Addis Ababa University
አዲስ፡አበባ፡ዩኒቨርሲቲ

SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

ETHIO TELECOM AIRTIME CREDIT RISK PREDICTION USING MACHINE LEARNING

By:

Selahadin Nurga

Advisor: Million Meshesha (Ph.D.)

JUNE, 2023

Addis Ababa, Ethiopia

Addis Ababa University
አዲስ፡አበባ፡ዩኒቨርሲቲ

SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

A Thesis Submitted To The College Of Natural And Computational Science Of Addis Ababa University In Partial Fulfilment Of The Requirement For The Degree Of Masters Of Science In Information Science And System (Information Science)

By: Selahadin Nurga

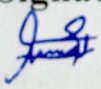
Name and signature of Members of the Examining Board

Million Meshesha (Ph.D.)	million	04/09/2023
Advisor	**Signature**	**Date**
Marta yifiru ((Ph.D.)		04/09/2023
Examiner	**Signature**	**Date**
Tibebe beshah (Ph.D.)		04/09/2023
Examiner	**Signature**	**Date**

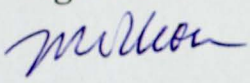
JUNE, 2023

DECLARATION

I, the undersigned, declare that the research presented in this thesis paper, titled "Ethio Telecom Airtime Credit Risk Prediction Using Machine Learning," is the result of my own original work and intellectual endeavors. The completion of this thesis has been guided by the expert supervision of my advisor, Dr. Million Meshesha. I further confirm that this thesis has not been previously submitted for any form of academic recognition at the university level. In line with academic integrity, all external resources employed in this study have been duly acknowledged and referenced.

Name	**Signature**
Selahadin Nurga	

I, as a university advisor, have granted my approval for the submission of this thesis for examination.

Name	**Signature**
Million Meshesha (Ph.D.)	

ACKNOWLEDGMENT

I want to start by thanking Allah (God) for giving me the strength, guidance, and blessings I needed for this journey. Your unwavering support has been the foundation of my perseverance and the source of my inspiration. I'm forever grateful for your divine wisdom and grace.

Next, I'm grateful for my advisor, Dr. Million Meshesha (Ph.D.), who mentored and guided me in this research work. My heartfelt appreciation goes out to him for his expertise, patience, and encouragement which helped shape my research and academic growth.

I also want to thank ethio telecom and the staff for giving me access to data that helped me conduct thorough research and reach meaningful conclusions.

I'm also grateful to my wife, Zebiba Negash, for her understanding, love, and support that sustained me during challenging times. Her constant encouragement pushed me beyond my own expectations.

Finally, I want to thank everyone who supported me - family, friends, and colleagues - with their contributions big or small. I'm highly grateful for your unwavering support and encouragement. To everyone who played a role in my academic and personal growth, you have my deepest gratitude.

ABSTRACT

Prepaid mobile consumers can use airtime credit to use telecom services even after their balance has run out and pay for it later. Users will find this service useful, and operators also make more money from it. But there's also a chance that subscribers won't pay their credits back. This study's main focus is on how machine learning techniques are applied to ethio telecom's airtime credit service customers to evaluate credit risk. Ethiopia's top telecom company, ethio telecom, needs to manage credit risk well to maintain financial stability and customer satisfaction. The company can identify customers who are more likely to default on their airtime credit by using accurate credit risk prediction, which enables proactive measures to lower risks and boost financial performance. The historical customer data included in this study include customer profile data, call detail data, loan history data, and usage data. Data preprocessing techniques are used before model training to handle missing values, encode categorical variables, and reduce features, ensuring the quality and consistency of the dataset.

In order to predict the credit risk associated with airtime, this study used supervised machine learning algorithms. Four different machine learning algorithms, including Naive Bayes classifiers, logistic regression, random forests, and k-nearest neighbors, were trained and tested using a dataset of 1,168,000 ethio telecom prepaid subscribers. Performance evaluation metrics like accuracy, precision, recall, and F1-score are used to assess each model's efficacy. Using class-balancing strategies, the models' robustness and generalizability are also validated.

According to experimental results, the random forest algorithm has successfully predicted airtime credit risk with 99% accuracy. In order to identify customers who are highly likely to default on their airtime credit, ethio telecom is able to take preventative actions with the help of this developed model, like adjusting credit limits.

The study's findings can fill the gap on how credit risk is predicted in the telecommunications industry and demonstrate how machine learning can improve risk management strategies and financial performance. The proposed method can be used as a foundation for the development of automated credit risk prediction systems for ethio telecom and other comparable organizations, resulting in enhanced decision-making processes and reduced financial losses.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background

Nowadays the need in staying competitive in the business has increased interest in risk and risk management in the scientific community. The phrase "risk" is currently employed in a number of different economic, social, and scientific contexts. However, the term 'risk' is widely used in finance, banking, insurance, and medicine. A few researchers attempt to construct a general theory of risk, however there is still a theory of risk situated in a specific context, such as insurance and banking. There is a need for research into risk factors in business processes to improve business outcomes. Risk research has become increasingly important for companies whose business revolves around providing services. For example, telecommunication service is an area that needs research to identify the risk factors in order to stay in the business with a competitive advantage. The telecommunication sector is one of the most competitive markets because it is constantly changing with the introduction of new technologies and business models. When a company knows a great deal about its customers, it can offer them better deals that give it a competitive edge. Identifying and managing risk in the telecommunications industry helps to minimize losses and increase profits[1].

In the telecommunications sector, special attention is paid to churn analysis, fraud detection, customer segmentation, optimal use of telecommunications infrastructure, etc. using data analysis techniques. In recent years, the credit risk analysis of individual and corporate customers in the activation process has also become important from the point of view of the telecommunications company's operational process. The term "credit risk" is used in both every day and scientific terms, however, Interpretation differs depending on the type of economic activity. The term 'credit risk' usually describes the risk of not achieving the target. The importance of credit risk is not the same across different sectors such as banking, insurance, telecommunications, energy, industry, and the public sector. For the telecommunications sector, credit risk means reduced potential profits, cash flow shortfalls, and financial difficulties that can lead to company bankruptcy. The main aim of this telecommunication sector is to help people stay in touch with each other and build strong connectivity networks with people around the globe using the network's airtime[2].

If someone wants to connect their mobile phone to a telecommunications network, they must purchase airtime so that they can use airtime on that telecommunications network. Airtime gives access to network services, voice, mobile data, SMS, and more. In developing countries, airtime is quickly becoming a staple of the growing middle class. For many prepaid customers, the lack of airtime to communicate or load data plans poses a challenge. This presents an opportunity for several mobile network operators (MNOs) operating in emerging markets to provide their subscribers with short-term airtime loans at a reasonable interest rate. The potential for this new service is to raise their average income per user[3].

In order to top up someone's account after running out of airtime, a telecom subscriber must go out and purchase a recharge card. But in some cases, a customer may be unable to leave the house to purchase a recharge card, particularly late at night or while traveling or he/she may not have the money to do that. Due to low balances, a client cannot make a call in for any needs, costing the network service providers and the customer a chance to make money. For those consumers who have used up their airtime but are unable to instantly go out and acquire recharge cards, network service providers now provide solutions. By sending text messages to the short numbers, customers can request to use the airtime credit feature. By recharging their accounts, customers will pay back this airtime credit loan.

1.1.1 Airtime credit

As stated by Iorliam et al [4], airtime credit is defined as "a recharge card or top-up card bought by a mobile phone user to aid him/her recharge/top-up his/her phone. This could assist the user in making calls, sending short message service (SMS), and obtaining bundles for browsing purposes". The real amount of time spent talking on a smartphone is known as airtime. It is the amount of time spent talking on a mobile device. However, airtime credit is a credit that can be used to purchase a specific amount of airtime (talk time).

A service offered by mobile network operators (MNO) called airtime credit can also be described as a type of service designed to raise customer happiness and bring in more money for Internet service providers (ISPs)[5].

Different names and explanations of what airtime credit loan means have been provided by various network service providers. For example, MTN offers Xtratime, Glo offers Borrow Me Credit, 9mobile offers Easy Credit, and Airtel offers Extra Credit. When a qualifying customer's account

balance is too low to normally support an ongoing call, MTN Xtratime offers airtime on credit. The airtime received on credit may be utilized for any fee-based activity, especially on the MTNN network. Glo-Borrow-me Credit is a service where active globacom prepaid customers are expected to claim airtime when airtime is running low and repay it with their next charge within a specified period of time. Easy credit by 9mobile is a service that offers more flexible top-up options to all prepaid clients. On the mobile network, they can borrow airtime for voice, data, and extra services and pay back the loan when they reload. Extra credit by Airtel is a service that provides credit advances to prepaid consumers with a three-day repayment period. Calls to other networks can be placed using the airtime. All users who have a three-month minimum Airtel network subscription are eligible for this service[6].

1.1.1.2 Ethio telecom airtime credit service

Using a contract with ethio telecom, Hikma electronic PLC and Credok communication technology partners launched airtime credit service in Ethiopia on August 31, 2018. Three parties were involved in the provision of the airtime credit service. These are the subscriber, a VAS (Value Added Service) provider, and the mobile service operator[7].

In 2018, ethio telecom introduced a pre-paid airtime credit service for its users. According to a 2020 report, this service is heavily utilized with two million users accessing it each month and a monthly loan value of 1.1 billion birr. In 2020, ethio telecom reported an estimated net revenue of 200 million birrs from the airtime credit service. This service provides convenience and versatility to users by offering access to telecom services even when their balance is low or zero, without the need to purchase a voucher card. To activate the service, customers must send an SMS with one of three keywords (A, L, or C) to the number 810 and follow the instructions provided. A credit scoring method is used to determine a user's eligibility before activating the service and providing a loan, which must be repaid within 180 days. Only active prepaid subscribers can access the airtime advance service, and just like a regular recharge the borrowed credit can be used to make voice calls, purchase internet or SMS bundles, use value-added services, or transfer the balance to another user. The airtime credit service requires a minimum of 5 birr and a maximum of 100 birr, with subsequent charges applied during a user's subsequent top-up. Overall, the airtime credit service has become a valuable revenue stream for ethio telecom while enabling greater convenience and accessibility for its users. If subscribers have paid off all prior debt, they may

borrow as often as they would want. According to the rule of ethio telecom credit service provider, eligible customer for airtime credit service (ACS) is the one which fulfill the following prerequisites[8]:

- The service user must be active prepaid mobile service customers,
- Customers whose service number is not barred, blacklisted, or suspended,
- Customers who have fully paid their previous airtime credit and
- All prepaid mobile customers who have been on ethio telecom network for at least three months with a minimum airtime recharge history of 15 birr per month.

However, as the market becomes more crowded, identifying eligible customer for airtime credit service becomes more challenging to attract new clients. Customers who take advantage of these deals are extremely valuable since they help the business make money. Airtime credit has a lot of benefits when loaded, but some people choose not to pay their bills or loans, which could result in significant losses for the business. To manage such kind of losses of business, companies can take safeguards by utilizing machine learning techniques to reduce this risk. Machine learning techniques is used to find clients who might not pay their bills on time. Analysis of the activation data is necessary to determine the customer's level of credit risk.

1.1.1.3 Machine learning as a solution

Machine learning, a subset of artificial intelligence, has grown significantly over the past decade when it comes to data analysis and computing, which often enables programs to perform intelligently. Machine learning is typically referred to as the most well-liked newest technologies in the fourth industrial revolution and provides systems with ability to learn and improve from experience automatically without being explicitly programmed. Machine learning algorithms are therefore essential for effectively analyzing these data and creating the related practical applications. The features and type of the data, as well as the success of the learning algorithms, determine the effectiveness and efficiency of a machine learning solution. Techniques including supervised, unsupervised, semi-supervised and reinforcement learning are available in the field of machine learning algorithms to efficiently create data-driven systems[9].

Many scholars have developed many algorithms, often referred to as methodologies, to carry out machine learning solutions. Random Forest, Naive Bayesian, k-Nearest Neighbor, k-Means, decision tree, Logistic Regression, etc. are a few examples. A variety of domains can benefit from applying machine learning. For instance, using machine learning technique can be employed to extract valuable information from time-series data, websites, temporal and spatial data, educational data, business data, telecom data, medical data, science data, engineering data, etc.[10].

One of the earliest sectors to use machine learning was the telecommunications sector. This is most likely due to the fact that telecommunications businesses often produce and retain massive amounts of high-quality data, have a significant client base, and operate in a dynamic environment with intense competition. This sector uses machine learning to enhance marketing campaigns, spot fraud, and maintain their networks more effectively. Due to the vastness of their data sets, as well as the sequential and temporal nature of their data, these businesses also confront a variety of machine learning and data mining issues[11].

For example, from the massive amount of data, call detail record data's, which include a description of each call, are a sort of data that telecommunications firms store on the calls that travel through their networks. This meant that there were billions of call detail records that were easily accessible for machine learning because call detail records are maintained online for several months. Applications for marketing and fraud detection can benefit from call detail data. A lot of consumer information is also kept on file by telecommunications firms, including billing information as well as data gathered from third parties, such credit score data. To enhance the outcomes of machine learning, this data is frequently supplemented with data particular to telecommunications. This study also demonstrates the use of machine learning techniques in the telecommunications industry to forecast airtime credit risk. Machine learning can be used for airtime credit risk prediction because it allows for the creation of predictive models that can analyze large amounts of data and identify patterns and trends that may not be immediately apparent to humans.

1.2 Motivation of the study

The motivation for this work comes from the recognition that, as 5G and Internet of Things (IoT) applications generate massive amounts of data, the need for effective data-driven algorithms has become critical. Approximately 63.5% of the current telecommunications infrastructure has begun to invest in and test the application of ML algorithms in operational network and business choices. Annual investments in the telecommunications sector are expected to reach $36.7 billion by 2025 [12]. According to the statement and recommendation of the study by Monika and Andrzej [2], which is based on credit risk handling in the telecommunication sector, the failure of the consumer to pay his/her bills is the cause of the telecom industry losses. This results in a lack of cash flow for the business and has a negative effect on it by generating bad debt. Bad debt is extremely risky and may be the cause of the company's financial downfall. Therefore, it is essential for businesses to identify threats and implement preventive measures. As a service provider, there are over 57 million mobile subscribers in ethio telecom. From this amount of telecoms subscriber, there is a large number of customers that use an airtime credit service. But the company offering this airtime credit service should have good prior knowledge about the customer's willingness to repay the loan or not. Machine learning algorithms provide a practical answer for the telecom industry's massive amounts of data to be turned into insights. Therefore, the study on airtime credit service contributes to a better understanding of how air time credit service should be offered to the customers to make the telecommunication industry more successful.

1.3 Statement of the problem

The telecommunications industry is among the most competitive market area because it adapts quickly to new technology and business models. The more insight a business has into its clients, the more appealing of a deal it can present, giving it a competitive edge over rivals.

According to Govind and Manish [13] , in the telecommunications sector, fair pricing, good customer service, and coverage are important variables that can significantly impact customer satisfaction. Among these variables, offering services like airtime credit is the one which keep clients happy and satisfied. The credit airtime service allows prepaid mobile subscribers to obtain airtime even if the customer has low or no credit. Additionally, in most cases, subscribers cannot top up their accounts late at night or earlier in the morning. Providing throughout this time credit loans for airtime, can increase user convenience and increase the revenue generated by users. An

appropriate evaluation of the applicants' trustworthiness is necessary for the granting of airtime credit (loans) by the telecommunications sectors.

In fact, when customer become a defaulter (stopped making loan repayment), the sector suffers losses of principal and interest, the interruption of cash flow and an increase in collection costs. One of the key processes for seeking to evaluate, measure, and control the risks emerging from the Probability of default is the implementation of credit risk management (CRM) systems in the telecommunications industry. If we take the case of ethio telecom report, as of 2019, more than 4.5 million subscribers had unpaid loans after ethio telecom announced its credit airtime service for prepaid mobile consumers on September 1, 2018. The total loan balance for these subscribers at that time is roughly 2.7 million Birr. One can easily understand that this is a significant quantity of outstanding debt, which the lender may find quite concerning [5].

This is a result of the lack of a system that addresses the issue of knowing the trend, pattern, and probability distribution of airtime lending and recovery. Knowing the trend and pattern have the power not only to assist the vendors to know repayment of loan mechanism, but it also helps on knowing how much funds to be injected into the airtime bucket. So, there should be a special treatment of customers with a high airtime credit risk that allows the company to minimize its financial losses. It is also very important to have all the necessary information regarding identifying the best features that used to indicate a customer to grant with airtime credit services or not. Proper identification of the customer and scoring his risk level lets the company to lower its credit risk. Nowadays, telecommunication companies started to make several attempts to control risk management. The problem addressed in this study revolves around the effective prediction of airtime credit risk. To achieve accurate predictions, it is essential to identify suitable attributes and determine the most appropriate machine learning algorithms for airtime credit predictions. Additionally, evaluating the performance of optimal models in predicting airtime risk is crucial to understand their effectiveness and practical applicability.

Oliyad [5] and Shashu[7] both used a supervised machine Learning algorithms to predict mobile airtime credit risk in 2019 and 2017 respectively. But on predicting airtime credit risk in telecommunication sector, the performance of the prediction algorithms applied by these study was impacted since the researchers did not take a consideration of potentially relevant information of customer usage data attributes such as, repay_amount, operater_type, late_pay_fee,

off_peak_usg_minute, inter_usg_minute , repay_poundage, sms_local_usage , sms_inter_usage , init loan_amt ,the amount of fee used for local SMS, and more are not into consideration to predict airtime credit risk. There is also an issue on the advancement of service, technology and an increase data size in the telecom industry within this time period gap. According to Oliyad [5], a research on credit risk in financial institutions cannot be used for predicting airtime credit risk because there are significant differences in the nature of loans, such as collateral agreement, loan amount, risk assessment, technology dependence, micro loan size, convenience, and trust-based lending.

The problem addressed in this study also try to identify suitable machine learning algorithms for airtime credit predictions and evaluate the performance of optimal models in predicting airtime risk. In the context of airtime credit, it is crucial to determine the most appropriate machine learning algorithms that can effectively predict credit risk. With a wide range of algorithms available, it becomes essential to explore and identify the algorithms that are specifically suitable for airtime credit predictions. By selecting the most appropriate algorithms, accurate predictions can be made regarding customers' creditworthiness, enabling better decision-making processes and risk management strategies.

Furthermore, assessing the performance of optimal models is essential to understand their predictive capabilities in predicting airtime risk. By evaluating the accuracy, precision, recall, and other performance metrics, the study aims to determine how well these models perform in differentiating between potential non-defaulters and defaulters. This evaluation will provide insights into the effectiveness of the models and their practical applicability in managing airtime credit risk.

Overall, the aim of this study is to build a machine learning model for airtime credit risk prediction by identifying suitable attributes, selecting appropriate machine learning algorithms, and evaluating the performance of optimal models. The findings of this research will provide valuable insights for telecommunications companies, such as ethio telecom, in managing credit risk effectively and making informed decisions regarding airtime credit.

1.4 Research Questions

The following research questions are investigated and answered in this study

- What are the attributes suitable for airtime credit risk prediction?
- Which machine learning algorithms are suitable for airtime credit predictions?
- How the optimal models perform with prediction of airtime risk?

1.5 Objective of the study

This study's general and specific objectives are mentioned below.

1.5.1 General objective

The main objective of this study is to Explore the development of a predictive model using machine learning algorithms that can accurately predicts airtime credit risk of ethio telecom.

1.5.2 Specific objectives

By aligning the specific objectives with deliverable-oriented outcomes, this research will contribute to the development of a reliable and accurate predictive model for airtime credit risk prediction, providing valuable insights for risk assessment and decision-making in the context of ethio telecom. The Specific objectives of this research, in order to achieve the stated general objective are mainly concentrates on the following key points:

1. Conduct a comprehensive literature review on credit risk prediction: This objective involves reviewing relevant literature to identify and analyze existing methods and techniques for predicting credit risk. The outcome will be a comprehensive understanding of the current state of research in this field, enabling the identification of best practices and potential gaps for further investigation.
2. Collect and analyze consumer data for modeling and analysis: This objective aims to gather and analyze important consumer data that is necessary for building the predictive model. This includes data related to customer behavior, transaction history, and any other relevant variables. The outcome will be a well-curated dataset that serves as the foundation for the model construction and analysis.

3. Identify key attributes for enhancing airtime credit risk prediction: This objective involves identifying the specific attributes or variables that significantly impact the prediction of airtime credit risk. By conducting thorough exploratory data analysis and statistical techniques, the objective is to determine the most influential attributes. The outcome will be a set of key attributes that enhance the model's capability to accurately predict airtime credit risk.
4. Select the most appropriate machine learning algorithms for constructing an optimal model: This objective focuses on evaluating and selecting the machine learning algorithms that are best suited for constructing an optimal predictive model. Through performance evaluations metrics, the objective is to determine the algorithms that yield the highest predictive accuracy and efficiency. The outcome will be the selection of the most suitable algorithms for model construction.
5. Evaluate the performance of the proposed predictive model: This objective aims to assess the performance of the developed predictive model using appropriate evaluation metrics such as accuracy, precision, recall. The objective is to determine the model's effectiveness in predicting airtime credit risk and to provide insights into its practical applicability. The outcome will be a comprehensive evaluation of the model's performance and its potential for real-world implementation.

1.6 Scope and Limitations of the study

The purpose of this study is on creating models that can determine the likelihood that customers will default on their airtime usage payments or not. In order to make wise choices about credit risk management in the context of ethio telecom's airtime services, this scope includes analyzing historical data, using machine learning algorithms, and creating predictive models.

1.7 Significance of the Research

For ethio telecom, the application of machine learning to forecast airtime credit risk is extremely important. Ethio telecom can lower the risk of default, effectively manage its credit portfolio, and boost overall revenue by predicting the credit risk of customers prior to granting airtime credits. Additionally, by offering a solution that enables telecoms to precisely identify high-risk customers and offer them a tailored experience, this study will aid in the development of Ethiopia's telecom sector.

The research also has a significant impact on policy makers, researchers as to serve as a starting point if they need to implement machine learning models using telecommunication databases to get the hidden knowledge and to make the industry more competitive edge by identifying loyal customers. As a result, this study is being carried out in order to provide adequate details on airtime credit risk by employing machine learning algorithms to discover hidden knowledge based on ethio telecom's historic data. Specifically, the significance of this study for ethio telecom include increasing and identifying loyal customers, providing quality service for the customers and to applying data driven solution. And the study also can help ethio telecom on the way of how to apply machine learning for data driven decisions to predict airtime risk.

