FRAUD DETECTION AND PREDICTION WITH MACHINE LEARNING APPLICATIONS

ALPEREN SAYAR

MEF UNIVERSITY

AUGUST 2023

MEF UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MASTER'S IN INFORMATION TECHNOLOGIES

M. Sc. THESIS

FRAUD DETECTION AND PREDICTION WITH MACHINE LEARNING APPLICATIONS

Alperen SAYAR

ORCID NO: 0000-0001-6089-2547

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

AUGUST 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 as a graduation project Master's Degree in Information Technologies.

Name and Surname: Alperen SAYAR

Signature:

ABSTRACT

FRAUD DETECTION AND PREDICTION WITH MACHINE LEARNING APPLICATIONS

Alperen SAYAR

M.Sc. in Information Technologies

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

AUGUST 2023, 36 Pages

The main purpose of this study is to determine the fraudulent activities on transactions 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 transaction and connection data of the customers, and then to perform predictive models for the target. A hit rate of around 79% was achieved in XGBoost and CATBoost models, which are classification model algorithms. In this way, it is aimed to directly detect fraudulent activities on a trasnaction basis by acting in a more effective, efficient and correct approach after detecting the customer that shows high potential to make fraud.

Keywords: Machine learning, transaction, feature extraction, feature selection, fraud, graph, redis.

Numeric Code of the Field: 92404

ÖZET

YAPAY ÖĞRENME YÖNTEMLERİ İLE DOLANDIRICILIK TAHMİNİ VE TESPİTİ

Alperen SAYAR

Bilişim Teknolojileri Tezli Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Tuna ÇAKAR

Ağustos 2023, 36 sayfa

Bu çalışmanın temel amacı, faktoring sektöründe faaliyet gösteren bir şirketin müşterilerinin işlemleri üzerindeki dolandırıcılık faaliyetlerini tespit etmek ve buna bağlı olarak müşterilerin geçmiş işlem ve bağlantı verilerine dayalı keşifsel veri analizi ile ölçülebilir parametreler yakalamaktır. ve ardından hedef için tahmine dayalı modeller gerçekleştirmek. Sınıflandırma modeli algoritmaları olan XGBoost ve CATBoost modellerinde %79 civarında isabet oranı elde edilmiştir. Bu sayede dolandırıcılık yapma potansiyeli yüksek müşteri tespit edildikten sonra daha etkin, verimli ve doğru bir yaklaşımla hareket edilerek işlem bazında dolandırıcılık faaliyetlerinin doğrudan tespit edilmesi amaçlanmaktadır.

Anahtar Kelimeler: Makine öğrenimi, işlem, özellik çıkarma, özellik seçimi, dolandırıcılık.

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

LIST OF FIGURES

LIST OF TABLES

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 using of data science methods has expanded, especially in finance. The main reason for this is that the benefits of data-based decision making ways is particularly works very well. They did a study on how it affects anomaly detection performance. They rated firms according to how strongly they used data when making decisions on customer suspicious and fraudulent activities. Statistically, the more data-driven a firm is, the more productive it is. As can be seen, the impact of data-driven decision making is too applicable. The ultimate goal of data science is to improve anomaly detection with suitable data. In doing so, it uses scientific ways, calculations, algorithms and environments. There are some important considerations when making these applications. First of all, it is important to correctly detect the dataset to be used to solve the problem. Then you need to determine the right model and structure is suitable for the data.

Many businesses use data science teams to gain a competitive edge. Data science can benefit businesses in a variety of ways. As Hurricane Frances approached the US state of Florida, executives at Wal-Mart Stores started to work to turn the situation in their advantage. This is an example of how a retailer and a telecoms firm employ data science. The company's information chief retrieved information on how sales had altered since Hurricane Charley struck the region a few weeks ago. We noticed a rise in the sales of a few products, noting if this increase occurred nationwide or just in the states affected by hurricanes. According to the analysis, it was certain that some products would see up to a 7-fold rise in sales in the case of a hurricane and that these products would be stockpiled. Making decisions based on data allowed the company to enhance its sales rate in this way.

According to research, MegaTelCo, one of the largest telecommunications providers in the USA, is losing 20% of its clients whose contracts have expired. Finding new clients is extremely challenging in the saturated market that is telecommunications today. A business must invest in incentives to draw clients, and when those clients leave, it loses money. Additionally, it is considerably more expensive to acquire new consumers than it is to keep an existing one. The reason why

businesses desire to keep their outdated Customers should be considered while allocating marketing funds. Data scientists must choose which client to keep in this situation. The data science team selects this model after configuring the issues and creating a model. These two instances show two distinct categories of data science judgments. In the first illustration, conclusions were drawn from the data, and choices were chosen by looking at recurrent occurrences. In the second scenario, tens of millions of data points were evaluated at various scales, and people who met the required criteria were chosen. We can possibly gain enormous benefits by applying this skill to millions of clients in the population, the more we can enhance our technique of determining how beneficial it will be to focus on a certain customer.

Each industry has implemented automatic decision-making at a distinct rate and with its own characteristics. Large-scale systems have been put in place by banks to manage data-driven fraud control decisions. Online advertisements, casino rewards programs, automatic Netflix and Amazon recommendations, and online advertising are more examples. We have so far discussed the usefulness of data science for making decisions. As a result, the data science team and responsible business executives need to be in frequent contact. Businesses that lack a basic understanding of what data scientists do are significantly at a disadvantage. Losses for businesses can result from wasting time and money on projects or, even worse, from making poor decisions.

2. Related Work

Today, information is created by processing and transforming massive amounts of data from a variety of sources. Based on this information, it is possible to identify fraudulent activity. It varies from expert systems in that it is data-based and looks to be a more objective tool. However, numerous ways are feasible employing big data techniques, not just for decision-making processes in current systems but also for the creation of new service and business models. Therefore, a significant paradigm and perspective shift has been brought about by the newly introduced data-based method. Now, methods based on machine learning and artificial intelligence, in addition to big data analytics, take center stage. Different techniques are highlighted in this context when it comes to the calculations and evaluations conducted in B2B fraud detection systems. A multi-axis strategy is necessary for a service structure that will cover B2B procedures because it will cover a wide range of topics. This

transdisciplinary approach within the big data framework is undergoing this change process. In addition to the research framework and methodologies, analytics has brought about a number of advances and opportunities in practice. From a different angle, the main objective of B2B processes, within the framework of artificial intelligence technologies, big data analytics, data mining, and data science, is to determine more appropriate fraud detection strategies and to make more accurate decisions by providing more suitable suspicious transactions for the financial sector. This kind of data-driven strategy can also bring to the fore certain social and ethical aspects as well as many perspectives that haven't yet been put on the agenda. Inside the parameters of this thesis.

An attempt is being made by a financing corporation to have, on average, over a thousand distinct businesses cash checks each month. This amount exceeds or equals the 400 thousand mark in the annual balance statement. Finding high-potential clients, particularly those with whom the company is initially in contact, is one of the most important concerns for the business. In spite of gaining clearance for various reasons, one-third of potential clients who first contact this business do nothing. Price and service rank among the most crucial elements, according to field research, even though there may be other factors such as duration, service, and price that contribute to this decision. In the context of this thesis, those with high customer potential are identified among the applicants who are contacted for the first time using data-driven modeling techniques, and various marketing strategies are used, such as making a more appropriate offer (offering a price reduction), conducting a specific search for the relevant customer candidates, and placing this call by a more knowledgeable customer representative. Aim to boost transaction rates, particularly for this target market. As a result, by improving the rate of application acceptance, the company is able to directly contribute to its short-term profitability as well as boost its chances of doing business with customers that have significant medium-term potential. Long-term profitability will rise as a result.

