University of Warsaw
Faculty of Economic Sciences

Kamil Matuszelański
Student's book no.: 387078

Modeling customer churn in the context of e-commerce retail business

Second cycle degree thesis
specialty: Data Science and Business Analytics

The thesis written under the supervision of
dr. hab. Katarzyna Kopczewska, prof.ucz.
Department of Statistics and Econometrics
WNE UW

Warsaw, August 2021

Declaration of the supervisor

I declare the following thesis project was written under my supervision and I state that the project meets all submission criteria for the procedure of academic degree award.

Date Signature of the Supervisor

Declaration of the author (authors) of the project

Aware of legal responsibility, I declare that I am the sole author of the following thesis project and that the project I submit is entirely free from any content that constitutes copyright infringement or has been acquired contrary to applicable laws and regulations.

I also declare that the below project has never been subject of degree-awarding procedures in any school of higher education.

Moreover I declare that the attached version of the thesis project is identical with the enclosed electronic version.

Date Signature of the Author

Summary

This study is an attempt to propose a model for customer churn prediction in the case of an e-commerce retail store operating in Brazil. Among the sets of features included in the models were transaction, location, geodemographic and perception variables. The results show that transaction features from the previous purchase are the main factor driving the customers' loyalty

Key words

churn analysis, customer relationship management, topic modeling, geodemographics

Area of study (codes according to Erasmus Subject Area Codes List)
Economics (14300)

The title of the thesis in Polish

Modelowanie odpływu klientów w biznesie e-commerce

TABLE OF CONTENTS

CHAPTER I. Introduction…….. 2
CHAPTER II. Literature review……. 4
2.1. Customer churn…… 4
2.2. Explainable Artificial Intelligence…….. 8
CHAPTER III. Dataset description…… 10
3.1. Data sources…… 10
3.2. Quantitative analysis……... 12
CHAPTER VI. Methods description…….. 20
4.1. Modeling methods…… 20
4.2. Features preprocessing……. 21
4.3. Variables selection methods……. 31
CHAPTER V. Results……. 35
5.1. Results of the pre-modeling phase……... 35
5.2. Performance analysis……... 36
5.3. Model's working explanations……. 42
CHAPTER VI. Summary…… 52
APPENDIXES…… 53
Appendix A - Spatial join of census data to the main dataset……. 53
Appendix B - reviews topics……... 54
Appendix C - table of lift values for selected quantiles…….. 59
REFERENCES…… 61

CHAPTER I

Introduction

Maintaining high customer loyalty is a challenge that most of the businesses face. Multiple studies have shown that retaining customers is more profitable than acquiring new ones. In Customer Relationship Management (CRM) field, churn prediction is a very active area of research. Most of the previous studies were conducted in the industries operating in contractual setting, where the churn rate is not that big, for example telecom or banking.

This study is aimed at predicting loyalty of the customers of an e-commerce retail shop operating on the Brazilian market. A challenge that to the best of the author's knowledge was not addressed in the previous studies is churn prediction in a industry, in which a very low rate of the customers stay with the company. In the case of this study almost 97% of the customers don't decide to make a second purchase. Also, this study is a first approach to predict customer loyalty not using a rich customer's history, but only the first transaction. Because of that, usage of non-conventional data sources was needed. That is why, in this study besides transaction data about the customer, I also analyze customer's geodemographic environment obtained from census data, as well as information from customer's perception about the purchase in form of textual reviews.

From a technical point of view, I have used a Machine Learning approach. I have tested out 2 classification algorithms, namely XGBoost and Logistic Regression. To obtain a meaningful information from the text reviews, I have used topic modeling technique. To be able to answer hypotheses about the influence of variables on the target in the case of XGBoost modeling, which is an unexplainable, black-box model, I have made extensive use of XAI techniques. To assess the importance of particular sets of features I have used Permutation Variable Importance, while for assessing the strength and direction of the influence - Partial Dependence Profile technique.

CHAPTER II

Literature review

Introduction

In the first section of this chapter a literature review of previous studies regarding customer loyalty churn prediction is presented. The second section describes shortly the field of Explainable Artificial Intelligence, and advantages of usage of such approach in Machine Learning modeling.

2.1 Customer churn

Customer Relationship Management is defined as a process, in which the business manages its interactions with the customers using data integration from various sources and data analysis (Bardicchia 2020). Oliveira (2012) specifies 4 areas in which CRM approaches can be of use and what questions do they aim to answer:

- Customer identification (acquisition) - who can be a potential customer?
- Customer attraction - how can one make this person a customer?
- Customer development - how can one make a customer more profitable?
- Customer retention - how can one make the customer stay with the company?

The last one is the main focus of this study.

Improving the loyalty of the customer base is profitable to the company. This has its source in multiple factors, the most important one being the cost of acquisition. Multiple studies have shown that retaining customers costs less than attracting new ones (Dick and Basu 1994; Gefen 2002; Buckinx and Poel 2005). Moreover, there is some evidence that loyal customers are less sensitive to the competitor's actions regarding price changes (Achrol and Kotler 1999; Choi et al. 2006).

There are 2 basic approaches for the company to deal with customer churn. The first one is an "untargeted" approach. The company seeks to improve its product quality and relies on mass advertising to reduce the churn. The other way is a "targeted" approach - the company tries to address aim their marketing campaigns at the customers that are more likely to churn (Burez and

Poel 2007). This approach can be divided further, by how the targeted customers are chosen. The company can target only those that have already decided to resign from a further relationship. For example, in contractual settings, this can mean canceling the subscription or breaching the contract. The other way to approach the churn problem is to try to predict, which customers are likely to churn soon. This has the advantage of having lower cost, as the customers that are about to leave are likely to have high demands from the last-minute deal proposed to them (Tamaddoni Jahromi et al. 2010).

As pointed out by Tamaddoni Jahromi et al. (2010) in their literature review, most of the studies concerning churn prediction were done in contractual settings. In other words, churn was defined as the client resigning from using the company's services by canceling the subscription or breaching the contract. Such a way to specify the churn is different from the businesses in which the customer doesn't have to inform the company about resigning.

One problem that arises in the non-contractual setting is the definition of churn. As there is no clear moment that the customer decides not to use the company's services anymore, it has to be specified by the researcher based on the goals that one has to achieve from the churn analysis. Oliveira (2012) defined partial churners as the customers not making new purchases in the retail shop for the next 3 months. A different approach was used by Buckinx and Poel (2005). All the customers that had the frequency of purchases below average were treated as "churners" since these customers were shown to provide little value to the company.

Customer churn prediction

If the company can successfully predict, which customers are most likely to leave, it can target them with a retention-focused campaign. Contrary to targeting all of the customers with such a campaign, focusing on the customers that are most likely to leave leads to a reduction of the cost of the campaign.

Churn prediction fits well with the framework of classification, as the variable that one would like to predict is binary (churn-no churn). However, not only such binary prediction is valuable for later retention campaign efforts. As noted by Wai-Ho Au, Chan, and Xin Yao (2003), equally important is that the machine learning model can predict the likelihood of the

customer leaving. After such prediction, the customers can be ranked from the most to the least likely to churn.

This has two benefits. First, the company can decide what percentage of the customers to target in the retention campaign and is not bound by how many customers the model will predict as potential churners. Second, the company can decide how strong the targeting should be based on the likelihood to leave. For example, based on cost-benefit analysis of various targeting approaches, one could decide that for the top 10% of the most "risky" customers the company should offer big discounts for the next purchase, while for the top 30% - only send an encouraging email.

The churn prediction task can be decomposed into 2 main important aspects that one has to tackle. First is the decision about a specific Machine Learning model that gives the best performance. The second is deciding on the model formula - in other words, deciding about which variables should be included in the model and what should be the form of the relationship.

Machine Learning models for churn prediction

In previous churn prediction studies multiple machine learning algorithms for prediction were tested out (for an overview see Verbeke et al. (2011)). The two most widely used techniques are Logistic Regression (LR) and Decision Trees. An important feature of both of them is that they are relatively simple, and because of that the way they make predictions can be assessed by a qualified expert (Paruelo and Tomasel 1997). However, these two methods often give sub-optimal results compared to more advanced and recent approaches like Neural Networks or Random Forests (Murthy 1998; Oliveira 2012). Moreover, this was shown not only in the case of churn prediction setting but also in more general benchmarks that used multiple datasets and comparison metrics (Caruana and Niculescu-Mizil 2006).

Recently, the XGBoost algorithm (Chen et al. 2015) has been gaining popularity in multiple prediction tasks. XGBoost's main strengths are the ability to infer non-linear relationships from the data, and relative speed, which allows the researcher to try out multiple hyperparameters and decide on the best ones (Chen et al. 2015). Because of that, it is considered a go-to standard for machine learning challenges, and very often solutions based on it achieve the best results in various competitions and benchmarks (hcho3 2020). In the context of churn

prediction, XGBoost was used by Gregory (2018). It achieved superior performance compared to other techniques, specifically Logistic Regression and Random Forests.

Variables used in previous churn prediction studies

Previous churn prediction studies used a variety of variables to include in the model formulation. Buckinx and Poel (2005) divided them into 3 broad categories - behavioral, demographic, and perception.

The first category of variables tries to describe how the customer has interacted with the company before. Typical features belonging to this category are recency, frequency, and monetary value, which constitute the basis of the RFM customer segmentation framework. These features are used in multiple studies (Oliveira 2012; Bhattacharya 1998), and typically accompany more complex variables that are the main focus of particular studies. Besides these basic features, some studies focus on other areas of customer behavior. For instance, dummies indicating particular products that the customer has bought in previous purchases were shown to be a useful predictor for churn prediction (Buckinx and Poel 2005; Athanassopoulos 2000).

The second category of features used in churn prediction constitutes of demographic variables about the customer, such as age, gender or address. Such variables were shown to be good predictors of customer churn in multiple studies (for an overview see Verbeke et al. (2011)). However, the availability of such predictors to use in modeling is very often limited for multiple reasons. In non-contractual settings, customers don't have to always provide such data to the company. Moreover, usage of such personal data can be in some cases considered unethical, and lead to predictions biased against particular age or gender.

Another way to include demographics data in the churn prediction model was shown by Yu Zhao et al. (2005). They successfully used the census data obtained from the statistical office for particular regions that the customer is residing in.

The last category of variables used for churn prediction specified by Buckinx and Poel (2005) is customer perception about the company. According to Kracklauer, Passenheim, and Seifert (2001), customer satisfaction is the most important factor driving customer retention. However, although such features could have potentially high predictive power, they are usually hard to observe and quantify meaningfully. The most widely used approach is asking the

customers for direct feedback using questionnaires or providing a way to post a review on the purchase. This kind of feedback can be obtained in different forms, one of them being textual reviews. A couple of previous studies were aimed at extracting meaningful features from such reviews using different text mining methods. De Caigny et al. (2020) have used text embedding approach, while Suryadi (2020) - simple tf-idf technique. In both studies, the results using such methods were superior compared to the models without including such information.

2.2 Explainable Artificial Intelligence

Introduction

While deciding on the type of Machine Learning algorithm, one usually faces the explainability-performance trade-off (Nanayakkara et al. 2018). More flexible models, like bagging, boosting or neural networks, very often present superior performance to less flexible approaches. On the other hand, their predictions cannot be explained as easily as in the case of for example Decision Trees or Linear Regression.

Explainable Artificial Intelligence (XAI) is a set of tools aimed at explaining predictions of these highly flexible models. This area started gaining popularity among Machine Learning researchers to somehow transfer the advantages of simple models to the approaches that provide superior performance.

Doshi-Velez and Kim (2017) specifies some of the machine learning model's traits that can accompany typical requirement of achieving the best accuracy:

- fairness - whether the algorithm is biased against a particular gender, age, race, etc.
- robustness - whether the algorithm can provide correct predictions when the parameters change
- trust - whether the final users of the algorithm trust the model's predictions

Machine learning practitioners when deciding on the methodology to apply have to assess which of the requirements are important in a particular task. For example, in CRM settings the trust in the model's predictions is way less important than in medical areas, but still can be crucial for a wide adoption of modeling across the company. On the other hand, sometimes the

explainability is important only for the person developing the model, to understand its limitations and be able to improve upon it.

The tools of XAI can help in addressing the aforementioned issues, without losing the usual performance gain from black-box models. For an extensive overview of existing XAI methods, see Biecek and Burzykowski (2021).

XAI in marketing

Research on Explainable Artificial Intelligence in Marketing domain is not very developed. To the best of the author's knowledge, the only study touching the subject of XAI in the context of marketing is by Rai (2020). In their commentary, they specify potential areas for future research in this field:

- understanding, what are acceptable requirements regarding explainability compared to accuracy in different marketing tasks
- making AI trustworthy - to understand how the eagerness to use AI system's predictions grows in the company when various explainability tools are made available to the end-users
- How model explanations should be presented to various groups of system's users. For example, a Machine Learning expert is interested in very detailed and complex explanations, while the company's customer may simply want a 1 sentence summary of what was taken into account while making predictions

CHAPTER III

Dataset description

Introduction

This chapter is aimed at describing the datasets used in this study. In the first section the data sources and avaliable variables are specified, while in the second - an Exploratory Data Analysis is conducted.

