MINISTRY OF SCIENCE AND EDUCATION
REPUBLIC OF AZERBAIJAN

KHAZAR UNIVERSITY

GRADUATE SCHOOL OF SCIENCE, ART AND TECHNOLOGY

Field of study and code:	60631
Speciality:	Computer Engineering

Master Student

Habibullayeva Naila Habil

“Security Issues of Big Data in the Banking Sector and Analytical Approach”

MASTER THESIS

Supervisor:	Dr. Amir Masoud Rahmani
Advisor:	Dr. Behnam Kiani Kalejahi

December – 2023

ABSTRACT

This study highlights the banking sector as a rich source of data that includes customer data, financial transactions and economic indicators. The main focus is on effectively applying and training machine learning techniques using this abundance of data. In particular, it aims to determine which model gives better results by examining various machine learning models for risk detection in the mentioned data set.

In this context, the main objectives of the study are:

- Applying machine learning techniques for risk detection using a rich data set in the banking sector.
- To effectively train these techniques and adapt them to the data set.
- Performing risk detection using various machine learning models, especially the logistic regression model.
- To determine which model performs better by making comparisons between these models.

This thesis study includes codes written in Python programming language and focuses on the logistic regression model. Despite the broad spectrum of machine learning, the focus of the study is on using a logistic regression model to address the complexity and uncertainty in the dataset.

Python's flexible and comprehensive structure provides an ideal environment for analyzing large data sets and developing machine-learning models. In this thesis, the logistic regression model was chosen and the focus was on the effectiveness of this model in predicting the probability of getting a loan and evaluating future risks.

However, this study is not limited to logistic regression only; it also explores the potential of other tools and libraries within Python's machine-learning ecosystem. Logistic regression may be the basic model of the study, but the infrastructure provided by this thesis may open a door to researchers and professionals who want to apply different models and techniques in the future.

In conclusion, this thesis not only includes the practice of coding in Python but also provides a foundation for developing data analysis and machine learning models, emphasizing the logistic regression model. This not only ensures clarity of the methodological approach of the thesis but also may act as a foundation for further studies.

However, this thesis addresses not only the current situation but also changing risks. An important goal of this work is to identify potential risks in the industry and offer a proactive approach against these risks. By identifying problems that may arise in the future, it manages to ensure that the banking sector has a more resilient and adaptable structure.

As a result, this thesis focuses on exploring whether big data and machine development models are a policy that can strengthen risk management strategies in the banking sector. The findings can provide valuable insights to industry professionals, regulators, and researchers to more effectively prepare for outperformance.

6 chapters, 29 subchapters, an Introduction, Conclusion, Future work, Advantages and disadvantages and References are all included in the thesis.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

INTRODUCTION

The features of big data, such as volume, diversity, and speed, are often used to characterize it. The term volume denotes the extent of the datasets. The fact that datasets originate from multiple sources and are in different forms is referred to as diversity. In contrast, speed necessitates rapid data production, collection, and analysis.

Multiple industries utilize big data. For example, it is used in the financial sector to watch and control the risks of customer trades, in the healthcare sector to analyze patient data, and in the retail sector to examine consumer behavior. The analysis of large datasets differs from conventional data analysis methods. It helps create useful insights from information by using technologies like big data analytics, deep learning, AI/ neural networks, and data mining.

Big data uses a lot of information to get the most exact and complete results possible. But there are also some possible risks to using big data. Here are some dangerous habits in big data:

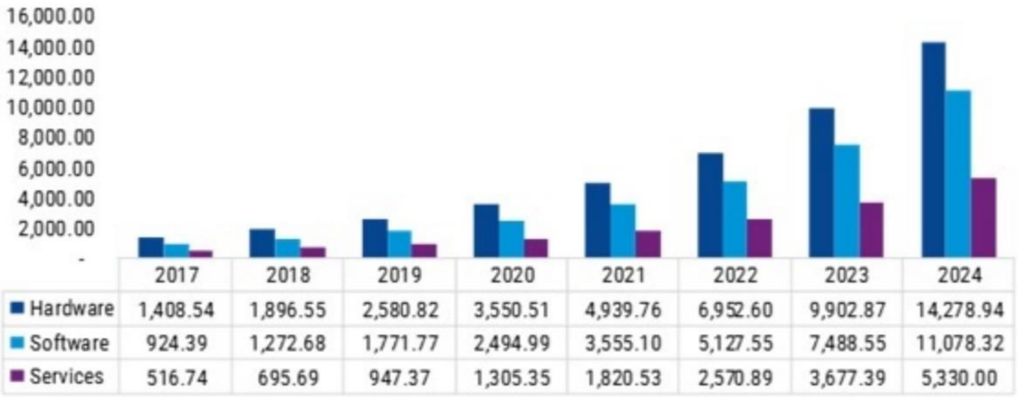
- Concerns about data security: Personal and private information can be found in big data. If this data is lost or stolen, institutions and users could be in a lot of trouble. The use of big data can be hurt by problems with data protection, computer threats, or the theft of data.
- Misconceptions: Analyzing big data means putting together different pieces of information to come to a decision. However wrong or partial data can lead to confusion and wrong conclusions. Also, mistakes made during the research lead to wrong results.
- Breaches of data privacy: Big data includes information about customers and other private information. If this information is used or shared without permission, it could be a breach of users' privacy.
- Bias and discrimination: The effects of big data are built on statistics from the past. But it may have biased or unfair information about the past. So, when analyzing a lot of data, skewed data is likely to change the results.
- The human factor: Analyzing big data is often a technical and complicated process. Incorrect results can be caused by wrong data entry, wrong parameter selection, or wrong analysis.

In conclusion, big data analytics is a useful tool, but it can be dangerous if it isn't used right. Experts in data analysis should be in charge of it so that correct data can be collected, analyzed, and used. Also, problems of data protection and privacy should be taken into account. The banking sector employs big data technology to enhance customer experience, mitigate risks, and optimize business outcomes through the analysis of substantial volumes of customer data. The aforementioned information can be procured from various sources such as customer profiles, bank statements, payment histories, credit scores, social media, and other analogous sources.

The utilization of big data has the potential to provide banks with enhanced insights into customer behavior, trends, and requirements. By means scrutinizing this data, financial institutions can discern the requirements of their clients, furnish superior customer assistance, and enhance customer allegiance. The utilization of big data analytics can improve the risk management practices of banks. Financial firms can use big data analytics models to enhance their comprehension of clients' credit risk. The models utilized in this context employ an analysis of various data sources such as social media, web searching, payment pasts, and transaction histories to assess the credit risk of customers.

The banking sector is one of the areas where machine learning technology is used most intensively. It has been possible to address challenges such as the need to generate and process large amounts of data, risk management, fraud detection, improving customer experience, and creating more efficient business processes thanks to machine learning. The banking industry can derive numerous advantages from big data analytics, which include enhanced customer experience, improved risk management, and superior business outcomes. Consequently, it is of utmost significance for financial institutions to gather, retain, and scrutinize client information by utilizing big data technologies.

In recent years, businesses around the world, including banks, have begun aggressively using new machine-learning techniques to increase their competitiveness in the customer acquisition market. Due to the increasing quantity of data and widespread contact with high-performance computing and computer resources, ML has enabled businesses to greatly improve customer experience. The expansion of worldwide Machine Learning marketing during the latest years is seen in Figure 1.



	2017	2018	2019	2020	2021	2022	2023	2024
■ Hardware	1,408.54	1,896.55	2,580.82	3,550.51	4,939.76	6,952.60	9,902.87	14,278.94
■ Software	924.39	1,272.68	1,771.77	2,494.99	3,555.10	5,127.55	7,488.55	11,078.32
■ Services	516.74	695.69	947.37	1,305.35	1,820.53	2,570.89	3,677.39	5,330.00

Figure 1. Global ML market. Source: datafloq.com.

The research technique used in this work is primarily grounded on credible, publically accessible worldwide historical and contemporary experiences, research findings, analysis, and databases. Numerous specialist scientific publications, pertinent institutional records (including those from regulatory agencies), and publicly accessible papers pertinent to this topic will be taken into consideration, among other things. The study considers both large and developed banking sectors and bank attributes, as well as small and emerging ones. The thesis's comparative overview, analysis, and conclusions are the main points of emphasis in this work. This is the rationale for the examination of the practical applications of Artificial Intelligence and ML in banking risk management, in addition to the anticipated enhancements.

The main goal of this thesis is to introduce the reader to the current machine learning and fraud cases in the banking sector and the challenges of training big data by machine learning algorithms and the specific fraud cases that can be solved in these cases. Additionally, an ML algorithm was developed during the implementation phase to demonstrate a real-world situation. The results of this thesis will provide an IT professional with regular information regarding the use of machine learning for credit. A marketing professional will learn the latest information about digital trends, machine learning, and algorithms. Any small bank employee as well as a bank employee who is new to this sector can apply the project presented in this thesis and use it as a starting point for the application of machine learning.

Research Questions:

RQ1: What are the fundamental security concerns and challenges associated with big data?

RQ2: What is the information about the business dataset, particularly in the financial/ banking industry? How data is collected, what instruments are utilized, etc.

RQ3: Which machine learning algorithms are used for identifying and predicting security issues in the banking industry?

RQ4: What are examples of anomaly actions in the banking industry? Understanding and defining fraud acting.

RQ5: What levels does each data-based transaction go through in the banking sector, and how does each transaction that is predicted to cause security problems to end?

1. LITERATURE REVIEW

There has been a surge in curiosity regarding the application of analytics for big data and machine learning in banking sector fraud analysis over the past few years. We will look at a few of the most important works on this topic in this summary of the literature.

A study conducted by Johnson and Smith (2020) [1] examined the importance of machine learning models in the banking industry and how these models are used in the field of fraud detection. In this comprehensive research, the most current developments in the sector were evaluated by focusing on various academic theses and articles.

The prominent work of Brkić et al. [2] described a specific machine-learning algorithm aimed at preventing credit card fraud. This algorithm aims to detect signs of fraud by analyzing customer spending habits, geolocation data, and unusual activities. In the tests, it was determined that this special algorithm has higher accuracy rates than traditional methods.

The work of Lee et al. [3] made a significant contribution to the use of deep learning techniques on large data sets. The ability of deep learning algorithms to effectively analyze multiple variables, such as customer account activity, demographic data, and previous shopping habits, has increased fraud detection performance. This study has been successful in detecting more complex fraud scenarios than traditional methods.

Garcia and Martinez [4] evaluated various machine learning models, such as decision trees and support vector machines, in a study on credit card fraud prevention. In particular, they showed that the use of attributes such as customer spending, geographic location, and device usage are effective in detecting fraud.

Chen et al. [5] examined the role of big data analytics in analyzing bank customers' online behavior. Using machine learning algorithms, they developed a model to detect anomalies in customer online behavior and identify potential fraud situations in advance.

Kim and Park [6] aimed to provide a bibliometric analysis and discussion on research trends of ML for mental health in social media. By identifying anomalies in customer habits and healthy movements, they concluded that this model strengthens customer relationships and reduces fraud.

G. S. Vaishnavi Nath Dornadula [7] conducted a study to detect fraudulent transactions on customer accounts. In this study, a model was developed that identifies specific patterns of fraudulent transactions using techniques such as time series analysis and support vector machines.

Chang and Wu [9] researched preventing fraudulent activities on mobile banking applications. In this study, a deep neural network and linguistic processing model were created to identify irregularities in customer contacts and promptly take action in possible fraud scenarios.

Shalini et al. [13] research addresses the use of blockchain technology. In this study, a model was proposed on how machine learning techniques can be integrated to monitor financial transactions on the blockchain and increase security.

Takao Terano et al. [12] conducted a study to detect fraudulent transactions made through bank ATMs. In this study, a model was developed to identify fraud cases by focusing on factors such as customer ATM usage habits and geographical location data.

Kshetri and Voas et al. [11] proposed a study that introduces ABISS, an agent-based simulator that can be used to analyze customer wandering patterns and shopping patterns in a supermarket. The simulator enables "virtual experiments" to be carried out by altering different shop operations and retail business characteristics. The simulation model's construction is discussed in the article, which also demonstrates how the layout of the shop, and the placement of in-store advertising and recommendation systems affect consumer flow and sales.

M. Brouwer et al. [20] carried out a comparison research between Support Vector Machines, Random Forests, Decision Trees, and Logistic Regression to detect credit card fraud. Their findings highlighted the superiority of ensemble methods like Random Forest in achieving higher accuracy and lower false positive rates.

Geron et al. [19] employed Bayesian Regularized Neural Networks for credit card fraud detection. Their approach demonstrated competitive performance compared to traditional machine learning algorithms, showcasing the potential of neural networks in this context.

L. Breiman [22] introduced the concept of the Random Forest algorithm, which has since been widely adopted for credit card fraud detection due to its ability to handle large and imbalanced datasets effectively. The algorithm's ensemble approach and robustness against overfitting have contributed to its popularity in the field.

Pozzolo et al. [27] explored the application of deep learning techniques, including ANNs, for credit card fraud detection. Their results demonstrated that deep learning models, especially ANNs with multiple hidden layers, achieved superior performance compared to traditional machine learning methods.

S. Yadav et al. [29] proposed a novel deep learning-based approach using a stacked autoencoder to detect credit card fraud. Their technique showcased exceptional accuracy and robustness against adversarial attacks, reaffirming the effectiveness of ANNs in this domain.

H. Carrascosa et al. [37] employed Long Short-Term Memory (LSTM) networks, a specialized type of ANN, for detecting fraud in credit card transactions. Their results revealed that LSTM networks were capable of capturing temporal dependencies in transaction sequences, leading to improved fraud detection performance.

2. FRAMEWORK AND PROBLEM STATEMENT

2.1 Purpose and Objectives

The banking sector is known to face numerous security challenges. However, the primary focus of this project is to use machine learning methods to classify users for credit transactions and to forecast fraudulent activities that may impact transaction time.

This thesis's primary objective is to establish machine learning algorithms to identify and prevent fraudulent activities. This thesis examines the machine learning techniques that are taught and used under supervision to detect anomalies that occur during banking transactions (such as money transfers and credit applications), detect missing information in user data, and analyze fraudulent activities. explains the efficacy and efficiency of various techniques.

