Machine Learning for churn prediction

Kamil Matuszelański

# 1 Introduction

Konspekt, co opisać:

* Intro do problemu - co to churn, zalety predykcji
* Main hypothesis - machine learning modeling can be helpful in ranking customers by their propensity to churn

Contributions:

* Use a wide range of predictors from different categories
* Use topic modeling to extract info from text review - first usage in churn prediction ever
* Use geodemographic variables - in previous studies the results were ambiguous whether they improve prediction
* First that extensive usage of XAI tools to explain model’s predictions in churn - previous works only variable importance or more basic algorithms like LR

# 2 Literature review

## 2.1 Customer churn

### 2.1.1 Introduction

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) provides 4 areas in which CRM approaches can be of use:

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

The last one is the main focus of this study.

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

There are 2 basic approaches for the company to deal with customer churn. First is “untargeted” approach. The company seeks to improve its product quality and relies on mass advertising to reduce the churn. The other way is “targeted” approach - the company tries to address aim their marketing campaigns at the customers that are more likely to churn (Burez and Poel 2007). This approach can be divided further, by the way in which targeted customers are chosen. The company can target only these ones that have already made a decision to resign from further relationship. For example, in contractual settings this can mean canceling the subscription or breaching the contract. The other way to approach churn problem is to try to predict, which customers are likely to churn in the near future. This has an advantage of having lower cost, as the customers that are about to leave are likely to have higher demands from the last-minute deal proposed to them (Tamaddoni Jahromi et al. 2010).

As pointed out by Tamaddoni Jahromi et al. (2010) in their literature review, most of the studies concerning churn prediction were done in contractual settings. In other words, churn was defined as the client resigning from using the company’s services by canceling the subscription or breaching the contract. Such 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 non-contractual setting is the definition of churn. As there is no clear moment that the customer decides not to use the company’s services anymore, it has to be specified by the researcher based on the goals that one has to achieve from the churn analysis. Oliveira (2012) defined partial churners as the customers not making new purchases in the retail shop for the next 3 months. A different approach was used by Buckinx and Poel (2005). All the customers that had frequency of purchases below average were treated as “churners,” since these customers were shown to provide little value to the company.

### 2.1.2 Customer churn prediction

#### 2.1.2.1 Introduction

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

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

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

Churn prediction task can be decomposed into 2 main important aspects that one has to tackle. First is the decision about specific Machine Learning model that gives the best performance. 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.

#### 2.1.2.2 Machine Learning models for churn prediction

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

Recently, XGBoost algorithm (Chen et al. 2015) has been gaining popularity in multiple prediction tasks. XGBoost main strengths are 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 as a go-to standard for machine learning challenges, and very often solutions based on it achieve the best results in various competitions and benchmarks (hcho3 2020). In the context of churn prediction, XGBoost was used by Gregory (2018). It achieved superior performance compared to other techniques, like Logistic Regression and Random Forests.

#### 2.1.2.3 Variables used in previous churn prediction studies

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

##### 2.1.2.3.1 Behavioural features

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

Some of the variables studied previously are available only in some of the domains. For example, one possibility of analyzing customer behaviour in the context of e-commerce shopping is analyzing the customer behaviour while interacting with the company’s website. Koehn, Lessmann, and Schaal (2020) analyzed such click-stream sequential data using Recurrent Neural Networks. They found that such information can serve as a good predictor of customer churn. Similar studies were performed by Berger and Kompan (2019) and Yu et al. (2011), however the click stream data was aggregated and preprocessed manually, instead of using sequential modeling approach.

##### 2.1.2.3.2 Demographics features

Second category of features used in churn prediction are 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 overview see Verbeke et al. (2011)). However, availability of such predictors to use in modeling is very often limited from multiple reasons. In non-contractual settings, customers don’t have to always provide such data to the company. Moreover, usage of such personal data can be in some cases considered unethical, and lead to predictions biased against particular age or gender.

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

##### 2.1.2.3.3 Perception features

Last category of variables used for churn prediction specified by Buckinx and Van den Poel (2005) is customer perception about the company. According to Kracklauer, Passenheim, and Seifert (2001), customer satisfaction is the most important factor driving customer retention. However, although such features could have potentially high predictive power, they are usually hard to observe and quantify meaningfully. Most widely used approach is asking the customers for direct feedback using questionnaires or providing a way to post a review on the purchase. This kind of feedback can be obtained in different forms, one of them being a textual reviews. Couple of previous studies were aimed at extracting meaningful features from such reviews using different text mining methods. De Caigny et al. (2020) have used text embedding approach, while Suryadi (2020) - simple tf-idf technique. In both studies the results using such methods were superior compared to the models without including such information.

## 2.2 Explainable Artificial Intelligence

TODO: Na tą chwilę tylko konspekt. Jest świetna książka z mnóstwem przykładów i cytowań do wykorzystania na temat XAI: <https://ema.drwhy.ai/introduction.html>

* Explainability-performance trade off
* What is XAI
* Benefits of explainability
* Review of methods
  + dataset level
  + instance level
* XAI in marketing
  + 2 papery które znalazłem:
  + <https://www.sciencedirect.com/science/article/pii/S1094996820300888?casa_token=Odeol3YA4U4AAAAA:NCkaJ-yDb53CB6JWnezV-MZpcI6kHb2D_XpTapz7DDih72a6U3N3Kcxn28IIiUqrmOnSnOhf5Q> - wyzwania dla marketingu przy stosowaniu AI - nie tylko o XAI
  + <https://link.springer.com/article/10.1007/s11747-019-00710-5> - Głównie opisane zalety korzystania z XAI w modelach do marketingu

### 2.2.1 Explainability-performance trade off

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 boosting or neural networks, usually present superior performance to more basic approaches. On the other hand, their predictions cannot be explained as easily as in the case of for example Decision Trees…

# 3 Dataset description

## 3.1 Olist

TODO: Opis firmy w oderwaniu od datasetu, zamienić podpunkty na paragraf

* (<https://www.kaggle.com/olistbr/brazilian-ecommerce> access 14.03.2020)
* Brazilian company
* The dataset was published by the company for public use
* Contains information about 100 thousand orders made on the e-commerce shop site from 2016 to 2018
* Very rich dataset - contains information about the order, customer and review

In particular, there were 96180 transactions (96%) from the customers that never previously bought in this shop.

In this study I am mostly interested in analysing these transactions, and trying to predict just after first transaction, if the customer is likely to buy second time. Changing the customer attitude after buying for the second time is out of the scope of this study. One reason is the lack of data to properly conduct modeling. The other is that usually making customer buy for the second time is the hardest. In particular in this e-commerce store, in the group of the customers that bought for the first time, only 3.2% of them will buy for the second time. However, in the group of the customers that already bought for the second time, 8.6% will buy third time. The same measure is 18.7% for going from third to fourth time. This is a proof that the very first step of retaining the customer is the most important one, and further it is easier and easier to stop the churn.

The primary key in the dataset in the case of almost all features is order number. However, as in the final dataset I’m including the full information only about the first order of each customer, one observation is equal both to one order and one customer.

