CUSTOMER TRANSACTION PREDICTIVE MODELING VIA MACHINE LEARNING ALGORITHMS

SEYİT ERTUĞRUL

MEF UNIVERSITY
JANUARY 2023

MEF UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MASTER'S IN INFORMATION TECHNOLOGIES

M.Sc. THESIS

CUSTOMER TRANSACTION PREDICTIVE MODELING VIA MACHINE LEARNING ALGORITHMS

Seyit Ertuğrul

Orcid No: 0000-0003-0828-7336

Asst. Prof. Dr. Tuna ÇAKAR

JANUARY 2023

ACADEMIC HONESTY PLEDGE

I declare that all the information in this study is collected and presented in accordance with academic rules and ethical principles, and that all information and documents that are not original in the study are referenced in accordance with the citation standards, within the framework required by the rules and principles.

Name and Surname: Seyit ERTUĞRUL
Signature:

ABSTRACT

CUSTOMER TRANSACTION PREDICTIVE MODELING VIA MACHINE LEARNING ALGORITHMS

Seyit ERTUĞRUL

M.Sc/MA in Information Technologies

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

January 2023, 41 Pages

The main purpose of this study is to determine the behavior and characteristics of the customers of a company that is active in the factoring sector, and accordingly, to capture measurable parameters with exploratory data analysis based on the historical data of the customers, and then to perform predictive models for the target. A hit rate of around 80% was achieved in SVM and Extra Trees models, which are classification model algorithms. In this way, it is aimed to directly contribute to the transaction volume on a business basis by acting in a more effective, efficient and correct approach after approving the check that shows high potential, that is, the customers who are likely to accept it after the offer is made as a business.

Keywords: machine learning, customer order frequency, feature extraction, feature selection.

Numeric Code of the Field: 92404

ÖZET

YAPAY ÖĞRENME YÖNTEMLERİ İLE MÜŞTERİ İŞLEM TAHMİNİ MODELİ

Seyit Ertuğrul

Bilişim Teknolojileri Yüksek Lisans Programı

Tez Danışmanı: Asst. Prof. Dr. Tuna ÇAKAR

Ocak 2023, 41 sayfa

Bu çalışmanın temel amacı faktoring sektöründe aktif olarak faaliyet gösteren bir şirketin müşterilerinin davranışlarını ve özelliklerini saptayabilmek, buna bağlı olarak da müşterilerin geçmiş verilerinden yola çıkarak, keşifçi veri analiziyle ölçülebilir parametreler yakalayabilmek ve akabinde hedefe yönelik tahminsel modellemeler gerçekleştirebilmektir. Sınıflandırma modeli algoritmalarından SVM ve Extra Trees modellerinde %80 seviyesi üzerinde isabet oranı yakalanmıştır. Bu sayede yüksek potansiyel gösteren, yani sorgulattığı çeki, işletme olarak onayladıktan ve teklif yapıldıktan sonra kabul etme ihtimali yüksek olan müşterileri tahmin edilmesi daha etkin, verimli ve doğru yaklaşımlar içerisinde hareket edip aksiyon alarak, işletme bazında işlem hacmine doğrudan katkısı sağlanması amaçlanmıştır.

Anahtar Kelimeler: Makine öğrenmesi, müşteri işlem sıklığı, müşteri işlem ihtimali, öznitelik çıkarma, öznitelik seçme.

Bilim Dalı Sayısal Kodu: 92404

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my thesis professor, Asst. Prof. Tuna Çakar, for his unwavering support and guidance throughout this thesis.

TABLE OF CONTENTS

ABBREVIATIONS

AUC	: Area Under the ROC Curve
FP	: False Positive
FN	: False Negative
FPR	: False Positive Rate
GLM	: Generalized Linear Model
GMM	: Gaussian Mixture Models
IQR	: Interquartile Range
KNN	: K-Nearest Neighbors
LASSO	: Least Absolute Shrinkage and Selection Operator
LGBM	: Light Gradient-Boosting Machine
LOGOCV	: Leave-One-Group-Out Cross Validation
MAR	: Missing at Random
MCAR	: Missing Completely at Random
MDS	: Multidimensional Scaling
MICE	: Multiple Imputation by Chained Equations
MNAR	: Missing Not at Random
PCA	: Principal Component Analysis
ROC	: Receiver Operating Characteristic
SVM	: Support Vector Machines
TN	: True Negative
TP	: True Positive
TPR	: True Positive Rate
T-SNE	: T-Distributed Stochastic Neighbor Embedding
VIF	: Variance Inflation Factor
XGB	: Extreme Gradient-Boosting Algorithm

INTRODUCTION

1. Purpose of Thesis

In recent years, the demand for data science has increased, especially in finance and online advertising. The main reason for this is that the benefits of data-based decision making have been conclusively demonstrated. They did a study on how it affects firm performance. They rated firms according to how strongly they used data when making decisions. Statistically, the more data-driven a firm is, the more productive it is. The difference is associated with a 4-6% increase in productivity. As can be seen, the impact of data-driven decision making is too great to ignore. The ultimate goal of data science is to improve decision making with available data. In doing so, it uses scientific methods, processes, algorithms and systems. Data-driven decision making (DDD) refers to basing decisions entirely on analysis of data, not on intuition. There are a few important considerations when making these applications. First of all, it is important to correctly select the data to be used to solve the problem. Then you need to determine the right model for which you will use this data.

Many firms work with data science teams for competitive advantage. Data science can have multiple benefits for companies. To exemplify how a store and telecommunications company use data science; As Hurricane Frances approached the US state of Florida, executives at Wal-Mart Stores began to work to turn the situation in their favor. The firm's chief of information pulled out data on what sales had changed when Hurricane Charley hit the state a few weeks ago. We observed an increase in sales of some products, noting whether this increase was only in hurricane-hit states or was there an overall increase. According to the analysis, it was certain that in the event of a hurricane, the sales of some products increased up to 7 times, and these products were stocked. In this way, the company increased its sales rate by making data-based decisions.

MegaTelCo, one of the major telecommunications companies in the USA, is losing 20% of its customers whose contracts have expired, according to research. Since telecommunications is now a saturated market, it is very difficult to find new customers. A company has to spend on incentives to attract customers, and when the customer leaves, it loses revenue. In addition, acquiring new customers is much costly than retaining an old customer. That is why companies want to retain their old customers and allocate a marketing budget for them. Here, the task of data scientists is to decide which customer to retain. The data science team, which configures the problems and builds a model, decides on this model. These two examples illustrate two different kinds of decisions of data science. In the first example, discoveries were made within the data and decisions were made by examining recurring events. In the second example, evaluations were made at different scales from tens of millions of data and individuals who provided the desired characteristics were selected. The more we can improve our practice of predicting how profitable it will be to focus on a particular customer, the more we can potentially reap huge benefits by applying this ability to millions of customers in the population.

Each industry has adopted automatic decision making with its own characteristics and at different rates. Banks have implemented large-scale systems to manage data-driven fraud control decisions. There are also examples of online advertising, casinos' rewards programs, Amazon and Netflix's automatic suggestions. So far, we have talked about how effective data science is in terms of decision making. Therefore, responsible business people who make decisions with the data science team need to be in constant communication. Firms where business people don't understand what data scientists are doing are at a significant disadvantage. Wasting time and money on projects, or even worse, making wrong decisions can lead to losses for companies.

2. Related Work

Today, data obtained in masses in many different fields are processed and transformed into information, and this provides a chance to make decisions based on this information. In terms of being based on data, it differs from expert systems and appears as a more objective tool. However, using big data methods, different approaches are possible not only for decision-making processes in existing systems, but also for the development of new service and business models. Therefore, the newly entered data-based approach has caused a major paradigm and perception change. At this point, apart from big data analytics, machine learning and artificial intelligence-based methods come to the fore. In this context, different methods come to the fore regarding the calculations and evaluations made in B2B marketing processes. A service structure that will cover B2B processes requires a multi-axis approach since it will cover many different areas. This transformation process, which is taking place in this multidisciplinary perspective within the framework of big data analytics, has brought many innovations and opportunities in practice along with the research framework and approaches. From another point of view, in B2B processes, the primary goal is to determine more appropriate marketing strategies and to make more accurate decisions by providing more suitable marketing processes for businesses within the framework of artificial intelligence technologies, big data analytics, data mining and data science. A data-driven approach revealed in this way can also mean that different social and ethical elements come to the fore, along with different perspectives that have not yet been brought to the agenda. Within the scope of this thesis, it is aimed to develop a forecasting model that can be directly applied within the framework of B2B marketing management.

This thesis is an attempt by a finance company to have an average of over a thousand individual companies come and cash a check per month. This number reaches or even exceeds the level of 400 thousand in the annual balance sheet. One of the most critical issues for the company is the identification of high potential customers, especially those who are in contact for the first time. One third of the prospective customers who contact this company for the first time do not take any action despite receiving approval for different reasons. Although there may be different reasons for this preference such as price, service and duration, field studies show that price and service are among the most important factors. Within the scope of this thesis, among the applicants who are in contact for the first time, those with high customer potential are determined using data-driven modeling approaches, and different marketing strategies are used, such as providing them with a more suitable offer (offering a price reduction), making a special search for the relevant customer candidates, and making this call by a more experienced customer representative. is to increase the transaction rate especially for this target audience. Therefore, by increasing the rate of admission of applications, it both directly contributes to the short-term profitability of the company and increases the possibility of doing business with customers whose potential is determined to be high in the medium term. That will be increase profitability in the long-term.