Big data statistics can assist organizations in achieving improved business outcomes. By obtaining a deeper comprehension of consumer behavior, for instance, they can improve their targeting. In addition, data analytics assists businesses in making more effective choices and can enhance business processes. You will learn how to construct a machine learning model susceptible to spotting anomaly behaviors by the end of this module. In addition, you will learn how to manage class imbalances that occur in any data set, as well as how to choose a model and tune its hyperparameters.

This thesis can discuss the vulnerabilities of the banking industry's security systems and how fraudulent activities are conducted. After the thesis, this evidence can assist banks in developing more robust and efficient security systems, or enlighten any IT professional about the fielding possibilities of machine learning algorithms. The project that is described in this thesis may serve as a jumping-off point for the implementation of applied machine learning algorithms in any medium or small business, including marketing departments.

2.2 Big Data and Data Mining

Big data refers to information data sets that are too large, complex, and diverse to handle the processing power of traditional databases. This data is typically created and stored at high speed. Big data has 3 types, there are Structured, Unstructured, and Semi-Structured data.

Data mining is a discipline in which statistical, mathematical, and computational techniques are used to discover hidden info and patterns within huge data sets. Data mining helps businesses gain a competitive advantage.

For machine learning, this data contains information that drives the learning process of an algorithm. Machine learning includes the processes of collecting accurate and qualified data, processing this data, and creating and training the model. Therefore, data quality and processing processes are critical for a successful machine learning application.

- Training Data: The data set on which the machine learning method is experienced. This data set is applied for the model to learn a task. Examples include input data and target outputs.
- Validation Data: It is a separate data set applied to monitor the model's functionality during training and make adjustments. The model is tested on validation data.
- Test Data: It is the data set in which the performance of the model after training is measured. The model makes predictions on the test data and compares them with the actual results.
- Features: These are the entries in the data set. Features represent data points that enter the model's learning process. For example, features such as the size of a house, its location, and the number of rooms can be used to estimate the price of a house.
- Labels: These are the target outputs in the dataset. The learned model tries to predict these labels. For example, in an image dataset, labels may represent the objects contained in the images.
- Data Preprocessing: It is the process of bringing the data into a form that the model can use. This process may include operations such as data cleaning, filling in missing data, normalization, and coding.
- Data Splitting: Separating the data into training, validation, and test data is important to evaluate the performance of the model.
- Overfitting and Underfitting: Overfitting or underfitting problems are important issues that must be prevented during the training process of the model. These issues can cause the model to misbehave on training and testing data.
- Data Collection: Data greatly affects the success of the model. Insufficient or inaccurate data can negatively impact the quality of the model.

Data Mining has several algorithms and they are:

a. Decision Trees: Used for classification and prediction.
b. Support Vector Machines (SVM): Used to solve classification and regression problems.
c. Clustering Algorithms: Used to group data points with similar characteristics.

d. Logistic Regression: Used to model the relationship between two variables.

At the same time, data mining is used in many areas, for example:

a. Marketing: Customer segmentation, product recommendations, and campaign analysis.
b. Health Services: Disease diagnosis, treatment planning, and hospital data analysis.
c. Finance: Credit risk assessment, stock price prediction, and fraud detection.

2.3 Machine learning background

Machine learning is a branch of artificial intelligence that refers to the ability of a computer system to learn through data-driven experiences. Essentially, machine learning is the process of training a computer with data and algorithms so that it can solve a specific task or problem without human intervention. Machine learning includes these key components:

- Data: The data that underlies machine learning projects. Data are the sources of information on which learning algorithms are trained and the results are evaluated. This data can be text, images, audio, numbers, or other types.
- Algorithms: Machine learning involves the use of various mathematical and statistical algorithms. These algorithms analyze data, learn its properties, and use them to predict or classify the results.
- Model Training: The machine learning model learns based on data. The training process involves the model examining the data, extracting features, and predicting desired outcomes.
- Model Evaluation: The trained model is evaluated on a separate dataset to test its performance. Different metrics can be used to evaluate how well or poorly the model works.
- Model Tuning: It involves the process of tuning algorithms and parameters to improve the performance of the model.

Machine learning is used in many different application areas. Examples include image recognition, natural language processing, recommendation systems, automated driving technologies, financial forecasting, healthcare analytics, and many more. Machine learning helps automate tasks such as big data analysis, pattern recognition, and prediction and solve complex problems.

Machine learning is an ever-evolving field, and new algorithms and techniques are constantly being developed. In this way, it provides more application areas and data analysis opportunities, which play an important role in business and scientific research.

2.3.1 Data in Machine Learning

Financial institutions can utilize various corporations' databases to manage and analyze data. Several instances can be cited as examples:

- The Oracle database is a widely adopted technology among numerous prominent financial institutions globally. The financial services industry frequently opts for the Oracle database owing to its dependable nature and efficient functionality.
- Microsoft SQL Server is a widely used database system that is employed by numerous financial institutions to fulfill their data management requirements. SQL Server is recognized as a robust tool for managing and analyzing data.
- IBM DB2 is a widely utilized database system in the banking industry to manage extensive data sets. The financial services sector frequently opts for DB2 due to its capacity for expansion and dependability.

In the banking field, user data is stored in the database. They are mainly used in the databases of large companies such as SQL, Oracle, and IBM. Transactions at the end of each day are collected in the database, and the development team writes them themselves.

Machine learning models must predict any score in the data, so the target is defined a year in advance, that is, what this model should analyze, the test is defined and data is collected.

Any machine learning algorithm relies on data as its primary source of input. Data offers insights into customer behavior and serves as the foundation for training machine learning algorithms and collecting valuable information. Let's begin by providing a precise definition of data.

According to the Cambridge Dictionary, data is defined as information, specifically facts or statistics, collected for analysis, consideration, and utilization in support of decision-making. It can also refer to information in an electronic format that computers can process. In the realm of machine learning, data is of paramount importance. It is shown in Figure 1.1.

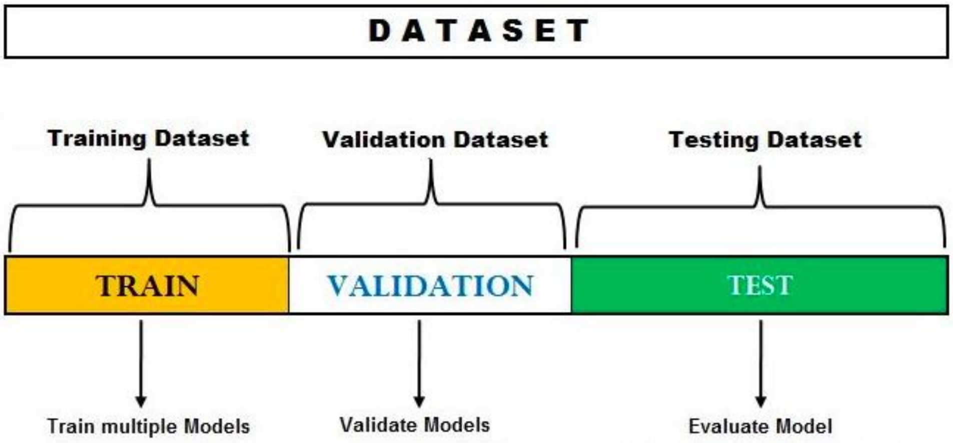


Figure 2.1 Dataset prepare process. Source: Turing.com

It encompasses the collection of data that can be analyzed or measured to train a machine-learning model. The quantity and quality of data used for training and testing significantly impact the performance of a machine-learning model. Data may originate from different sources, comprising databases, spreadsheets, and APIs, and can come in different formats such as numerical, categorical, or time-series data. Machine learning algorithms leverage data to unearth patterns and relationships between input parameters and desired outcomes, which can then be applied to tasks involving categorization or prediction.

The concept of data has existed for a long time and is well-established. Big Data is a term used to describe vast and complex datasets that cannot be effectively processed or analyzed using traditional data processing tools. It encompasses an immense amount of structured, semi-structured, and unstructured data generated from various sources, including social media, sensors, transactions, and other digital activities. This term also describes the technologies and methodologies used to store, process, and analyze these massive datasets to extract valuable insights and knowledge.

Big data is often defined by data scientists using the four Vs

1. Volume: This pertains to the enormous amount of information derived from sources like social media, IoT devices, sensors, and transactions. Big Data is characterized by its large volume, necessitating scalable storage solutions for effective management and processing.
2. Velocity: Big Data is often created in real-time or almost real-time, therefore it has to do with the pace at this data accumulates and demands real-time or near-real-time handling and evaluation.

3. Variety: This encompasses the diverse types and formats of data originating from many resources, such as data that is unstructured, semi-structured, and structured. Handling Big Data requires a wide range of data processing and analysis techniques.
4. Veracity: This involves the accuracy and reliability of the data generated. Big Data is often incomplete, inaccurate, or inconsistent, necessitating advanced data cleaning, filtering, and quality assurance techniques to ensure accuracy and reliability.

Data organization methods include two primary categories emphasized in data science: structured and unstructured data.

- Structured data follows a fixed format with a predefined schema, making it easily manageable and analyzable using traditional data processing tools and technologies. Examples of structured data in Big Data include transactional data, customer data, and financial records, typically stored in relational databases. Specialized tools and technologies like Hadoop, Spark, and NoSQL databases are employed to manage and process huge volumes of structured data in Big Data. Advanced analytics tools like machine learning and data mining are used to extract insights from structured data, aiding organizations in data-driven decision-making.
- Unstructured data lacks a predefined structure or format and encompasses data types such as text, images, videos, audio, social media posts, and other digital content. Unstructured data is generated in large volumes and presents challenges in storage, management, processing, and analysis using traditional tools. The rise of Big Data has led to an exponential increase in unstructured data, offering both opportunities and challenges for businesses. Extracting insights from unstructured data necessitates advanced methods such as deep learning, machine learning, and conversational speech extraction.

2.3.2 Algorithms in Machine Learning

For ML algorithms and client behavior analysis, data is a crucial component. The definition of an algorithm and its function in an ML pipeline are covered in this section. An algorithm in ML refers to a set of rules and instructions that are used to train a model to recognize patterns in data. It is a mathematical formula or a sequence of steps that enable an ML model to learn from data and make predictions or decisions based on that learning. In ML, algorithms are used to transform input data into useful output by discovering relationships, patterns, and trends in the data.

There are four main types of algorithms in ML that can be divided by their purposes. It is shown in Figure 1.2:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-supervised Learning
4. Reinforcement Learning

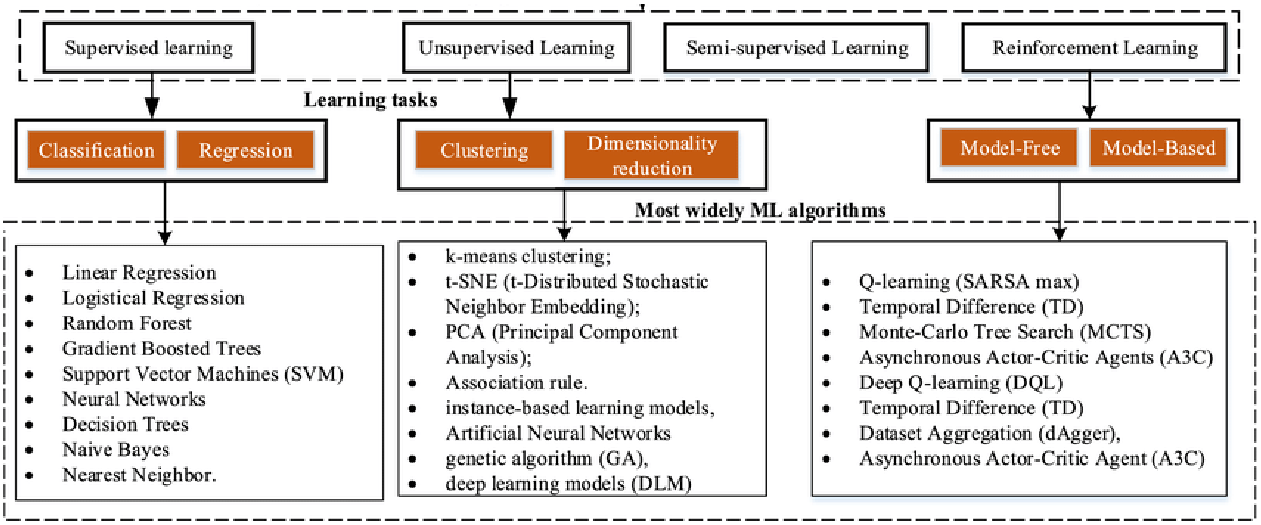


Figure 2.2 Types of ML algorithms. Source: www.researchgate.net

Supervised Learning: In supervised learning, a labeled dataset containing input feature pairs matched with matching output or target variable is used to train a model. Using the labeled data supplied during the training phase, the algorithm in this paradigm learns a mapping function that links the input characteristics to the output variable.

A fundamental method in machine learning, supervised learning is used for problems like regression and classification, where the model is trained to make judgments or predictions based on input data. The suitability of the selected model and the caliber of the training data are the two main factors that affect how well-supervised learning works. Apart from what we said, there are other supervised models (Table 1.1).

