UNIVERSITY OF TARTU
Institute of Computer Science
Data Science curriculum

Marge Maidla

UTILISING MACHINE LEARNING AND RFM ANALYSIS FOR CUSTOMER RETENTION IN AN ONLINE GROCERY DELIVERY STARTUP

Master's thesis
(15 ECTS)

Supervisors: Maarja Pajusalu, MSc
Elena Sügis, PhD

Tartu 2023

Abstract

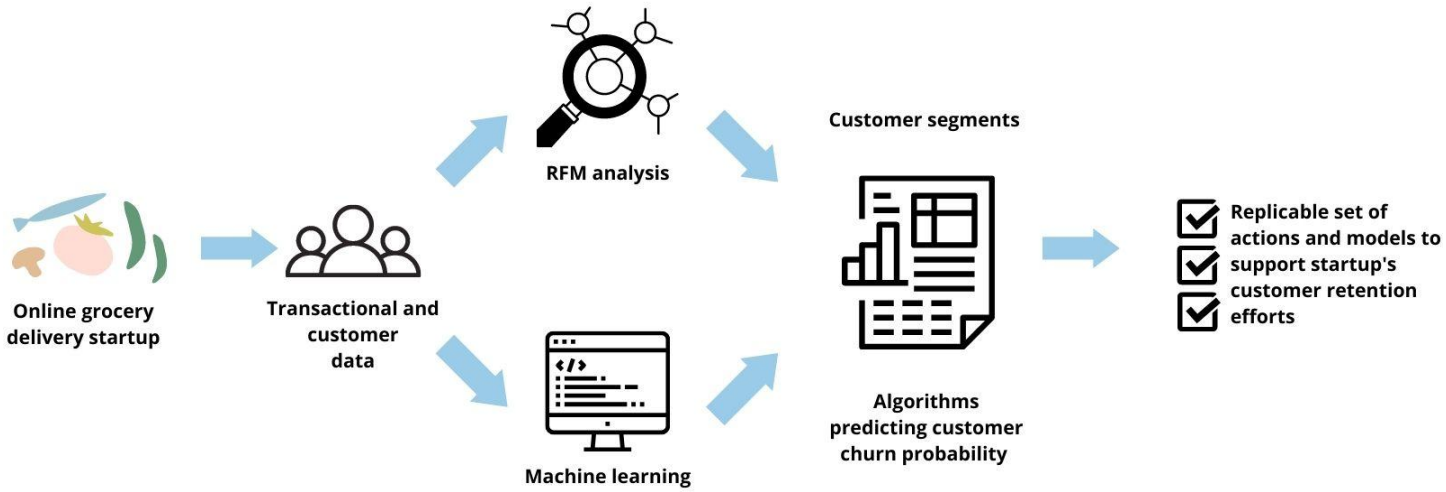
Utilising machine learning and RFM analysis for customer retention in an online grocery delivery startup

Retaining customers is one of the key steps towards a financially successful company. Online delivery businesses need to focus especially hard on retaining customers who they have already managed to convert as consumers have more and more competitors to turn to. Despite available tools and methods, recognising a startup's uniqueness is vital for designing tailored approaches to address customer churn. This thesis is conducted based on data from an early-stage grocery delivery startup and focuses on providing an actionable framework for its management supporting them with retention efforts. Descriptive analysis methods such as Recency, Frequency and Monetary (RFM) analysis and conventional machine learning such as Logistic Regression, Decision Tree, Random Forest and XGBoosting algorithms have been implemented. The RFM analysis showed that the case study company has an almost equal number of customers who are loyal supporters and those who need activation. The best machine learning results were obtained by applying the XGBoost algorithm to predict customer churn. Additionally, the results of this work have implications for the company's everyday operations by providing a practical and easily interpretable framework for the company's management to evaluate customer churn going forward as well.

Keywords:
Churn prediction, RFM, machine learning

CERCS: P176 Artificial Intelligence

UTILISING MACHINE LEARNING AND RFM ANALYSIS FOR CUSTOMER RETENTION IN AN ONLINE GROCERY DELIVERY STARTUP



UNIVERSITY OF TARTU
Institute of Computer Science

#UniTartuCS

Supervisors: Maarja Pajusalu, MSc
Elena Sügis, PhD

Author: Marge Maidla
Data Science (MSc), 2023

Lühikokkuvõte

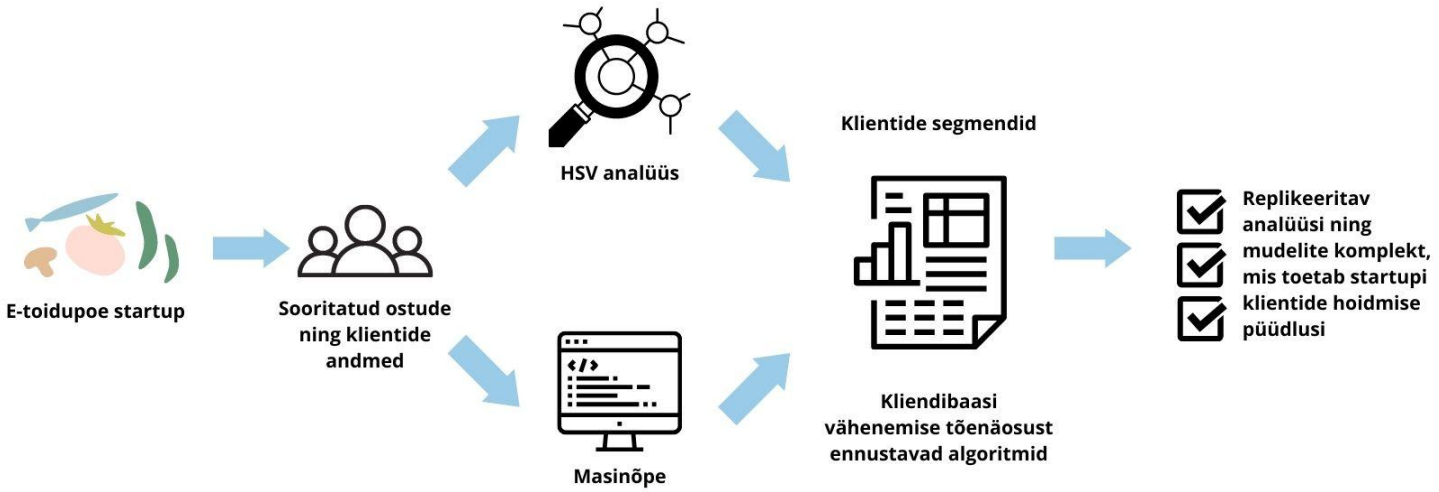
Masinõppe ja HSV analüüsi kasutamine klientide hoidmiseks e-toidupoe startupi näitel

Klientide hoidmine on üks kõige olulisemaid eesmärke ettevõtete jaoks, kes püüdlevad finantsiliselt jätkusuutlike majandusmudelite poole. Kuna e-poodide hulgas on pakkumist palju ning klientidel valikut rohkelt, võiksid ettevõtted eriti hoolikalt klientide hoidmisega tegeleda. Vaatamata juba eksisteerivatele tööriistadele ja meetoditele on startupi omapärade arvesse võtmine klientide hoidmise lahenduste disanimisel oluline. Antud magistritöö baseerub varajase faasi e-toidupoe andmetel. Töö eesmärgiks on luua raamistik, mis abistab ettevõtte juhatusel ja ka teistel sarnastel varajase faasi ettevõtetel klientide hoidmist planeerida. Töös on rakendatud kirjeldavaid analüüsimeetodeid nagu Hiljutisuse, Sageduse ning Väärtuse (*Recency, Frequency and Monetary*) (HSV) (*RFM*) mudelit ja masinõppe mudeleid nagu logistiline regressioon, otsustuspuu, otsustusmets ja XGBoost. HSV analüüsi põhjal on juhtumiuuringus osalenud ettevõttel peaaegu võrdselt kliente, kes on ettevõtte lojaalsed püsikliendid ning neid, kes vajavad aktiveerimist. XGBoost saavutas masinõppe mudelitest parima tulemuse. Antud töö tulemused toetavad ettevõtte igapäevast tegevust, pakkudes ettevõtte juhtkonnale praktilise ja kergesti tõlgendatava raamistiku, et hinnata klientide hoidmist ka edaspidi.