3. Problem Definition

The creation of a data-driven model is a strategy that will enable the identification of clients with a higher potential for financial fraud by looking at and evaluating the transaction records, customer traits, and behavior within the context of

this thesis project. As a result, the system will learn the broad traits of suspicious clients from the records that are now available. Therefore, it is anticipated that the project's scope will allow for the necessary changes, which will directly affect the company's annual profitability. Data science, big data analytics, and machine learning are the major methodologies and technologies that will be used in this R&D project.

4. Analysis of the Problem

Every month, on average, over a thousand different businesses try to defraud the financial company where the project is being carried out. The annual reports place this number at roughly 4,000. In spite of gaining clearance for various reasons, one-third of potential clients who first contact this business do nothing. Field studies have shown that pricing and service are among the criteria, even if there may be other elements such as transaction type, service, and duration that contribute to this preference. To determine the transaction rate specifically in this target audience, different detection strategies are used within the context of this study, such as identifying high customer potential for fraud who are in contact for the first time with data-driven modeling approaches and presenting a more suitable action for preventing this, making a special call to the relevant customer candidates, and making this call by a more senior customer representative. Specifically intended to grow. Therefore, boosting the percentage of application acceptance will both directly increase the company's short-term profitability and increase the likelihood that doing business with customers who have a high potential for fraud or other questionable activity will be avoided.

5. Thesis Contribution

Although they have received approval from prospective customers who have not yet taken any action, only 47.2% of the customers who have made transactions in the past, or those who we refer to as being "old," actually convert their approved applications into transactions. The reasons given by customer prospects for choosing not to conduct business with the organization include pricing, time, service, system issues, choosing to do business with another factoring company, and other factors. Therefore, the group that the business should focus on most is the potential customers who have already contacted it but have not yet taken any action. Even in the realistic

scenario, where the potential customers from this contact end up becoming customers, there is a net profit potential of more than 5,000 TL per day and an increase in net profit of roughly 1 million TL in the annual balance sheet. This is based on the average of the optimistic and pessimistic scenarios. When especially taken into account for the factoring industry, achieving this growth in net profit correlates to significant amounts. Consequently, focusing on identifying suspicious activity and stopping it will help the organization identify fraud for the first time. This is one of the techniques the company can take to boost its yearly turnover and profit margin. As a result, it is anticipated that this project's successful completion will result in a quantifiable outcome.

6. Aim of this Project

Which customer traits will be employed and assessed to provide the most accurate and effective outcomes is the main issue for the categorization of customers, as it is frequently emphasized in the relevant academic literature. Therefore, any client information that will be assessed and utilized as the foundation for factoring will be investigated as part of this research. The goal of this research is to identify persons who have a high likelihood of engaging in fraudulent activity and determine what traits they possess, what customer profile traits they exhibit, and how these elements might be quantified for consumers who pose little risk. This project will be able to use the "Right action for the right person" technique [6] when these categories are successfully completed.

7. The Outcome of this Thesis

As a result of this study, an estimation model will be created that, based on the first touch features of the model to be developed, will make an estimation about the fraud potential of the customer, especially in the context of high-danger customers (within the bounds of transaction type and frequency). In the second step, this model is made adaptive, and the required data is sent to the pertinent units along with the automatic processing and completion of these procedures and the new data recording and transaction process. The major goal of this R&D project is to design a model that will be used to boost the number and rate of transactions for clients who are 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. Consequently, the recently adopted data-based strategy has resulted in a significant paradigm and perception shift [1]. Now, approaches based on machine learning and artificial intelligence, in addition to big data analytics, take center stage. Different techniques are highlighted in this context when it comes to the calculations and assessments conducted in B2B Fraud detection systems. A multi-axis strategy is necessary for a service structure that will serve B2B procedures since it will cover a wide range of topics [1]. In addition to the study framework and methodologies, this fraud process, which is occurring from a multidisciplinary perspective within the context of big data analytics, has brought forth several advances and opportunities in practice [2]. From a different angle, the main objective of B2B processes, within the framework of artificial intelligence technologies, big data analytics, data mining, and data science, is to determine more appropriate fraud detecting strategies and to make more accurate decisions by providing businesses with more suitable fraudulent processes [2].

When a loan transaction is accepted and completed, the factoring company will transfer the funds into the customer's account after the client applies to them for financing. The factoring business takes its repayment from the drawer's account on the due date. Three steps make up the transaction process. The customer and the drawer (buyer of the deal) have transacted in Step 1. Instead of paying the customer (seller) in cash for this transaction, the drawer instead writes a check with a future maturity date. Think about a check that is refundable. Some customers started engaging in questionable behavior in order to get their money back, so-called fraud, to maintain

the company's reputation. Step 2 involves the customer (the seller of the deal) presenting this check to a factoring business and asking for cash in return for a certain interest deduction. When the check matures in Step 3, the factoring business will deduct the payback from the drawer's account rather than the client's account. The strength or financial standing of the drawer on the due date, as well as their ability to guard against fraud, determine the factoring company's ability to collect the payback. Therefore, the factoring firm should assess the applicant's financial and behavioral characteristics at the time of application in order to identify the state of customer transactions and stop fraud before it connects. Because the drawer, who made the purchase on the first day of business and not the customer to whom the factoring firm made the payment on that day, must have adequate cash in his or her account when the check is due. In order to protect itself, the factoring firm must have a system that looks at the client and the drawer in the transaction.

Another essential element is the connection between a customer looking for a bank loan and the bank. There are just two parties involved in this transaction: the client and the bank. The customer is required to repay the bank on the loan's due date (maturity or installment due date) before the due date customer for paying debts. The bank sends cash to the client's account on the day the loan will be utilized. Customers are allowed to pay off their obligations in questionable and illegal methods. The information provided by the client at the time of application or the customer's interactions with the bank over the course of the previous period are the most crucial factors to consider when evaluating a bank. In a factoring transaction, there are additional differences between the client and drawer profiles. Between January 2021 and August 2022, 65.514 consumers requested a greater bank limit from the factoring business. 35% of these consumers had bank limits at the time of application that are less than 40,000 TL, a little amount. There are 124,507 drawers with comparable programs. Compared to customers, just 5% of buyers have a bank limit of 40.000 TL, while 47% have one of one million TL or more.

Limit utilization among customers is higher than that of drawers. Demonstrating that clients are more successful in using their financial constraints and have a higher demand for financing. Factoring is a product that is substantially more expensive than bank loans. The client's capacity to obtain a bank loan will determine

this. The factoring company acquired 26% of its total 15,291 clients during the previous 12 months (with whom it did transactions for the first time in 2021). Banks are therefore unable to evaluate the risk associated with these consumers. Due of this, banks refuse to extend loans to some customers. As of the date of the bank loan application, 39% of these consumers had no bank loan limit. As a result, they use factoring financing, a far more expensive good, to manage their cash flows. In the factoring sector, drawers should be segregated in addition to clients given the discrepancy in factoring bank operations and the fact that the customer and drawers of a factoring transaction have different characteristics.

1.2. Financial Fraud Managements

The issue of fraud has become a top concern for many firms, particularly in the financial sector, yet the traditional methods for spotting fraud are no longer reliable. To detect complex fraud, new techniques are needed; RedisGraph, a high-performance graph database, could be able to assist. With the use of neural networks, RedisGraph can accurately and successfully detect fraudulent transactions in extensive and complicated scenarios. Businesses often combine Python and Oracle Databases, which offer robust data management and on-demand AI computing, to construct systems for identifying fraud. These technologies may be used to develop fraud detection systems that can identify fraudulent activities in real time. But today's fraud techniques have advanced, making the use of only these tools ineffective. In this study, a proof of concept for using RedisGraph-powered neural networks to identify financial fraud is shown. It highlights the need of properly combining Python with Oracle Database to build and implement real-time systems capable of effectively identifying fraudulent actions.