Table 2.1 The supervised ML models

Methods	Advantages	Disadvantages

Logistic Regression	The result of this method is easy to understand and it can be practically useful in establishing the relationships between other events	It can accept a large number of independent variables. Therefore, the parameter estimation procedure of logistic regression can lead to inaccurate estimates of parameters. Because there is data too large
Support Vector Machine	● Durable to noise. ● Easy to command. ● It can be used to model non-linear. ● Overfitting is unlikely to occur.	● Training is slow ● Difficult to determine optimal parameters when training data is ● Difficult to understand the structure of the algorithm
Naïve Bayes	● Not difficult to understand ● It takes less computational time ● No effect of the order on training	● Mutually exclusive ● Variables must be statistically independent ● Less accruable because attribute and class frequencies affect the accuracy
Decision Tree	● Not difficult to understand ● Fast to forecast ● No effect of the order on training	● Error in the training set can lead to wrong the result final decision ● Mutually exclusive
Random Forest	● Not difficult to understand ● Parameters can be set easily, therefore, eliminating the need for pruning the trees	● Data including categorical variables with a difficult number of level

		● Random Forest is biased in forever of those attributes

In supervised learning, sets of inputs and their corresponding accurate outputs make up the training data. The main goal is to obtain a function that can anticipate new, unknown inputs with accuracy. Reducing the difference between the expected and actual outputs is the aim. Several crucial phases are often included in the supervised learning process:

- **Data Gathering and Processing:** The first stage comprises collecting and getting ready the training data, which includes transforming, cleaning, and making it appropriate for the model.
- **Feature Selection and Extraction:** Next, pertinent characteristics are taken out of the input data, and the most important ones are selected to be included in the model.
- **Model Selection and Training:** The next step is to choose a suitable model architecture and use the labeled data to train it. The training process of the model involves modifying its weights and biases to minimize the deviation between the expected and actual output.
- **Model Assessment**: Following training, the model, it is assessed using a separate validation dataset to gauge its accuracy and performance. If necessary, the model is refined and retrained.

The working principle of the supervised learning model is shown in Figure 2.3.

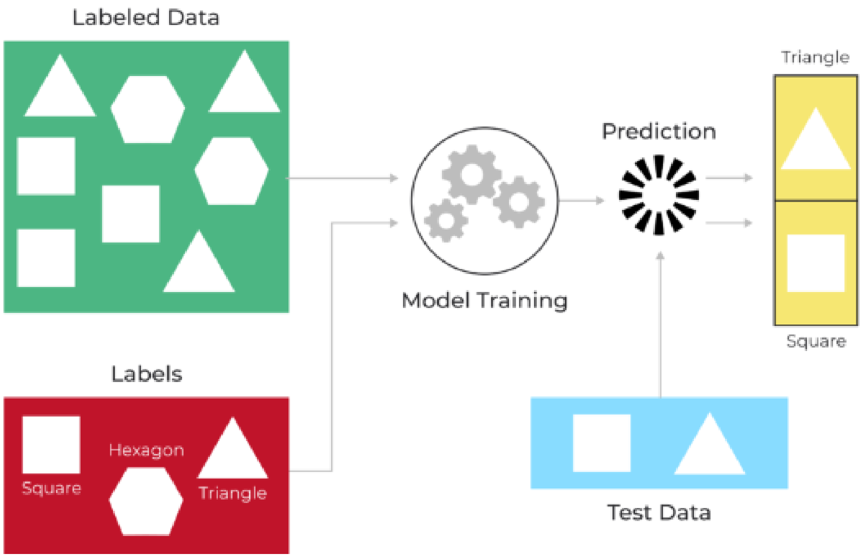


Figure 2.3 Workflow of Supervised Learning algorithm workflow. Source: www.postindustria.com

Unsupervised Learning method: Structure recognition is done via unsupervised learning techniques and structures in data in a way that does not require a labeled output or target variable. Using past data to deduce underlying hidden patterns is known as unsupervised machine learning. The types advantages and disadvantages of unsupervised algorithms are shown in Table 2 below.

Table 2.2 Unsupervised Learning algorithm, and their advantages and disadvantages

Algorithm	Advantages	Disadvantages
K-Means	• Simple and fast. • Effective on large datasets.	• Requires pre-specifying the number of clusters • Sensitivity to the initial choice of cluster centers.
Hierarchical Clustering	• Reveals a hierarchical structure	• Computationally expensive on large datasets • Longer runtime.

	• Useful for understanding relationships between clusters.	
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	• Can identify clusters of varying shapes. • Robust to noise.	• May struggle with clusters of uneven density • Critical parameter selection.
Gaussian Mixture Model (GMM)	• Flexible model that adapts to cluster shapes • Capable of soft clustering.	• Sensitivity to the initial choice of parameters. • Sensitive to the complexity of the dataset.
t-SNE (t-distributed Stochastic Neighbor Embedding)	• Useful for displaying extremely dimensional information in a place with fewer dimensions • Groups similar objects together.	• Computationally expensive • Two-dimensional visualizations can be challenging to interpret.
Principal Component Analysis (PCA)	• Used for dimensionality reduction and feature selection • Useful for understanding the fundamental structure of data.	• Assumes linearity in the data. • Sensitive to outliers.

In this method, a machine learning model searches the data on its own for any patterns, structure, similarities, or contrasts. Human involvement is not required beforehand. These algorithms are used for various purposes:

- **Clustering:** Unsupervised learning algorithms are used by grouping data based on similarity metrics. This helps identify natural groupings within the data. For example, customer segments, disease subtypes, or network anomalies.
- **Dimensionality Reduction:** Data complexity may be decreased via the utilization of unsupervised learning, that is, it is the process of turning high-dimensional data into a lower-dimensional representation. This can make it easier to visualize, analyze, or process data.
- **Anomaly Detection:** It is possible to utilize unsupervised learning can be used to determine rare and unusual data points (anomalies) that do not conform to the norm in the dataset. It is especially useful for fraud detection, network security, and quality control.
- **Density Estimation:** Unsupervised learning can estimate the stochastic density product of the data, which can help understand the data distribution and model probability.
- **Market Basket Analysis:** This is a method used in the retail industry and is used to identify patterns of items found together in customer transactions. This is useful for product recommendations and inventory management.

Unsupervised learning algorithms are valuable for exploratory data analysis, data preprocessing, and finding hidden structures within data. It is especially useful when you do not have labeled data or a specific target variable and you want to understand the internal structure of the data. It is shown in Figure 2.4.

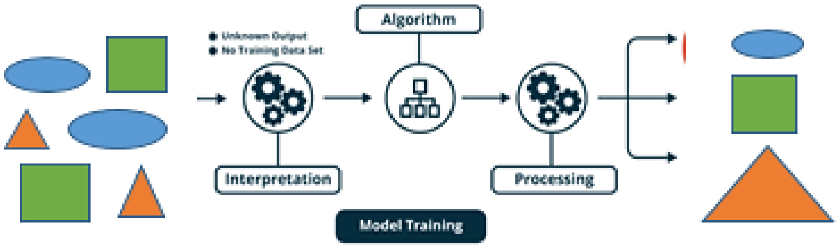


Figure 2.4 Unsupervised Learning models workflow. Source: www.mygreatlearning.com

As we see in the picture above, the machine does not know what the objects are based on their similarities (color, shape, etc.) and groups them into different categories. It is used to find the structure of a particular data set.

Semi-supervised Learning: An increasingly novel and less common kind of machine learning is called semi-supervised learning, which mixes a major quantity of unmarked data with a minimal quantity of labelled data during training. Between supervised learning (using labelled practice data) and unsupervised learning (using unlabeled practise data) is semi-supervised learning.

Numerous practical uses exist for semi-supervised learning. In many sectors, labeled data are scarce. The labels (target variable) could be hard to get because they need costly, time-consuming investigations, specialized equipment, or human annotators. Self-training, one of the most well-liked semi-supervised learning techniques, co-training, and generative models such as autoencoders and variational autoencoders. These algorithms use the labeled data to initially train the model and then use the unstructured data to fine-tune the model. Figure 2.5 describes the process working principle.

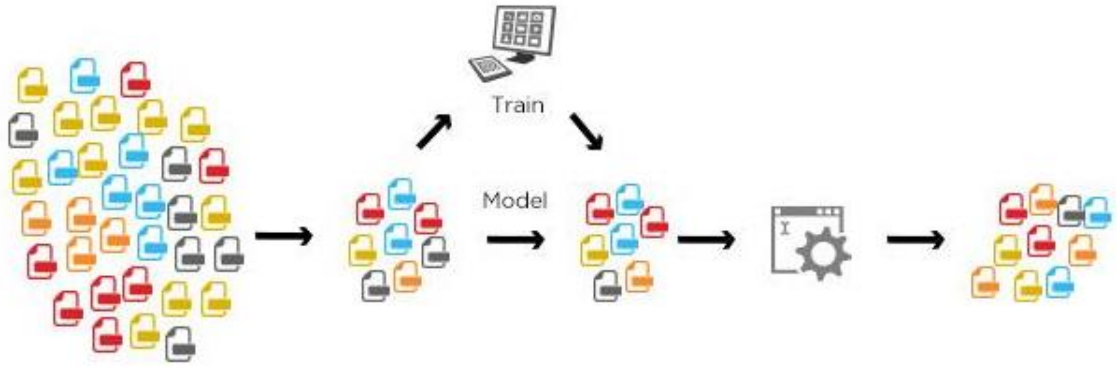


Figure 2.5 Semi-supervised ML algorithm workflow. Source: www.medium.com

Overall, semi-supervised learning is a promising approach to machine learning that can help improve the accuracy and generalization of models while reducing the cost and effort of obtaining labeled data.

Reinforcement Learning: Reinforcement Learning is a type of machine learning algorithm where an agent learns to take actions in an environment to maximize a reward signal. The goal of reinforcement learning is to develop an optimal policy or decision-making strategy that maximizes the cumulative reward over time.

In reinforcement learning, the agent interacts with an environment through a series of actions and observations. The environment provides the agent with feedback in the form of a reward signal, which indicates the quality of the action taken by the agent. The agent's objective is to learn a policy that maximizes the expected cumulative reward over time. It is shown in Figure 2.6.

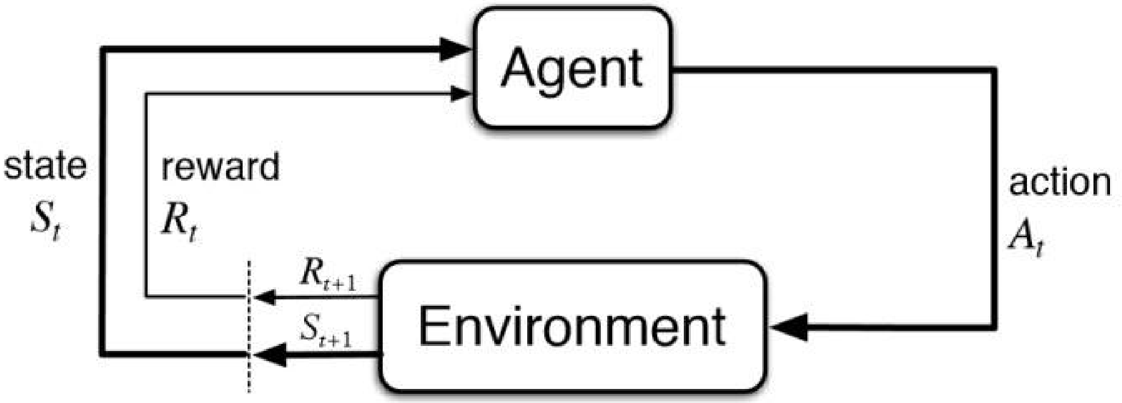


Figure 2.6 Basic Diagram of Reinforcement Learning algorithm.

Source: www.analyticsvidhya.com

Reinforcement learning is different from supervised learning in that the training data is not explicitly labeled. Instead, the agent learns from the feedback provided by the environment. The agent uses a trial-and-error approach to explore the environment and learn from its experience.

Reinforcement learning has various applications, including robotics, game-playing, and recommendation systems. For example, in robotics, reinforcement learning can be used to train robots to perform complex tasks such as navigation, grasping, and manipulation. In-game playing, reinforcement learning can be used to develop strategies for playing games such as chess and Go. In recommendation systems, reinforcement learning can be used to optimize personalized recommendations for users.

Reinforcement learning algorithms include Q-learning, SARSA, and policy gradient methods. The choice of algorithm depends on the type of environment, the reward signal, and the learning objective.

2.3.3 Machine learning Pipeline process

The data analytics process in machine learning is a process that includes a series of data processing and model-building steps. It takes data, prepares it, trains the model, and finally makes predictions. The data analytics process may vary depending on the complexity and data structure of the project, but in general, these steps provide a guide. The following steps explain the process of data analytics in detail:

- Data Collection and Cleansing: Data collection involves collecting data from different data sources depending on the purpose of your project. Data cleaning is performed to detect and correct missing or erroneous data in the data set.
- Data Discovery and Analysis: Examining the data set is important to understand the structure and properties of the data. This step is accomplished using data visualization.
- Data Preprocessing: During the data preprocessing stage, various operations are performed to make the data suitable for model training. For example, feature engineering can be done, categorical data can be digitized and data can be normalized.
- Data Partition: The data set is usually divided into training, validation, and testing data. This allows the model to use separate datasets for training and evaluation.
- Model Selection: It is important to choose a model that suits the needs and data structure of your project. Example models include Logistic Regression, Decision trees, XGBoost support vector machines, and deep learning models.
- Model Tutorial: It is used to train the selected model on training data. The model analyzes data and learns patterns.
- Model Evaluation: Validation data is used to evaluate the performance of the model. Common evaluation metrics include accuracy, intercept squared error (MSE), mean absolute error (MAE), precision, sensitivity, and F1 score.
- Model Adjustment: Hyperparameter tuning can be done to improve model performance. This includes learning rate, network structure, network depth, etc. It includes optimizing parameters such as.
- Model Production: After training the model, you can use it to predict real-world data.
- Interpretation of Results: The results of the model should be interpreted to understand how it impacts your business problem. It is important to translate results into business decisions.
- Deployment and Maintenance: If you want to use the model within a product or service, it may be necessary to deploy and maintain it on an ongoing basis. Updates should be made regularly with new data.