3.1 Data sources

In this study, I have used data from 2 sources. The main one is e-commerce store transactions data. Olist company is operating in Brazil, and the dataset was made available online for public use[1]. This dataset was enhanced by census data obtained from the Brazilian Statistical Office[2].

Transaction dataset

The Olist company dataset contains information about 100 thousand orders made on the e-commerce shop site from 2016 to 2018. Besides technical variables indicating keys to join multiple tables from the dataset, it also contains the following features groups:

- payment value - the value of the order in Brazilian Reals
- transportation value
- number of items the customer bought in a particular order
- review of the order - after the finished order the customer can provide the review of the order in 2 forms - 1-5 score or textual review. In the dataset codebook, the authors stated that not all of the customers in real life put any review, but this dataset was sampled in such a way that the records without 1-5 review were excluded. On the contrary, the textual review is filled only in ~50%. The data about 1-5 review can be included in the models as-is. The textual

[1] https://www.kaggle.com/olistbr/brazilian-ecommerce access 14.03.2020

[2] https://sidra.ibge.gov.br/tabela/3548 access 26.09.2020

review requires however more advanced preprocessing, which is described in the chapter *Methods description* of this study.

- location of the customer - the main table containing customer information contains the 5-digit ZIP code of the customer's home. The company provided also a mapping table, in which each ZIP code is assigned to multiple latitude/longitude coordinates. Probably this was done because of anonymization reasons - so that one cannot connect the customer from the dataset with the exact house location. To obtain an exact one-to-one customer-geolocation mapping, to each zip code I have assigned the most central geolocation from the mapping table. To obtain the most central point, I have used the Clustering Around Medoids algorithm with only one cluster and ran the algorithm separately for each ZIP code.
- products bought - the dataset contains information about how many items there were in the package, as well as the product category of each item - in the form of raw text. In total there were 74 categories, but the top 15 accounted for 80% of all the purchases. To limit the number of variables used in the modeling process, I have decided to change the label of all the least popular categories to "others."

The main goal of this study is to try to predict just after the first transaction if the customer is likely to buy for the second time. **In the dataset there were 96180 transactions (96.6%) from the customers that never previously bought in this shop.**

Geodemographic dataset

Demographic statistics were obtained from Instituto Brasileiro de Geografia e Estatística web service. In this study, I have used the data obtained from the 2010 general census. The dataset is available in aggregation to microregions (a Brazilian administrative unit, it has a similar level of aggregation to NUTS 3 European classification). 558 microregions were available. In particular, I have chosen the following 36 variables from the dataset:

- total population of the microregion - 1 variable
- age structure - a percentage of people in a particular age bin (with the width of the bins equal to 5 years) - 20 variables
- percentage of people living in rural areas and urban areas - 2 variables
- percentage of immigrants compared to total microregion population - 1 variable

– earnings structure - share of the people that earn between x0*minimum_wage and x1*minimum_wage - 11 variables

3.2 Quantitative analysis

Univariate analysis

In table 1, statistics about customer's orders divided by sequential order number are presented. In the whole dataset, 96 thousand orders were the customer's first order. Then, the number of orders falls abruptly, and there are only 47 orders in the dataset that were customer's 5th or later order. The mean value of the transaction does not change with the order number. This means that if the company can make the customer place a second order, it would gain about the same revenue as from the customer's first order. In the last column, percentages of stage-to-stage movement are presented. For example, the probability that customers who bought one time will also buy a second time is 3.18%. The same value, but from second to third order is 8.56%. This means that encouraging the customer to buy for the second time is the hardest task the company faces. With the next purchases, the customers are becoming more and more loyal.

Table 1. Sequential orders analysis

Order number	No. of orders	Mean value	Proportion from previous stage
1	96180	161	-
2	3060	150	3.18%
3	262	152	8.56%
4	49	197	18.70%
5 or more	47	101	-

On the plot 1 the density estimation of the values of payment (left) and transport (right) for each order are presented. I have used the Kernel Density Estimation technique to smoothen the plots. As the distribution is highly right-skewed, I have logarithmed the values. The density plot is grouped by the fact whether the particular customer also placed a second order later. It can be seen that for both variables the 2 densities almost overlap. This means that payment value and

transportation cost probably would not be good predictors in an univariate approach - although maybe they can be interacted with other features and start having predictive power.

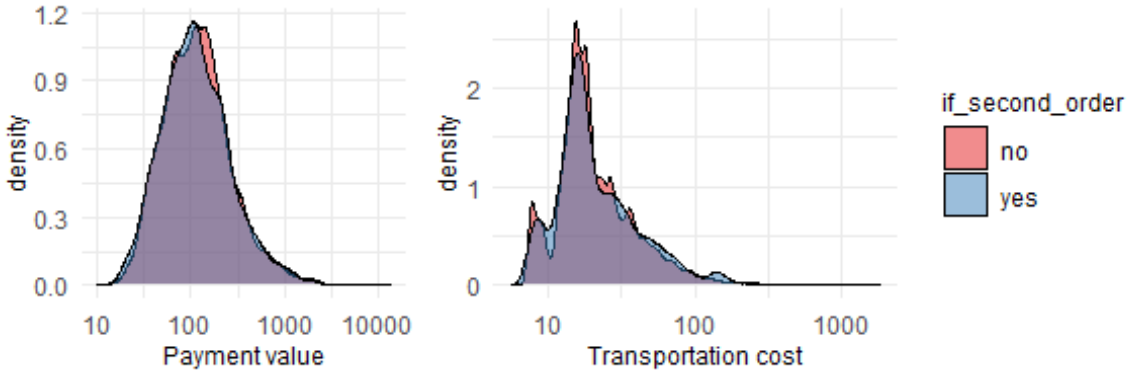


Figure 1: Payment value and transportation cost. x-axis is log-transformed.

An interesting thing to check is whether the value of the ordered products and the transportation cost are correlated. Pearson correlation between these two is 0.504, meaning that the value of the items ordered somehow influences the rest of the costs. I have also plotted these two against each other on the figure 2. Again, I have logarithmed both axes.

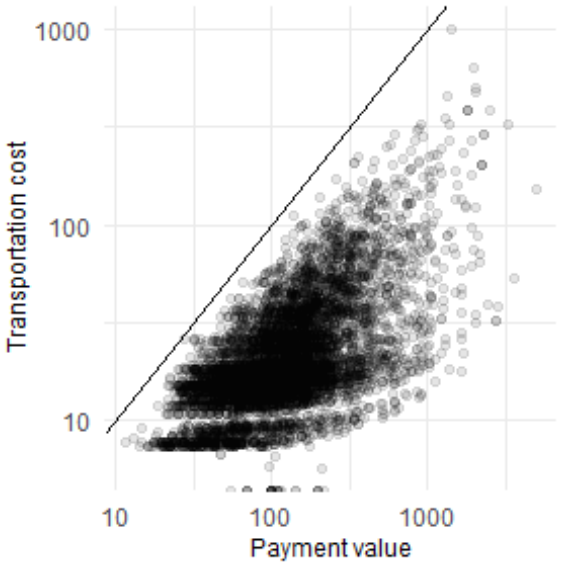


Figure 2: Scatterplot of transportation value and order value. Both axes are logarithmed for better plot clarity.

The relationship is very clear here. For the particular value of the package, the transportation fee is seldom bigger than the value itself. I have added a line with a slope of 1 to highlight that. This probably comes from the company's policy - that it limits transportation cost on purpose because customers wouldn't buy the company's products if the transportation would cost more than the product itself.

On the plot 3 percentages of orders that were given x stars in the review are shown. On the right subplot percentages of the customers that made a second order are presented. Most of the reviews are positive - the scores 4 and 5 make up for 75% of the whole dataset. Another thing worth noticing is the tendency to the negative score polarization - if the customer is unsatisfied with the order, it is more likely to give 1 score than 2.

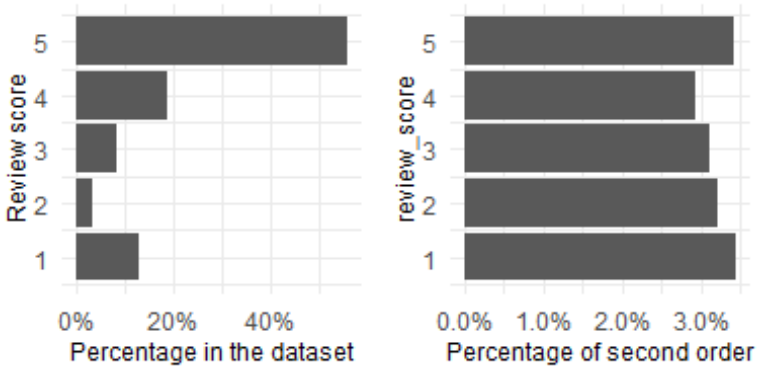


Figure 3: Number of orders grouped by 1-5 review of the purchase (left) and percentage of such orders that resulted in second order (right).

The relationship between making a second order and the review score for the first one is somehow surprising. One would expect that if the clients are unsatisfied for the first time, they will never buy in this store again. In the case of this dataset, it is the opposite - the customers that gave one-star reviews are also the most likely to make the second order. It is worth noting is that the differences between the groups are very small - between 2.9% for review 4 (smallest one), and 3.45% for review 1. One can wonder if this can come simply from random reasons, and that the review score does not influence the probability to come back at all. In particular, the difference between the percentages for the scores 1 and 5 (0.003%) is that small that it is most likely for random reasons.

On the plot 4, analysis of the number of items in the order is presented. I have binned all orders with the number of items more than 4 to one category. On the left subplot is shown the percentage share in the full dataset, while on the right one - percentage of the customers that put second order after ordering x items for the first time.

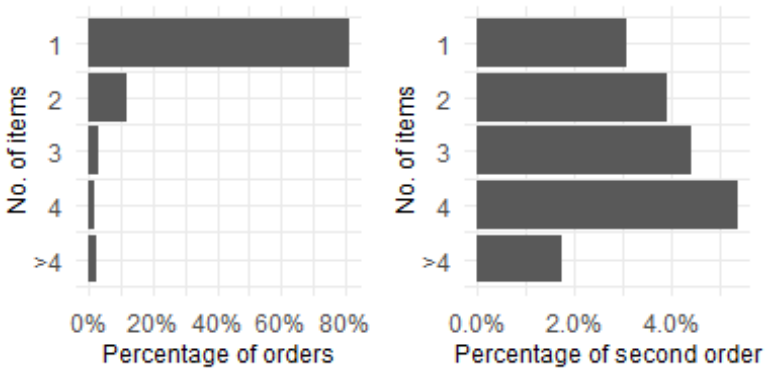


Figure 4: Number of items in an order (left) and percentage of orders that resulted in second order grouped by the number of items (right).

A trend is visible - the more items the customer has bought in the first order, the more likely she is to also put the second order. This difference is pretty strong - between 1 and 4 items the percentage increase in the response is 100%. For more items than 4, this relation is not visible anymore, however, these orders make up for a very small percentage of the dataset.

In the table 2, summary statistics about product categories are presented. The most popular category, “bed, bath and tables” accounts for 12% of all items bought in the shop. The table is ordered by the percentage of the customers that in first purchase bought particular category and later decided to buy in the shop for the second time. The difference in the percentages is visible. For “the best” category, it is 13.8%, while for the worst one - only 1.3%. This is a very promising result and a signal that the dummy variables indicating product category can serve as important features in the modeling phase.

Table 2. Product categories

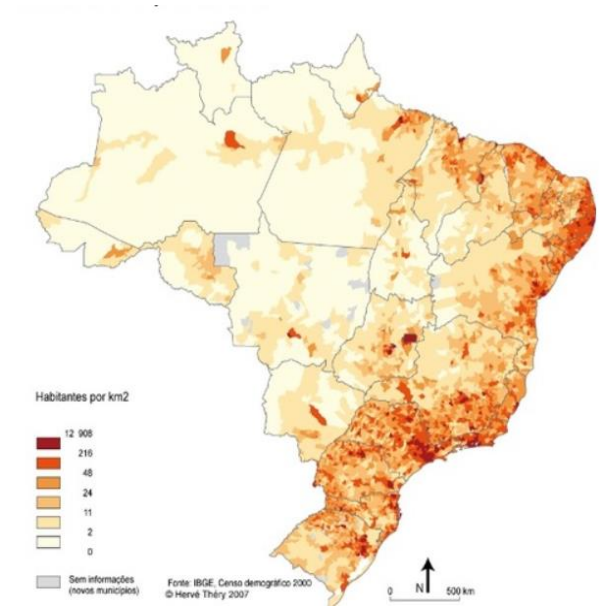
Product category	No. items	Percentage	Percentage of second order
bed_bath_table	7509	11.4%	13.8%
furniture_decor	5801	8.8%	11.5%
sports_leisure	6170	9.4%	9.4%
health_beauty	6996	10.6%	7.4%
computers_accessories	5601	8.5%	6.7%
housewares	5047	7.7%	5.8%
watches_gifts	4475	6.8%	3.8%
telephony	3512	5.3%	3.5%
garden_tools	3432	5.2%	3.4%
auto	3316	5.0%	2.9%
toys	3250	4.9%	2.6%
perfumery	2792	4.2%	2.6%
cool_stuff	3041	4.6%	2.0%
baby	2530	3.8%	1.9%
electronics	2423	3.7%	1.3%

Spatial analysis

In the picture below, a map of Brazil's population density is presented[3]. The most densely populated areas are located in the Southern part of the country. There, also the biggest cities like São Paulo and Rio de Janeiro are located. Another populated area is on the Eastern coast. The North-Western part of the country is the least populated. The distribution of the customers

[3] source: https://www.gifex.com/detail2-en/2018-12-15-15407/Population_density_of_Brazil.html

follows this density very closely (with a correlation of 93%), that is why I did not include the map of customers density.