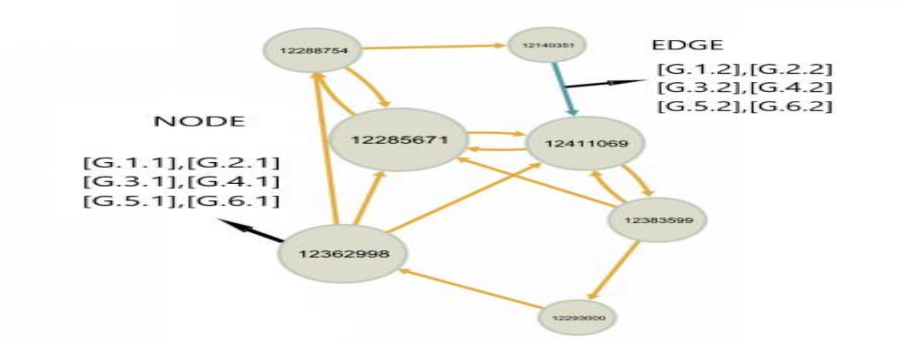

Figure 1.1: Basic Edge-Node Structure

1.3. Financial Fraud Applications

Due to the rise in digital transactions and the evolving tactics used by fraudsters, detecting and combating fraud has become a significant problem for organizations of all kinds. It's possible that conventional fraud detection techniques are no longer sufficient. However, developments in artificial intelligence and machine learning have created new opportunities for identifying and stopping fraud [3]. Combining neural networks with RedisGraph, a high-performance in-memory graph database, is a fresh and potent method that may be applied. The main goal is to show how RedisGraph, neural networks, and other machine learning algorithms can work well together to produce a useful and cutting-edge fraud detection solution. We will also look at how these methods are used in modern firms and organizational procedures.

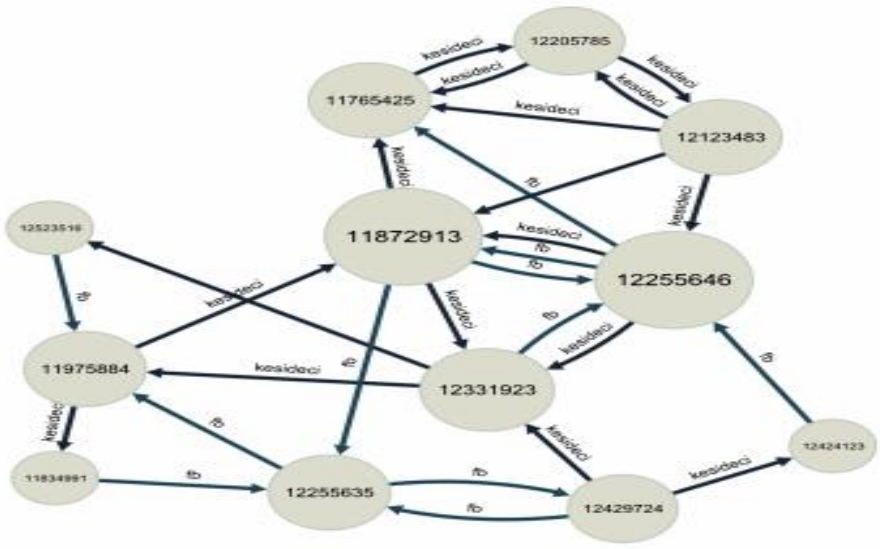

Figure 1.2: 12 Connection Possible Fraudulent Structure

1.4. Data Collection Phase

Many analytical modeling projects start with a flat dataset, create a predictive model for an interest goal measure (such churn, fraud, or default), and then evaluate the model on a different out-of-sample dataset. The data are typically implicitly believed to be independent and distributed evenly. Recent studies have questioned this assumption and looked at how people use the many social networks that. Links between them may have an impact [3]. Numerous social behavior patterns may be

seen. Social influence is the term used to describe when a person's contacts with other people have an effect on how they behave [3], [3]. Some social behavior may also be caused by other confounding variables (external, for instance) [4]. Network learning tries to incorporate social behavior patterns into prediction models so that the benefits of coordinated client behaviors may be utilized efficiently [5]. The network, which comprises of nodes and edges, is a crucial input for any social network learning activity. In certain situations, the idea behind these networks is rather straightforward. Consider the example of telecom fraud prediction. It is obvious that the network might be constructed using the data from the CDR. Significant social network effects were shown in an earlier study to help predict telecom churn [6]. Another example is the detection of credit card fraud, which is accomplished by connecting merchants and credit cards to form a network. In this setting, significant social network effects have also been seen [7]. There is strong agreement among credit scoring researchers and practitioners that credit ratings and borrower default behavior are correlated [7].

It has been established that this interdependency has a considerable negative impact on small and medium-sized firms [7]. One of the main challenges to understanding network effects or default propagation in credit ratings is the definition of the network itself. First attempts have been made to build client networks in online peer-to-peer lending. For example, Lin et al. [8] shown how contacts made online with people who don't default can improve credit ratings. Freedman and Jin [8] also issued a warning that internet connections alone would not give reliable information on credit-worthiness and might potentially be changed [8] to support their results. Using information obtained from social media networks and collected from Facebook profiles, For microfinance, De Cnudde et al. [9] developed credit rating models. According to their research, explicit networks of friends who interact are more predictive than explicit networks of friends who do not, even though implicit networks of people with similar behavior are superior to both explicit friendship networks. Social networks are already used in the commercial world to assess creditworthiness.

Not Only SQL (NoSQL) and Structured Query Language (SQL) are two categories of database systems with different uses. While NoSQL databases handle unstructured or semi-structured data in a flexible manner, SQL databases store and manage data in relational tables with set restrictions. While NoSQL databases perform

better in applications needing speed and adaptability, such big data processing and real-time data processing, SQL databases are better suited for complicated queries and massive amounts of specialized data. For instance, Table 1 compares the performance of NoSQL and SQL databases using the results of a test with 155,000 nodes and 126,000 edges as characteristics.

1.5. Algorithmic Approaches to Fraud Detection

Fraud detection is viewed as a classification challenge using various algorithms for analytical purposes [10]. Support vector machines (SVMs), artificial neural networks (ANNs), and extreme learning machines (ELMs), among others, are commonly compared to logistic regression (LR) as a benchmark [10]. The usage of ensemble classifiers, which integrate many classifiers for better performance, has increased in recent studies [11]. Beyond accuracy, other factors in classifier selection include complexity and the cost of misclassification [12], which have an impact on the application and usage of fraud-detection algorithms. Additionally, there has been an upsurge in interest in recent research on techniques for preventing fraudulent activities through feature selection or activity detection [12], [12].

In general, standard or alternative data formats are used for fraud detection. Two categories of information, such as demographic data and financial history, such as loan inquiries, are frequently used to estimate the likelihood of fraud. The unavailability of these types of information for people who are financially excluded has led to an increase in the use of alternative data in fraud detection. Many companies with operations in developing countries have started to provide digital credit services in recent years, pre-screening potential borrowers based on data from their mobile phone usage [13]. A sort of digital credit [13] built a fraud detection algorithm with the goal of selecting customers to migrate from prepaid to postpaid mobile phone subscriptions. 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 control-bureau (Buyers for whom credit bureaus have scant data). The area under the receiver operating curve (AUC) (a measure of fraud predisposition) demonstrates that when financial history (including bank account and credit card activity) and mobile phone usage data were combined, a significant rise in fraud susceptibility was seen [14]. (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 fraud predisposition, as shown by the area under the receiver operating curve (AUC) [14].