Researchers can use the current study as a base for conducting further study towards designing and implementing an intelligent system that automatically predict airtime risk.

1.8 Thesis organization

In this thesis, Chapter One serves as the Introduction, providing the background, motivation, problem statement, research questions, objectives, scope of the study. Chapter Two is the Literature Review, which presents an overview of machine learning and related works in the field. Chapter Three focuses on Data Preparation and explaining the proposed framework. Chapter Four is dedicated to the Experiment, detailing the experimental setup, the dataset used, and presenting the experimental results for each algorithm (Random Forest, Logistic Regression, Naïve Bayes classifiers, and K-Nearest Neighbors). The chapter also includes a discussion of the findings. Finally, Chapter Five concludes the thesis with a summary of the research findings.

CHAPTER TWO

LITERATURE REVIEW

A survey of both conceptual and empirical literature was undertaken in this chapter to gain a conceptual knowledge of machine learning and the field in which it might be used. Then, in order to understand how machine learning might be used in credit risk prediction, earlier research on the issue is also reviewed.

2.1 Overview

Information is currently recognized as being extremely significant in the economic world and influencing how businesses make decisions in light of the modern economy's rapid expansion. Finding information that is beneficial to the company has therefore become crucial[23]. As a result of the development of information technology, large databases and massive volumes of data have been generated in a variety of sectors. As a result of research in databases and information technology, a method for storing and exploiting these massive volumes of data for future decision-making has evolved[24].

Credit risk assessment is a vital aspect of decision-making, aiming to predict the probability of borrowers defaulting on their obligations. Traditional methods of credit risk analysis have relied on manual evaluation and predetermined rules. However, with the advancement of technology, machine learning algorithms have emerged as powerful tools to enhance the accuracy and efficiency of credit risk prediction[25].

Machine learning leverages computational models and algorithms to automatically analyze vast amounts of data and identify patterns or relationships that may not be evident to human analysts. In the context of credit risk, machine learning algorithms can efficiently process diverse information such as borrower characteristics, financial history, and transactional dat. By incorporating machine learning techniques, financial institutions can gain deeper insights into credit risk factors and make more informed decisions in lending and risk management.[26]. Today the information industry has access to a vast amount of data. But until this data is transformed into valuable information, it is of no use. Therefore, it is essential to examine this vast volume of data

and draw out relevant information using machine learning. Beside information extraction, Machine learning requires additional procedures, including data preparation, data wrangling, analyze data, train the model, test the model, and deployment.

2.2 Credit risk

Credit risk is a term used to describe the likelihood that a borrower will default on a loan or fail to meet their financial obligations. It is a crucial factor in modern economies, as it affects the stability of financial institutions and the overall health of the economy. Accurate estimation of credit risk is of utmost importance for the stability of the financial system. Inaccurate credit risk estimation, can have severe systemic consequences. As a result, lenders dedicate significant resources to predict the creditworthiness of borrowers and formulate lending strategies that effectively mitigate their risks. Traditionally, credit risk has been measured using statistical methods and manual auditing. However, recent advances in financial artificial intelligence have led to the development of machine learning-driven credit risk models. These models use statistical, machine learning, and deep learning techniques to estimate credit risk[25].

The credit risk in the telecommunication sector, refers to the risk of customers not paying their bills on time or at all. In the context of telecommunication, airtime credit risk can be defined as the risk of customers not paying for the airtime or data usage they have consumed. This can result in financial losses for telecommunication companies, which is why it is important to identify customers who may be at a higher risk of not paying their bills[2].

According to B. Dushimimana et.al, airtime credit risk refers to the risk associated with lending airtime to subscribers who may not be able to repay the loan. Credit scoring is a method used to assess the likelihood of a customer's ability to repay a loan based on various features associated with the customer. These features encompass the customer's historical and financial data, including factors such as loan repayment history, frequency of transactions, and source of income. By analyzing these features, credit scoring models aim to quantify the creditworthiness of individuals and provide insights into their repayment capabilities.

2.3 Overview of Machine Learning

In many ways, machine learning (ML) affects how we live our daily lives. We frequently ask smartphones with ML capabilities to recommend wonderful places or to lead us through an unfamiliar area. ML techniques are already common tools in many disciplines of research and engineering. Machine learning applications are changing people's lives at an increasing rate and scale[27]. Machine learning has emerged as one of the most important subjects among development organizations searching for new ways to exploit data assets to assist the business obtain a new degree of understanding. Organizations can use proper machine learning models to continuously predict changes in the business so that they can best predict what's next. Because data is continually being added, machine learning models ensure that the solution is always up to date.[28].

Machine learning is a developing technology, that allows computers to automatically learn from previous data[29]. Machine learning is a concept that can be used to describe learning from the past (in this case, historical data) to enhance performance in the future. This field only focuses on autonomous learning techniques. The term "learning" describes the automatic adjustment or enhancement of an algorithm based on prior experiences without any external human input[30]. By providing the data required for a computer to train and adjust effectively when exposed to new data, machine learning advances the technique to a more sophisticated level. In order to improve its ability to understand incoming data and generate more useful findings, it focuses on extracting information from incredibly massive sets of data by using different statistical metrics to detect and identify underlying patterns[30]. In research on artificial intelligence, machine learning is crucial. One cannot consider an intelligent system to be truly intelligent if it is incapable of learning[31].

Automatically creating models is a major goal of machine learning research. A model is a pattern, plan, representation, or description that demonstrates how a system or concept works, such as a rule that must be followed to perform a mathematical operation and produce a specific result, a function from sets of formulae to formulae, or a pattern that can be used to create things or parts of things from data[32]. In the study of Martin et al[33] which studied on banking risk management using machine learning noted that one disadvantage of machine learning is that its processes are more "black box," with occasionally unexpected outcomes. They are also said to be sensitive to outliers, which leads to overfitting of the data and incorrect predictions. They are also touted as

having the advantages of being a better fit for non-linear correlations between the explanatory and explained factors, as well as having the advantage of being able to use a wider range of variables, which tends to increase accuracy. Over the past few years, an amazing number of ML algorithms have been developed and introduced. Not all of them have a large used by users. One was replaced by another because some of them did not satisfy or resolve the issue[30].

2.4 Machine learning types

According to their characteristics, modes of learning, and methods for using data, Machine learning algorithms can be classified into one of four categories. Supervised, Unsupervised, semi-supervised, and reinforcement learning [9].

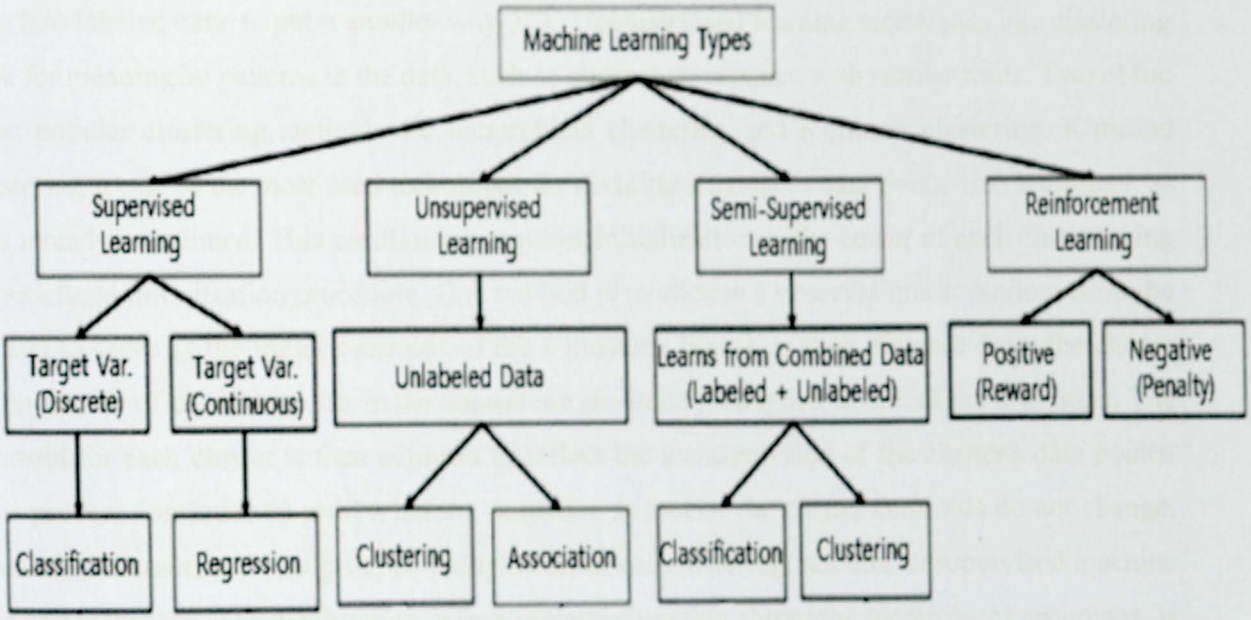


Figure 1: Type of Machine learning with data used and algorithms [9]

Supervised machine learning algorithms are type of machine learning which able to build general patterns and hypotheses and forecast future instances with the aid of externally supplied examples. Its goal is to classify data using knowledge from the past. It investigates and analyzes the training data and attempts to deduce a function that can be applied to mapping new examples. In a best-case situation, it will enable the algorithm to accurately estimate the class labels for cases that have not yet been seen. The learning algorithm must be able to reasonably generalize from training data to new scenarios[34]. Supervised Learning is the process of calculating and changing the error to

obtain the desired output. This learning process is based on the comparison of computed output and expected output. In supervised learning, the computer incorporates both the examples' labels or objectives and their actual information. The labels on the data help the computer correlate the features. This means that supervised learning necessitates training on labeled data that contains both desired inputs and outputs[35].It's similar to training a baby to walk up until the child learns to walk independently by holding the child's hand, showing the infant how to put forward a foot and walk by example, and so on[36] .

In Unsupervised learning, Instead of using labeled or training data, It try to group some random data according to the behaviors of the data. If there is some commonality in the data, it will categorize according to behavior and similarities. It simply represents an effort to turn unlabeled data into labeled data, to put it another way[35]. Unsupervised learning techniques like clustering look for meaningful patterns in the data, such as clusters of samples with similar traits. Two of the most popular clustering methods are hierarchical clustering and K-means clustering. K-means clustering is one of the most used techniques for dividing a dataset's data points into k clusters, as was already mentioned. This partitioning requires initialization of the center of each cluster using a predefined initialization procedure. One method is to choose k observations at random from the dataset to serve as the initial centroids of the k clusters. Based on their distance from the cluster centroids, all of the data points in the dataset are iteratively assigned to one of the k clusters. The centroid for each cluster is then adjusted to reflect the average value of the cluster's data points. This process continues up until a halting condition is met or the cluster centroids do not change. Unlabeled datasets are also grouped using hierarchical clustering, another unsupervised machine learning technique. The dendrogram, a tree-like structure that shows the hierarchy of groupings, is created using this technique. The number of clusters need not be predetermined, as we did with the K-Means algorithm[35][22][37].

Semi-supervised learning is a type of machine learning that focuses on using both labeled and unlabeled data to perform certain learning tasks. It allows for the utilization of vast amounts of unlabeled data available in various use cases in conjunction with normally smaller collections of labeled data. It is situated conceptually between supervised and unsupervised learning. The majority of semi-supervised learning research is concerned with classification. Semi-supervised classification approaches are very beneficial when there is a scarcity of labeled data. In those

conditions, developing a trustworthy supervised classifier could be difficult. In application domains where labelled data is expensive or difficult to gather, it is difficult to develop a trustworthy supervised classifier.[38].

Rewarding desired behaviors and/or punishing undesirable ones are the foundations of the machine learning training method known as reinforcement learning. A reinforcement learning agent can typically perceive and interpret its environment, act, and learn through mistakes[39].

This research focuses on the uses of supervised learning by employing classification algorithms. As a result a detail discussion of supervised machine learning is given below.

2.5 Supervised machine learning algorithms

Classification algorithms are employed for supervised learning. Classification is the systematic process of creating classification models for training and test data sets. The Classification algorithm is a Supervised Learning technique that is used to categorize new observations on the basis of training data. A program learns from the dataset or observations provided and then classify additional observations into a number of classes or groups in classification. The categorization method produces a category rather than a value as an output variable. Because it is a supervised learning technique, the Classification algorithm employs labeled input data. As a result, it includes both input and output information. The primary objective of a classification algorithm is to determine the category of a given dataset, and these algorithms are primarily used to forecast the results for categorical data. A classification model is created using the training data set and this categorization model may then automatically categorize the data into various groups[40][41][31].

Depending on the dataset being used, one can apply a variety of classification algorithms. Here are a few of the most popular machine learning classification algorithms.

- Decision Tree Classification
- Random Forest Classification
- K-Nearest Neighbors
- Support Vector Machines
- Logistic Regression
- Naïve Bayes

2.5.1 Decision Tree Classification

One of the classification methods is the decision tree, which classifies data using splitting criteria. The decision tree is a tree structure that resembles a flowchart and is used to categorize instances by ordering them according to the values of their attributes (features). An attribute in a classification instance is represented by each and every node in a decision tree. Every branch represents a test result, and each leaf node has the class label. According to the instance's feature value, a classification is made [42].

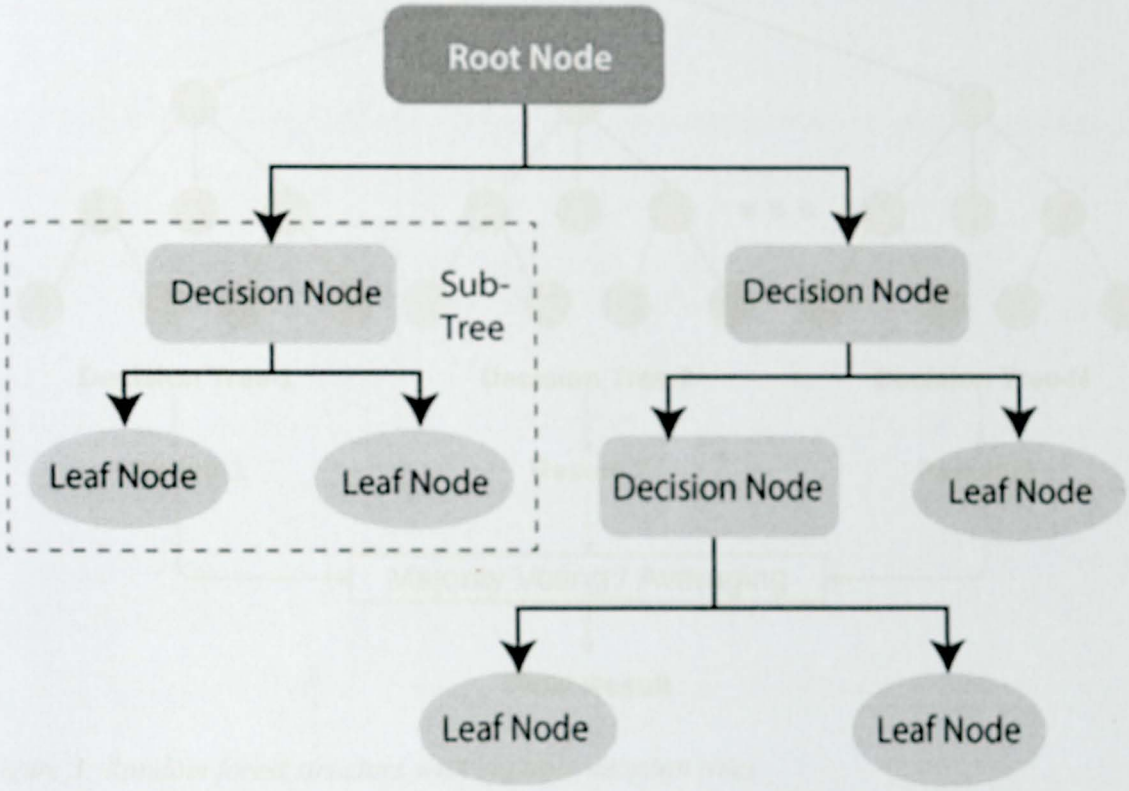


Figure 2: an example of a decision tree structure

Finding the attributes that divides training data most effectively can be done using a variety of techniques, including information gain, gain ratio, Gini index, etc. The most popular method for creating decision trees involves top-down greedy method partitioning, which begins with the training set and iteratively seeks a split feature that maximizes a particular local criterion. The classification rule for the data set is produced by the decision tree. The ID3, C4.5, and CART fundamental algorithms are frequently employed[42][9].

2.5.2 Random Forest

A random forest is an example of the bootstrap aggregation or bagging family of ensemble learning techniques. Bagging's major objective is to deal with the high variance of models, which causes overfitting and a lack of generalizability.

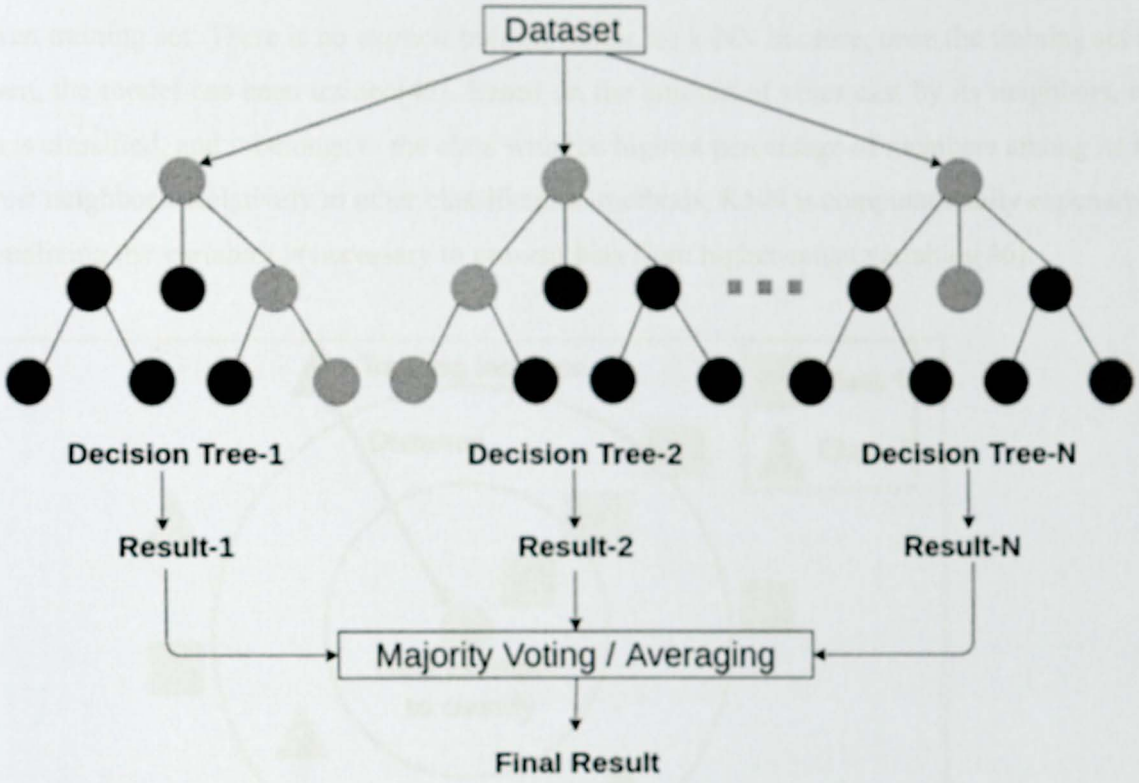


Figure 3: Random forest structure with multiple decision trees

A group of decision trees make form a random forest model. Using a categorization by majority vote tasks and a mean or avarage for regression tasks, the output of these decision trees is merged for a given data point to produce a forecast that performs better than the output of any one decision tree alone. The trees are uniquely trained using random training set subsamples. A training set is created for each decision tree in a random forest using random sampling and replacement with data from the initial training set. Because each decision tree was trained separately using a different set of random samples from a training dataset, combined decision trees perform better in terms of predictions than a single decision tree. Because it uses randomization to ensure that each tree is unique from the others, this model is known as a random forest. This algorithm works well for

both categorical and continuous variables and can be applied to problems involving classification and regression[43][22][9].

2.5.3 K nearest Neighbor

k-nearest neighbors (KNN) is one of the supervised machine learning technique that can be used to tackle both classification and regression issues[44]. K-NN is a more straightforward method of classification algorithm, which classifies data based only on the nearest neighbor (or neighbors) in a given training set. There is no explicit training phase for k-NN because, once the training set is chosen, the model has been trained[45]. Based on the amount of votes cast by its neighbors, an item is classified, and it belongs to the class with the highest percentage of members among its k-nearest neighbors. Relatively to other classification methods, KNN is computationally expensive. Normalizing the variables is necessary to prevent bias from higher-range variables[46].

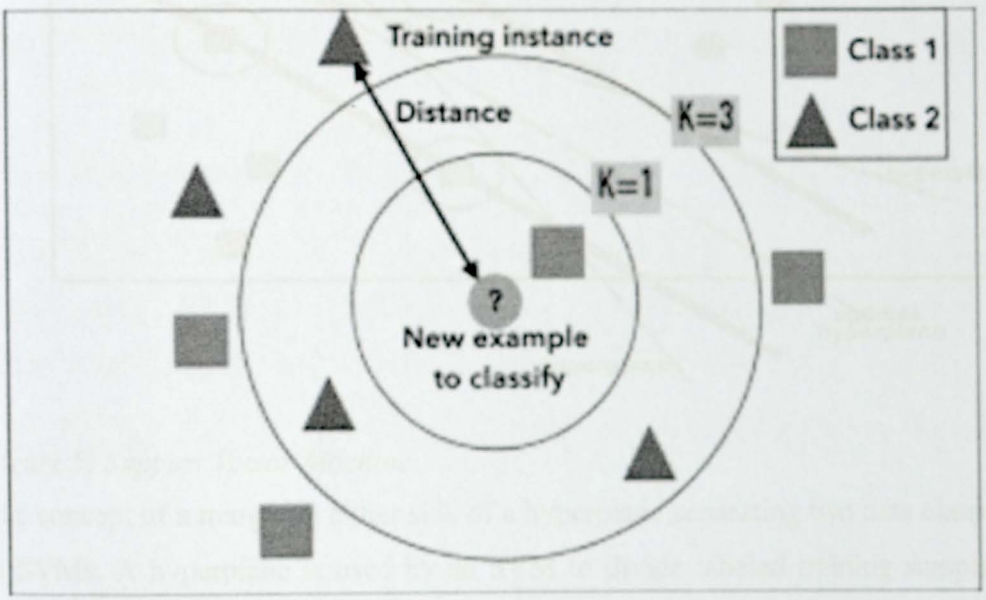


Figure 4: Classification on new object using KNN

Although k-Nearest Neighbors classification algorithm is one of the simplest methods in machine learning and fairly straightforward to understand and put into practice, this approach has found extensive application in a range of disciplines, including anomaly detection, semantic search, and recommendation systems[30].