On the maps in the figure 5, basic statistics about the spatial distribution of the features are presented - in aggregation to microregion level. Such binning is relatively coarse - because of that, some of the statistics can be not reliable in the regions with a very small number of customers. That is why I have decided to remove from the map these microregions, in which the number of customers was less than 5.

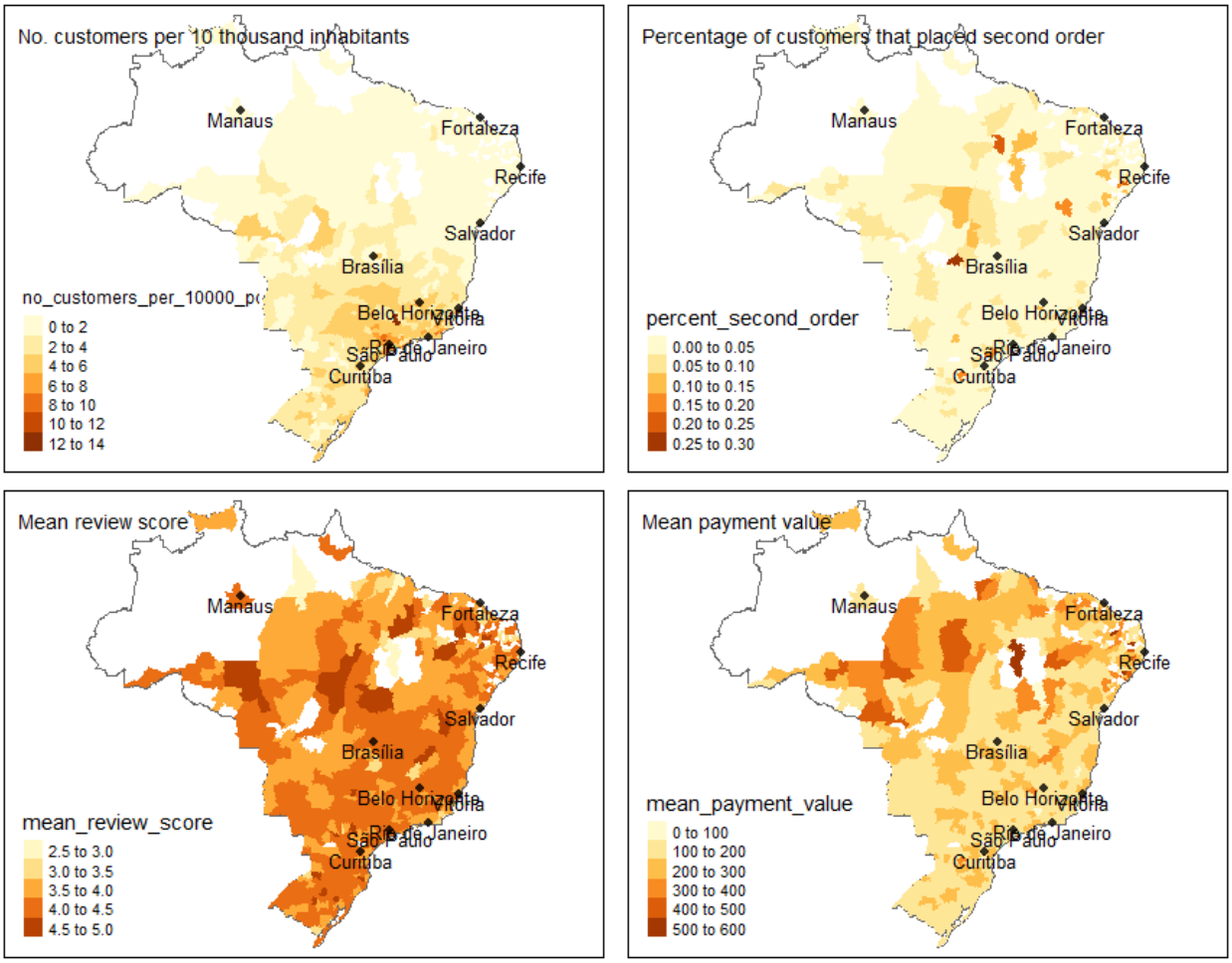


Figure 5: Cartograms presenting customers statistics by microregion.

The top-left map shows the number of customers per 10 thousand inhabitants. It is visible that bigger shares of customers appear in the southern part of the country, concentrated in the triangle between São Paulo, Rio de Janeiro, and Belo Horizonte agglomerations.

The top-right map shows percentages of customers that placed a second order in each microregion. It is could be argued that in the northern part of the country the people this percentage is a bit higher. However, this relationship is rather weak. The same can be said about the mean review score (bottom-left map) - there is no clear pattern visible.

Mean transaction value (bottom-right) is bigger in the northern, more desolated part of Brazil (because of the Amazon Rainforest). One explanation could be that in these parts deliveries of the packages are more complicated/expensive/take more time, and thus the customers are more eager to place one bigger order than few small ones. Another possibility is

that in the northern part the competition between e-commerce sites is smaller, and thus the customers are pushed to buying more items at one supplier.

Summary

Main takeaways of Exploratory Data Analysis are as follows:

- There is a very low percentage of loyal customers (only 3.3% of customers placed the second order), meaning that the classification problem is highly imbalanced
- Most of the variables are very skewed - order value, number of items bought (80% of orders have only one item) and review score (60% of the ratings are 5-star)
- Product category variable looks very promising as a predictor - with mean target variable at 3%, some of the products categories have as much as 11% of customers that bought for the second time. Also, the number of items in the order placed gives some differentiation in terms of value of the independent variable.

CHAPTER IV

Methods description

Introduction

The methodology used in this study can be divided into 3 broad categories:

- Machine Learning Modeling methods - choice of model, cross-validation, upsampling, etc.
- Preprocessing applied to the variables present in the dataset
- Methods used for variable selection.

In the following sections I have described these categories in greater detail.

4.1 Modeling methods

In this study, I have compared Logistic Regression and XGBoost models. The reasons for the choice of these particular models are as follows. Logistic Regression is relatively simple and explainable and was used in the task of churn modeling in previous studies (Nie et al. 2011; Dalvi et al. 2016). On the other hand, the XGBoost model was shown to give superior performance in all kinds of modeling using tabular data, also in the context of churn prediction (Gregory 2018). It can also learn non-linearities and interactions between the variables on its own, contrary to LR where such features should be introduced to the model manually.

Regarding cross-validation, I have used a simple train-test split of the dataset, with 70% of the observations belonging to the training dataset. On the training dataset, I have searched for optimal hyperparameters using 2-fold cross-validation on the training dataset. I have defined search space simply as a grid of all possible combinations of the hyperparameters.

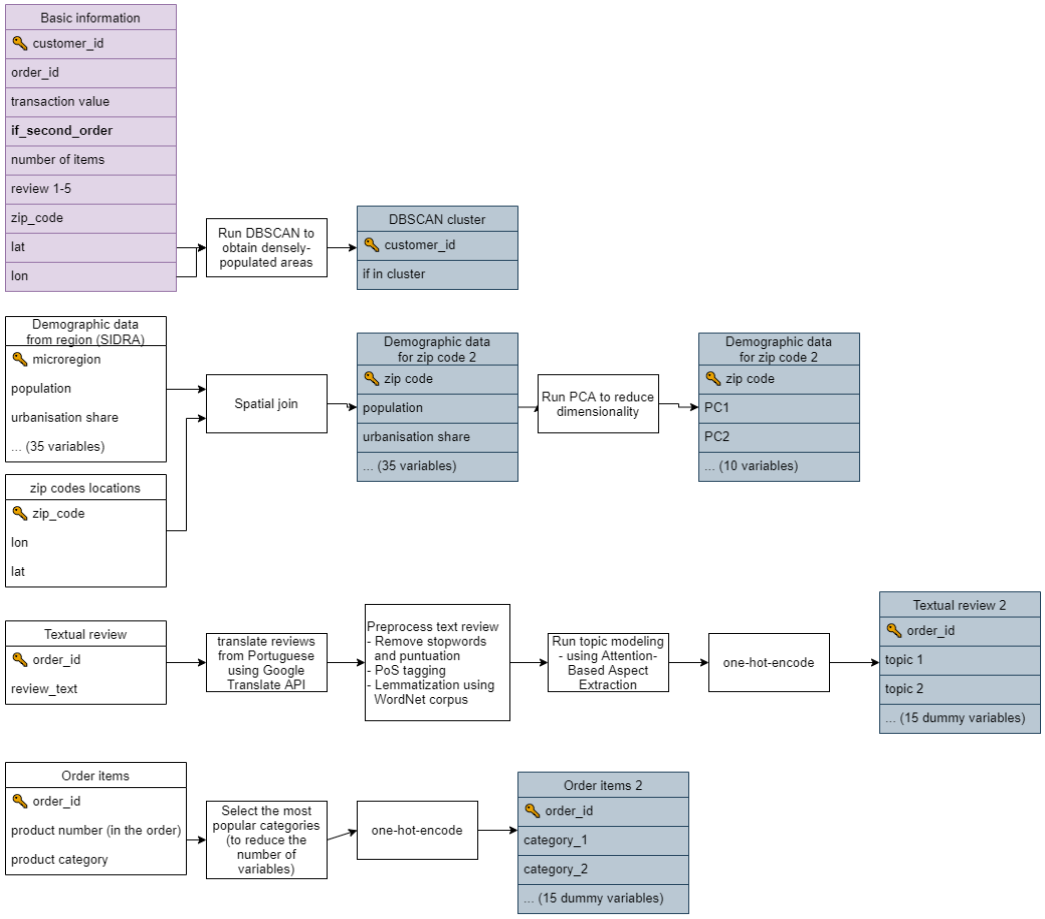
One important problem with this dataset is its very high target classes imbalance. Only 3% of the customers have decided to buy for the second time. To handle this issue I have used upsampling of the minority class on the training dataset to obtain equal class proportions. Also, the choice of an appropriate metric to optimize is very important in an imbalanced dataset, as some metrics (like accuracy) are very biased in these cases. That is why I have decided to optimize the Area-Under-Curve metric, as it weights the performance of the minority and majority classes equally.

4.2 Features preprocessing

In this study I have separated 3 groups of variables analyzed:

– behavioral (first transaction) features
– location features
– perception features.

On the diagram below a summary of preprocessing applied to all the parts of the dataset is presented. All the tables on the left-hand side are coming directly from Olist (4 tables) and Statistical Office sources (1 table - demographic data). The purple table is the primary one, the features from this table were combined with all the remaining sets of variables. The final tables after preprocessing each of the parts of the dataset are shown in gray. In the modeling phase, I have used a simple join of the basic table, and the remaining ones, separately from each other (e.g. basic information + order items, basic information + DBSCAN cluster, etc.).



4.2.1 Behavioral features

Behavioral predictors can be defined as the variables quantifying previous actions of the customer. In most cases, this narrows down to the data about previous transactions and previous interactions with the company. Behavioral information about customer's interactions with the company was shown to be an important predictor in churn prediction (for an overview see Schmittlein and Peterson (1994)).

The widest range of behavioral variables up in churn prediction setting up to date was used by Buckinx and Poel (2005). Besides 7 variables meant for encompassing frequency and monetary value, they also included variables indicating total spending divided by categories of the products available in an e-commerce shop. They found that all 3 categories of variables are statistically significant and bring improvement to the model's predictions. In particular, they found that bigger customer spending leads to the customer's desire to keep being a company's customer. Besides that, the categories that the customer has been buying in the previous purchases also have been shown to influence the customer's decision to stay. This is in line with findings from the previous studies (Athanassopoulos 2000; Mozer et al. 2000). One possible explanation of churning based on categories bought suggested by Mozer et al. (2000) is that the satisfaction of purchasing a particular category is low - no matter if because of the high price or the low quality of the product bought.

In this study, from the category of behavioral variables, I have included information about the monetary value of the first purchase, delivery cost, number of items bought and categories of items that the customer bought. The specification of the behavioral variables used is slightly different from previous studies. Namely, I have excluded Recency and Frequency from the set of predictors. The reason is that in this study I am interested in predicting customer loyalty just after the first purchase. Because of that, these variables require some time to pass since the purchase and can't be calculated or used.