The Olist dataset has the following features groups:

* payment value transportation value - value of the order in Brazilian Reals excluding the transportation cost
* number of items the customer bought in particular order
* review of the order - after the finished order the customer can provide the review of the order in 2 forms - 1-5 score or textual review. In the dataset codebook the authors stated that not all of the customers in real life put any review, but this dataset was sampled in such a way that the records without 1-5 review were excluded. On the contrary, the textual review is filled only in ~50%. The data about 1-5 review can be included to the models as-is. Textual review requires however more intense preprocessing, which is described in the *methods* section of this study.
* location of the customer - the main table containing customer information contains 5-digit ZIP code of the customer’s home. The company provided also a mapping table, in which each ZIP code is assigned to multiple latitude/longitude coordinates. Probably this was done because of anonimisation reasons - so that one cannot connect the customer from the dataset with the exact house location. To obtain an exact one-to-one customer-geolocation mapping, to each zip code, I have assigned the most central geolocation from the mapping table. To obtain the most central point, I have used Clustering Around Medoids algorithm with only one cluster, and ran the algorithm separately for each ZIP code.
* products bought - the dataset contains information about how many items there were in the package, as well as 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 the variables for the modeling process, I have decided to change the label of all the least popular categories to “others.”

## 3.2 SIDRA

The dataset about the population statistics was obtained from Instituto Brasileiro de Geografia e Estatística web service called SIDRA (<https://sidra.ibge.gov.br/tabela/3548> access 26.09.2020). In this study I have used the data obtained from 2010 general census. The dataset is avaliable in aggregation to microregions (a Brasilian administrative unit, it has similar level of aggregation to NUTS 3 european classification). 558 microregions were avaliable. In particular, I have chosen the following 36 variables from the dataset:

* total population of the microregion - 1 variable
* age structure - percentage of people in particular age bin (with the width of the bins equal to 5 years) - 20 variables
* percentage of people living in rural areas and urban areas - 2 variables
* percentage of immigrants compared to total microregion population - 1 variable
* earnings structure - share of the people that earn between x0\*minimum\_wage and x1\*minimum\_wage - 11 variables

## 3.3 EDA

#### 3.3.0.1 Statistics in a table

Basic statistics

| Variable | min | Q1 | median | mean | Q3 | max |
| --- | --- | --- | --- | --- | --- | --- |
| payment\_value | 0.00 | 64.86 | 112.40 | 177.80 | 190.96 | 13664.08 |
| review\_score | 1.00 | 4.00 | 5.00 | 4.02 | 5.00 | 5.00 |
| geolocation\_lat | -36.61 | -23.59 | -22.93 | -21.23 | -20.16 | 42.18 |
| geolocation\_lng | -72.67 | -48.11 | -46.63 | -46.20 | -43.66 | -8.58 |
| no\_items | 1.00 | 1.00 | 1.00 | 1.34 | 1.00 | 21.00 |
| sum\_freight | 0.00 | 14.16 | 17.99 | 26.58 | 27.81 | 1794.96 |

TODO: opisać? Czy do appendixa??

#### 3.3.0.2 if\_second\_order

TODO: wkleić te info gdzieś do tekstu o sidra

* że 97% (?) klientów nie kupuje drugi raz
* że 50% klientów kupuje max do miesiąca po ostatnim zakupie (tabela z kwantylami?)

#### 3.3.0.3 Payment value

On the plot 1 the values of payment for each order are presented. I have used the Kernel Density Estimation technique to smoothen the plot. As the distribution is highly right-skewed, I have logarithmed the values. The density plot is grouped by the fact whether the particular customer also created a second order later. It can be seen that the 2 densities almost overlap. This means that payment value would not be a good predictor in an univariate approach - although maybe it can be interacted with other features and start having predictive power.

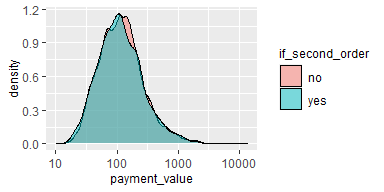


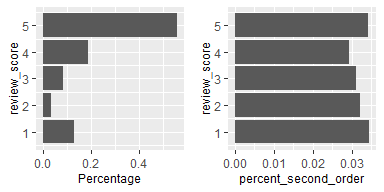
Figure 1: aaa

#### 3.3.0.4 Review score

On the plot .. 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 for her to give the lowest review.

The Relationship between making a second order and review score for the first one is somehow surprising. One would expect that if the client is unsatisfied for the first time, she will never buy in this store again. In the case of this dataset it is the opposite - the customers that gave one-star review 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 most likely for random reasons.

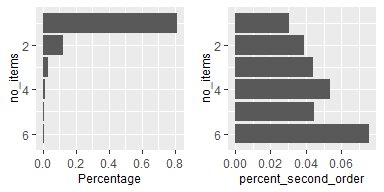
One should bear in mind that the observations avaliable in the dataset are not the complete customers data that the Olist company has. Rather, they were somehow sampled. The dataset authors claim that it represents the customers base in a complete manner. However, there was some sampling bias introduced while creating the dataset that is based on the value of the review score. Namely, in real life case customers do not have to provide star review of every order. The authors of the dataset sampled the orders database in such a way that they excluded the orders, for which the review was not given. One should bear in mind that the analysis of review score is incomplete because of that - one would wonder if there are factors that influence the customer to provide the review, and the very fact of providing the review changes the probability to buy for the second time for that particular customer.



#### 3.3.0.5 Items - numbers and percentage

On the plot .. analysis of number of items in the order is presetned. There were also orders with number of items above 6, however they make up for 0.2% of the dataset only, that is why I excluded them for clarity of the plot. 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.

A trend is clearly visible - the more items the customer has bought in the first order, the more likely she is to also put the second order. This difference is pretty strong - between 1 and 6 items the percentage increase in the response is almost 150%. However, one should bear in mind that big orders are extremely rare - 93% of the customers buy one or two items only.

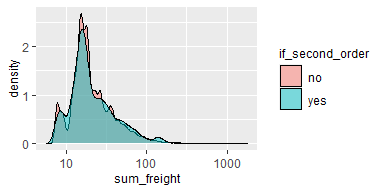


#### 3.3.0.6 Transportation cost

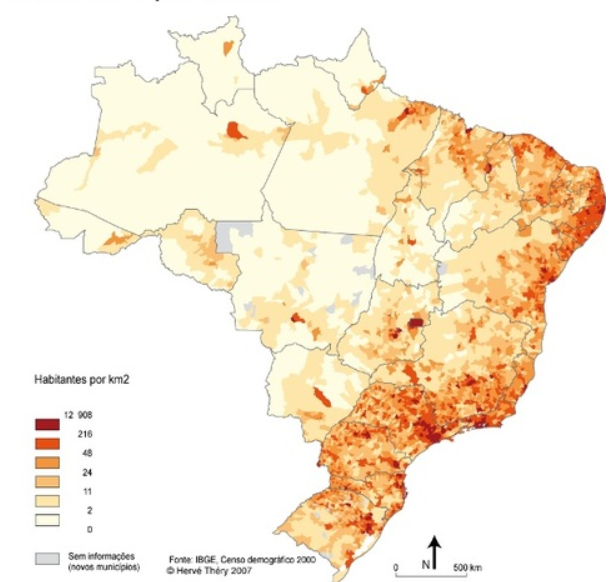
Sum of the final transportation costs that the customer had to make. The distribution is highly right-skewed, so for clarity I have log-transformed the values.

No clear distinction between two groups of observations are apparent.

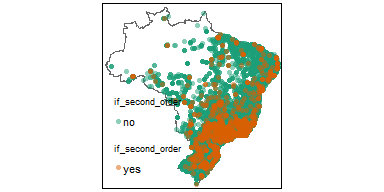
An interesting thing to check is the relationship between the value of the ordered products and the transportation cost. Pearson correlation between these two is 0.5, meaning that the value of the items ordered somehow influences the rest of the costs.

