University of Warsaw
Faculty of Economic Sciences

Kamil Matuszelański
Student's book no.: 387078

Modeling customer churn in 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

The main goal of this study was to propose a model for customer churn prediction in the case of an e-commerce retail store operating in Brazil. Two Machine Learning models were tested and compared, namely Logistic Regression and Extreme Gradient Boosting. Among the sets of features included in the models were transaction, location, geodemographic and perception variables. Importance of each of the set for customer loyalty was assessed using Explainable Artificial Intelligence techniques. 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

Introduction

Maintaining high customer loyalty is a challenge that most of the businesses face. Multiple studies (Dick and Basu, 1994; Gefen, 2002; Buckinx and Poel, 2005) 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. In such industries, it was shown that customer churn can be successfully predicted using Machine Learning approach (Nie et al., 2011; Dalvi et al., 2016; Gregory, 2018; Buckinx & Poel, 2005).

This study is aimed at predicting loyalty of the customers of an e-commerce retail shop operating on the Brazilian market. The main dataset used was shared publicly by the Olist company, and contains a very detailed information about 100 thousand orders made in the store from 2016 to 2018. Value of the order, products bought, textual review of the order and customer location are among available variables. This dataset was enhanced by adding some information from the Brazil Census data. This data contains information at a very detailed spatial level, and thus could be combined with customer's location to obtain information about his geodemographic environment.

A challenge that to the best of the author's knowledge was not addressed in the previous studies is churn prediction in an industry, in which a very low rate of the customers stay with the company. In the case of the company analysed, 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.

Main research hypothesis that this study tries to assess is as follows: **customer loyalty can be predicted using Machine Learning modelling using the data from the customer's first purchase only**. Besides this main hypothesis, couple of secondary hypotheses are tested. These are aimed at checking the influence of particular factors on the customer loyalty and are rooted in the results of the previous studies. These factors are the amount of money spent on the first purchase (Buckinx and Poel, 2005), categories of products bought by the customer (Mozer et al., 2000), customer location (Long et al., 2019), geodemographic data (Yu Zhao et al.,2005), rural or urban location of the customer (Jha, 2003), as well as review provided by the customer (Kracklauer, Passenheim and Seifert, 2001; De Caigny et al., 2020)

From a technical point of view, Machine Learning approach was used. Two classification algorithms were tested out, namely XGBoost and Logistic Regression. To obtain a meaningful information from the text reviews, three topic modelling techniques were tested out, both using generative statistical approach, as well es neural network modelling. To be able to answer hypotheses about the influence of variables on the target in the case of XGBoost modelling, which is an unexplainable, black-box model, XAI techniques were used. To assess the importance of particular sets of features Permutation Variable Importance was applied, while for assessing the strength and direction of the influence - Partial Dependence Profile technique.

This paper is structured as follows. First chapter is aimed at reviewing previous studies regarding churn prediction and usage of Explainable Intelligence techniques in this area. Second chapter contains a detailed description of the data sources used in later analyses, as well as Exploratory Data Analysis, performed to gain intuition about the features used in Machine Learning modelling. Third chapter presents methods used for feature generation, feature preprocessing as well as Machine Learning process description. A framework for testing influence of various factors on customer loyalty is also presented. In the fourth chapter, the results obtained from the modelling process are presented, as well as answers for research hypotheses are provided.

CHAPTER I

Customer churn in CRM and modelling - literature review

In the first section of this chapter a literature review of previous studies regarding customer loyalty churn prediction is presented. The second section is aimed at reviewing previous studies with regards to the variables analysed. The third section describes shortly the field of Explainable Artificial Intelligence, and advantages of usage of such approach in Machine Learning modelling.

1.1 Customer retention

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, customer retention, 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 two 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 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 cancelling 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 cancelling 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 must 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. **In the case of this study, a customer is a churner if he never bought in the shop again after the first purchase.**

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 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 et al. (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 "riskiest" 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 & 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 & Niculescu-Mizil, 2006).

Recently, the Extreme Gradient Boosting (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 (Nielsen, 2016). 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.

1.2 Variables used in previous churn prediction studies

The choice of appropriate variables to analyse is important not only for creating the best performing model. It is also beneficial for gaining intuition about the factors influencing customer churn, which can be used in other areas of CRM, not necessarily modelling. Previous churn prediction studies used a wide variety of variables to include in the model formulation. Buckinx and Poel (2005) divided them into three broad categories:

- Behavioural – describing how the customer has interacted with the company previously
- Demographic – describing the customer in terms of his inherent characteristics, independent of his interactions with the company
- Perception – describing how the customer is rating his previous interactions with the company.

The studies in which these sets of variables were used are reviewed in the following three sections.

1.2.1 Behavioural features

Behavioural 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 with the company. This type of data is very easy to obtain in most of the companies, as it is needed for accounting purposes and very often is already analysed in some areas of the company. Moreover, variables such as transaction value are understandable even by non-experts. Such data was shown to be an important predictor in churn prediction in multiple studies (for an overview see Schmittlein and Peterson, 1994).

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. The widest range of behavioural variables in churn prediction setting up to date was used by Buckinx and Poel (2005). Besides seven 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 these categories of variables are statistically significant and bring improvement to the model's predictions. Moreover, 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 that was 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.

To assess the validity of previous findings regarding the behavioural variables in an e-commerce retail context, 2 hypotheses were tested in this study. (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.**

1.2.2 Demographic and location features

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 modelling 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 race, age, or gender.

In the case of the dataset analysed in this study, the only demographic feature available is customer location. Lee and Bell (2013) argue that customer location and his neighbourhood is an important factor to consider in CRM analyses. There are multiple ways to include this kind of spatial information in modelling churn prediction. In this study, three approaches are applied:

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

Previous studies using these approaches are reviewed in the following sections.

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 behaviour than the zip code.*" On the other hand, Buckinx and Poel (2005) showed that such dummies were significant in the case 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 analysed - the region variables indicated countries rather than postcodes.

Llave, López, and Angulo (2019) also 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 shown to be significant.

A hypothesis worth checking in the case of this study is **if the propensity to churn can be explained directly 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 obtained from the census data were significant in the churn prediction model.

In this study, the following hypothesis is assessed regarding geodemographics features: **The demographic and social structure of the customer's environment can serve as a useful predictor of churn tendency.**

Rural vs. urban customer location

Generally, there is a consensus among researchers that there is a difference in customer behaviour 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 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.**

1.2.3 Perception features

Customer perception of the company is considered an important factor driving customer loyalty (Kracklauer et al., 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 potentially 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).

Ways of analysing 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 two studies used the data from textual reviews for churn prediction. De Caigny et al. (2020) used text embedding approach, while Suryadi (2020) - simple tf-idf technique.

Lucini et al. (2020) specifies two natural language processing areas that are usually used to extract insights from customer reviews, namely topic modelling 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, the focus is put 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. **A hypothesis is that customer perception expressed both with the numeric review, as well as with textual review can be useful predictors of customer loyalty.**

1.3 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 modelling 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 one sentence summary of what was considered while making predictions.
-

***

From the review of the previous literature presented in this chapter, one can conclude that churn prediction using machine learning approach can bring big benefits to the company, by facilitating customer targeting and obtaining richer information about the customers. In general, using non-linear, more advanced modelling approaches was shown to provide superior results. However, such techniques suffer from lack of explainability compared to simple linear or rule-based modelling. To address this issue, methods of XAI can be used.

Regarding the factors influencing customer churn, the easiest ones to use and the most covered in the previous literature are the features informing about customer's previous transactions. The evidence regarding usage of spatial dimension is mixed, but its usage is not that challenging and should be assessed. The least research was done in the area of reviews analysis. It requires advanced preprocessing, and is not guaranteed to improve the results. However, in the previous studies customer satisfaction was shown as an important factor influencing customer loyalty. Also, inferring the topic of the textual review can be also usable in other areas of CRM, for example for on-line monitoring of customer satisfaction from the company's service.

CHAPTER II

Dataset description

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

2.1 Data sources

This study utilizes data from two 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 the second dataset - 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 two forms - 1-5 score or textual review. In the dataset codebook, the authors stated that not all 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 review

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

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

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 also provided 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 the most central geolocation from the mapping table was assigned. To obtain the most central point, Clustering Around Medoids algorithm was used, 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 modelling process, the label of all the least popular categories was changed 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, the data obtained from the 2010 general census was used. 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, the following 36 variables were chosen 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

2.2 Quantitative analysis

Univariate analysis

In Tab.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. Because of that, in the case of this study, **only the first customer's purchase is analysed.**

Table 1. Sequential orders analysis

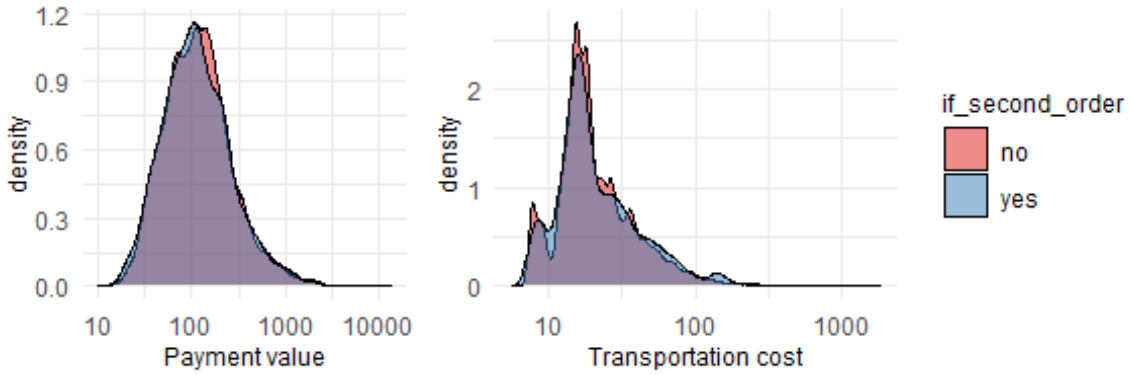
Order number	No. of orders	Mean value	Proportion from previous stage
1	96180	161	-
2	3060	150	3.18%
3	262	152	8.56%
4	49	197	18.70%
5 or more	47	101	-

Source: Own calculations based on transaction database

In Fig.1 the density estimation of the values of payment (left) and transport (right) for each order are presented. To smoothen the plots, the Kernel Density Estimation technique was used. As the distribution is highly right-skewed, the values were logarithmed. The density plot is grouped by the fact whether the customer also placed a second order later. For both variables the two

densities almost overlap. This means that payment value and transportation cost probably would not be good predictors in a univariate approach - although maybe they can be interacted with other features and start having predictive power. It can be beneficial for modelling process to use a Machine Learning method that can infer such interaction on its own, as no previous research states what features should be interacted with payment value to improve the model's performance.