In the case of the company analyzed in this study, the products sold belong to 74 distinct categories. At the same time, the top 15 categories account for 80% of all purchases. Because of potential problems with generalization and also slower model training, I have decided to bin the

least popular ones as a new category "other." Then, I have used the one-hot-encoding approach to create a numeric representation, with the "other" category set as a base level.[4]

To assess the validity of previous studies' findings regarding the behavioral variables in an e-commerce retail context, I have tested 2 hypotheses. (1) **The amount of money spent on the first purchase positively influences the customer's probability of buying for the second time.** (2) **Categories of products bought by the customer can influence the customer's probability to stay with the company.**

4.2.2 Location features

Lee and Bell (2013) argues that customer location and its neighborhood is an important factor to consider in CRM analyses, even in e-commerce settings.

There are multiple ways to include spatial dimension in modeling. In this study, I have analyzed 3 broad approaches that were used in previous studies:

- directly including location variables (geographical coordinates, zip code, region indicator dummies, etc.)
- analyzing neighborhood that the customer resides in (demographical statistics about the region)
- classifying customers by living in an urban or rural area

Direct inclusion of spatial variables

To the best of the author's knowledge, no studies on churn prediction conducted before included raw geographic coordinates in the model formulation. Rather, usually dummy variables indicating the administrative regions were used. There is no consensus on whether such data can improve the predictions. Verbeke et al. (2012) argued that "the number of times a customer called the help desk will most probably be a better predictor of churn behavior than the zip code." On the other hand, Buckinx and Poel (2005) showed that such dummies were significant in the case

[4] Because in one order there can be multiple product categories, it is not guaranteed that there will be only one "1" entry per each row as in the classical one-hot-encoding method.

of the Neural Network model, but not in Random Forest. Also, Long et al. (2019) found that these dummies are significant. However in that case, a different spatial extent was analyzed - the region variables indicated countries rather than postcodes.

Llave, López, and Angulo (2019) used geolocation data in the context of churn prediction for an insurance company. They took a different approach to operationalizing customer location. Instead of including dummies indicating the customer's region, they calculated the distance between the customer and the closest insurance agent. Such variable was significant.

In this study, I have simply included longitude/latitude data about each customer directly to the model formulation. This is to assess **if the propensity to churn can be explained by customer location**.

Geodemographics

Geodemographics is the "analysis of people by where they live" (Harris, Sleight, and Webber 2005). In this paradigm, it is assumed that people living in the same area share similar characteristics, like their social status, income, etc.

As pointed by Singleton and Spielman (2014), geodemographic features were mostly used in the studies regarding public sector areas, mainly public health and law enforcement. Publicly available research in the usage of geodemographics in the context of marketing, or specifically churn prediction is almost non-existent. This has its reasons in the confidential nature of research done in individual companies (Webber 2004). The only publicly available study was conducted by Yu Zhao et al. (2005). They found that geodemographic features were significant in the churn prediction model.

A hypothesis I would like to check is **if the social structure of the customer's environment can serve as a valuable predictor of churn tendency.**

In total, I have included 35 demographic features for the microregion from which the customer is - age structure, percentage of the population in an urban area, income structure, number of immigrants. These features were obtained from the Brazilian statistical office[5].

Geodemographic dimension in this study is relatively high dimensional. At the same time, one would expect that the information can be somehow compressed because lots of the variables represent very similar concepts (for example there are 20 variables encoding only age structure). Because of that, I have decided to process this part of the dataset using Principal Components Analysis. This can potentially bring some improvements in the process of Machine Learning modeling, as training the model on a smaller, compressed dataset is more resource-efficient and at the same time was shown to improve the modeling results in some cases (Howley et al. 2005; Dutkiewicz, Terhal, and O'Brien 2021).

One decision regarding PCA transformation is whether to use a standard version or the one with rotated loadings (Corner 2009). The trade-off between these two methods is that the rotated loadings version allows for an interpretation of the loadings, but is less optimal in a sense that the variance along each loading is not maximized. I have decided that a standard one would be more suitable in the case of this study because the explainability of the input variables to the model is not as important as correctly representing the features in lower-dimensional space and thus preserving as much valuable information as possible for the modeling phase.

Rural vs. urban customer location

Generally, there is a consensus among researchers that there is a difference in customer behaviors between rural and urban areas (Sun and Wu 2004). In particular, a couple of studies in the FMCG sector have found that rural customers tend to be more loyal to the previously chosen company (Jha 2003; Sharma and Singh 2021). The potential reason for such finding provided by the authors is a smaller choice of other options in the rural shops compared to urban ones. However, up to date, there were no studies that were meant to assess the differences between

[5] Joining the data coming from this source and main transaction dataset proved to be challenging. The details of such spatial join are presented in Appendix A.

customer loyalty in urban and rural areas but aimed at the e-commerce sector. The findings from the FMCG sector do not have to translate directly, as in an online setting the customers are generally not limited by the availability of the brand in their area.

A hypothesis worth checking is **if the tendency to churn is dependent on whether the customer is living in a densely populated area.**

There are 2 possible ways to conclude if a particular customer is living in an urban or rural area. One is simply checking if the customer's coordinates are inside the city's administrative boundaries. Such an approach does not guarantee that this customer is really living in a densely populated area - because of the fact that administrative boundaries do not have to reflect actual boundaries (for example, because of fast suburbanization spilling to previously village areas).

Another way is inferring the population density in the area from empirical data. This way, one gets more reality-reflecting densely populated areas classifications. As was shown before in the dataset review, the number of customers per microregion highly correlates with population density in this area. Because of that, it can be argued that also in a smaller scale of analysis than microregions such correlation will be also evident. This leads to a conclusion that the company's customers' locations can be used as a proxy for population density, so it can be used for classifying densely and sparsely populated areas.

In this study, I have used the Density-Based Spatial Clustering with Noise (DBSCAN) algorithm for the task of rural vs. urban areas classification. This clustering algorithm besides assignment to a particular cluster can also detect noise points. Because of that, the assignments have a natural interpretation. When the point belongs to any cluster it means that this customer is living in a densely populated area, while the points decoded by DBSCAN as noise are the customers living in more isolated places.

DBSCAN has 2 parameters to be decided before running the algorithm. These are the minimal number of points lying close to each other that are needed to constitute a cluster (k), and maximal distance, at which one considers the points to lay close to each other (*epsilon*). A typical rule-of-thumb for deciding k and epsilon parameters is to first set k, and then plot k-nearest-neighbors distances. Epsilon should be then decided based on *elbow point*, where the line is bending. However, when the features are geographical coordinates, epsilon is actually a physical

distance between two locations. That is why one can set what should be more reasonable criteria for constituting clusters.

In my work, I have decided that the minimal number of customers in the cluster is 100, and the maximum distance between the customers in one cluster is 50 kilometers. For the location of Brazil on the geoide, this transfers roughly to epsilon=0.2.

4.2.3 Perception features

Customer perception of the company is considered an important factor driving customer loyalty (Kracklauer, Passenheim, and Seifert 2001). Unfortunately, customer satisfaction is an immeasurable variable. Different proxies can be however included in the model, and usually gathering such data requires conducting customer surveys. Oliveira (2012) specifies possible dimensions of such survey: "overall satisfaction, quality of service, locational convenience and reputation of the company."

In e-commerce settings, an industry-standard is to provide a way for the customers to express their opinions about the purchase (Lucini et al. 2020). The company has to decide, in how structured way it would like to collect them. Text reviews can provide way richer information about the customer experience, as they are not limited to describing the experience in predefined dimensions. On the other hand, extracting meaningful information from sometimes millions of text reviews is a very challenging task to which no universally acclaimed solutions exist (Felbermayr and Nanopoulos 2016; Yabing Zhao, Xu, and Wang 2019).

In the case of the dataset analyzed in this study, there are 2 proxies of customer perception available. One is a customer review on a scale from 1 to 5. The other is a textual review of the purchase. Using numeric review in the modeling is straightforward and doesn't require further explanation. In the next sections, I have described the preprocessing of textual reviews in greater detail.

Ways of analyzing textual reviews

As stated before, text reviews can potentially serve as a rich source of information about customer satisfaction. Although text mining for customer reviews in general is an active field of research, usage of such information in the context of churn prediction is way less covered. To the

best of the author's knowledge, only 2 studies used the data from textual reviews for churn prediction. De Caigny et al. (2020) have used text embedding approach, while Suryadi (2020) - simple tf-idf technique.

Lucini et al. (2020) specifies 2 natural language processing areas that can be used to extract insights from customer reviews, namely topic modeling and sentiment analysis. The first one is meant to answer the question "what the review is about?" while the second - "what is the perception contained in this review?" A combination of these two dimensions can help answer the question, which areas of customer experience are rated positively, and which need improvement.

In the case of this study, I have focused only on extracting the topic from the review. The reason is that information about whether the experience of the customer was positive is already contained in a numeric review. **My hypothesis is that both the numeric review, as well as topic of the textual review can be useful predictors of customer loyalty.**

Previous research in topic modeling

Undoubtedly the most popular model for inferring the topic of a text is Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). The method is based on assumption that each document is a mixture of a small number of topics. At the same time, each topic can be characterized by a distribution of words frequency.

Hong and Davison (2010) argues that short texts (as in the case of customer reviews) comprise of a very small amount of topics, usually only one. Because of that, LDA should not be used in such settings as its assumptions are violated. This claim is supported by an empirical study of short texts from Tweeter, in which LDA has failed to find informative topics.

The drawbacks of LDA in the setting of short texts were addressed by Yin and Wang (2014) . They used the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model, which is an improvement over typical LDA. The main difference compared to the basic algorithm is an introduction of assumption, that each text comprises only one topic. The authors show that this algorithm provides superior performance compared to the basic LDA technique in the context of short texts.

More modern approaches to topic modeling were also developed recently. A milestone in the whole NLP field was inventing an efficient way to embed words in a vector space while preserving their meaning, namely word2vec (Mikolov et al. 2013). On a basis of this method, He et al. (2017) presented an Attention-based Aspect Extraction[6] model. At first, words embedding using the Word2Vec model is created. After that, for each text in the corpus, attention weight for each word is computed using a neural network with an attention layer. Then, an embedding of the whole sentence is created by computing an average for all words embedding. The words are weighted by their attention weights. The last step of the procedure is creating an encoder-decoder model for learning sentence aspect embedding. The reconstruction of the sentence is the linear combination of aspect embeddings, and aspect embeddings are learned by mapping sentence embedding to a lower-dimensional space.

Other studies using the embedding technique were conducted by Tulkens and Cranenburgh (2020) and Luo et al. (2019). In both of the studies, the algorithms presented outperformed the LDA method in the task of short text topic modeling.

Text reviews preprocessing in this study

In this study, I have tried and evaluated 3 algorithms for topic modeling:

- Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) - because it is a go-to standard for topic recognition.
- Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (Yin and Wang 2014) - as this method is an improvement over LDA, meant especially for short texts. This is true in this case, as most of the reviews are just a couple of words long.
- Attention-Based Aspect Extraction (He et al. 2017) - this method is also meant for short texts, and at the same time, it uses the most modern, state-of-the-art NLP techniques. Besides that, in the original paper, the authors worked in a similar domain of internet text reviews.

Various preprocessing steps were needed to apply all 3 aforementioned algorithms:

[6] Words “Aspect” and “Topic” are often used interchangeably in the NLP literature

- **Translation of the reviews from Portuguese to English.** Olist e-commerce store is operating only in Brazil. That is why most of the reviews are written in Portuguese. I have used Google Translate API to change their language to English. This is to facilitate not only understanding the reviews, but also the NLP tools available for the English language are more advanced than for other languages.
- **Removal of stopwords and punctuation.**
- **Lemmatization** using WordNet lemmatizer (Fellbaum 1998) combined with Part-of-Speech tagger. This step is needed to limit the number of words in the vocabulary. Thanks to the Part-of-speech tagger, the lemmatizer can change the form of the word on a more informed basis, and thus apply correct lemmatization to more words.

Later steps of the preprocessing were different for each of the algorithms.

For LDA and Gibbs Sampling, only **converting lemmatized reviews into vector format** was needed. In the case of LDA, the count-vectorizing approach was applied, with removing of words that appeared in less than 0.1% of reviews. In the case of Gibbs Sampling, the same preprocessing is done internally by the training function from the package. In both of these cases after vectorization, one should obtain a matrix with n rows and k columns, where n is the number of observations in the original dataset, while k - the size of the vocabulary.

Very different preprocessing was required in the case of Attention-Based Aspect Extraction. The neural network architecture proposed by the authors requires simply lemmatized reviews in a textual format as the output. Then, one of the layers of the network is meant to embed the currently preprocessed word. These embeddings are not learned during the network training, they should be trained beforehand instead. The authors of the paper propose the Word2vec technique (Mikolov et al. 2013) for learning embeddings. Following their guidelines I have used this method, setting the dimensionality of the vector space to 200. I have also applied the word window of 10. After applying word2vec on this dataset, I have obtained the matrix with m rows and 200 columns, where m stands for the number of words in the dataset, and 200 is the dimensionality of the vector space chosen as a hyperparameter.