3. Problem Definition

Within the scope of this thesis project, an approach that will enable the identification of customers with higher potential by examining and analyzing the records, characteristics and behaviors of company customers is the development of a data-driven model. Therefore, the general characteristics of high-potential customers will be determined through the available records and taught to the system. It will be possible to direct the marketing team to more efficient channels if the high potential leads that have just been in the first contact can be identified correctly and successfully. Therefore, it is thought that necessary improvements will be made within the scope of this project, which will have a direct impact on the company's annual profitability. This R&D project is mainly planned to use techniques and technologies such as data science, big data analytics and machine learning.

4. Analysis of the Problem

On average, more than a thousand single companies attempt to cash checks per month to the finance company where the project is implemented. This number is around 400 thousand in the annual balance sheet. One third of the prospective customers who contact this company for the first time do not take any action despite receiving approval for different reasons. Although there may be different reasons for this preference such as price, service and duration, field studies show that price and service are among the factors. Within the scope of this study, different marketing strategies such as identifying high customer potential among the applicants who are in contact for the first time with data-driven modeling approaches and presenting a more suitable offer, making a special call to the relevant customer candidates, and making this call by a more senior customer representative, are used to determine the transaction rate especially in this target audience. specifically aimed to increase. Therefore, by increasing the rate of admission of applications, it will both directly contribute to the short-term profitability of the company and increase the probability of doing business with customers with high potential in the medium term.

5. Thesis Contribution

While an annual average of 47.2% of the customers who have made transactions, which we describe as old, turn their approved applications into transactions, they do these transactions, but the rest do not, despite the fact that they have received approval from the prospective customers who have not taken any action before. The reasons for the customer candidates who do not prefer to make transactions are recorded as the reasons for not making transactions with the company due to price, time, service, system problem, preferring another factoring firm and other reasons. Therefore, the group with the highest potential that the company should target are the potential prospects who have already contacted it but have not taken any action before. Even in the realistic scenario (considering the average of the optimistic and pessimistic scenarios), there is a net profit potential of more than 15 thousand TL per day, and a net profit increase potential of approximately 4 million TL in the annual balance sheet, even in the realistic scenario where the prospective customers in this contact become customers. Achieving this increase in net profit corresponds to large numbers when considered specifically for the factoring sector. Therefore, one of the strategies that the company can use to increase its annual turnover and profit margin is to focus on these first-time leads and get them to trade for the first time. Therefore, as a result of the successful completion of this project target, it is expected to produce a measurable result.

6. Aim of this Project

As it is frequently mentioned in the related academic literature, the biggest problem observed for the classification of customers is which customer characteristics will be used and measured to produce the most accurate and most effective results. Therefore, within the scope of this study, all customer data that will be evaluated and used on the basis of factoring will be examined. The aim of this project is to find high-potential people and what characteristics they have, what customer profile characteristics they show, as well as these factors can be measured for low-potential customers. As a result of the successful completion of these classifications, this project will gain the ability to apply the "Right action for the right person" methodology [6].

7. The Outcome of this Thesis

As the output of this study, an estimation model will be developed that will make an estimation about the potential of the customer, based on the first touch features of the model to be developed, especially in the context of high potential customers (within the scope of transaction capacity and frequency). In the second stage, this model is made adaptive and these processes are automatically processed and finalized together with the new data recording and transaction process, and necessary information is provided to the relevant units. As mentioned above, 50% of the old customers and 30% of the new customers go to the transaction stage. The main expected benefit of this R&D project is to increase the number and rate of transactions for customers driven by the model to be developed for customers making transactions for the first time.

1. THEORETICAL BACKGROUND

1.1. General Framework

One of the most important building blocks of today's businesses and the main factor that can ensure their existence under intense competition conditions is data-oriented approaches. Data obtained in masses in many different fields are processed and transformed into information, and it provides a chance to make decisions based on this information [1]. In the light of this information, by creating the mechanism and structuring the process from raw material to product correctly, producing meaningful results from the output allows for output that provides added value and sustainable success. In this context, new generation analytical methods differ from expert systems in that they are based on data and appear as a more objective tool. Therefore, the newly entered data-based approach has caused a major paradigm and perception change [2]. At this point, apart from big data analytics, machine learning and artificial intelligence-based methods come to the fore. In this context, different methods come to the fore regarding the calculations and evaluations made in B2B marketing processes. A service structure that will cover B2B processes requires a multi-axis approach since it will cover many different areas [3]. This transformation process, which is taking place in this multidisciplinary perspective within the framework of big data analytics, has brought many innovations and opportunities in practice along with the research framework and approaches [4]. From another point of view, in B2B processes, it is the primary goal to determine more appropriate marketing strategies and to make more accurate decisions by providing more suitable marketing processes for businesses within the framework of artificial intelligence technologies, big data analytics, data mining and data science [5].

In factoring, the client applies to the factoring firm for financing, and if the loan transaction is accepted and consummated, the money is deposited into the customer's account. On the due date, the factoring company deducts its payback from the account of the drawer. The transaction procedure consists of three phases. In Step 1, a transaction has occurred between the client and drawer (buyer of the trade). In exchange for this transaction, instead of paying the client (seller) in cash, the drawer issues a check with a future maturity date. Consider a check that guarantees payback 90 days after the exchange. Using this strategy, the buyer agrees to pay the seller the price of the products 90 days after the transaction with the customer (seller). In Step 2, the client (seller of the transaction) presents this check to a factoring company and requests cash in exchange for a specific interest deduction. In this application procedure, the drawer is never engaged; only the client and the factoring company interact. In Step 3, when the check is due at maturity, the factoring company will deduct the payback from the account of the drawer, not the client. Therefore, the capacity of the factoring company to collect the payback relies on the strength or financial condition of the drawer on the due date. Therefore, the factoring company should evaluate the financial and behavioral aspects of the applicant at the time of application (especially on the risk assessment side). Because when the check is due, it is essential that the drawer, who purchased the products on the first day of trade, has sufficient funds in his or her account (not the customer to whom the factoring company made the payment on the first day). The client, however, will approve or reject the interest deduction that the factoring provider would apply. Therefore, the factoring company must have a system that examines both the client and the drawer in the transaction.

The link between a client seeking a bank loan and the bank is also crucial component. This transaction consists of just two parties: the bank and the client. The bank transfers funds to the client's account on the day the loan will be used, and the customer is obligated to return the bank on the loan's due date (maturity or installment due date). The most important aspect in evaluating a bank is the customer's information at the time of application or the customer's transactions with the bank throughout the prior period. There are also discrepancies between the profiles of the client and the drawer in a factoring transaction. Between January 2021 and August 2022, 87.514 clients applied to the factoring company for a higher bank limit. 40% of these clients' bank limitations at the time of application are below 50,000 TL, which is a negligible sum. Similar programs include 154,507 drawers. In contrast to clients, just 6% of purchasers have a bank limit of less than 50.000 TL, while 57% have a bank limit of 1,000,000 TL or more.

Customers' ratio of limit utilization is greater than that of drawers, indicating that customers have a greater demand for finance and use their limits more successfully. A product that is much costly than bank loans is factoring. This is based on the ability of the client to get a bank loan. 29% of the total 17,091 clients (with whom the factoring firm conducted transactions for the first time in 2021) were acquired during the previous 12 months. As a result, banks are unable to do a risk assessment on these clients. This causes banks to withhold credit limits from certain clients. 39% of these clients do not have a bank loan limit as of the date of application for bank loans. Therefore, they need factoring finance, a much costly product, to manage their cash flows. Considering the disparity in factoring bank procedures and the fact that the customer and drawers of a factoring transaction have distinct profiles, in the factoring industry, drawers should be divided in addition to customers.

1.2. Risk Management of Financial Institutions

The management of the financial institution must take on more risk in order to enhance profits for its stockholders. Interest rate risk, market risk, credit risk, off-balance-sheet risk, operational and technological risk, foreign currency risk, nation or sovereign risk, liquidity risk, liquidity risk, and insolvency risk are just a few of the hazards that banks must deal with. The success of a bank depends on its ability to effectively handle these risks. In addition, because of these dangers and the part banks play in financial systems, they are under regulatory scrutiny [6]. Because of the multiple risks that might come from a bank's diverse operations, authorities mandate that banks maintain capital. Since their creation in 1998, the Basel criteria for calculating capital needs have grown and changed. Each of the major risk kinds requires capital. The biggest risk that banks have historically faced is credit risk, which often calls for the most capital. While operational risk is the chance of losses due to internal system faults or outside occurrences, market risk is the risk that results largely from a bank's trading operations. Most major banks compute economic capital in addition to regulatory capital, which is based on a bank's models rather than directives from regulators [7]. Credit, market, and operational risks are the three basic categories of hazards that banks must deal with. Other risks include liquidity, business risk, and reputational risk. Banks aggressively manage, assess, and monitor these risks through risk management.