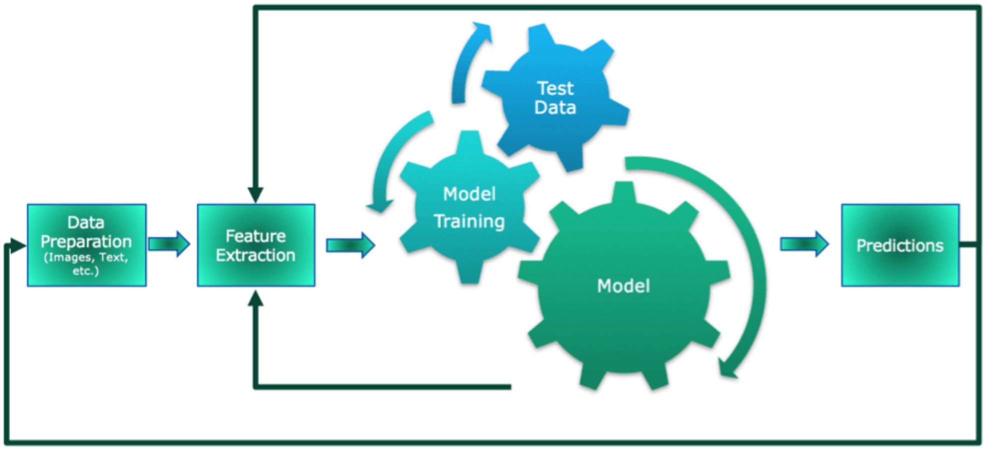


Figure 2.7 Standard ML pipeline. Source: www.blog.westerndigital.com

The machine learning pipeline can be automated using various tools and frameworks such as Apache Spark, TensorFlow, and Scikit-learn. Automation of the pipeline enables the entire process to be performed faster, with more accuracy, and with less human intervention.

2.3.4 Data Collection and Preparation Process

Data collection is the first step in the ML pipeline and is a process in which information obtained from various sources is stored in a suitable format. This phase includes identifying data sources and collecting the necessary data to train the ML model. The data collection process can be time-consuming and the quality of the data collected can significantly impact the accuracy of the model. Therefore, it is of great importance to choose data sources and ensure that the data collected is reliable, relevant, and representative of the problem area.

Methods include web scraping, API calls, sensor data, surveys, and public data sources. These methods include extracting information using a variety of tools and platforms, pulling data from third-party APIs, collecting data from sensors, collecting information through surveys, and obtaining information from public databases. Collected data is usually stored in databases or file systems. The data is then processed to prepare for analysis and transformed for use in machine learning processes.

Data preparation is a crucial step in the machine learning pipeline, as it involves cleaning, transforming, and organizing the raw data to make it suitable for analysis. Data preparation ensures that the data is of high quality, consistent, and relevant to the machine-learning task at hand. The following are the main steps involved in data preparation:

- **Data cleaning:** This involves removing irrelevant or duplicate data, dealing with missing or null values, and handling outliers.
- **Data transformation:** This involves converting the data into a suitable format for analysis. It includes steps such as encoding categorical variables, normalizing or scaling numerical variables, and reducing the dimensionality of the data.
- **Data splitting:** This involves dividing the data into a training set and a test set. The training set is used to train the machine learning model, while the test set is used to evaluate its performance.
- **Data augmentation:** This involves generating new data from the existing data to increase the size and diversity of the dataset. It is especially useful when the dataset is small or imbalanced.
- **Data validation:** This involves checking the quality and consistency of the data. It includes steps such as cross-validation, which ensures that the model is robust to variations in the data, and outlier detection, which identifies data points that are significantly different from the rest of the data.

The quality of the data used to train a machine learning model has a significant impact on its performance. Therefore, it is important to devote sufficient time and resources to data preparation to ensure that the data is of high quality and suitable for the intended purpose.

2.3.5 Feature extraction

Feature extraction is a process in machine learning that involves selecting and transforming the most relevant features from the raw data to improve the performance of a machine learning model. Feature extraction is important because it helps to reduce the dimensionality of the data while retaining the most important information.

Feature extraction involves the following steps:

- Feature selection: This involves selecting the most relevant features from the raw data. The selection of features is based on their relevance to the problem at hand, their correlation with other features, and their ability to discriminate between different classes or categories.
- Feature transformation: This involves transforming the selected features into a new representation that is more suitable for analysis. The transformation can be linear or nonlinear and can involve techniques such as scaling, normalization, principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE).
- Feature engineering: This involves creating new features from the existing ones to improve the performance of the model. Feature engineering can be done manually or automatically and involves techniques such as polynomial features, interaction terms, and feature crosses

Feature extraction can be done manually or automatically. Manual feature extraction is time-consuming and requires domain knowledge and expertise. Automatic feature extraction, on the other hand, uses machine learning algorithms to learn the most relevant features from the raw data. Automatic feature extraction is useful when the number of features is large or when domain knowledge is limited.

Feature extraction is a crucial step in machine learning as it helps to reduce the dimensionality of the data while retaining the most important information. Feature extraction can be done manually or automatically, depending on the size and complexity of the dataset and the availability of domain knowledge.

2.3.6 Model selection

Model selection is the process of selecting the most appropriate machine learning algorithm for the given task. Model selection is a crucial step in the machine learning pipeline, as the choice of algorithm can have a significant impact on the performance of the model. The following are the main factors to consider when selecting a machine-learning algorithm:

- Type of problem: The choice of algorithm depends on the type of problem, i.e., classification, regression, clustering, or anomaly detection. Each problem requires a different type of algorithm.
- Size and complexity of data: The choice of algorithm also depends on the size and complexity of the data. For example, decision trees and random forests work well for small datasets, while deep learning algorithms work well for large and complex datasets.

- Accuracy and interpretability: Some algorithms such as decision trees and logistic regression are highly interpretable, while others such as deep learning algorithms are less interpretable but more accurate.
- Training time and resource requirements: The choice of algorithm also depends on the available resources such as time, computational power, and memory. Some algorithms such as linear regression are fast and require less computational power, while others such as deep learning algorithms are slow and require a lot of computational power.
- Hyperparameters: Each algorithm has hyperparameters that need to be tuned to achieve optimal performance. The choice of algorithm also depends on the ease of tuning the hyperparameters.

Some commonly used machine learning algorithms include linear regression, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, and neural networks. The choice of algorithm depends on the specific requirements of the task at hand. It is common practice to experiment with multiple algorithms and compare their performance before selecting the best one.

2.3.7 Model evaluation and training

The term “ML model” refers to the model artifact created by the training process. The ML algorithm works by examining a large number of samples to build a predictive model and trying to minimize losses. This loss is a constant value that measures the predictive ability of the model. The loss value decreases with improved model predictions. The goal of the training procedure is to find a combination of weights and biases that provides the smallest loss of a given algorithm for all samples.

After the stage in which the data is prepared and the appropriate algorithm is selected, come the steps in which the model is trained and its performance is evaluated on the test data set. In the model training phase, the selected algorithm learns the patterns in the training data set. Model parameters are iteratively adjusted to minimize the error between predicted values and actual values in the training data set.

To evaluate the performance of the model, model validation, model evaluation, and model improvement steps are followed. These steps include parameter adjustments to improve the model through evaluations on validation and test data sets. Finally, the model is ready for deployment when its performance is satisfactory. Deployment involves the process of making the model available in various ways, such as embedding it in a software application, offering it as a web service, or distributing it as an API.

The performance of the model depends on factors that are called hyperparameters. Here are some commonly used model evaluation factors:

- **Accuracy:** It expresses the ratio of correctly predicted samples to the total number of samples. However, accuracy alone may not be sufficient in unbalanced class distributions.
- **Precision:** It expresses the ratio of predicted examples belonging to a certain class among all predicted examples belonging to that class. High precision means the false positive rate is low.
- **Recall (Recall or Sensitivity):** It expresses the ratio of real examples belonging to a certain class among all real examples belonging to that class. A high recall means the false negative rate is low.
- **F1 Score:** It is used to achieve the balance between precision and recall. It may be preferred over accuracy, especially in unbalanced class distributions.
- **ROC Curve and AUC (Receiver Operating Characteristic Curve and Area Under the Curve):** It is used to evaluate the performance of classification models. AUC refers to the area under the ROC curve, and the closer it is to 1, the better the model performs.
- **Complexity Matrix:** It is a matrix containing actual and predicted classes. Other metrics can be calculated based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values.

To achieve the best performance, it is important to carefully evaluate the model's performance and make parameter adjustments iteratively.

2.3.8 Prediction

The prediction stands as the ultimate phase in the machine learning pipeline, where the adept model comes into play to forecast outcomes on fresh, unseen data. Following the model's training and assessment on the test dataset, it transitions to making predictions on real-world data. The pivotal steps encompassed in the prediction process are:

- Model application: Post-data preparation, the proficient model is employed to make predictions on new data. The model takes input data and produces the anticipated target variable as output.
- Prediction evaluation: Following the prediction phase, the model's efficacy is scrutinized using diverse performance metrics like accuracy, precision, recall, and F1 score. As the model's performance on new data may vary from its performance on the test data, meticulous evaluation becomes imperative.
- Model refinement: In instances where the model's performance falls short, adjustments to the model parameters are made, and the training process is reiterated until the desired performance level is attained.
- Deployment: After the model undergoes training and evaluation on fresh data, it is ready for deployment through diverse channels, including integration into a software application, functioning as a web service, or deployment as an API.

Prediction constitutes the fundamental application of machine learning across sectors such as finance, healthcare, marketing, and engineering. The accuracy of predictions hinges on data quality, algorithm selection, and hyperparameters. It is crucial to systematically evaluate and refine the model's performance to achieve optimal results.

2.3.9 Visualization and Insights

For a more comprehensive grasp, let's delve into the definition: "Data visualization is the act of transforming information into a visual format, like a map or graph, to simplify data comprehension and extract insights with ease".

The primary objective of data visualization is to present information in an easily understandable manner. Visuals and forms stand out as one of the most comprehensible languages for humans. Moreover, they facilitate a deeper comprehension of the outcomes of model predictions. Despite the remarkable progress in data science and machine learning, it remains paramount to consider user expectations and experiences.

Effective visualization and insights play a crucial role in conveying the outcomes of machine learning models to stakeholders and decision-makers. These insights contribute to identifying novel patterns and relationships within the data, enhancing model accuracy, and guiding decision-making across diverse domains such as finance, healthcare, marketing, and engineering.

While the realm of machine learning encompasses a wide array of topics, this chapter offers a succinct overview of ML and its pipeline. It serves as a pivotal resource for attaining a more profound understanding of this thesis.

3. RESEARCH HAZARDOUS CASES IN BANKING

3.1 Big data and Machine Learning in the banking sector

The banking sector is one of the areas where machine learning technology is used most intensively. Thanks to machine learning, it has become possible to address challenges such as the need to generate and process large amounts of data, risk management, fraud detection, improving customer experience, and creating more efficient business processes.

Many banks use machine learning-supported chatbots and virtual assistants to improve customer service and provide personalized experiences. These AI-based systems serve as an effective tool to quickly respond to customer requests and perform basic banking transactions. Additionally, machine learning in customer relationship management provides valuable insights into customer behavior, preferences, and needs, so banks can make tailored product and service offers to their customers.

Below are examples of the use of machine learning in the banking sector:

- **Credit risk management**: is one of the key focuses of the banking industry and this is where machine learning makes a big difference. Going beyond traditional methods, machine learning algorithms analyze large amounts of data and provide more accurate results in evaluating loan applications. These algorithms, which can predict customer payment performance using historical credit data, help minimize risks while accelerating credit approval processes.
- **Fraud detection**: is a major problem that banks struggle with. While detecting fraud using traditional methods can be difficult and time-consuming, machine learning offers fast and effective solutions. These algorithms, which identify suspicious transactions by detecting anomalies in customer accounts, allow banks to take stronger action against fraud. The system can automatically identify suspected fraud by monitoring customers' account activity, habits, and unusual behavior. For example, abnormal situations such as large withdrawals or purchases made from geographical locations that differ from customer habits are identified by the algorithm and flagged as transactions with a high probability of fraud. By examining such transactions in more detail, the bank can prevent fraud situations in advance.
- **Segmentation:** Machine learning enables banks to segment customers based on data analysis and run marketing campaigns directed to the right target audience. Providing

personalized product offers based on customer behavior increases customer satisfaction and strengthens customer loyalty.

- **Credit Assessment**: Banks can determine customers' credit risk by using machine learning when evaluating loan applications. By using historical financial data and customer profiles, automated loan approval processes can be improved. Machine learning takes into account many factors such as customer history, financial situation, and risk profile when used to evaluate loan applications. Past loan payments, debt status, income level, and other credit risk indicators are analyzed by the algorithm. Machine learning algorithms can predict the probability of a particular customer's loan repayment based on historical data, and based on these predictions, banks can make their loan approval processes more efficient. This speeds up the lending process and helps make better decisions.
- **Risk Analysis and Market Forecasts:** Helps banks improve their investment strategies and portfolio management. This technology, which can predict future trends by analyzing complex data in financial markets, enables more informed and data-driven investment decisions.
- **Customer Service**: Machine learning can be used in customer service and support processes. For example, natural language processing algorithms can be used to understand customer questions and route them to the right departments. Natural Language Processing (NLP) and other machine learning techniques can be used in customer service and support processes. They can be used to understand the questions customers want to ask, analyze customer feedback, and find solutions. Automated response systems can increase customer satisfaction by responding to customer requests faster.
- **Account Management:** Machine learning can be used to manage transactions on customer accounts. Processes such as automatic money transfers, regular bill payments, and automatic savings accounts can improve banks' operational efficiency by providing better service to customers.

The use of machine learning in the banking sector brings with it some difficulties. In particular, customer data confidentiality, security, and ethical issues are among the important issues. Banks are required to use machine learning ethically by protecting customer data and complying with appropriate regulations. Additionally, the accuracy and transparency of the algorithms is also an important consideration. Because false or misleading results could undermine customer confidence or affect financial stability. Because it helps banks gain a competitive advantage and increase customer satisfaction by offering a data-focused, customer-specific, and more efficient approach.

However, it should not be forgotten that issues such as data security, ethics, and transparency should also be seriously addressed with the use of this technology.

3.2 Dangerous situations and risk factors in the banking sector

In the banking sector, it is customary to evaluate the likelihood of loan approval for clients through the utilization of pre-existing machine learning algorithms, particularly in the context of lending transactions. The algorithms assess various factors to determine the likelihood of loan approval for customers.

The bank's management's commitment to increasing returns for its stakeholders comes at the expense of heightened exposure to risk. Banks encounter a spectrum of risks, encompassing interest rate fluctuations, market instability, creditworthiness uncertainties, off-balance-sheet vulnerabilities, technology and operational hazards, foreign exchange rate volatility, sovereign or country-specific uncertainties, liquidity fluctuations, and insolvency concerns. The effective management of these risks stands as a pivotal factor in determining a bank's overall performance.