Figure 1: Payment value and transportation cost



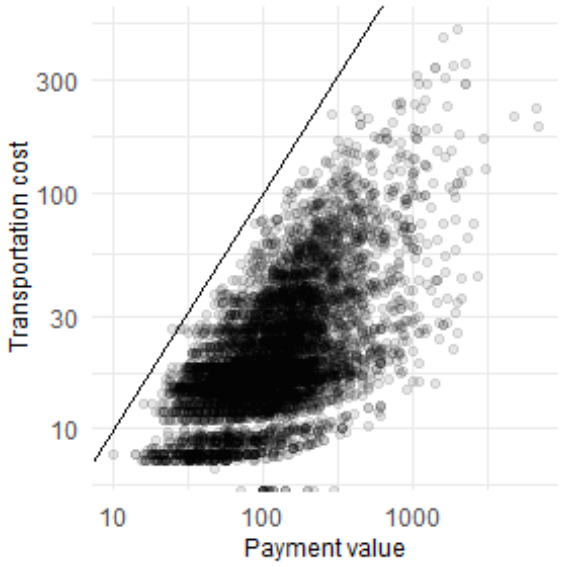
Note: x-axis is log-transformed

Source: Own calculations based on transaction database

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. These two plotted against each other in the Fig.2. Again, both axes were logarithmed.

The relationship is very clear here. For the particular value of the package, the transportation fee is seldom bigger than the value itself. A line with a slope of 1 was added 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.

Figure 2: Scatterplot of transportation value and order value

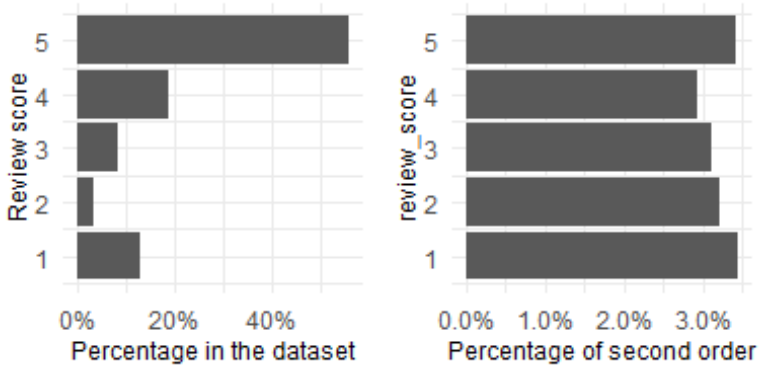


Note: Both axes are logarithmed for better plot clarity.

Source: Own calculations based on transaction database

In the Fig.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).

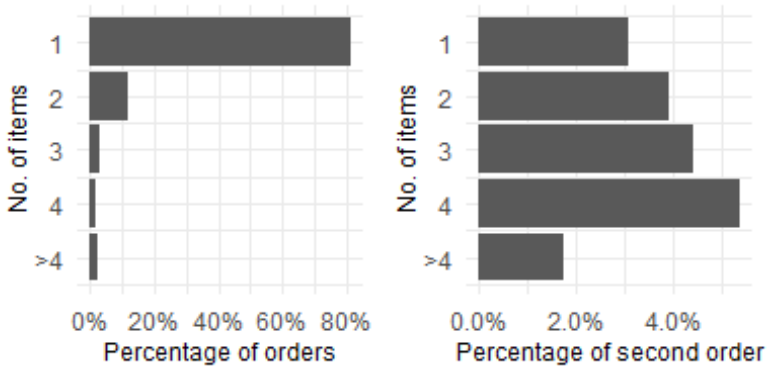


Source: Own calculations based on transaction database

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 were 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.

In the Fig.4, analysis of the number of items in the order is presented. All orders with the number of items more than 4 were binned 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).



Source: Own calculations based on transaction database

A trend is visible - the more items the customer has bought in the first order, the more likely he 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 Tab.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 modelling phase.

Table 2. Product categories

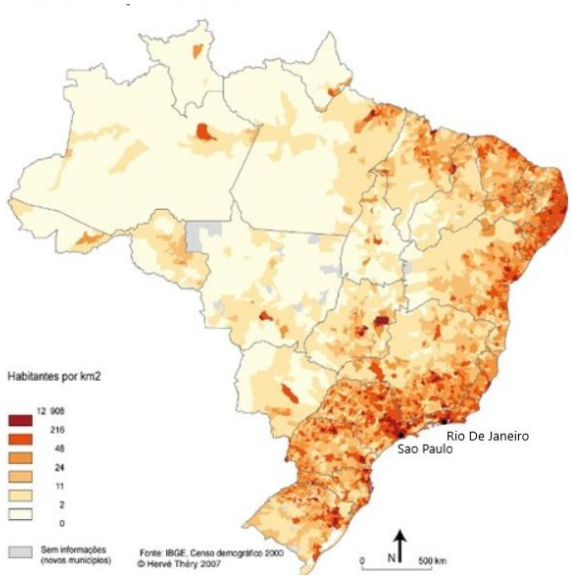
Product category	No. items	Percentage	Percentage of second order
bed_bath_table	7509	11.4%	13.8%
furniture_decor	5801	8.8%	11.5%
sports_leisure	6170	9.4%	9.4%
health_beauty	6996	10.6%	7.4%
computers_accessories	5601	8.5%	6.7%
housewares	5047	7.7%	5.8%
watches_gifts	4475	6.8%	3.8%
telephony	3512	5.3%	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%

Source: Own calculations based on transaction database

Spatial analysis

In the database, there is information about approximate customer's location. In this section analysis of spatial dimension is conducted. First, some context about Brazil's geography is introduced. In the Fig.5, a map of Brazil's population density is presented. 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 follows this density very closely (with a correlation of 93%), that is why the map of customers density is not included.

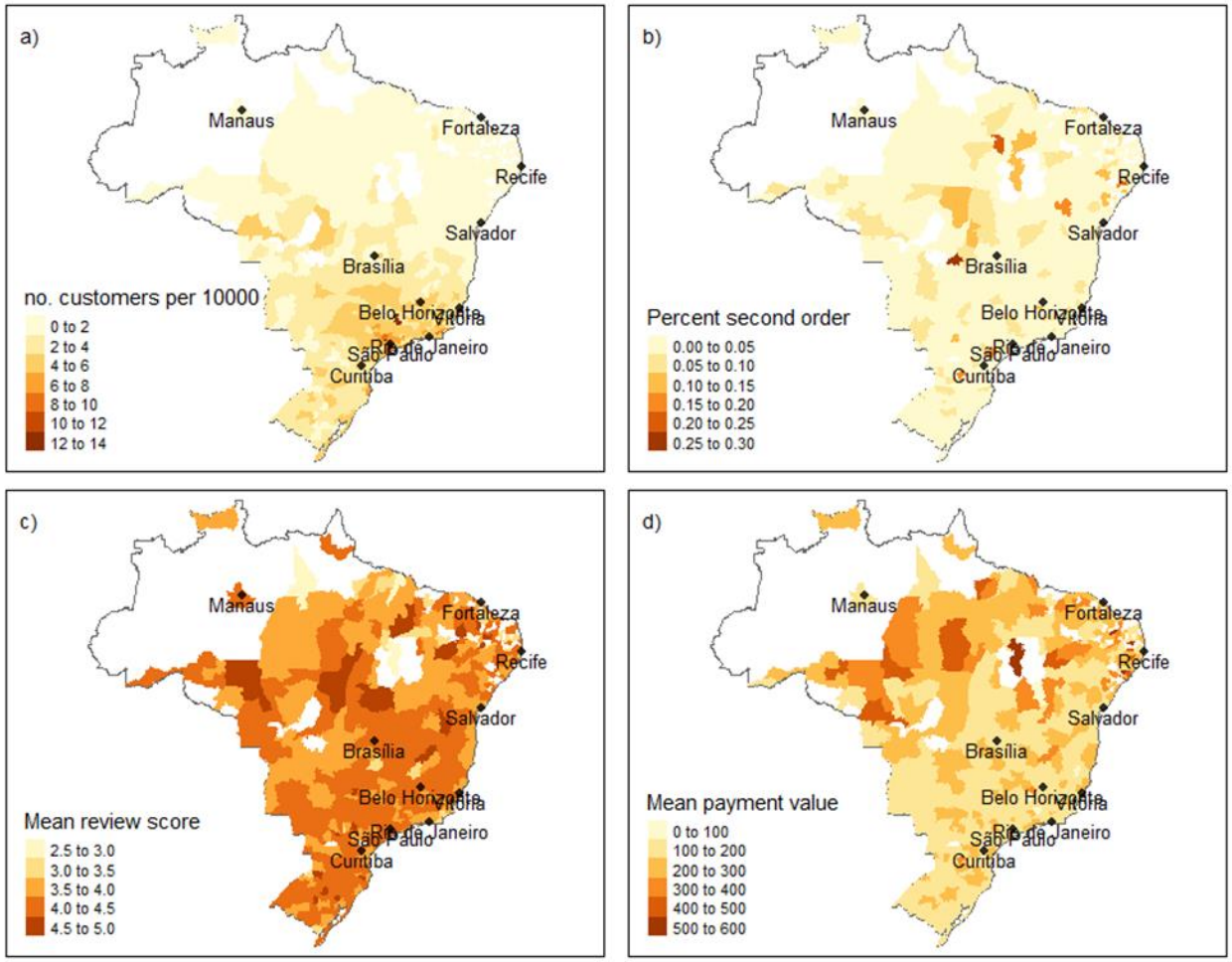
Figure 5: Map of Brazil's population density