Market risk is the possibility of suffering losses "due to changes in the level or volatility of market prices" [9]. Interest rate risk, equities risk, foreign currency risk, and commodity risk are all examples of market risk. The potential loss brought on by changes in interest rates is known as interest risk. The possible loss resulting from a negative change in a stock's price is known as equity risk. The risk that changes in currency exchange rates may affect a bank's asset or liability values is known as foreign exchange risk. The potential loss brought on by a negative change in the price of commodities owned is known as commodity risk. Credit is the possibility of a bank losing money if a borrower does not fulfill their commitments (interest, principal amounts). The biggest risk that banks face is credit risk [8]. Banks are permitted to use the internal ratings-based approach to credit risk under the Basel Accord, since they are able to create their own internal credit risk models for estimating potential loss. The three main risk factors that need to be calculated are exposure at default, loss given default, and probability of default (PD) [10].

Asset liquidity risk and financing liquidity risk are the two types of liquidity risk, which are managed separately from the other concerns. When a transaction cannot be completed at the current market prices—which might happen as a result of the position's size in relation to the typical trading lot size—a bank is exposed to asset-liquidity risk. The failure to satisfy cash flow commitments is referred to as funding liquidity risk, often referred to as cash flow risk [11]. Banks must set up a strong framework for managing liquidity risk that would guarantee sufficient liquidity is maintained, including the capacity to resist various stress situations. Implementing a robust approach for the identification, assessment, monitoring, and management of liquidity risk [12]. According to BCBS, operational risk is a "fundamental part of risk management" for banks and is described as the risk of loss brought on by "inadequate or failing internal processes, people, and systems, or by external events." Strategic and reputational risk are not included in this formulation but legal risk is. It is seen to be a natural part of all banking operations, activities, procedures, and systems [13]. Operational risk, which is more commonly referred to as non-financial risk, was reported in the annual reports in a variety of ways and contained a number of subrisks. Among many other things, it included risks related to fraud, cyber security, clients' products and business practices, information and resiliency risk, risks related to money laundering and financial crime, risks related to vendors and outsourcing, risks related to technology, and risks related to business disruption.

As an alternative to using the current literature to evaluate the risks particular to banks, an examination of bank annual reports was conducted. A taxonomy of the many risk types that banks normally aim to manage as part of their business and the strategies and tools in use were plotted based on the review. To ascertain which risk categories were explicitly covered by these banks in their annual reports, ten of the top institutions were examined. The study also included a list of the particular instruments, methods, or elements of the risk management framework that were in use. The list of banks contained a representative from each area, including an Asian bank in addition to US, European, and mostly worldwide active institutions. Additionally, a variety of banking business lines were handled by these institutions, including investment banking, securities trading, client or retail banking, and corporate banking. The main risks, which comprised credit risk management, market risk management, liquidity risk, and operational risk, were basically the same despite variations in how they were handled and presented, including sub-risks.

1.3. Application Credit Scoring

A new applicant's likelihood of making the required payments on time during a specified outcome period—typically 12 or 18 months—is predicted by application credit scoring algorithms. Traditional models would include covariates (or inputs into a machine learning model) like years at address, years in employment, income, age, and credit bureau data like repayment history on prior loans both at that institution and other institutions, the proportion of the population in the postcode that default, etc., that were measured at the time of application [14]. Accounts that have been open for a long enough time for the analyst to evaluate aspects of their use, such as the sum still owing after six months and the average spending on the account over the previous three months, are subject to behavioral scoring models [15]. Additional factors include application and bureau as well as the variables in both types of models may be categorized as socio-demographic and financial [16]. While certain variables may have a category for missing values, others may cause an application to be refused.

A score cannot be obtained and, in the case of a credit application, it is frequently declined if a model includes a variable related to, for instance, the number of credit lines open in the last three months or whether an account has defaulted in the last 12 months, but there is no data for a new (or existing) customer for that variable.

Applications with these kinds of missing information are occasionally referred to as having "no file" or "a thin file." Such factors are present in a relatively large percentage of scoring models. For instance, according to Jennings (2015), a person must have at least one active credit line from the previous six months in order to receive a FICO score [17] and a similar point was mentioned by several other researchers [18, 19, 20, 21]. This is especially prevalent in lower-income nations where a sizable percentage of persons lack credit histories.

While it is possible that the reason you haven't had credit in the past is because earlier credit risk assessments indicated you posed too great a danger for a lender to approve a loan, this isn't always the case. Persons who immigrate to a nation, certain recent college graduates, people who don't utilize an existing financial account, and occasionally even people who have never applied for a loan may not have enough credit history [22]. Since the late 2000s, researchers have experimented with using covariates other than traditional financial and socio-demographic variables to see if their inclusion, either instead of or in addition to conventional variables, increases predictive accuracy or not [23], [24]. Very various forms of information-related variables have been employed. Another study focused on psychometric characteristics and email usage-related variables. There is a dearth of work that examines how well psychometric factors predict outcomes, and much of the empirical material deals with loans to micro-business owners. An earlier empirical study that used a lab experiment discovered that impatience and default were connected [25].

1.4. Data Collection Phase

Numerous analytical modeling exercises begin with a flat dataset, develop a predictive model for a goal measure of interest (such as churn, fraud, or default), and then assess the model on a separate out-of-sample dataset. It is frequently implicitly assumed that the data are independent and uniformly distributed. Recent research challenged this presumption and examined how users of the many social networks that link them might affect one another [26]. There are several social behavior patterns that may be seen. One of these is homophily, which is the strong propensity for people to associate with someone they believe to be similar to themselves in some manner. When people's interactions with other people have an impact on their conduct, this is known as social influence [27], [28]. Other (external, for example) confounding variables may also be responsible for some of the social behavior [29]. In order to effectively utilize the effects of coordinated client behaviors, network learning aims to integrate social behavior patterns into prediction models [30]. Any social network learning activity requires the network, which consists of nodes and edges, as a critical input. The concept of these networks is rather simple in some contexts. Take churn prediction in telecom as an illustration. The network may clearly be built based on information contained in the CDR. An earlier study discovered considerable social network effects for telecom churn prediction [31]. Another illustration is the detection of credit card fraud, where a network is created by linking businesses and credit cards. Strong social network effects have also been seen in this context [32]. Researchers and practitioners of credit scoring firmly believe that a correlation exists between borrower default behavior and credit scores [33].

Small and medium-sized businesses have been shown to be significantly impacted by this interdependency [34]. The defining of the network itself is one of the major obstacles to understanding network effects or default propagation in credit rating. In online peer-to-peer lending, first attempts have been made to create networks between clients. For instance, Lin et al. [35] showed how online connections with those who don't default raise credit scores. Additionally, Freedman and Jin [36] cautioned that internet connections on their own might not provide accurate information regarding credit-worthiness and might even be altered [37] in order to support their findings. Using data from Facebook accounts that was taken from social media networks, De Cnudde et al. [38] created credit rating models for microfinance. Their findings imply that while implicit networks of persons with similar behavior are superior to both explicit friendship networks, explicit networks of friends who communicate are more predictive than explicit networks of friends who do not. Technology businesses like LenddoEFL, which utilize social media connections to measure people's default risk, are already using social networks in the business world to evaluate creditworthiness [39]. Using Wei et al formulation's of the potential value of credit scores obtained with networks - for example, based on social media or calls - and how strategic tie-formation might affect these scores, the interest in using call networks as a new Big Data source for credit scoring has grown more recently. Even while the study is particularly intriguing in light of the Chinese government's proposal for a social credit system [40], it is mainly theoretical and does not provide a crucial empirical assessment of the suggested models [41]. The possibility of call networks as a different data source for credit scoring is also suggested by recent press coverage of specialized smart-phone programs that assess people's creditworthiness using the vast amounts of data generated by their handsets [42], [43]. The majority of these research have either examined the possibility of CDR-induced social networks in credit scoring or have concentrated on the use of social networks in the context of social media.

There is a wealth of literature on the analysis of CDR [44]. The concept of utilizing CDR data for credit scoring comes from the presumption that how individuals use their phones is a good indicator of their manner of life and level of economic activity. According to prior studies, call networks created using CDR data to connect people who are in contact with one another produce social networks that can be used for both descriptive and predictive studies on issues such as age, gender, ethnicity, language, economic factors, geography, urbanization, and epidemics [45–50]. Some of the findings especially demonstrate the unequal distribution of income and debt as well as the stronger bonds that exist between members of the same socioeconomic level. Furthermore, some of the studies examined the customer revenue distribution inside a call network and showed that both high and low revenue customers are generally associated with other high and low revenue customers.

1.5. Algorithmic Approaches to Credit Scoring

Credit scoring is viewed as a categorization issue using a variety of algorithms for analytical reasons [51]. Commonly used as a benchmark for comparison with other, more sophisticated techniques like support vector machines (SVMs), artificial neural networks (ANNs), and extreme learning machines (ELMs) is logistic regression (LR) [52]. Ensemble classifiers, which integrate numerous classifiers for better performance, have become popular in recent research [53]. Beyond accuracy, other factors in classifier selection include complexity and the cost of misclassification [54], which have an impact on the usage and implementation of credit-scoring algorithms. Additionally, more attention has been placed recently in research on how to increase profitability through feature selection or profit scoring [55], [56].