Additionally, given the multifaceted risks entailed and the pivotal role banks play within the financial sector, they draw regulatory scrutiny. Each major risk category necessitates a distinct allocation of capital. Credit risk has traditionally been the most substantial peril faced by banks, often demanding the most significant capital allocation. Market risk primarily emanates from a bank's trading activities, whereas operational risk centers on potential losses stemming from internal system failures or external events. In addition to regulatory capital calculations, most major banks also compute economic capital based on their internal models rather than relying solely on regulatory directives. It is shown in Figure 3.1.

The probability of a customer acquiring a loan can be assessed by analyzing variables such as credit score, previous loan repayments, level of indebtedness, employment situation, and income level. The aforementioned procedures have the potential to be executed in an automated manner through the utilization of a pre-existing machine-learning algorithm.

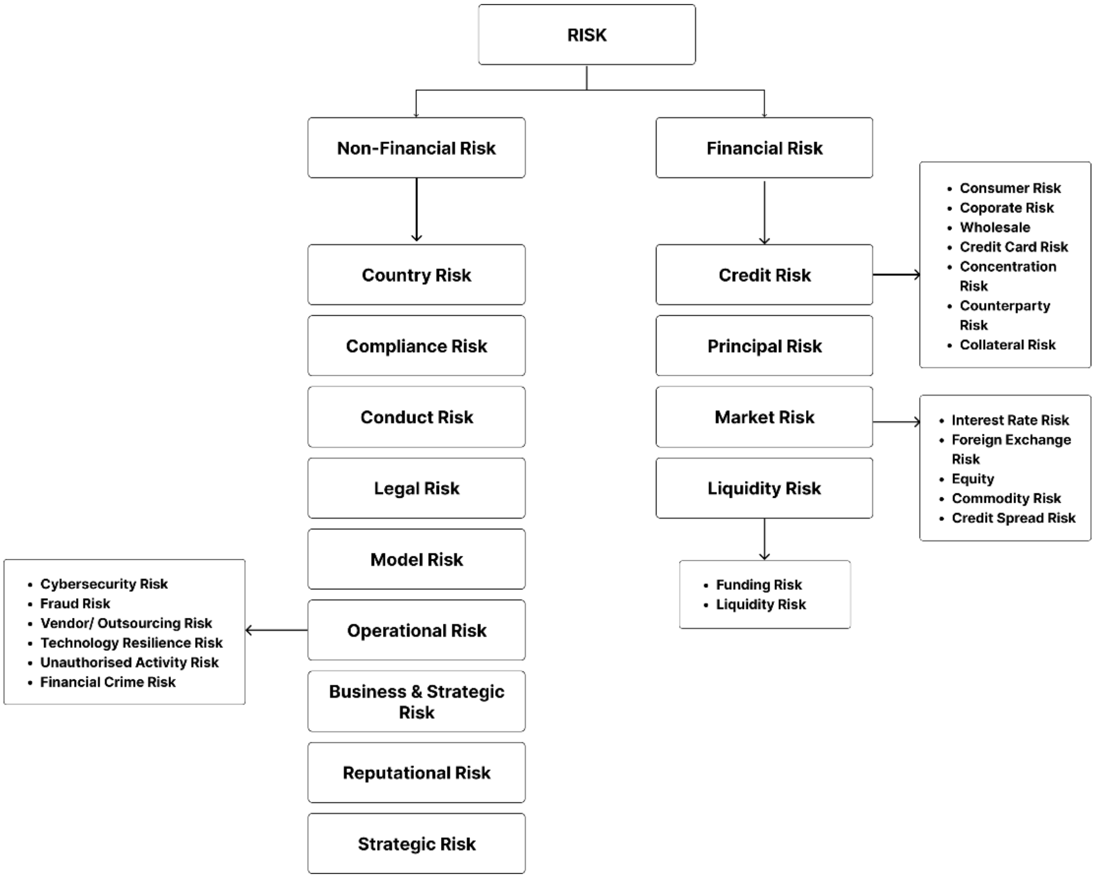


Figure 3.1 Risk Types in Finance

Generally, Machine learning algorithms that have already been taught are often employed to detect fraud in the banking sector. The aforementioned algorithms possess the capability to scrutinize and identify instances of deceitful conduct such as:

- Identity theft- refers to the act of unlawfully obtaining an individual's personal information and utilizing it to conduct financial transactions.
- Credit card fraud- refers to the act of acquiring credit card details and utilizing them to make unauthorized purchases.
- ATM fraud -refers to the illicit act of accessing automated teller machines (ATMs) through hacking or obtaining card information to withdraw funds.
- Money laundering- refers to the act of conducting financial transactions that simulate legitimate activities using funds or assets that have been acquired through criminal means.

The implementation of machine learning algorithms can facilitate the detection of fraudulent activities in the banking sector. Instances where an individual submits multiple credit card applications or attempts to access a single account from distinct IP addresses may be identified as fraudulent behavior, prompting bank authorities to be notified accordingly. Furthermore, the algorithm can identify fraudulent activity through the detection of unexpected or substantial monetary transfers within an individual's account.

Furthermore, a multitude of procedures within the banking industry, such as identifying fraudulent activity and managing customer relationships, can be classified utilizing machine learning algorithms. The utilization of algorithms for the analysis of extensive data sets enables the detection of fraudulent activities or customer behavior. There are many dangerous situations related to big data in the banking sector. Sometimes these can be identified and resolved, and sometimes we may need the help of artificial intelligence for this. Yes, artificial intelligence is used a lot in this sector, as it helps us everywhere. Below, we can see in which dangerous situations artificial intelligence comes to our aid.

3.2.1 Understanding Credit Card Fraud Situation

For various banks, the strategy of retaining high-income customers is the primary goal of these institutions. However, banking fraud poses a significant threat to this goal. This phenomenon, which is a worrying situation for both banks and customers, can cause financial losses, lack of trust and credibility problems.

According to the Nilson Report, banking frauds worldwide are expected to exceed $30 billion by 2020. With the spread of digital payment channels, the possibility of fraudulent transactions being carried out through new and various methods increases.

In the banking industry, using machine learning to detect credit card fraud is not only a trend but also a necessity to put proactive monitoring and fraud prevention mechanisms in place. With effective machine learning instead of time-consuming manual reviews, these institutions can analyze transactions faster, minimize costly chargebacks, and prevent rejections of legitimate transactions.

3.3 ML Modelling in the Azerbaijan Banking Sector

McKinsey methods find application in several banks in Azerbaijan, incorporating the utilization of Machine Learning (ML) models. While it is feasible to manually assess credit availability for customers by navigating through a set of steps, this proves arduous for banks dealing with thousands of customers due to the considerable time and resource constraints. Consequently, ML modeling emerges as a pivotal solution in these instances.

Notably, within several banks, the IBM Operational Decision Manager (ODM) platform plays a crucial role in forecasting. The credit risks are categorized by degree, customers are segmented, and diverse credit scoring methodologies, primarily application and behavioral scoring, are implemented. The procedural flow involves the drafting of a report that is subsequently forwarded to the data sector for integration. This report undergoes testing in Excel and is then segregated by customer segments.

Following this segmentation, customers are identified based on their financial number or VÖEN code, distinguishing them by company or personal credit. Subsequently, the scoring process takes place, followed by imposing restrictions. For instance, a customer might be affiliated with a blacklisted company, indicative of fraudulent activities within that company. The code examples, as illustrated earlier, demonstrate the examination of the customer's credit history (e.g., instances of being 20 days late in a month), with all these rules meticulously defined using IBM ODM.

The culmination of this process results in the establishment of a model, and based on the outcomes, the customer's loan application is either accepted or rejected. Additionally, there is a manual review of suspicious customer information at times, recognizing that instances of fraud may occur, underscoring the importance of a human-centric approach in certain scenarios.

3.4 Ml in the banking sector to predict data risks

As we have reported in previous titles, machine learning algorithms are used in the financial and banking sectors in customer services, security, credit assessment, fraud detection, customer behavior analysis, and many other areas. One of these is risk management. As we know, the bank works with customer data and the risk rate is always high in big data. Unsupervised and supervised algorithms come into play to manage this and even predict the risk. Risk management and risk detection procedures play an important role in the daily operations of banks and financial institutions. And below we can see the steps of a typical risk management and risk detection procedure.

- Data Collection: The first step is to collect the data necessary for risk management. This data may include customer information, credit history, financial status, transaction history, credit card transactions, and other related data.
- Data Cleaning and Organizing: It is important to clean and organize the collected data. Missing or incorrect data is corrected and data transformations are made when necessary.

- Data Analysis: Data analysis is used to identify risk factors and risk predictions. Machine learning algorithms are used at this stage. For example, logistic regression or random forests can be used to predict and classify risk factors.
- Risk Assessment: Risk assessment is used to evaluate customers' credit risk or other financial risks. At this stage, credit scores and risk scores are created. Algorithms can classify customers' risk levels and decisions can be made based on specific risk tolerances.
- Fraud Detection: Banks use data analysis and machine learning algorithms to prevent fraud. Algorithms that detect abnormal transaction patterns identify potential fraud attempts.
- Risk Monitoring and Management: Risk management means monitoring, analyzing, and managing risks. This ensures that risk is kept at acceptable levels. If risks exceed a certain threshold, necessary actions are taken.
- Model Update and Improvement: Risk management models should be constantly updated and improved. Models are updated as data changes, new risk factors are identified, or to improve the performance of the algorithms.
- Reporting: As part of the risk management process, regular reports are provided to regulators and senior management. These reports describe the status and management of risks.
- Decision Making: Finally, the risk management procedure involves making decisions based on the risk detection results. Decisions such as approving loan applications, setting credit limits, and handling fraud cases are made in this step.

This procedure is a general description of a financial institution's risk management process. Machine learning algorithms play an important role in data analysis and risk prediction stages and help make better decisions.

4. METHOD

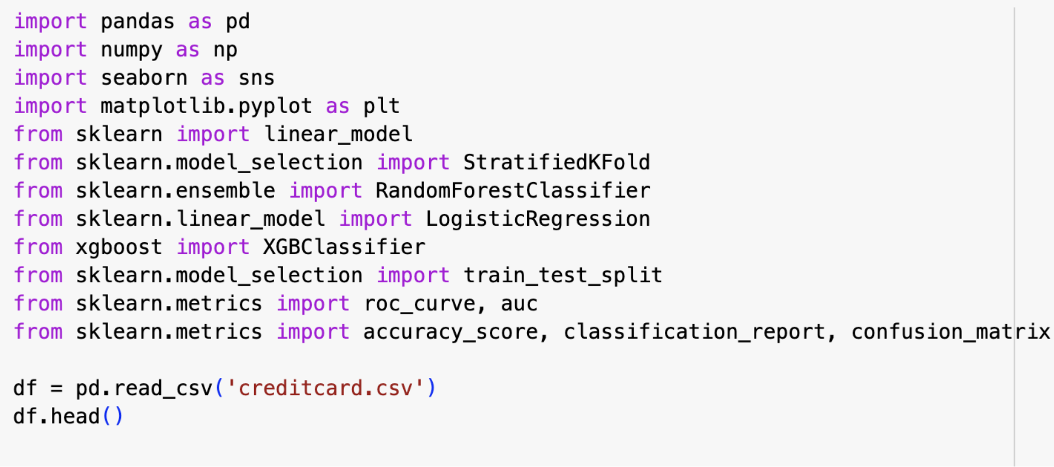
This section provides an explanation of the approach used in the project. a synopsis of the information and its processing steps prior to categorization. Next, the model's setup and performance are discussed, and then the process for determining the key recall and ROC-AUC variables is shown.

Our project aims to detected customer’s fraud situations on the credit card. Together, we aim to evaluate existing models, choose the most suitable one, and personalize these models. Rather than creating a new model, we plan to optimize our solution using industry best practices and taught models.

4.1 Libraries Setup

This machine learning modeling example uses Anaconda, the author's favorite data science distribution. Python 3.11.0, the most current and stable version of the Python language, was chosen to form the basis of this project. In addition, the Jupiter Notebook Integrated Development Environment (IDE) for the development of the project is a feature provided within the Anaconda package, and thus coding processes can be managed more efficiently and effectively.

Before starting our project, we need to add to our Python environment using the "import" option to add the necessary libraries. This step ensures that the various modules required by the project are included in the system, thus ensuring the smooth execution of the project. The import process allows us to easily integrate the extensive libraries provided by Anaconda into our project. This is important to reduce the complexity of the project and optimize the development process.