Source: https://www.gifex.com/detail2-en/2018-12-15-15407/Population_density_of_Brazil.html [access 05.06.2021]

On the maps in the Fig.6, 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 these microregions, in which the number of customers was less than 5 were removed from the map.

Figure 6: Cartograms presenting customers statistics by microregion: a) no. customers per 10 thousand inhabitants, b) percentage of customers that placed second order, c) mean review score, d) mean payment value



Note: regions with less than 5 customers were removed as the summary statistics would be biased

Source: Own calculations based on transaction and customer geolocation databases

The top-left map (see Fig.5a) 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 (see Fig.5b) shows percentages of customers that placed a second order in each microregion. It 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 (Fig.5c) - there is no clear pattern visible.

Mean transaction value (Fig.5d) 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.

***

This chapter showed description of the two data sources used in this study, as well as basic statistical analysis of the variables available in these sources. There are few main conclusions from the Exploratory Data Analysis. Firstly, there is a very low percentage of loyal customers - only 3.3% of customers placed the second order. This finding means that the classification problem is highly imbalanced, and appropriate Machine Learning techniques for handling this issue should be used. They are described in greater detail in the next chapter. Second, 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. This also gives primary evidence to support the hypothesis, that the products bought by the customer influence his loyalty. The number of items in the placed order also gives some differentiation in terms of value of the independent variable. This finding means that including this variable in the machine learning model specification can potentially lead to better model performance. Findings from this chapter can directly help in CRM efforts by providing intuition about loyal customers – for example, it can be already stated that customers buying products from certain categories don’t make a second purchase. This should be investigated, as maybe these products are of inferior quality and this is the reason why the customers leave.

CHAPTER III

Methods description

Main hypothesis of this study is that customer loyalty can be successfully modelled. Secondary hypotheses are aimed at testing if there is influence of various predictor variables on the customer churn. To answer the main hypothesis, Machine Learning modelling should be performed. This requires the following steps:

- Preprocessing the variables present in the dataset so that they can be included in the model
- Defining Machine Learning modelling methods to be used, in particular choice of the metric to be optimized and the type of the model
- Training the model using various sets of variables, and selection of the independent variables that maximize the performance of the proposed model.

Then, to answer secondary hypotheses, the models obtained from the previous step can be analysed, to check which variables they use in making predictions about the unobserved variables.

The methodology used in this study can be divided into four broad categories loosely following the above procedure:

- Methods used in preprocessing applied to the variables present in the dataset
- Methods used for variable selection
- Machine Learning modelling methods - choice of model, cross-validation, upsampling, etc.
- Methods used for answering the hypotheses about the strength of variable's influence.

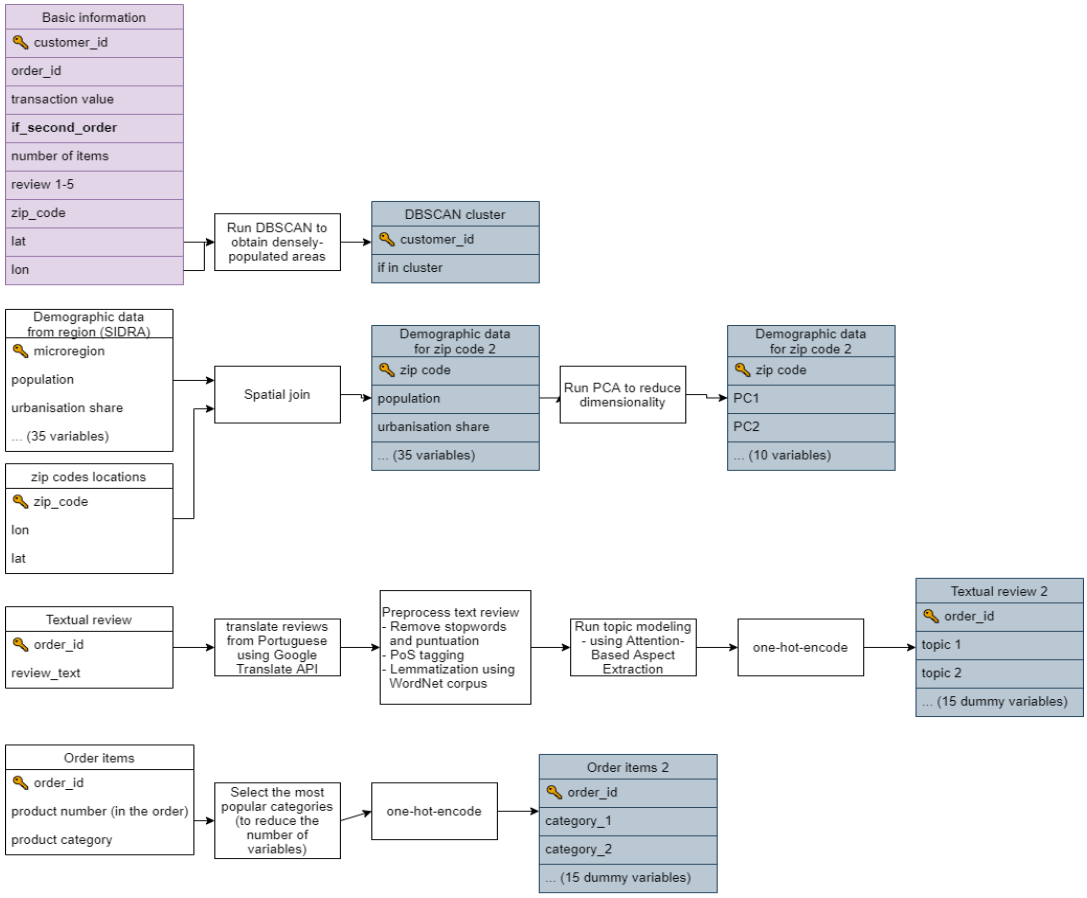
In the following sections these categories are described in greater detail.

3.1 Features preprocessing

On the diagram below (Fig.7) a summary of preprocessing applied to all the parts of the dataset is presented. A detailed description of the preprocessing steps is presented in the following subsections. 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 grey. In the modelling phase, a

simple join of the basic table, and the remaining ones was performed (basic information + order items, basic information + DBSCAN cluster, etc.).

Figure 7: Diagram depicting transformations on the source tables applied to obtain the dataset suitable for modelling purposes



Source: Own work, diagram was created using draw.io software

In this study, three separate groups of variables are analysed (with details described in following sections):

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

3.1.1 Behavioural features

In this study, from the category of behavioural variables, information about the monetary value of the first purchase, delivery cost, number of items bought and categories of items that the customer bought are included into model formulation. Value of the purchase as well as product categories are of main interest, as these were also included in the models from the previous studies, and hypotheses about their influence on customer churn should be assessed.

Value of the purchase can be directly inserted to the model as it doesn't require any preprocessing. In the case of product category, there are two steps that have to be taken. First, some of the products are very rare in the dataset, and thus should be binned into one category, because of potential problems with generalization and slower model training. Second, this variable has to converted to numeric format. Thus, all product categories except the top 15 most popular ones (responsible for 80% of purchases) were binned as a new category "other." Then, the one-hot-encoding approach was utilized to create a numeric representation, with the "other" category set as a base level.[3]

3.1.2 Location features

There are multiple ways to include spatial dimension in modelling. As mentioned before, three broad approaches that were used in previous studies are analysed:

- directly including location variables
- analysing geodemographical data
- classifying customers by living in an urban or rural area

To assess if the propensity to churn can be explained by customer location, simply longitude/latitude data about each customer were directly included to the model formulation. Preprocessing applied to the next two sets of variables is described in the following sections.

[3] 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.

Geodemographics

In total, 35 demographic features for the microregion from which the customer is were included - age structure, percentage of the population in an urban area, income structure, number of immigrants[4]. These features helped to check if the social structure of the customer's environment can serve as a valuable predictor of churn tendency.