Concerning topic models training, I have searched for optimal hyperparameters for all 3 models based on grid search. For LDA, I have tested a varying number of topics that the model

has to learn (3, 5, 10, and 15). For GSDMM, 2 parameters influence topics coherency in each "cluster." I have run the algorithm for all 16 combinations of both parameters chosen from the values 0.01, 0.1, 0.5, and 0.9. For Attention-Based Aspect Extraction, I have manipulated the number of topics to learn, from the values 10, 15. Unfortunately, as this last model takes a very long time to run (around 3 hours per one set of hyperparameters), I have limited the number of hyperparameters checked compared to the LDA model.

The evaluation of topic extraction is a hard task, as no model-agnostic metrics that can be compared between different models exist. The only reasonable method is human inspection. That is why after running every model I have verified the obtained topic for coherency (whether reviews inside one topic are similar) and distinctiveness (whether there are visible differences between modeled topics).

4.3 Variables selection methods

To summarise, from variable preprocessing I have obtained these 6 sets of features:

- basic information - the value of the purchase, geolocation in raw format lat/lng, the value of the package, number of items in the package, review score (6 variables)
- geodemographic features for the region from which the customer is - age structure, percentage of the population in an urban area, income structure, number of immigrants (35 variables)
- geodemographic features transformed using PCA - (10 variables/components)
- indicator whether the customer is in an agglomeration area obtained from DBSCAN on location data (1 variable)
- product categories that the customer has bought in the purchase (15 dummy variables)
- main topic that the customer has mentioned in the review (15 dummy variables).

An approach used by Oliveira (2012) for an assessment of the new feature previously untested in the churn prediction was to compare 2 models, one containing only basic RFM features, and the other RFM features and also this new feature. I have used a similar approach. Namely, first I have included basic features that didn't require any preprocessing. This model served as a baseline. Then, for each of the sets of features that I have computed, I have estimated

a model containing these features + basic features. Lastly, I have created one model containing all the variables. This resulted in the following 7 feature sets tested:

- basic features
- geodemographic + basic features
- geodemographic with PCA + basic features
- agglomeration + basic features
- product categories + basic features
- review topic + basic features
- all variables - (with geodemographic features transformed with PCA)[7]

Automatic feature selection - Boruta algorithm

To test if the approach with including whole sets of features to the training set is an optimal one, I have also tested one method of automatic feature selection, namely a Boruta algorithm (Kursa, Rudnicki, and others 2010). It is widely popular among machine learning practitioners (Kumar and Shaikh 2017). The algorithm belongs to the category of wrapper feature selection algorithms, and a Random Forest algorithm is usually used as an machine learning method. It works as follows. At first, all features from the original dataset are randomly permuted. This way, one obtains a dataset with close-to-zero predictive power. Then, the resulting features are added to the original dataset and the Random Forest model is trained.

This model has a built-in feature importance measure, which is usually Mean Decrease Impurity (MDI). After running the model, for each of the original features, MDI is compared against all MDI scores for shadow features. If for any original variable the score is less than the one from any of the shadow features, the variable gets a "hit."

[7] I have not run the model containing all variables with demographic features without PCA preprocessing. There are 2 reasons for that - one is that number of variables in this set is very big, which poses performance reasons - model training simply would take a very long time. The other is that the model with only included PCA demographic variables performed better than the full set of variables.

The above procedure is repeated for a preassigned number of iterations. Finally, important features that should make it to the final model are the ones that obtain fewer hits than preassigned value.

After gaining knowledge about the variables that should make it to the model, I have trained the XGBoost classifier using these features. The rest of the fitting procedure (cross-validation, up-sampling, hyper-parameters, etc.) stayed the same as in the rest of the approaches.

One should have in mind that the Boruta algorithm is very time-consuming. The minimal number of runs recommended by the method authors is 100, and one run consists of fitting a Random Forest model to the whole dataset with doubled number of features (because of added shadow features). In the case of this analysis, model computation took about 12 hours on a modern laptop. Although other wrapper algorithms also require an iterative fitting of the model, they usually start with fitting the model to one variable, in the next iteration to 2, and so on up to k features. On the other hand, the Boruta algorithm in each iteration fits the model to 2*k features (original and shadow features).

CHAPTER V

Results

Introduction

In this chapter the results of the analyses conducted in this study are presented. First section covers the outcome of the methods that were used for the dataset preprocessing. In the second section the performance of the churn prediction models tested in this study is presented and analyzed in a greater depth. In the last section the models are analyzed to verify the hypotheses about the direction of the influence of predictors on the customer churn.

5.1 Results of the pre-modeling phase

Topic modeling

I have manually assessed the topics obtained from LDA, Gibbs Sampling and Aspect Extraction methods. The only meaningful output was produced by the last method. The results of the topic modeling with examples of reviews for each topic and proposed topic labels are presented in Appendix B.

DBSCAN for long/lat data

In the case of epsilon equal to 50 kilometers range, DBSCAN found 78 clusters and 18 thousand noise points. After a quick visual inspection of the points colored by cluster association, I have concluded that the boundaries of the clusters overlap with bigger cities' boundaries, which proves that the clustering has discovered valuable information from the data.

I have also tried running the clustering with the epsilon obtained from using the elbow method, namely 2. Only 2 clusters were created, of which one of them contained 99.97% of observations. The probable reason for that is that 2 would roughly translate to 500 kilometers of cluster range, which is a way too large radius to model information about the population density on a country level.

To the final dataset, I have added only one variable indicating whether the point belongs to the cluster or is one of the noise points.

PCA for demographic data

Cumulative variance explained by each of the consecutive loadings is presented on the plot 6. I have decided to leave the 10 most informative PCA eigenvectors. These account for 97.3% of the explained variance.

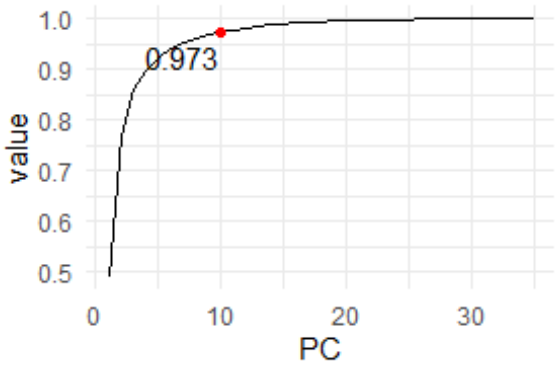


Figure 6: Explained cumulative variance by each of the PCA loadings.

Boruta feature selection

The Boruta algorithm concluded that from all 47 variables only the 14 variables indicating topics are non-relevant. One should notice that even using automatic feature selection, the algorithm has dropped the whole category of variables, meaning that the approach of manually setting sets of variables to include in the model is also "recommended" by the algorithm.

5.2 Performance analysis

AUC metric analysis

Table 3 shows the performance of the XGB models using various sets of variables. The best AUC score on the test set is obtained by the model containing basic features combined with dummies indicating product categories that the customer has bought during the first purchase. AUC is greater than 0.5, which means that the model has predictive power better than random guessing.

Table 3. AUC values for XGBoost model

Model	AUC test	AUC train	Performance drop vs. the best model
product_categories	0.6505	0.9995	0.00%
all_with_pca	0.6460	0.9997	-0.68%
boruta_no_topics	0.6426	0.9998	-1.20%
agglomeration	0.6382	0.9993	-1.88%
topics	0.6353	0.9992	-2.34%
basic_info	0.6338	0.9991	-2.56%
demographics_pca	0.6323	0.9996	-2.80%
demographics	0.6254	0.9995	-3.86%

The second-best model is the one containing all variables, with demographic variables transformed with PCA. It is worth noticing that this model also contains the features containing product categories information, so similar performance is not a surprise. The percentage drop in AUC is very small (0.6%). The model with only basic information is about 2.5% worse.

The score of the subset of features selected by the Boruta algorithm using AUC on the test set is 0.646 - less than the model including all variables. This means that using the Boruta algorithm did not bring additional predictive power to the model. At the same time, this means that the variables indicating reviews topics seem to be relevant for the model performance.

Another thing worth noticing is the fact that AUC on the train set is almost 1 in every model. These values are worrying because this means that the models are highly over-fitted, and that generalization problems can be present. The XGBoost model has some built-in parameters that can be used as regularization strategies, like the maximum tree depth of a single tree trained and a number of iterations. In search of a less over-fitted model, I have tweaked these parameters in cross-validation. However, although in some cases I was able to make the model overfit less, the performance in 2-fold cross-validation was still the best with highly over-fitted models.

I have also created a table 4 containing similar information, but this time for the Logistic Regression model. The main finding is that even the best LR model (containing product categories and basic features) is worse than the worst XGBoost model (0.586 vs. 0.625, respectively). This means that linear modeling is in general very poorly suited for this prediction task.

Table 4. AUC values for Logistic Regression model

Model	AUC test	AUC train	Performance drop vs. the best model
lr_product_categories	0.5862	0.5922	0.00%
lr_all_with_pca	0.5813	0.5960	-0.84%
lr_basic_info	0.5535	0.5529	-5.58%
lr_demographics	0.5492	0.5632	-6.31%
lr_demographics_pca	0.5482	0.5606	-6.48%
lr_agglomeration	0.5464	0.5532	-6.79%

AUC values for the test set oscillating below 0.6 mean that the model is very poorly fitted to the data. For the worst model containing only the agglomeration feature, it is at the value of 0.546. It is that close to the level of random classifier (0.5), that one could even argue that this model does not have any predictive power.

An interesting remark is that judging my AUC values, both LR and XGBoost select the same 2 models as the best ones - namely the one with product categories and with all variables. From the fact that 2 such different models arrived at the same conclusion in terms of which variables should be included, this means that these variables simply provide the biggest predictive power, regardless of the model used.

Comparison of performance for *agglomeration* set of features is particularly interesting. In the XGBoost model, this feature is rated as the 3rd best one (after excluding Boruta set to compare meaningfully with LR table). In the LR case, it is scored as the worst one. One possible

explanation is that it's because of the inherent ability of XGBoost to create interactions between variables, while these interactions should be included in the LR model manually.

From the perspective of CRM, the most important result of the modeling procedure is that the created model has predictive power in the task of churn prediction. This means that using the model's predictions the marketing department can understand which of the customers are most likely to place the second order and can be encouraged further. And on the other hand, which customers have a very low probability to buy, and thus the company can restrain from losing money on targeting them.

The AUC scores in the above tables are only point estimates. From such information, one cannot tell whether the performance would still be the same for a slightly different test set. This is especially crucial in the case of this study, as the differences between all the XGBoost models are not that big.

A standard way to compare the models' performance in a more robust way is using a bootstrap technique. I have sampled with replacement observations from the test set and calculated the AUC measure. Specifically, I have done 100 re-sample rounds, and the models chosen were the best one (product categories + basic information), second-best (all variables), and the one with only basic information as a benchmark. The figure 7 shows density estimates of these 3 empirical AUC score distributions.

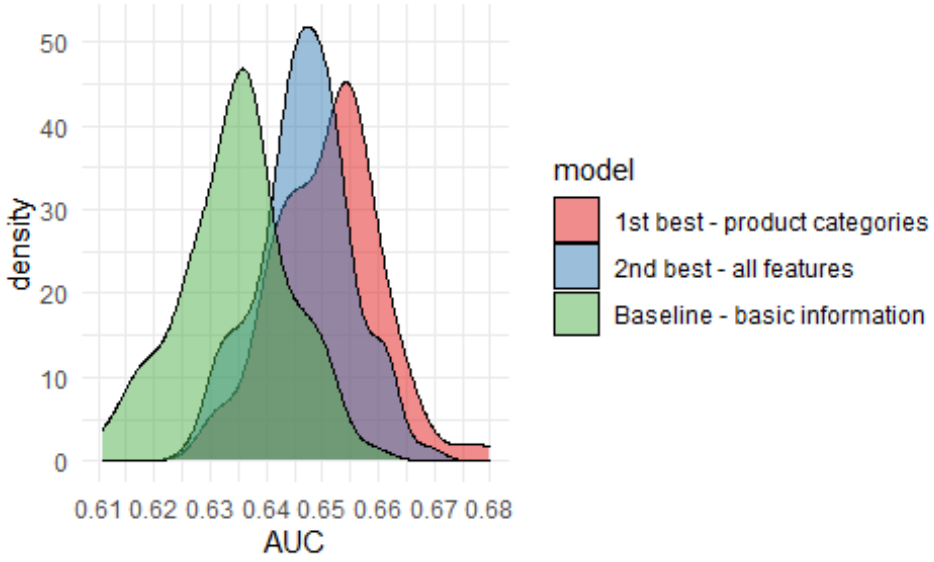


Figure 7: Boostrap AUC estimates for 3 XGBoost models.