It was recommended that alternative score factors for those who are financially disadvantaged be integrated with data on mobile application usage for credit evaluation [15]. Mobile applications provide another source of alternative data, according to who utilized information from mobile financial activities to create a credit-evaluation approach for those without bank accounts [15]. The associated research findings [15] recommended employing a mobile application to acquire information from social media for credit scoring in order to simplify data gathering for financial firms. In order to improve the accuracy of fraud detection systems, the researchers [16] proposed a method for recursively adding client network data. It was recommended that alternative score factors for those who are financially disadvantaged be integrated with data on mobile application usage for credit evaluation [16]. Mobile applications provide another source of alternative data, according to who utilized information from mobile financial activities to create a credit-evaluation approach for those without bank accounts [17]. The associated research findings [18] recommended employing a mobile application to acquire information from social media for credit scoring in order to simplify data gathering for financial firms. In order to improve the accuracy of fraud detection systems, the researchers [19] proposed a method for recursively adding client network data.

1.6. Applications of Machine Learning Methods

The mining of predicted data is the most major application of machine learning [20]. Machine learning algorithms use a set of attributes to make predictions based on a dataset. These qualities might be continuous, categorical, or binary, depending on the particular problem and data. There are generally two types of learning: supervised learning and unsupervised learning. if every has a label This process is known as supervised learning and involves a dataset instance. The goal of supervised learning is to locate newly discovered unlabeled data. This is achieved by using labeled data during training to figure out the description of classes, which is subsequently utilized to label newly discovered data [20]. Classification problems and regression problems

are the two subdivisions of supervised learning. By using a labeled training set, classification aims to predict labels from unlabeled data. The methods KNN, decision trees, random forests, and support vector machines are frequently used to solve classification problems. On the other hand, regression is utilized to understand the relationship between the dependent and independent variables. These three regression algorithms—linear regression, logistic regression, and polynomial regression—are the most basic ones. A few of the assessment metrics used to analyze the effectiveness of supervised learning systems are accuracy, f1-score, receiver operating characteristics (ROC), and confusion matrix. The ROC curve evaluates a model's propensity to forecast for a binary classification job based on two criteria.

Without labels, unsupervised learning makes predictions. Unsupervised learning is dominated by clustering, anomaly detection, and density estimation [21]. Unlabeled data are used in clustering to classify patterns. The labels used are data-driven [65]. Applications for clustering include recommendation engines, image segmentation, and dimensionality reduction. These problems are frequently addressed using K-means, DBSCAN, agglomerative clustering, and affinity propagation [21]. Anomalies can also be found using DBSCAN and K-means. Outliers may also be found using GMM, minimal covariance determinant (fast-MCD), isolation forest, local outlier factor (LOF), and one-class SVM [21]. Density estimates may also come from unsupervised learning. For data visualization and analysis, it calculates the probability density function of the random process that generated the dataset. Density is estimated via GMM and DBSCAN [68]. These characteristics need to be understood in order to build a better algorithm that integrates many algorithms to enhance one another. In other cases, it could be challenging to pinpoint a particular algorithm that offers the highest accuracy or another type of score criteria. Ensemble learning is what is used in these situations to combine two or more classifiers [22]. Popular ensemble strategies include voting, bootstrap aggregating (bagging), pasting, boosting, and stacking. Single classifiers may not perform as well as voting method classifiers. The classifier diversity used in this ensemble technique increases accuracy by producing a variety of errors. The bagging strategy combines numerous subsets of the original dataset to train predictions with a single algorithm. Sampling with replacement is necessary for bagging. To sample without replacing is to paste [22]. Increasing ensemble learning techniques teach predictions to gradually correct one another. The

most common ones are Adaptive, Gradient, Extreme Gradient, LightGBM, and CatBoost [23]. Stacking algorithms (stacked generalization) train a model from a collection of algorithms rather than averaging classifier predictions. The final prediction is produced by a Meta learner or blender using the classifier outputs as training data [23].

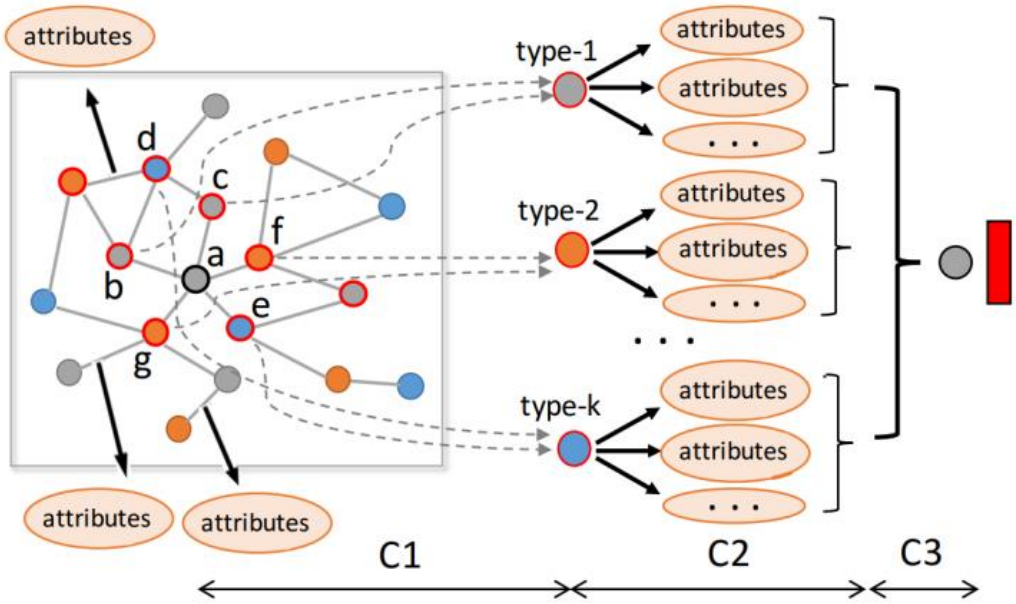

Figure 1.3: Feature Extraction

2. DATA PREPARATION AND PREPROCESSING

2.1. Data preparation and Data Preprocessing

This stage involves getting the data ready and processed so that it can be used with machine learning models. Depending on the issue at hand, it may include a number of sub-steps.

2.2. Data Collection Phase

The collection of pertinent dataset is the initial stage in solving a machine learning problem. It is crucial to obtain the most informative aspects while gathering it. It is possible by utilizing the domain-specific expertise of an expert. The only alternative, if an expert is not accessible, is to apply sheer force, making use of all features. This method's drawback is that it generates noise and missing values, which necessitates significant data cleaning and preprocessing [24]. The dataset at hand is typically rife with mistakes and noise. Real-world data must be updated since it is not always accurate. To prepare the dataset for usage in algorithms, a hierarchy of issues has been identified and has to be addressed. First thing to look at is the presence of impossible values inputted in features [25]. We recognize a value as being impossible, for example, if the relevant characteristic is predicted to have binary values but one instance of it has a discrete value. The best time to solve them is during the data intake process, allowing for correction. However, if entering the right values is impractical, they can simply be considered as a missing value category and eliminated from the dataset [26]. The fact that no values have been entered into an instance of a feature [26] is the following issue to be examined. This problem can be solved in a number of ways, the majority of which will be covered in the sentences that follow. Finally, the dataset contains a few characteristics that are irrelevant [27]. They are just disregarded and excluded from the dataset.

2.3. Data Cleaning

One of the most crucial phases of data mining is data cleansing. The most important step in data preparation is identifying and cleaning up damaged data. Unrepaired filthy data would produce erroneous analytics and faulty models. Error detection and correction are the two steps that typically make up data cleaning. There

are quantitative and qualitative techniques to define mistakes that may be used for error identification. While qualitative strategies utilize rules, limits, and patterns to address data flaws, quantitative techniques use statistical tools to find outliers [28].

2.4. Outlier Detection

Evaluate improbable values next [29]. Values that significantly deviate from those in the sample are included in this category [30]. Observations that differ from the data are another way to describe outliers [30]. Data cleaning on a variable-by-variable basis is one method. The peculiar probability distribution of these outliers helps to identify them. Depending on the domain and distribution, they are larger than or equal to one standard deviation away from the mean. These numbers are often eliminated from the dataset because they might be the result of technical issues, alterations in system behavior, fraud, mistakes made by humans or instruments, or variations in the population [31]. For the third reason, it's feasible that values near a distribution's tails, where it is more scattered and volatile than previously believed [31], are correct. They might also reflect actual facts that are within one cluster but belong to another cluster or label [32]. To create a precise, accessible model, they typically also eliminated material.