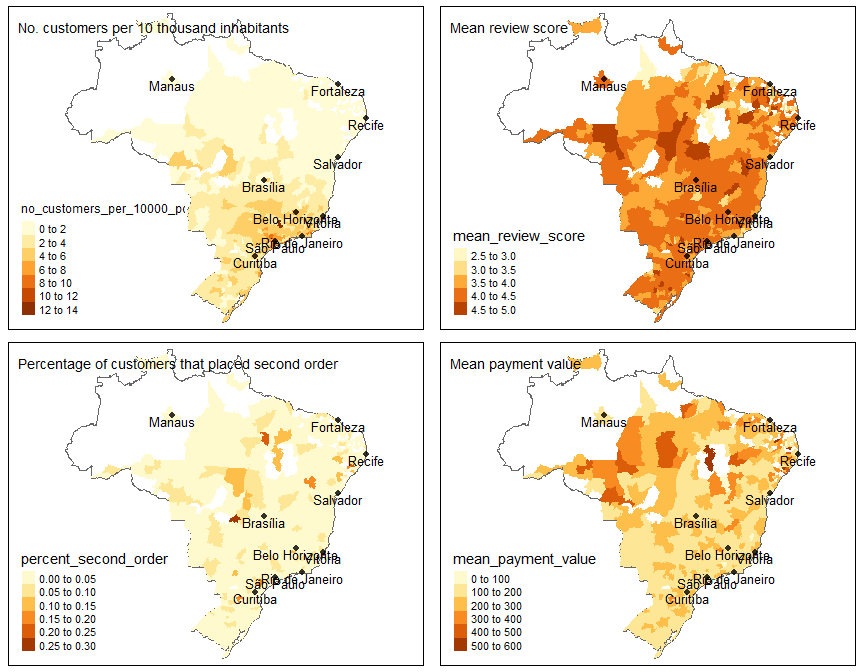


#### 3.3.0.7 Maps



On the picture .., map of Brazil population density is presented (source: <https://www.gifex.com/detail2-en/2018-12-15-15407/Population_density_of_Brazil.html>). The most densely populated areas are located in souther 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. North-western part of the country is the least populated. The distribution of the customers (as expected) follows this density, that is why I did not include the map of customers density.





On the figure .., basic statistics about spatial distribution of the features are presented - in aggregation to microregion level. Such binning is relatively coarse - because of that, some of the statistics can be not reliable in the regions with a very small number of customers. That is why I have decided to remove from the map these microregions, in which number of customers was less than 5. Because of hight correlation between number of customers in the region and total population (~93%), a more meaningful statistic than total number of customers is the number of customers per 10 thousand of inhabitants. It is presented on the plot .. . 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.

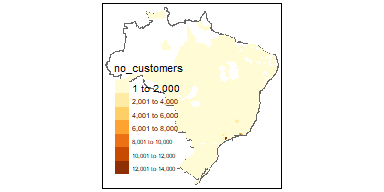
Mean transaction value is bigger in the northern, more desolated part of Brazil. One explanation could be that in these parts deliveries of the packages are more complicated/expensive/take more time, and thus the customers are more eager to place one bigger order than few small ones. Other possibility is that in the northern part the competition between e-commerce sites is smaller, and thus the customers are pushed to buiyng more items at one supplier.

It is could be argued that in the northern part of the country the people that placed second order are a bit higher. However, this relationship is rather weak. Similar things can be said about mean review score - there is no clear pattern visible.

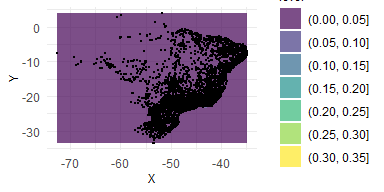
TODO: Dalsze mapy raczej do wyrzucenia, wprost wynikają z gęstości populacji ??

Number of customers

wprost wynika z gęstości populacji (korelacja 0.93), raczej do wyrzucenia

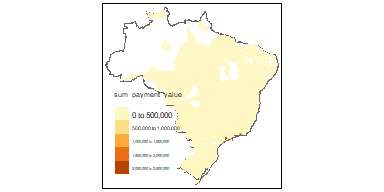


KDE of customers



prawie nic nie widać, raczej do wyrzucenia (??)

Total transactions value



Tak jak z ilością customerów, duża korelacja po prostu z gęstością zaludnienia, raczej do wyrzucenia

#### 3.3.0.8 Product categories

In the table .., 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 clearly visible. For “the best” category, it is 13.8%, while fot the worst one - only 0.029. This is a very promising result, and a signal that the dummy variables indicating product category can serve as important features in the modeling phase.

Product categories

| category | no\_items | percentage | percent\_second\_order |
| --- | --- | --- | --- |
| bed\_bath\_table | 7509 | 0.114 | 0.138 |
| furniture\_decor | 5801 | 0.088 | 0.115 |
| sports\_leisure | 6170 | 0.094 | 0.094 |
| health\_beauty | 6996 | 0.106 | 0.074 |
| computers\_accessories | 5601 | 0.085 | 0.067 |
| housewares | 5047 | 0.077 | 0.058 |
| watches\_gifts | 4475 | 0.068 | 0.038 |
| telephony | 3512 | 0.053 | 0.035 |
| garden\_tools | 3432 | 0.052 | 0.034 |
| auto | 3316 | 0.050 | 0.029 |
| toys | 3250 | 0.049 | 0.026 |
| perfumery | 2792 | 0.042 | 0.026 |
| cool\_stuff | 3041 | 0.046 | 0.020 |
| baby | 2530 | 0.038 | 0.019 |
| electronics | 2423 | 0.037 | 0.013 |

# 4 Methods description

## 4.1 Introduction

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

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

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

## 4.2 Modeling methods

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

Regarding cross-validation, I have used a simple train-test split of the dataset, with 70% of the observations belonging to the training dataset. On the training dataset, I have searched for optimal hyperparameters using 3-fold cross-validation on training dataset. I have defined search space simply as a grid of all possible combinations of the hyperparameters.

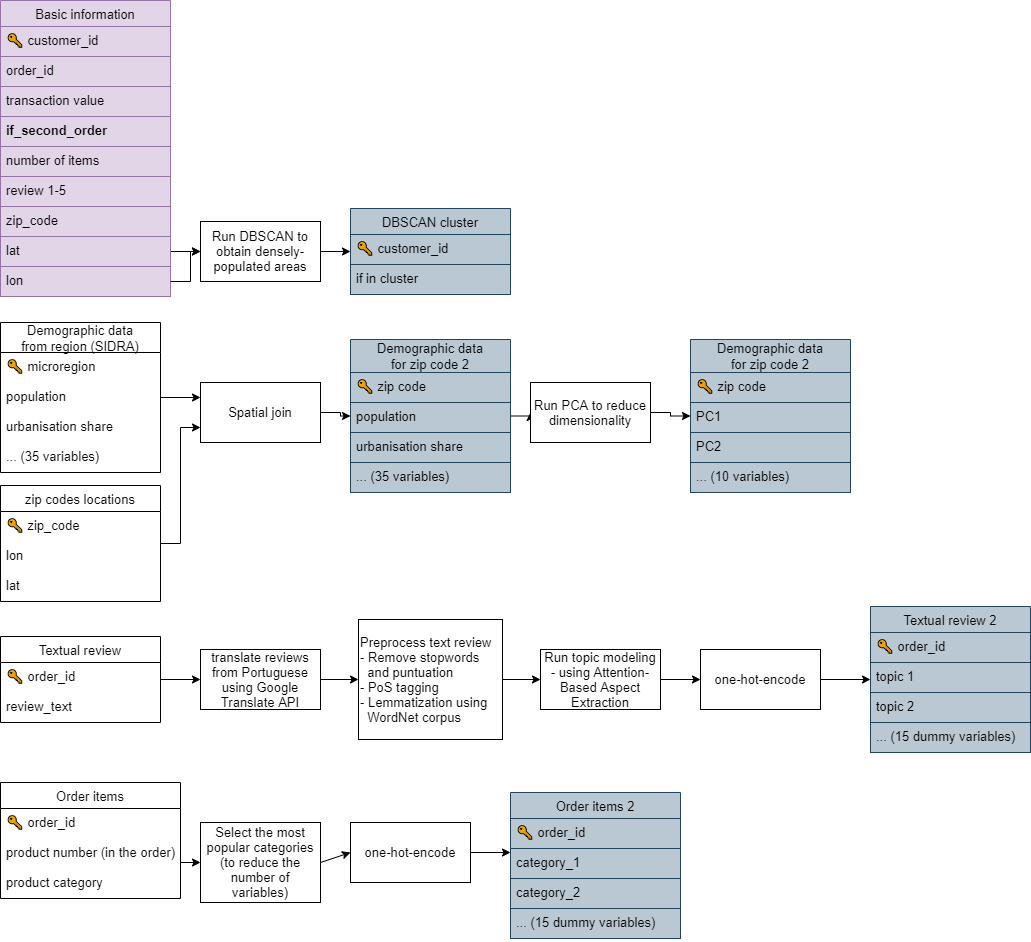
One important problem with this dataset is its very high target classes imbalance. Only 3% of the customers have decided to buy for the second time. To handle this issue I have used upsampling of the minority class on the training dataset to obtain equal class proportions. Also, the choice of an appropriate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