Võtmesõnad:
Klientide lahkumise ennustamine, HSV, masinõpe

CERCS: P176 Tehisintellekt

MASINÕPPE JA HSV ANALÜÜSI KASUTAMINE KLIENTIDE HOIDMISEKS E-TOIDUPOE STARTUPI NÄITEL



1. Introduction **7**
2. Theoretical Background **9**
2.1 CRISP-DM Standard 9
2.2 Churn 10
2.3 Recency, Frequency, Monetary Value Analysis 10
2.4 Machine Learning Algorithms 12
2.4.1 Logistic Regression 13
2.4.2 Decision Tree 13
2.4.3 Random Forest 14
2.4.4 XGBoost 15
2.5 Dataset Balancing Method SMOTE 16
2.6 Explainability 16
2.7 Model Performance Evaluation Metrics 16
2.7.1 Confusion Matrix 17
2.7.2 Precision, Recall, and F1 Score 17
2.7.3 ROC - AUC 18
3. Experiments **19**
3.1 Company Introduction and Understanding Business Question 19
3.2 Data Acquisition and Description 19
3.2.1 Data Description for RFM Analysis 20
3.2.2 Data Description for Machine Learning Modeling 21
3.3 Data Preprocessing 22
3.4 Modeling 22
3.5 Evaluation 23
3.6 Deployment 23
4. Results **24**
4.1 RFM Analysis Results 24
4.2 Machine Learning Modeling Results 25
5. Conclusion **29**
References **30**
Glossary **32**
Appendix **33**
I. Additional Figures 33
II Source Code 34
III License 35

1. Introduction

While the competition is growing, retaining customers is one of the most crucial tasks companies have in order to succeed (Ahmad and Buttle 2002). Moreover, retaining existing customers, especially in saturated markets, is important as it has been reported that acquiring new customers can be five to six times more expensive than retaining already acquired customers. In addition to that, long-term customers would generate higher profits as they tend to purchase more from the company over time (Verbeke et al. 2012). For example, it has been found that a 5% increase in customer retention produces more than a 25% increase in profits (Reichheld and Schefter 2000).

In order to retain customers, it's important to prevent them from churning (Bijmolt et al. 2010). Understanding customer retention is a widely researched area. However, early-stage startups often lack resources like money, time, and workforce to conduct complex analyses. The case study company is an early-stage grocery delivery startup that is precisely in this position. They have accumulated a fair amount of data, but they have not had the resources to analyse their customer base in detail and understand which customers might be churning nor have they had previous experience with churn prediction using machine learning.

Having the latter in mind, the main aim of this work at hand is to provide an interpretable framework for the management of the case study company. The framework shall allow them to make educated data-informed decisions and plan their customer retention campaigns. For example, customers who are more likely to churn could be offered an incentive to make them stay. In addition to that, understanding additional insights about customers' features and characteristics could be used to take preventive measures to avoid churn rather than cure it.

The goal of the work is fulfilled by complementary assessing churn both with statistical interpretability algorithms and with interpretable machine learning algorithms. This work utilizes real transactional data and customer historical data from the case study company. For statistical analysis, RFM (Bult et al. 1995) framework is applied. The RFM analysis offers a way of grouping customers based on their consumption recency, frequency, and monetary value. This helps to understand which customers have been less active and could be more subject to churn and which ones are loyal. While the RFM method provides a good insight into groups of customers, it is, however, limited as it takes into account only three features and hence might neglect other important variables explaining customer behaviour. On the other hand, the relative simplicity of RFM analysis could make it an accessible tool for an early-stage startup. In addition, conventional machine learning such as Logistic Regression, Decision Trees, Random Forest and XGBoosting algorithms have been implemented in order to identify customers with an increased likelihood of churning. The mentioned models have been demonstrated to be effective for churn prediction, including in studying online grocery retailers (Tamaddoni et al. 2016).

The thesis consists of five chapters. In the theoretical chapter, the CRISP-DM method is introduced as well the theoretical background is covered for churn and its prediction methods, RFM analysis, and machine learning models along with evaluation techniques. The third chapter explains the methodologies of the experiments. The fourth chapter will cover the results and the fifth chapter will conclude with the main findings of this study.

2. Theoretical Background

In this chapter, the main methods used in the thesis will be covered as well as explanations are given why those methods were chosen. In addition, recent studies on churn prediction will be addressed along with an overview of RFM analysis.

2.1 CRISP-DM Standard

The goal of this thesis is to provide an easily interpretable framework for the management of the case study startup on how to retain their customers. In order to understand the process of a data science experiment better, especially for people who do not have previous background in data science or machine learning, providing a process methodology supporting the understanding and replication of the framework is a necessary foundation.

A widely used data science-focused methodology which was conceived by Chapman and his co-authors in 1996 and published in 2000 was chosen for both conducting the experiments and providing it as a tool for the management of the case study company. The methodology is called the Cross Industry Standard Process for Data Mining (CRISP-DM) and it consists of six phases (Chpampan et al 2000) which are introduced below in Figure 1.

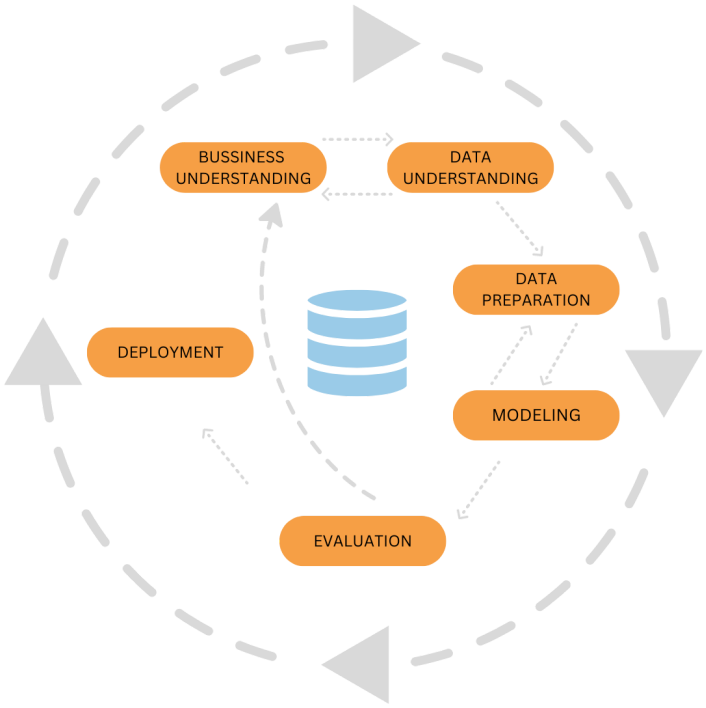


Figure 1: CRISP-DM iterative cyclic process

1. **Business understanding** – this is a preliminary phase that focuses on understanding the business itself based on which the goals of the experiment will be determined and a plan addressing those goals will be produced
2. **Data understanding** – this phase starts with the initial data collection and proceeds with steps in order to understand the data, its quality, deficiencies, and first insights
3. **Data preparation** – this phase is likely repeated multiple times while the data is cleaned, transformed, and a final dataset is produced
4. **Modeling** – in this phase, different models are tested, and as different techniques have specific requirements, circling back to the previous data preparation step might be needed
5. **Evaluation** – in this phase, the models are evaluated and the ones that best meet the business goals are chosen
6. **Deployment** – in this phase, the final report with the results and recommendations is provided, and the maintenance plans are developed

2.2 Churn

Churn can be defined as a loss of customers from the customer base. Churn predictions are often defined as binary classification problems which aim to assign customers to class 1 as a churner and class 0 as a non-churner (Stripling et al. 2017).

In order to group customers as churners or non-churners, a threshold should be defined based on which a customer becomes a churner or stays a non-churner. There is no uniformly accepted way to define churn in a non-contractual setting between the company and the user as different authors have defined the threshold differently. Oliveira et al. (2012) grouped their experiment data over periods of three months and classified a customer as a partial churner if he or she ordered less than 40% in the next period compared to the reference period. In this work, however, we adapt more recent churner and non-churner definitions proposed in the study conducted by Tamaddoni et al. (2016). If the customer placed an order within two months of their last order, they were considered non-churners and if not, they were considered churners. A detailed description is provided in section 3.2.2.