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 [32] As a result, a bank may identify a potentially problematic customer at an early stage and take appropriate action to stop further loans from being granted or to terminate the customer's credit limit to prevent excessive credit utilization. A population study could identify tall outliers. It is therefore typical and may occur depending on the study. Fraudulent surveillance cameras are warned by outliers If the alert was correctly raised, outlier data may be saved somewhere else to improve detection. The typical class is taught in semi-

supervised detection or identification, and the model picks up on irregularities. Through the use of normal and pre-labeled data, this method creates a normality border. The model may be trained and improved with each new piece of data. All normal data must be provided for generalization. In contrast to type 2 data, anomalies are not required for training. This is advantageous for fault identification since it would be too expensive to train the model using anomalous data obtained by causing engine damage. It would be too expensive to train the model by damaging the engine for outlier identification in aircraft engines. New fraud types may be handled incorrectly by fraud detection systems By simulating normalcy, the model may detect fresh fraud before it reaches the usual range. The model may detect new fraud until it falls within the typical range by modeling normality.

2.5. Methods of Outlier Removal

The most efficient technique to deal with outliers is to display a feature that can identify negative values that occur frequently. This feature is a strong and useful tool. Methods for identifying outliers rely on two factors. In order to model the data distribution and identify outliers for a clustering, classification, or recognition model, first develop a technique. A decent topic should be chosen second. All distribution densities should be accommodated in neighborhood selection. The majority of outlier identification methods have the same roots but different names. Outlier, novelty, strangeness, noise, deviation, and exception mining are a few examples. Outliers are found using statistics, neural networks, and machine learning. To enhance their model, some algorithms pick one or more of these fields.

2.6. Statistical Methods for Outlier Removal

Statistical techniques were first used to detect outliers. Some of the earliest strategies are univariate, while others are only applicable to unidimensional data sets. Grubbs has a one-dimensional approach. To get a Z value, the difference between the mean value of the attribute and the requested value is divided by the standard deviation of all values. User-supplied parameters are no longer required in this procedure because of data-generated parameters. The model becomes more statistically representative as there are more data values [33]. The application of statistical models is constrained to quantitative real-valued and quantitative ordinal data sets. The

appropriateness grows Statistical techniques were first used to detect outliers. Some of the earliest strategies are univariate, while others are only applicable to unidimensional data sets. Grubbs has a one-dimensional approach. To get a Z value, the difference between the mean value of the attribute and the requested value is divided by the standard deviation of all values. User-supplied parameters are no longer required in this procedure because of data-generated parameters. The model becomes more statistically representative as there are more data values [34]. The application of statistical models is constrained to quantitative real-valued and quantitative ordinal data sets. If data modifications are large, appropriateness lengthens processing time [35]. Univariate and multivariate outliers may be found using informal box plots. Box plots display the median, upper quartile, lower extreme, and higher extreme [35]. They search for peculiar values in collections of categorical data and consider both symmetric and asymmetric distributions. Outliers can be observed visually. Effectiveness is improved by selecting upper and lower thresholds 1.5 times the interquartile range (IQR) from the upper and lower quartiles, which contain the lower and higher outliers. Univariate outlier elimination looks for outliers in each variable. 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. Many distance measures, like Euclidean distance, solely give location information, making them inappropriate for data collecting.

2.7. Machine Learning-based Outlier Removal Methods

Machine learning systems can detect categorical data outliers that are hidden from most statistical methodologies. The C4.5 decision tree is used to find categorical outliers because decision trees, which do not require data distribution or features, may find outliers more quickly than statistical approaches. Decision tree approaches are less flexible and gradual than rule-based systems, which may be modified or changed to discover outliers. Characteristics will be assessed following data analysis. Here, the emphasis will be on the variables' missing data. Data missing for two reasons. Loss of data without recovery. Machine learning is hampered by missing data. Thus, knowing how to complete missing data is essential. A preliminary investigation will

be followed by a scenario analysis of missing data. These include the MCAR, MNAR, and MAR states, all of which have been supported by literature.

2.8. Manage Inadequate Data

Partial or missing values are present in most data sets. Missing values may be the result of omission, misplacement, non-applicability, or the data set designer's refusal to provide a value [36]. One must look into their unpredictable nature in order to come up with a decent strategy for handling irrelevant and lost values. Three randomness kinds (MAR) exist: missing entirely at random, missing not at random, and missing at random. MCAR is brought on by mistakes such subject measurement loss. Therefore, the likelihood of absence is unrelated to other qualities. This type of missing value may be handled in a variety of ways, unlike MNAR data, where missing values are linked with unobserved information about other subject qualities. Missing values in a data collection that cannot be handled by a general method may result from errors in experiment design [37]. MAR occurs when the result or other predictors depend on the missing data. A feature is considered to be MAR if missing values are present after a device breakdown but not during regular use.

There are several approaches to fix problems with missing values and partial data. The experiment might be redone or resampled to address MNAR problems. MCAR and MAR scenarios assign missing values based on external variables or internal properties. Single and multiple imputation are used for missing data. In order to eliminate bias, data is imputed rather than deleted. The best course of action in MNAR data could be to delete incomplete occurrences [38]. Mean, distribution, regression, and KNN are examples of single imputation techniques. Based on the distribution of the data, missing values are replaced using mean, median, or mode imputation. The data distribution may be distorted by a sizable number of missing values, which is only one of the method's many shortcomings. Without altering their structure, distributions can be used to fill in blanks [39]. Regression approaches are more complex since they use imputation and additional variables. A linear connection between the MAR and MNAR data and the imputed feature is required for unbiased imputation [40].

K neighbor distance is used in KNN single imputation to impute comparable values. The computing cost of this approach increases. Imputed data sets are averaged by several imputations. This lessens single-imputation biases in imputed data that is uncertain. To lessen uncertainty and single-imputation mistakes, it runs several valid imputations and averages them. In this approach, a data set is input n times, assessed, and then combined [40]. The most common technique for dealing with missing data is multiple imputation by chained equations, or MICE. To choose the best approach, it makes advantage of the correlation between the traits that are imputed and others.

2.9. Modeling Framework

This section covers the steps that prepare the data for use in prediction models, including data cleaning, outlier identification and removal, imputation for missing values, data clustering, one-hot encoding, and standardization. It will be described how the modeling framework was developed. Due to the usage of five algorithms, three different feature extraction approaches, and a wrapper method in addition to feature selection, it would have been quite difficult and disorganized to compare all of their findings without the creation of a modeling framework. Consequently, six basic frameworks were established: the primary model without feature extraction, the model with PCA, the model with t-SNE and the model with Isomap, the model with frontal alpha asymmetry transformation, and the model with feature selection. All ML algorithms were applied to the main model before the best method was chosen to be applied to feature extraction models. The best model between feature extraction and main models was then picked to apply feature selection to further enhance score. The machine learning (ML) techniques that were applied to the primary model will be discussed in the next section along with information about their characteristics and hyper-parameter adjustment.

3. RESULTS

The outcomes of each section written in the data collection, model construction, and model validation procedures are presented in this chapter. The project's target outcomes were chosen within its parameters based on the requests and suggestions of the appropriate units. The following are the intended project deliverables in accordance with the set outputs: calculating the likelihood that clients who try fraud for the first time will be caught while making a transaction throughout the course of the next 3, 6, and 12 months.