## 4.3 Variables preprocessing

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

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

On diagram .. a summary of preprocessing applied to all the parts of the dataset is presented. All the tables on the left-hand-side are coming directly from Olist (4 tables) and SIDRA sources (1 table - demographic data). 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 of each of the parts of the dataset are shown in gray. In the modeling phase, I have used simple join of the basic table, and the remaining ones, separately from each other (e.g. basic information + order items, basic information + DBSCAN cluster etc.).



### 4.3.1 Behavioural features

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

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

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

In the case of company analyzed in this study, the products sold belong to 74 distinct categories. At the same time, top 15 categories account for 80% of all purchases. Because of potential problems with generalization and also slower model training, I have decided to bin the least popular ones as a new category “other.” Then, I have used one-hot-encoding approach to create numeric representation, with “other” category set as a base level.[[1]](#footnote-64)

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

### 4.3.2 Location features

#### 4.3.2.1 Ways of including spatial dimension to the churn prediction

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

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

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

#### 4.3.2.2 Direct inclusion of spatial variables

To the best of authors 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 whether such data can improve the predictions. Verbeke et al. (2012) argued that “the number of times a customer called the help desk will most probably be a better predictor of churn behavior than the zip code.” On the other hand, Buckinx and Poel (2005) showed that such dummies were significant in the case of Neural Network model, but not in Random Forest. Also, Long et al. (2019) found that these dummies are significant. In these case however, a different spatial extent was analyzed - the regions variables indicated countries rather than postcodes.

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

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

#### 4.3.2.3 Rural vs. urban customer location

Generally, there is a consensus among researchers that there is a difference in customer behaviours between rural and urban areas (Sun and Wu 2004). In particular, couple of studies in 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 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 e-commerce sector. The findings from FMCG sector does not have to translate directly, as in online setting the customers are generally not limited by the availability of the brand in their area.

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

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

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

In this study I have used Density-Based Spatial Clustering with Noise (DBSCAN) algorithm for the task of rural vs. urban areas classification. This clustering algorithm besides assignment to 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 the customer is living in a densely-populated area, while the points decoded by DBSCAN as noise are the customers living in more isolated places.

DBSCAN has 2 parameters to be decided before running the algorithm. These are the minimal number of points laying close to each other that are needed to constitute a cluster (*k*), and maximal distance, at which one considers the points to lay close to each other (*epsilon*). A typical rule-of-thumb for deciding k and epsilon parameters is to first set k, and then plot k-nearest-neighbors distances. Epsilon should be then decided based on *elbow point*, where the line is bending. However, when the features are geographical coordinates, epsilon is actually a physical distance between two locations. That is why based on expert knowledge from the company one can set what should be more reasonable criteria for constituting clusters.

In my work I have decided somehow arbitrarily that minimal number of customers in the cluster is 100, and the maximum distance between the customers in one cluster is 50 kilometers. For the location of Brazil on the geoide, this transfers roughly to epsilon=0.2. However, this kind of decision in a real company setting can (and should be) consulted with the company’s domain experts in marketing.

#### 4.3.2.4 Geodemographics approach

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

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

In total, I have included 35 demographic features for the microregion from which the customer is - age structure, percentage of population in urban area, income structure, number of immigrants. These features were obtained from Brazilian statistical office SIDRA. Joining of the data coming from SIDRA and OLIST sources proved to be challenging. The details of such spatial join are presented in Appendix A.

Geodemographic features in this study are relatively high dimensional (35 variables). At the same time, one would expect that the information can be somehow compressed, because lots of the variables represent very similar concept (for example there are 20 variables encoding only age structure).

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

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 rotated loadings allows for interpretation of the loadings, but is less optimal in a sense that the variance along each loading is not maximized. I have decided that a standard one would be more suitable in case of this study, because the explainability of the input variables to the model is not as important as correctly representing the features in lower-dimensional space and thus preserving as much valuable information as possible for the modeling phase.

### 4.3.3 Perception features

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

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

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

#### 4.3.3.1 Ways of analyzing textual reviews

As stated before, text reviews can potentially serve as a rich source of information about customer satisfaction. Although text mining for customer reviews in general is an active field of research, usage of such information in the context of churn prediction is way less covered. To the best of author’s knowledge, only 2 studies used the data from textual reviews for churn prediction. De Caigny et al. (2020) have used text embedding approach, while Suryadi (2020) - simple tf-idf technique.

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

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

#### 4.3.3.2 Previous research in topic modeling

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

Hong and Davison (2010) argues that short texts (as in the case of customer reviews) comprise of very small amount of topics, usually only one. Because of that, LDA should not be used in such settings as its assumptions are violated. This claim is supported by 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 Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model, which is improvement over typical LDA. Main difference compared to the basic algorithm is an introduction of assumption, that each texts comprises of only one topic. The authors show that this algorithm provides superior performance compared to basic LDA technique in the context of short texts.

More modern approaches to topic modeling were also developed recently. A milestone in the whole NLP field was inventing an efficient way to embed words in a vector space while preserving their meaning, namely word2vec (Mikolov et al. 2013). On a basis of this method, He et al. (2017) presented an Attention-based Aspect Extraction[[2]](#footnote-72) model. At first, words embedding using Word2Vec model is created. After that, for each text in the corpus, attention weight for each word is computed using neural network with an attention layer. Then, embedding of the whole sentence is created by computing an average for all words embedding. The words are weighted by their attention weights. Last step of the procedure is creating encoder-decoder model for learning sentence aspect embedding. The reconstruction of the sentence is the linear combination of aspect embeddings, and aspect embeddings are learned by mapping sentence embedding to a lower dimensional space.

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

#### 4.3.3.3 Text reviews preprocessing in this study

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

* Latent Dirchlet 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 the similar domain of internet text reviews.