2.3 Recency, Frequency, Monetary Value Analysis

RFM analysis enables companies to segment their customers into groups based on their purchasing behaviour. According to those segments, companies can target specific groups who are more likely to respond to retention activities. One of the advantages of RFM analysis is that customers can be segmented effectively using a small number of features. (Achyar 2019)

Using a small number of features could easily be executable for early-stage startups who might not collect enough data for more complex analysis. According to Koch (2008), The RFM analysis leverages the Pareto 80/20 principle which states that 80% of the returns can be achieved by 20% of the input. Following the Pareto principle while working on customer retention efforts and focusing on the 20% of the most recent customers who order more frequently and pay more, it might be possible to achieve 80% of the desired results.

The RFM features are defined below (Bult et al. 1995):

- **Recency** - how much time (days, weeks, months, etc.) it has been since a customer made their first and their last order (in the experiment defined as the reporting date)
- **Frequency** - how many times a customer has made an order between their first order date and the reporting date
- **Monetary value** - how much money has the customer spent until the reporting date

RFM score can be calculated by performing quantile calculation and dividing the data into five equal groups. Every customer gets a score from one to five where five is the highest score. The higher the value, the higher the score for frequency and monetary value while the lower the recency value, the higher the score. A low recency value means that the customer has been active and has recently made an order. The individual Recency, Frequency, and Monetary scores are then concatenated to derive the overall RFM score for every customer. Customers who buy frequently, have recently made an order and usually spend a lot of money, would get a score of 555 where the Recency equals five, Frequency equals five and Monetary equals five. Inversely, customers who ordered a long time ago, order rarely and spend less, will get a score of 111 where Recency equals one, Frequency equals one and Monetary equals one. (Lin et al. 2010)

The visual representation of RFM analysis can be concluded by dividing customers into 11 segments (Bloomreach 07.05.2023). Table 1 below provides an overview of the segments along with the definition and scoring options.

Table 1. Customer segments according to their scores (Bloomreach 07.05.2023)

Customer Segment	Definition	Scores
Champions	Customers who made an order recently, order often and have high order value	555, 554, 544, 545, 454, 455, 445
Loyal	Customers who order regularly and are responsive to promotions	543, 444, 435, 355, 354, 345, 344, 335
Potential Loyalists	Recent customers with	553, 551, 552, 541, 542, 533,

	considerable spending	532, 531, 452, 451, 442, 441, 431, 453, 433, 432, 423, 353, 352, 351, 342, 341, 333, 323
New Customers	Customers who bought most recently	512, 511, 422, 421 412, 411, 311
Promising	Customers who spent frequently with considerable spending, but the last purchase was some time ago	525, 524, 523, 522, 521, 515, 514, 513, 425,424, 413,414,415, 315, 314, 313
Need Attention	Customers who have above-average recency, frequency and monetary values. They have not bought recently	535, 534, 443, 434, 343, 334, 325, 324
About To Sleep	Customers with below average recency, frequency, and monetary values. They might be lost if they are not reactivated	331, 321, 312, 221, 213, 231, 241, 251
Cannot Lose Them	Customers who have often made high-value orders, but have not made an order for a long time	155, 154, 144, 214,215,115, 114, 113
At Risk	Customers who have spent a lot and purchased often. However, their last order was a long time ago.	255, 254, 245, 244, 253, 252, 243, 242, 235, 234, 225, 224, 153, 152, 145, 143, 142, 135, 134, 133, 125, 124
Hibernating	Customers with smaller and infrequent purchases in the past, but have not purchased anything in a long time	332, 322, 233, 232, 223, 222, 132, 123, 122, 212, 211
Lost	Last purchase was a long time ago	111, 112, 121, 131,141,151

2.4 Machine Learning Algorithms

One of the focus areas of this work is to utilise suitable machine learning algorithms which based on the training data evaluate whether the customer will be a churning customer or not. The case study company and similar companies can utilise the best-performing algorithms to develop their customer retention activities. This subsection describes the machine learning algorithms that were applied in this work.

2.4.1 Logistic Regression

The Logistic Regression algorithm is often used for classification. The logistic function is an S-shaped curve that for a given set of input variables estimates the probability of an outcome, such as the probability of whether a customer would churn or not. Logistic regression is considered easy to implement and can be used as a baseline model to compare other models against it.

The form of the logistic function that outputs a probability y for one input variable x is the following:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1}$$

In order to fit the logistic curve to a dataset, we need to solve for βo (the intercept) and $\beta 1$ (the slope for x) coefficients. For that, the maximum likelihood estimation can be used which maximises the likelihood a given logistic curve would output the observed data. The gradient descent procedure is used in turn to apply maximum likelihood in an optimal way. (Nield 2022)

2.4.2 Decision Tree

The Decision Tree model resembles a tree-like structure and it divides a dataset into smaller subsets. By starting from the root, a feature is evaluated and one of the two branches is selected. This procedure is repeated until a final leaf is reached, which normally represents the classification target. The core algorithm for building Decision Trees is called Iterative Dichotomizer 3 (ID3). ID3 uses Entropy and Information Gain to construct a Decision Tree. (Grus 2019)

The Entropy of a dataset is the measure of disorder in the target feature of the dataset. Information Gain calculates the reduction in the entropy and measures how well a given feature classifies the target classes. The feature with the highest Information Gain is selected as the best one.

Entropy for a dataset S is calculated as follows:

$$H(S) = -p_1 \log_2 p_1 - \ldots - p_n \log_2 p_n \tag{2}$$

where,

n is the total number of classes in the target column

p is the probability of a class.

Information Gain for a feature column A is calculated as follows:

$$IG(S, A) = \text{Entropy}(S) - \sum((|S_v| / |S|) * \text{Entropy}(S_v)) \quad (3)$$

where,

S_v is the set of rows in S for which the feature column A has value v, $|S_v|$ is the number of rows in S_v and likewise $|S|$ is the number of rows in S. (Mitchell 1997)

ID3 construction steps are as follows:

1. Calculate the Information Gain of each feature.
2. Consider that all rows do not belong to the same class, and split the dataset S into subsets using the feature for which the Information Gain is at maximum.
3. Create a Decision Tree node using the feature with the maximum Information Gain.
4. If all rows belong to the same class, mark the current node as a leaf node with the class as its label.
5. Repeat for the remaining features until there are no more features or the Decision Tree has all leaf nodes.

Decision Trees can work efficiently with un-normalised datasets because their internal structure is not influenced by the values assumed by each feature. However, Decision Trees are sensitive to imbalanced classes and can yield poor accuracy when a class is dominant. (Bonaccorso 2018)

2.4.3 Random Forest

The Random Forest model is an elaborated version of a Decision Tree model which consists of a large number of individual Decision Trees that operate as an ensemble. Each individual tree in the Random Forest produces a class prediction and the class with the most votes becomes the model's prediction (Breiman 2001).

The Random Forest learning algorithm uses the bagging ensemble method to improve the stability and accuracy of the model. It works by first creating different copies of the training data. After that, the weak learner is applied to each copy to obtain multiple weak models and then combine them together. (Burkov 2019)

Random Forest classifier consists of a tree-structure classifier (Breiman 2001):

$$\{h(x,\Theta_k), k = 1,...\} \quad (4)$$

where,

$\{\Theta k\}$ are independent identically distributed random vectors

x is the input at which each tree casts a unit vote for the most popular class.

2.4.4 XGBoost

Gradient boosting is another type of ensemble learning model and it is based on the Decision Tree algorithm and boosting method. The idea behind boosting method is that first a model is built on the training dataset after which a second model is built to address the errors present in the first model. The model receives the mistakes of the previous model and tries to improve the model by learning from those mistakes. Finally, XGBoost creates an ensemble of Decision Trees, which sums the predictions of the leaves of different trees to form a model. The XGBoost algorithm searches for a solution to the following sum of functions: (Chen et al. 2016):

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{i=1}^{K} f_k(\mathbf{x}_i),\ f_k \in F \tag{5}$$

where,

$F = \{f(x) = w_{q(x)}\}(q: \mathrm{R}^m \rightarrow T,\ w \in \mathrm{R}^T)$ is the space of regression trees

q represents the structure of each tree that maps an example to

the corresponding leaf index

T is the number of leaves in the tree

Each f_k corresponds to an independent tree structure *q* and leaf weights *w*

w_i represents the score on *i*-th leaf.

In order to learn the best function, the regularised objective is minimised:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{6}$$

where,

l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i

Ω penalizes the complexity of the model which helps to smooth the final learnt weights to avoid over-fitting.