2.5.4 Support Vector Machine (SVM)

Strong regularization characterizes this supervised machine learning approach, which can be applied to classification or regression problems. Their use of kernels, the sparseness of the solution, and the capacity control obtained through margin manipulation, the quantity of support vectors, etc., are what define them. Parameters that are unaffected by the size of the feature space, govern the system's capacity. A detailed knowledge of an SVM can be surprisingly challenging despite its basic simplicity, as is the case with many machine learning approaches[30].

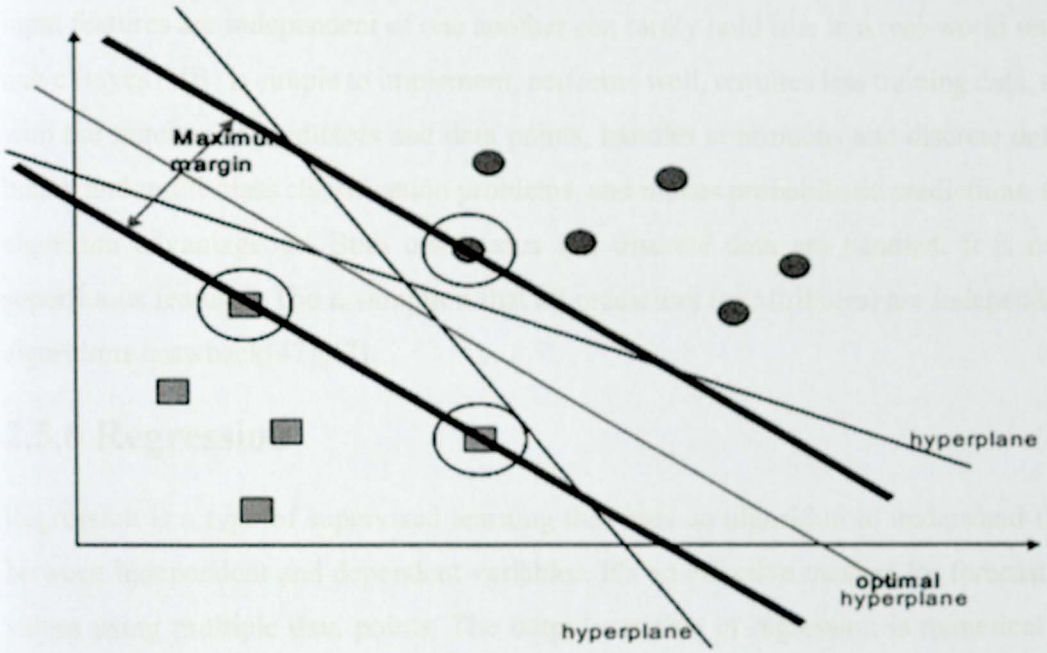


Figure 5: Support Vector Machine

The concept of a margin in either side of a hyperplane separating two data classes is fundamental to SVMs. A hyperplane is used by an SVM to divide labeled training samples into classes, if possible, while maximizing the "margin," or the distance between the classes and the hyperplane. The data in an SVM is often mapped to a higher dimensional space, which enables hyperplane separation. We can work in a higher dimensional space using the so-called "kernel trick" without suffering a major performance hit[45][15][30].

2.5.5 Naïve Bayes

This algorithm is constructed using the Bayes theorem and it is a fundamental classifier for probabilistic classification. The naive Bayes classifier (NBC) is a basic classifier used for probabilistic classification and it has been developed based on the Bayes theorem. The NB model is referred to as "naive" since it is based on the assumption that each feature is statistically independent of the other features and irrelevant to them in the training set. The assumption that all input features are independent of one another can rarely hold true in a real-world setting. Because naive Bayes (NB) is simple to implement, performs well, requires less training data, scales linearly with the number of predictors and data points, handles continuous and discrete data, can handle binary and multi-class classification problems, and makes probabilistic predictions, this makes the algorithm advantageous. Both continuous and discrete data are handled. It is not reactive to superfluous features. The assumption that all predictors (or attributes) are independent makes the algorithms drawback[47][43].

2.5.6 Regression

Regression is a type of supervised learning that uses an algorithm to understand the connection between independent and dependent variables. It's an effective method for forecasting numerical values using multiple data points. The output variable in regression is numerical (continuous), meaning that we train the f(x) hypothesis to provide continuous output(y) given the input data (x). The regression technique is utilized in the prediction of numbers, size, values, etc. since the output contains information about the actual number. Linear regression, support vector regression, and Poisson regression are examples of popular regression algorithms. The well-known regression algorithm called linear regression is an algorithms that finds the relationships and dependencies between variables. It depicts a relationship between a continuous scalar dependent variable (also known as a label or target in machine learning terminology) and one or more explanatory variables (also known as independent variables, input variables, features, observed data, observations, attributes, dimensions, data point, etc.[48][22].

2.5.7 Logistic regression

One of supervised approach for solving a classification problem is logistic regression. It delivers the binomial outcome, which represents the likelihood that an event will occur or not, based on the values of the input variables. A logistic function, often known as the sigmoid function in mathematics, is frequently used in logistic regression to estimate the probabilities. Mostly, this algorithm is used to solve problems of industry scale. Given that the result of a logistic regression is a probability score, in order to utilize it to solve a business problem, one must define customized performance measures in order to achieve a cutoff that can be used to categorize the target. The advantages of logistic regression are considered to be ease of regularization, computational efficiency, and efficiency from a training perspective. Input feature scaling is also not necessary. The use of a logistic regression method may become limiting if there are large number of features/variables present or if the variables are highly correlated. Another limitation of this algorithm is the assumption of linearity between the dependent and independent variables[47][49].

2.6 Related works

Here under related works done, to investigate airtime credit risk, majority of the works are done by foreign scholars by applying data mining and machine learning.

The article of Monika and Andrzej [2] highlights credit risk concerns as well as techniques for identifying clients who might stop making payments following the activation processing of telecom industry sector. The paper discusses the application of data mining methods in the telecommunication sector for credit risk prediction and management. The authors present several models and methods for identifying customers with high credit risk, based on activation data and payment behavior. The models are constructed for individual and business customers, as well as for different types of services. The paper emphasizes the importance of credit risk management in the telecommunication sector to prevent bad debt and financial losses. The authors suggest that credit risk management can be improved by implementing deposit policies and predicting customer inability to repay the debt after the debt collection process. The paper also discusses the use of activation models in the telecommunication sector to predict customer non-payment and prevent financial debt. For their data analysis the researchers exploited a database from a telecommunications provider that contains data on both individual and commercial clients who entered into a contract with the company between January 1, 2007, and March 31, 2008. The

database contained 53433 observations which represent customer's population. This population consists of 48663 willing payers of invoices (good customers) and 4770 clients whose contracts were terminated for nonpayment (bad customers). The models use decision trees and variable importance measures to identify high-risk customers and their common characteristics. The authors suggest that these models can be used to prevent financial debt by identifying high-risk customers and implementing targeted interventions. The paper presents misclassification rate charts based on training and validation data, which assess how well the model fits the data. The authors recommend using decision trees and variable importance measures to identify high-risk customers and their common characteristics. They suggest future research on analyzing activation models based on different split criteria and predicting customer churn. Overall, the paper concludes that the models presented in the paper can help telecommunication companies make informed decisions and prevent financial losses due to credit risk. The authors suggest that future research should focus on improving the accuracy of the models and analyzing activation models based on different split criteria. The weaknesses in the literature identified in the paper "Credit Risk Handling in the Telecommunication Sector" is, beyond the focus on activation models based on classification trees, there is a need for a comparative examination of various data mining methodologies or models for credit risk analysis in the telecommunications sector.

Bernard et al [3] explored airtime credit service (ACS) that enables consumers to readily acquire airtime on a credit basis using an empirical examination of more than three million loans from more than 41 thousand clients. Their study began with a meta-analysis of earlier works that looked at how credit scoring algorithms are built and helped to choose the pertinent elements and model structures that have worked well. Their goal was to develop a suitable quantitative model for forecasting loan outcomes using consumer behavior on the mobile network and loan-related financial data.

They discovered two separate mechanisms in the airtime lending sector for their investigation. In the first, MNOs provide customers with airtime loans while taking on the risk of defaulted loans. In this mechanism, a subscriber who runs out of airtime can borrow money with the assumption that it will be paid back within five days, up to the amount they have added to their account in the previous seven days. The second approach entails a collaboration between the MNO and a third-party lender, whereby the MNO grants access to the customers and mobile network while the risk is passed to the third party. According to their discoveries, the amount of data exchanged with the

lenders is restricted in such an environment since MNOs have a propensity to protect the privacy of their clients. As a result, lenders are less knowledgeable about their clients than financial institutions are. In this system, the credit amount is paid to the MNO by the third party lender before the consumer makes a repayment. Therefore, in the event of a default, the third party is out the entire sum. The lender must make sure that the revenues from loans that are successfully repaid outweigh the losses from non-performing loans in order to continue to be financially sustainable. The research described in this publication is based on this second airtime lending mechanism. The researchers used data analysis, feature selection, evaluation and cross-validation as a methodology. They conduct an exploratory examination of the data and provide a summary of the variables that will be used to create a credit scoring model. The average loan term, total number of loans, and average use amount for both non-defaulting and defaulting customers are summarized in these statistics. In feature engineering part, the researchers divides the features into three categories: customer behavior, loan details, and customer details (Age, Gender).The following features are used for their study: loan amount, recharge frequency, usage amount, activation date, loan application date, loan due date, and total amount used each month. The study was just focused on the loan detail and customer behavior from the list of features above.

They employed binary classifiers for the two discrete outcomes of client behavior repayment or default. A decision tree, a logistic regression, and random forests were three different machine learning model structures that were taken into consideration. The classification models are evaluated using a confusion matrix, with positive (negative) results indicating repayment and default, respectively. With an accuracy of 82.3%, Random Forest was the top classifier, showing the superiority of nonlinearity and an ensemble technique. The authors used a single dataset obtained from only one company and they did not include or used a demographic information for the examination.

V. Dengov[50] presents a statistical credit risk analysis model for Russian telecommunications companies in his article titled "Credit Risk Analysis for the Telecommunication Companies of Russia: A Statistical Model." This model outlines the potential utilization of statistical techniques in evaluating credit risk. The article discusses the state of the Russian telecommunications sector today and the significance of credit risk analysis for these businesses.

In the opening section of the article, a general overview of Russia's rapidly expanding telecommunications sector is given. According to the author, it is crucial for telecommunications

companies to accurately assess credit risk in order to maintain financial stability as the industry continues to grow. The article's statistical model is based on logistic regression analysis, a method that is frequently employed in credit risk analysis. To determine the likelihood of default for a specific telecommunications company, the model makes use of a number of different variables, including financial ratios.

In-depth details on the model's variables and their calculations are provided in the article. The significance of data quality and data preprocessing for creating precise statistical models is also covered by the author. The author offers detailed examples of financial ratios and other elements pertinent to this sector. This means that, compared to a more general model, the statistical model is more applicable to Russian telecommunications companies. The focus on data quality and data preprocessing is another aspect of this article's strength. The author admits that a careful preprocessing procedure and high-quality data are necessary for creating an accurate statistical model. This demonstrates the significance of effective data management in credit risk analysis.

The article does, however, have some restrictions. One drawback is that the article's statistical model is solely based on financial ratios. Financial ratios are a crucial part of credit risk analysis, but they might not fully reflect the creditworthiness of a telecommunications company. Other elements that might be important to credit risk analysis, like market circumstances or the competitive environment, are not covered in the article.

In conclusion, Dengov's study, "Credit Risk Analysis for the Telecommunication Companies of Russia: A Statistical Model," offers an important framework for carrying out credit risk analysis within the Russian telecom sector. The article highlights the significance of preprocessing and data quality, and it offers specific instances of variables that are pertinent to this sector. Although the model suggested in the article has some drawbacks, it offers a useful starting point for additional research in this field.

According to Oliyad[5], offering clients services like airtime credit is one way to make them happy and satisfied. But when it comes to debt repayment, this company encounters special problems. The author undertakes this study to try to bridge this gap applying a data mining technique for predicting airtime credit risk. The experiment was carried out using an open source data mining tool WEKA. The author used different supervised classification algorithms to determine which model performed the best. These algorithms included the Nave Bayes, Multilayer Perceptron, Logistic Regression, and J48 Decision Tree. The algorithms were tested using 86, 024 instances

and eleven attributes from the ethio telecom prepaid subscriber's usage data. 10-fold cross validation and percentage split test options were applied. To assess the effectiveness of the models the author used several metrics like Confusion matrix, ROC area, accuracy, precision, recall, and f-measure. Based on experimental result of the author, the J48 decision tree-based model outperformed the other classifiers with an accuracy rate of 98.56%. The accuracy of the model created using Logistic Regression is 97.17%. While the accuracy of the Multilayer Perceptron and Naive Bayes classifiers was 96.7622% and 94.63%, respectively. Finally the author indicated that, among all features, data usage is the main attribute that showed the potential prediction power. According to the study's findings, factors such as topping up channel, and voice usage have great predictive power. The publication also notes the researcher's desire to improve the models used in the study by including new features such as subscriber loan history.

In the study of Shashu [7] , one of the value added services (VAS) provided by MNOs is the airtime credit service, which enables prepaid mobile users to top up their airtime credit at any time and from any location. However, many subscribers typically do not repay their loans on time or they fail. The author undertakes this study to try to bridge this gap by applying "Machine Learning Based mobile airtime credit risk prediction using customer Profile and Loan Information". The researcher highlighted that, when prepaid mobile subscribers are unable to recharge their service number, they mostly use airtime credit. This service both rise subscriber happiness and income earned by the telecom provider. Even if this service increases customer satisfaction and telecom revenue, the authors claim that there is no assurance that the prepaid customers would repay the credit they borrowed, hence using this service carries some risk. To solve this issue, and to evaluate the impact of the suggested strategy, the author carried out an experiment with data from 90,000 mobile subscribers. The researcher used four machine learning algorithms for the experiment: decision tree (DT), logistic regression (LR), random forest (RF) and multilayer perceptron (MLp). According to the researchers experiment, the results show that the mix of existing features employed by Oliyad [5] and new variables together increases the prediction of airtime credit risk performance for all algorithms and J48 decision tree have scored best from all. Even though the study shows that the combined feature set improves accuracy, it doesn't go into detail about how the new features were identified or explore. It would have been better also, if the author had provided more details about the study's data. The author could have, for instance, explained the collection and preparation of the data.

Author name	Title (Domain)	ML algorithm used	Issue addressed	Research Gap
Monika and Andrzej i[2]	Credit risk handling in telecommunication	Decision Trees	Evaluating and projecting credit risk from the time a customer account is opened until it is closed due to nonpayment	There is a need for a comparative examination of various data mining methodologies or models
Bernard et al [3]	Creating a Credit Score Model for Airtime Loans	logistic regression (LR), decision tree (DT), Random Forest (RF)	Creating a machine learning model for airtime lending	The need for access to more variables to improve the performance of credit scoring models
Victor[50]	"Credit Risk Analysis for the Telecommunication Companies"	Statistical method	Evaluating credit risk models and create an awareness for telecom based industries.	• The statistical model proposed in the article is based solely on financial ratios. While financial ratios are an important component of credit risk analysis, they may not capture all of the relevant information about a telecommunications company's creditworthiness. • The article does not discuss other factors that may be relevant to credit risk analysis, such as market conditions or competitive landscape.
Oliyad[5]	Airtime Credit Risk	decision tree classifier, Naïve Bayes, Logistic Regression, multilayer Perceptron	Predicting the risk of defaulting in the airtime credit service for prepaid mobile subscribers.	• Unable to incorporate a features that used to boost the accuracy this include the loan history of the subscriber
Shashu[7]	Mobile Airtime Credit Risk Prediction	logistic regression, random forest and j48 decision tree, MLP	Predicting airtime credit risk using machine learning	• It doesn't dig into detail about how the features were identified or explored. • Did not provided more details about the study's data collection and preparation.

Table 1: Comprehensive Review of Related Work: A Summary of Key Findings in Table Format

As the reviewed related literature indicates, different credit risk prediction models are built, but these different models of credit risk estimation by themselves indicate that no one model can provide reliable and accurate results by itself. This is due to the gaps in data size used and duration of the study, algorithms selection, ignoring the necessary variables or attributes, and measuring the accuracy of the model using only training data. As a result of taking into account all of the research gaps, this study resolves the problems of airtime credit risk prediction.

CHAPTER THREE

METHODOLOGY

3.1 Overview

In this section the study explains the machine learning objectives, data sources, and procedure employed to create the models of machine learning. The purpose of this study, as stated at the beginning, is to create a model that can help in predicting the airtime credit risk of borrowers utilizing machine learning techniques. As a result, the experiment is carried out using a chosen machine learning approach. To create machine learning models for this work, supervised learning techniques of classification algorithms are used. This is because, supervised learning algorithms are well suited for credit risk prediction tasks where labeled data is available and they can efficiently learn from historical patterns and provide precise predictions for new instances based on those learned patterns.

3.2 Research design

The design of this study is experimental research design. It is a scientific approach to conducting research, involves manipulating one or more actions on independent variables and then observing the impact on a dependent variable. To systematically examine how well various machine learning models perform in predicting credit risk in ethio telecom, an experimental design for airtime credit risk prediction using machine learning has been developed. The experimental research design was chosen for the airtime credit risk prediction using machine learning because it offers a methodical way to assess the performance of various machine learning models and feature selection techniques in predicting credit risk in ethio telecom and allows the researchers to establish causal relationships between the independent and dependent variables. To conduct an extensive experiment, this, the research follows various machine learning process, starting with problem identification and ending with model prediction and validation. The methodology for this research is adapted from previous studies[14][15] and the process is broken down into the following steps, as illustrated in below figures.

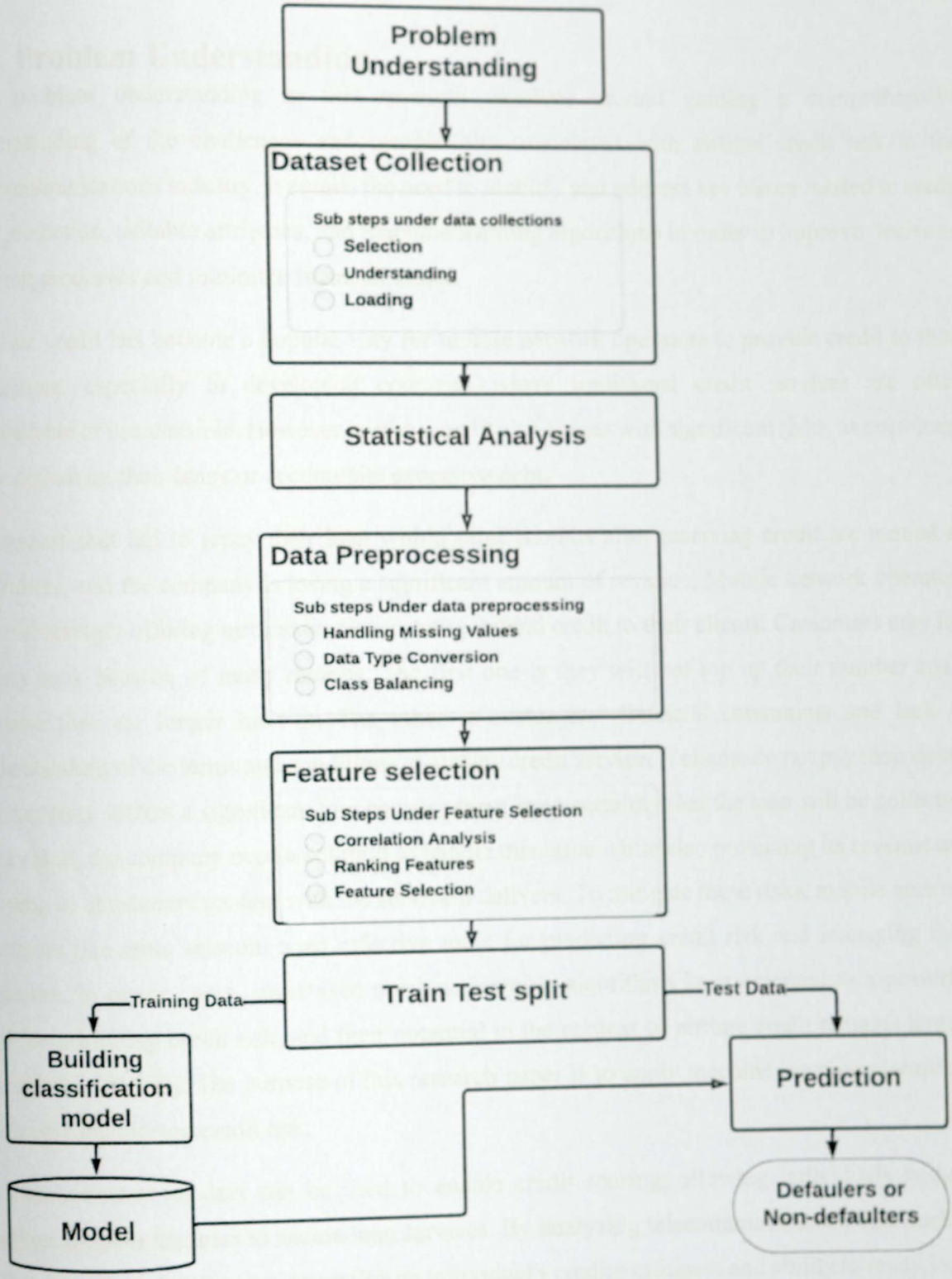


Figure 6: General Methodology of the Study

3.3. Problem Understanding

The problem understanding in this research revolves around gaining a comprehensive understanding of the challenges and complexities associated with airtime credit risk in the telecommunications industry. It entails the need to identify and address key issues related to credit risk prediction, suitable attributes, and machine learning algorithms in order to improve decision-making processes and minimize financial losses.

