LoyaltyCard	- 0.2389***	- 0.2339***	- 0.2344***	- 0.2309***
DeliveryCost	- 0.1777***	- 0.1768***	- 0.1759***	- 0.1742***
TimeSinceFirstPurchase	+ 0.1497***	+ 0.1484***	+ 0.1486***	+ 0.1490***
SuppierBrand	- 0.1063***	- 0.1088***	- 0.1091***	- 0.1059***
DeliverySubsTotal	- 0.0944*	- 0.0941*	- 0.0936*	- 0.0915*
SocioEconomicStatus	-0.0708***	- 0.0659**	- 0.0683***	- 0.0682***
Promotions	+ 0.0656***	+ 0.0567**	+ 0.0557**	+ 0.0538**
NumberSubcategories	- 0.0508*	- 0.0547*	- 0.0545*	- 0.0609*
DeliveryIncidences	+ 0.0202	- 0.0186	+ 0.0191	+ 0.0187
FoodSubstitutions		- 0.0362*		
FoodStockouts		- 0.0353.		
PerishablesSubstitutions		- 0.0272		
PerishablesStockouts		- 0.0173		
FoodFailures			+ 0.0449**	
PerishablesFailures			+ 0.0276.	
Stockouts				+ 0.0674***
Substitutions				+ 0.0629***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

Appendix B

Variables coefficients' from Model 3

Table B.1: Coefficients regarding Model 3: RFM

	Model 3.0	Model 3.1	Model 3.2	Model 3.3
Intercept	- 1.3099***	- 1.3104	- 1.3106***	- 1.3113***
TimeBetweenFirstLastPurchase	- 0.9771***	- 0.9693	- 0.9698***	- 0.9647***
NumberOrders	- 0.7567***	- 0.7622	- 0.7633***	- 0.7709***
TimeSinceFirstPurchase	+ 0.1461***	+ 0.1443	+ 0.1440***	+ 0.1449***
ChangeTotalSpending	- 0.4331***	- 0.4290	- 0.4286***	- 0.4267***
FoodSubstitutions		+ 0.0519**		
FoodStockouts		+ 0.0463**		
PerishablesSubstitutions		+ 0.0280.		
PerishablesStockouts		+ 0.0181		
FoodFailures			+ 0.0613***	
PerishablesFailures			+ 0.0286.	
Substitutions				+ 0.0782***
Stockouts				+ 0.0759***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

Appendix C

Evaluation and Hyperparameters for GBM models

Table C.1: Hyperparameters for the top AUC scoring GBM model

Hyperparameter	Value
Score tree interval	5
Number of trees	278
Maximum depth	8
Minimum rows	10
Learning rate	0.005
Row sample rate	0.6
Column sample rate	0.4
Column sample rate per tree	1
Minimum split improvement	0.0001

Table C.2: Performance metrics for the GBM models

	AUC	Accuracy	Recall	Specificity
Model 1	0.8665	0.7960	0.8987	0.6288
Model 2	0.8664	0.7853	0.9002	0.6040
Model 3	0.8662	0.7964	0.8988	0.6271
Model 4	0.8660	0.7894	0.8978	0.6144
Model 5	0.8660	0.7948	0.9044	0.6222