Various preprocessing steps were needed to apply all 3 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. I have used Google Translate API to change the language of them to English. This is to facilitate not only understanding the reviews, but also the NLP tools available for 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 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 case of LDA, count-vectorizing approach was applied, with removing of words which 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 python package. In both of these cases after vectorization one should obtain a matrix with n rows and k columns, where n is number of observations in the original dataset, while k - 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 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 learnt during the network training, they should be trained beforehand instead. The authors of the paper propose Word2vec technique (Mikolov et al. 2013) for learning embeddings. Following their guidelines I have used this method, setting the dimensionality of the vector space to 200. I have also applied the word window of 10. After applying word2vec on this dataset, I have obtained the matrix with m rows and 200 columns, where m stands for number of words in the dataset, and 200 is the dimensionality of the vector space chosen as a hyperparameter.

Concerning topic models training, I have searched for optimal hyperparameters for all 3 models based on grid search. For LDA, I have tested varying number of topics that the model has to learn (3, 5, 10 and 15). For GSDMM, there are 2 parameters that influence topics coherency in each “cluster.” I have run the algorithm for all 16 combinations of both parameters chosen from the values 0.01, 0.1, 0.5 and 0.9. For Attention-Based Aspect Extraction, I have manipulated with the number of topics to learn, from the values 10, 15. Unfortunately, this last model takes a very long time to run (around 3 hours per one set of hyperparameters), I have limited the number of hyperparameters checked compared to LDA model.

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

## 4.4 Variables selection methods

### 4.4.1 Sets of conceptual features (?? lepszy tytuł)

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

* basic information - value of the purchase, geolocation in raw format lat/lng, 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 population in 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 assesment of the new feature previously untested in the churn prediction was to compare 2 models, one containing only basic RFM features, and the other RFM features and also this new feature. I have used a similar approach. Namely, first I have included basic features that didn’t require any preprocessing. This model served as a baseline. Then, for each of the sets of features that I have computed, I have estimated a model containing these features + basic features. Lastly, I have created one model containing all the variables. This resulted in the following 7 feature sets tested:

* basic features
* geodemographic + basic features
* geodemographic with PCA + basic features
* agglomeration + basic features
* product categories + basic features
* review topic + basic features
* all variables - (with geodemographic features transformed with PCA)[[3]](#footnote-77)

### 4.4.2 Automatic feature selection - Boruta algorithm

To remove the human judgement about which predictors should be used, I have also used a Boruta algorithm for feature selection (Kursa, Rudnicki, and others 2010). It is widely popular among machine learning practitioners (Kumar and Shaikh 2017). The algorithm belongs to category of wrapper feature selection algorithms, and Random Forest algorithm is usually used as an underlying algorithm. It works as follows. At first, all features from the original dataset are randomly permuted. In this way one obtains a dataset with close-to-zero predictive power. Then, the resulting features are added to the original dataset and Random Forest model is trained.

Random Forest model have a built in feature importance measure, that is usually Mean Decrease Impurity. After running random forest model, for each 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 shadow features, the variable gets a “hit.”

Above procedure is repeated for preassigned number of iterations. Finally, important features that should make it to the final model are the ones that obtain less hits than preassigned n.

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

One should have in mind that 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 medium-class modern laptop. Although other wrapper algorithms also require an iterative fitting 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 Boruta algorithm in each iteration fits the model to 2\*k features (original and shadow features).

# 5 Results

## 5.1 Results of the pre-modeling phase

### 5.1.1 Boruta feature selection

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

### 5.1.2 Topic modeling (TODO)

TODO: that only one model produced meaningful results (Description of the table czy tabelka do appendixa??)

### 5.1.3 DBSCAN for long/lat data

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

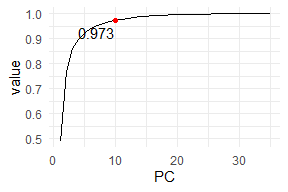
I have also tried running the clustering with the epsilon obtained from using the elbow method, namely 2. Only 2 clusters were created, of which one of them contaied 99.97% of observations. The probable reason for that is that 2 would roughly translate to 500 kilometers of cluster range, which is way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

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

### 5.1.4 PCA for demographic data

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

TODO: Zostawić wykres czy wywalić??



## 5.2 Performance analysis

### 5.2.1 Introduction (TODO)

### 5.2.2 AUC performance tables (XGB, LR)

AUC values for XGBoost model

| name | AUC\_test | AUC\_train | AUC\_perc\_performance\_drop |
| --- | --- | --- | --- |
| product\_categories | 0.6505 | 0.9995 | 0.0000 |
| all\_with\_pca | 0.6460 | 0.9997 | -0.0068 |
| boruta\_no\_topics | 0.6426 | 0.9998 | -0.0120 |
| agglomeration | 0.6382 | 0.9993 | -0.0188 |
| topics | 0.6353 | 0.9992 | -0.0234 |
| basic\_info | 0.6338 | 0.9991 | -0.0256 |
| demographics\_pca | 0.6323 | 0.9996 | -0.0280 |
| demographics | 0.6254 | 0.9995 | -0.0386 |

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

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 worse for about 2.5%.

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

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

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

Table 1:

| name | AUC\_test | AUC\_train | AUC\_perc\_performance\_drop |
| --- | --- | --- | --- |
| lr\_product\_categories | 0.5862 | 0.5922 | 0.0000 |
| lr\_all\_with\_pca | 0.5813 | 0.5960 | -0.0084 |
| lr\_basic\_info | 0.5535 | 0.5529 | -0.0558 |
| lr\_demographics | 0.5492 | 0.5632 | -0.0631 |
| lr\_demographics\_pca | 0.5482 | 0.5606 | -0.0648 |
| lr\_agglomeration | 0.5464 | 0.5532 | -0.0679 |

AUC values for the test set oscillating below 0.6 mean that the model is very poorly fitting to the data. For the worst model containing only 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 wonder if this model has any predictive power at all.

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 to 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 XGBoost model this feature is rated as the 3rd best one (after excluding Boruta set to compare meaningfully with LR table). In LR case it is scored as the worst one. One possible explanation is that it’s because of inherent ability of XGBoost to create interactions between variables, while these interactions should be included in LR model manually.

From the perspective of CRM, the most important result of the modeling procedure is that the created model has predictive power in the task of churn prediction. This means that using 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.

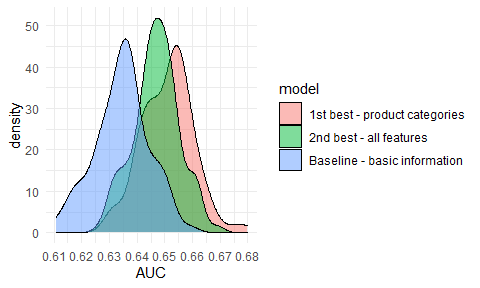
From the comparison of Logistic Regression and Extreme Gradient Boosting results, the latter one was shown to be superior over the other. Logistic Regression interpretability is an important feature that XGB lacks. Knowing what are the factors that make the customer more likely to stay is important not only for increasing trust about model predictions. Also, such information can serve as an important source of knowledge that can be passed to non-technical marketing workers in the company and potentially help them in their daily work.

* TODO: which parts of the dataset should be observed after putting model to production (drift analysis)

#### 5.2.2.1 Bootstraping AUC values

The AUC test value in the above table is shown for the whole test set. However, these point estimates cannot tell whether the performance would still be the same for sightly different test set. This is especially crucial in the case of the 2 best models - as the difference in AUC values is so small that it is not clear if it is reproducible, or just obtained by random chance.

A standard way to compare the models performance is using bootstrap technique to obtain different test sets. I have sampled with replacement observations from the test set and calculated AUC measure for the models of choice. Specifically, I have done 100 re-sample rounds, and the models chosen were the best one (product categories + basic information), second best (all variables), and the one with only basic information. The picture .. shows density estimates of these 3 empirical AUC score distributions.