2.5 Dataset Balancing Method SMOTE

There are different methods that can be used to improve data that is imbalanced. For example a) random over-sampling of classes with lower frequency, b) Random under-sampling of classes with higher frequency, and c) Synthetic Minority Over-sampling Technique (SMOTE) on classes with lower frequency. The first two above-mentioned methods have both advantages and disadvantages. Random over-sampling randomly adds more minority observations by replication, which results in no information loss. This might cause potential overfitting due to the replication of the same information. Random under-sampling randomly removes the majority of class observations. This method may help with balancing the dataset, but the removed observations could carry important information and the loss of those might lead to biased results. SMOTE is a modified over-sampling method that creates new synthetic observations based on minority class observations and its nearest neighbours. While SMOTE lowers the effect of overfitting and does not lead to loss of information, it may increase the overlapping of classes. (Lim 2020)

2.6 Explainability

Machine learning model explainability is important for churn prediction. Knowing the most important factors and their impact on model decisions would increase trust in the result and speed up model adoption into practice. It might be challenging to transparently describe how the machine learning model has arrived at a certain outcome (Maan 2023).

Model explainability is a wide topic in machine learning research. In the frames of this work, we are limiting ourselves to the extraction of the most influential features. There are several ways to estimate feature importance. As this study covers multiple machine learning algorithms, a model-agnostic permutation importance method is used. The permutation importance for a single feature measures the decrease in a model score when that feature value is randomly permuted. The disadvantage of the permutation method is that it relies on the independence of the different features. As a result, multicollinearity between different features should be checked before applying the classifiers. (Pitman 2022)

2.7 Model Performance Evaluation Metrics

Model evaluation is an integral part of the machine learning model development process. There are different performance metrics that can be utilised for classification tasks like churn prediction. This subsection describes the model performance methods and metrics used in this study.

2.7.1 Confusion Matrix

The confusion matrix (Figure 2) is a table indicating the performance of the model by showing the number of correct and incorrect predictions categorised by type of response. The columns represent the instances that belong to a predicted class while the rows refer to the instances that actually belong to that class (ground truth). The diagonal cells of the matrix show the number of correct predictions, and the off-diagonal cells show the number of incorrect predictions. A confusion matrix is a visual tool that helps to spot instances where the model might fail. (Saleh 2018).

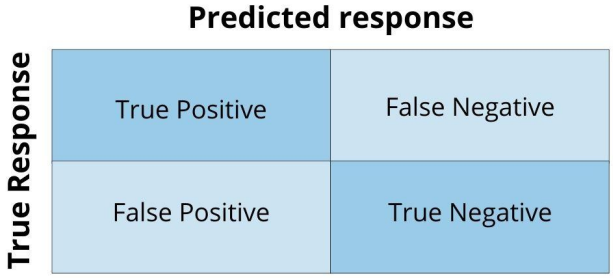


Figure 2. Confusion matrix for binary classification

2.7.2 Precision, Recall, and F1 Score

A confusion matrix is used for the computation of the major model performance metrics such as Precision, Recall and F1 score. Precision, Recall and F1 Score are commonly used in evaluating classification models. The Precision Score measures the model's ability to correctly classify positive labels (the label that represents the occurrence of the event) by comparing it to the total number of instances predicted as positive. This is represented by the ratio between the True Positives and the sum of the True Positives and False Positives, as shown in the following equation: (Bruce et al. 2020)

$$Precision = True\ Positive\ /\ (True\ Positive + False\ Positive) \quad (7)$$

The Recall (Sensitivity) measures the number of correctly predicted positive labels against all positive labels. The metric measures the strength of the model to predict a positive outcome - the proportion of the 1s that it correctly identifies and is represented as follows as the ratio between True Positives and the sum of False Negatives and True Positives: (Bruce et al. 2020)

$$Recall = True\ Positive\ /\ (False\ Negative + True\ Positive) \quad (8)$$

The harmonic mean of Precision and Recall is used to calculate an F1 Score which is a measure of performance of the model's classification ability. An F1 Score is calculated as follows: (Molin 2019)

$$F1\ Score = 2 * (Precision * Recall)/\ (Precision + Recall) \quad (9)$$

2.7.3 ROC - AUC

The Receiver Operating Characteristics (ROC) curve (Figure 4) plots recall (sensitivity or True-positive rate) on the y-axis against specificity (False-positive rate) on the x-axis. The ROC curve shows the trade-off between recall and specificity. The dotted diagonal line corresponds to a classifier whose results are presented as a random chance. An extremely effective classifier will have a ROC that touches the upper-left corner - it will correctly identify 1 class without misclassifying 0 classes as 1s. The ROC curve can be used to produce the area underneath the curve (AUC) metric. AUC is the total area under the ROC curve. The larger the value of AUC, the more effective the classifier is. An AUC of 1 indicates a perfect classifier: it predicts all the 1s correctly and it does not misclassify any 0s as 1s. (Bruce et al. 2020)

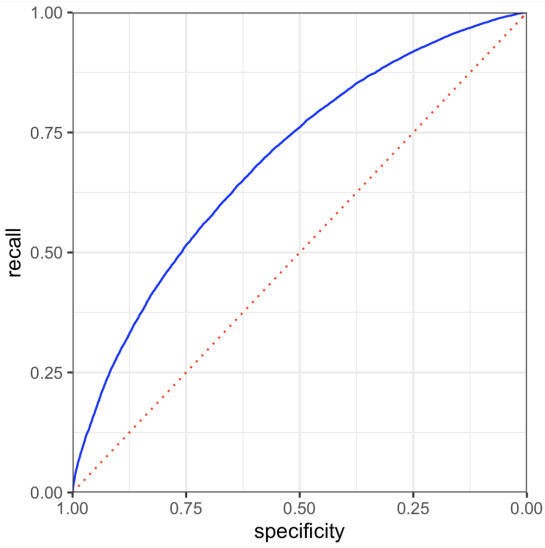


Figure 4. ROC curve (Bruce et al. 2020)

3. Experiments

The experiments were conducted following the CRISP-DM methodology. This chapter covers all 6 phases of the CRISP-DM methodology with the limitation regarding the maintenance plan mentioned in phase 6 since it is out of the scope of this thesis. In addition, this work covers two separate experiments: 1) RFM analysis and 2) machine learning modeling.

The practical experiments are conducted using Python programming language. The code is shared via the Google Colaboratory Notebook as this is considered to be the most convenient option for the case study company.

Main libraries used:

- Pandas 1.5.3
- Scikit-learn 1.2.2
- Numpy 1.22.4
- Imblearn 0.10.1
- Matplotlib 3.7.1
- Seaborn 0.12.2

3.1 Company Introduction and Understanding Business Question

This thesis uses data from an Estonian grocery delivery startup which aims to connect local producers to consumers while cutting out all the middlemen. They offer a weekly online grocery shopping solution.

In the early stages, startups might lack enough data to conduct complex analyses. In addition, they might also lack resources - both capital and workforce to focus on gathering insights from their data. The case study company had already gathered a fair amount of data, but the team did not have a chance to conduct any experiments yet. The management of the company is interested to know how to retain their customers better as they had not done any analysis or modeling to understand that before. As a result, both the RFM analysis and churn modeling were chosen so that the case study company can use the results and plan for retention activities.

3.2 Data Acquisition and Description

The data was gathered from the company's database tool Metabase using SQL queries. Prior to assembling the data, the management of the company explained the data structure and relevant tables. Based on the conversations features were defined and the dataset was derived.

3.2.1 Data Description for RFM Analysis

The grain of the dataset gathered for the RFM analysis is on the transactional level and consists of the following data:

1. ***user_id*** (unique customer identification number) INT
2. ***order_date*** (the date of the order) INT
3. ***total_sum*** (the price customer paid for the order) INT

The derived dataset consists of 40113 observations. The needed preparation in order to analyse the data in the right format was already done while assembling the data using SQL. The data was assembled into a per-user data frame with corresponding recency, frequency and monetary values (Figure 3).