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, this part of the dataset was processed using Principal Components Analysis. This should bring some improvements in the process of Machine Learning modelling, as training the model on a smaller, compressed dataset is more resource-efficient and at the same time was shown to improve the modelling results in some cases (Howley et al., 2005).

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. The standard version is 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 modelling phase.

Rural vs. urban customer location

The last hypothesis to check regarding spatial dimension and customer churn that was also assessed in the previous studies is whether loyalty of the customers changes based on whether they live in a rural area.

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

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. It is worth noting that even if such inferred feature will not give improvement in churn modelling, it can be appended into CRM system tables, and used in other kinds of customer analysis.

In this study, Density-Based Spatial Clustering with Noise (DBSCAN) algorithm was used for the task of rural vs. urban areas classification. DBSCAN is a density clustering method. Intuitively, this means that a points belongs to a cluster, when the points are densely located around this point. And other way, the points in which neighbourhood there aren't many points are considered as noise.

The algorithm 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 (ϵ). A detailed working of the algorithm is as follows:

1. Create a list *points_to_visit* with all the points from the dataset
2. Assign all points as noise
3. Select a point *x* randomly from *points_to_visit* and remove it from the list. Check, how many points have distance to it less than ϵ. If this number is more than k, a new cluster is created that includes all these points. Assign all these points to a list *cluster_points*. For each of the points from this list, repeat recursively step 3, until *cluster_points* does not contain any points.

4. After the previous cluster was created, select the next random point from *points_to_visit* and repeat step 3, until all points were visited.

DBSCAN, 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.

A typical rule-of-thumb for deciding k and ϵ 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, ϵ is actually a physical distance between two locations. That is why one can set what should be more reasonable criteria for constituting clusters.

In this work, the minimal number of customers in the cluster was set to 100, and the maximum distance between the customers in one cluster is 50 kilometres. For the location of Brazil on the globe, this transfers roughly to epsilon=0.2.

3.1.3 Perception features

In the case of the dataset analysed in this study, there are two 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. These two can help validate the hypothesis that the customer satisfaction is important factor in churn prediction. Using numeric review in the modelling is straightforward and doesn't require further explanation. In the next sections, the preprocessing of textual reviews using topic modelling is described in greater detail. It is worth noting that this study is a first attempt to use the topics inferred from the reviews in churn prediction. Similarly to inducing the indicator of rural vs. urban area described before, generated reviews topics can help in other CRM efforts. For example, the topics of the reviews can be included in a reviews monitoring system, and enable marketing specialists to analyse customer's reviews in a more automated manner.

Previous research in topic modelling

A short review of previous research aimed at analysing customer's topics reviews is presented below.

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) argue that short texts (as in the case of customer reviews) comprise of a very small number 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 modelling 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[5] model.

Other studies using similar techniques were conducted by Tulkens and Cranenburgh (2020) and Luo et al. (2019). In both studies, the algorithms based on words embedding outperformed the LDA method in the task of short text topic modelling.

[5] Words "Aspect" and "Topic" are often used interchangeably in the NLP literature

Text reviews preprocessing in this study

In this study, three algorithms for topic modelling were tried and evaluated. After applying each of these methods, one obtains assignments of each of the reviews to one of the topics. Such assignment can then be one-hot-encoded and included in the model specification. The models tested are as follows:

- 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.

The working of each of the methods is described in the following sections.

Latent Dirichlet Allocation is a generative statistical model. Its main assumptions are as follows:

- Consider a text corpus consisting of D documents. Each document D has N words, that belong to the vocabulary V. There are K topics.
- Each document can be modelled as a mixture of topics. For document D, it can be characterized by a distribution of topics θ_D, that is coming from Dirichlet family of probability distributions. Each topic has its distribution of words φ_k, that is coming from Dirichlet family. Then, a genrative process aimed at obtaining a document D of length of N words $w_{1,\dots N}$ is as follows:
- To generate word at a position i in the document:
    - Sample from the distribution of topics θ_D, and obtain assignment of word w_i to one of the topics $k = 1, .., K$. This is to obtain information, from which of the topics the word should be sampled.
    - Sample from the distribution of words in topic φ_k, and obtain the word to be inserted at position i.

The parameters of θ_D for each document D, as well as φ_k for each of the topics, should be learned using some method of statistical inference. Most of the practical implementations of the

algorithm are based on Expectation Maximization method. This iterative approach is aimed at finding the local maximum of the likelihood function for the analysed dataset.

Second method, Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture, is very similar to LDA approach. However, one important difference in assumptions is present, that is, **each document has words only from one topic**. This assumption is changed, because the authors claim that usually in the case of short texts, only one topic is present. This leads to the following generative process. To generate a document D:

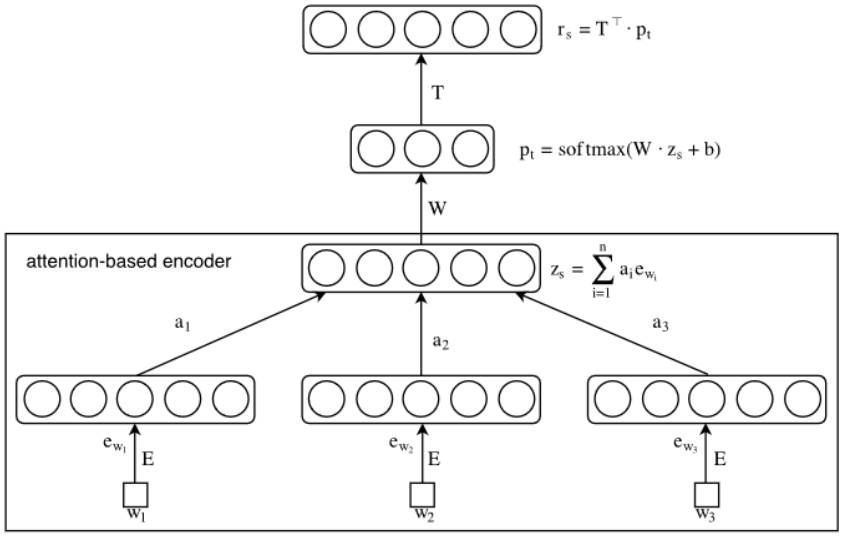
– Sample from the distribution of topics θ_D, and obtain assignment of the document to one of the topics $k = 1, .., K$.
– Sample all words from the topic distribution φ_k

Attention-Based Aspect Extraction takes a very different approach to topic modelling compared to the aforementioned methods. It is not based on a statistical model, but rather on neural network modelling. The following steps describe the model architecture presented visually in the Fig.8. For each document from the corpus:

5. Calculate word embeddings $e_{w_1}, e_{w_2}, e_{w_3}$, ... with dimensionality d for each of the words from the vocabulary based on the whole corpus. From this point, one obtains an assignment of the word w to the feature vector e_w in the feature space R^d.
6. Obtain document embedding z_s. This is done by averaging the embeddings of all the words from the document. Average is weighted by attention weights a_1, a_2, a_3 given to each of the words. These weights are to be estimated during the model training, and can be though of as a probability, that the particular word is a right word to focus on to correctly infer the main topic of the document. It is worth noticing that document embeddings share the same feature space as the word embeddings.
7. Then, calculate p_t using softmax non-linearity and linear transformation W. This vector p_t is of the same dimensionality as the number of aspects to be learned, and can be thought of as a representation of the probability, that the sentence is from the particular aspect. By taking the biggest probability of this vector, one can obtain the assignment to the particular topic.
8. Increase the dimensionality of vector p_t to the original dimensionality d by transforming it with aspect matrix T. Vector r_s is obtained.

9. The training of the model is based on minimising the reconstruction error between the vectors z_s and r_s.

Figure 8: A diagram presenting Attention-Based Aspect Extraction model architecture



Source: He et al., 2017

The authors claim that this method is able to capture better, more coherent topics than LDA and its improvements. The main specified reason is as follows. In LDA, all of the words are assumed to be independent. Thus, information about the "closeness" of the meanings of words is lost, and has to be learned by assigning similar words to the same topics. This is a hard task that LDA is not optimized to. On the contrary, by using the words embedding approach, the relationships between the words are known *apriori* by the model and can be built upon. For example, even without knowing the topics present in the corpus, one would expect that the words "cow" and "milk" should indicate the same topic with high probability. Such information is present in the words embeddings that this method uses.

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

– **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. Google Translate API was used 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 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, in this paper this method is used, with the dimensionality of the vector space set to 200, and the word window is set to 10. After applying word2vec on this dataset, a matrix with m rows and 200 columns was obtained, 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, optimal hyperparameters for all 3 models were searched based on grid search. For LDA, a varying number of topics that the model has to learn was tested (3, 5, 10, and 15). For GSDMM, 2 parameters influence topics coherency in each “cluster.” The algorithm was run 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, the different numbers of topics to learn were tested out, 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), the number of hyperparameters checked needed to be limited.

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 the obtained topics were verified for coherency (whether reviews inside one topic are similar) and distinctiveness (whether there are visible differences between modelled topics). To the modelling phase, only the one-hot-encoded topics obtained from the best model were used.