Generally speaking, the types of data utilized for credit rating are traditional or alternative. Demographic data and financial history, such as loan enquiries, are two types of information that are typically utilized for credit rating. The use of alternative data in credit scoring has increased as a result of the lack of these forms of information for those who are financially excluded. In recent years, a number of businesses with operations in underdeveloped nations have started to provide digital credit services, pre-screening potential borrowers based on information from their mobile phone usage [57]. For the purpose of choosing clients to switch from prepaid to postpaid mobile phone subscriptions, a type of digital credit [57] established a credit-scoring algorithm. The frequency and length of conversations were used as behavioral characteristics from mobile phone usage to forecast defaults. For thin-file debtors, it was discovered that these indications outperformed credit-bureau information (borrowers for whom credit bureaus hold limited information). When financial history (including bank account and credit card activity) and mobile phone usage data were combined, a considerable increase was observed in credit-scoring-model performance, as shown by the area under the receiver operating curve (AUC) [58].

The integration of mobile application usage data for credit evaluation with alternative score variables for people who are financially disadvantaged was suggested [59]. Also, who used data from mobile financial transactions to develop a credit-evaluation method for unbanked persons [60], mobile applications offer another source of alternative data. In order to make data gathering for financial organizations simpler, the related research findings [60] suggested using a mobile application to gather information from social media for credit scoring. The researchers [61] suggested a technique for recursively adding client network data to increase the accuracy of credit-scoring algorithms.

1.6. Applications of Machine Learning Methods

The most significant use of machine learning is the mining of predictive data [62]. A collection of features is utilized by machine learning algorithms to create predictions based on a dataset. Depending on the specific issue and data, these characteristics might be continuous, categorical, or binary. In general, there are two learning methods: supervised and unsupervised learning. If labels exist for each instance of a dataset, we refer to this as supervised learning. In supervised learning, the objective is to identify newly encountered unlabeled data. Using labeled data in training to determine the description of classes, which is then used to label freshly encountered data [63] accomplishes this. The two subcategories of supervised learning are classification issues and regression problems. Classification is training a labeled set and attempting to predict those labels in unlabeled data. Popular classification problem-solving algorithms include KNN, decision trees, random forests, and support vector machines. Regression, on the other hand, is used to comprehend the link between dependent and independent variables. These are the most fundamental regression algorithms: linear regression, logistic regression, and polynomial regression. Accuracy, f1-score, Receiver Operating Characteristics (ROC), and confusion matrix are a few of the assessment measures used to evaluate the performance of supervised learning systems. Based on two criteria, the ROC curve analyzes the prediction ability of a model for a binary classification task.

Unsupervised learning predicts without labels. Clustering, anomaly detection, and density estimation dominate unsupervised learning [64]. Clustering utilizes unlabeled data to categorize patterns. Data-driven labels are used [65]. Recommendation engines, picture segmentation, and dimensionality reduction are clustering applications. K-means, DBSCAN, agglomerative clustering, and affinity propagation are often used for these issues [66]. K-means and DBSCAN can also identify anomalies. GMM, minimum covariance determinant (fast-MCD), isolation forest, local outlier factor (LOF), and One-class SVM also identify outliers [67]. Unsupervised learning may also estimate density. It estimates the probability density function of the random process that created the dataset for data display and analysis. GMM and DBSCAN estimate density [68]. To create a better algorithm that combines many algorithms to complement one another, these features should be grasped. In other scenarios, it may be hard to identify a single algorithm that delivers the greatest accuracy or other score metric. In these cases, merging two or more classifiers is called ensemble learning [69]. Voting, bootstrap aggregating (bagging), pasting, boosting, and stacking are prominent ensemble techniques. Voting method classifiers may outperform single classifiers. This ensemble method's classifier diversity improves accuracy by causing diverse sorts of mistakes. Bagging approach trains predictors using one algorithm on multiple subsets of the original dataset. Bagging requires sampling with replacement. Pasting is sampling without replacement [70]. Boosting ensemble learning approaches train predictions progressively to correct each other. Adaptive, Gradient, Extreme Gradient, LightGBM, and CatBoost are the most popular

[71]. Instead of aggregating classifier predictions, stacking algorithms (stacked generalization) train a model from an ensemble of algorithms. A meta learner or blender uses classifier results as training data to create the final prediction [71].

2. DATA PREPARATION AND PREPROCESSING

2.1. Data preparation and Data Preprocessing

This step involves the preparation and processing of the data to get it to be implemented on machine learning models. It can contain several sub-steps depending on the problem at hand.

2.2. Data Collection Phase

The very first step of a machine learning problem is to gather relevant dataset. While gathering it, it is essential to collect the most informative features. It can be done by using an expert's knowledge in that domain. If an expert is not available, then the only option is to use brute-force, using every feature available. The downside of this method is that it comes with noise and missing values, which requires notable data cleaning and preprocessing [72]. In most cases, the dataset at hand is full of noises and errors. Real-world data does not come perfect, so it needs to be corrected. A hierarchy of problems has been proposed to be dealt with to make the dataset ready to be used in algorithms. First thing to look at is the presence of impossible values inputted in features [72]. For instance, if the relevant feature is expected to have binary values, but one instance of it has a discrete value, then we identify it as an impossible value. They can be solved ideally while inputting phase of the data, so that they can be corrected. However, if it is impossible to enter the correct values, they can be simply treated as a missing value category to be removed from the dataset [72]. Next problem to be looked at is that no values have been inputted in an instance of a feature [73]. There are several methods that solve this issue which most of them will be mentioned in upcoming paragraphs. Finally, some features that are irrelevant are present in the dataset [74]. They are simply ignored and left out of the dataset.

2.3. Data Cleaning

Data cleaning is one of the most important steps in data mining. Detecting and repairing dirty data is the crucial part of data preprocessing. If dirty data left unrepaired, it would lead to inaccurate analytics and unreliable models. Data cleaning usually consists of two phases: error detection and error repairing. For detection of the errors, quantitative and qualitative approaches exist in defining those errors. While quantitative techniques employ statistical methods to be used in outlier detection, qualitative techniques implement rules, constraints, and patterns to handle errors in the data [75].

2.4. Outlier Detection

Next, evaluate unlikely values [72]. This category includes values that vary greatly from others in the sample [73]. Outliers may also be defined as observations that deviate from the data [73]. Variable-by-variable data cleansing is one approach. These outliers are identified due to their unusual probability distribution. They are greater than one standard deviation from the mean in a normal distribution, or more depending on the domain and distribution. These values are usually removed from the dataset because they may be caused by mechanical flaws, system behavior changes, fraudulent activities, human error, instrument error, or population variances [73]. For the third explanation, it is possible that values at the tails of a distribution, where it is more dispersed than previously thought and hence more variable [73], are right. They may also represent veridical facts that belong to a cluster or label but are within another cluster [73]. In most cases, they also removed material to construct an accurate, public-friendly model.

Some jobs require finding outliers. Safety-critical settings where an outlier causes abnormal operating conditions, such as an aircraft engine's rotation fault or a nuclear power plant failure, may have environmental consequences. Another may find a system invader that needs rapid attention. A factory production line may identify an outlier by frequently comparing the properties of a typical product against those of newly created items to uncover faults and decrease mistake costs. Monitoring a customer's credit card use to detect a rapid change in usage pattern, which may indicate a stolen card, may help discover fraudulent activity. Outliers are identified by comparing time series of consumption data [73]. Loan applications are screened for fraud or troublemakers using outlier detection. Thus, a bank may recognize a potentially troublesome client early on and act accordingly to prevent future loans from being made or to terminate the customer's credit limit to minimize increased credit use. A demographic research may reveal tall outliers. Thus, it is normal and may arise depending on study. Outliers warn fraudulent surveillance cameras. To enhance detection, outlier data may be stored elsewhere if the alarm was properly triggered.

Semi-supervised detection or identification teaches the usual class and the model learns to spot irregularity. This approach induces a normalcy boundary using normal and pre-labeled data. Each fresh data set may train the model and tweak it. For generalization, all normal data must be available. Anomalies, unlike type 2 data, are not needed for training. This is beneficial in defect detection, where it would be too costly to obtain anomalous data by inflicting engine damage to train the model. Outlier detection in aviation engines would be too costly to train the model by destroying the engine. Fraud detection systems may mishandle emerging fraud types. The model may detect new fraud until it falls within the typical range by modeling normality.

2.5. Methods of Outlier Removal

The best way to handle outliers is to show a feature to recognize negative values that arise in a regular pattern, which is also a powerful and effective tool. Outlier identification methods depend on two elements. First, find a method that can model the data distribution and detect outliers for a clustering, classification, or recognition model. Second is picking a good topic. Neighborhood selection should fit all distribution densities. Most outlier detection systems have similar origins but different names. Examples include outlier, novelty, oddity, noise, deviation, and exception mining. Statistics, neural networks, and machine learning are used to identify outliers. Some algorithms choose one or more of these fields to improve their model.