Airtime credit has become a popular way for mobile network operators to provide credit to their customers, especially in developing countries where traditional credit services are often unavailable or inaccessible. However, airtime credit also comes with significant risks, as customers may default on their loans or accumulate excessive debt.

Customers that fail to repay their loan within three months after receiving credit are termed as defaulters, and the company is losing a significant amount of revenue. Mobile network operators are increasingly offering airtime credit service to extend credit to their clients. Customers may fail to pay back because of many reasons. The first one is they will not top up their number again because they no longer have it. The other is customers' financial constraints and lack of understanding of the terms and conditions of airtime credit service. If clients do not pay their debts, the company suffers a significant loss because there is no certainty that the loan will be collected. As a result, the company need a solution to handle this issue while also protecting its revenue and keeping its consumers content with the service it delivers. To mitigate these risks, mobile network operators like ethio telecom need effective tools for predicting credit risk and managing their exposure. In recent years, supervised machine learning algorithms have emerged as a powerful tool for predicting credit risk, and their potential in the context of airtime credit remains largely unexplored [51][52]. The purpose of this research paper is to apply machine learning algorithms for predicting airtime credit risk.

Telecommunications data can be used to enable credit scoring, allowing individuals without traditional credit histories to access loan services. By analyzing telecommunications data such as call and text logs, lenders can determine an individual's creditworthiness and ability to repay loans [53].

In the telecom sector, customer data includes records of calls, texts, data traffic, application program (App) usage, borrower demographic data (age, education, gender, location etc.), loan application details, and loan repayment history. These data are reliable since they are routinely captured, and they are plentiful because mobile phones are used frequently and hence have a lot of data on them. Moreover, data on customer usage has already been utilized to examine consumer habits, economic traits, and user behavior. With all these benefits, customer usage data has a significant potential to predict loan default[51].

Another study also supports that, mobile phone usage data can be a useful predictor of credit risk. From those several mobile phone usage variables, including call duration, SMS count, and data usage, were significantly associated with credit risk[54].

Recharge datasets are important for telecom companies as they provide insights into customer behavior and usage patterns. By analyzing this data, telecom companies can identify trends in recharge behavior, such as the most popular recharge amounts, the most common payment methods, and the times of day when customers are most likely to recharge their phones[51].

Accordingly, this research investigates the use of customer behavior, and transaction history to effectively predict credit risk using machine learning algorithms. To do this, we must take into account the factors or variables that can affect credit risk in order to fully understand the challenge of estimating credit risk for users of airtime credit. Customer behavior, transaction history, and income level are examples of these variables. Creating effective predictive models for anticipating credit risk and controlling exposure an attempt is made to better understand these factors and their relation with credit risk.

3.4 Data Collection Method

The methods for data gathering or collection from the many systems used by the organization are discussed in this part. The way in which the dataset are integrated and processed are also discussed. The dataset for this study was gathered from subscriber data for ethio telecom's prepaid mobile service, from the databases of the company's business administration systems. This study focuses on examining airtime credit risk prediction using a dataset from a specific three-month period, spanning from October to December 2022. The reason for collecting data only within this limited timeframe is attributed to the data retention policy of the organization. As per Ethio Telecom's data

retention policy, information is retained for a maximum of 45 days to three months before being archived.

Customer profile data

The first dataset collected from one of the telecom database system is customer profile data. This dataset in ethio telecom refers to information that is collected and maintained about a telecom customers, such as customer type, education, gender, age, religion and service number. By analyzing this data, we can gain insights into customers' behavior and preferences to improve customer service and develop new products and services. However, ethio telecom handles this data with care and respect customers' privacy by adhering to data protection laws and regulations. As a result of this, after combining the dataset using a service number as a primary key, the attribute "service number" is not used in the analysis.

Loan information dataset

The second dataset gathered from telecom database system is Loan information dataset. The loan Information data in ethio telecom refers to the data generated by the telecom operators regarding the loan or credit facilities provided to their customers. This data includes information on the amount of loan, the repayment period, the due date for repayment, the outstanding balance, and the payment history of the borrower. This loan information data is used in this study to determine the creditworthiness of customers and to assess their ability to repay the loans or credit facilities provided by the operator. This information is typically used to determine whether a customer is eligible for a loan or credit facility, as well as to set the terms and conditions of the loan.

In general Loan Information data is an important dataset to manage credit risk and to minimize the risk of default by telecom customers. By analyzing this data, operators can identify customers who are at a higher risk of default and take appropriate measures to mitigate that risk, such as offering more favorable repayment terms or reducing the amount of credit offered.

Usage detail dataset

The other dataset collected from their database system is customer's usage detail dataset. A customer usage dataset in ethio telecom data is a collection of data that records the way in which customers use telecommunication services. This data includes information on customer behavior,

such as the number of calls made, the duration of each call, the amount of data used, and the number of text messages sent. The customer usage dataset is valuable to ethio telecom as it can be used to identify trends and patterns in customer behavior. For example, it can help providers to identify which services are most popular, which customers are using the most data, and which areas have the highest call volumes. This information can be used to optimize network capacity, develop new services, and tailor marketing efforts to better meet the needs of customers.

Call detail dataset (recharge dataset)

Finally the fourth data type gathered from the telecom is call detail dataset (recharge dataset). A dataset for mobile phone recharge transactions is referred to as a recharge dataset in the telecom sector. The phone number being recharged, the amount being recharged, and the date and time of the transaction are normally included in this dataset.

3.4 Data construction (merging)

Merging a dataset is important because it allows data analysts and researchers to combine data from different tables create a larger, more comprehensive dataset and they can identify patterns, relationships, and trends that would not be apparent if the data were analyzed separately. A raw data captured from various environments are heterogeneous, complex, imperfect, and of a huge scale, which brings us many challenges to transform them into useful information[55].

Since the dataset collected from ethio telecom are from different database system, the researcher in this study merged them into one data frame using pandas merge () function. The researcher in this has also faced a challenge in merging the different datasets since the dataset has different shapes and some of the dataset sizes are also too large to read and process using known data analytic tools and the normal machines at hand. The way how the dataset are merged is described as follows.

For the case of customer usage data, which has a file name of "customer_usage_data.csv", it is difficult to read and know how many instances and attributes it has due to its size of 36GB. The researcher also use a remote server with a RAM capacity of 90GB and try to read and process the data, but still, the server cannot handle reading and processing the whole data. To solve the problem, the researcher first use a Linux terminal to split the dataset into three parts. Before splitting the data frame, the number of rows in the file/ data frame is counted by line using the

code `wc -l customer_usage_data.csv` in the terminal and it gives a result of 297,289,997 rows. After counting the number of rows, the data frame is split into three parts using the code `split -n l/3 customer usage data.csv`. After splitting the data frame, they individually imported into Python and duplicated values in all columns of the same rows are removed. Finally, all the split datasets are concatenated in columns into one data frame with assigning a name customer_usage_data.csv and it has 115,421,629 rows and it is saved into working directories to be merged with another data frame.

The other data set called loan information data is also difficult to read and process using the known machines and tools at hand since it has 95,786,870 rows. But the server that the researcher used before for the case of customer usage data has read it but takes a couple of minutes to import to Python. After reading the dataset, duplicated instance in all columns were removed. Finally, the loan information data is dropped to 30,980,644 rows.

Recharge information (call details) data has a 5,259,892 rows and 4 columns which is relatively a small data set and merged together on service number with loan information data and a new dataset with a name "loan_and_recharge.csv" is created and saved to working directory. This new dataset, which is loan_and_recharge.csv has a 131,698, 319 rows, and 14 columns. At last, customer profile data with a shape of 9,036,696 rows and 7 columns is reduced by removing duplicate rows in all columns commonly and the dataset shape is changed to 6,174,513 rows and 7 columns.

In the end, the service number, often known as the access number, was used as the primary key to link all the information and put them together in one table by taking a random sample of 200,000,000 instances from each dataset. This newly derived dataset has a 2000,000 rows and 33 columns as indicated in the table below. Now the rest of all data preprocessing is done using the newly derived dataset.

Table 2: Attribute description of the dataset

Attribute Name	Data Types	Description
SERVICE NUMBER	Numeric	identity used to uniquely identify a subscriber
CUST_TYPE_NAME	String	Individual or Enterprise
NET BUSI_TYPE_NAME	String	Network type (LTE,WCDMA)
STATUS NAME	String	The status of service number(Active, Idle, Barring, Predeactivated, Suspend)
CUST_AGE	Numeric	The age of the customers
GENDER NAME	String	Gender name (Male or Female)
EDUCATION_NAME	String	This shows academic rank of the user
OPER_TYPE	String	It can be by transfer or recharge.
LOAN BALANCE_TYPE	Numeric	This simple a code given to loan data
INIT LOAN_AMT	Numeric	Initially a customer takes the loan amount
INIT_LOAN_POUNDAGE	Numeric	Initial specific fee or charge related to loans or borrowing money
LOAN POUNDAGE	Numeric	specific fee or charge related to loans or borrowing money
LATE_PAY_FEE	Numeric	Shows the fee paid late
REPAY_POUNDAGE	Numeric	How much is repay per defend period of time
LOAN_PENALTY	Numeric	Penalty after grace time or given loan time
LOAN_PENALTY_LEFT	Numeric	Repay penalty left
LOAN GRADE	Numeric	Level of the loan, (silver, gold, platinum)
DECODE	String	Mobile type: - Prepaid, postpaid, and hybrid.
RECHARGE AMNT	Numeric	Birr recharged by customers
OFF_PEAK_USG_MINUTE	Numeric	Minute used in off peak hour
OFF PEAK_AMT_ETB	Numeric	Birr collected in off peak hour (3:00PM-12AM)
PEAK_USG_MINUTE	Numeric	Minute used in peak hour (7:00AM-3:00PM)
PEAK_AMT_ETB	Numeric	Money collected in peak time
INTER_USG_MINUTE	Numeric	Internet usage in minute
INTER_AMT_ETB	Numeric	Internet amount in birr
SMS_LOCAL_USAGE	Numeric	National SMS usage
SMS_LOCAL_FEE_ETB	Numeric	SMS national collected birr
SMS_INTER_USAGE	Numeric	international SMS usage
SMS_INTER_FEE_ETB	Numeric	SMS international collected birr
SMS_LOCAL_FEE_ETB.1	Numeric	SMS national collected birr
DATA_USAGE_MB	Numeric	Amount of data used in Mb
DATA_REVENUE_ETB	Numeric	Birr collected from data

3.5 Statistical Analysis

This step is generally used to get meaningful insights into the collected data set via data visualizations and analytical findings. This stage will assist in determining whether the data set contains any missing values, identifying categorical and numerical features, and more.

The following figure shows the result of a code for obtaining the data types of each attributes in the data set. According to the result, seven attribute have a categorical (based on python output categorical means object) data type and the rest are numerical by types

```
df.dtypes
```

```
SERVICE_NUMBER                                          int64
CUST_TYPE_NAME                                         object
NET_BUSI_TYPE_NAME                                     object
STATUS_NAME                                            object
CUST_AGE                                              float64
GENDER_NAME                                            object
EDUCATION_NAME                                         object
OPER_TYPE                                              object
LOAN_BALANCE_TYPE                                       int64
INIT_LOAN_AMT                                           int64
INIT_LOAN_POUNDAGE                                      int64
LATE_PAY_FEE                                            int64
LOAN_POUNDAGE                                           int64
REPAY_AMT                                               int64
REPAY_POUNDAGE                                        float64
LOAN_PENALTY                                            int64
LOAN_PENALTY_LEFT                                       int64
LOAN_GRADE                                              int64
DECODE(T.SGMT_TYPE,0,'PREPAID',1,'POSTPAID',2,'HYBRID')  object
RECHARGE_AMNT                                         float64
OFF_PEAK_USG_MINUTE                                     int64
OFF_PEAK_AMT_ETB                                        int64
PEAK_USG_MINUTE                                         int64
PEAK_AMT_ETB                                            int64
INTER_USG_MINUTE                                        int64
INTER_AMT_ETB                                           int64
SMS_LOCAL_USAGE                                         int64
SMS_LOCAL_FEE_ETB                                       int64
SMS_INTER_USAGE                                         int64
SMS_INTER_FEE_ETB                                       int64
SMS_LOCAL_FEE_ETB.1                                     int64
DATA_USAGE_MB                                           int64
DATA_REVENUE_ETB
```

Correlation aids in the discovery and removal of redundant or closely related features. When two attributes are highly associated, they may provide similar information and it may not be required to include them in the analysis. We can decrease the dataset's complexity and increase computational efficiency by detecting and eliminating unnecessary features. Correlation coefficients quantify how strongly two variables are related. When two variables are correlated, it indicates that when the value of one changes, the value of the other tends to change in a specific way. Understanding that relationship is beneficial because the value of one variable can be used to predict the value of the other variable.[20]. Therefore to remove attributes that have similar information, the correlation of the dataset in this study is also explored. To do this, the function df.corr ( ) is used to determine the pairwise correlations of all columns in a data set. It gives the following correlations between each numerical attributes. Figure 11 following depicts the pairwise correlations.

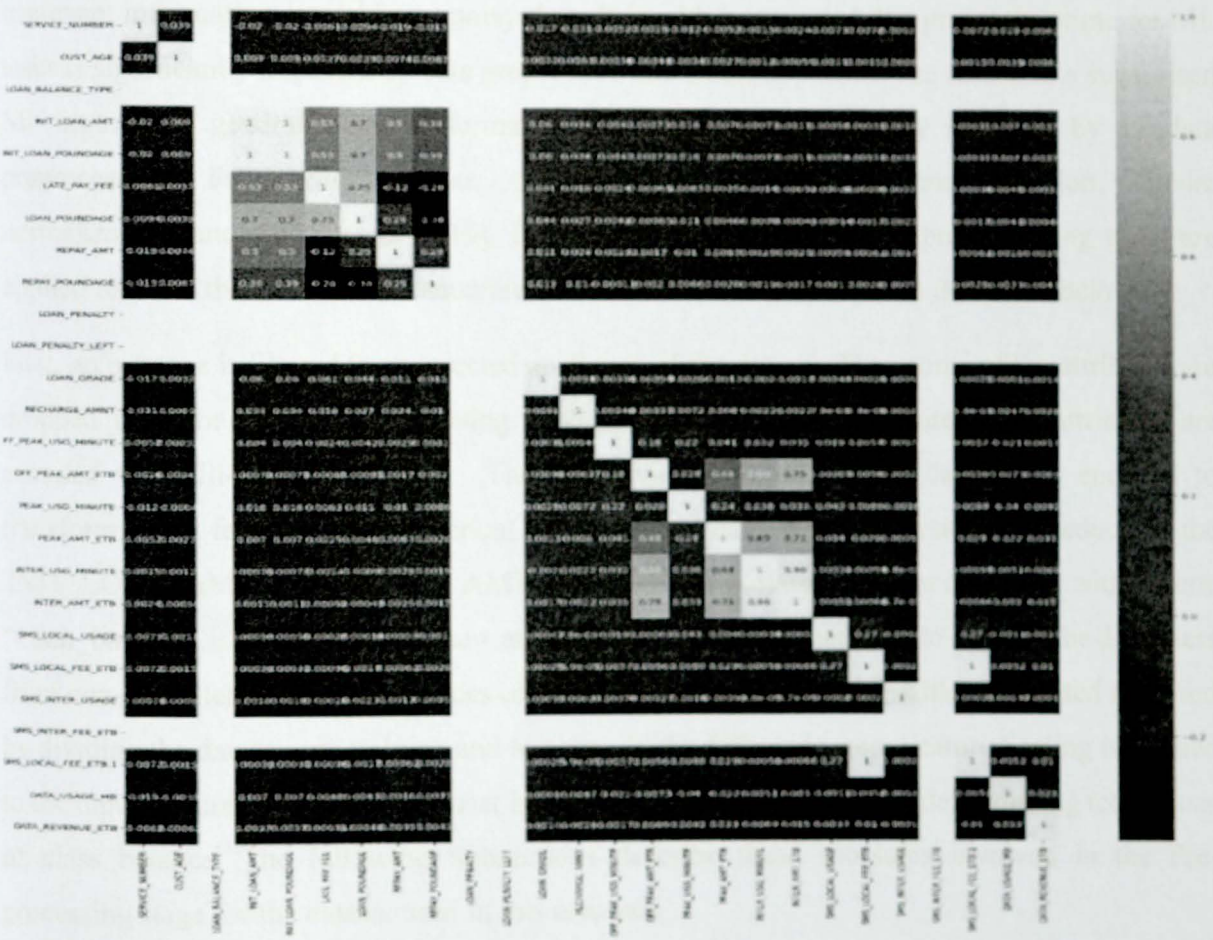


Figure 7: Correlation analysis of the attributes

Based on the result the attribute with a name "Initial loan amount" and "Initial loan poundage" strongly correlated with a correlation value of 1. This is because both attribute has identical values in each rows of instance. The other attributes "INTER_USG_MINUTE "and "INTER_AMT_ETB" has also strongly correlated with a value of 0.97. As indicated from the above attributes labeled with a white colors have no correlation at all. This is because all values or instance of each of the attributes are identical throughout the rows. These attributes include "LOAN_BALANCE_TYPE", "LOAN_PENALTY", "LOAN_PENALTY_LEFT" and "SMS_INTER_FEE_ETB"

3.6 Data Pre-Processing

Machine learning's (ML) effectiveness on a specific task depends on a number of factors parameter. The representation and quality of instance data are therefore of utmost importance. Knowledge discovery during the training phase is more challenging if there is a lot of redundant, irrelevant information available, or noisy data. It is widely accepted that processing time for ML tasks is significantly impacted by data preparation and filtering phases. The results of a supervised ML algorithm's generalization performance are frequently significantly impacted by the data preprocessing. Pre-processing data includes feature extraction and selection, feature normalization, and data cleansing[19]. In this research, different data preprocessing tasks are applied to make the dataset more accurate for the model and this tasks are described below.

First, an instance is filtered from selected attributes of the dataset. Then some of the attributes are dropped from the dataset. The missing values of the numerical and categorical attributes are encoded with fill-forward method. The Categorical Features of the dataset are encoded to transform these features into numerical values. After data type conversion, by deducting the 'INIT_LOAN_AMT' from 'REPAY_AMT attributes, a new class feature or attributed with a name "Loan_balance" is created and this new attribute was used as a class label to identify the defaulters from non-defaulters .Then the matrices of features and the predictor variable are created followed by dividing the dataset into training and test sets. In the following step, Feature Scaling is applied to the input features. Finally, the dataset is balanced using the over and under sampling techniques of class balance. The following Subsections describe these sub-steps involved in the Pre-processing stage for the dataset used in this research.

3.7 Data Reduction

At the start, during the data collection time, the ethio telecom data has included both the instance 'Individual' and 'Enterprise' customers in the "CUST_TYPE_NAME" attribute. But at processing stage only the instance "individual" is filtered by assuming the probability of 'Enterprise' to take a loan is too less. Similarly, in this research only customers whose service numbers are active is considered. The attribute with a name "STATUS NAME" has four different instance (labels), namely 'Active', 'Idle', 'Barring', 'Predeactivated', 'Suspend'. From this labels of instance, for this research, only 'Active' instance are filtered. Similarly the studies mainly concerns only customer that pays for services upfront, before they are using any of telecom service this is we call it is prepaid. This is because, ethio telecom gives a service of airtime credit for only activated prepaid subscribers. To filter both "Individual", "Active" and "prepaid" instance from "CUST_TYPE_NAME" , "STATUS NAME" and "DECODE" attributes respectively, the following python code is used.

```
df= df.loc[(df["CUST_TYPE_NAME"]=='Individual') & (df["STATUS_NAME"]=='Active')& \
    (df["DECODE(T.SGMT_TYPE,0,'PREPAID',1,'POSTPAID',2,'HYBRID')"]=='prepaid')]
```

After filtering instances, the features "CUST_TYPE_NAME" and "STATUS NAME had unique value in the feature columns, so they are dropped from the dataset. Similarly the features "LOAN_BALANCE_TYPE","LOAN_PENALTY","LOAN_PENALTY_LEFT",'SMS_INTER_FEE_ETB',and "SMS_INTER_USAGE" had a unique values in the features columns and as a result they are dropped from the dataset. Dropping those feature which have a unique values will be beneficial for the implemented Machine Learning models as it only contains only one value, so the Algorithms will not learn anything from this feature.

To collect and integrate data, the attribute "service number" is used as a primary key. However, due to privacy concerns and its insignificance in this study, it was eventually deleted. There is a duplicated feature with a name "SMS_LOCAL_FEE_ETB.1", and this is also removed.