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.model_selection import StratifiedKFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

df = pd.read_csv('creditcard.csv')
df.head()
```

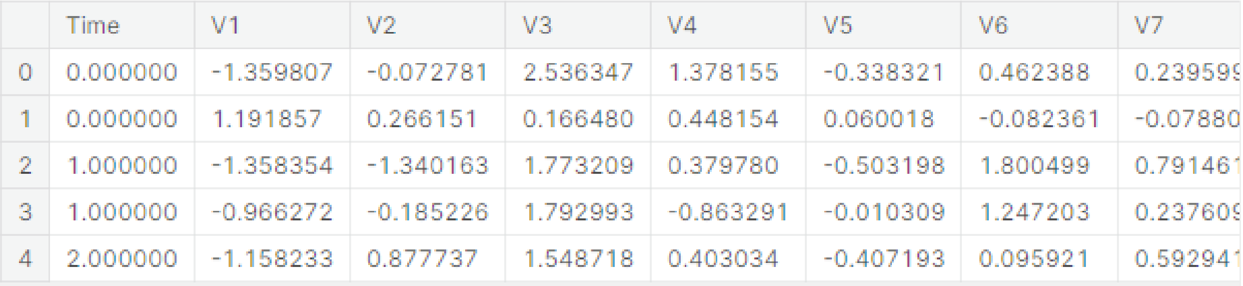
In addition to the basic libraries we imported at the beginning of the project, we further enriched our project by using the following popular libraries.

- **NumPy:** İt is a fundamental library for scientific computing and provides powerful tools for manipulating multidimensional arrays. It is often used to improve performance in numerical calculations.
- **scikit-learn:** Scikit-learn is a library for the rapid development and implementation of machine learning models. It contains many algorithms such as classification, regression, clustering, and dimensionality reduction.
- **pandas**: It is an effective and popular library used for data analysis and manipulation. Ideal for operating on data frames (DataFrames) and series.
- **matplotlib**: Matplotlib is a widely used library for creating plots, including line plots, scatter plots, histograms, and more. It can create various charts for data visualization.
- **missingno**: The library is a Python library used to visualize and understand missing data in data sets. This library offers a variety of graphs and visualizations to better understand, analyze, and handle missing data.
- **seaborn**: It is a Python library used for data visualization and is generally based on matplotlib. Seaborn is designed to create simpler and more aesthetic graphics and offers powerful tools, especially for statistical data visualization.

For modeling and tuning purposes, Scikit-Learn is essentially the recommended package for creating models. The most popular modeling tool among Python libraries is Scikit-Learn, which is particularly user-friendly and compatible with Pandas. In this case, the Pandas library was used to generate the Logistic Regression model, and Scikit-Learn was used to fine-tune it. Accompanying this, the seaborn and scikit-learn libraries were used to generate the model. Since these libraries are compatible with the modeling libraries stated above, Matplotlib, Seaborn, and Scikit-plot are often the tools of choice for data visualization.

4.2 Data Description

The data set may include a bank's customer base and financial transactions. To put it in more detail, a data set includes various demographic information of customers (age, gender, income level, etc.), financial history, credit card transactions, credit history and other potential risk factors. The dataset used here includes credit card transactions made by European cardholders in September 2013 and was retrieved from the Kaggle web library. There have been 2,84,807 transactions here in 2 days and only 492 of them are fraudulent. Since the data set is very complex and unbalanced, it is necessary to organize it and make it more trainable before creating a model. We downloaded this data set as a CSV file from Kaggle and prepared it for use in our project. A part of the data set we use is shown in Figure 4.1.



	Time	V1	V2	V3	V4	V5	V6	V7
0	0.000000	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599
1	0.000000	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.07880
2	1.000000	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461
3	1.000000	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609
4	2.000000	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941

Figure 4.1 A Piece of the Dataset

Data processing consists of six stocks as seen in Figure 4.2. During the data preprocessing stage, we must consider six key dimensions to control our data quality. By evaluating our data set on these dimensions, we can make a decision about the usability of the data. First, the data must be complete and have no missing values. Second, the dataset must be in the same format when representing values in attributes. Third, there should be no conflicting information between values that could mislead the analysis. In step number four, the accuracy and timeliness of the dataset must be ensured. Fifth, attention should be paid to duplicate values in the dataset. Finally, it is important to check for missing data or missing references.

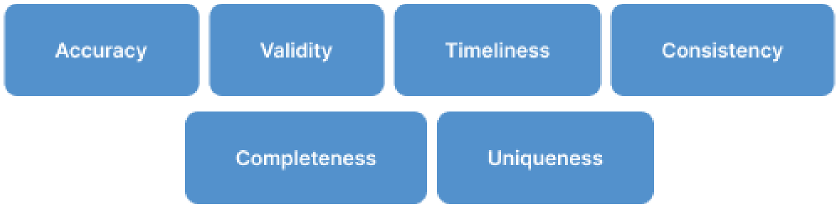
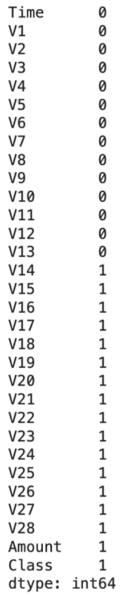


Figure 4.2 Data Processing

It is important to address missing values in our data set and fill in these gaps. For this purpose, we can analyze missing data using the *missingno* package. Let's make our data set more consistent by filling in the empty values with the mean and median. Then, we list how much data value is missing.



```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V13       0
V14       1
V15       1
V16       1
V17       1
V18       1
V19       1
V20       1
V21       1
V22       1
V23       1
V24       1
V25       1
V26       1
V27       1
V28       1
Amount    1
Class     1
dtype: int64
```

As we mentioned above, there is a limited amount of missing data in many columns in our dataset. To handle this situation and fill in missing values, we use mean and mode values. We make our dataset more consistent and meaningful by filling the NaN (Not a Number) values in the dataset with mean and mode values on a column-by-column basis.

We can delete or edit all missing values so that there are no missing values in our data set. Thus, our data set is ready to be trained for the model.

As we mentioned above, since this data set is variable, we deleted some lost data there. But sometimes the format of some data may not be the same as others. Therefore, we can convert categorical variables into numerical format.

This transformation converts the various data types in our dataset into a homogeneous numerical representation, allowing it to be seamlessly integrated into a wider range of analytics and modeling. In this way, we can make the data set ready for further analysis and discovery studies.

4.2.2 Data Visualization

Since we completed the missing data before, the absence of missing values in the data set makes modeling even easier. After deleting the deficiencies in this data, the number of transactions naturally decreased and there are more legitimate transactions than fraud cases, that is, there are

25837 non-fraud cases. We know that the number of previous legitimate transactions was over 2 million. In this section, we need packages such as *"matplotlib*" and *"seaborn*" to understand and visualize the visual information in our data set. These libraries offer powerful tools to express data more clearly through graphs and visual elements. In this way, we will be able to visually explore and understand the relationships, distributions, and other important features between the features of our data set.

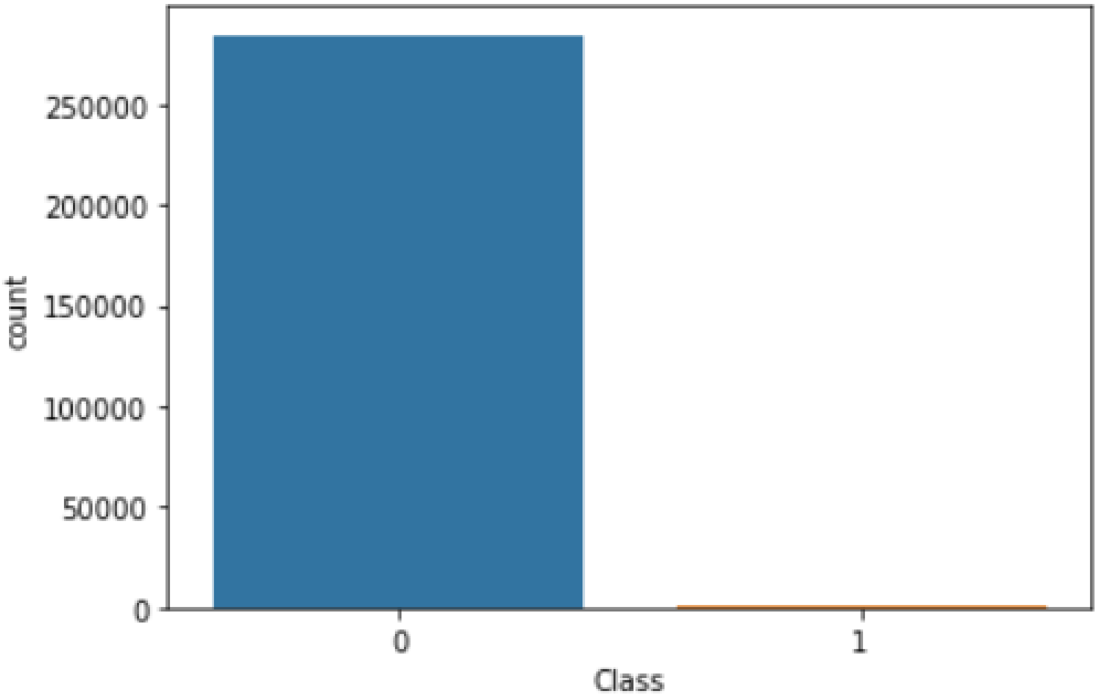


Figure 4.3 Fraud and Non-fraud cases comparing bar

As we see from Figure 4.3, the dataset has very high class imbalance. There are only 30 of 25837 records labeled as fraudulent transactions. The distribution percentage belonging to the majority class is 99.66%, and the distribution percentage belonging to the minority class is 0.34%.

4.2.3 Data matrices modelling

In this section, we learn extra information about data and its properties. Before starting our analysis, it is very important to understand the type of features and data. First of all, we are exploring object-type data. So, let's create a custom function to understand the types and numbers of unique values in columns.

In this section, designing a model data matrix to estimate the credit risk of customers in the project, the data matrix is used to include certain characteristics, usually including the customer profile and financial history. All of the independent data for each customer creates matrix X. The following data is shown in this matrix:

Customer Profile: Age, Gender, Marital Status, Education Level

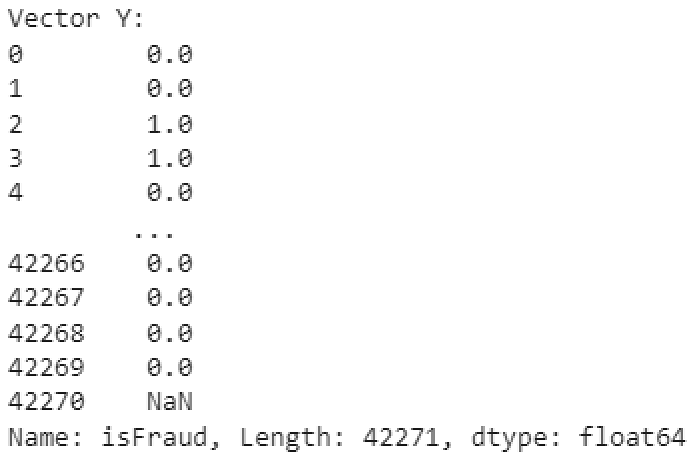
Financial Status: Income, Employment Status (Employed, Unemployed, Self-Employed, etc.), Debt Status, Asset Status (House, Vehicle, etc.)

Credit History: Credit Score, Past Loan Payments, Credit Card Transactions, Existing Credit Limits

Borrowing Status: Number of Loan Applications, Current Loan Amounts, Loan Installments

Other Factors: Market Conditions (economic situation, interest rates, etc.), Business Situation (general situation in the industry).

Dependent data: **Credit Risk** (1, Y: High Risk, 0, N: Low Risk). In the data I use here, the hanging data for the Y vector matrix is the Loan status.