2.6. Statistical Methods for Outlier Removal

Outlier detection started using statistical methods. Some of the initial techniques work only on unidimensional data sets, whereas others are univariate. Grubbs' approach is one-dimensional. It divides the difference between the attribute's mean value and the query value by the standard deviation of all values to get a Z value. Data-generated parameters eliminate the necessity for user-supplied parameters in this method. A higher amount of data values makes the model more statistically representative [74]. Statistical models work well with quantitative real-valued and quantitative ordinal data sets, limiting their applicability. Appropriateness increases processing time if data transformations are significant [73]. Informal box plots may find univariate and multivariate outliers. Box plots show the lower extreme, lower quartile, median, upper quartile, and higher extreme [74]. They find unusual values in categorical data sets and work for symmetric and asymmetric distributions. Outliers may be seen by sight. Choosing upper and lower thresholds 1.5 times the interquartile range (IQR) from the upper and lower quartiles, where the lower and higher outliers are, is more effective. Univariate outlier removal examines each variable for outliers. If there are too many univariate outliers that correspond to a lot of data, arrange them by frequency and delete the most frequent ones. Univariate outlier detection relies on data ordering, usually ascending, to obtain the five-number summary. Multivariate data have no complete ordering. A reduced sub-ordering technique addresses this problem. A distance metric converts multivariate observations into scalars. The Mahalanobis distance is optimal for multivariate outlier detection since it includes interdependence between features and finds odd value combinations. Many distance measures, like Euclidean distance, solely give location information, making them inappropriate for data collecting.

2.7. Machine Learning-based Outlier Removal Methods

Most statistical methods cannot identify categorical data outliers, but machine learning systems can. Categorical outliers are identified using the C4.5 decision tree, since decision trees may identify outliers faster than statistical methods since they do not need data distribution or features. Rule-based systems may be updated or altered to identify outliers, making them more flexible and incremental than decision tree techniques. After data analysis, characteristics will be evaluated. The variables' missing data will be the focus here. Missing data has two reasons. Data loss without recovery. Missing data hinders machine learning. Thus, how to fill in missing data is crucial. A missing data scenario analysis will follow a preliminary research. These include literature-proven MCAR, MNAR, and MAR states.

2.8. Manage Inadequate Data

Most data sets have partial or missing values. Missing values may be due to forgetfulness, loss, inapplicability, or the data set designer's unwillingness to give a value [72]. To find a good way to handle not applicable and lost values, one must investigate their unpredictability. Missing fully at random, missing not at random, and missing at random are three randomness types (MAR). Accidents like subject measurement loss cause MCAR. Thus, absence probability is independent of other characteristics. Unlike MNAR data, where missing values are correlated with unobserved information about other subject traits, this kind of missing value may be handled in many ways. Errors in experiment design may cause missing values in a data collection that cannot be handled by a universal approach [74]. MAR happens when missing data depend on the outcome or other predictors. If a feature has missing values after a device breakdown but not during normal operation, it is MAR.

Missing value incomplete data problems may be resolved in numerous ways. Resampling or rerunning the experiment may fix MNAR issues. MCAR and MAR scenarios impute missing values depending on other factors or their own characteristics. For missing data, single and multiple imputation are utilized. Data is imputed instead of erased to reduce bias. In MNAR data, deleting incomplete occurrences may be the best option [74]. Single imputation methods include mean, distribution, regression, and KNN. Mean, median, or mode imputation replaces missing values based on data distribution. This method has several drawbacks, such that a significant number of missing values may distort the data distribution. Distributions impute missing data without changing their shape [74]. Regression techniques incorporate more variables and imputation, making them more sophisticated. For unbiased imputation, MAR and MNAR data must be linearly connected to the imputed feature [74].

KNN single imputation employs k neighbor distance to impute similar values. This method is computationally expensive ask rises. Numerous imputation averages imputed data sets. This reduces single-imputation biases in ambiguous imputed data. Thus, it executes numerous valid imputations and averages them to reduce uncertainty and single-imputation errors. This method involves imputing a data set n times, assessing it, then consolidating it [74]. Multiple imputation by chained equations (MICE) is the most used method for handling missing data. It uses the correlation between imputed attributes and others to select an appropriate method.

2.9. Modeling Framework

This section includes all phases such as data cleaning, outlier detection and removal, imputation for missing values, data clustering, one-hot encoding, and standardization) that makes the data ready to be put into prediction models. An explanation of how the modeling framework was created will be explained. Since there were five algorithms and three feature extraction methodologies were used and additionally a wrapper method as feature selection, it would have been a little bit complicated and unoriented to compare all their results with each other, unless a modeling framework was designed. Therefore, there were six main frameworks defined i.e., main model without feature extraction, model with PCA, model with t-SNE and model with Isomap, model with frontal alpha asymmetry transformation and model with feature selection. In the main model, all ML algorithms were used and then the best algorithm was selected to be used in feature extraction models, and the best model among feature extraction and main models was chosen to apply feature selection to further improvement of score. In the next section, ML algorithms that were used in the main model will be mentioned indicating their properties and their hyperparameter tuning.

3. RESULTS

This chapter presents the results of each part written in data collection, model development and validation methodologies. Within the scope of the project, the objective outputs of the project were determined in line with the requests and opinions of the relevant units. In line with the specified outputs, the expected outputs for the project are as follows: Predicting the number of transactions and volumes of customers in the next 3, 6, 12 months periods estimating the probability of making a transaction for customers who make their first inquiries.

3.1. Predicting Customers' Next Transaction times

Obtaining the specified outputs at the end of the project process, the project process has been planned in line with the above-mentioned items. Within the scope of the project, the outputs of which were determined, analytical approach methods were determined. During the analytical approach process, three main actions were decided. The methodology and framework of the work throughout the project was determined. Then, the technological tools to be used were decided. It is planned to use Oracle database systems used within the company for the database for technological tools. Python programming language was used for data analysis, processing and modeling stages. Necessary meetings were held with the company's customer analytics and sales units in order to achieve the determined business outputs. As a result of the meetings, data that can be used for the project were determined in line with the findings. The determined data has been recorded, documents on other data owned by the company on the basis of customers have been examined, and all data that can be used have been documented.

The data determined and needed in line with the project outputs were drawn from the Oracle database of a factoring company, which is officially in use, by writing SQL queries. Risk center data obtained from the Credit Registration Office (CRO), including customer risk information of the data to be used, and other financial information of customers were used. The data coming from CRO includes customer information such as factoring grades, current debt status, number of credit institutions, risk information, limit information and credit information under follow-up. In addition to customer risk information, customers' inquiries at this company, number of transactions, total contributions, etc. information has been added to the dataset. In addition to this information collected in the data set, other attributes suggested as part of business information will be added to the data set by using attribute engineering after the data analysis and as a result of the meetings with the customer analytics unit and sales units. Information on these data is detailed in the data understanding and preprocessing section.

The obtained data were then converted to CSV format and stored in the database. Later, data files in CSV format were converted to Parquet file type due to the large space of the data in CSV format and the slowness of reading and writing operations on it. Although it is stored in different file types, the problem of data files taking up too much space has been solved by establishing a direct connection between the database programming language, since the data stored in different file types takes up a lot of space due to the changes made on the data and the change of the data set as a result of the updates. Using the cx_Oracle library, a connection was established between the Python programming language and the Oracle database. Thanks to this connection, high efficiency has been achieved both in terms of the space occupied by the file systems and the speed of reading the files, and at the same time, the resources have been made more efficient.

3.2. Feature Selection and Engineering

After the missing value analysis was completed, different methods were used for both categorical and numerical values to fill in the data. These are univariate and multivariate filling methods, respectively [74]. Average assignment, last value assignment, previous value assignment and random value assignment methods were used to fill in numerical data from univariate filling methods. In order to fill the categorical variables, the process of assigning the most repeated class values in the variable and assigning 'missing' values to the missing values were performed [78]. In addition to these methods, multivariate filling methods such as kNN filling method, Hot-Deck filling method, MICE (multivariate imputation chained equations) and MissForest-extends methods was also used [74].

Cardinality status for categorical variables was also analyzed. The cardinality represents the number of classes of categorical variables. Cardinalities affect multiple conditions. First of all, most machine learning models do not accept categorical variables, but it causes excessive compatibility problems in 'Tree' based algorithms [75]. Apart from this, it is likely to cause many operational problems. Analyzing and resolving the high cardinality situation has provided a positive effect on model performance [75]. Since the distribution of categorical variables between classes is very important, rare classes within categorical variables were identified and these classes will be combined and written as a single class or included in the analyzed class. As a result of the analyzed categorical variables, encoding was achieved. Three different methods were used for encode operation. Traditional methods, monotonic relationships and alternative methods. Traditionally, One-hot encoding, frequency encoding and ordinal tag encoding methods were used. Monotonically ordered variable analysis, WOE (Weight of Evidence) and Mean encoding methods will be used. Alternatively, Binary encoding, Feature Hashing and Rare Labels encoding methods were used.