3.1. Predicting Customers' Next Transaction times

The project process has been prepared in accordance with the aforementioned components in order to provide the desired outcomes at the conclusion of the project process. Analytical approach techniques were established within the parameters of the project, the outputs of which were established. Three primary activities were chosen during the analytical approach procedure. The structure and methods for the project's whole work were established. Then, choices were made on the technology to be employed. The company's internal Oracle database systems will be utilized as the database for technical tools. For the phases of data analysis, processing, and modeling, Python programming language was utilized. To accomplish the predetermined business objectives, meetings with the company's customer analytics and sales teams were necessary. The sessions led to the determination of data that may be utilized for the project in accordance with the results. All potentially useful data have been captured, together with the determined data, papers on other company-owned data based on customer information.

By creating SQL queries, the data identified and required in accordance with the project outcomes were extracted from an officially in use Oracle database of a factoring firm. Data from the risk center, including customer risk information for the data to be utilized and other financial information about clients, were collected from the Credit Registration Office (CRO). Information about the customers, including factoring grades, current debt status, the number of credit institutions, risk information, limit information, and credit information under follow-up, is included in the data from CRO. The dataset now includes data on customer risk, enquiries from consumers at

this business, number of transactions, total donations, etc. After the data analysis and as a consequence of the meetings with the customer analytics unit and sales units, more characteristics indicated as a part of business information will be added to the data set by employing attribute engineering, in addition to the information already included in the data set. The section on data interpretation and preparation goes into great detail about this data.

After being collected, the data were converted to CSV format and entered into the database. Due to the CSV format's huge storage requirements and sluggish reading and writing speeds, data files were later migrated from CSV to Parquet file types. The issue of data files taking up too much space, despite being stored in various file types, has been resolved by creating a direct connection between the database programming language and the database, as the data stored in various file types takes up a lot of space due to the updates made to the data and the resulting changes to the data set. A connection was made between the Python programming language and the Oracle database using the cx_Oracle module. The resources have been used more effectively while also achieving great efficiency thanks to this link in terms of the amount of space used up by file systems and the speed at which files are read.

3.2. Feature Selection and Engineering

Different techniques were utilized for both category and numerical values to fill in the data once the missing value analysis was finished. These filling techniques, which are univariate and multivariate, respectively [41]. Numerical data from univariate filling techniques were filled in using the average assignment, last value assignment, prior value assignment, and random value assignment methods. The most often occurring class values in the variable were assigned, and'missing' values were assigned to the missing values, in order to fill the categorical variables [42]. Multivariate filling techniques, such as the kNN filling method, Hot-Deck filling method, MICE (multivariate imputation chained equations), and MissForest-extends approaches, were also utilized in addition to these techniques [43]. The data were filled in using a variety of approaches for both category and numerical values once the missing value analysis was finished. These filling techniques, referred to as single and multiple variables, respectively [44]. The numerical data from univariate filling techniques were filled in using the average assignment, last value assignment, prior

value assignment, and random value assignment methods. The most frequent class values in the variable were assigned, and'missing' values were assigned to the missing values, in order to fill the categorical variables [45]. These techniques were employed in addition to multivariate filling techniques including the kNN filling method, Hot-Deck filling method, MICE (multivariate imputation chained equations), and MissForest-extends approaches [45].

Additionally, the cardinality status of categorical variables was examined. The number of categorical variable classes is represented by the cardinality. Multiple circumstances are impacted by cardinalities. First off, categorical variables are often not accepted by machine learning models, but they significantly complicate 'Tree' based algorithms' compatibility [45]. In addition to this, it is likely to result in several operational issues. The performance of the model has improved as a result of the analysis and resolution of the high cardinality scenario [46]. Rare classes within categorical variables were found, and as the distribution of categorical variables within classes is crucial, these classes will either be amalgamated and reported as one class or included in the class being studied. Encoding was accomplished as a consequence of the examined category variables. For the encode procedure, three alternative approaches were taken. Monotonous interactions, conventional approaches, and unconventional approaches. One-hot encoding, frequency encoding, and ordinal tag encoding were the conventional approaches. We'll employ WOE (Weight of Evidence), mean encoding, and monotonically ordered variable analysis. Alternative encoding techniques included Binary, Feature Hashing, and Rare Labels.

In order to examine the distributions of the variables, probabilistic distribution techniques were also applied. For both discrete and continuous variables, many techniques will be employed. We examined the Poisson and Binomial distributions for discrete variables. The skewness test and Gaussian distribution were applied to continuous data. The logarithmic transform, reciprocal transform, square root transform, exponential transform, box-cox transform, and yeo-johnson transform techniques were employed to standardize the variables as a result of the experiments conducted. Outliers were examined to bring the variables closer to normalcy after the data distribution analysis. Model performance was directly impacted by the analysis

and treatment of outliers. The LOF (local outlier factor) approach was then applied to all variables after performing quantitative analysis for outlier analysis.

3.3. Feature Scaling

In order to scale the data, the procedures of standardization, mean normalization, min-max scaling, resilient scaling, and scaling to unit length were applied. These techniques have been shown in the literature to provide good outcomes.

3.4. Model Development

The data set was split into three categories for the model creation phase: training, validation, and test set. Within three to six months of the model's training, a test set will be chosen from the data. Prior data will only be used for the training set. The outcome of the data set generated for modeling will be used to develop many models. Models based on machine learning will be employed. Regression, classification, and clustering models will all be used in accordance with the goals. Three-month, six-month, and twelve-month periods will be used for the study and estimation of fraud. Additionally, modeling will be done to predict whether or not the consumers would engage in their subsequent fraudulent acts. Both these estimates and the analysis will be done using classification models. Following that, the number of transactions and potential fraud acts that customers will commit over the course of the following three, six, and twelve months will be computed. For process estimates and contributions, regression models will be employed. In addition to the regression models that will be used, the models for estimating the number and volume of transactions and suspicious potential for the customers' next 3, 6 and 12 months will be changed to low, medium, and multi potential, and regression models were not used by converting them into classification problems.

For modeling, three distinct machine learning techniques were used. These include the techniques known as classical learning, ensemble learning, and artificial neural networks. Both supervised and unsupervised learning will be used to investigate traditional learning. "Bagging" and "Boosting" techniques will be used in group learning. The bagging approach involves selecting various subsets of the training data at random and training each classifier on those subsets. The process is asynchronous.

The training set is broken down into smaller portions, and the best model is chosen after being established with each one. Using this strategy will be more reasonable if the data set has bias and will result in a solution to the over-compatibility issue. It is better to utilize if there is substantial variation in the data set compared to the Bagging approach. The outcome of one classification serves as the input data for the other classification in the boosting procedure. It operates in synchrony. The model is generated separately for each of the subsets of data, and one model serves as an input for the other The modeling causes the data weights to shift and the weights of the wrongly discovered observations to be raised. Classification models included K-Nearest Neighbors (kNN), Logistic Regression, Decision Trees, Random Forests, CatBoost, LightGBM, XGBoost, and Neural Networks.

Since the modeling procedures must be run several times and the outcomes compared, these activities will be automated. A software that takes input from the user, trains it, and outputs the results as a report was created in order to automate it.

Because modeling costs rise along with data quantity, the model becomes more sophisticated and more prone to overfitting issues as the number of independent variables rises. For classification issues, the variable selection approach was applied in this. Additionally, three distinct approaches were used to choose the variables. These include filtering techniques, embedding techniques, and forward and backward iterative algorithms. The forward and backward iterative algorithms use addition or subtraction of all independent variables from the model to test every conceivable combination of independent variables in an effort to provide the best possible outcome.

By examining the correlation between the independent variable and the dependent variable, the filtering process employs statistical methodologies and applies feature selection. It can lower the overall model score even if it has a relatively low computational cost [47]. Filtering and iterative characteristics are combined in embedded approaches. The classification algorithms themselves make use of it. Regression techniques like Lasso, Ridge, and Elastic Net punish overfitting. In terms of price and pricey in terms of filtering, this approach is less expensive than iterative methods [48].