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 AUC, but this should be investigated more thoroughly. That is why I have used Kolmorogov-Smirnov test to check if the empirical distributions can come from the same probability distribution. I have run the test twice using 2 alternative hypotheses. First one with H1: auc\_best =/= auc\_2nd\_best, and the second one: H1: auc\_best > auc\_2nd\_best.

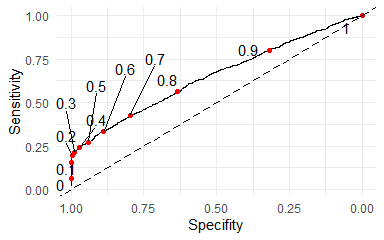
The p-value for the first hypothesis is 0.0014 This means that with the level of significance 0.05, 0.01 the performance of the models is distinguishable. At the same time, p-value with ‘greater’ hypothesis is 7^{-4}. This means that at the levels of significance 0.05, 0.01 one can say that the performance of the first model (only product categories) is better than of the second one (all variables).

Another reason to choose the model for usage in in 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 an important reason why the more complex approach should be used, the simpler is usually better. In this case, using simpler approach has the following advantages for the usage in CRM context:

* Faster inference about the new customers - especially in an online prediction setting when the predictions have to be done on the fly
* The predictions are easier to interpret. Not only the amount of factors taken into account is lower, but also the predictions are made only on the basis of the features that can be directly obtained from the purchase data. On the contrary - variables containing topic information or demographic features are somehow biased and uncertain. In the case of the review topic data it is because of the model imperfections, and in the case of demographic data because of too big generalization
* Easier model retraining and serving the final model in the IT infrastructure of the company.

#### 5.2.2.2 ROC curve

TODO: dokładnie ta sama infromacja jest w lift curve. Pytanie czy zostawić roc curve ??



## 5.3 Analysis of potential revenue - lift etc.

### 5.3.1 Lift analysis - introduction

TODO: to że prawdopodobieństwa z modelu i prawdopodobieństwa empiryczne z posortowanego rankingu klientów się rozjeżdżają wynika z tego że model jest nieskalibrowany - przewidywane prawdopodobieństwa nie są prawdopodobieństwami [link](https://towardsdatascience.com/pythons-predict-proba-doesn-t-actually-predict-probabilities-and-how-to-fix-it-f582c21d63fc). Nie wiem jak to opisać (i czy opisać) ??

An ultimate goal of customer churn prediction is gaining information, which customers are most likely to buy for the second time. From these insights, the CRM experts can make an informed decision which customers are the most likely to respond positively to targeting efforts.

As it was stated earlier in this study, a suitable class of Machine Learning models for churn prediction are the ones that are outputing probabilities of the good result. From these, one can create a ranking of customers, in which they are sorted by their likelihood to buy 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 lift analysis, and is a go-to tool for measuring the performance of targeting campaigns. Such information is also very easily understandable by the CRM experts without deep knowledge of statistics and machine learning.

A potential usages of customer ranking based on probability to churn and lift analysis include (but are not limited to):

* One strategy of marketing targeting based on lift information is targeting the customers that are the most susceptible to buying second time. This strategy can however generate unnecessary cost of trying to encourage the customers that would buy anyway. In an improved strategy, one can focus on the customers from the middle of the ranking.
* One can also vary the targeting strategy based on different cutoffs from the ranking. For example don’t do anything with top 1%, send encouraging email to top 10%, send discount to top 20%. This can serve as a way to optimize costs of more expensive targeting measures.
* It can serve as a baseline for various marketing tools in the paradigm of A/B testing - instead of computing the performance of the test group vs. control group, one can replace the performance of control group with more informed guess about the performance, namely the probability obtained from the ranking list.
* Another interesting insight that can be obtained from a well designed A/B testing experiment is the how the conversion rate of each of the customers group changes after performing some action, for example sending an encouraging email or giving a discount for the next purchase.
* A typical approach for judging the effectivity of a targeting campaign is comparing cost of particular targeting effort with potential revenue that can be gained. Only the approaches that still give some positive results should be continued. One could expect that the bigger probability outputed by the model, the more susceptible is the customer to be affected by targeting. This can be tested by using some targeting tools (e.g. gving discount) on a sample of customers selected from each quantile, and compare the performance with the control group per quantile. (TODO: przeformułować ten podpunkt bo chyba jest trochę niejasny)

The way to approach the customers is however in the hands of company’s marketing experts. Only information about the ranking of customers probabilities to buy second time is not enough to add value for the company. To this, know-how about effectiveness of various marketing tools and cost analysis of each channels that the company can influence customer’s choices is needed. Such information is often already available in the marketing departments, if not based on previous market research, then coming from the intuition of the experienced marketing department employees.

TODO: - Customer Lifetime Value for new customers - probability to buy second time from the model \* average buy second time

### 5.3.2 Lift analysis

Later, churn rates of both groups should be compared. *Lift* is defined as test group performance divided by control group performance.

1. Calculate proportion of the customers that have bought for the second time (share\_second)
2. Split the dataset into 100 sub-datasets. The split is done on the basis of quantile of the response from the model y\_hat.
3. Calculate true response rate for each subset i (share\_second\_model\_i)
4. Calculate lift value for each subset i (share\_second\_model\_i/share\_second)

Suppose the marketing campaign is meant to target 5% of the customers that has already made the first order. This equals to 1598 customers from the test set. If one would choose the customers to target randomly, on average in this group would be 3.3% of customers that would do the second order anyway. Using the output of the churn prediction, this score would be 17%, which means a performance gain of 416%. One should bear in mind that targeting the customers that would place the second order anyway is useless. However, because the model is able to predict if the customer will buy second time with higher accuracy, it can be assumed that it can also correctly predict the customers that are likely to “almost buy.”

TODO: tutaj przydałaby się argumentacja z discrete choice. Zakładamy że wynik modelu (albo rankingu) to latent variable, a decyzja czy klient kupuje jest bazowana na jakimś progu. Wtedy kiedy wiemy że model poprawnie przewiduje tych którzy kupili, to możemy zakładać że poprawnie przewiduje też tych którzy “prawie” kupili ??

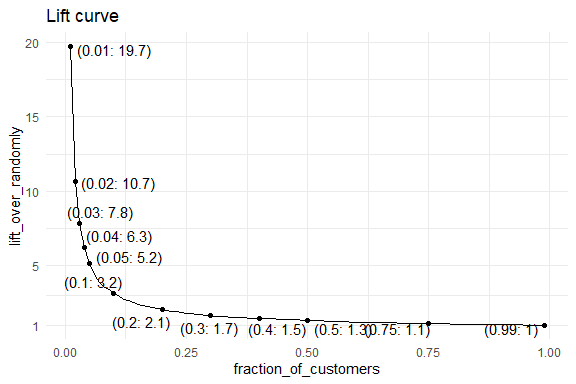
#### 5.3.2.1 Table with selected quantiles

TODO: wykres i tabela pokazują prawie te same informacje, tabelę może dać do appendixa??