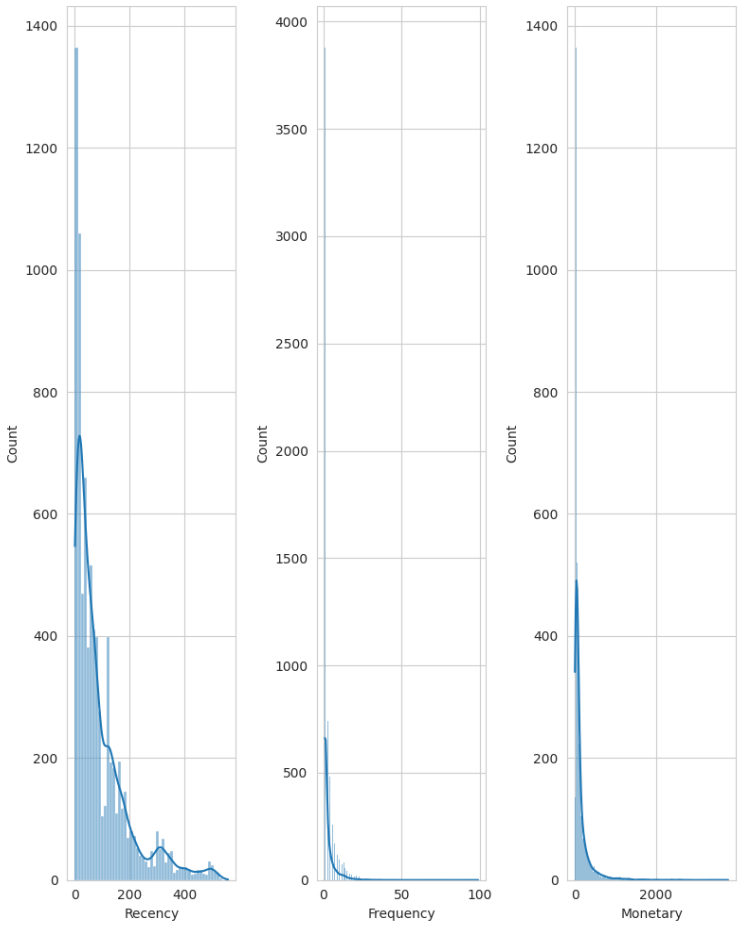


Figure 3: Distributions for Recency, Frequency and Monetary Values

Most of the customers have made orders within the last 150 days while the majority of them have made only one order. Most of the customers have spent around 100 euros while there are some customers who have spent more than 2000 euros (Figure 3).

3.2.2 Data Description for Machine Learning Modeling

The features for machine learning modeling were extracted based on management's insights. For example, payments below ten euros have been disregarded as well as orders from their own team members. The delivery was considered a late delivery if the courier was more than twenty minutes late as the startup promised a delivery +/- fifteen minutes from the estimated arrival time. In addition to that, the startup was issuing a refund for specific products whose weight and price were estimated wrong. Those refunds were disregarded too.

In order to label customers as churned or not-churned, we have applied the approach proposed by Tamaddoni et al. (2016). Customers were sorted ascending by the average time between orders, and the 90th percentile was identified. 90% of customers in the dataset placed an order within 2 months - this was used as a decision point where the customer was either considered as a churner or not. So if the customer placed an order within two months of their last order, they were considered non-churners and if they did not place an order, they were considered churners.

The grain for machine learning models is on the customer level and the data used for modeling includes nine following features:

1. ***service_area_id*** (number representing a certain delivery area) INT
2. ***number_of orders*** (number of orders per customer) INT
3. ***total_order_value*** (customer total order value) INT
4. ***avg_time_between_orders_days*** (the average time between customer orders in days) INT
5. ***avg_order_value*** (the average customer order value) INT
6. ***has_had_refund*** (shows if the customer has ever had a refund) INT
7. ***has_had_late_delivery*** (if the order had been more than 10 minutes late) INT
8. ***joining_month*** (the month when the customer signed up) INT
9. ***last_order_date_month*** (the date of the last order) INT

The customer-level data for the machine-learning experiment consisted of 4158 observations. The given dataset is imbalanced as 714 (17%) observations were determined as non-churners and 3444 (83%) observations as churners. All feature distributions can be found in Appendix under the Additional Figures section.

3.3 Data Preprocessing

The RFM analysis does not require extensive data preparation. The needed preparation in order to analyse the data in the right format was already done while assembling the data using SQL. However, for machine learning modeling, data needed to be prepared.

The following preprocessing steps were conducted in order to prepare data for machine learning algorithms:

1. Datetime data were converted into numerical data
2. Multicollinearity was detected and one highly correlated feature was removed
3. The imbalanced dataset was balanced with the SMOTE method

Eight features remained after the data preprocessing steps that were used in four different machine learning algorithms.

Additionally, the preprocessed dataset was split into training and testing sets correspondingly 80% and 20%. Machine learning models were trained on the training set using a cross-validation method. Evaluation of model performance was done on a test set.

3.4 Modeling

Logistic Regression was selected as the baseline model for the machine learning experiment as multiple different authors studying churn have demonstrated previously (Tamaddoni et al. 2016). The baseline model is compared against other, more complex models in order to understand whether the latter exhibits considerable advantages in terms of explainability and performance. Explainability is essential considering that the case study company does not have any prior experience with machine learning.

According to different studies, the following models were applied after the baseline model:

1. Decision Tree
2. XGBoost
3. Random Forest

During the modeling step, an ensemble of all models mentioned before was created as well. However, the ensemble did not perform better compared to the individual models and hence the need for studying it further was not considered justified.

In addition, a Grid Search (Pedregosa et al. 2011) method was applied in order to find the best-performing hyperparameters during cross-validation.

3.5 Evaluation

All developed machine learning models' performances were evaluated on a test set. In an imbalanced dataset, the majority class may dominate the performance metric, leading to inaccurate results. For instance, accuracy can be misleading since a model that always predicts the majority class will have high accuracy but will perform poorly on the minority class. (Branco et al. 2015)

ROC AUC, which is not dependent on the proportions of the predicted classes, and F1 were selected to evaluate model performance as they are commonly used as main performance metrics on imbalanced datasets (Bruce et al. 2020). Additionally, model precision and recall are reported. To understand the business impact of the model performance, it is also essential to consider the number of predicted false positives and false negatives. False positives, where the model predicts that a customer is a churner but he or she is not, can result in unnecessary retention efforts and costs. False negatives, where the model fails to predict that the customer is a potential churner, can lead to lost revenue.

The selected metrics, ROC AUC, F1 score, precision, and recall, provide a comprehensive view of the model performance and can guide the development of a retention strategy that minimizes the impact of false positives and false negatives.

For instance, a model with high precision and low recall would suit the customer retention strategy of a company that wants to minimize the costs of false positives, while a model with high recall and low precision would fit better for a company that wants to minimize the costs of false negatives. By considering the business customer retention perspective in addition to the model performance metrics, the company can make informed decisions about the churn prediction model.

3.6 Deployment

We provided the company with an interactive report in the form of an executable Google Colaboratory Notebook. The report includes both the RFM analysis and the selected machine learning models. These two approaches are complementary and provide a comprehensive analysis of customer purchasing behaviour and their likelihood to churn.

The Google Colaboratory Notebook provides an interactive and easily accessible format for the report. It allows the company to run the code upon necessity and explore the results, making it a valuable tool for decision-making. By providing an executable file, we ensure that the company can access the report and use it to inform their business strategies.

It is important to ensure that the model performance on new customer data is consistent with the results obtained during modeling. This can be achieved by monitoring the model performance regularly and making necessary adjustments if required.

4. Results

This chapter covers the results of the practical experiments. The RFM analysis and machine learning modeling are covered in separate subsections below.

4.1 RFM Analysis Results

The RFM analysis results (Figure 5) show that the biggest (20%) group of customers belong to the segment of "Champions". Those customers are the ones who are valuable customers with high recency, who buy rather often and have high order values. The second (16%) biggest segment belongs to the "Hibernating" customers. Those customers are the ones who ordered a long time ago, did not spend much nor did they order often. The third (12%) and fourth (12%) biggest segments stand for "Lost" and "Potential Loyalists". The former segment stands for customers who have ordered a long time ago and likely it is not worth trying to activate them anymore. The latter segment stands for the customers with relatively high recency and considerable order values. In addition, there is 3% of the customers in the "Cannot Lose Them" segment. Customers in this segment have often made high-value orders in the past but do not do it any more.