3.2 Variables selection methods

To summarize, from variable preprocessing the following six sets of features were obtained:

- 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 two models, one containing only basic RFM features, and the other RFM features and also this new feature. In this study a similar approach was. Namely, first only basic features that didn't require any preprocessing were included in the model formulation. This model served as a baseline. Then, for each of the sets of the computed features, a model containing these features + basic features was estimated. Lastly, one model containing all the variables was also trained. This resulted in the following seven 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)[6]

Automatic feature selection - Boruta algorithm

To test if the approach with including whole sets of features to the training set is an optimal one, one method of automatic feature selection was also tested, 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 a 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."

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, both XGBoost and Logistic Regression classifiers were trained using these features. The rest of the fitting

[6] The model containing all variables with demographic features without PCA preprocessing was not trained. There are two 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.

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).

3.3 Modelling methods

In this study, Logistic Regression and Extreme Gradient Boosting (Chen et al., 2015) models are compared. 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 modelling 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 modelling 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. **It is expected that XGBoost should give better performance, but at the cost of losing direct explainability.**

XGBoost is based on a principle of boosting. A general idea of the working of the algorithm is as follows. Firstly, a weak model (typically classification tree) is fitted to the data. Then, predictions are made using this model, and residuals are obtained. Next model is fitted with the original independent variables, but the dependent variable is the residuals obtained from the previous model. This is repeated for sufficiently many times, and the final output of the model is the prediction made by the last model.

Regarding validation procedure, a simple train-test split of the dataset was used, with 70% of the observations belonging to the training dataset. On the training dataset, optimal

hyperparameters were chosen using 2-fold cross-validation on the training dataset. The search space is defined 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, upsampling of the minority class on the training dataset was used 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 against minority class in these cases. For example, if the dataset contains 99% of observations from the majority class, one can simply use a classifier that always predicts the majority class and obtain 99% of accuracy – although none of the observations from minority class was classified correctly. That is why the Area-Under-Curve (AUC) metric was optimized, as it weights the performance of the minority and majority classes equally.

AUC metric is based on Receiver Operating Statistic (ROC) curve. This curve is created by plotting true positive rate against false positive rate for various cut-offs of the response variable. AUC metric is defined as area under curve of the ROC metric. It has its range between 0 and 1 (the more the better). Value of 0.5 means that the model is no better than random guessing - and thus has zero predictive power. Value of 1 is obtained by the model correctly classifying all observations.

It is worth noticing that usage of imbalanced metric should not change the results in the case where the minority class is upsampled to obtain equal proportions. However, in the case of this study, upsampling was applied only on the training set. Because of that, although the results judging by accuracy and AUC on the training set should be similar, on the test set AUC is a better choice. AUC was used in both model training and evaluation on the test set to maintain consistency.

3.4 Answering the hypotheses about feature's influence using Explainable Artificial Intelligence

As stated before, in this study Logistic Regression and XGBoost algorithms for the task of churn prediction are utilized. 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. On the contrary, XGBoost 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, Explainable Artificial Intelligence (XAI) techniques have to be employed. The results of using such approach can also help in gaining intuition about what factors the loyal customers have in common.

In this study, 2 techniques of XAI are used, namely Permutation-based Variable Importance (VI) and Partial Dependence Profile (PDP). The first one can help in answering 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?". Both techniques are described in the following paragraphs.

For the Variable Importance assessment, Permutation method was used (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 must 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 must 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. The values of AUC for the feature f (AUC_f) are scaled 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 of the scaled metric 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.

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

Besides Permutation-based Variable Importance, Partial Dependence Profile (Friedman, 2000) was also used for testing feature's influence on the dependent variable. It is aimed at discovering strength and direction of the feature's influence on the model response for all observations on average. PDP is based on the Cateris Paribus technique, so this method should be described first. 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.

***

This chapter was aimed at describing the methodology used in this study. Machine Learning modelling was used for predicting customer churn. Two ML methods were used, namely XGBoost and Logistic Regression. The first one was shown in the previous studies to be superior in terms of prediction quality, however Logistic Regression is a simpler, explainable model.

There were multiple transformations applied to the dataset. First, the data coming from national census was joined, to include demographic surrounding of the customer. This data was of high dimensionality, so PCA method was applied to reduce the number of features. Second, in the previous studies it was shown that the customers coming from rural areas are more loyal. To be able to assess that, the information whether the customer is living in a densely-populated area was inferred using DBSCAN method on the geographical coordinates of the customer. Third, customer's text reviews were mined using topic modelling approach. It is worth noting that this study is the first one using topic modelling in the task of churn prediction. Three methods were used and compared, namely Latent Dirichlet Allocation, Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture and Attention-Based Aspect Extraction. The first one belongs to the category of generative statistical models, and is the one used the most in the previous NLP studies. However, it was shown that it doesn't cope well in the case of short texts. Second method is an improvement over LDA aimed at this problem. The last one, Aspect Extraction is based on a very different approach of neural network modelling, and has an advantage that it takes into account the

“meanings” of words inferred using text embedding approach. In theory, this should give better topic assignments.

It is worth noting that even if inferring population density, as well as topic review would be successful, it does not mean that these inferred variables will have an impact on the customer loyalty prediction. However, the results of these feature generation methods can be still valuable for the company, because these variables can be appended to the company’s database and serve as an input for other CRM analyses.

To infer about the importance of the variables included in the XGBoost models, as well as the direction of the influence on the customer loyalty, Explainable Artificial Intelligence techniques were used. Specifically, two XAI algorithms were applied, Permutation-based Feature Importance as well as Partial Dependence Profile.

CHAPTER IV

Results

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 analysed in a greater depth. In the last section the models are analysed to verify the hypotheses about the direction of the influence of predictors on the customer churn.

4.1 Results of the pre-modelling phase

Customer perception analysis - topic modelling

Topic modelling was aimed at extracting meaningful information from the customer's text reviews. Resulting topic assignments should help in validating the hypothesis that customer perception is important for his propensity to churn. Moreover, such data can be used in other parts of CRM, for example for live monitoring of customer satisfaction.

The topics obtained from LDA, Gibbs Sampling and Aspect Extraction methods were manually assessed. For the first two methods, the topic assignments were not coherent – the models were not able to infer topics meaningfully. The only reasonable output was produced by the last method. For some of the inferred topics, all the reviews had a similar content – for example one topic with reviews praising fast delivery ("On-time delivery"), and another – containing short positive message about the purchase ("OK"). An interesting remark is that "spam" reviews (e.g., "vbvbsgfbsbfs", "Kskskskk") were also classified into one topic. From this one can conclude that the topics were correctly inferred by the Aspect Extraction method, and the variables indicating topic assignments can potentially give some improvement to the Machine Learning model. The results of the topic modelling with examples of reviews for each topic and proposals of topic labels are presented in Appendix B.

Attention-Based Aspect Extraction was superior to Latent Dirichlet Allocation and its improved version and probably can discover better topics in the case of short texts. Nevertheless, usage of LDA is considerably easier, as this method is widely popular, with good coverage of documentation and a number of easy to apply implementations. Aspect Extraction requires some

level of expertise regarding neural network modelling to apply it. The available implementation needs some changes in the code to be able to work on other dataset than the one from the original study. Besides that, neural network model training takes couple of hours, while for LDA - only twenty minutes.

Inferring densely populated areas - DBSCAN

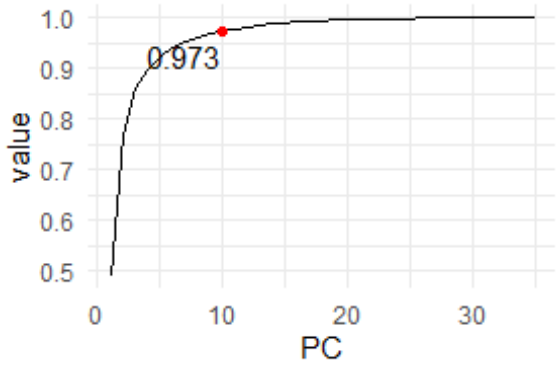
DBSCAN clustering technique was used to obtain information, whether the customer is living in rural (sparsely populated) or urban (densely populated) area. Such customer's assignments features can then be included in the machine learning model formulation, to validate the hypothesis that customers from rural areas are less prone to churn.

After a quick visual inspection of the points coloured by cluster association inferred by DBSCAN, it was visible that that the boundaries of the clusters overlap with bigger cities' boundaries, which proves that the clustering inferred densely-populated areas assignment correctly. Besides including this variable in the model formulation, it can also be appended to existing data sources in the company and used as an input to other CRM analyses.

PCA for demographic data

Dimension reduction (using PCA method) was applied to all 36 geodemographic features to reduce the number of features the models will have to learn from, while keeping most of the information from the original set of features. Cumulative variance explained by each of the consecutive loadings obtained from PCA method is presented in the Fig.9. Ten most informative PCA eigenvectors account for 97.3% of the explained variance. Such a high value of variance explained means that applying PCA transformation was successful in terms of data compression and information preservation. Thus, ten eigenvectors were included as independent variables in the modelling phase. This means that to include geodemographic information in the model, only 10 instead of 36 additional features must be included, greatly reducing model complexity and model training time.

Figure 9: Explained cumulative variance by each of the PCA loadings.



Source: Own calculations based on PCA method results

4.2 Performance analysis

AUC metric analysis

Tab.3 shows the performance comparison of all XGBoost models tested out in this study. As stated in the previous chapter, these models differ only in terms of different sets of independent variables used. The dependent variable, an indicator whether the customer has placed a second order, stayed the same in all the models tested. 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.

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.