Probabilistic distribution methods were also used in order to analyze the distributions of the variables. More than one method will be used for discrete and continuous variables. Binomial and Poisson distributions for discrete variables were tested. For continuous variables, the Gaussian distribution and skewness test were performed. In order to normalize the variables as a result of the tests performed, logarithmic transform, reciprocal transform, square root transform, exponential transform, box-cox, and yeo-johnson transform methods were used. After the data distribution analysis, outliers were analyzed to bring the variables closer to normality. Analysis and handling of outliers had a direct impact on model performance. Quantitative analysis was first performed for outlier analysis and then the LOF (local outlier factor) method was used for all variables.

3.3. Feature Scaling

For data scaling, standardization, mean normalization, min-max scaling, robust scaling and scaling to unit length methods, which have been found to give successful results in the literature, were used.

3.4. Model Development

For the model development phase, the data set was divided into 3 parts, training validation and test set. The test set will be selected from the data within 3-6 months before the model will be trained. Data before these dates will be reserved for the training set. More than one model will be created from the result of the data set created for modeling. Machine learning models will be used for modeling. Clustering, classification and regression models will be applied in line with the objectives. For customer abandonment analysis (Churn), estimation will be made for periods of 3 months, 6 months and 12 months. In addition to this, modeling will be done whether the customers will perform their next transactions or not. Classification models will be used for these estimations and for the analysis performed. Then, the number of transactions and contributions that customers will perform in the next 3 months, 6 months and 12 months will be calculated. Regression models will be used for process estimation and contributions. In addition to the regression models to be used, in case the expected accuracy values cannot be reached and their usability decreases, the models for estimating the number and volume of transactions for the next 3, 6 and 12 months of the customers will be changed to low, medium and multipotential, and regression models were not used by converting them into classification problems.

Three different machine learning methods were applied for modeling. These are: Classical Learning, Ensemble Learning and Artificial Neural Networks methods. Classical learning will be examined under the name of supervised and unsupervised learning. In ensemble learning, “Bagging” and “Boosting” methods will be applied. With the bagging method, each classifier is trained with randomly selected different subsets of the training set. It works asynchronously. The training set is divided into sub-parts and the model is established with all of them and the best model is selected. If there is a bias in the data set, it will be more logical to use this method and it will produce a solution against the over-compatibility problem. Unlike the Bagging method, it is more appropriate to use if there is high variance in the data set. In the boosting method, the output of each classification is the input data of the other classification. It works synchronously. In the data sets divided into subsets, the model is built respectively and one model works as the input of the other. As a result of the modeling, the weights of the observations that are detected incorrectly are increased and the data weights are changed. K-Nearest neighbors (kNN), Logistic regression, Decision trees, Random Forests, CatBoost, LightGBM, XGBoost and Neural Networks were used for classification models.

These processes will be automated as the modeling processes need to be repeated multiple times and the results should be compared. In order to automate it, a program that receives data from the user and trains this data and presents the results in the form of a report were developed. As the data sizes increase, the modeling costs also increase, and accordingly, the increase in the number of independent variables increase the complexity of the model and lead to overfitting problems. For this, the variable selection method was used for classification problems. Three different methods were also applied for variable selection. These are forward and backward iterative algorithms, filtering methods and embedded methods. Forward and Backward Iterative Algorithms: This method aims to achieve the best result by trying all possible combinations of independent variables by adding or subtracting all independent variables from the model, respectively.

The filtering method uses statistical approaches and applies feature selection by looking at the correlation between the independent variable and the dependent variable. Although it is very cheap in terms of computational cost, it can reduce the overall model score [76]. Embedded methods combine the attributes of filtering and iterative methods. It is used by the classification algorithms themselves. Lasso, Ridge and Elastic Net regression methods penalize the overfitting problem. This method is cheaper than iterative methods in terms of cost and expensive in filtering methods [76].

3.5. Model Hyperparameter Optimization

Since Random Search tries random hyper parameter combinations, the total number of parameter combinations is less than the grid search method. This reduces the time complexity and makes it difficult to find the best parameters [76]. Grid Search is the automation of manually performed hyperparameter optimization. When the grid search method tries all parameter combinations in order, it increases the time complexity [76]. The Halving Grid Search method is an optimized version of the grid search method. The halving grid search method searches through the predicted parameters by applying the sequential division method. It starts by testing all parameters on these sample sets by packing all observations into small packages, and recursively selects the best parameters by combining small sample packages [77]. Finally, the threshold value setting method will be applied. With this method, it is aimed to make unbalanced data sets more useful. In particular, the above-mentioned AUC-ROC curve will be used as a support for this topic. In general, the threshold value is 0.5 in all machine learning models. Adjusting the threshold value will not change the model score, but will directly affect the score distribution between classes [78]. This threshold calculation part will play an important role, especially since the datasets in the process of working on the models, such as predicting the transaction status of customers who come into contact with Tam Finans for the first time, are unbalanced.

3.6. Model Evaluation and Selection

For model evaluation, the evaluative metrics mentioned above will be used on the basis of each different model. Due to the large number of measurement metrics, the part with the automated object-based measurement metrics comparison model created for the modeling part will be integrated to facilitate the calculation of these metrics. Measurement metrics will be automated and the best model will be selected in this way. Afterwards, hyperparameter tuning was done for the best selected models. Hyperparameter optimization provides a basis for reaching the best parameters by training or optimizing the models according to certain coefficients in the process of creating the models. Especially since different models will be established in line with different targets, very different hyper-parameters will appear. At this point, the parameters frequently used in the literature will be given priority. Three different methods will be used in the hyper parameter tuning process. These are: Random-Search, Grid Search and Halving Grid Search methods.

3.7. Model Deployment, Integration and Maintenance

The created model will be transmitted to the software unit using version control systems. The model will be put live on servers that are currently active. During this process, server-client architecture will be used. The machine learning model will be sent to the generated API. Flask will be used as an API. The API interacts full-time with the database both to obtain the model results and to record the findings obtained by the machine learning model. The results will be distributed over the ready user so that they can be used by users and customer representatives. The results to be inspected will be performed on the basis of the customer code. In order to obtain the outputs produced as a result of the model, the desired outputs will be reached by sending the customer code to the live machine learning model. A control group will be created during the deployment phase. The information of the customers who make a new move due to the active inquiries and transactions of the customers will be stored and integrated at the end of the day. Since there are too many customer movements during the day, the query process must continue uninterrupted, and all new movements will be added to the data set at the end of the day in terms of cost and server load, and the model will be made operational again, and new results will be produced.

The outputs obtained during the feedback process are used by the customer representatives. As a result of the results of the outputs of the models and the feedback received from the customer representatives, the cases that may be faulty will be examined and the revision cases will be taken into consideration. Afterwards, the measurement and deployment phases will be continued again, and the workflow will continue in the modeling, measurement, deployment and feedback loop. In fact, it is planned to receive ideas, opinions and suggestions from CRMs, which are the target user group, at every stage. However, at this stage, these notifications will be made in a much more regular and targeted way, with the aim of directly increasing the system usability and improving the user experience.

4. DISCUSSION AND CONCLUSION

Afterwards, cross-validation was performed to ensure correct and reliable outputs as a result of model setup. For the cross-validation method, 80% of the data is divided as training and 20% as testing. Since the data set is unbalanced, stratified sampling was used and the cross-validation value was chosen as 5. Since the data set is unstable, the direct accuracy rate was not used to analyze the model results. For the analysis of model accuracies, complexity matrix, balanced accuracy, geometric mean, dominance, imbalance index, ROC-AUC curve, Precision-Recall curve were used. In particular, the ROC-AUC curve and the Precision-Sensitivity curves are used to set the threshold value. By adjusting the threshold value, the accuracy difference between the classes is reduced. After modeling measurements, hyperparameter tuning (optimization) was performed with halving grid search for the models that gave the best results. Made for Support Vector Machines, Logistic Regression, Random Forests and XGBoost.

Support Vector Machines (SVM) gave the best results for the individual customer dataset, while the XGBoost classifier gave the best results for commercial customers. The average F1-Score for real customers was 79%, while the average F1-Score for commercial customers was above 75%. Since the score values of the dominant target variable are high, the threshold value for real companies was determined as 0.46 using the ROC-AUC curve. On the other hand, the threshold value of commercial companies was determined as 0.38. The models developed as a result of the study are currently being used within the company, and within the next 3 months, the results will be checked and the necessary adjustments will be made again according to the situation, and this will continue in this life cycle. It can be said that sensitivity scores are more important since high potential customer candidates can be found as the final target within the scope of this project [80].

Accurate estimation was achieved for sensitivity scores, ie how many of those actually labeled as high potential. Another is Precision scores, i.e. how many of those predicted as high-potential leads are truly high-potential [81]. Approaching these two scores from different perspectives shows that different actions can be taken. Within the scope of this study, it can be stated that sensitivity scores come to the forefront due to the fact that high potential customer candidates can be found as the final target [82].

Since there is no similar model actively used in the organization, it has a very innovative effect when evaluated from the company's point of view, and the developed model has added value since it has been put into practice. It is planned to direct the marketing team to more efficient channels if the high potential customers who have just made the first contact can be identified correctly and successfully. Therefore, it is anticipated that necessary improvements will be made within the scope of this study, which will have a direct impact on the company's annual profitability.