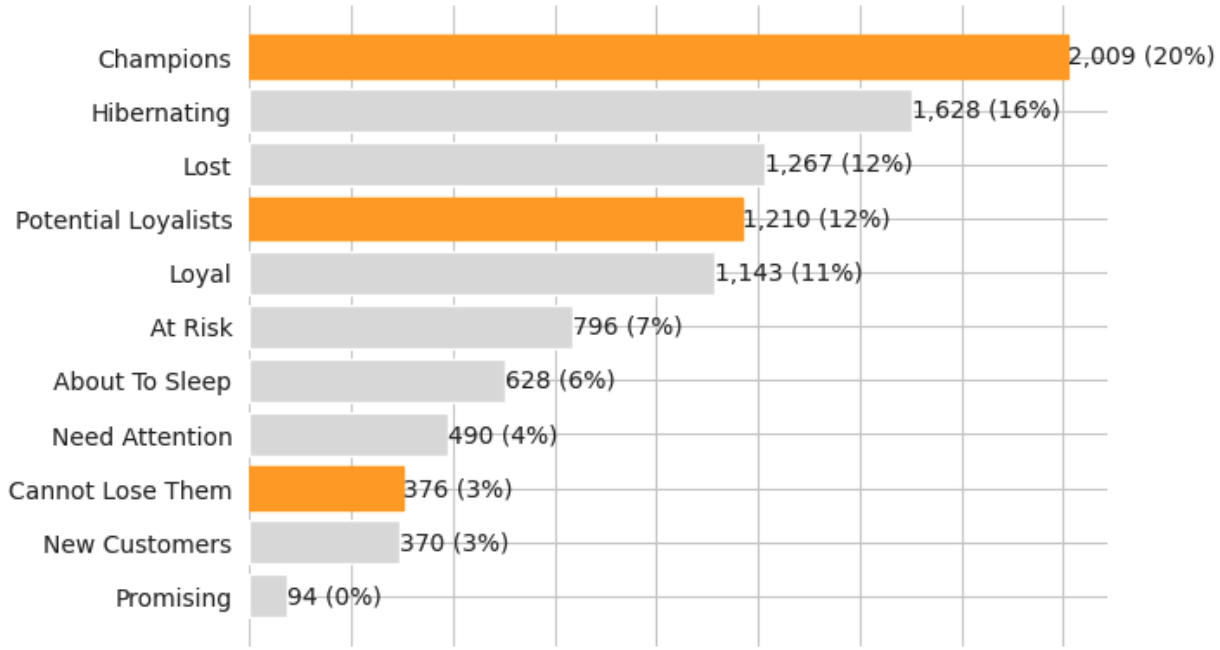


Figure 5: Customer counts and shares in different RFM segments

The average customer in the most desired segment "Champions" spends around 530 euros, makes 12 orders and does it recently. The average customer in the "Hibernating" segment spends around 41 euros and makes one order. Even though they are the second biggest group, the

"Champions" and the "Potential Loyalists" segments together stand for 3219 customers out of 10011 observed customers in this experiment and should be retained as well as possible.

Table 2. Average values for "Champions", "Hibernating", "Lost", "Potential Loyalists" and "Cannot Lose Them" segments

Segment	Recency mean	Frequency mean	Monetary mean
Champions	31	12	530
Hibernating	161	1	41
Lost	334	1	25
Potential Loyalists	64	1	51
Cannot Lose Them	305	2	119

In addition, even if the "Cannot Lose Them" segment makes up only 3% of the customer base, targeting them with a suitable marketing campaign might be relevant as the average customer in this segment spends 119 euros (Table 2).

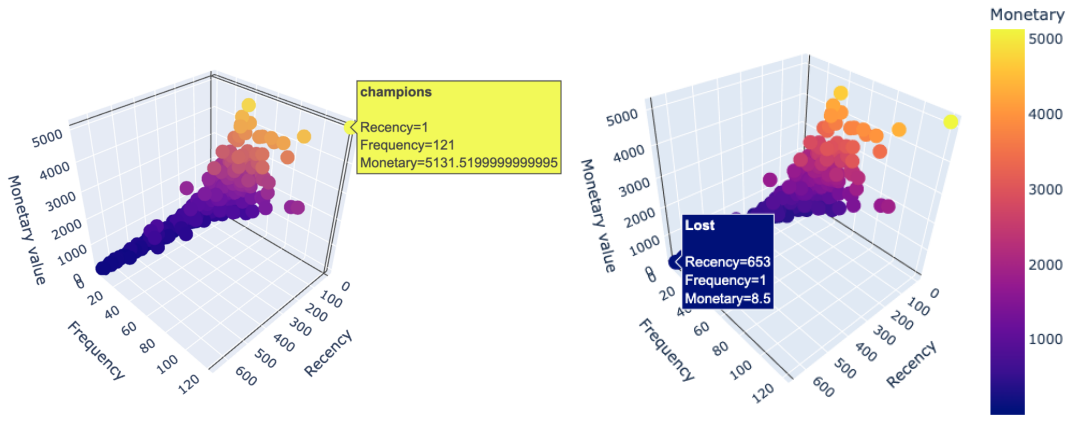


Figure 6: Highest-spending customer in the "Champions" segment and one of the lowest-spending customers in the "Lost" segment

The highest spender in the "Champion" segment has made orders for more than 5000 euros during the observation period while the lowest-spending customer in the Lost segment has made orders worth of only nine euros (Figure 6).

4.2 Machine Learning Modeling Results

The classification results in the form of confusion matrices for all four models used in the experiment are shown in Figure 7. The confusion matrix for Logistic Regression indicates that

the model predicted correctly 165+644 instances and 20+3 incorrectly (20 churners were wrongly predicted as non-churners and 3 non-churner was wrongly predicted as churner). The confusion matrix for the Decision Tree indicates that the model predicted correctly 150+654 instances and incorrectly 10+18 (10 churners were wrongly predicted as non-churners and 18 non-churners were wrongly predicted as churners). The confusion matrix for XGBoost indicates that the model predicted correctly 161+662 classes and 2+7 classes incorrectly (2 churners were wrongly predicted as non-churners and 7 non-churners were wrongly predicted as churners). The confusion matrix for Random Forest predicts 154+655 instances correctly and 9+14 incorrectly (9 churners were wrongly predicted as churners and 14 non-churners were wrongly predicted as churners).

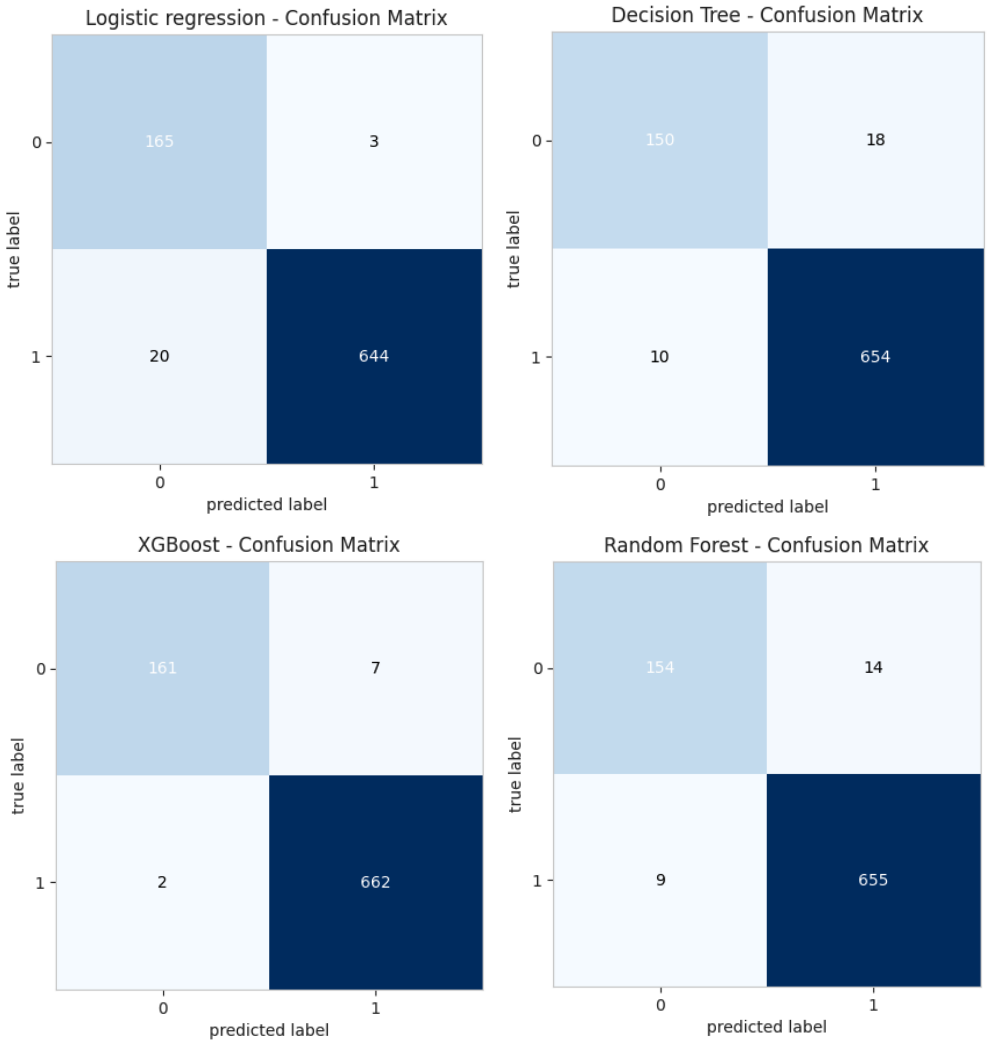


Figure 7: Confusion matrices of Logistic Regression, Decision Tree, XGBoost and Random Forest models. Class churners is indicated as 1 and class non-churners as 0.