Table 3. AUC values for XGBoost model

Model with included basic variables and…	AUC test	AUC train	Performance drop vs. the best model
Product categories	0.6505	0.9995	0.00%
All remaining variables	0.6460	0.9997	-0.68%
Ones selected by Boruta algorithm[7]	0.6426	0.9998	-1.20%
Population density indicator	0.6382	0.9993	-1.88%
Review topics	0.6353	0.9992	-2.34%
nothing more	0.6338	0.9991	-2.56%
Geodemographics (with PCA)	0.6323	0.9996	-2.80%
Geodemographics (without PCA)	0.6254	0.9995	-3.86%

Note: The models are ordered by the performance on the test set. The last column shows percentage change of performance compared to the first (best) model.

Source: Own calculations based on AUC scores obtained after applying XGBoost model

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, these parameters were tweaked in cross-validation. However, although in some cases it was possible to make the model overfit less, the performance in 2-fold cross-validation was still the best with highly over-fitted models.

[7] 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.

Tab.4 contains similar information as the previous one, 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 modelling is in general very poorly suited for this prediction task.

Table 4. AUC values for Logistic Regression model

Model with included basic variables and…	AUC test	AUC train	Performance drop vs. the best model
Product categories	0.5862	0.5922	0.00%
All remaining variables	0.5813	0.5960	-0.84%
Ones selected by Boruta algorithm	0.5801	0.5912	-1.05%
Review topics	0.5639	0.5595	-3.81%
nothing more	0.5535	0.5529	-5.58%
Geodemographics (without PCA)	0.5492	0.5632	-6.31%
Geodemographics (with PCA)	0.5482	0.5606	-6.48%
Population density indicator	0.5464	0.5532	-6.79%

Note: The models are ordered by the performance on the test set. The last column shows percentage of performance drop compared to the first (best) model.

Source: Own calculations based on AUC scores obtained after applying Logistic Regression model

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.

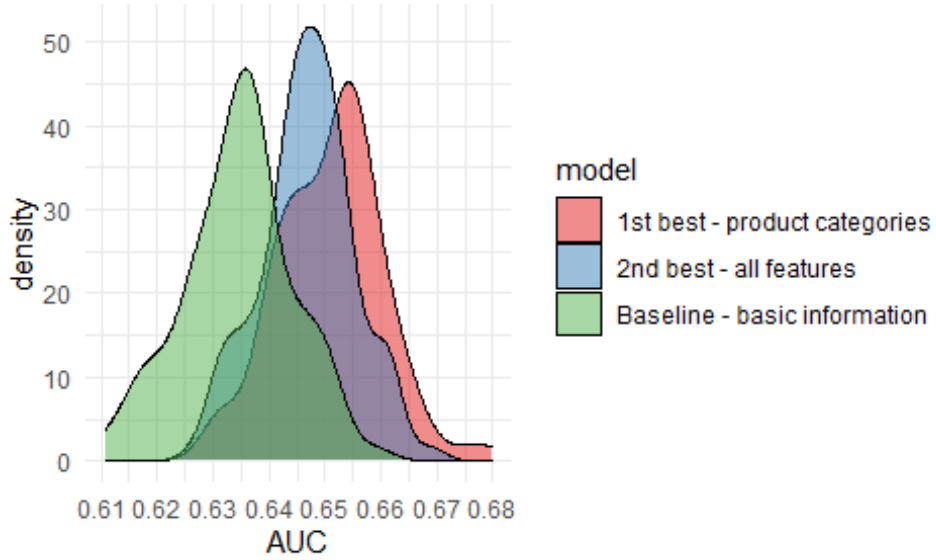
Comparison of performance for *agglomeration* set of features is particularly interesting. In the XGBoost model, this feature is rated as the 3rd best one (after excluding Boruta set to compare meaningfully with LR table). In the LR case, it is scored as the worst one. One possible explanation is that 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 modelling 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. Observations from the test set were sampled with replacement, and the AUC measure was calculated. Specifically, 100 re-sample rounds were performed, 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 Fig.10 shows density estimates of these 3 empirical AUC score distributions.

Figure 10: Bootstrap AUC estimates for 3 XGBoost models.



Source: Own calculations

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 a Kolmorogov-Smirnov test[8] was used, to check if the empirical distributions can come from the same probability distribution. The test was run 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).

[8] Kolmorogov-Smirnov (K-S) test is a non-parametric test to assess whether two empirical samples come from the same distribution. K-S statistic is calculated based on the largest distance between empirical distribution functions of both samples. Then, this statistic is compared against Kolmorogov distribution. Null hypothesis in this test is that two samples come from the same underlying distribution.

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 must 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 modelling 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.

In such an analysis, one typically divides the customers population into segments, defined as top x% of the ranking outputted by the targeting model. A procedure for calculating lift metric, for example for top 5% of customers is defined as follows:

1. Sample 5% of all the customers. Calculate share of these customers (*share_random*), that have positive response (that truly bought for the second time).

2. Using machine learning model, predict the probability to buy for the second time for all the customers. Then, rank these customers by the probability and select the top 5% with the biggest probability. Calculate share of these customers (*share_model*), that have positive response.

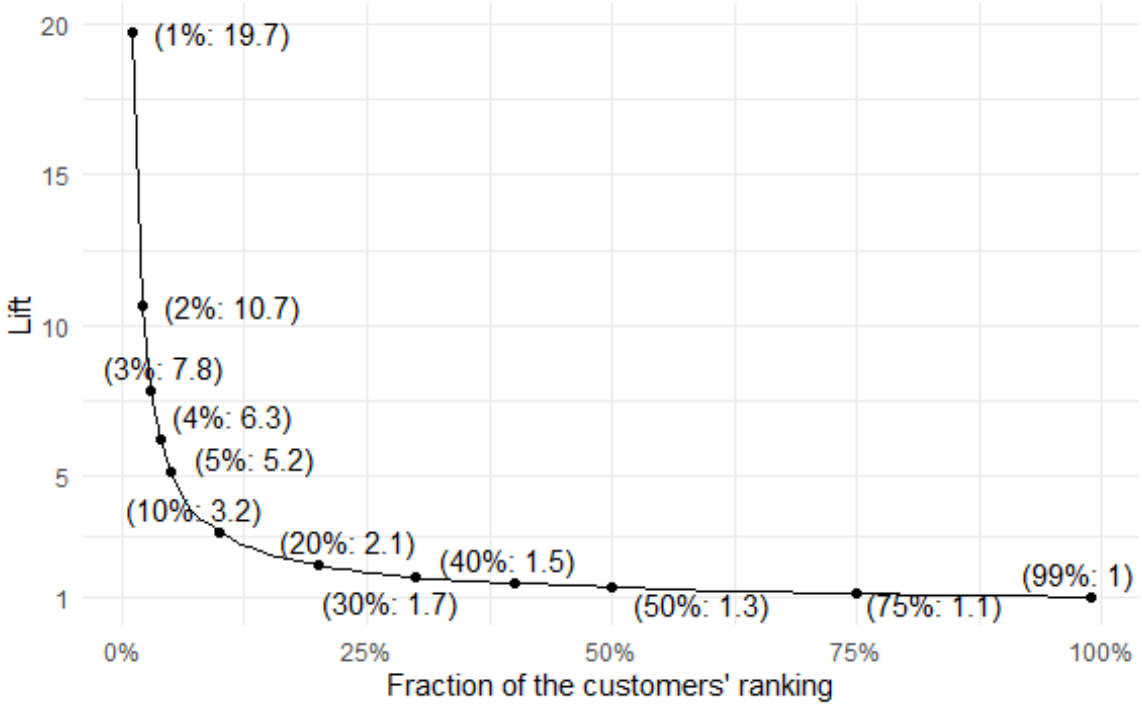
3. Calculate the lift measure as *share_model/ share_random.* If the lift value is equal to 1, this means that machine learning model is no better in predicting top 5% of the best customers than

random guessing. The bigger the value, the better the model in the case of this top 5% segment. For example, if the lift metric is equal to 3, this means that the model is three times better at targeting promising customers than random targeting.

Such calculation can be repeated for multiple customers segments, typically defined by top x% of the ranking. CRM experts can then consider lift values for various segments, combine this insight with targeting cost and make an informed decision about what percentage of the customers should be targeted.

A convenient way to visualise lift metric for multiple segments at once is a lift curve[9]. It is presented in the Fig.11. 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.

Figure 11: Lift curve



Source: Own calculations based on XGBoost predictions

[9] A table presenting the same information as the plot is included in Appendix C.

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 target half of 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.

4.3 Answering research questions about feature's influence on customer loyalty

In this section, features' influence on the customer loyalty are tested using XAI techniques – which were described in greater detail in the previous chapter. First, Permutation-based Variable Importance is employed, to check if particular sets of features have any influence on the model's predictions, and if they have, how strong. Second, Partial Dependence Profile technique is used to check the direction of this influence.