```
Vector Y:
0        0.0
1        0.0
2        1.0
3        1.0
4        0.0
        ...
42266    0.0
42267    0.0
42268    0.0
42269    0.0
42270    NaN
Name: isFraud, Length: 42271, dtype: float64
```

Figure 4.4. Example of vector Y

A large part of this data set was used as training data. The remaining small part is reserved as test data.

Customer Information: Customer personal information, demographic details such as their age, gender, marital status. Information about their economic situation, such as their income level, occupation and education level.

Financial Transactions: Credit card usage, spending habits and shopping history. Detailed information about bank account movements, deposits and loans.
Credit History and Risk Factors: Customers' past credit purchases, credit scores and debt situations. Cases of delay in previous payments and the frequency of these situations.

4.3 Setup Data Classification

The applied algorithm will evaluate customers on a weekly basis and predict credit risk probabilities. This approach allows customers to review their credit score and identify target groups, providing users with information before initiating treatment. However, data processing and modeling processes are computationally costly and there are some delays in recording data into the system, so daily customer evaluation is not always possible. Therefore, observations in the entire data set are evaluated based on the scoring date, a week before the treatment is sent. This process only occurs during the training phase and once the model is implemented the scoring date will reflect the current date. The training phase of the classification model can be divided into five parts:

- Data collection and processing
- Training, Validation and Testing phase
- Incomplete sampling in the training set
- Use a validation set to fine-tune the model after training on it.
- Using a test set to evaluate the most successful model

To successfully estimate fraud risk probabilities, data is divided into several pieces for several different purposes. There are two main areas of partitioning, namely train-test separation and sampling in unbalanced classes. So the data will be divided into a training set, validation set and testing set. The notable point is that the partitioning is done after combining the X and Z data matrices; and this means that no historical information is lost for any observation. Most machine learning methods use the ability to split data into a training and validation set to evaluate performance. For example, it is common to randomly split a dataset into 80% training and 20% validation data with the corresponding response within a predetermined time frame. However, since the purpose of the thesis project is to predict the future risk of customers' financial data, we will use 70% for training and 30% for validation and testing (15%-15%).

4.3.1 Classification Models

Three different classifiers were used to predict fraud risk. First a single Decision Tree and secondly a Logistic Regression classifier and our 3rd model, XGBoost was defined in previous sections. But if we give some detailed information:

Decision Tree Model: Decision trees create a tree-like model that represents the relationships between a series of decisions and their outcomes. Decision trees usually consist of many stages. Each stage represents a decision point on a feature in the dataset. The formula is generally as follows:

$$f(x) = \sum_{i=1}^{M} ci \, . \, I \, (x \in R \, i)$$

Here:

f(*x*): Input data sample and predicted output for x.

M: Number of terminal nodes in the tree.

R_i: The region where the samples in each terminal node are located.

C_i : Estimated output for each region.

The decision tree model is built based on a pre-labeled training dataset. It learns the relationships between features and labels in this dataset.

Logistic Regression Classifier: Although it is generally used for binary classification, it can also be applied to multi-class classification problems. Logistic regression focuses on predicting the probability of an event using a combination of independent variables. Logistic regression uses the sigmoid (logistic) function. The sigmoid function converts any real number from the range (−∞, +∞) to the range (0, 1).

In the training phase, the logistic regression model learns the parameters that best predict the probability of the dependent variable being 1.

In the prediction phase, the model makes probability predictions for new data samples using the learned parameters. The threshold value is generally considered to be 0.5, probabilities greater than this value are assigned to class 1, and those less than this value are assigned to class 0. The performance of the model is evaluated using metrics such as classification accuracy, sensitivity, sensitivity (recall), and specificity.

XGBoost Model: XGBoost (Extreme Gradient Boosting) is a powerful tree-based machine learning algorithm specifically designed to achieve high performance on structured data sets. XGBoost has a number of enhancements and improvements based on the gradient boosting method. XGBoost is an advanced version of the gradient boosting algorithm. Gradient boosting creates a strong model by combining weak learners (usually decision trees). XGBoost optimizes this process and provides faster training times. XGBoost reduces training times by effectively using parallel processing capabilities. Additionally, it has a code base specifically designed for fast calculations.

XGBoost has been an algorithm often used successfully in competitions and industrial applications. It is popular due to its high performance on various types of problems and its ability to obtain effective results on large-scale data sets.

Two models were trained with and without the X matrix and then compared between the two setups. This is because the variables in These features are called simulated variables because they are possible future outcomes, but are actually unknown until the actual context is created. Models were trained with and without these features to demonstrate the impact of the simulated variables. Not only the independent effect but also the interaction effect with customer demographics at Z were considered. How this is resolved in terms of scoring is explained in detail in Section 4.4The models were then evaluated by looking at the ROC-AUC score. Model metrics from the Confusion Matrix were also used to evaluate performance. Predicted probabilities for training epochs were visualized via distribution plots to see the separation of predicted probabilities between the two classes.

ROC curve AUC value: ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) is a metric used to evaluate the performance of classification models. ROC-AUC is a reliable metric for evaluating the performance of classification models and is an effective tool for understanding how the model performs at different cutoff points. The ROC curve shows the relationship between sensitivity and specificity, and the AUC represents the area under the ROC curve.

The ROC curve is a graph of the classification model showing sensitivity and specificity values at different cutoff points. Examining the curve to understand the performance of the model is useful to understand how a classifier performs at different threshold values.

Sensitivity: It refers to the true positive rate. That is, it is the ratio of true positives to the total number of positives.

Specificity: It refers to the true negative rate. That is, it is the ratio of true negatives to the total number of negatives.

AUC (Area Under the Curve): It refers to the area under the ROC curve. The AUC value is a number between 0 and 1. A high AUC value indicates that the model has good performance. If the

AUC is very close to 1, it means that the model is performing perfectly. It evaluates sensitivity and specificity values in a balanced way in unbalanced data sets.

Confusion Matrix: The confusion matrix represents a 2*2 matrix. TPR, TNR, FPR and FNR are the four outputs of the matrix. Features, accuracy, and false alarm rate can all be calculated using classification models. Next we will see what suits us best to detect credit card fraud:

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Sensitivity = TP/ (TP+FP)

F1 score = 2((sensitivity * recall) / (sensitivity + recall))

TP - True Positive

TN - True Negative

FP - False Positive

FN - False Negative

4.3.2 Tuning

Logistic Regression and Decision Tree models were evaluated with a random layer approach for use in the cross-validation process, these details are explained in detail in Appendix A.1. The Holdout Verification method was preferred in this evaluation process. Holdout validation divides the data set into three basic parts: training, validation and test set. The model learns on the training and validation set and then evaluates its performance on the test set.

The main goal in this process was to maximize the AUC (Area Under Curve) value of the model. However, in achieving this goal, attempts have been made to avoid the problem of overfitting. Overfitting refers to the situation where the model overfits the training data and loses the ability to generalize to real-world data. Therefore, a balanced approach was adopted to improve the AUC performance of the model and at the same time efforts were made to avoid overfitting.

Because holdout validation divides the data set into three parts, the training and validation process performs the training of the model while the test set provides an objective assessment of performance. In this way, how well the model generalizes to real-world data can be more robustly evaluated. As a result, the cross-validation process of Logistic regression and Decision tree models adopted a careful balancing approach in order to obtain an optimal AUC value while minimizing the tendency for the model to overfit.

4.4 Model Scoring and Training

If any bank wants to predict the likelihood of fraud before giving a specific response to a customer, the features in matrix X will not be available. Because the observations in this matrix are observation specific and are not specified until the communication is sent, live scoring is not possible. To resolve this during scoring, all combinations of a subset of features in X are scored as separate observations. To compare performance with and without observation-specific features, models were also trained without features in X. When the model begins to lack performance in predicting test data, the model will be retrained on data from more recent periods. The frequency of scoring and retraining can be adjusted, leading to flexibility in the use of the model. Scoring is done with the following basic steps:

- **Data Collection and Processing:** First, collect and process data from the current date up to a certain period of time. This may include customer transactions, credit history, account activities, and similar data.
- **Model Training:** Train a fraud scoring model using the features you specify. This often involves using machine learning algorithms.
- **Loading the Model:** Integrate the trained model into a system or infrastructure. This means the model can interact with live data and take action in real time.
- **Scoring Process:** Score all possible combinations using your model. This may include live transactions, new customer applications, or transactions within a specific time period. The model will score the likelihood of fraud for a particular customer.

When the model begins to perform less accurately, retrain the model by following the steps to trin the model as described in Section 4.3

5. RESULTS

This text focuses on the results presented. Results were made with simulated variables for three models. Models without simulated features are also offered. Classification performance varies depending on the selected threshold value. ROC AUC curves are shown for each of the 3 models (Logistic Regression, Decision Tree and XGBoost). Models were also evaluated according to confusion matrix and classification report data.

5.1 XGBoost Modelling

An unspecified decision tree serves as the standard model. The graphs below are the ROC curve and AUC score for the decision tree model.

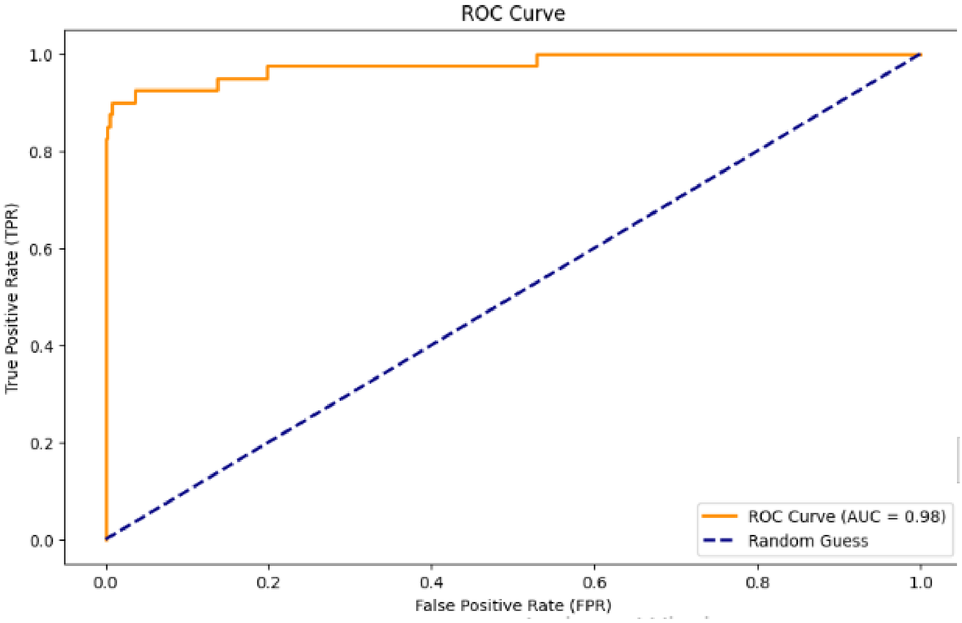


Figure 5.1 ROC curve of XGBoost model

The *Accuracy rate* score of this model: `0.9992888544091026`. The Table 1 displays the adjusted confusion matrix and classification report for the whole decision tree model.

Table 1. a) Confusion matrix and b) Classification report

Table 5.1. Confusion matrix and Classification reports of XGBoost modelling

	0	1
0	15427	1
1	10	30

	precision	recall	f1-score
0	1	1	1
1	0.97	0.75	0.85

Apparently the model has been maintained to a fairly high accuracy. However, since there is class imbalance, the accuracy value alone may not fully reflect the performance of the model. Confusion Matrix and Classification Report outputs provide a more detailed evaluation. Based on these outputs, the model shows a decrease in recall for class 1. Since it is important to be able to better recognize class 1 examples given the class imbalance, we performed a recall-oriented evaluation to improve the model. Of course, it is especially important to consider precision and recall values, because accuracy can be misleading in case of class imbalance.

5.2 Logistic Regression Modelling

Same as the first modeling technique, in the Decision tree model, 20% data was used for testing and 80% data was used for training. And its RoC curve and AUC score are shown graphically.

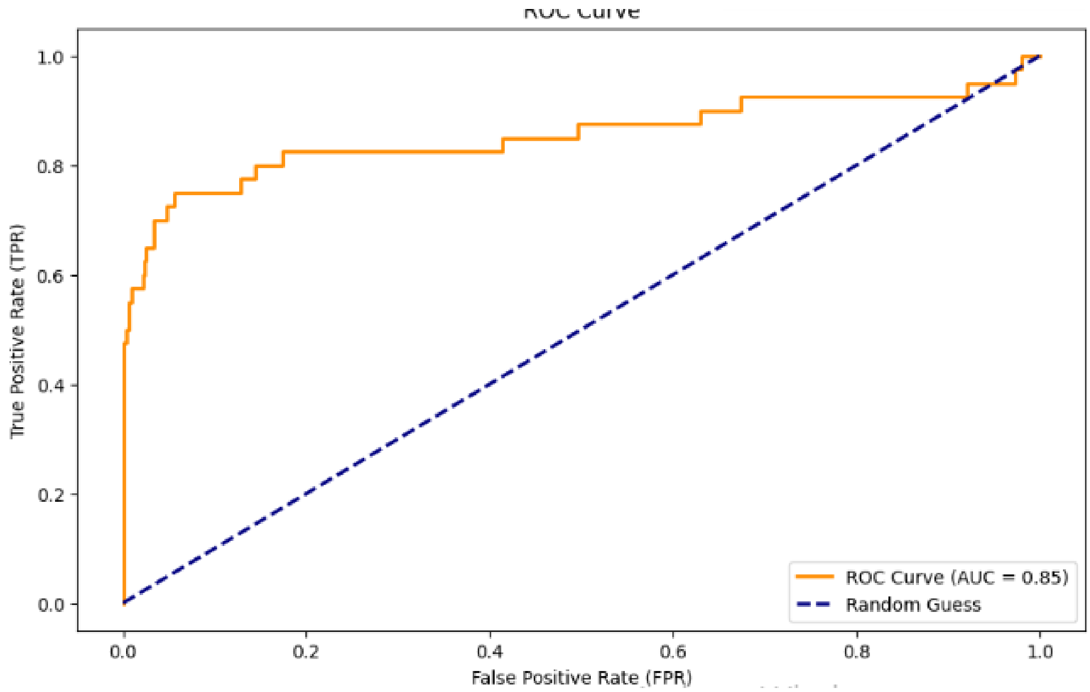


Figure 5.2 ROC curve of Logistic Regression model

Confusion matrix and Classification report table is as follows:

Table 5.2 Confusion matrix and Classification reports of Logistic Regression modelling

	0	1
0	15416	12
1	23	17

	precision	recall	f1-score
0	1	1	1
1	0.59	0.42	0.49

This is the final *Accuracy rate* score of this model: 0.997737264028963

5.3 Decision Tree Modelling

Decision tree models create a binary structure by dividing the probabilities they predict down to class purity. This means that the predictions are only 0 or 1. In this context, drawing a general distribution of probabilities makes little sense; however, significant differences between classes may be of greater interest. Since the predicted probabilities have a binary nature, the process of determining the classification threshold occurs independently of these probabilities. This provides flexibility in the process of determining class labels and allows evaluating the performance of the model in a more general context.

In Decision Tree models, estimated probabilities express the probability that an observation belongs to a particular class. These probabilities are calculated based on the class distribution at the end nodes of the tree. The scikit-learn library is often used in Python to train and use a Decision Tree model to obtain estimated probabilities.

We can see the ROC curve of these predictions below. At this stage, "ROC curve area=0.84" indicates that the classification ability of the model has been improved and its overall performance is good.

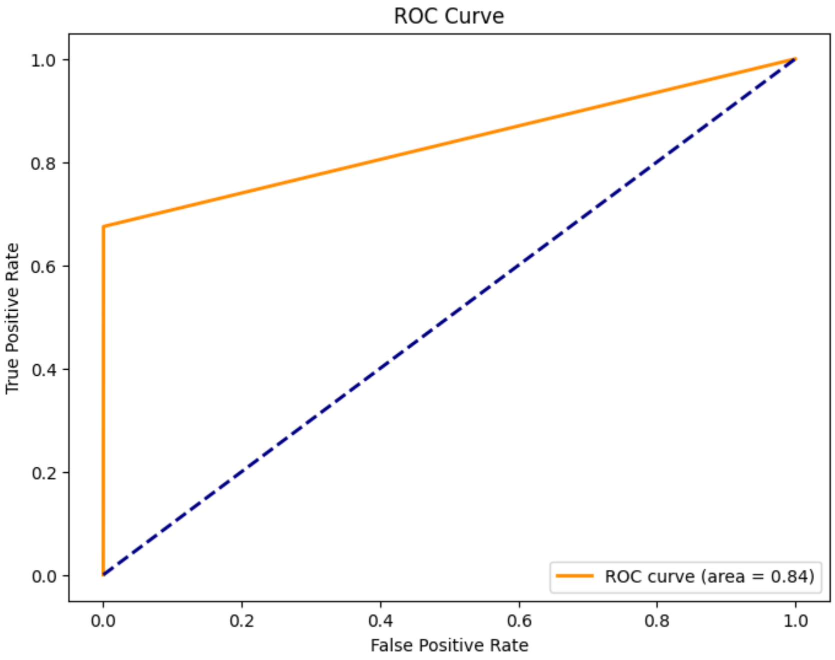


Figure 5.3 ROC curve of Decision Tree model

Table 5.3 Confusion matrix and Classification reports of Decision Tree modelling

	0	1
0	15423	5
1	13	27

	precision	recall	f1-score
0	1	1	1
1	0.84	0.68	0.75

This is the final *Accuracy rate* score of this model: `Accuracy: 0.9988363072148952`

6. DISCUSSION

As seen in other sections, we trained and tested our 3 machine learning models in accordance with the data set. And we got different results for each model.

XGBoost Model: Considering the high accuracy (Accuracy: 0.9993) and low false positive rates, your XGBoost model looks quite successful. However, considering the class imbalance, f1-score and recall values are also important. These metrics show the model's performance in the minority class. The recall value may be low, in which case we should focus on better classifying examples in the minority class.

Logistic Regression Model: High accuracy (Accuracy: 0.9977) but low recall and f1-score values in the minority class. This shows that the model has difficulty in correctly classifying the minority class. We can improve this situation by further tuning or feature engineering.

Decision Tree Model: High accuracy (Accuracy: 0.9988) and better recall and f1-score values in the minority class. Your Decision Tree model also seems generally successful. However, there may be a risk of overfitting, we can review hyperparameters such as tree depth to control this situation.

The results are generally successful, but it is important to understand the weaknesses of each model and work on them to improve them.

6.1 Models improving