3.5. Model Hyper-parameter Optimization

The total number of parameter choices is fewer than the grid search approach since Random Search tries random hyper parameter combinations. Finding the ideal settings becomes challenging as a result, decreasing time complexity [48]. The automation of manually carried out hyper-parameter optimization is called grid search. The grid search method's temporal complexity increases when all parameter combinations are tested sequentially [48]. The grid search approach has been improved using the Halving Grid Search technique. The sequential division approach is used in the halve grid search method to browse through the expected parameters. By grouping all observations into tiny packages, it begins by evaluating each parameter on these sample sets, then iteratively chooses the optimal parameters by mixing small sample packages [49]. The threshold value setting technique will then be used. This approach aims to increase the utility of uneven data sets. The AUC-ROC curve stated above will be utilized in particular to support this issue. In all machine learning models, the threshold value is typically 0.5. The model score will not change if the threshold value is changed, but the score distribution between classes will be significantly impacted [49]. This threshold calculation step will be crucial, especially in light of the imbalanced datasets used to build the models, which include forecasting the transaction status of consumers contacting Tam Finans for the first time.

3.6. Model Evaluation and Selection

On the basis of each unique model, the evaluative criteria indicated above will be employed for model evaluation. The section with the automated object-based measurement metrics comparison model developed for the modeling part will be included to simplify the computation of these metrics due to the enormous number of measurement metrics. Automation of measurement metrics will be used to choose the optimal model. For the best models, hyperparameter tweaking was then carried out. By training or improving the models in accordance with certain coefficients when building the models, hyperparameter optimization offers a foundation for arriving at the best parameters. Particularly since several models will be developed in accordance with various aims, a wide range of hyper-parameters will manifest. At this stage, emphasis will be placed on the criteria that are often utilized in the literature. The

procedure of hyper parameter tweaking will involve three different approaches. These include the Random-Search, Grid-Search, and Halving Grid-Search techniques.

3.7. Model Deployment, Integration and Maintenance

Using version control systems, the produced model will be sent to the software unit. On active servers, the model will be made operational. Server-client architecture will be applied during this procedure. Sending the machine learning model to the created API. The API chosen will be Flask. The API communicates continuously with the database, retrieving model results and logging results from the machine learning model. The ready user will get the findings, which users and customer service personnel can utilize. On the basis of the customer code, the results that need to be reviewed will be carried out. Sending the client code to the live machine learning model will enable the desired outputs to be obtained as a consequence of the model. The deployment step will involve the formation of a control group. At the conclusion of the day, the data of the customers who make a new move as a result of their ongoing enquiries and transactions will be kept and combined. The query process must go without interruption since there are too many client moves throughout the day. All new movements will be added to the data set at the end of the day in order to save costs and server load, and the model will then be made operational once more, generating fresh results.

The customer service agents make advantage of the outputs from the feedback process. The instances that might be flawed will be investigated, and the revision cases will be taken into account, based on the outcomes of the models' outputs and the feedback provided by the client representatives. The workflow will then continue in the modeling, measurement, deployment, and feedback loop, with the measurement and deployment steps being repeated. The intended user group, CRMs, are really expected to provide ideas, comments, and suggestions at every level. However, at this point, these messages will be sent much more frequently and specifically, with the intention of enhancing the system's usefulness.

DISCUSSION AND CONCLUSION

Cross-validation was then carried out to guarantee that the model configuration would lead to accurate and trustworthy outputs. 25% of the data is used for testing and 75% is used for training in the cross-validation approach. The cross-validation value was set at 5, and stratified sampling was utilized because the data set is uneven. The outcomes of the model were not examined using the direct accuracy rate since the data set is unreliable. Complexity matrix, balanced accuracy, geometric mean, dominance, imbalance index, ROC-AUC curve, and Precision-Recall curve were employed for the investigation of model accuracies. The Precision-Sensitivity curve and the ROC-AUC curve are utilized in particular to determine the threshold value. The accuracy discrepancy between the classes is diminished by modifying the threshold value. After modeling measurements, the models that produced the best results underwent hyper-parameter tweaking (optimization) by halved grid search. Made for XGBoost, Logistic Regression, Random Forests, and Support Vector Machines.

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 77%, while the average F1-Score for commercial customers was above 73%. Since the score values of the dominant target variable are high, the threshold value for real companies was determined as 0.44 using the ROC-AUC curve. On the other hand, the threshold value of commercial companies was determined as 0.36. The study's models are presently in use inside the organization, and during the following three months, the findings will be verified and any required modifications will be made once again in accordance with the circumstances. This process will continue throughout the life cycle. Since the project's end goal may be identified as having a high potential for fraud, it can be argued that sensitivity ratings are more significant [50].

For sensitivity scores—how many of them were truly classified as very suspicious—accurate assessment was accomplished. Another is precision scores, which measure the proportion of leads that are actually high-suspicious against those that are anticipated to be [81]. varied strategies can be used, as seen by the varied ways these two scores were analyzed. In the context of this study, it can be said that sensitivity ratings become more important given that high-fraud potential clients are

frequently the intended target [50]. When viewed from the perspective of the corporation, the produced model has added value since it has been employed, and there is no active usage of a comparable model in the organization, thus it has a very unique effect. If the high potential clients who have recently initiated contact can be accurately and effectively identified, it is intended to drive the marketing team to more effective channels. As a result, it is predicted that within the parameters of this research, the required changes will be implemented, directly affecting the yearly profitability of the organization.

Cross-validation was then carried out to guarantee that the model configuration would produce accurate and trustworthy outcomes. For the cross-validation procedure, 25% of the data is used for testing and 70% for training. The cross-validation number was set at 5, and stratified sampling was utilized due to the imbalanced nature of the data set. The outcomes of the model were not examined using the direct accuracy rate since the data set is erratic. Complexity matrix, balanced accuracy, geometric mean, dominance, imbalance index, ROC-AUC curve, and Precision-Recall curve were utilized for the investigation of model accuracies. The threshold value is determined in particular by the ROC-AUC and Precision-Sensitivity curves. The accuracy gap between the classes is narrowed by changing the threshold value. Following modeling measurements, halve grid search was used to do hyper-parameter tweaking (optimization) for the models that produced the best results. Made for XGBoost, Random Forests, Logistic Regression, and Support Vector Machines.

For the dataset of individual consumers, Support Vector Machines (SVM) produced the best results, but the XGBoost classifier produced the best results for business clients. Real clients had an average F1-Score of 77%, whereas commercial customers had an average F1-Score of more than 73%. The threshold value for actual businesses was calculated using the ROC-AUC curve to be 0.46 due to the high score values of the dominating target variable. The threshold value for commercial businesses, however, was set at 0.36. The study's models are presently in use inside the organization, and during the following three months, the findings will be verified and any required modifications will be made once again in accordance with the circumstances. This process will continue throughout the life cycle. Since high-risk

fraudulent clients may be identified as the project's ultimate aim, it can be stated that sensitivity ratings are more significant.