Algorithms were evaluated based on their ROC-AUC Score, Precision, Recall, and F1 Score metrics (Table 3). As displayed in Table 3, the XGBoost algorithm outperformed the rest of the

tested algorithms with a ROC-AUC score of 99.95%. The baseline algorithm Logistic Regression had the second-best result with a ROC-AUC score of 99.68%. The Random Forest algorithm with a result of 99.55% outperformed The Decision Tree algorithm which received the lowest score of 96.74%. The XGBoost algorithm has the best Precision (99%), Recall (96%) and F1 Score (97%).

Table 3. Machine Learning Algorithms and their ROC-AUC scores

Algorithm	ROC-AUC Score	Precision	Recall	F1 Score
Logistic Regression	0.9968	0.89	0.98	0.93
Decision Tree	0.9674	0.94	0.89	0.91
XGBoost	0.9995	0.99	0.96	0.97
Random Forest	0.9955	0.94	0.92	0.93

Figure 8 displays the ROC curves for the four algorithms used in this study. XGBoost classifier touches almost the upper-left corner of the graph indicating that the algorithm will correctly identify churners without misclassifying non-churners as churners.

Based on the performance metrics, it can be concluded that XGBoost outperforms the other three algorithms in terms of ROC-AUC Score, Precision, Recall, and F1 Score. Logistic Regression shows competitive results, especially in terms of recall, while Random Forest demonstrates a good balance between precision and recall. Decision Tree, although relatively less effective compared to the other algorithms, still offers acceptable performance. These results allow us to suggest that for the given dataset all of those could be considered for practical implementation.

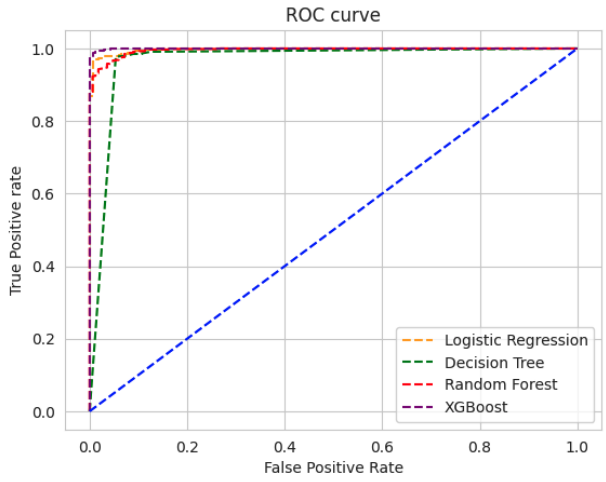


Figure 8: ROC curve of each classification algorithm. The dashed blue line in the centre of a figure demonstrates the ROC curve of a random classifier.

The most important features that contribute to predicting the churn for the XGBoost algorithm are shown in Figure 9.

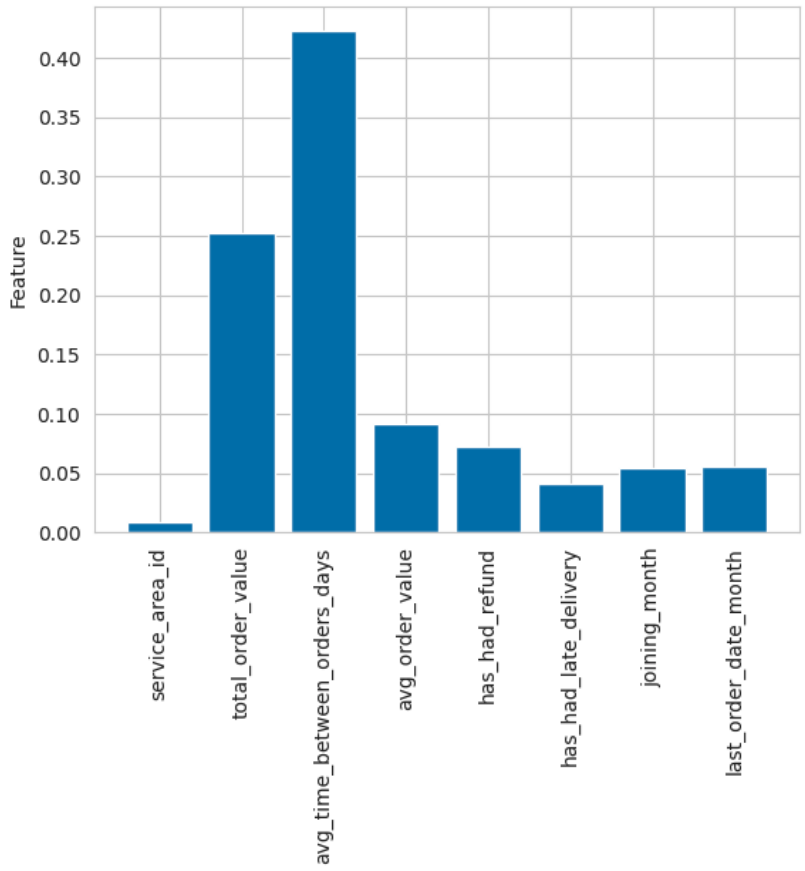


Figure 9: The most important features that contribute to predicting the churn with the XGBoost algorithm

Important features according to the XGBoost algorithm were:

- *average_time_between_orders_days* which indicated the average time between customer orders counted in days
- *total_order_value* which indicates the total value customer has spent in the case-study company
- *avg_order_value* which indicates the average value of customer order.

5. Conclusion

The RFM analysis of customer purchasing behaviour and churn predictions can be complementary when it comes to offering an easily interpretable solution for early-stage startups. The RFM on the one hand is a simple, yet illustrative and insightful tool which helps to understand customer segments and the potential to churn. Machine learning modeling on the other hand offers a different way to look at the churn and take into account more features that might play an important role in understanding customer churn.

According to the RFM analysis, the case study company data shows that they have an almost equal number of customers who need activating and those who are loyal supporters. The management of the case study company can take that and other information regarding their customer segments into account when developing their retention activities. The best machine learning results were obtained by applying the XGBoost algorithm which performed just slightly better than the baseline Logistic Regression and Random Forest. However, complexities deriving from a highly imbalanced dataset and algorithm interpretability are problems to overcome while conducting such an experiment. Due to easier interpretability, Logistic Regression can be considered as a model for deployment.

To deploy the RFM analysis and machine learning models and integrate them into case study startup operations, we recommend using an executable Google Colaboratory Notebook. This format allows for easy access and interactive exploration of the analysis results, making it a valuable tool for decision-making.

In the future, in addition to RFM analysis, customer segmentation techniques like Principal component analysis (PCA) and KMeans clustering could be used. Furthermore, uplift modeling which is a predictive modeling technique helping to identify which customers are most likely to respond positively to a particular marketing campaign could be used. This requires additional data on previous marketing campaigns the case study company has conducted in the past. Relating to the machine learning experiment, additional algorithms like the Support Vector Machine could be tested. In addition to that, additional features could be added to the dataset and test how the results differ.

References

Achyar A. Customer Segmentation on Online Retail using RFM Analysis: Big Data Case of Bukku.id. 2019. doi:10.4108/eai. 1-4-2019.2287279.

Ahmad K and Buttle M. Customer retention management: A reflection of theory and practice. Marketing Intelligence & Planning 20 (3):149-161, 2002, doi:10.1108/02634500210428003.