3.8 Handling missing values

Incomplete data is an unavoidable problem in dealing with most of the real world data sources. To solve this incomplete data, there are a number of methods for handling missing values. Method of Ignoring Instances with Unknown Feature Values, Most Common Feature Value, Concept Most Common Feature Value, Mean substitution, Regression or classification methods[19]. To fill the missing values in the dataset, this study used a fill-forward method, in which each missing data point was replaced with the most recent observed value. To do this, first, the total number of missing values in each columns are checked and the following result are gained.

```
df.isnull().sum()# ['CUST_AGE','GENDER_NAME', 'EDUCATION_NAME',
```

```
NET_BUSI_TYPE_NAME          0
STATUS_NAME                 0
CUST_AGE                 3877
GENDER_NAME              3727
EDUCATION_NAME         126276
OPER_TYPE                   0
INIT_LOAN_AMT               0
INIT_LOAN_POUNDAGE          0
LATE_PAY_FEE                0
LOAN_POUNDAGE               0
REPAY_AMT                   0
REPAY_POUNDAGE              0
RECHARGE_AMNT               0
OFF_PEAK_USG_MINUTE         0
OFF_PEAK_AMT_ETB            0
PEAK_USG_MINUTE             0
PEAK_AMT_ETB                0
INTER_USG_MINUTE            0
INTER_AMT_ETB               0
SMS_LOCAL_USAGE             0
SMS_LOCAL_FEE_ETB           0
DATA_USAGE_MB               0
DATA_REVENUE_ETB            0
```

As it is indicated with above output of python code, a feature CUST_AGE has 3877 missing values, GENDER_NAME has 3727 and education has 126276 missing values. From these feature, CUST_AGE is integer in data types and "GENDER_NAME" and "EDUCATION_NAME" are a string in data type. To fill these missing values of the attributes a, forward fill method is applied for both type of data types. Forward filling involved imputing the data using the last time point with available data[56]. To fill the missing data using the forward fill method the following python code is used.

```
# Forward-Fill
df= df.fillna(method="ffill")
```

According to the statement of the expertise in ethio telecom, the telecom data collection system is set up to register Ethiopian Birr values in multiples of 10,000. As a result, the attribute that indicates customers' loan information, like"INIT_LOAN_AMT","INIT_LOAN_POUNDAGE","LATE_PAY_FEE","LOAN_POUNDAGE","REPAY_AMT","REPAY_POUNDAGE","OFF_PEAK_AMT_ETB","PEAK_AMT_ETB","INTER_AMT_ETB","SMS_LOCAL_FEE_ETB","DATA_USAGE_MB" and "DATA_REVENUE_ETB"has a value in multiple of 10,000 as shown the sample below.

INIT_LOAN_AMT	LOAN_AMT	LOAN_POUNDAGE	REPAY_AMT
50000	0	0	50000
100000	40000	10000	60000
250000	100000	25000	150000
150000	0	15000	150000
250000	0	25000	100000
500000	500000	50000	0
100000	100000	10000	0

To get the real value of loan information data, these attribute values are divided by 10,000 and the returned value is changed to integer types to get the real values using the python code given below

```
df[["INIT_LOAN_AMT","INIT_LOAN_POUNDAGE","LATE_PAY_FEE","LOAN_POUNDAGE","REPAY_AMT","REPAY_POUNDAGE",\
   "OFF_PEAK_AMT_ETB","PEAK_AMT_ETB","INTER_AMT_ETB","SMS_LOCAL_FEE_ETB","DATA_USAGE_MB","DATA_REVENUE_ETB"]]\
=df[["INIT_LOAN_AMT","INIT_LOAN_POUNDAGE","LATE_PAY_FEE","LOAN_POUNDAGE","REPAY_AMT","REPAY_POUNDAGE",\
    "OFF_PEAK_AMT_ETB","PEAK_AMT_ETB","INTER_AMT_ETB","SMS_LOCAL_FEE_ETB","DATA_USAGE_MB","DATA_REVENUE_ETB"]]\
.div(10000).astype("int")
```

3.9 Encoding Categorical Features

The feature engineering process is a crucial component of machine learning. It is crucial to convert the categorical features into numerical values this is because the algorithms that will be used can only read numerical values[57].

At this stage of this research, The raw dataset is processed by get_dummies() function of sklearn to change all the categorical variables into dummy variables. The categorical attributes of the dataset in this study are NET_BUSI_TYPE_NAME, STATUS_NAME, GENDER_NAME, EDUCATION_NAME and OPER_TYPE. To change these categorical features in to numerical, the following python code and package are used.

```
df=pd.get_dummies(df,df.columns[df.dtypes == 'object'])
```

After converting all category attributes to numeric values, a target variable named "Loan_Balance" is created by calculating the difference between the initial loan amount (INIT_LOAN_AMT) and the repayment amount (REPAY_AMT). Rows with a difference value (the value of Loan_balance) equal to zero are classified as non-defaulters, whereas those with a difference value different from zero are classified as defaulters. This labeling technique can be justified, since the difference in loan and repayment amounts directly reflects a customer's financial accountability and adherence to loan repayment obligations. Customers who have completely repaid their loans (zero difference) demonstrate responsible financial behavior and fulfill their obligations, making them suitable for the non-defaulter category. Another justification can be as assessing default status based on the difference between loan and repayment amounts aligns with industry practices and commonly used indicators of loan repayment behavior. It allows for consistency and comparability with existing credit risk assessment methodologies used by financial institutions and lending organizations.

As discussed before, the value of the target class attribute have a values 0 indicating customers who repay the loan they taken and values different from zeros indicating customers who did not repay the loan amount in a given period of time. In this research a two class label values are created to indicate defaulters and non-defaulters. Therefore, all values different from zeros are changed to category of label 1 and the rest of the values are considered as a non-defaulters with a category of 0. To label the class attribute of "Loan_balance", the following python code is used

```
df2['Loan_Balance'] = df2['Loan_Balance'].apply(lambda x: 1 if x != 0 else x)
```

After labeling the class label the following count of both the defaulters (with a class label of 1) and non-defaulters (with a class label of 0) are obtained.

```
df["Loan_Balance"].value_counts()
```

```
1    848326
0    321581
```

In the below image, It is shown that the distribution of the target variable "Loan balance" and with this distribution the majority of the class 73% with a label of 1 is those defaulter and minority of the distribution 23% with a label value 0 is a non-defaulters.

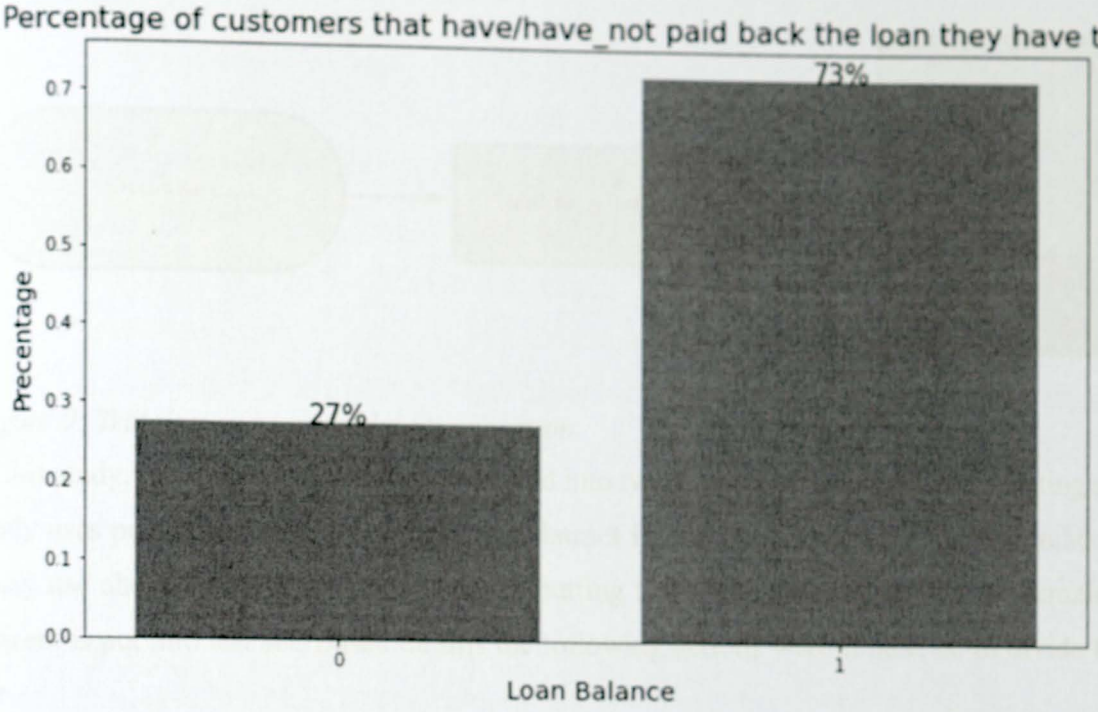


Figure 8: Defaulters and non-defaulters in percentage

3.10 Forming Features and Target Matrices

Following class labeling process the matrix of features to be used as input variables as well as the target variable will be taken into account for model building is formed. There for the input variable features and the target matrices is taken into x and y variables as follows.

```
x=df.drop('Loan_Balance',axis=1)
y=df['Loan_Balance']
```

Definition of training and testing sets

A training set is a portion of a data set used to fit (train) a model for prediction or classification of values that are known in the training set, but unknown in other (future) data. The training set is used in conjunction with validation and/or test sets that are used to evaluate different models[18].

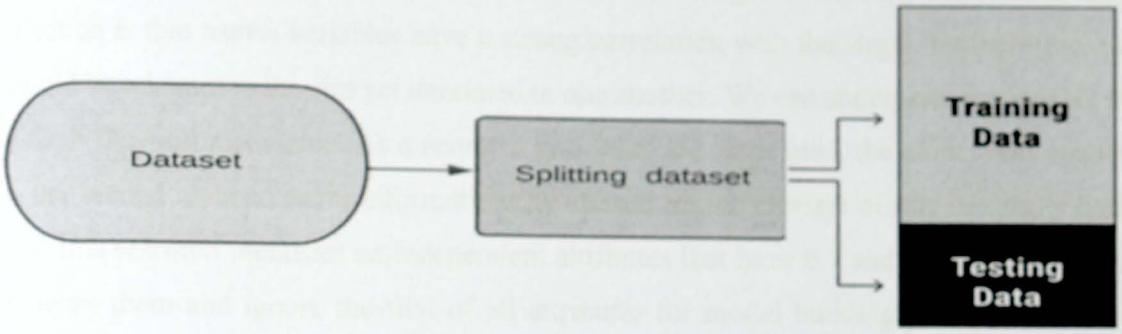


Figure 9: Training and testing data preparation

In this study, the preprocessed data are divided into two pieces, a training set and a testing set. The study uses percentage split for splitting the dataset into train and test set. The researcher in this study use about 80 percent of the data and putting them into training set and the remaining 20 percent is put into test set. Based on this the following actions were conducted to divide the data set

For separating dependent independent data

```
X = df.drop("Loan_Balance",axis=1)
y = df["Loan_Balance"]
```

For dividing data to train and test data

```
from sklearn.model_selection import train_test_split

# Splitting the data into train and test using train_test_split function
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=422,stratify=y)
X_train.shape, X_test.shape
```

3.11 Feature selection Process

Correlation is needed to drop one of the highly correlated independent attributes because keeping both attributes could result in multi collinearity, which could lead to unstable and unreliable model estimates and make it challenging to understand the impact of each independent variable on the dependent variable.

Pearson's correlation method is selected because it is a well-known filter method that may determine the direction and degree of a linear relationship between two variables. The linear relationship between two or more variables is measured via correlation. We can predict one

variable from the other using correlation. The reasoning for utilizing correlation for feature selection is that useful variables have a strong correlation with the target. Furthermore, variables should be relevant to the aim yet unrelated to one another. We can anticipate one variable from the other if they are associated. As a result, if two traits are associated, the model only requires one, as the second adds no more information[20]. Based on the concept of this Pearson's correlation rule, this research identifies an independent attributes that have 0.7 and higher correlation values between them and ignore the first of all attributes for model building process. To do this first dependent and independent variables were separated and this variables were divided to train and test data with a ratio of 80:20 as indicated above and then the correlation of the training data is calculated. Based on this method a list of independent attributes with a name 'INIT_LOAN_POUNDAGE', 'INTER_AMT_ETB', 'INTER_USG_MINUTE', and 'LOAN_POUNDAGE' are dropped from the dataset. This ignoring of the highly correlated attribute data is also applied to the test data similarly.

The following python code were used to do all the above process

For finding the correlation between attribute of the training data

```
X_train.corr()
```

For selecting highly correlated features and to remove the first feature that is correlated with the other feature the following function is used

```
def correlation(dataset, threshold):
    col_corr = set()  # Set of all the names of correlated columns
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if (corr_matrix.iloc[i, j]) > threshold:
                colname = corr_matrix.columns[i]  # getting the name of column
                col_corr.add(colname)
    return col_corr

corr_features = correlation(X_train, 0.7)
len(set(corr_features))
```

A list of the correlated features with a threshold value of 0.7 and above are displayed as follows

```
corr_features
```

```
{'INIT_LOAN_POUNDAGE', 'INTER_AMT_ETB', 'INTER_USG_MINUTE', 'LOAN_POUNDAGE'}
```

Finally correlated data are dropped from both the training and test data of columns and to drop those correlated attributes from both training and test data, the following code is used.

```
X_train.drop(corr_features,axis=1)
X_test.drop(corr_features,axis=1)
```

Therefore after dropping the correlated data attributes from the training dataset, the sample of the preprocessed data is shown in the below table.

CUST_AGE	INIT_LOAN_AMT	LOAN_AMT	REPAY_AMT	REPAY_POUNDAGE	RECHARGE_AMNT	OFF_PEAK_USG_MINUTE	OFF_PEAK_AMT_ETB	P
42	25	0	10	2	25	73	6128	
20	5	4	1	0	5	0	0	
44	5	5	0	0	45	739	0	
34	15	4	10	0	60	0	0	
60	10	0	0	1	20	0	0	

3.12 Model selection

The type of algorithm that will work well is determined by different data sets with varying types of variables and the amount of instances. According to the no theorem of [21], there is no single learning method that will outperform other algorithms across all data sets. At this step, the best performing learning algorithm is studied, which depends on the type of problem to be solved and the type of data at hand. In this work, classification algorithms are taken into account. Random Forest, Logistic Regression, Nave Bayes, K-nearest neighbor, and SVM are the classification techniques used. These algorithms were selected based on their strengths in handling the specific characteristics of credit risk prediction in the telecom industry based on previous studies.

3.13 Training

Here using supervised classification algorithm described before, the model has been trained using the data to enhance its capability. The dataset is separated into two parts: training and testing. For this study, the training/testing split is about 80/20. The training dataset is used for training purpose and the testing dataset is used for the testing purpose. Training dataset is fed to the learning algorithm. And this learning algorithm finds a mapping between the input and the output and generates the model.

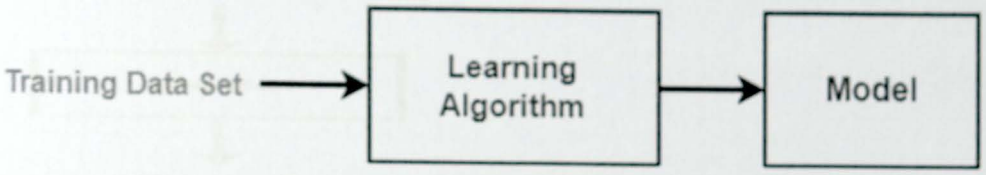


Figure 10: Training a machine learning model using training data

3.13.1 Tool used

For this study Python is used to carry out experimentation to present the work. Python offers a number of machine learning algorithms that make the implementation process simpler. One can create and integrate their own Java code with it, and it offers several libraries and frameworks that provide GUI capabilities and can be used in conjunction with machine learning. It is free software that may be downloaded using The Python Software Foundation License (PSF License) agreement[21].

The methods used for this airtime credit investigation can be summed up as follows. The dataset from ethio telecom is collected in the first phase. The most relevant attributes/fields/features are determined by asking the experts in the subject. Also different study are gathered to identify and determine the relevant features of a customer based data in telecommunication industries. Brute-force method which is the simplest one that separates the most relevant/informative qualities by measuring everything accessible is also used. Additionally, a feature selection process is used to find and remove as many redundant, irrelevant, and unnecessary attributes as possible. In this study, some data preprocessing is also done. Since the dataset contains a missing feature values, and some categories that need to be transformed into dummy variables, preprocessing is used. To conduct the experimentation and show the work, a Python is employed.

3.14 Evaluation

Model evaluation is the process of analyzing the performance and strengths and weaknesses of a machine learning model using a variety of evaluation metrics. The preserved testing dataset is used to assess the model's performance at this stage. Through this evaluation, the model can be tested against data that has never been used for training[22]. For this study, the Performance is evaluated using metrics including recall, precision and accuracy and F-measure. To calculate such metrics confusion matrices is required.

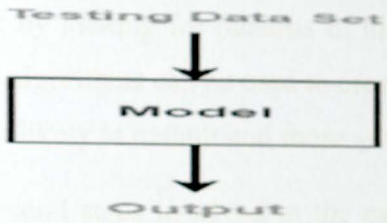


Figure 11: Model evaluation using test data

CHAPTER FOUR

EXPERIMENT

4.1 Overview

Using historical data and machine learning algorithms, it is possible to predict the risk of airtime credit default by looking for patterns in historical data. In order to accomplish this, supervised classification algorithms can be used to create models that can precisely categorize customers into those who are likely to default and those who are not.

There are several steps involved in the experimentation process with supervised classification algorithms. To clean, transform, and balance the dataset, data preparation is done first as described before. Following that, the prepared dataset will be used to train a number of supervised classification algorithms, including Naive Bayes, Random Forest, Logistic Regression, and KNN algorithms. After the models have been trained, performance of each model is evaluated using metrics like accuracy, precision, recall, and f1-score. The true positive, false positive, true negative, and false negative rates for each model are also displayed using the confusion matrix. At the conclusion of the experiment, the final model for estimating airtime credit risk is chosen based on its performance.

4.2 Experimental setup

In this study, a comprehensive experimental setup was conducted to evaluate the performance of four supervised classification algorithms (Logistic Regression, Random Forest, KNN, and Naive Bayes) in predicting the risk of airtime credit.

In the first phase, four experiments were conducted using the initial unbalanced dataset. This allowed for a thorough assessment of the algorithms' predictive capabilities without any data balancing techniques.

Moving to the second phase, eight experiments were carried out on balanced datasets using two data balancing techniques: under-sampling and over-sampling. Each of the four classification

algorithms was applied to these balanced datasets to assess how well the data balancing techniques enhanced their performance. This phase aimed to determine the effectiveness of the balancing techniques in improving the algorithms' ability to predict credit risk accurately.

In addition to the detailed experimentation in the first and second phases, the study also included additional experiments for comparison purposes. This experiment (experiment 13) focused on evaluating the performance of the machine learning algorithms using a different dataset attributes previously used by other researchers. This provided an opportunity to compare the results obtained from the previous attributes with the outcomes achieved on the new dataset, providing valuable insights into the generalizability and robustness of the algorithms.

Furthermore, a separate experiment (experiment 14) was conducted on the new dataset after removing certain attributes that were utilized in creating the class labels without any balancing techniques and compared the result with the experiment before removing the attributes. This allowed for an investigation into the impact of excluding specific attributes on the performance of the classification algorithms, shedding light on the significance of these attributes in predicting credit risk accurately.

By employing this comprehensive experimental setup, the study aimed to identify the most effective algorithm for constructing a credit risk prediction model and to assess the impact of different data balancing techniques and attribute selection on the algorithms' performance. Overall, the experimental setup comprised detailed investigations in the first and second phases, while the additional experiments were conducted for the purpose of comparison and analysis.

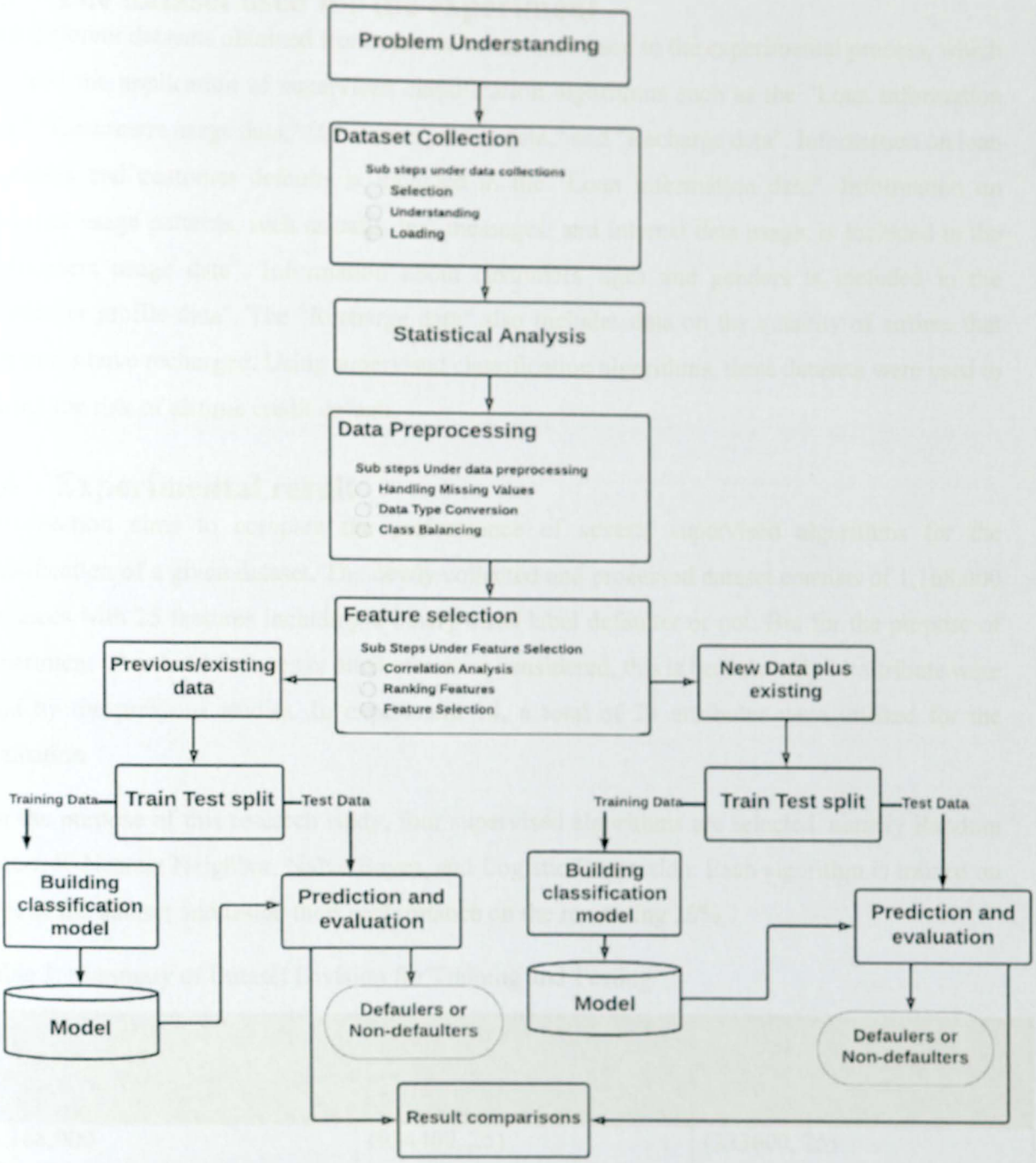


Figure 12: Experimental design of the research

4.3 The dataset used for the experiment

Four different datasets obtained from ethio telecom were used in the experimental process, which involved the application of supervised classification algorithms such as the "Loan information data," "Customers usage data," "Customer profile data," and "Recharge data". Information on loan payments and customer defaults is included in the "Loan information data". Information on customer usage patterns, such as calls, text messages, and internet data usage, is included in the "Customers usage data". Information about customers' ages and genders is included in the "Customer profile data". The "Recharge data" also includes data on the quantity of airtime that customers have recharged. Using supervised classification algorithms, these datasets were used to predict the risk of airtime credit default.