The curve for the model with basic features is standing out of the others. However, the difference between 1st-best and 2nd-best models is not as clear - it looks like the better model has slightly better density curve shape, but this should be investigated more thoroughly. That is why I have used a Kolmorogov-Smirnov test to check if the empirical distributions can come from the same probability distribution. I have run the test twice using 2 alternative hypotheses. First one with H1: auc_best =/= auc_2nd_best, and the second one: H1: auc_best > auc_2nd_best.

The p-value for the first hypothesis is 0.0014. This means that with the level of significance 0.05, 0.01 the performance of the models is distinguishable. At the same time, p-value with 'greater' hypothesis is 0.0007. This means that at the levels of significance 0.05, 0.01 one can say that the performance of the first model (only product categories) is better than that of the second one (all variables).

Another reason to choose the "smaller" model for usage in the production setting is Occam's razor heuristic. The model with product categories has 21 variables, while the one with all variables included - 47. If there is no important reason why the more complex approach should be used, the simpler is usually better. In this case, using a simpler model has the following advantages for the usage in the CRM context:

- Faster inference about the new customers - especially in an online prediction setting when the predictions have to be done on the fly
- The predictions are easier to interpret
- Easier model training (and retraining, if the model's performance will drop with time)

Lift metric analysis

Typically, the output of churn prediction modeling is used in customer targeting campaigns.

An ultimate goal of customer churn prediction is gaining information, which customers are most likely to place a second order. More specifically, one has to create a ranking of customers, in which they are sorted by their likelihood to buy for the second time. For each cumulative part of the ranking (top 1% of customers, top 10%, etc.), one can compute, which percentage of this part is truly buying for the second time. This type of approach is called a lift analysis and is a go-to tool for measuring the performance of targeting campaigns. Such information is also very easily understandable by CRM experts without deep knowledge of statistics and machine learning.

From these insights, the CRM experts can make an informed decision which customers are the most likely to respond positively to targeting efforts.

On the plot 8, a lift curve is presented. On the x-axis, the fraction of the *top* customers ranked by probability to buy for the second time is presented. On the y axis, a lift value for this quantile is shown.

The shape of the plot resembles the one of the function 1/x. Values of lift are very big for the smallest percentage of the best customers to target, and they are getting smaller very quickly. This means that the more customers the company would like to target based on the model prediction, the less marginal effects it would get from the usage of the model. For example, for the top 1% of the customers, the model can predict retention 18.7 times better than the random targeting approach. For the top 5%, it is still very effective, being 4.2 times better. If one would like to taget half oth the customers, the improvement over random targeting is 0.3 (130%). Although this value is less impressive than for smaller percentages, it is still an improvement over random targeting.

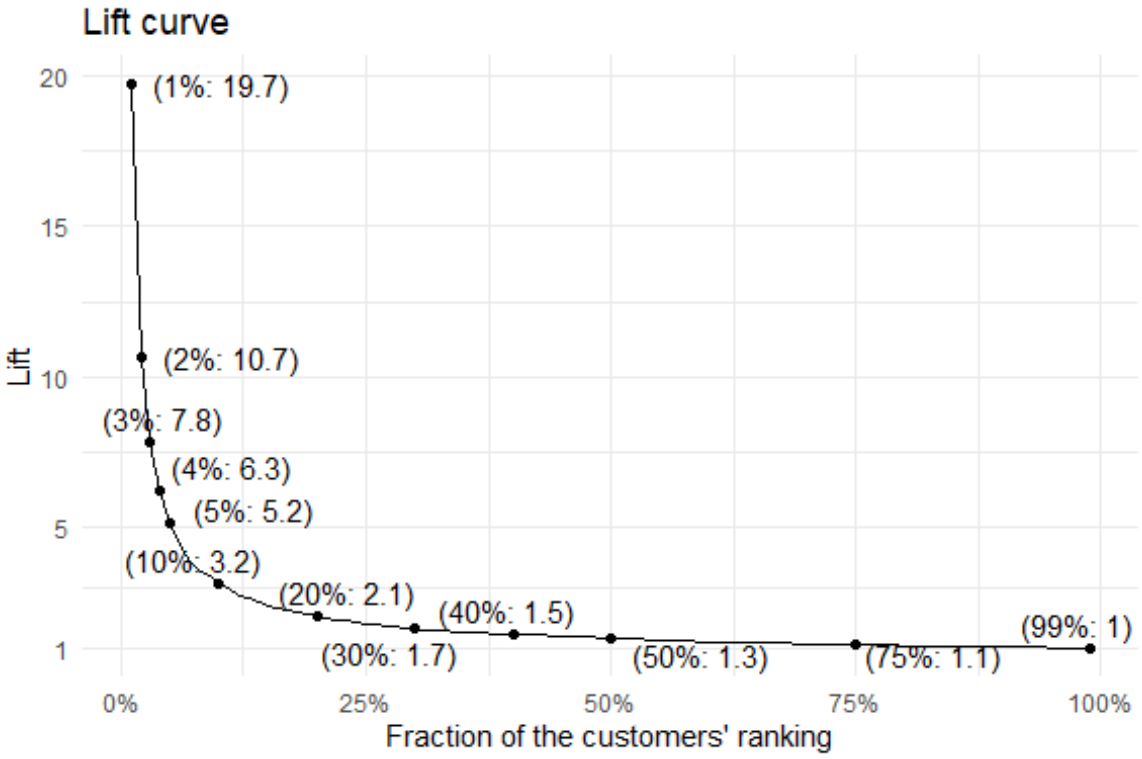


Figure 8: Lift curve.

5.3 Model's working explanations

In this study, I have analyzed Logistic Regression and XGBoost algorithms for the task of churn prediction. Logistic Regression is an interpretable model by design - one can simply look at the model coefficients and infer about strength and direction of a particular feature influenced on the final prediction. However, as shown in the previous sections it is inferior to XGBoost with regards to predictions quality. XGB is a *black-box model*, meaning that its structure is too complex to be directly inspected. To be able to test hypotheses about importances and direction of influence for model prediction, XAI tools have to be used.

In this study, I have used 2 techniques of XAI, namely Variable Importance (VI) and Partial Dependence Profile (PDP). The first one can answer the question "which variables (or categories of variables) influence the predictions the most?" while the second - "What is the direction and strength of this influence?"

Influence of features groups on customer churn

In this study for the feature importance assessment, I have used a Permutation method (Biecek 2018). There were 2 reasons for that choice. First, XGBoost model does not have a model-specific Variable Importance measure (as for example Mean Decrease Gini in the case of Random Forest), and thus a model-agnostic method has to be used. Second, Permutation Feature Importance allows to test not only feature importance of one variable at a time, but also sets of variables.

The method is based on model performance changes when random permutations are applied to predictor variables. Because of the feature values are "exchanged" between the observations, they stop bringing any information to the model (because they are random). If a particular feature is heavily used by the model in obtaining predictions, then the model's performance will drop by a large amount. Similarly, if a feature is not used by the model at all, when it will be shuffled the model's performance won't change. Such operation can be easily generalized to sets of features - one simply has to permute more features at once instead of just one.

A scaling can be applied to the resulting AUC scores per each feature to facilitate interpretation. It is specified in the formula 1. I have scaled the values of AUC for the feature f (AUC_f) to the 0-1 range based on 2 quantities - AUC for the model without any variables' permutations (AUC_{full}), and 0.5 (AUC score for random classifier). Then interpretation is as follows. If a score for a particular feature is close to 1, this means that the model after excluding this feature starts behaving like a random classifier, so this feature is extremely important. On the other hand, if the score is close to 0, this means that the model performance didn't change at all, so the feature is unimportant.

$$score_f = 1 - \frac{AUC_f - 0.5}{AUC_{full} - 0.5} \qquad (1)$$

I have examined variable importances for two of the XGB models from this study - one for the best set of variables (with included product categories), and another - with all variables included. The reason to check variable importances also for the model with all features is that it can answer the questions about the significance of the particular sets of variables.

To the left of the figure 9 the variables importances from the XGB model for the best variables (with included product categories) are presented. The most important one is the transportation cost. Also, high importance scores are obtained by value of the payment and vanilla geolocation variables.

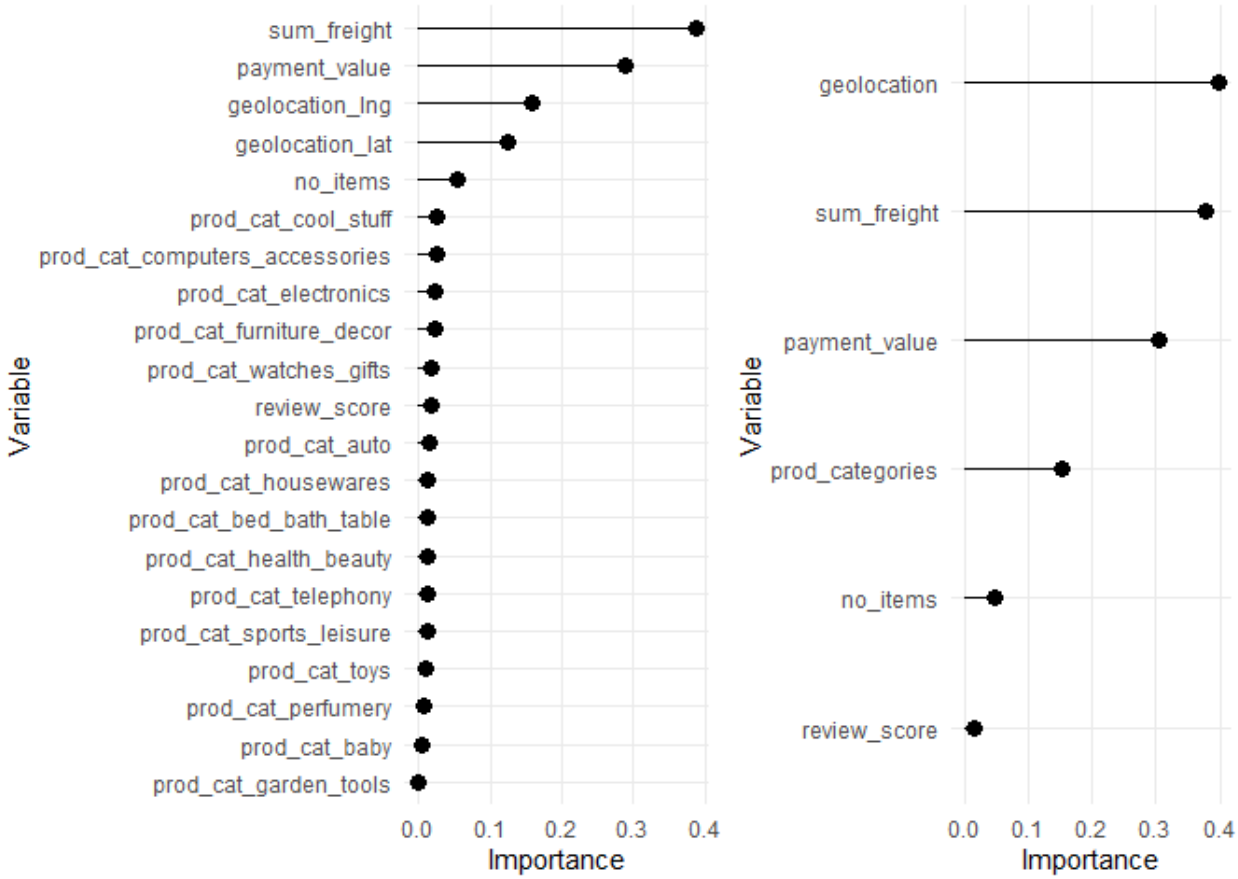


Figure 9: Variable importance plots for the model with included product categories. Left subplot shows single variables, while right one - binned product categories and geolocation features.

Most of the dummies indicating product categories are in the latter part of the ranking. One could wonder, why despite these features are relatively unimportant variables, they lead to a 2.5% gain in AUC compared to the model without them.

This is because conceptually all of the dummies indicating product categories encode one information, these variables' importances should be treated jointly. The same can be argued about geographic coordinates. To account for this, I have used a feature importance for these variables

sets (“geolocation” and “prod_categories”) instead of individual feature importance. This information is presented in the right subfigure.

After this operation, it can be seen that product categories gained in relative importance - now they are the 4th variable. Also, the geolocation variables set became more important than payment value and transportation cost.

In the figure 10 variables’ categories importances for the model with all variables are presented. I have grouped the variables into 3 categories:

- variables describing the first transaction of the customer - payment value, product categories, etc.
- variables describing perception - namely 1-5 review and dummies for a topic of the textual review
- “geo” variables - with 3 subgroups:
  - variables describing demographics of the region that the customer is in
  - raw location - simply longitude/latitude coordinates
  - density - variable indicating whether the customer is in a densely populated area.