Table 2:

| fraction\_of\_customers | no\_customers\_in\_bin | prob\_cutoff | score | lift\_over\_randomly |
| --- | --- | --- | --- | --- |
| 1% | 320 | 0.75 | 0.65 | 19.71 |
| 2% | 640 | 0.64 | 0.35 | 10.66 |
| 3% | 959 | 0.58 | 0.26 | 7.84 |
| 4% | 1279 | 0.53 | 0.21 | 6.26 |
| 5% | 1598 | 0.49 | 0.17 | 5.16 |
| 10% | 3196 | 0.38 | 0.10 | 3.16 |
| 20% | 6392 | 0.26 | 0.07 | 2.07 |
| 30% | 9587 | 0.19 | 0.05 | 1.65 |
| 40% | 12783 | 0.14 | 0.05 | 1.48 |
| 50% | 15978 | 0.10 | 0.04 | 1.35 |

#### 5.3.2.2 Lift curve



On the plot .., lift curve is presented. On the x axis, the fraction of the *top* customers judging by probability to buy second time is presented. On the y axis, lift value for this quantile is shown. The shape of the plot resembles the one of the function 1/x. Values of lift are very big for the smallest percentage of the best customers to target, and they are getting smaller very quickly. This means that the more customers the company would like to target based on the model prediction, the less marginal effects it would get from the usage of the model. For example, for top 1% of the customers, the model can predict retention 18.7 times better than the random targeting approach. For top 5%, it is still very effective, being 4.2 times better. However, the improvement for half of the customers is only 0.3. Although this value is less impressive, it is still an improvement over random targeting.

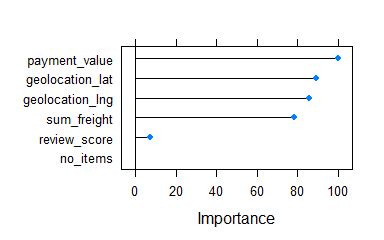
## 5.4 Model explanations

### 5.4.1 Introduction

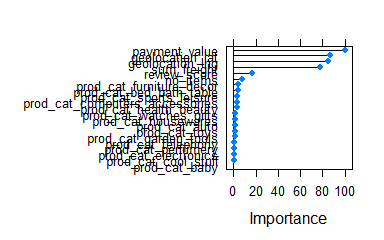
Despite superior performance in lots of machine learning tasks, the biggest drawback of Extreme Gradient Boosting models is its black-box nature, meaning lack of interpretability of predictions and parameters. This problem can be mitigated by usage of Explainable AI tools. One of such tools is ceteris-paribus analysis. The idea is to set all variables except the variable(s) of interest as constants, and create a new set of observations by manipulating only one or to variables. Such analysis can give a similar information to the values of coefficients in Logistic Regression models.

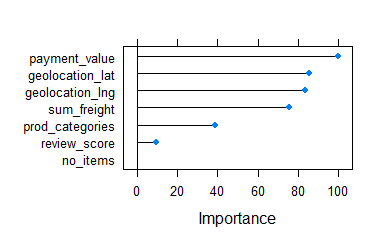
(TODO: opisać trochę więcej)

#### 5.4.1.1 Variable importances

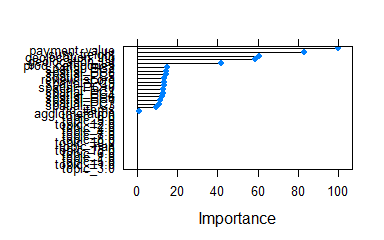


On the plot .. the variables importances from XGB model for basic variables are presented. The best one is the value of the payment. Also, high importance scores are obtained by vanilla geolocation variables and the transportation cost for the value. Surprisingly, the review score of the purchase is not very important variable. Number of items bought has the smallest variable importance. One should remember that although the shown variable importance is equal to 0, this doesn’t mean that the variable was not used in growing trees in XGB model. Rather, the variable importances are simply scaled to the range 0-100 on the plot.

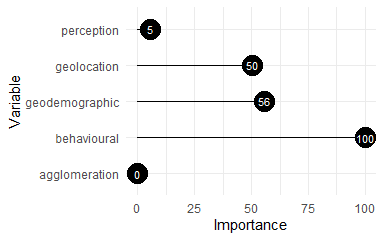




Second variable importance plot contains importances for the best model - with included product categories. Top 6 best predictors are the basic information, while the variables indicating product categories are the least important ones. Just from this information one could wonder, why despite product categories are relatively unimportant variables, they lead to 2.5% gain in AUC compared to the model without them. The reason is that all product categories variables in reality encode one variable. That is why variable importances of these variables should be treated jointly, which can be done simply by summing up the importances and scaling all importances to 0-100 range. This is presented on the plot .. . Now, product categories has become 5th most important variable, with importance bigger than review score.

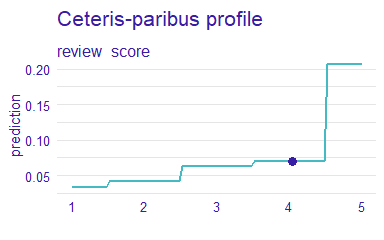


All categories of features at once: (TODO: poc VarImpPlot z ggplota, do zastanowienia czy uspójnić z resztą i w ogóle które zostawić)

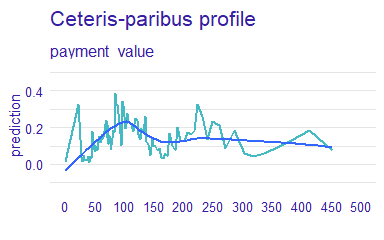


#### 5.4.1.2 Cateris paribus plots

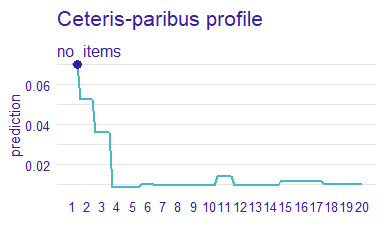
In my analysis I have decided to obtain the ceteris paribus profiles for mean values of each variable (Biecek 2018). Unfortunately this method is not able to meaningfully analyse dummy variables, that is why I have excluded the variables indicating product categories. Another variables requiring special treatment are the ones indicating geographic location. For them, I have created a 2-way ceteris paribus estimates and plotted them over the Brazil map. This way it is possible to see if it is correlated with presence of bigger cities or small population density (as in the Amazonia area).



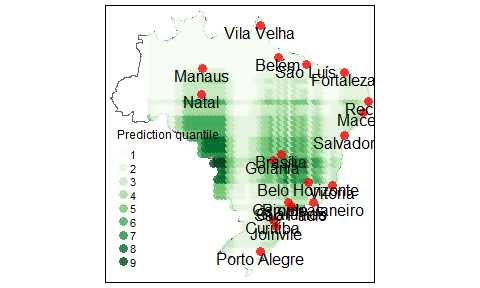
The shape of the ceteris paribus profile for the review score variable for the mean observations has an expected shape. The better the review, the it is probable for the client to purchase for the second time. What is also worth noticing is the fact that the probability increase is non-linear with respect to review score. The biggest probability increase is visible from the score 4 to 5. It is often the case in the customer reviews setting, observed also in the area of recommendation engines [(TODO: znaleźć lepsze cytowanie niż xkcd, z wykładu o recommenders)](https://xkcd.com/1098/).



Model response for ceteris paribus profile is non-monotonous. To facilitate drawing conclusions I have included smoothing line used *loess* technique. From analysis of this smoothed model response one can say that the model response is increasing to the point of around 100. This value is a median payment value in the whole dataset. After this threshold the probability to buy second time is falling up to the level of around 175, and then rise a little bit again.



Number of items bought in the first purchase negatively influences the probability of the second purchase. This effect is visible only for the smallest numbers of items, while the model response for no\_items>4 stays roughly the same.