4.4 Experimental result

This section aims to compare the performance of several supervised algorithms for the classification of a given dataset. The newly collected and processed dataset consists of 1,168,000 instances with 25 features including a binary class label defaulter or not. But for the purpose of experiment 13 only 11 features or attributes were considered, this is because only 11 attribute were used by the previous studies. In experiment 14, a total of 24 attributes were utilized for the evaluation

For the purpose of this research study, four supervised algorithms are selected, namely Random Forest, K-Nearest Neighbor, Naïve Bayes, and Logistic Regression. Each algorithm is trained on 80% of the dataset and tested their performance on the remaining 20%.

Table 3: Summary of Dataset Division for Training and Testing

Total Number of Instance	Number of Train Data (Rows, columns)	Number of Test Data (Rows, columns)
1,168,000	(934400, 25)	(233600, 25)

To evaluate the performance of each algorithm, this research uses various metrics such as accuracy, precision, recall, and F1-score.

4.4.1 Random Forest:

Three experiments are conducted using random forest such as modeling using the original imbalanced data, under-sampled data, and over sampled data. The result obtained in each are presented as follows.

Experiment 1: Random Forest model before applying the class balance technique

Random Forest is implemented using the imbalanced data that was partitioned as train and test data in the previous and the output is indicated below. Based on that data, the random forest model's output shows that it had an accuracy of 0.99, correctly classifying 99.99 of the instances in the dataset. The model correctly identified almost all instances of each class with producing a small false positives and no false negatives, and this is also evidenced by the precision, recall, and F1-score for both classes (0 and 1) being the highest. The model's strong performance is also supported by the fact that the macro-averaged and weighted-average F1-scores are both 0.99.

The accuracy is 0.99994

Classification Result:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	64205
1	1.00	0.99	0.99	169395
accuracy			0.99	233600
macro avg	0.99	0.99	0.99	233600
weighted avg	0.99	0.99	0.99	233600

According to the confusion matrix analysis, the model correctly classified every single one of the 64,205 instances of class 0 (True Positive). The model classified 14 of the instances of class 1 as class 0 (False Positive), while correctly classifying 169,381 of them (True Negative).

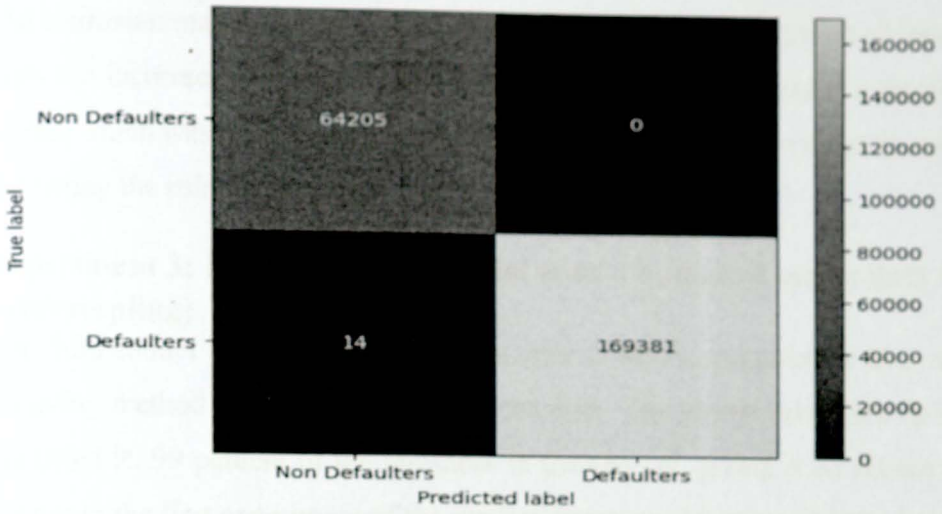


Figure 13: confusion matrix analysis of Random Forest model before applying class balance technique

Experiment 2: Random Forest model with a balanced target (using Under-sampling)

The accuracy of the random forest output using balanced data by applying under-sampling technique is slightly less than the accuracy of the random forest model with unlabeled data. This suggests that even though accuracy is not as accurate as the first model, it still performs well in predicting the class labels. The classification report shown in the table reveals that, although slightly lower than the first model, the precision, recall, and f1-score for the minority class are still very high as shown below.

```
The accuracy is  0.9910488013698631
classification result
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     64205
           1       1.00      0.99      0.99    169395

    accuracy                           0.99    233600
   macro avg       0.98      0.99      0.99    233600
weighted avg       0.99      0.99      0.99    233600

confussin matrix result
[[ 64205      0]
 [  2091 167304]]
```

The confusion matrix also indicate that the model correctly identified all instances of the minority class and incorrectly identified some instances of the majority which is the defaulters. Overall, the model, which was trained using the under-sampling technique, is accurate and performs well when predicting the minority and majority class.

Experiment 3: Random Forest model with a balanced target data (using oversampling)

The third model is a random forest classifier is also experimented after applying model over-sampling method to handle the unbalanced data. The model result shown below table correctly predicted 99.99 percent of the instances in the test set, giving it an accuracy of 0.9999 which is similar to the first experiment of the random forest model using imbalanced data.

The accuracy is 0.99993

Classification Result:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	64205
1	1.00	1.00	1.00	169395
accuracy			0.99	233600
macro avg	0.99	1.00	0.99	233600
weighted avg	0.99	0.99	0.99	233600

Confusion Matrix Result:

[[64205 0]

[15 169380]]

The classification outcomes demonstrate that precision and recall for both classes (0 and 1) are high, demonstrating that the model correctly identified every instance of both classes. The model's

precision and recall are perfect for both classes, as shown by the fact that the f1-score for class 0 is 0.99 and class 1 is 1.00.

To conclude based on the result of the three experiments, using unbalanced data, the first experiment produced a highly accurate model that correctly identified almost every instance of both classes. The accuracy was slightly lower in the second experiment, which balanced the target class using under-sampled data, but it still performed well in predicting class labels. A highly accurate model similar to the first experiment, with perfect precision and recall for both classes was produced in the third experiment by using over-sampling to balance the target class. Overall, these experiment show both before applying class balance technique and after applying over-sampled technique has a similar and slightly higher accuracy than the under-sampled model experiment.

4.4.2 Logistic Regression:

A logistic regression model is also implemented using the same dataset used in the random forest algorithm. Modeling with the original unbalanced data, under-sampled data, and over-sampled data were all tested using logistic regression also. Following is a presentation of the findings from each experiment.

Experiment 4: Logistic Regression model before applying class balance technique

The results of the logistic regression model indicate that the model performed fine, though slightly less accurate than the previous random forest model. The model correctly classified about 93.9% of instances, according to the reported accuracy of 0.939. The classification report offers details on each class's precision, recall, and F1-score (0 and 1).

The precision for Class 0 (non-defaulters) is 0.88, meaning that 88 percent of instances that were predicted to be in Class 0 were correctly identified. The recall is 0.91, indicating that 91% of all instances were correctly classified as Class 0 by the model. The harmonic mean of recall and precision, or the F1-score, is 0.89 for Class 0.

```
THe accuracy is  0.939486301369863
classification result
              precision    recall  f1-score   support

           0       0.88      0.91      0.89     64205
           1       0.96      0.95      0.96    169395

    accuracy                           0.94    233600
   macro avg       0.92      0.93      0.92    233600
weighted avg       0.94      0.94      0.94    233600
```

A precision of 0.96 for Class 1 indicates that 96 percent of instances predicted as Class 1 (defaulters) were correctly classified by the model. The recall is 0.95, indicating that the model correctly classified Class 1 for 95% of all instances.

The model's performance for each class is thoroughly broken down in the confusion matrix. A total of 58,290 instances for Class 0 (Non-defaulters) were correctly predicted by the model, but 5,915 instances were incorrectly assigned to Class 1 (false negatives). While the model correctly identified 161,174 instances of Class 1 as true Negatives, it incorrectly labeled 8,221 instances as Class 0 (false positives).

In conclusion, the logistic regression model has a good accuracy as the random forest model did but with a marginally lower level of accuracy of random forest. For both classes, it displays a respectably high F1-score, recall, and precision, indicating efficient classification abilities. The number of incorrect classifications is higher than with the random forest model, with false negatives for Class 0 and false positives for Class 1 being the most common.

The plot of the confusion matrix below displays how many of the model's predictions were right and wrong when compared to the test set's actual results. The result of this confusion matrix can be used to calculate metrics such as accuracy, precision, recall, and F1 score, which are described above that are used to evaluate the performance of a classification model.

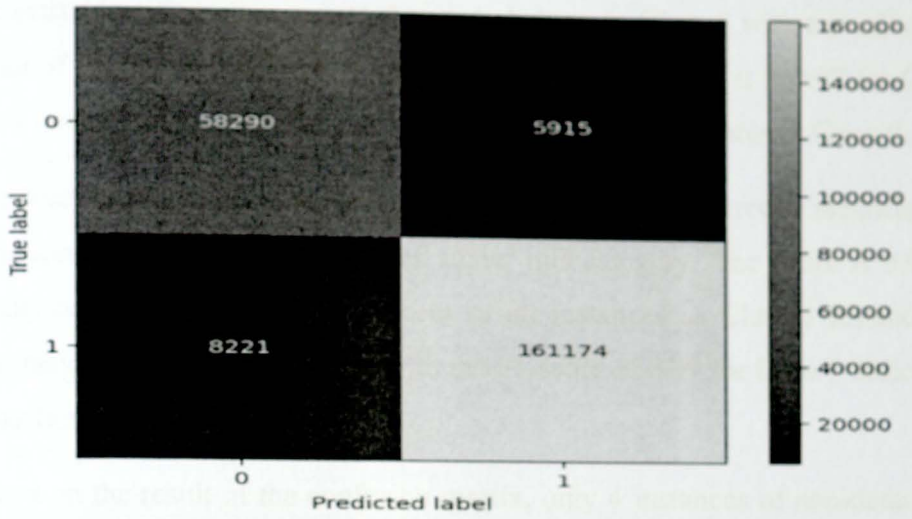


Figure 14: confusion matrix analysis of Logistic Regression before applying class balance technique

Experiment 5: Logistic Regression model with a balanced target (using under-sampling)

The outcome of the logistic regression model after using under-sampling strategies to balance the classes suggests that the model performs significantly better than the prior logistic regression model. According to the model's increased accuracy, instances were correctly classified in 98.28 percent of cases as shown below.

```
The accuracy is  0.9827953767123287
classification result
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     64205
           1       1.00      0.98      0.99    169395

    accuracy                           0.98    233600
   macro avg       0.97      0.99      0.98    233600
weighted avg       0.98      0.98      0.98    233600

confussin matrix result
[[ 64201      4]
 [  4015 165380]]
```

The classification report provides that, the precision for non-defaulters (Class 0) is 0.94, meaning that 94% of instances that were predicted to be in Class 0 were correctly identified. The model

correctly identified all instances that truly belong to Class 0 with a recall of 1.00. The harmonic mean of recall and precision is known as the F1-score, and it is 0.97 for Class 0. These metrics show exceptional performance in correctly classifying instances as Class 0.

The precision for Class 1 is 1.00, meaning that the model correctly identified Class 1 for 100% of instances that were predicted to fall under this category. The recall is 0.98, indicating that the model correctly identified 98 percent of all instances as Class 1 instances. A high degree of accuracy in classifying instances with the F1-score of 0.99 for Class 1 indicates that it is a Class 1 classification.

Based on the result of the confusion matrix, only 4 instances of non-defaulters or Class 0 (false negatives) were incorrectly classified as defaulters (Class 1) by the model, which correctly predicted 64,201 instances of the class. The model correctly identified 165,380 instances of defaulters (Class 1) and misclassified 4,015 instances as Class 0 (false positives) in Class 1.

In conclusion, the logistic regression model performs noticeably better than the prior logistic regression model after balancing the classes using under-sampling techniques. It performs well in terms of accuracy, recall, precision, and F1-score for both classes. The majority of instances in both classes are correctly identified by the model, leading to a few misclassifications. These findings indicate that under-sampling strategies can effectively address class imbalances and improve the model's ability to classify data.

Experiment 6: Logistic Regression model with a balanced target data (Over-sampling)

After balancing the classes using over-sampling techniques, the logistic regression model result shows a significant improvement in performance over the prior logistic regression model without applying any class balance techniques. The model's accuracy has increased indicating that roughly 97.97 percent of instances were correctly classified.

```
The accuracy is  0.9796618150684931
classification result
              precision    recall  f1-score   support

           0       0.93      1.00      0.96     64205
           1       1.00      0.97      0.99    169395

    accuracy                           0.98    233600
   macro avg       0.97      0.99      0.98    233600
weighted avg       0.98      0.98      0.98    233600

confussin matrix result
[[ 64020    185]
 [  4566 164829]]
```

The Python output provides a classification report for each class (non-defaulters and defaulters), in terms of precision, recall, and F1-score metrics in-depth. Accordingly, a precision of 0.93 for Class 0 indicates that 93% of instances predicted to be in Class 0 were correctly identified. The recall is 1.00, indicating that all instances that actually belong to Class 0 were recognized by the model. For Class 0, the F1-score, which represents the harmonic mean of precision and recall, is 0.96. According to these metrics, instances that belong to Class 0 can be identified with a high degree of accuracy. On the other hand, the precision for Class 1 is 1.00, meaning that the model correctly identified Class 1 for 100% of instances that were predicted to fall under this category. The recall is 0.97, indicating that the model correctly classified Class 1 in 97 percent of all instances. The F1-score for Class 1 is 0.99, which indicates excellent accuracy in classifying instances.

The model's effectiveness for each class is thoroughly broken down in the confusion matrix. The model classified 185 instances as Class 1 (false positives) while correctly predicting 64,020 instances for Class 0. The model correctly identified 164,829 instances for Class 1 and misclassified 4,566 instances as Class 0 (false negatives).

In conclusion, the logistic regression model performs significantly better than the unbalanced logistic regression model and slightly less than under-sampled method model after balancing the classes using over-sampling techniques. For both classes, it achieves high accuracy and displays excellent precision, recall, and F1 score. The majority of instances in both classes are correctly identified by the model, leading to a few misclassifications. These findings show that class

imbalances can be addressed and the model's ability to classify data is improved by over-sampling techniques.

4.4.3 Naïve Bayes classifiers

The original imbalanced data, the under-sampled data, and the over-sampled data were also used in an experiment to create a model using Naive Bayes. The implemented naïve bays classifier model before applying any class balance technique using the training and test data has the following result and each metrics are described below

Experiment 7: Naïve Bayes classifiers model before applying class balance technique

The outcome of the Naive Bayes classifier model without the use of class balance techniques points to a moderate level of success in classifying the data. According to the reported result of the model output, 82.67% of instances were correctly classified.

```
THe accuracy is  0.8266823630136987
classification result
              precision    recall  f1-score   support

           0       0.77      0.53      0.63     64205
           1       0.84      0.94      0.89    169395

    accuracy                           0.83    233600
   macro avg       0.80      0.73      0.76    233600
weighted avg       0.82      0.83      0.82    233600
```

When we describe the detail of the classification result for each class, the precision for non-defaulter (Class 0) is 0.77, indicating that 77 percent of instances that were predicted to be in Class 0 were correctly identified. The recall is 0.53, showing that the model correctly classified Class 0 in 53% of all instances. The harmonic mean of recall and precision, known as the F1-score, is 0.63 for Class 0. According to these metrics, Class 0 instances are correctly identified with a relatively lower rate of accuracy.

When we come to the second class which is the defaulters, the precision is 0.84, meaning that 84% of instances that were predicted to be in Class 1 (defaulters) were correctly classified by the model. The recall is 0.94, indicating that the model correctly identified 94% of all instances as being in Class 1.

As we can see from the confusion matrix result, the model classified 30,139 instances as Class 1 (false negatives) for Class 0, while correctly predicting 34,066 instances for Class 0. In terms of Class 1, the model correctly predicted 159,047 instances while misclassifying 10,348 instances as Class 0 (false positives).

In conclusion, the Naive Bayes classifier model achieves a moderate level of performance in classifying the data without the use of class balance techniques. It struggles to accurately identify instances as Class 0 because of its relatively lower Class 0 precision, recall, and F1 score. With high precision, recall, and F1-score, the model performs well in predicting instances that belong to Class 1. Overall performance appears to be acceptable, but there is a chance for improvement, especially in correctly classifying instances that belong to Class 0. The confusion matrix plot is given below in figure 15.

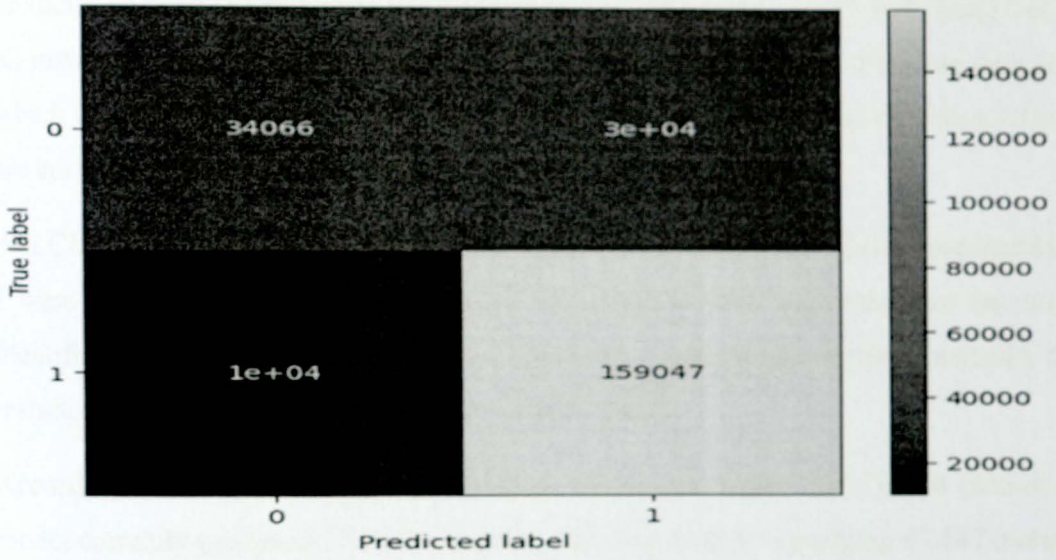


Figure 15 : confusion matrix analysis of Naive Bayes classifier before applying class balance technique

Experiment 8: Naïve Bayes classifiers model with a balanced target (Under-sampling)

The Naive Bayes classifier experiment result based on the under-sampling technique of class balance consists of performance metrics from a Naive Bayes classifier model used on a dataset. After using under-sampling techniques to balance the classes, the Nave Bayes classifier model result suggests relatively lower performance in classifying the data. 68.23% of instances were classified correctly, according to the reported accuracy of 0.68.

```
The accuracy is  0.6822902397260274
classification result
              precision    recall  f1-score   support

           0       0.39      0.27      0.31     64205
           1       0.75      0.84      0.79    169395

    accuracy                           0.68    233600
   macro avg       0.57      0.55      0.55    233600
weighted avg       0.65      0.68      0.66    233600

confussin matrix result
[[ 17018  47187]
 [ 27030 142365]]
```

Based on the result, the precision for Class 0 is 0.39, meaning that only 39% of instances that were predicted to be in Class 0 were correctly identified. The recall is 0.27, indicating that only 27% of all instances were correctly classified as Class 0 by the model. The F1-score for Class 0 is 0.31, which is the harmonic mean of precision and recall. According to these metrics, Class 0 instances are correctly identified with a relatively lower rate of accuracy.

For Class 1, the precision is 0.75, meaning that 75% of instances that were predicted to be in Class 1 were correctly classified by the model. The recall is 0.84, indicating that the model correctly identified 84% of all instances as Class 1 instances. A reasonable degree of accuracy in classifying instances according to the F1-score for Class 1 is 0.79.

According to the confusion matrix result of the model output, for Class 0 (non-defaulters), the model correctly predicted 17,018 instances while incorrectly classifying 47,187 instances as Class 1 (false negative). The model correctly identified 142,365 instances for Class 1 and incorrectly labeled 27,030 instances as Class 0 (false positives) in Class 1.

In conclusion, the Nave Bayes classifier model performs relatively worse at classifying the data than the original model after using under-sampling techniques to balance the classes. It did not correctly identify instances that belong to Class 0, as evidenced by its lower Class 0 precision, recall, and F1 score. But the model predicts instances that belong to Class 1 with higher precision, recall, and F1-score. The overall accuracy indicates limited success in accurately classifying the data, especially in correctly classifying instances belonging to Class 0.

Experiment 8: Naïve Bayes classifiers model with a balanced target data (Over-sampling)

This output is the result of a Naive Bayes classifier model that was trained on a dataset that had the class distribution balanced using oversampling methods. After balancing the classes with oversampling techniques, the Naive Bayes classifier model result accuracy is 68.1%.

```
The accuracy is  0.6808818493150685
classification result
              precision    recall  f1-score   support

           0       0.46      1.00      0.63     64205
           1       1.00      0.56      0.72    169395

    accuracy                           0.68    233600
   macro avg       0.73      0.78      0.68    233600
weighted avg       0.85      0.68      0.69    233600

confussin matrix result
[[64205     0]
 [74546 94849]]
```

The classification report based on performance measure metrics like precision, recall, and F1-score of each class (non-defaulter and defaulters) is described in detail as follows. For Class 0 (non-defaulters), the precision is 0.46, meaning that out of all instances predicted to be in Class 0, 46% were correctly classified. The recall of Class 0 is 1.00, indicating that all instances that actually belong to Class 0 were recognized by the model. The harmonic mean of recall and precision, or the F1-score, is 0.63 for Class 0. When classifying instances as Class 0, these metrics demonstrate a good level of accuracy.