Sensitivity ratings, or the percentage of people who were genuinely classified as high potential fraud, were accurately estimated. Another is precision scores, which measure the proportion of leads that are really anticipated to be high-potential fraud. diverse strategies can be used, as seen by the diverse approaches taken to these two scores. In the context of this study, it can be said that sensitivity ratings become more important given that high-risk clients are frequently the intended target. When viewed from the perspective of the corporation, the produced model has added value since it has been employed, and there is no active usage of a comparable model in the organization, thus it has a very unique. If the highly suspicious clients who have recently initiated contact can be accurately and effectively recognized, it is intended to steer the marketing team to more effective channels. As a result, it is predicted that within the parameters of this research, the required changes will be implemented, directly affecting the yearly profitability of the organization. It is planned to direct the marketing team to more efficient channels if the high suspicious 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 Unscaled, Standard Scaler, Min-Max Scaling, Max Abs, Robust Scaling, Yeo-Johnson, Gaussian-PDF, Uniform-PDF, and L2 normalization methods are the ones that were used. Four measurement variables were chosen to compare the cluster outcomes. These include the Elbow Method, Calinski Harabasz Index, Davies Bouldin Index, and Silhouette Score. The average

distance between a data point and all other data points in the cluster, as well as the average distance to other data points in the cluster that are closest to the given data point, is known as the silhouette score. The Silhouette Score should be high and ranges from -1.1 since it gauges how well a cluster fits into the data inside its own cluster in comparison to other clusters. To gauge the degree of average similarity between several categories, the Davies Bouldin Score is employed. It is more crucial to have a low score since Davies Bouldin gauges how similar distinct clusters are, and the lowest value is 0. To determine the ideal number of clusters inside a cluster, the Elbow approach is utilized. With this approach, random cluster number values are chosen, and each cluster number value is then used to apply K-Means. Each point (piece of data) in a cluster is measured with respect to its distance from the center of gravity, and the value at which the average distance decreases the quickest is chosen. When compared to other clusters, the Calinski-Harabasz approach evaluates how similar a set of data is to its own cluster. The distances between the data points in a cluster and the cluster center are used to assess the similarity in this case. The cluster centers' separation from the spherical center determines how it is divided. In contrast to the elbow technique, the best cluster to choose is the location of the quickest upward breakdown.

The clustering procedure has been automated using an object-based automated clustering paradigm. With this model, it is no longer necessary to repeat the process several times or create alternative models for various parameters. The clustering approach underlies the model's operation. The input parameters to the model include the model that you wish to create, the data collection that will be utilized, and its parameters. The raw data is scaled in 8 different ways with the purpose of obtaining the optimal scaling technique and the most clusters possible. Scaling the raw data first, then setting up the model with Min-Max Scaling, Max-Abs, and Robust Scaling, There were four methods used: Yeo-Johnson, Gaussian-PDF, L2 Normalization, and Uniform-PDF. The cluster model was then developed and trained using the scaled data. The score metrics Silhouette, Davies Bouldin, and Calinski Harabasz were then computed for the developed model. The data frame structure contains the calculated metric scores. The score metrics were then put up against one another. Following this comparison, the variation graphs were created by taking the metrics' differences.

Finally, graphs representing the resulting comparison and metric score discrepancies are produced.

The traditional machine learning techniques focus on correlation and pattern analysis, in conclusion. For causal analysis, it might not be enough. To develop estimates utilizing these causal effects, it is crucial to comprehend the relationships between the independent variables, as well as those between each independent variable and the dependent variable. We will also apply 'Causal Inference' approaches in addition to conventional machine learning techniques. We'll compare the outcomes of "Causal Inference" techniques versus conventional models. Additionally, the results of the 'Causal Inference' approaches will be used to create more robust and trustworthy machine learning models. 'Do Why' and 'EconML' libraries created by two independent Microsoft teams will be utilized for the 'Causal Inference' approach. The 'Causal ML' libraries created by Uber will also be utilized in addition to these libraries.

Table 1: Score Table

Number of Features	Metrics	Random Forrest	XGBoost	SVM
1500	Precision	0.62	0.7	0.75
	Recall	0.47	0.62	0.81
	F1-Score	0.51	0.65	0.78
2500	Precision	0.63	0.72	0.73
	Recall	0.47	0.66	0.87
	F1-Score	0.54	0.69	0.8
6000	Precision	0.66	0.71	0.77
	Recall	0.48	0.72	0.87
	F1-Score	0.54	0.68	0.8
12000	Precision	0.67	0.77	0.79
	Recall	0.55	0.80	0.89
	F1-Score	0.6	0.77	0.84

REFERENCES

[1] S. Kaheh, M. Ramirez, K. George, Using Concurrent fNIRS and EEG Measurements to Study Consumer's Preference, 2021

[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] C. Alexander, The Present and Future of Financial Risk Management **"***Journal of Financial Econometrics*, Volume 3, Issue 1, Winter 2005, Pages 3–25,**"**

[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] J. Préfontaine**,** J. Desrochers**,** L**.** Godbout The Analysis Of Comments Received by The BS Principles For Sound Liquidity Risk Management And Supervision, 2010

[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] K. A. Kaminski, T. Sterling Wetzel, & L. Guan, (2004). Can financial ratios detect fraudulent financial reporting? Managerial Auditing Journal, 19(1), 15-28

[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] K. A. Kaminski, T. Sterling Wetzel, & L. Guan, (2004). Can financial ratios detect fraudulent financial reporting?. Managerial Auditing Journal, 19(1), 15-28

[18] M. R. Dileep, A. V. Navaneeth, & M. Abhishek, (2021, February). A novel approach for credit card fraud detection using decision tree and random forest algorithms. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 1025-1028).

[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] Shiyang Xuan, et al. "Random forest for credit card fraud detection." 2018 IEEE 15th international conference on networking, sensing and control (ICNSC). IEEE, 2018.

[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] V. Dheepa, & R. Dhanapal, (2012). Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft computing, 2(4), 391-397.

[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] J. Chen, Z. Liu, J.Wu, Y. Yang, (2019). A comparative study of credit card fraud detection: Supervised and unsupervised approaches. Information Sciences, 501, 271-283.

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

[27] Y. Li, & S. Manoharan, (2013, August). A performance comparison of SQL and NoSQL databases. In 2013 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM) (pp. 15-19).

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

[29] X. Jiang, P. Ji, & S. Li, (2019, August). CensNet: Convolution with Edge-Node Switching in Graph Neural Networks. In IJCAI (pp. 2656- 2662).

[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] C. Liu, J. Wu, W. Liu & W. Hu, (2021). Enhancing graph neural networks by a high-quality aggregation of beneficial information. Neural Networks, 142, 20-33.

[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] F. Ameer, M. K. Hanif, R. Talib, M. U. Sarwar, Z. Khan, Zulfiqar, K., & A. Riasat, (2019). Techniques, tools and applications of graph analytic. International Journal of Advanced Computer Science and Applications, 10(4).

[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] G. ShefaliPatil, & A. Bhatia, (2014). Graph databases-an overview. 1Student, ME Computers, Terna College of Engg, Navi Mumbai, 2, 657- 660.

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

[37] N. Martinez-Bazan, S. Gómez-Villamor, & F. Escale-Claveras, (2011, April). DEX: A high-performance graph database management system. In 2011 IEEE 27th International Conference on Data Engineering Workshops (pp. 124-127). IEEE.

[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] S. Khatri, A. Arora, & A. P. Agrawal, (2020, January). Supervised machine learning algorithms for credit card fraud detection: a comparison. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 680-683).

[41] H. Rashid, (2020). Prospects of digital financial services in Bangladesh in the context of fourth industrial revolution. Asian Journal of Social Science, 2(5), 88-95.

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

[43] E. Elliott, & E. Elliott, (2021). Structured Streaming. Introducing. NET for Apache Spark: Distributed Processing for Massive Datasets, 171-184.

[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] A. Smid, R. Wang, & T. Cerny, (2019, September). Case study on data communication in microservice architecture. In Proceedings of the Conference on Research in Adaptive and Convergent Systems (pp. 261- 267).

[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] J.G. Shi, Y. B. Bao, F. L. Leng, & G. Yu, (2008, December). Study on log-based change data capture and handling mechanism in real-time data warehouse. In 2008 international conference on computer science and software engineering (Vol. 4, pp. 478-481).

[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] J. Kreps, N. Narkhede, & J. Rao, (2011, June). Kafka: A distributed messaging system for log processing. In Proceedings of the NetDB (Vol. 11, No. 2011, pp. 1-7).

[50] A. Simitsis, P. Vassiliadis, & T. Sellis, (2005, April). Optimizing ETL processes in data warehouses. In 21st International Conference on Data Engineering (ICDE'05) (pp. 564-575).