In the case of geolocation data, I have created 2-d ceteris paribus estimates .. and visualized it on the map. To facilitate the analysis I have marked 10 most populated Brazilian cities. They are mostly located along the coast, with the exception of Brasilia (new capital of the country) and Manaus (the biggest city in Amazonia region). It can be seen that the predictions are the highest in two distinct large spots - one having center close to Brasilia, and the other one same latitude, but closer to the western country border.

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

#### 5.4.1.3 Break down of average prediction

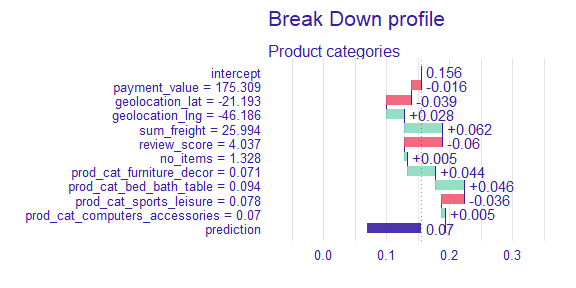
(TODO: opisać bardziej ogólnie metodę we wstępie i po co robić, na tą chwilę też opisy wyników są trochę nieczytelne, do poprawy)

Variable Attributions via Sequential Variable Conditioning - algorithm (Biecek 2018):

For breakdown of one particular observation x\_1:

1. Calculate prediction for all of the observations and calculate mean. This value is conceptually equivalent to intercept in Linear Regression
2. Change value of one variable in all observations and set it as the one from observation x\_1
3. Calculate prediction and check how much it has changed from the value obtained from intercept
4. Repeat for the rest of the variables.

Because the variables are set to a constant sequentially, the recommended algorithm is to perform a greedy search for the variable that once set constant results in the biggest change in prediction. This approach however does not facilitate comparisons between the plots as the order of the variables will change. That is why I have decided to have a fixed order of the variables, with variable importance as the main criterium.

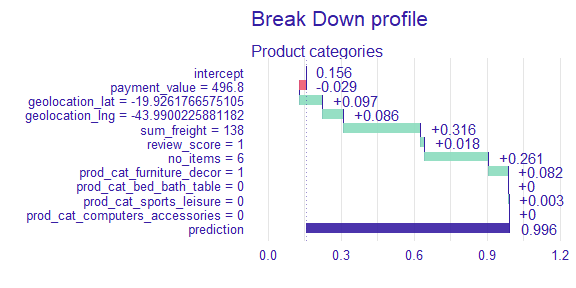


On the plot .. such break down additive plot is presented. The way to interpret it is as follows:

1. Obtain prediction for all of the observations. Calculate the mean prediction (*intercept*), in this case 0.156.
2. For all observations in the dataset, set the value of the variable *payment\_value* to 175. This is the value from the mean observation.
3. Obtain prediction for all of the observations. Calculate the mean prediction and substract it from the *intercept*. In this case the change is -0.016. This means that *payment\_value* equal to the value from the mean observation lowers the average probability that the customer will buy second time.
4. Select next variable, (*geolocation\_lat*) and repeat the process. In this case the predicted probability drops even further, to 0.101 (0.156-0.016-0.039)
5. Go through all variables. The last value on the plot, *prediction*, contains true prediction for the mean observation.

One interesting remark from this plot is that the low response of the model does not come from the fact that all variables influence it negatively. Rather, that for the variables that increase the response, there are other variables that lower the response for a similar amount and that the influences cancel out.

The prediction for the average customer at the probability is very low (0.07). This is caused by the fact that there is big imbalance in the data, and in the test dataset used for analyzing the model results there is no imbalance correction applied. From the perspective of marketing targeting more important is gaining some knowledge not about an average customer, but the ones that are the most probable to buy next time. That is why I have also created a breakdown plot for the observation that is scored as the most probable to buy second time.



In the case of the prediction for the best customer, almost all variables influence the prediction positively. The only variable doing otherwise is payment value. The influence is however minor compared to the other variables. This negative influence can be analyzed in connection with ceteris-paribus plot for that variable analyzed before. On the plot … it can be seen that the model response has a small drop for the values around 500.

The rest of the variables influence the score positively. The biggest gain is from the value of the transport, however number of items equal to 6 has also a large influence.

# 6 Summary

# 7 Appendixes

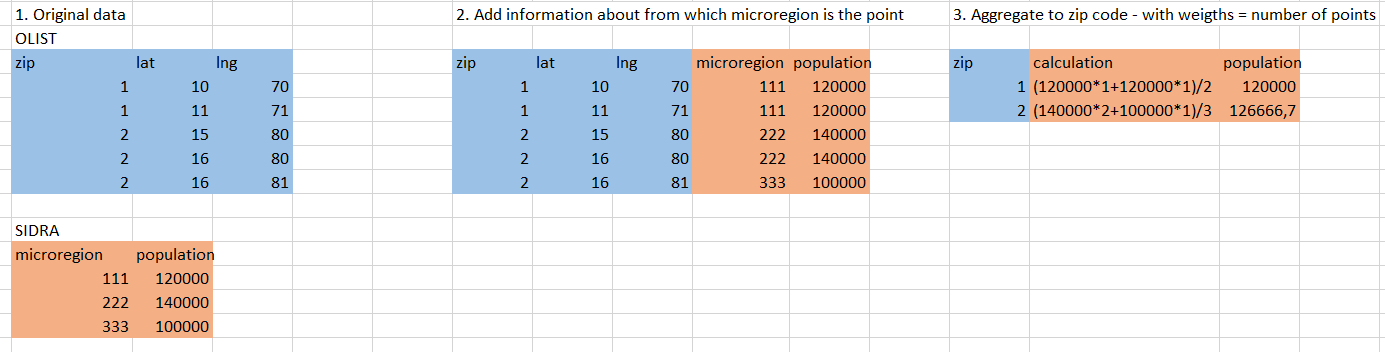
## 7.1 Appendix A - Spatial join of SIDRA and Olist sources

Joining of the data coming from SIDRA and OLIST sources proved to be challenging. There were multiple reasons for that:

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

My approach was as follows:

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



(TODO: Zrobić tabelki w R a nie w excelu, pewnie też lepiej opisać)

# 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.

———. 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.

Berger, P., and M. Kompan. 2019. “User Modeling for Churn Prediction in e-Commerce.” *IEEE Intelligent Systems* 34 (2): 44–52. <https://doi.org/10.1109/MIS.2019.2895788>.

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>.

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.

———. 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.

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. https://doi.org/<https://doi.org/10.1016/j.ejor.2003.12.010>.

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.

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

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

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

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

———. 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.

Koehn, Dennis, Stefan Lessmann, and Markus Schaal. 2020. “Predicting Online Shopping Behaviour from Clickstream Data Using Deep Learning.” *Expert Systems with Applications* 150: 113342.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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>.

Yu, Xiaobing, Shunsheng Guo, Jun Guo, and Xiaorong Huang. 2011. “An Extended Support Vector Machine Forecasting Framework for Customer Churn in e-Commerce.” *Expert Systems with Applications* 38 (3): 1425–30.

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.

1. Because of the fact that in one order there can be multiple product categories, it is not guaranteed that there will be one “1” value per each row as in classical one-hot-encoding method. [↑](#footnote-ref-64)
2. Words “Aspect” and “Topic” are often used interchangably in NLP literature [↑](#footnote-ref-72)
3. I have not run the model containing all variables with demographic features without PCA preprocessing. There are 2 reasons for that - one is that number of variables in this set is very big, what poses performance reasons - model training simply would take a very long time. Assuming that the PCA set of features would give better score can be also supported by the fact that the model with only included PCA demographic variables is performing better than the full set of variables. [↑](#footnote-ref-77)