Afterwards, cross-validation was performed to ensure correct and reliable outputs as a result of model setup. For the cross-validation method, 80% of the data is divided as training and 20% as testing. Since the data set is unbalanced, stratified sampling was used and the cross-validation value was chosen as 5. Since the data set is unstable, the direct accuracy rate was not used to analyze the model results. For the analysis of model accuracies, complexity matrix, balanced accuracy, geometric mean, dominance, imbalance index, ROC-AUC curve, Precision-Recall curve were used. In particular, the ROC-AUC curve and the Precision-Sensitivity curves are used to set the threshold value. By adjusting the threshold value, the accuracy difference between the classes is reduced. After modeling measurements, hyperparameter tuning (optimization) was performed with halving grid search for the models that gave the best results. Made for Support Vector Machines, Logistic Regression, Random Forests and XGBoost.

Support Vector Machines (SVM) gave the best results for the individual customer dataset, while the XGBoost classifier gave the best results for commercial customers. The average F1-Score for real customers was 79%, while the average F1-Score for commercial customers was above 75%. Since the score values of the dominant target variable are high, the threshold value for real companies was determined as 0.46 using the ROC-AUC curve. On the other hand, the threshold value of commercial companies was determined as 0.38. The models developed as a result of the study are currently being used within the company, and within the next 3 months, the results will be checked and the necessary adjustments will be made again according to the situation, and this will continue in this life cycle. It can be said that sensitivity scores are more important since high potential customer candidates can be found as the final target within the scope of this project.

Accurate estimation was achieved for sensitivity scores, ie how many of those actually labeled as high potential. Another is Precision scores, i.e. how many of those predicted as high-potential leads are truly high-potential. Approaching these two scores from different perspectives shows that different actions can be taken. Within the scope of this study, it can be stated that sensitivity scores come to the forefront due to the fact that high potential customer candidates can be found as the final target. Since there is no similar model actively used in the organization, it has a very innovative effect when evaluated from the company's point of view, and the developed model has added value since it has been put into practice. It is planned to direct the marketing team to more efficient channels if the high potential customers who have just made the first contact can be identified correctly and successfully. Therefore, it is anticipated that necessary improvements will be made within the scope of this study, which will have a direct impact on the company's annual profitability.

One of the expected results in line with the project outputs is to estimate the probability of making transactions for customers who contact the company for the first time and make inquiries. This was chosen for both preliminary findings and project initiation across the entire project. Customers who contacted this factoring company for the first time between March 2019 and February 2022 were selected. Segmentation has been done. There are too many differences between data values before segmentation. Among the clustering methods, first of all, the K-Means method was used. The K-Means method is sensitive to scaling, and for this reason, the data were scaled with different methods before clustering. The applied methods are: Unscaled, Standard Scaler, Min-Max Scaling, Max Abs, Robust Scaling, Yeo-Johnson, Gaussian-PDF, Uniform-PDF and L2 normalization methods. In order to compare the cluster results, 4 measurement metrics were determined. These are: Silhouette Score, Davies Bouldin Index, Calinski Harabasz Index and Elbow methods. The Silhouette Score is the average distance between a data point and all other data points in the cluster, the average distance to other data points in the cluster closest to that data point. Since the Silhouette Score measures how well a cluster fits into the data within its own cluster compared to other clusters, it should be high and its values vary between -1.1. The Davies Bouldin Score is used to measure the average similarity between different categories. Since Davies Bouldin measures the similarity between different clusters, it is more important to be low and the lowest score is 0. The Elbow method is used to find the optimal number of clusters in a cluster. For this method, random cluster number values are selected and then K-Means are applied using each cluster number value. The distance from the center of gravity of each point (data) in a cluster is found and the value where the average distance falls the fastest is selected. The Calinski-Harabasz method is a measure of how similar a data is to its own cluster compared to other clusters. The similarity here is estimated based on the distances from the data points in a cluster to the cluster center. Its segregation is based on the distance of the cluster centers from the spherical center. The optimum cluster selection, on the contrary to the elbow method, is the place where the fastest upward breakdown occurs.

An object-based automated clustering model has been developed to automate the clustering process. With this developed model, it is freed from the burden of repeating multiple times and setting up different models for different metrics. The model works based on a clustering method. The model you want to set up and the data set used together with its parameters are the input parameters to the model. The raw data is scaled in 8 different ways, and it is aimed to obtain the optimum number of clusters with the best scaling method. After first scaling the raw data, after the model setup: Min-Max Scaling, Max-Abs, Robust Scaling, Yeo-johnson, Gaussian-PDF, L2 Normalization and Uniform-PDF were applied. Afterwards, the cluster model was created and trained with the scaled data. For the created model, then the score metrics Silhouette, Davies Bouldin and Calinski Harabasz were calculated. The calculated metric scores are stored in the data frame structure. Then the score metrics were compared with each other. As a result of this comparison, the variation graphs were obtained by taking the differences of the metrics. Finally, the obtained comparison and metric score differences graphs are visualized.

In conclusion, the classical machine learning methods work on pattern analysis and correlation. It may be insufficient for causal analysis. It is very important to better understand and understand the causality between the independent variables and between each other and the dependent variable, and to make estimations using these causal effects. In addition to classical machine learning methods, 'Causal Inference' methods will also be used. The results obtained from classical models and 'Causal Inference' methods will be compared. In addition, more healthy and reliable machine learning models will be obtained with the outputs obtained from the 'Causal Inference' methods. For the 'Causal Inference' method, 'Do Why' and 'EconML' libraries developed by 2 different teams of Microsoft will be used. In addition to these libraries, 'Causal ML' libraries developed by Uber will be used.

REFERENCES

[1] M. P. Çakir, T. Çakar, Y. Girisken, and D. Yurdakul, “An investigation of the neural correlates of purchase behavior through fNIRS,” *Eur J Mark*, vol. 52, no. 1–2, pp. 224–243, Feb. 2018, doi: 10.1108/EJM-12-2016-0864.

[2] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing”, *Decision Support Systems*, 62, 22-31, 2014.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] Warburton, K, “Deep learning and education for sustainability”, *International Journal of Sustainability in Higher Education*, 4(1), 44-56, 2003.

[5] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, and M. Karsai, “Socioeconomic correlations and stratification in social-communication networks”, *Journal of The Royal Society Interface*, 13(125):20160598, 2016.

[6] A. Saunders, M. M. Cornett, and P. A. McGraw, “Financial Institutions Management: A Risk Management Approach” *New York: McGraw-Hill*, 2006.

[7] H. John, “Risk Management and Financial Institutions”, *New York: John Wiley and Sons*, vol. 733, 2012.

[8] R. Apostolik, C. Donohue, P. Went, and Global Association of Risk Professionals, “Foundations of Banking Risk: An Overview of Banking, Banking Risks, and Risk-Based Banking Regulation”, *New York: John Wiley*, 2009.

[9] J. Philippe, “Value at Risk: The New Benchmark for Managing Financial Risk”, *New York: McGraw-Hill*, 2007.

[10] Basel Committee on Banking Supervision. 2005a. Guidance on Paragraph 468 of the Framework Document. *Basel: Bank for International Settlements*.

[11] Basel Committee on Banking Supervision. 2006. Minimum Capital Requirements for Market Risk. *Basel: Bank for International Settlements*.

[12] Basel Committee on Banking Supervision. 2008. Principles for Sound Liquidity Risk Management and Supervision. *Basel: Bank for International Settlements*.

[13] Basel Committee on Banking Supervision. 2011. Principles for the Sound Management of Operational Risk. *Basel: Bank for International Settlements*, pp. 1–27.

[14] L. C. Thoma, J. Crook, and D. Edelman, Credit Scoring and Its Applications. *London: Siam*, 2017.

[15] L. C. Thomas, “A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers”, *International Journal of Forecasting*, 16(2), 149-172, 2000.

[16] S. Meier and C. Sprenger, “Present-Biased Preferences and Credit Card Borrowing”, *American Economic Journal: Applied Economics,* 2(1), 193-210, 2010.

[17] A. Jennings, “Expanding the credit eligible population in the USA: a case study. , Edinburgh” Presentation at *the Credit Scoring and Credit Control XIV Conference*, 2015.

[18] S. Agarwal, S. Alok, P. Ghosh, and S. Gupta, “Financial inclusion and alternate credit scoring for the Millenials: role of big data and machine learning in Fintech. Business School”, *National University of Singapore Working Paper,* SSRN 3507827, 2015.

[19] P. Carroll and S. Rehmani, “Alternative Data and the Unbanked Oliver Wyman Report”, 2017.

[20] K. P. Brevoort, P. Grimm, and M. Kambara, “Credit invisibles and the unscored”, *Cityscape*, 18(2), 9-34, 2016.

[21] J. San Pedro, D. Prosperpio, and N. Oliver, “Mobiscore: towads universal credit scoring from mobile phone data”. In Ricci F. and Bontcheva K. and Coulan O. and Lawless S. (eds) *User Modelling, Adaptation and Personalisation, 23rd International Conference*, UMAP 2015, Dublin. Proceedings, 2015.

[22] C. J. Makela, T. Punjavat, and G. I. Olson, “Consumers’ credit cards and international students”, *Journal of Consumer Studies and Home Economics*, 17, 173- 186, 1993.