Bijmolt T. H. A., Leeflang P. S. H., Block F., Eisenbeiss M., Hardie B. G. S., Lemmens A. and Saffert P. Analytics for Customer Engagement. Journal of Service Research, 13 (3), 2010, 341-356.

Bloomreach Software. RFM Segmentation.
URL: https://documentation.bloomreach.com/engagement/docs/rfm-segmentation (07.05.2023)

Bonaccorso G. Machine Learning Algorithms - Second Edition. Packt Publishing. 2018.

Branco P., Torgo L. and Ribeiro R. P. A Survey of Predictive Modelling under Imbalanced Distributions. 2015. https://doi.org/10.48550/arXiv.1505.01658

Breiman L. Random Forests. Machine Learning, 45, 5–32. 2001.

Bruce P., Bruce A, Gedeck P. Practical Statistics for Data Scientists, 2nd Edition. O'Reilly Media, Inc. 2020.

Bult J. R and Wansbeek T. Optimal Selection for Direct Mail. 1995.

Burkov A. The Hundred-page Machine Learning Book. 2019, IS- BN: 9781999579500.

Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, https://doi.org/10.1145/2939672.2939785

Crus J. Data Science from Scratch, 2nd Edition. O'Reilly Media, Inc. 2019.

Koch R. The 80/20 Principle: The Secret to Achieving More with Less. New York: Doubleday. 2008

Lim T.. RFM and Classification Predictive Modelling to Improve Response Prediction Rate. 2020, doi: 10.1109/ZINC50678.2020.9161800

Lin, S.-Y., Wu H.-H. . A review of the application of RFM model. African Journal of Business Management, 2010, 4(19):4199-4206.

Maan J. and Maan H. Customer Churn Prediction Model using Explainable Machine learning. International Journal of Computer Science Trends and Technology (IJCST), 2023, Volume 11 Issue 1, Jan-Feb 2023

Mitchell M. T. Machine Learning. 1997

Molin S. Hands-On Data Analysis with Pandas. Packt Publishing. 2019

Nield T. Essential Math for Data Science. O'Reilly Media, Inc. 2022

Olivera V. L. M. Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques. 2012, Corpus ID: 167771876

Pedregosa et al., Scikit-learn: Machine Learning in Python. JMLR 12, 2011, pp. 2825-2830

Pitman D., Munn M. Explainable AI for Practitioners. O'Reilly Media, Inc. 2022

Reichheld F., and Schefter P. E-Loyalty: Your Secret Weapon on the Web. Harvard Business Review 78(4). 2000

Saleh H. Machine Learning Fundamentals. Packt Publishing. 2018

Simmons et al.The next S-curve of growth: Online grocery to 2030. McKinsey & Company. 2022.
URL:https://www.mckinsey.com/industries/retail/our-insights/the-next-s-curve-of-growth-online-grocery-to-2030 (07.05.2023)

Stripling et al. Profit Driver Decision Trees for Churn Prediction. 2017.
URL: https://doi.org/10.48550/arXiv.1712.08101

Tamaddoni A., Stakhovych S. and Ewing M. Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. Journal of Service Research. 2016. doi: 10.1177/1094670515616376

Verbeke et al. A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. European Journal of Operational Research, 2012, 218(1):211-229. doi:10.1016/j.ejor.2011.09.031

Glossary

Google Colaboratory Notebook A Jupyter notebook that runs in the cloud and is highly integrated with Google Drive, making them easy to set up, access, and share. 19, 23, 29, 34

Imblearn On open source, MIT-licensed library relying on scikit-learn (sklearn) which provides tools when dealing with classification with imbalanced classes. 19

Matplotlib Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. 19

Non-contractual A relationship between a customer and a business where there is no binding contract between the parties. 10

Numpy NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. 19

Pandas Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. 19

Python Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation via the off-side rule. 19

Scikit-learn (Sklearn) is a library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistency interface in Python. 19

Seaborn Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. 19

Appendix

I. Additional Figures

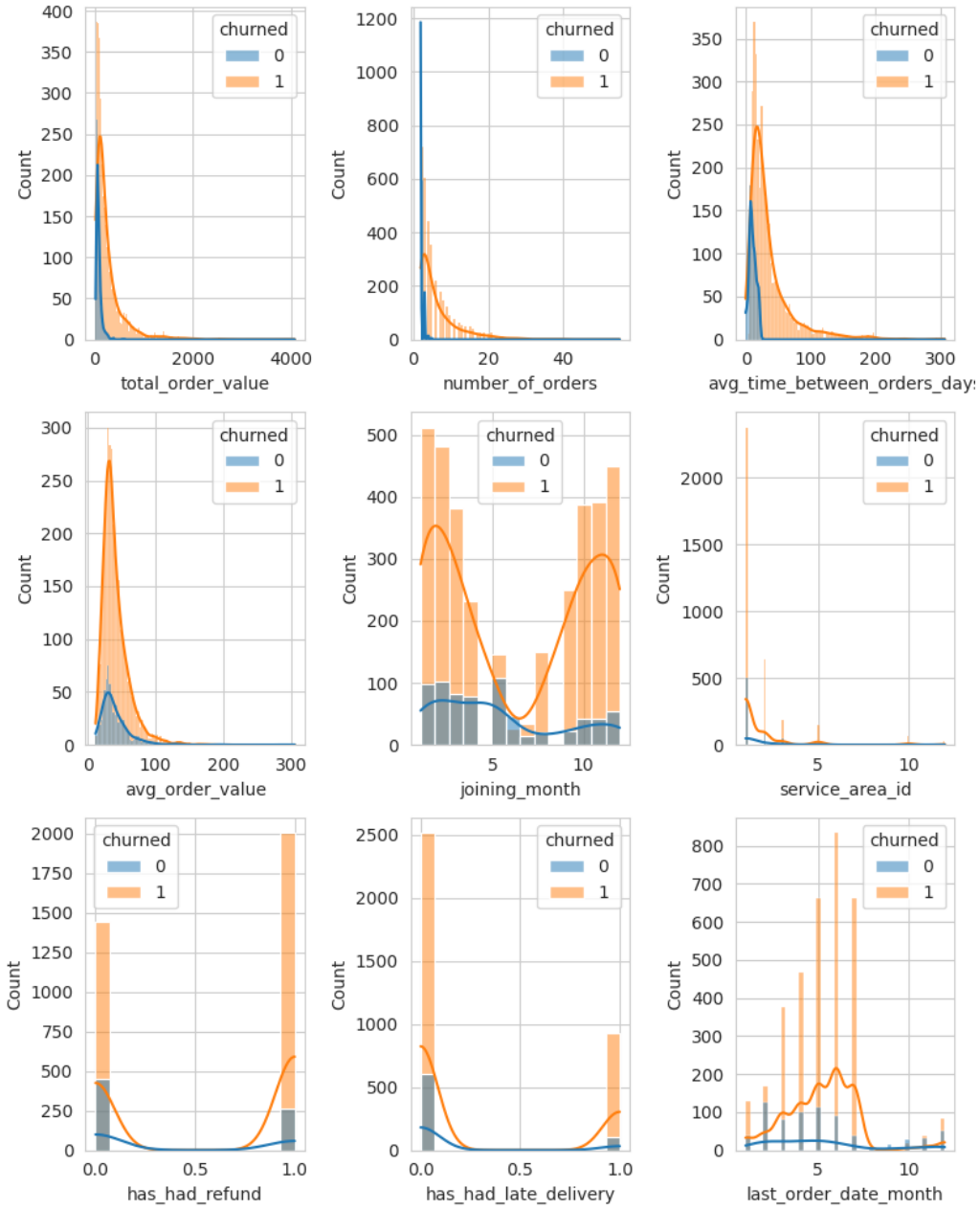


Figure 10. All feature distributions from the Machine Learning experiment

II Source Code

The analysis performed as part of the thesis has been carried out using the Python programming language. The source code is located in the Google Colaboratory Notebook, which can be accessed from the following link:

https://colab.research.google.com/drive/15ZWO9MA8mDeVVxGJjozO2Hporvc_wagc?usp=sharing

III License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Marge Maidla,
(author's name)

1. grant the University of Tartu a free permit (non-exclusive licence) to:

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Utilising machine learning and RFM analysis for customer retention in an online grocery delivery startup,
(title of thesis)

supervised by Maarja Pajusalu and Elena Sügis
(supervisors' name)

2. I grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from 09/05/2023 until the expiry of the term of copyright.

3. I am aware that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Marge Maidla
09/05/2023