For Class 1, the precision is 1.00, meaning that 100% of instances predicted as Class 1 were correctly classified by the model. The model correctly identified 56 percent of all instances as being in Class 1 (recall), according to this estimate. The Class 1 F1-score is 0.72, which indicates a fair degree of accuracy in classifying instances.

The confusion matrix result shows that there were no false negatives in the model's predictions for Class 0, which included all 64,205 instances. The model correctly identified 94,849 instances for Class 1 and misclassified 74,546 instances as Class 0 (false positives) in that category.

As a result of using over-sampling techniques to balance the classes, the Naive Bayes classifier model performs inconsistently when classifying the data. It successfully identifies instances that belong to Class 0 with a high degree of precision and recall. Lower recall and F1-score, on the other hand, show that the model is less successful at predicting instances that belong to Class 1. The overall accuracy indicates a modest level of success in accurately classifying the data, but there is still room for improvement, especially in correctly identifying instances that belong to Class 1. To improve the model's performance, additional research and examination of alternative class balance methods or model adjustments may be required.

In summary, the Naive Bayes classifier model was tested with different class balance techniques. Without any techniques, the model achieved moderate performance but struggled to identify instances in Class 0 due to lower precision, recall, and F1 score. Both Under-sampling techniques and over-sampling techniques resulted in worse overall performance.

4.4.4 K-Nearest Neighbors

An experiment using K-Nearest Neighbors were conducted with, imbalanced data, under-sampled balanced data and over-sampled balanced data. The results of all the tree experiment are discussed as follows.

Experiment 10: K-Nearest Neighbors model before applying class balance

The KNN (K-Nearest Neighbors) model was applied first to an unbalanced dataset without the use of any class-balancing techniques. The model's accuracy is 0.903, which shows that the model is performing well in predicting the outcomes without balancing the dataset as shown below.

```
The accuracy is  0.9034417808219178
classification result
              precision    recall  f1-score   support

           0       0.82      0.82      0.82     64205
           1       0.93      0.93      0.93    169395

    accuracy                           0.90    233600
   macro avg       0.88      0.88      0.88    233600
weighted avg       0.90      0.90      0.90    233600

confussin matrix result
[[ 52908  11297]
 [ 11259 158136]]
```

We can see from the classification report that class 0 has a precision of 0.82, which indicates that 82 percent of the predicted instances of class 0 are accurate. The model can recognize 82 percent of the instances of class 0 in the dataset, as indicated by the fact that the recall for class 0 is also 0.82. The harmonic mean of recall and precision yields 0.82 as the f1-score for class 0.

For the case of class 1, the precision for class 1, which is 0.93, indicates that 93 percent of the predicted instances of class 1 are accurate. The model can recognize 93 percent of the instances of class 1 in the dataset, as indicated by the recall for class 1 is 0.93. The model is doing very well at predicting class 1 based on the f1-score of 0.93 for class 1.

The confusion matrix demonstrates that the model predicts both classes quite well. It accurately predicts 158136 instances of class 1 and 52908 instances of class 0, respectively. On the other hand, it predicts 11259 instances of class 0 as class 1 and 11297 instances of class 1 as class 0, both incorrectly.

In conclusion, the KNN model works remarkably well at predicting the results when used on an unbalanced dataset. It is capable of correctly classifying instances into each of the two classes because it has high precision and recalls for each class. The model predicts correctly on the provided dataset without using any class balancing techniques, as shown by its overall accuracy of 0.903, which shows its effectiveness.

Experiment 11: K-Nearest Neighbors model applying class balance (Under-sampling)

An under-sampled dataset that has been balanced using the KNN (K-Nearest Neighbors) model has been also used. After balancing the dataset using under-sampling, the model's accuracy is 0.732, indicating that it performs moderately well but drastically decreased when compared to the previous model KNN.

```
The accuracy is  0.7316909246575343
classification result
              precision    recall  f1-score   support

           0       0.51      0.95      0.66     64205
           1       0.97      0.65      0.78    169395

    accuracy                           0.73    233600
   macro avg       0.74      0.80      0.72    233600
weighted avg       0.84      0.73      0.75    233600

confussin matrix result
[[ 60775   3430]
 [ 59247 110148]]
```

According to the classification report, class 0 has a precision of 0.51, meaning that 51 percent of the predicted instances of this class are in fact true. Since the model can recognize 95% of the instances of class 0 in the dataset, class 0 has a recall of 0.95. The precision for class 1 is 0.97, which indicates that 97 percent of the instances of class 1 that are predicted are actually true. The model can recognize 65 percent of the instances of class 1 in the dataset, as indicated by the recall for class 1 which is 0.65. Since the model is successful in predicting class 1 with an f1-score of 0.78, this class is predicted well by the model.

The confusion matrix demonstrates that the model is somewhat predicting both classes. It correctly predicts 60775 instances of class 0 and 110148 instances of class 1 data. However, it predicts 3430 instances of class 1 as class 0 and 59247 instances of class 0 as class 1 incorrectly. In conclusion, the KNN model is doing a fair job of predicting the results when applied to a dataset that has been balanced using the under-sampling technique.

Experiment 12: K-Nearest Neighbors model applying class balance (Over-sampling)

The KNN (K-Nearest Neighbors) model has been applied to a dataset that has been balanced using the over-sampling technique. The model's accuracy is 0.899, which shows that it is successfully forecasting the results after using over-sampling to balance the dataset.

```
The accuracy is  0.8988356164383562
classification result
              precision    recall  f1-score   support

           0       0.77      0.91      0.83     64205
           1       0.96      0.90      0.93    169395

    accuracy                           0.90    233600
   macro avg       0.86      0.90      0.88    233600
weighted avg       0.91      0.90      0.90    233600

confussin matrix result
[[ 58267   5938]
 [ 17694 151701]]
```

We can see from the classification report that the precision for non-defaulters (class 0) is 0.77, which indicates that 77 percent of the predicted instances of class 0 are actually true. The recall for class 0 is 0.91, which indicates that 91 percent of the instances of class 0 in the dataset can be recognized by the model. As the harmonic mean of recall and precision, the f1-score for class 0 is 0.83. In the case of class 1, the precision is 0.96, which indicates that 96% of the instances of class 1 that are predicted are in fact true. The recall for defaulters (class 1) is 0.90, indicating that 90% of the instances of class 1 in the dataset can be recognized by the model. The model is doing very well at predicting class 1 as evidenced by the f1-score of 0.93 for class 1.

The model's overall performance for both classes is represented by the 0.86 macro average of precision, recall, and f1-score. The overall performance is strong, as evidenced by the weighted average, which accounts for the class imbalance, of 0.91. The confusion matrix also demonstrates that the model does a good job of predicting both classes. It correctly predicts 151701 instances of class 1 and 58267 instances of class 0, respectively. It predicts 5938 instances of class 1 as class 0 and 17694 instances of class 0 as class 1 incorrectly.

The over-sampled model has the highest accuracy when compared to the results of the unbalanced and under-sampled models. In conclusion, the KNN model applied to a dataset balanced using the over-sampling technique is doing well in terms of outcome prediction. It can accurately classify instances into their respective classes because it has high precision and recall for both classes. The model successfully balances the dataset using over-sampling, as evidenced by its accuracy of 0.899, which demonstrates its efficacy.

In summary, a KNN model was used in three experiments to evaluate how well it performed with various class balance strategies. A high accuracy of 90% among all was displayed by the unbalanced dataset. However, after under-sampling, the accuracy dropped to 0.732. An accuracy of 0.899 after oversampling which is nearly similar to the first model of KNN, demonstrated that forecasting with a balanced dataset was successful. The following table lists the results of the supervised classification algorithms on various datasets (existing and newly added attributes) and balancing methods applied.

Table 4: A summarized result of the experimental setup and Accuracy obtained

No	Algorithm	Experimental Setup	Accuracy
1	Random Forest	Before Applying class Balancing	0.999
2	Random Forest	Balanced Target Data (Under-sampling)	0.991
3	Random Forest	Balanced Target Data (Over-sampling)	0.999
4	Logistic Regression	Before Applying class Balancing	0.939
5	Logistic Regression	Balanced Target Data (Under-sampling)	0.982
6	Logistic Regression	Balanced Target Data (Over-sampling)	0.979
7	Naïve Bayes	Before Applying class Balancing	0.826
8	Naïve Bayes	Balanced Target Data (Under-sampling	0.682
9	Naïve Bayes	Balanced Target Data (Over-sampling)	0.680
10	KNN	Before Applying class Balancing	0.903
11	KNN	Balanced Target Data (Under-sampling	0.731
12	KNN	Balanced Target Data (Over-sampling)	0.898

Experiment 13: Performance Evaluation of Machine Learning Algorithms: Novel/New Attributes versus Existing Attributes

The purpose of experimenting with the performance evaluation of machine learning algorithms in terms of novel/new attributes versus existing attributes is to assess the impact and effectiveness of incorporating new or additional features in the machine learning models. By comparing the performance of algorithms using novel/new attributes (features that were not previously considered or included) against algorithms using only existing attributes, researchers aim to determine the potential improvements or insights that can be gained from these new attributes. This experimentation helps in understanding whether the inclusion of novel attributes leads to better predictive performance, higher accuracy, improved precision, recall, or F1-score, among other metrics. It also allows us to identify the strengths and weaknesses of different algorithms

when trained on these different attribute sets.in consideration with this the result of the experiment is summarized in the Table 5 below.

Table 5: Performance Comparison of Algorithms Using Novel/New Attributes versus Existing Attributes

		Performance of evaluation metrics				
Dataset Used	**Algorithms**	**Accuracy**		**Precision**	**Recall**	**f1-score**
Dataset with new attributes distinct from previous study	Random Forest	98.40%	class 0	0.95	0.99	0.97
			class 1	1.00	0.98	0.99
	Logistic Regression	94.70%	class 0	0.91	0.90	0.90
			class 1	0.96	0.96	0.96
	Naïve Bayes	68.4%	class 0	0.47	1.00	0.64
			class 1	1.00	0.57	0.72
	KNN	97.3%	class 0	0.93	0.97	0.95
			class 1	0.99	0.97	0.98
Dataset consisting only of existing attributes from previous study	Random Forest	69.2%	class 0	0.41	0.27	0.32
			class 1	0.75	0.85	0.80
	Logistic Regression	72.5%	class 0	0.25	0.00	0.00
			class 1	0.73	1.00	0.84
	Naïve Bayes	72%	class 0	0.29	0.01	0.03
			class 1	0.73	0.99	0.84
	KNN	70.3%	class 0	0.45	0.35	0.40
			class 1	0.77	0.84	0.80

Experiment 14: Performance Evaluation of Machine Learning Algorithms before and after removing attributes that were utilized in creating the class labels.

The purpose of experimenting with the performance evaluation of machine learning algorithms before and after removing attributes used in creating the class labels is to assess the impact of attribute removal on the predictive power and performance of the models. By removing attributes that were utilized in creating the class labels, researchers can determine whether these attributes are redundant or unnecessary for the predictive task at hand. This experimentation helps in understanding the importance and contribution of specific attributes in the classification process. The primary goal is to evaluate how the removal of these attributes affects the performance metrics of the machine learning algorithms. By comparing the performance before and after attribute removal, researchers can determine if the removal improves the accuracy, precision, recall, F1-score, or other evaluation measures

Table 6: Performance Evaluation of Machine Learning Algorithms Before and After Attribute Removal

		Performance of evaluation metrics				
Dataset Used	**Algorithms**	**Accuracy**		**Precision**	**Recall**	**f1-score**
Before Removing the attributes	Random Forest	99.9%	class 0	0.99	1.00	0.99
			class 1	1.00	0.99	0.99
	Logistic Regression	93.9%	class 0	0.88	0.91	0.89
			class 1	0.96	0.95	0.96
	Naïve Bayes	82.6%	class 0	0.77	0.53	0.63
			class 1	0.84	0.94	0.89
	KNN	90.3.3%	class 0	0.82	0.82	0.82
			class 1	0.93	0.93	0.93
After Removing the attributes	Random Forest	80.4%	class 0	0.65	0.63	0.64
			class 1	0.86	0.87	0.87
	Logistic Regression	76.4%	class 0	0.65	0.32	0.43
			class 1	0.78	0.93	0.85
	Naïve Bayes	68.7%	class 0	0.47	0.91	0.62
			class 1	0.95	0.60	0.74
	KNN	75.3%	class 0	0.55	0.52	0.54
			class 1	0.82	0.84	0.83

4.5 Discussion of the result

The discussion section aims to provide a comprehensive analysis of the findings presented in the study regarding airtime credit risk prediction. The research questions of attribute suitability, suitable machine learning algorithms, and model performance are addressed. Furthermore, this discussion section critically evaluates the study's contributions and provides insights into the advancements made in comparison to previous works.

This study aimed to build a machine learning model that can predict airtime credit risk in ethio telecom by evaluating the performance of various machine learning algorithms. Specifically, the study assessed the accuracy, precision, recall, F1-score, and other metrics of Naive Bayes, Random Forest, Logistic Regression, and KNN, four supervised classification algorithms. To conduct the analysis, the study utilized an ethio telecom dataset comprising loan information, customer usage data, customer profile data, and recharge data.

The initial unbalanced dataset served as the basis for analysis in the first stage of the experiment, allowing to establish a baseline and identify any potential bias. Subsequently, the algorithms were re-evaluated in the second phase by employing under- and over-sampling techniques to produce balanced datasets.

The findings revealed significant variations in the performance of different algorithms and data balancing strategies. Based on their performance using both balanced and unbalanced datasets, the Random Forest algorithm stood out as the top performer among logistic regression, Naïve Bayes, and KNN.

The Random Forest model exhibited exceptional accuracy, precision, recall, and F1 scores. With an accuracy of 99.9%, it showcased robust performance for both classes, misclassifying only 15 instances of the positive class according to the confusion matrix. Even after applying under-sampling to balance the target data, the Random Forest model maintained an accuracy of 0.99, demonstrating high precision, recall, and F1 scores for both classes. However, there was a higher number of misclassifications for the negative class, totaling 2,091 instances. Over-sampling was also effective, resulting in an accuracy of 0.99 with only 15 instances of misclassification for the negative class.

Logistic Regression also yielded promising results, with an accuracy of 94% before applying any class balance techniques. Both classes demonstrated relatively high precision, recall, and F1 scores. However, 5,915 instances of the positive class and 8,221 instances of the negative class were misclassified according to the confusion matrix. The accuracy improved to 0.9828 after applying under-sampling to balance the target data, leading to a significant reduction in misclassifications. Similarly, oversampling enhanced the accuracy to 0.9797, with high precision, recall, and F1 scores for both classes. However, the confusion matrix indicated 4,566 instances of the negative class and 185 instances of the positive class that were misclassified.

The Naive Bayes classifier exhibited an accuracy of 0.8267 before employing any class balance methods. Although the precision, recall, and F1 scores were relatively high for the positive class, they were lower for the negative class. The confusion matrix revealed a considerable number of misclassifications for both classes. Applying under-sampling led to a drop in accuracy to 68.2%, accompanied by lower precision, recall, and F1 scores for both classes. Similarly, oversampling resulted in an accuracy of 68%, with higher precision, recall, and F1 scores for the negative class

but not for the positive class. These results indicate that Naive Bayes may not be the optimal solution for airtime credit risk prediction, as both under-sampling and oversampling techniques led to decreased performance.

K-Nearest Neighbors (KNN) achieved an accuracy of 90.3% without applying any class balance techniques. The F1 score, recall, and precision were relatively high for both classes. However, a significant number of misclassifications for both classes were evident in the confusion matrix. Under-sampling resulted in a decline in accuracy to 73%, accompanied by lower precision, recall, and F1 scores for both classes. Conversely, over-sampling improved the accuracy to 89.8%, with high F1 scores, recall, and precision for both classes. Compared to the original model, the confusion matrix revealed fewer misclassifications.

In comparison to other models, Naive Bayes demonstrated lower performance. The findings suggest that Naive Bayes may not be the best solution for airtime credit risk prediction due to the decreased performance observed with both under-sampling and oversampling techniques. On the other hand, K-Nearest Neighbors showed moderate to high performance depending on the class balance technique employed. Oversampling improved model accuracy and reduced misclassifications, while under-sampling led to decreased performance.

In addition to evaluating different attributes, the study experiment also aimed to assess the performance of machine learning algorithms using various dataset attributes. Specifically, the study compared the performance of novel/new attributes with existing attributes previously utilized by other researchers. Table 7 presents the evaluation metrics for four algorithms applied to both the new and existing attribute sets.

For the new attributes, Random Forest achieved the highest accuracy of 98.40%. It exhibited excellent precision and recall for both class 0 and class 1. Logistic Regression also performed well, achieving an accuracy of 94.70% with balanced precision, recall, and F1 scores. On the other hand, Naïve Bayes yielded a lower accuracy of 68.4% and demonstrated higher precision for class 1 but comparatively poorer performance in other metrics. KNN achieved an accuracy of 97.3% and demonstrated good precision and recall for both classes.

In contrast, when considering the existing attributes, the algorithms experienced a decline in performance compared to the new attributes. Logistic Regression achieved an accuracy of 72.5%

but exhibited low precision and recall values for class 0. Similarly, Naïve Bayes and KNN demonstrated reduced accuracy and inadequate precision and recall scores for class 0.

In Table 7, the experiment present the comparison results of the four algorithms with a combination of novel and previous attributes before and after the removal of certain attributes ('INIT_LOAN_AMT' and 'REPAY_AMT') used in creating class labels.

Before attribute removal, Random Forest achieved an impressive accuracy of 99.9% for classifying both classes, displaying high precision and recall values. Logistic Regression maintained an accuracy of 94.70%, with slightly lower precision and recall for class 0 compared to the initial results. Naïve Bayes exhibited consistent accuracy of 68.4% but displayed improved precision for class 0 after attribute removal. KNN maintained a similar accuracy of 97.3% and consistent precision and recall values for both classes. After attribute removal, Random Forest retained a high accuracy of 80.4%. However, Logistic Regression and Naïve Bayes showed a slight decline in accuracy. KNN maintained its accuracy at 70.3% and displayed consistent precision and recall values for both classes.

These results highlight the importance of attribute selection in machine learning algorithms. The performance evaluation on new attributes demonstrated superior results compared to the existing attributes. Attribute removal had varying effects on the algorithms, with Random Forest showing robust performance regardless of attribute changes. In contrast, Logistic Regression and Naïve Bayes experienced minor fluctuations in accuracy after attribute removal, while KNN maintained consistency.

In general in the conducted experiment, it was observed that the utilization of new attributes led to improved airtime credit prediction compared to the previously used attributes. Additionally, the combination of both new and previously used attributes further enhanced the performance of the prediction model. These findings suggest that incorporating new attributes and leveraging the existing ones together can significantly boost the accuracy and reliability of airtime credit prediction models.

It has been found through deep examination and testing that the random forest algorithm performs remarkably well in this study. Once random forest algorithm is selected based on its performance, the researcher attempts to learn more about the algorithm's inner workings by ranking the significance of the features it makes use of. Feature importance refers to a measure of the individual contribution of the corresponding feature for a particular classifier[58].

We were able to determine the relative importance of each feature in terms of how it affected the algorithm's ability to predict outcomes through this procedure. We were also able to determine the main factors and develop a deeper knowledge of the underlying patterns and relationships in the dataset by giving importance scores to the features. This research helps with feature selection by guiding us toward the most important variables and offers insightful information about how the random forest method operates. In this study, features are ranked using the "feature_importances" function from the Python sklearn package to identify the components that contributed to model performance. Figure 16 shows feature importance ranking of the attributes.

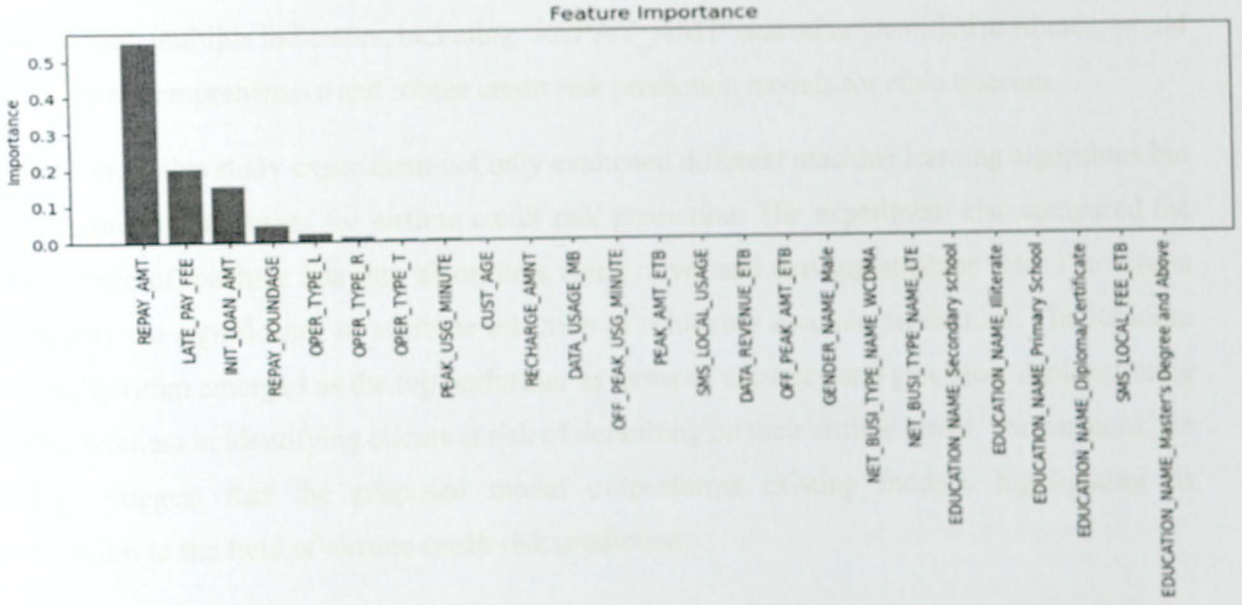


Figure 17: Feature importance of the selected algorithm

The uniqueness of the variables used in this research compared to previous studies lies in their ability to effectively predict airtime credit risk. The identified top features, namely REPAY_AMT, LATE_PAY_FEE, INIT_LOAN_AMT, REPAY_POUNDAGE, OPER_TYPE_L,

OPER_TYPE_R, and OPER_TYPE_T, demonstrate their significance in accurately assessing the credit risk of airtime users. These attributes were specifically selected and assessed in this study, distinguishing the research from previous works. In general, the study introduces a set of unique attributes that have not been extensively studied in previous works on airtime credit risk prediction. The study advances the existing knowledge by confirming the significance of these attributes and achieving superior predictive performance. The inclusion of these unique variables enhances the understanding and applicability of airtime credit risk prediction models, contributing to the field and providing valuable insights for telecom service providers, including ethio telecom.