Influence of features groups on customer churn

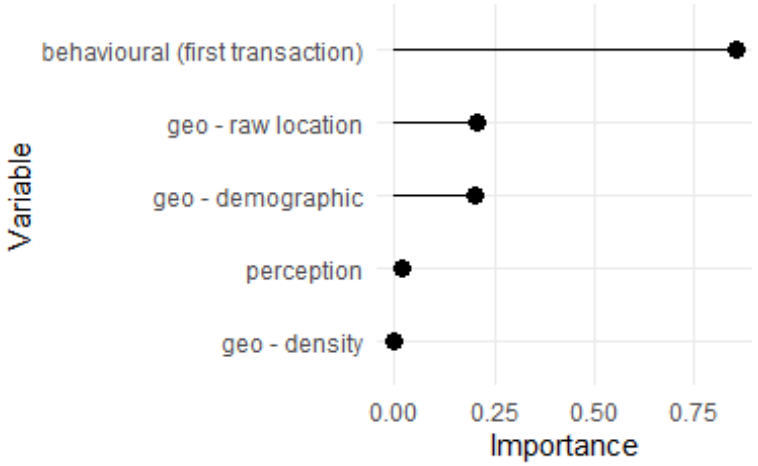
Variable Importances are assessed for two of the XGBoost models tested out in this study – one with all variables included, and another one - the model including the best set of variables (with included basic features and product categories). The reason to check variable importances for the model with all features is that it can answer the questions about the significance of the particular sets of variables.

In the Fig.12, variables' sets importances for the model with all variables are presented. The variables are grouped into 5 sets:

- 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 12: Variable importance plots for the model with all variables



Source: Own calculations

The best set of variables is the one containing behavioural 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 research hypotheses stated earlier. **Customer's propensity to churn depends on**:

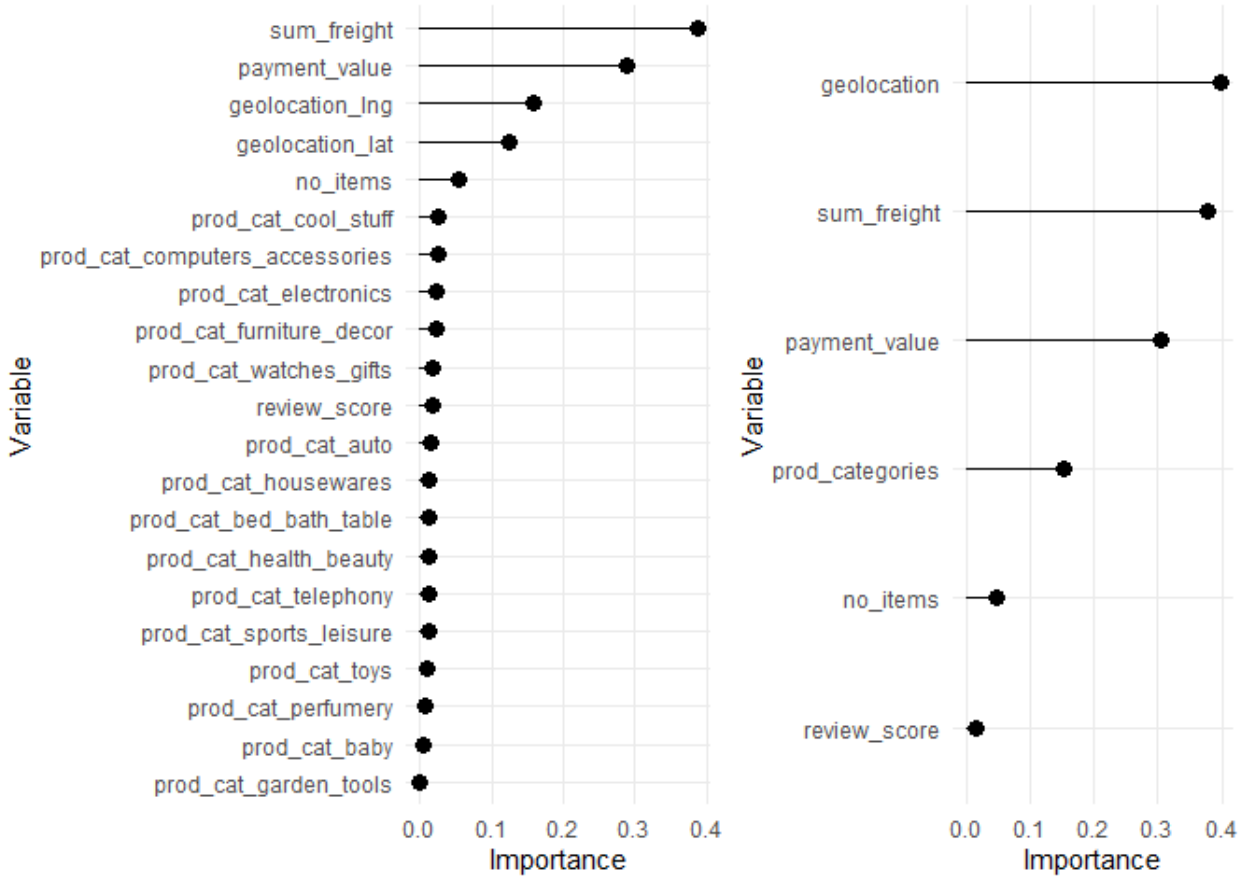
– Payment value for the first order, number of items bought, transport cost of the package
– Categories of the products bought
– Demographic environment of the customer
– Customer's location.

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

To the left of the Fig.13 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 13: Variable importance plots for the model with included product categories. Left subplot shows single variables, while right one - binned product categories and geolocation features.



Source: Own calculations

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 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, a feature importance for these variables sets

("geolocation" and "prod_categories") was used, instead of individual feature importance. This information is presented in the right subfigure.

After this operation, 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.

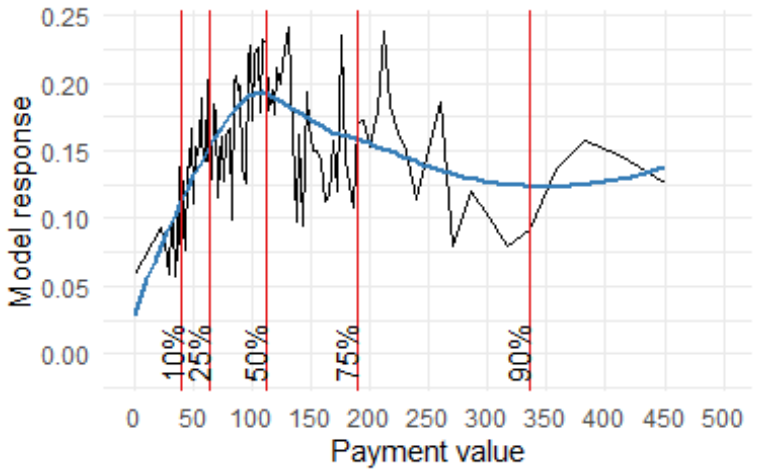
Direction and strength of features influence on customer churn

Below, PDP technique is used to test the direction and strength of the influence of various factors on customer churn. It is applied to the following features:

- Payment value for the first order
- Number of items purchased
- Customer's location
- Review score.

In the Fig.14 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 a smoothing line is provided (blue).

Figure 14: PDP plot for payment value of the purchase



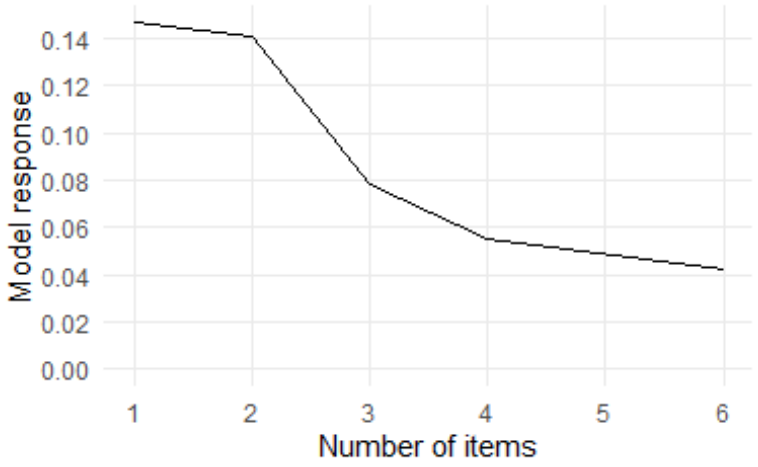
Note: Blue line is a smoothed PDP curve

Source: Own calculations

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.

In the Fig.15 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 must 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 feature on its own cannot serve as a very good discriminator of customer churn for most of the observations.

Figure 15: PDP plot for number of items in the customer's purchase



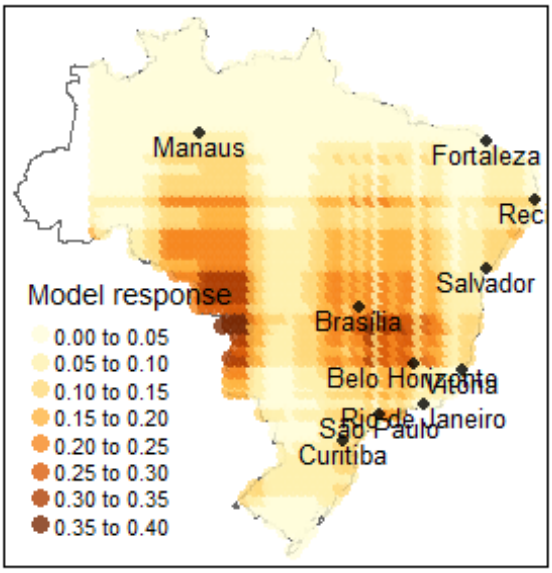
Source: Own calculations

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, a 2-d partial dependence profile was created and visualized it in the Fig.15. The predictions are the highest in two distinct large spots - one having its centre

close to Brasilia (new capital of the country), and the other one on the same latitude, but closer to the western country border.