[23] S. De Cnudde, J. Moeyersoms, M. Stankova, E. Tobback E., V. Javaly V. and S. Martens, “What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance”, *Journal of Operational Research Society,* 70 (3), 353-363, 2019.

[24] M. Oskarsdottir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, “The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics”, *Applied Soft Computing Journal*, 74, 26-39, 2019.

[25] S. Meier and C. Sprenger, “Present-Biased Preferences and Credit Card Borrowing”, *American Economic Journal: Applied Economics,* 2(1), 193-210, 2010.

[26] V. Barnett and T. Lewis, “Outliers in statistical data,” *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics,* 1984.

[27] D. B. Skalak and E. L. Rissland, “Inductive Learning in a Mixed Paradigm Setting.,” in *AAAI*, 1990, pp. 840–847.

[28] G. H. John, “Robust Decision Trees: Removing Outliers from Databases.,” in *KDD*, 1995, vol. 95, pp. 174–179.

[29] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, “A gentle introduction to imputation of missing values,” *J Clin Epidemiol*, vol. 59, no. 10, pp. 1087–1091, 2006.

[30] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of performance of data imputation methods for numeric dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[31] M. Oskarsdottir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens and J. Vanthienen. “Social network analytics for churn prediction in telco: Model building, evaluation and network architecture”, *Expert Systems with Applications 85*, 204-220, 2017.

[32] B. Ngonmang, E. Viennet, and M. Tchuente, “Churn prediction in a real online social network using local community analysis”. *In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2012), pp. 282–288, 2012.

[33] S. H. Park, S. Y. Huh, W. Oh, and S. P. Han, “A social network-based inference model for validating customer profile data”. *MIS Quarterly,* 36, 1217–1237, 2012.

[34] D. Van den Poel and B. Lariviere, “Customer attrition analysis for financial services using proportional hazard models”, *European journal of operational research,* 157, 196–217, 2004

[35] M. Lin, R. P. Nagpurnanand, and S Viswanathan. “Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending.” *Management Science* 59 (1): 17–35, 2013.

[36] S. Freedman and G. Z. Jin, “Learning by Doing with Asymmetric Information: Evidence from Prosper.com” NBER Working Paper #16855, 2010.

[37] G. Z. Jin and A. Kato, “Dividing Online and Offline: A Case Study” Review of Economic Studies 74(3): 981-1004, 2007.

[38] S. De Cnudde, J. Moeyersoms, M. Stankova, E. Tobback, V. Javaly, D. Martens, “What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance”, *J. Oper. Res. Soc.* 2019, *70*, 353–363.

[39] Y. Wei, P. Yildirim, C. Van den Bulte, and C. Dellarocas, “Credit scoring with social network data” *Market. Sci.*, *35*, 234–258, 2015.

[40] D. M. Sithigh, M. Siems. The Chinese Social Credit System: A Model for Other Countries? The Modern Law Review Vol. 82 (6), pp. 1034-1071, 2019.

[41] S. Hill, F. Provost, and C. Volinsky, “Network-Based Marketing: Identifying Likely Adopters via Consumer Networks” *Stat. Sci.* vol. *21*, 256–276, 2006.

[42] A. Shema. Effective credit scoring using limited mobile phone data. Proceedings of the Tenth International Conference, 2019.

[43] R. Emekter, Y. Tu, B. Jirasakuldech, M. Lu, “Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending”, *Appl. Econ.*, *47*, 54–70, 2015.

[44] V. D. Blondel, A. Decuyper, and G. Krings, “A survey of results on mobile phone datasets analysis”, *EPJ Data Sci.* 4, 10, 2015.

[45] G. Song, W. Bernasco, L. Liu, L. Xiao, S. Zhou, W. Liao, “Crime feeds on legal activities: Daily mobility flows help to explain thieves’ target location choices”, J. *Quant. Criminol.*, 35, pp. 831–854, 2019.

[46] W. D. Lee, M. S. Haleem, M. Ellison, and J. Bannister, “The influence of intra-daily activities and settings upon weekday violent crime in public spaces in Manchester, UK”, *Eur. J. Crim. Policy Res.*, 27, pp. 375–395, 2020.

[47] K. H. Grantz, H. R. Meredith, D. A. Cummings, C. J. E. Metcalf, B. T. Grenfell, J. R. Giles, S. Mehta, S. Solomon, A. Labrique, N. Kishore, et al, “The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology”, *Nat. Commun*., 11, 4961, 2020.

[48] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner, “Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study”, *Lancet Infect. Dis.*, 20, 1247–1254, 2020.

[49] B. Dewulf, T. Neutens, W. Lefebvre, G. Seynaeve, C. Vanpoucke, C. Beckx, and N. Van de Weghe, “Dynamic assessment of exposure to air pollution using mobile phone data”, *Int. J. Health Geogr.*, 15, 14, 2016.

[50] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, and M. C. González, “Sequences of purchases in credit card data reveal lifestyles in urban populations”, *Nat. Commun*., 9, 3330, 2018.

[51] Y. A. De Montjoye, L. Radaelli, V. K. Singh, “Unique in the shopping mall: On the reidentifiability of credit card metadata”, *Science*, *347*, 536–539, 2015.

[52] R. Borzekowski, E. Kiser, and S. Ahmed, “Consumers’ Use of Debit Cards: Patterns, Preferences, and Price Responses”, *Journal of Money, Credit, and Banking* 40: 149–172, 2008.

[53] J. Heckman, J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, Vol. 5, pp. 475–492, 1976.

[54] Y. S. Kim, S. Y. Sohn. "Managing loan customers using misclassification patterns of credit scoring model". Expert Systems with Applications Vol. 26 (4), pp. 567-573, 2004.

[55] P. Lynnette and Y. Cecilia, "Consumer Credit Assessments in the Age of Big Data", *Big Data in Finance*, 10.1007/978-3-031-12240-8, pp. 95-113, 2022.

[56] Y. Xia, Y. Li, L. He, Y. Xu, Y. Meng, "Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending", *Electronic Commerce Research and Applications*, 49, (101095), 2021.

[57] T. Gutierrez, G. Krings, and V. D. Blondel, "Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets", 2013.

[58] V. Frias-Martinez, C. Soguero-Ruiz C, E. Frias-Martinez, and M. Josephidou, "Forecasting socioeconomic trends with cell phone records", In: *Proceedings of the 3rd ACM symposium on computing for development.* ACM, New York, article no 15, 2013.

[59] S.Y. Sohn, D.H. Kim, and J.H. Yoon, "Technology credit scoring model with fuzzy logistic regression*", Appl Soft Comput J*, 43 (2016), pp. 150-158, 2016.

[60] A. Markov, Z. Seleznyova and V. Lapshin, "Credit scoring methods: Latest trends and points to consider". *The Journal of Finanace and Data Science*, Vol. 8, pp. 180-201, 2022.

[61] J.P. Onnela, S. Arbesman, M.C. González, A.L. Barabási, and N.A. Christakis, "Geographic constraints on social network groups", PLoS one, 6(4):e16939, 2011.

[62] A. F. Siegel, *Statistics and data analysis: an introduction*. Wiley, 1988.

[63] V. Barnett and T. Lewis, "Outliers in statistical data," *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*, 1984.

[64] D. B. Skalak and E. L. Rissland, "Inductive Learning in a Mixed Paradigm Setting.," in *AAAI*, 1990, pp. 840–847.

[65] G. H. John, "Robust Decision Trees: Removing Outliers from Databases.," in *KDD*, 1995, vol. 95, pp. 174–179.

[66] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "A gentle introduction to imputation of missing values," *J Clin Epidemiol*, vol. 59, no. 10, pp. 1087–1091, 2006.

[67] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of performance of data imputation methods for numeric dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[68] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art.,” *Psychol Methods*, vol. 7, no. 2, p. 147, 2002.

[69] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. “ O’Reilly Media, Inc.,”, 2019.

[70] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[71] L. Breiman, “Bagging predictors,” *Mach Learn*, vol. 24, no. 2, pp. 123–140, 1996.

[72] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.,

[73] D. W. Zimmerman, “A note on preliminary tests of equality of variances,” *British Journal of Mathematical and Statistical Psychology*, vol. 57, no. 1, pp. 173–181, 2004.

[74] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[75] Law E, “Impyute,” 2017. https://impyute.readthedocs.io/en/master/

[76] K. Kirasich, T. Smith, and B. Sadler, “Random forest vs logistic regression: binary classification for heterogeneous datasets,” *SMU Data Science Review*, vol. 1, no. 3, p. 9, 2018.

[77] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[78] V. Vapnik, *The nature of statistical learning theory*, Springer science & business media, 1999.

[79] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995.

[80] T. Bozkan, T. Çakar, A. Sayar, and S. Ertuğrul, “Customer Segmentation and Churn Prediction via Customer Metrics,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 2022, pp. 1–4.

[81] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann Stat*, pp. 1189–1232, 2001.

[82] A. Ercan, B. Karan, and T. Çakar, “Köpek Gezdirici Segmentasyonu / Dog Walker Segmentation,” in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 2022, pp. 1–4.