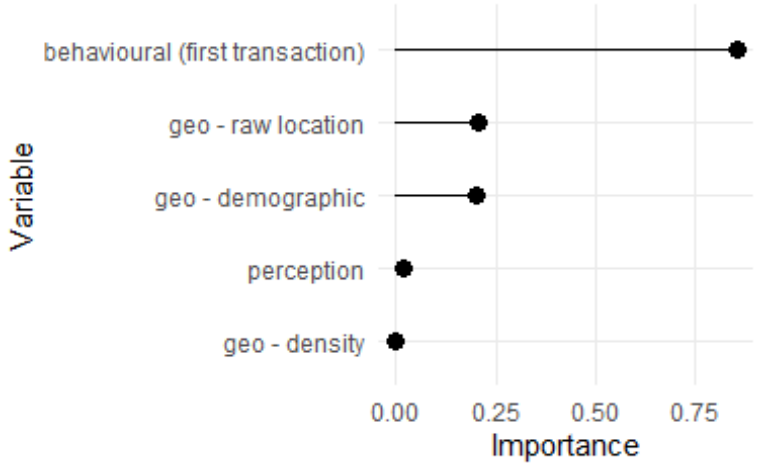


Figure 10: Variable importance plots for the model with all variables.

The best set of variables is the one containing behavioral features. The next 2 sets, namely geodemographic and spatial location have a similar influence. The lowest impact on the model predictions have the perception variables and the density population indicator.

From these values, one can validate the hypotheses stated earlier. **Customer's propensity to churn depends on**:

- Payment value for the first order, number of items bought, transport cost of the package
- Categories of the products bought
- Demographic environment of the customer
- Customer's location.

At the same time, customer's propensity to churn is not (or is only very mildly) influenced by the following factors:

- population density in the customer's area
- 1-5-star review of the purchase
- topic of the customer's textual review

In the following section, I have tried to answer the questions about the direction of the influence of predictors of customer churn.

Direction and strength of features influence on customer churn

Partial Dependence Profile is based on the Cateris Paribus technique. This technique is meant to perform a "what if" analysis for one observation and one feature. For this observation, the variable of interest is changed and the model predicts the response for each of these changes. Partial Dependence Profile is simply averaged value of such Cateris Paribus analysis for each of the observations from the dataset.

In the figure 11 Partial Dependence Profile for payment value is presented. The black line is the profile itself. To facilitate drawing conclusions from the PDP plot for payment value I have included a smoothing line (blue). Also, I have added vertical lines indicating quantiles of this variable's values in the dataset (red).

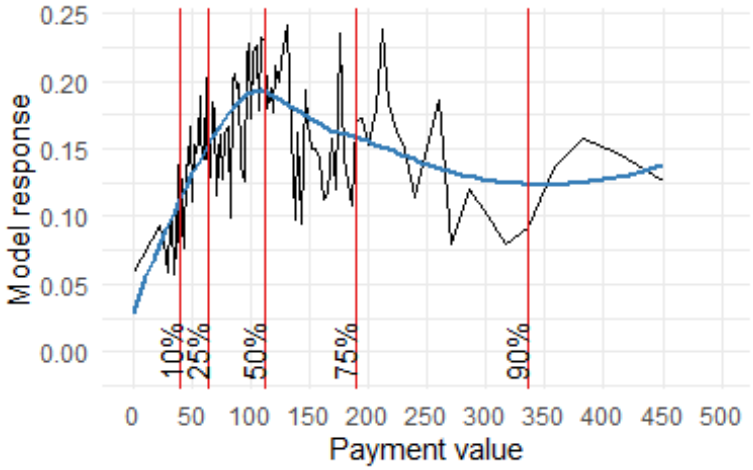


Figure 11: PDP plot for payment value of the purchase.

Model response for payment value is non-monotonous. From an analysis of smoothed model response, one can say that it is increasing to the point of around 100. This means that on average, until the payment value of 100, the bigger the payment value, the model is predicting a bigger probability of placing a second order by the customer. After this threshold of 100, the probability to buy for the second time is falling slowly.

On the figure 12 similar PDP plot but for the number of items is presented. The relationship between the number of items bought in the first purchase and the probability of the second purchase is negative. One has to remember that in 80% of the orders there is only one product, while in 10% - 2 items. At the same time, the drop in the model's response between 1 and 2 items is not very abrupt, meaning that this relationship is hard to use in modeling.

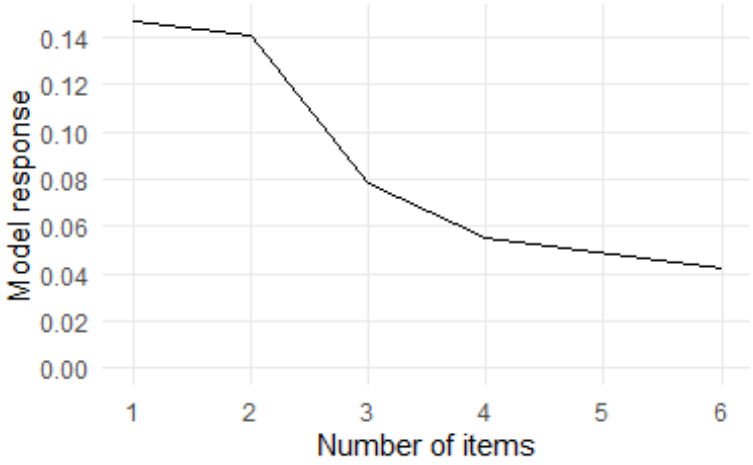


Figure 12: PDP plot for number of items in the customer's purchase.

For CRM, information about such a relationship can lead to the following trade-off. The more the customer buys in the first purchase, the bigger are the chances that they will not make a second purchase. This can have implications in cross-selling campaigns. The company can try to maximize the revenue from the first transaction by making the customer buy more, but then there is a bigger possibility that the customer will not make the second purchase.

In the case of geolocation data, I have created a 2-d partial dependence profile and visualized it on the figure 13. It can be seen that the predictions are the highest in two distinct large spots - one having its center close to Brasilia (new capital of the country), and the other one on the same latitude, but closer to the western country border.

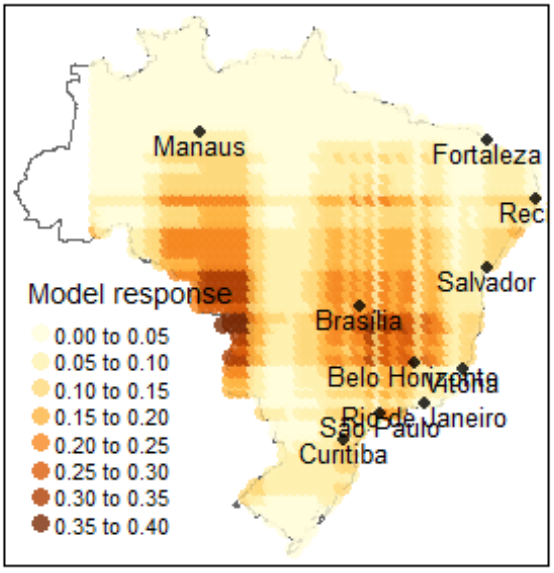


Figure 13: PDP plot for customer's location.

The predictions form a visible pattern in stripes. As was noticed by Behrens et al. (2018), it comes from the limitation of the model underlying the XGBoost method, that is decision trees. The vanilla decision tree algorithm works by partitioning the feature space on a discrete basis, and a typical output of that model on 2-d space is in the form of visible rectangles. And as XGBoost is consisting of stacked decision trees, the resulting partition pattern is a bit more complex, but still, decision-tree-typical artifacts are visible.

One the plot 14 PDP for review score is presented. Analysis of model responses in case of this variable should be treated with caution, as variable importance assessment showed it to be relatively non-important. However, I have still tried to analyze PDP plots, to check the expectation of this relationship being positive.

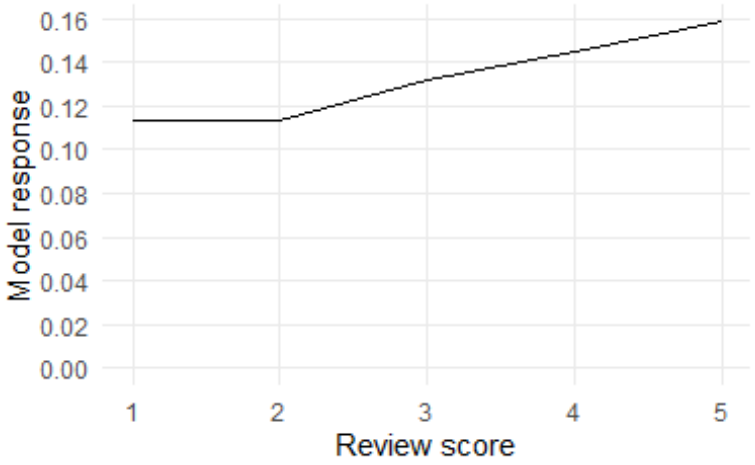


Figure 14: PDP plot for 1-5 review score.

The model response is relatively flat in reaction to changes in review score. For reviews 1 and 2, the response is not changing at all - meaning that it doesn't matter "how bad" the review is. Rather, that unsatisfied customers will not buy again in general. From 2 to 5 the model response expectedly monotonically increases.

From the analysis of both the variable's importance plot and the PDP plot, one can conclude that customer satisfaction positively influences customer's propensity to buy again. However, the strength of this relationship is very weak compared to other variables present in the model.

Summary

In this chapter the results of churn prediction were presented. After comparing a wide range of model's formulations, I have concluded that XGBoost method with variables encoding basic information about the purchase and included product categories had the best performance in terms of AUC metric. These results were robust with regards to permutation in the testing dataset.

I have also tested groups' of features influence on the model's prediction. Behavioral and geolocation features were the ones bringing the biggest improvement. Perception features - both 1-5 star review as well as textual review encoded using topic modeling did not have a big influence on the model.

CHAPTER VI

Summary

Main goal of this study was to propose a model for predicting customer loyalty in an e-commerce retail business. I have used 2 popular Machine Learning techniques, namely XGBoost and Logistic Regression. The dataset used in this study contains a wide range of variables, including transaction data, customer location, geodemographic context and customer perception about the previous purchase. To assess the importance of these features for the predictions I have applied a Permutation-based Variable Importance technique. The results reported show that transaction, location and geodemographic data are the most relevant predictors. On the contrary, customer perception proxied by the numeric review and the topic of the text review were shown to be not important. XGBoost techniqe showed its superiority over Logistic Regression judging by the Area-Under-Curve metric.

The results of this study can help companies in improving the profitability of their customer retention efforts. Because of the fact that the predictions can be made just after the first purchase, the targeting measures can be applied quickly as one does not have to gather the information about the customer for a long time.

Appendixes

6.1 Appendix A - Spatial join of census data to the main dataset

Joining of the data coming from Brazilian census and the e-commerce company soures proved to be challenging. There were multiple reasons for that:

- In the e-commerce dataset the spatial dimansion is encoded mainly in a form of ZIP codes, while in demographic dataset - in a form of microregions (a Brazilian administrative unit).
- The boundaries of zipcodes and microregions do not align.
- The customer's geolocation data has 3 columns - zip code and lat/lng coordinates. For each zip code there are multiple entries for coordinates. This probably means that the company has exact coordinates of each of their customers, but decided to not provide exact customer-location mapping in public dataset for anonimisation reasons. Because of that the boundaries of zip codes cannot be specified exactly and one has to rely on the particular points from this zipcode area.

My approach to the challenge of joining these two data sources was as follows:

1. For each of the points in geolocation dataset, establish in which microregion it is. Join the dataset for that region to OLIST geolocation dataset.
2. Group the dataset by zip code and calculate the mean of each of the features in the dataset. In this case this mean would be a weighted mean (with weight in form of "how many customers are in this area?")

1. Original data

OLIST

zip	lat	lng
1	10	70
1	11	71
2	15	80
2	16	80
2	16	81

SIDRA

microregion	population
111	120000
222	140000
333	100000

2. Add information about from which microregion is the point

zip	lat	lng	microregion	population
1	10	70	111	120000
1	11	71	111	120000
2	15	80	222	140000
2	16	80	222	140000
2	16	81	333	100000

3. Aggregate to zip code - with weigths = number of points

zip	calculation	population
1	(120000*1+120000	120000
2	(140000*2+100000	126666,7

6.2 Appendix B - reviews topics

Table 5. Topics Inferred by Attention-based Aspect Extraction

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
0	Mentions product	9720	3.2%	Reliable seller, ok product and delivery before the deadline.
				great seller arrived before the deadline, I loved the product
				Very good quality product, arrived before the promised deadline
1	Unsatisfied (mostly about delivery)	1439	2.8%	I WOULD LIKE TO KNOW WHAT HAS BEEN, I ALWAYS RECEIVED AND THIS PURCHASE NOW HAS DISCUSSED
				Terrible
				I would like to know when my product will arrive? Since the delivery date has passed, I would like an answer, I am waiting!
2	Short positive message	2270	3.6%	Store note 10

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				OK I RECOMMEND
				OK
3	Short positive message, but about the product only	1379	2.9%	Excellent quality product
				Excellent product.
				very good, I recommend the product.
4	Non-coherent topic	6339	3.6%	I got exactly what I expected. Other orders from other sellers were delayed, but this one arrived on time.
				I bought the watch, unisex and sent a women's watch, much smaller than the specifications of the ad.
				so far I haven't received the product.
5	Positive message but longer than topic 2	1194	4.5%	Wonderful

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				super recommend the product which is very good!
				Everything as advertised .... Great product ...
6	Problems with delivery - wrong products, too many/too little things in package	2892	3.8%	I bought two units and only received one and now what do I do?
				I bought three packs of five sheets each of transfer paper for dark tissue and received only two
				The delivery was split in two. There was no statement from the store. I came to think that they had only shipped part of the product.
7	Good comments about particular seller	4839	3.4%	Congratulations lannister stores loved shopping online safe and practical Congratulations to all happy Easter
				I recommend the seller ...