Figure 16: PDP plot for customer's location



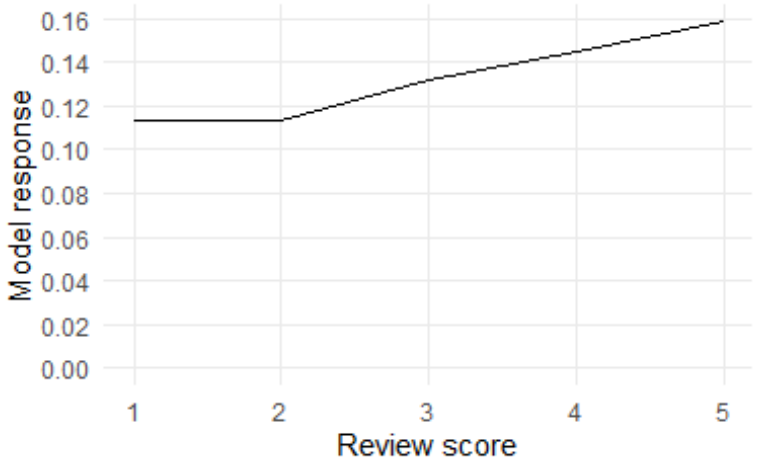
Source: Own calculations

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. A vanilla decision tree algorithm works by partitioning the feature space on a discrete basis, and a typical output of that model on a 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.

In the Fig.17, 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, it is still worth it to analyse PDP plots, to check the expectation of this relationship being positive.

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.

Figure 17: PDP plot for 1-5 review score



Source: Own calculations

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.

***

In this chapter it was shown that churn modelling can be successful even in case when only the data about customer's first purchase is available. XGBoost method had superior performance compared to Logistic Regression, most probably because it can infer non-linearities and interactions between the variables on its own. Using XAI methods, the characteristics of the customers that are most probable to make a second purchase were inferred. Such customers made a purchase with average value, bought less than three items and live in the area surrounding city of Brasilia – new country's capital. As of the categories of variables, behavioural 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 modelling did not have a big influence on the model predictions.

Three methods were used to enhance the dataset original dataset – joining census data, topic modelling of the reviews, and inferring the population density from the customer's area. The last two were inferred using dedicated algorithms from the original data.

In the task of topic modelling, Attention-Based Aspect Extraction was superior to Latent Dirichlet Allocation and its improved version. However, LDA is an easier to apply method, and thus, it should be tested out first. Only if the results are mediocre, Aspect Extraction should be tried and most probably it will improve the results.

Regarding customer density modelling, DBSCAN was able to produce meaningful results. The elbow method, that is a standard heuristic for choosing good model parameters, indicated one big cluster containing all the customers. Thus, parameters should be chosen based on empirical evaluation of the results.

Although both feature generation methods were successful, the results were shown to be not important for the task of churn prediction. However, such features can be joined to the existing CRM database to facilitate other analyses.

Summary

The goal of this study was to propose a model for predicting customer loyalty in an e-commerce retail business, as well as infer the factors driving customer loyalty.

The dataset used was obtained from Olist e-commerce retail company operating in Brazil. A wide range of variables was used in the modelling process, including transaction data, customer location, and customer perception about the previous purchase. This dataset was enhanced by adding the data from the national census. A distinctive characteristic of this company is that a very low percentage of customers decides to make a second purchase (3,3%), which made the modelling harder.

Various preprocessing and feature generation methods were applied to the dataset. Regarding feature generation, mining the texts of the reviews proved to be the most challenging. Latent Dirichlet Allocation, a go-to model for topic modelling, didn't give meaningful results. A complex, neural-network based Aspect Extraction method had to be used, for which only one draft implementation was available. In the end, although inferring the topic was successful, this feature was not important for model's predictions. The only advantage of extracting review topic can be enhancing the company's data warehouse by this new information, and using it for other CRM analyses than churn prediction.

Regarding the Machine Learning modelling used in this study, two models were tested, Logistic Regression and Extreme Gradient Boosting. XGBoost performed significantly better than Logistic Regression in loyalty prediction. This study also showed that it is possible to use the performance advantages of this more advanced method, and at the same time have an explainable model using Explainable Artificial Intelligence (XAI) methods. Specifically, two XAI algorithms were used, Permutation-based Feature Importance for assessing importance of the features, and Partial Dependence Profile for inferring the direction of features' influence. 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.

Main conclusion from this study important for CRM efforts is that customer churn can be successfully modelled and bring revenue to the company. More specifically, the predictions

obtained from the model can be used in the task of customer targeting, to target the customers that are the most possible to stay with the company.

Besides behavioural variables (that didn't require much resources investment for preprocessing), the best variables were the ones coming from the census data. This can be a hint that spatial analyses should be more invested in by the company, for example by hiring skilled experts or buying access to external spatial data sources. Such data providers are specialized in providing a very detailed and granular demographic data. Their functioning is based on combining information from different sources and as a result they can provide demographic data even to the resolution of neighbourhood.

There are some of the products categories that the customer bought in the first purchase that influence the propensity to churn negatively. It is possible that these categories are of lower quality, and the customers are simply unsatisfied with the purchase. This should be investigated further.

Discussion regarding improving the results

There are couple of areas that can be improved upon based on the results of this study. They are described in the following sections.

An area of research regarding usage of XAI that was not covered in this study was improving the Machine Learning model using the output of XAI methods. For example, it was shown using Partial Dependence Profile technique that the customers buying one or two items are more probable to make a second purchase than the ones that bought three or more times. This relationship was discovered using unexplainable XGBoost model. Such information can then be used to create a feature *if_less_than_3_items*, and this feature can be included to Logistic Regression model, and potentially improve its performance. After couple of iterations of using such technique, it is possible that Logistic Regression would become as good as XGBoost in terms of predictions quality. At the same time, Logistic Regression advantages would still be present, like inherent explainability and better speed of model training and prediction.

Spatial dimension in this study was induced by including raw coordinates and an indicator of population density in the area, as well as enhancing the dataset with demographic information. An approach that was not used, but potentially can bring prediction benefits is spatial autocorrelation analysis and Geographically Weighted Regression/Classification. Using the

methods from this family is based on the assumption that spatially close customers tend to have similar behaviours and characteristics. Another way of including spatial data is a technique called Euclidean Distance Fields, presented by Behrens (2018). It is worth noting that including raw geographical coordinates in the model (as in this study) is a special case of this method.

In this study, only two classes of Machine Learning models were tested. The reason for not including more models types was resource constraints – the training of all the models used in this study took around 30 hours. Performing models training and validation is easily parallelizable, so a potential improvement would be to use a multi-CPU cloud infrastructure and test more models in shorter time.

One should bear in mind that this study is aimed at proposing a well performing Machine Learning model. However, to be able to successfully integrate the model with the company's data environment, there are many more issues that have to be addressed. This includes (but is not limited to) writing input data validations, setting up the server hosting the model so that it can make predictions for newly coming customers, and creating a continuous evaluation system, so that the performance of the model can be monitored and checked if it doesn't deteriorate over time.

Appendixes

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

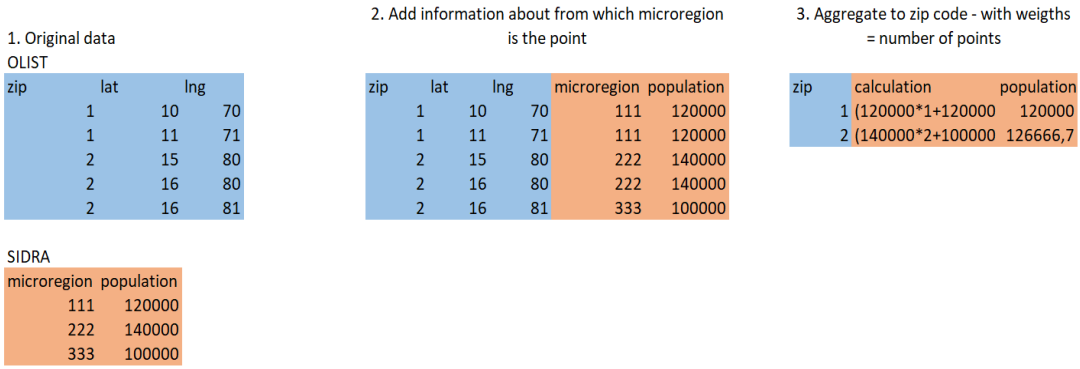
Joining of the data coming from Brazilian census and the e-commerce company sources proved to be challenging. There were multiple reasons for that:

- In the e-commerce dataset the spatial dimension 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 zip codes 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 anonymisation reasons. Because of that the boundaries of zip codes cannot be specified exactly, and one has to rely on the particular points from this zip code area.

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

10. For each of the points in geolocation dataset, establish in which microregion it is. Join the dataset for that region to OLIST geolocation dataset.
11. 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?") (An example is shown in the Fig.18)

Figure 18: Diagram depicting spatial join of Transaction and demographic databases



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.

Topic no.	Topic description	Number of reviews	percent_second_order	Example review
				Good product, but what came to me does not match the photo in the ad.
				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 ...

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*.

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.

Friedman, Jerome H. 2000. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics* 29: 1189–1232.

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.

Nielsen, Didrik. 2016. "Tree Boosting with Xgboost-Why Does Xgboost Win" Every" Machine Learning Competition?" Master's thesis, NTNU.

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.