While ethio telecom solely relies on the "REPAY_AMT" attribute for its credit risk prediction, the research highlights the importance of considering additional factors as described above. These supplementary indicators provide valuable insights into the likelihood of default. By incorporating these variables into the credit risk prediction model, ethio telecom can enhance the accuracy and effectiveness of its assessments. The research findings emphasize the need to broaden the scope of attributes considered in credit risk prediction, as relying solely on a single attribute may overlook crucial information that could significantly impact the prediction outcomes. Therefore, incorporating multiple indicators, including "REPAY_AMT" and other identified attributes, would lead to more comprehensive and robust credit risk prediction models for ethio telecom.

In conclusion, this study experiment not only evaluated different machine learning algorithms but also the suitable attributes for airtime credit risk prediction. The experiment also compared the performance of machine learning algorithms using novel and existing attribute sets. The results underscore the significance of attribute selection in achieving accurate predictions. The Random Forest algorithm emerged as the top performer in terms of accuracy and precision, demonstrating its effectiveness in identifying clients at risk of defaulting on their airtime credit. Furthermore, the findings suggest that the proposed model outperforms existing models, highlighting its contribution to the field of airtime credit risk prediction.

CHAPTER FIVE

5. CONCLUSION AND FUTURE RECOMMENDATION

5.1 CONCLUSION

In conclusion, this research focused on the development of a machine learning model for airtime credit risk prediction in ethio telecom. The study aimed to assess the suitability of different algorithms, identify key attributes for prediction, and evaluate the performance of the models.

The findings of this research highlight the efficacy of machine learning models in accurately predicting airtime credit risk. In the study, machine learning algorithms, namely Random Forest, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors, show varying degrees of effectiveness in predicting airtime credit risk. Among these algorithms, Random Forest emerged as the most robust and accurate model, exhibiting high precision, recall, and F1 scores for both classes. The performance of Logistic Regression and K-Nearest Neighbors was also noteworthy, although they exhibited some limitations in terms of precision and recall for specific classes. Naïve Bayes, on the other hand, demonstrated lower overall accuracy and decreased performance when using class balance techniques.

One of the significant contributions of this research is the identification of key attributes for airtime credit risk prediction. By incorporating many variables or attributes from customer profiles data, customer's usage patterns data, customers loan histories data, and customers recharge behavior, the models developed in this study demonstrated a high degree of predictive accuracy. The distinctive aspect of the variables employed in this research, in contrast to previous studies, is their remarkable predictive capability in assessing airtime credit risk.

The research also compared the performance of novel attributes with existing ones. The evaluation metrics showed that the models built using new attributes achieved higher accuracy rates compared to those using existing attributes. This highlights the importance of carefully selecting and incorporating relevant attributes in machine learning algorithms to enhance prediction performance.

However, it is essential to acknowledge certain limitations of this research. One limitation is the reliance on a limited dataset provided by ethio telecom, covering only three months of data. This timeframe may not capture the complete range of credit risk factors and could potentially limit the generalizability of the models.

In conclusion, the findings of this research contribute to the field of airtime credit risk prediction by demonstrating the efficacy of machine learning algorithms and identifying key attributes for accurate prediction. The models developed in this study offer valuable insights for ethio telecom and other telecom service providers in effectively managing credit risks associated with airtime usage.

5.2 Recommendations

This research results provide domain experts in the telecom industry with valuable insights and tools to improve their airtime credit risk management. The predictive models, attribute significance, and performance metrics obtained from our study enable domain experts to assess risks, make informed decisions, develop early warning systems, segment customers, and optimize collection strategies effectively. By leveraging this research findings, domain experts can enhance their credit risk management practices, minimize defaults, and improve financial performance

The following suggestions are proposed based on the airtime credit risk prediction using machine learning conducted in this study.

- Implement machine learning models: This method performed remarkably well in predicting the risk of airtime credit. Therefore, it is advised to add a machine learning model as a credit risk management system to be used by ethio telecom. Ethio telecom can improve its risk assessment capabilities and make better decisions about credit extension by utilizing the strengths of machine learning models, such as its capacity to handle complex data and provide accurate predictions.
- Update and improve the predictive model frequently: As consumer behavior and market dynamics change over time, it is crucial to routinely update and improve the predictive model. This can be accomplished by adding fresh data sources, identifying new patterns, and optimizing the machine learning algorithms. ethio telecom can guarantee the

effectiveness and relevance of its airtime credit risk prediction system by staying up to date with the most recent information and continuously improving the model.

- Monitor model performance and carry out periodic evaluations: It's crucial to regularly evaluate the predictive model's performance. This involves evaluating the model's accuracy, precision, recall, and F1 scores as well as contrasting the predictions with actual instances of credit default. In order to maintain the model's effectiveness and reliability in predicting airtime credit risk, regular evaluations help identify any potential flaws or areas for improvement.

5.3 Future works

Even though this study's use of machine learning to predict airtime credit risk, there are still many ways to explore and advance this areas. Future work can be thought about in the following areas:

- Exploration of advanced machine learning methods: Although this study used well-known machine learning algorithms like Random Forest, Logistic Regression, Naive Bayes, and K-Nearest Neighbors, there are many advanced algorithms available that can uncover new insight that were not captured by the algorithm used. Future studies might examine how well other algorithms, like gradient boosting, support vector machines, or deep learning models like neural networks, perform. Comparing how well these algorithms predict airtime credit risk can reveal new information and possibly result in models that are more reliable and accurate.

REFERENCES

[1] M. Lyra, A. Onwunta, and P. Winker, "Threshold accepting for credit risk assessment and validation," *J. Bank. Regul.*, vol. 16, no. 2, pp. 130–145, 2015, doi: 10.1057/jbr.2013.26.

[2] M. Szczerba and A. Ciemski, "Credit risk handling in telecommunication sector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5633 LNAI, pp. 117–130, 2009, doi: 10.1007/978-3-642-03067-3_11.

[3] B. Dushimimana, Y. Wambui, T. Lubega, and P. E. McSharry, "Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans," *J. Risk Financ. Manag.*, vol. 13, no. 8, p. 180, 2020, doi: 10.3390/jrfm13080180.

[4] I. Aamo, A. Myom, and Y. I. Shehu, "Airtime Credit Banking: From Two Applications to One Application," *J. Comput. Commun.*, vol. 05, no. 10, pp. 10–15, 2017, doi: 10.4236/jcc.2017.510002.

[5] O. Tarekegn, "Application of Data Mining Technique for Predicting Airtime Credit Risk : The Case of Ethio Telecom Application of Data Mining Technique for Predicting Airtime Credit Risk : The Case of Ethio Telecom," 2019.

[6] K. Garba and S. Sa, "Airtime Credit Loan and Service Charge / Fee by Telecommunications Service Providers in Nigeria : Islamic Law Perspective," *IOSR J. Bus. Manag.*, vol. 21, no. 8, pp. 8–14, 2019, doi: 10.9790/487X-2108040814.

[7] Shashu Brhane Berhe, "Machine Learning Based Mobile Airtime Credit Risk Prediction using Customer Profile and Loan Information," 2021.

[8] "Airtime Credit – ethiotelecom." https://www.ethiotelecom.et/air-time-credit/ (accessed Dec. 18, 2022).

[9] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[10] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *Int. J. Inf. Technol.*, vol. 12, no. 4, pp. 1243–1257, 2020, doi: 10.1007/s41870-020-00427-7.

[11] G. Weiss, "Data Mining in the Telecommunications Industry," *Encycl. Data Warehous. Mining, Second Ed.*, no. May, 2011, doi: 10.4018/9781605660103.ch076.

[12] H. H. H. Mahmoud and T. Ismail, "A Review of Machine learning Use-Cases in Telecommunication Industry in the 5G Era," in *16th International Computer Engineering Conference, ICENCO 2020*, Dec. 2020, pp. 159–163. doi: 10.1109/ICENCO49778.2020.9357376.

[13] G. S. Popli and M. Madan, "Determinants of Customer Satisfaction in Telecom Industry - A Study of Indian Telecom Industry," *SSRN Electron. J.*, no. May, 2013, doi: 10.2139/ssrn.2277570.

[14] P. C. Sen, M. Hajra, and M. Ghosh, *Supervised Classification Algorithms in Machine Learning: A Survey and Review*, vol. 937. Springer Singapore, 2020. doi: 10.1007/978-981-13-7403-6_11.

[15] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[16] G. Sudhamathy, "Credit risk analysis and prediction modelling of bank loans using R," *Int. J. Eng. Technol.*, vol. 8, no. 5, pp. 1954–1966, 2016, doi: 10.21817/ijet/2016/v8i5/160805414.

[17] B. Yang, R. Nazari, D. Elmo, D. Stead, and E. Eberhardt, "Data preparation for machine learning in rock engineering," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1124, no. 1, 2023, doi: 10.1088/1755-1315/1124/1/012072.

[18] "Machine Learning & Training Data: Sources, Methods, Things to Keep in Mind." https://labelyourdata.com/articles/machine-learning-and-training-data (accessed Dec. 17, 2022).

[19] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *Int. J. ...*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[20] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi:

10.3389/fenrg.2021.652801.

[21] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.

[22] F. Maleki, K. Ovens, K. Najafian, B. Forghani, C. Reinhold, and R. Forghani, "Overview of Machine Learning Part 1: Fundamentals and Classic Approaches," *Neuroimaging Clin. N. Am.*, vol. 30, no. 4, pp. e17–e32, 2020, doi: 10.1016/j.nic.2020.08.007.

[23] Q. Chang and J. Hu, "Research and Application of the Data Mining Technology in Economic Intelligence System," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/6439315.

[24] M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," vol. 1, no. 4, pp. 301–305.

[25] S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau, "Machine learning-driven credit risk: a systemic review," *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14327–14339, 2022, doi: 10.1007/s00521-022-07472-2.

[26] E. S. Priya and K. Anandhan, "An Overview of Data Mining - A Survey Paper," *Int. J. Mod. Comput. Sci.*, vol. 6, no. 1, pp. 2320–7868, 2018, [Online]. Available: https://www.researchgate.net/publication/344397023

[27] M. Fralick and K. R. Campbell, "The Basics of Machine Learning," *NEJM Evid.*, vol. 1, no. 5, 2022, doi: 10.1056/evide2200062.

[28] J. Mueller and L. Massaron, *Machine learning for dummies*. 2016.

[29] I. S. Ahmad, A. A. Bakar, M. R. Yaakub, and S. H. Muhammad, "A Survey on Machine Learning Techniques in Movie Revenue Prediction," *SN Comput. Sci.*, vol. 1, no. 4, pp. 472–478, 2020, doi: 10.1007/s42979-020-00249-1.

[30] V. V. S. Dileep*, N. Rishitha, R. Gummadi, and P. N. P, "DNA Sequencing using M achine L earning and D eep L earning A lgorithms," *Int. J. Innov. Technol. Explor. Eng.*, vol. 11, no. 10, pp. 20–27, 2022, doi: 10.35940/ijitee.j9273.09111022.

[31] X. Teng and Y. Gong, "Research on Application of Machine Learning in Data Mining," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 392, no. 6, 2018, doi: 10.1088/1757-899X/392/6/062202.

[32] T. Oladipupo, "Machine Learning Overview," *New Adv. Mach. Learn.*, no. February 2010, pp. 8–18, 2010, doi: 10.5772/9374.

[33] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, 2019, doi: 10.3390/risks7010029.

[34] I. S. Ahmad, A. A. Bakar, M. R. Yaakub, and S. H. Muhammad, "A Survey on Machine Learning Techniques in Movie Revenue Prediction," *SN Comput. Sci.*, vol. 1, no. 4, pp. 1–14, 2020, doi: 10.1007/s42979-020-00249-1.

[35] R. Kaur, "A Survey on Machine Learning Techniques with Applications".

[36] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016, doi: 10.1186/s13634-016-0355-x.

[37] "Hierarchical Clustering in Machine Learning - Javatpoint." https://www.javatpoint.com/hierarchical-clustering-in-machine-learning (accessed Dec. 17, 2022).

[38] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020, doi: 10.1007/s10994-019-05855-6.

[39] P. S. Helode, Dr. K. H. Walse, and Karande M.U., "An Online Secure Social Networking with Friend Discovery System," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 4, pp. 8198–8205, 2017, doi: 10.15680/IJIRCCE.2017.

[40] A. Mishra, B. B. Gupta, D. Perakovi, and F. J. G. Peñalvo, "A Survey on Data mining classification approaches," *CEUR Workshop Proc.*, vol. 3080, 2021.

[41] H. Leopord, W. K. Cheruiyot, and S. Kimani, "A Survey and Analysis on Classification and Regression Data Mining Techniques for Diseases Outbreak Prediction in Datasets," pp. 1–11, 2016.

[42] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma, "Machine learning techniques for data mining: A survey," *2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013*, no. December, pp. 1–7, 2013, doi: 10.1109/ICCIC.2013.6724149.

[43] A. Aldahiri, B. Alrashed, and W. Hussain, "Trends in Using IoT with Machine Learning in Health Prediction System," *Forecasting*, vol. 3, no. 1, pp. 181–206, 2021, doi: 10.3390/forecast3010012.

[44] M. Batta, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 18, no. 8, pp. 381–386, 2018, doi: 10.21275/ART20203995.

[45] M. Stamp, "A Survey of Machine Learning Algorithms and Their Application in Information Security," no. November, pp. 33–55, 2018, doi: 10.1007/978-3-319-92624-7_2.

[46] N. Ganatra, "An Experimental Study of Supervised Machine Learning Classifiers," vol. VI, no. 1441, pp. 1441–1448, 2019.

[47] S. Ray, "A Quick Review of Machine Learning Algorithms," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.

[48] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, no. December 2017, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[49] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green, "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods," *Acad. Pathol.*, vol. 6, 2019, doi: 10.1177/2374289519873088.

[50] V. Dengov, *Credit Risk Analysis for the Telecommunication Companies of Russia: Statistical Model*, vol. 2, no. August 2015. 2015. doi: 10.5593/sgemsocial2015/b22/s6.018.

[51] L. Ma, X. Zhao, Z. Zhou, and Y. Liu, "A new aspect on P2P online lending default prediction using meta-level phone usage data in China," *Decis. Support Syst.*, vol. 111, pp. 60–71, 2018, doi: 10.1016/j.dss.2018.05.001.

[52] D. Björkegren and D. Grissen, "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment," *World Bank Econ. Rev.*, vol. 34, no. 3, pp. 618–634, 2020, doi: 10.1093/wber/lhz006.

[53] ITU, "Use of telecommunications data for digital financial inclusion," 2021.

[54] H. Ots, I. Liiv, and D. Tur, "Mobile phone usage data for credit scoring," *Commun. Comput. Inf. Sci.*, vol. 1243 CCIS, no. January, pp. 82–95, 2020, doi: 10.1007/978-3-030-57672-1_7.

[55] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, 2020, doi: 10.1016/j.inffus.2019.12.001.

[56] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, and B. T. T. Yeo, "Predicting Alzheimer's disease progression using deep recurrent neural networks," *Neuroimage*, vol. 222, no. September, 2020, doi: 10.1016/j.neuroimage.2020.117203.

[57] I. C. Dipto, T. Islam, H. M. M. Rahman, and M. A. Rahman, "Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease," *J. Data Anal. Inf. Process.*, vol. 08, no. 02, pp. 41–68, 2020, doi: 10.4236/jdaip.2020.82003.

[58] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Appl. Sci.*, vol. 3, no. 2, pp. 1–12, 2021, doi: 10.1007/s42452-021-04148-9.

Appendix 1 Random Forest Algorithm Implementation Sample Code"

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
rc=RandomForestClassifier()
rc=rc.fit(X_train,y_train)
p=rc.predict(X_test)
print("The accuracy is ", accuracy_score(y_test,p))
print("classification result")
print(classification_report(y_test, p))
print("confussin matrix result")
print(confusion_matrix(y_test, p))
```

Appendix 2 Screenshot of Logistic Regression Algorithm Implementation Sample Code"

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
logr = LogisticRegression()
logr= logr.fit(X_train,y_train)
pl=logr.predict(X_test)
print("THe accuracy is ", accuracy_score(y_test,pl))
print("classification result")
print(classification_report(y_test, pl))
print("confussin matrix result")
print(confusion_matrix(y_test, pl))
```

Appendix 3 Screenshot of Naïve Bayes Classifier algorithm Implementation Sample code

```
# train a Gaussian Naive Bayes classifier on the training set
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
# instantiate the model
gnb = GaussianNB()
# fit the model
gnb.fit(X_train, y_train)
p22=gnb.predict(X_test)
print("THe accuracy is ", accuracy_score(y_test,p22))
print("classification result")
print(classification_report(y_test, p22))
print("confussin matrix result")
print(confusion_matrix(y_test, p22))
```

Appendix 4 Screenshot of K nearest Neighbor Algorithm Implementation Sample code

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 3)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
print("The accuracy is ", accuracy_score(y_test,y_pred ))
print("classification result")
print(classification_report(y_test, y_pred ))
print("confussin matrix result")
print(confusion_matrix(y_test,y_pred ))
```

Appendix 5 Screenshot of a code for implementation of Feature Selection with highly correlated feature removal function

```
def correlation(dataset, threshold):
    col_corr = set()  # Set of all the names of correlated columns
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if (corr_matrix.iloc[i, j]) > threshold:
                colname = corr_matrix.columns[i]  # getting the name of column
                col_corr.add(colname)
    return col_corr

corr_features = correlation(X_train, 0.7)
```

Appendix 6
Screenshot of a code for Feature Importance Analysis of the selected Model.

```
import matplotlib.pyplot as plt

# Get feature importances from the trained random forest model
importances = rc.feature_importances_

# Get the indices of features sorted by importance in descending order
sorted_indices = np.argsort(importances)[::-1]

# Get the names of the sorted features
sorted_feature_names = [feature_names[i] for i in sorted_indices]

# Plot the feature importances
plt.figure(figsize=(10, 6))
plt.bar(range(len(importances)), importances[sorted_indices])
plt.xticks(range(len(importances)), sorted_feature_names, rotation='vertical')
plt.xlabel('Feature')
plt.ylabel('Importance')
plt.title('Feature Importance')
plt.tight_layout()
plt.show()
```

Appendix 7 Sample Data Used in the analysis

CUST_AGE	INIT_LOAN_AMT	LATE_PAY	REPAY_AM	REPAY_PC	RECHARGE	OFF_PEAK	OFF_PEAK	PEAK_USA	PEAK_AM	SMS_LOG	SMS_LOG	DATA_USA	DATA_RE	NET_BUSI	NET_BUSI	GENDER_F	EDUCATIC	EDUCATIC	EDUC
42	25	0	10	2	25	73	0	70	0	0	0	0	0	1	0	1	0	1	
20	5	4	1	0	5	0	0	29	0	0	0	5	0	1	0	0	0	0	
44	5	5	0	0	45	739	0	86	0	0	0	0	0	0	1	1	0	0	
34	15	4	10	0	60	0	0	268	2	0	0	0	0	1	0	1	0	1	
60	10	0	0	1	20	0	0	464	0	0	0	0	0	1	0	0	0	0	
33	5	5	0	0	5	0	0	162	1	0	0	10	0	0	1	1	1	0	
31	25	5	10	0	5	338	0	932	0	24	0	7549	0	1	0	1	0	1	
25	15	0	15	1	5	0	0	423	3	0	0	0	0	0	1	1	0	1	
29	10	0	5	1	50	27	0	2156	0	0	0	46	0	0	1	1	0	0	
48	10	0	10	1	15	0	0	143	0	0	0	0	0	0	1	1	0	1	
75	25	0	25	0	50	0	0	0	0	4	0	0	0	0	1	1	0	1	
49	10	0	6	1	25	0	0	0	0	0	0	367	0	1	0	1	0	1	
31	15	0	5	1	65	0	0	68	0	0	0	0	0	1	0	0	0	0	
26	5	0	5	0	10	0	0	1	0	0	0	0	0	0	1	1	0	0	
34	25	0	25	2	20	64	0	75	0	0	0	0	0	1	0	1	0	1	
46	25	15	10	0	10	0	0	0	0	0	0	0	0	0	1	1	0	1	
25	5	0	5	0	15	0	0	434	3	0	0	0	0	0	1	1	0	1	
43	5	0	5	0	25	0	0	0	0	0	0	3617	0	0	1	1	0	0	
45	15	0	0	1	5	0	0	789	6	0	0	0	0	0	1	1	0	0	
41	15	5	10	0	25	0	0	0	0	0	0	2097	0	0	1	1	0	1	
29	50	0	0	5	40	0	0	2023	0	0	0	0	0	1	0	0	0	0	

Appendix 8 Sample description of the Attributes/features

Attribute Name	Data Types	Description
SERVICE NUMBER	Numeric	identity used to uniquely identify a subscriber
CUST_TYPE_NAME	String	Individual or Enterprise
NET BUSI_TYPE_NAME	String	Network type (LTE,WCDMA)
STATUS NAME	String	The status of service number(Active, Idle, Barring, Predeactivated, Suspend)
CUST_AGE	Numeric	The age of the customers
GENDER NAME	String	Gender name (Male or Female)
EDUCATION_NAME	String	This shows academic rank of the user
OPER_TYPE	String	It can be by transfer or recharge.
LOAN BALANCE_TYPE	Numeric	This simple a code given to loan data
INIT LOAN_AMT	Numeric	Initially a customer takes the loan amount
INIT_LOAN_POUNDAGE	Numeric	Initial specific fee or charge related to loans or borrowing money
LOAN POUNDAGE	Numeric	specific fee or charge related to loans or borrowing money