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				congratulations station ... always arrives with a lot of antecedence .. Thank you very much ....
8	Short message, mostly about quality of the product	3808	3.4%	But a little, braking ... for the value ta Boa.
				Very good. very fragrant.
				I loved it, beautiful, very delicate
9	non-coherent	1275	3.4%	The purchase was made easily. The delivery was made well before the given deadline. The product has already started to be used and to date, without problems.
				I hope it lasts because it is made of fur.
				I asked for a refund and no response so far
10	Short message, lots of times wrong spelling/random letters	15	9.1%	vbvbsgfbsbfs

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				I recommend ... mayor;
				Kskskskk
11	non-coherent	2614	2.5%	I always buy over the Internet and delivery takes place before the agreed deadline, which I believe is the maximum period. At stark, the maximum term has expired and I have not yet received the product.
				Great store for partnership: very fast, well packaged and quality products! Only the cost of shipping that was a little sour.
				I DID NOT RECEIVE THE PRODUCT AND IS IN THE SYSTEM I RECEIVED BEYOND PAYING EXPENSIVE SHIPPING
12	Praises about the product	2003	2.2%	very beautiful and cheap watch.
				Good product, but what came to me does not match the photo in the ad.

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				Beautiful watch I loved it
13	Short positive message about the delivery	1788	3.0%	On-time delivery
				It took too long for delivery
				super fast delivery .... arrived before the date ...

6.3 Appendix C - table of lift values for selected quantiles

Table 6. Lift values for selected quantiles. General probability of buying second time is 3.29%.

Fraction of customers	No. customers in bin	Probability in selected bin	Lift
1%	320	0.65	19.71
2%	640	0.35	10.66
3%	959	0.26	7.84
4%	1279	0.21	6.26
5%	1598	0.17	5.16
10%	3196	0.10	3.16
20%	6392	0.07	2.07
30%	9587	0.05	1.65
40%	12783	0.05	1.48
50%	15978	0.04	1.35

References

Achrol, Ravi S, and Philip Kotler. 1999. “Marketing in the Network Economy.” *Journal of Marketing* 63 (4_suppl1): 146–63.

Athanassopoulos, Antreas D. 2000. “Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior.” *Journal of Business Research* 47 (3): 191–207.

Bardicchia, Marco. 2020. *Digital CRM-Strategies and Emerging Trends: Building Customer Relationship in the Digital Era*.

Behrens, Thorsten, Karsten Schmidt, Raphael A Viscarra Rossel, Philipp Gries, Thomas Scholten, and Robert A MacMillan. 2018. “Spatial Modelling with Euclidean Distance Fields and Machine Learning.” *European Journal of Soil Science* 69 (5): 757–70.

Bhattacharya, CB. 1998. “When Customers Are Members: Customer Retention in Paid Membership Contexts.” *Journal of the Academy of Marketing Science* 26 (1): 31–44.

Biecek, Przemyslaw. 2018. “DALEX: Explainers for Complex Predictive Models in r.” *Journal of Machine Learning Research* 19 (84): 1–5. https://jmlr.org/papers/v19/18-416.html.

Biecek, Przemyslaw, and Tomasz Burzykowski. 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. CRC Press.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3: 993–1022.

Buckinx, Wouter, and Dirk Van den Poel. 2005. “Customer Base Analysis: Partial Defection of Behaviourally Loyal Clients in a Non-Contractual FMCG Retail Setting.” *European Journal of Operational Research* 164 (1): 252–68.

Burez, Jonathan, and Dirk Van den Poel. 2007. “CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services.” *Expert Systems with Applications* 32 (2): 277–88.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. “An Empirical Comparison of Supervised Learning Algorithms.” In *Proceedings of the 23rd International Conference on Machine Learning*, 161–68.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, and others. 2015. “Xgboost: Extreme Gradient Boosting.” *R Package Version 0.4-2* 1 (4): 1–4.

Choi, Duke Hyun, Chul Min Kim, Sang-Il Kim, and Soung Hie Kim. 2006. “Customer Loyalty and Disloyalty in Internet Retail Stores: Its Antecedents and Its Effect on Customer Price Sensitivity.” *International Journal of Management* 23 (4): 925.

Corner, Statistics. 2009. “Choosing the Right Type of Rotation in PCA and EFA.” *JALT Testing & Evaluation SIG Newsletter* 13 (3): 20–25.

Dalvi, Preeti K, Siddhi K Khandge, Ashish Deomore, Aditya Bankar, and VA Kanade. 2016. “Analysis of Customer Churn Prediction in Telecom Industry Using Decision Trees and Logistic Regression.” In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1–4. IEEE.

De Caigny, Arno, Kristof Coussement, Koen W. De Bock, and Stefan Lessmann. 2020. “Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network.” *International Journal of Forecasting* 36 (4): 1563–78. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.03.029.

Dick, Alan S, and Kunal Basu. 1994. “Customer Loyalty: Toward an Integrated Conceptual Framework.” *Journal of the Academy of Marketing Science* 22 (2): 99–113.

Doshi-Velez, Finale, and Been Kim. 2017. “Towards a Rigorous Science of Interpretable Machine Learning.” *arXiv Preprint arXiv:1702.08608*.

Dutkiewicz, Alicja, Barbara M. Terhal, and Thomas E. O’Brien. 2021. “Heisenberg-Limited Quantum Phase Estimation of Multiple Eigenvalues with a Single Control Qubit.” http://arxiv.org/abs/2107.04605.

Felbermayr, Armin, and Alexandros Nanopoulos. 2016. “The Role of Emotions for the Perceived Usefulness in Online Customer Reviews.” *Journal of Interactive Marketing* 36: 60–76.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Gefen, David. 2002. “Customer Loyalty in e-Commerce.” *Journal of the Association for Information Systems* 3 (1): 2.

Gregory, Bryan. 2018. “Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data.” *arXiv Preprint arXiv:1802.03396*.

Harris, Richard, Peter Sleight, and Richard Webber. 2005. *Geodemographics, GIS and Neighbourhood Targeting*. Vol. 8. John Wiley & Sons.

hcho3. 2020. “Awesome XGBoost.” https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. “An Unsupervised Neural Attention Model for Aspect Extraction.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Hong, Liangjie, and Brian D Davison. 2010. “Empirical Study of Topic Modeling in Twitter.” In *Proceedings of the First Workshop on Social Media Analytics*, 80–88.

Howley, Tom, Michael G Madden, Marie-Louise O’Connell, and Alan G Ryder. 2005. “The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data.” In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 209–22. Springer.

Jha, Mithileshwar. 2003. “Understanding Rural Buyer Behaviour.” *IIMB Management Review* 15 (3): 89–92.

Kracklauer, Alexander, Olaf Passenheim, and Dirk Seifert. 2001. “Mutual Customer Approach: How Industry and Trade Are Executing Collaborative Customer Relationship Management.” *International Journal of Retail & Distribution Management*.

Kumar, Smitha S, and Talal Shaikh. 2017. “Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest.” In *2017 International Conference on Computer and Applications (ICCA)*, 227–31. IEEE.

Kursa, Miron B, Witold R Rudnicki, and others. 2010. “Feature Selection with the Boruta Package.” *J Stat Softw* 36 (11): 1–13.

Lee, Jae Young, and David R Bell. 2013. “Neighborhood Social Capital and Social Learning for Experience Attributes of Products.” *Marketing Science* 32 (6): 960–76.

Llave, Miguel Ángel De la, Fernando A López, and Ana Angulo. 2019. “The Impact of Geographical Factors on Churn Prediction: An Application to an Insurance Company in Madrid’s Urban Area.” *Scandinavian Actuarial Journal* 2019 (3): 188–203.

Long, Hoang Viet, Le Hoang Son, Manju Khari, Kanika Arora, Siddharth Chopra, Raghvendra Kumar, Tuong Le, and Sung Wook Baik. 2019. “A New Approach for Construction of Geodemographic Segmentation Model and Prediction Analysis.” *Computational Intelligence and Neuroscience* 2019.

Lucini, Filipe R, Leandro M Tonetto, Flavio S Fogliatto, and Michel J Anzanello. 2020. “Text Mining Approach to Explore Dimensions of Airline Customer Satisfaction Using Online Customer Reviews.” *Journal of Air Transport Management* 83: 101760.

Luo, Ling, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. “Unsupervised Neural Aspect Extraction with Sememes.” In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5123–29. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/712.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv Preprint arXiv:1301.3781*.

Mozer, Michael C, Richard Wolniewicz, David B Grimes, Eric Johnson, and Howard Kaushansky. 2000. “Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry.” *IEEE Transactions on Neural Networks* 11 (3): 690–96.

Murthy, Sreerama K. 1998. “Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey.” *Data Mining and Knowledge Discovery* 2 (4): 345–89.

Nanayakkara, Shane, Sam Fogarty, Michael Tremeer, Kelvin Ross, Brent Richards, Christoph Bergmeir, Sheng Xu, et al. 2018. “Characterising Risk of in-Hospital Mortality Following Cardiac Arrest Using Machine Learning: A Retrospective International Registry Study.” *PLoS Medicine* 15 (11): e1002709.

Nie, Guangli, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. 2011. “Credit Card Churn Forecasting by Logistic Regression and Decision Tree.” *Expert Systems with Applications* 38 (12): 15273–85.

Oliveira, Vera Lúcia Miguéis. 2012. “Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques.” PhD thesis, Universidade do Porto (Portugal).

Paruelo, JoséM, and Fernando Tomasel. 1997. “Prediction of Functional Characteristics of Ecosystems: A Comparison of Artificial Neural Networks and Regression Models.” *Ecological Modelling* 98 (2-3): 173–86.

Rai, Arun. 2020. “Explainable AI: From Black Box to Glass Box.” *Journal of the Academy of Marketing Science* 48 (1): 137–41.

Schmittlein, David C, and Robert A Peterson. 1994. “Customer Base Analysis: An Industrial Purchase Process Application.” *Marketing Science* 13 (1): 41–67.

Sharma, Sakshi, and Maninder Singh. 2021. “Impact of Brand Selection on Brand Loyalty with Special Reference to Personal Care Products: A Rural Urban Comparison.” *International Journal of Indian Culture and Business Management* 22 (2): 287–308.

Singleton, Alexander D, and Seth E Spielman. 2014. “The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom.” *The Professional Geographer* 66 (4): 558–67.

Sun, Tao, and Guohua Wu. 2004. “Consumption Patterns of Chinese Urban and Rural Consumers.” *Journal of Consumer Marketing*.

Suryadi, D. 2020. “Predicting Repurchase Intention Using Textual Features of Online Customer Reviews.” In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 1–6. https://doi.org/10.1109/ICDABI51230.2020.9325646.

Tamaddoni Jahromi, Ali, Mohammad Mehdi Sepehri, Babak Teimourpour, and Sarvenaz Choobdar. 2010. “Modeling Customer Churn in a Non-Contractual Setting: The Case of Telecommunications Service Providers.” *Journal of Strategic Marketing* 18 (7): 587–98.

Tulkens, Stéphan, and Andreas van Cranenburgh. 2020. “Embarrassingly Simple Unsupervised Aspect Extraction.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3182–87. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.290.

Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. 2012. “New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach.” *European Journal of Operational Research* 218 (1): 211–29.

Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. 2011. “Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques.” *Expert Systems with Applications* 38 (3): 2354–64.

Wai-Ho Au, K. C. C. Chan, and Xin Yao. 2003. “A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction.” *IEEE Transactions on Evolutionary Computation* 7 (6): 532–45. https://doi.org/10.1109/TEVC.2003.819264.

Webber, Richard. 2004. “Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business.” Springer.

Yin, Jianhua, and Jianyong Wang. 2014. “A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering.” In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–42. KDD ’14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2623330.2623715.

Zhao, Yabing, Xun Xu, and Mingshu Wang. 2019. “Predicting Overall Customer Satisfaction: Big Data Evidence from Hotel Online Textual Reviews.” *International Journal of Hospitality Management* 76: 111–21.

Zhao, Yu, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren. 2005. “Customer Churn Prediction Using Improved One-Class Support Vector Machine.” In *International Conference on Advanced Data Mining and Applications*, 300–306. Springer.