The optimization of each model may vary depending on the type of model and the data set it is used on. So by reviewing all the improvement methods, we can deal with class imbalance for better classification of Minority class for these 3 models. This may include methods such as oversampling or undersampling. Why did we choose the SMOTE optimization method for all three models?

For the Logistic Regression Model: Since the recall and f1-score values in the minority class are low, it may be important to deal with class imbalance. Oversampling or trying different classifiers may improve this situation.

For Decision Tree Model: Although the performance of the Decision Tree model is generally good, the recall value in the minority class is a little low. We can fix this situation with techniques such as hyperparameter adjustments or sample augmentation.

For XGBoost Model: Although the XGBoost model shows high accuracy, its recall value in the minority class is low. We can make hyperparameter adjustments and try sample augmentation techniques to better classify the minority class.

If we come to such a conclusion after the research and experience, when we look at the results of the situation that we experienced for 3 models in the first place and trained on the data set, the XGBoost model was evaluated as the better among the models. However, when we retrained and checked it later using Smote technique, the Decision Tree model was evaluated as better. However, when we looked not only at the accuracy value but also at other values, there was very little difference between the first attempt and the re-improved values.
To find the statistical range of these results, we need to evaluate Precision and Confidence Intervals: Presenting your classification metrics (precision, recall, accuracy, etc.) with precision and confidence intervals shows how reliable the predictions are at a certain confidence level.

6.2 Results

Following the implementation of the SMOTE technique to enhance the model's refinement, our next crucial step involved a meticulous evaluation to ascertain the actual improvement achieved by each of these refined models. This evaluation necessitated a comprehensive training session for the models. Through this process, we aimed to discern and validate the effectiveness of the refinement introduced by the SMOTE technique.

The iterative nature of this methodology allowed us to not only enhance the model but also meticulously track and measure the extent of this enhancement. By retraining the models, we sought to instill a higher level of accuracy and reliability. Post-training, a thorough re-evaluation of the model's performance metrics was conducted. This encompassed a detailed examination of various aspects, including but not limited to precision, recall, and F1-score.

Our objective was to ensure that the refinement induced by the SMOTE technique translated into tangible and quantifiable improvements across these performance metrics. Rigorous checks were performed to validate if the models exhibited enhanced capabilities, particularly in handling imbalanced datasets, which is a critical aspect addressed by the SMOTE technique.

This comprehensive evaluation allowed us not only to gauge the immediate impact of the refinement but also to lay the foundation for a robust and continuously improving modeling process. The emphasis was not just on the implementation of a technique but on the sustained and validated enhancement of our models, ensuring their effectiveness in real-world applications.

Table 6.1 Final Results of three models

	Accuracy	***precision***	***recall***	***f1-score***
XGBoost	0.99%	0.99%	0.98%	0.99%
Logistic Regression	0.98%	0.99%	0.98%	0.99%
Decision Tree	0.99%	0.99%	0.99%	0.99%

In this section, we first improved them using the Smote technique and then retrained the 3 models. We double-checked the trained models and the results are shown in the Table 3.
These results reveal the precision and recall performance of each model. Additionally, the confidence intervals calculated for each metric indicate the statistical uncertainty of these performance measures. Here's what these results mean: These results show that the XGBoost model has a high success rate in predicting class 1 correctly. Sensitivity and recall values are around 0.89. Confidence intervals are also quite wide, indicating that the model's performance is within statistical uncertainty.

Logistic Regression Model Performance: While the Logistic Regression model shows low sensitivity for class 1, its recall value is high. The sensitivity value is particularly low and the confidence interval has a rather narrow range.

Decision Tree Model Performance: The Decision Tree model performs similar to the XGBoost model. Sensitivity and recall values are high, but confidence intervals are wide.

Statistical uncertainty may vary depending on the sample size and characteristics of your data set. Considering precision and recall values, XGBoost and Decision Tree models show similar performance. Both values are important here. However, the recall value shows the rate at which true positive samples are correctly predicted. In the case of credit card fraud, it is important not to miss the real positives (correctly detecting fraud). Therefore, a high recall value is desirable.

Therefore, as a result, Decision tree and XGBoost models with large recall values are in the same equation.

FUTURE WORK

In this thesis, we presented a detailed analysis on customer behavior using big data analytics and machine learning techniques. However, we think there is ample room for future research and development. In this section, we discuss potential future research directions, focusing specifically on risks in the banking sector and the predictive capabilities of machine learning.

One future work is to investigate the use of deep learning techniques in assessing risks. Deep learning is known for its ability to model and solve complex problems using neural networks. Today, deep learning has achieved significant success in many areas such as image recognition, natural language processing and speech recognition. Using deep learning techniques in risk assessment can provide more precise and in-depth insights.

Additionally, a promising area for future research is examining the use of data visualization techniques in risk analysis. Data visualization techniques such as heat maps, scatter plots, and line charts can be effective for revealing hidden patterns and insights in large data sets. By using data visualization techniques, risks can be better understood and more effective risk management strategies can be developed.

The use of unsupervised learning techniques in risk analysis is also an interesting direction for future research. Using these techniques, hidden relationships between risk and other variables can be revealed.

Examining the impact of different data sources on risk analysis is also a potential area of research. This thesis used data from a single source, but in, reality customer data for risk analysis may have come from a variety of sources such as social media, customer ,feedback and sales data. Understanding the impact of different data sources can help develop more accurate and comprehensive risk models.

Finally, exploring ethical dimensions may also be an important area for future research. With increasing concerns about privacy and protection of data used for risk analysis, it is important to develop ethical rules in this area.

This thesis presented a comprehensive review of risk analysis using big data analytics and machine learning. However, there are still many opportunities for future research. There is potential for further exploration and development in areas such as deep learning, data visualization techniques, unsupervised learning, disparate data sources, and the ethics of risk analysis.

ADVANTAGES AND DISADVANTAGES

As with any methodology, there exist both merits and demerits in employing these advanced techniques. In this section, we will examine the advantages and disadvantages of applying big data analytics and machine learning in the context of predicting risks and threats in banks.

Advantages:

- **Fast and Automated Decisions:** Machine learning models can predict customer risk by quickly analyzing large data sets. This enables faster and more automated decisions compared to manual evaluations.
- **More Precise Predictions:** Machine learning algorithms can make more precise predictions with the ability to detect complex relationships and patterns. This can help the bank assess customer risk more effectively.
- **Personalized Risk Assessment:** Machine learning can perform personalized risk assessments by building customer-specific models. By being better adapted to each customer's specific situation, these models can provide more accurate forecasts.
- **Data-Based Strategic Planning:** Insights obtained through machine learning can influence banks' future strategic planning processes. Accurate analysis can help banks develop and adapt their risk management strategies.

Disadvantages:

- **Data Quality and Representativeness:** Machine learning models are highly dependent on the quality and representativeness of the dataset. If the data set used contains missing or incorrect information, the accuracy of the model may be negatively affected.
- **Requirement for Human Intervention:** Building and training machine learning models should be done by human experts. Additionally, human expertise may be needed for the understandability and interpretability of the predictions produced by the model.
- **Privacy and Security Concerns:** Using customer data may raise privacy and security concerns. This may require compliance with stringent regulations on data protection and privacy.
- **Insufficient Model Understanding Ability:** Machine learning models can often be complex and difficult to understand. Understanding exactly how the model works can be challenging for banking professionals.

- **Biases in Training Data:** Biases in training data can also cause the model to make biased predictions. This may lead to the model being generalized and unfair to different groups.

By recognizing these pros and cons, banks can make more informed decisions about the strategic implementation of big data analytics and machine learning in the analysis of customer behavior. This awareness allows for a more balanced and judicious approach, addressing challenges while leveraging the advantages of these advanced technologies in the banking sector.

CONCLUSION

In conclusion, this study delves into the potential of machine development models in predicting and managing risks within the banking sector. The extensive analysis of dynamic and large datasets reveals the abundance of data in the banking industry, encompassing customer information, financial transactions, and economic indicators. The primary focus of this research is on utilizing machine learning techniques, particularly the logistic regression model, to predict the probabilities of credit currency reception.

The logistic regression model serves as a powerful tool, considering various variables in the dataset to predict positive or negative reactions indicative of credit score deterioration. The obtained accuracy rates, while critical, are not merely evaluative but serve as a benchmark for the model's effectiveness.

The study is not confined solely to logistic regression; it explores the potential of other tools and libraries within Pmachine-learninglearning ecosystem. While logistic regression forms the core of the research, the study lays the groundwork for future exploration of different models and techniques within the realm of machine learning.

Through the practice of coding in Python and the emphasis on the logistic regression model, this thesis provides a comprehensive foundation for developing data analysis and machine learning models. It not only ensures clarity in the methodological approach of the thesis but also paves the way for further research.

Importantly, this thesis does not only address the current risk landscape but also anticipates changing risks in the banking sector. By identifying potential future challenges, it contributes to building a more resilient and adaptable structure within the industry.

In its entirety, this thesis aims to answer whether big data and machine development models can fortify risk management strategies in the banking sector. The insights gained from this research can prove valuable to industry professionals, regulators, and researchers, offering a proactive approach to risk identification and management.

With 5 chapters, 29 subchapters, an introduction, a robust conclusion, and comprehensive references, this thesis stands as a thorough exploration of the intersection of machine learning and risk management in the banking industry.

REFERENCES

1. Johnson and Smith et al. Identifying patterns and predictors of lifestyle modification in electronic health record documentation using statistical and machine learning methods(2020). https://doi.org/10.1016/j.ypmed.2020.106061
2. Brkić et al. "Detecting CreditCard Fraud using selected Machine Learning Algorithms" (2019). doi: 10.23919/MIPRO.2019.8757212
3. Lee et al. “Applications of machine learning in addiction studies: A systematic review” (2019). https://doi.org/10.1016/j.psychres.2019.03.001
4. Garcia and Martinez, “Machine Learning Algorithms Detecting: Supervised, Unsupervised Learning” (2017).
5. Pin-Liang Chen et al. “Social Network and Consumer Behavior Analysis: A Case Study in the Shopping District “(2017) https://doi.org/10.1007/978-981-10-3187-8_84
6. Kim and Park et al. “Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer's disease” (2019). doi: 10.1016/j.nicl.2019.101811.
7. G. S. Vaishnavi Nath Dornadula et al. "Credit Card Fraud Detection using Machine Learning Algorithms" (2019). doi:10.1016/j.procs.2020.01.057
8. Prabha et al. Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach (2021) https://doi.org/10.1007/s12652-020-01961-9
9. Chien-Chang Hsu et al. “An Intelligent Interface for Customer Behaviour Analysis from Interaction Activities in Electronic Commerce” (2004) https://doi.org/10.1007/978-3-540-24677-0_33
10. Hala Z Alenzi. "Fraud Detection in Credit Cards using Logistic Regression," (2020).
11. Kshetri and Voas et al. “Blockchain in Developing Countries” (2016) https://doi.org/10.1109/MITP.2018.021921645
12. Takao Terano et al. “Agent-Based In-Store Simulator for Analyzing Customer Behaviors in a Super-Market” (2009) https://doi.org/10.1007/978-3-642-04592-9_31
13. Shalini et al. “Comparative Analysis of Clustering Techniques for Customer Behaviour” (2017) https://doi.org/10.1007/978-981-10-5699-4_71
14. Simon Denman et al. Identifying Customer Behaviour and Dwell Time Using Soft Biometrics (2012) https://doi.org/10.1007/978-3-642-28598-1_7
15. José-Ramón Segarra-Moliner et al. Engaging in customer citizenship behaviours to predict customer lifetime value (2022) https://doi.org/10.1057/s41270-022-00195-2

16. Dongyun Nie et al. From data acquisition to validation: a complete workflow for predicting individual customer lifetime value (2022). https://doi.org/10.1057/s41270-022-00197-0
17. Gershgorn, Dave. "ImageNet: the data that spawned the current AI boom — Quartz" (2017)
18. Canini et al. "Sibyl: A system for large scale supervised machine learning". Jack Baskin School of Engineering. UC Santa Cruz.(2016)
19. Geron, A. Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow. O'Reilly (2020). https://deepai.org/machine-learning-glossary-and-terms/random-forest
20. Brouwer et al. "Machine learning applications in radiation oncology: Current use and needs to support clinical implementation" (2020). https://doi.org/10.1016/j.phro.2020.11.002
21. Eibe Frank et al. "Technical Note: Naive Bayes for Regression" https://doi.org/10.1023/A:1007670802811
22. Breiman. "L. Random Forests. Machine Learning" (2001) https://doi.org/10.1023/A:1010933404324
23. Bhatore, S., Mohan, L. & Reddy, Y.R. "Machine Learning Techniques For Credit Risk Evaluation: A Systematic Literature Review" (2020). https://doi.org/10.1007/s42786-020-00020-3
24. Bustos, O. & Pomares-Quimbaya, A. "Stock Market Movement Forecast: A Systematic Review". (2020) https://doi.org/10.1016/j.eswa.2020.113464.
25. De Jesus D.P. & Besarria C.D.N. "Machine Learning and Sentiment Analysis: Projecting Bank Insolvency Risk". Research in Economics, (2023). 77(2), 226-238. DOI: 10.1016/j.rie.2023.03.001
26. Pozzolo et al. "Credit card fraud detection and concept-drift adaptation with delayed supervised information" (2015). doi: 10.1109/IJCNN.2015.7280527
27. Higgins, J.P.T., Thomas, J., Chandler, J., et al eds. "Cochrane Handbook for Systematic Reviews of Interventions": (2019) . https://training.cochrane.org/handbook
28. Yadav et al. "Deep convolutional neural network based medical image classification for disease diagnosis" (2019).
29. Huang, J., Chai, J. & Cho, S. "Deep Learning in Finance and Banking: A Literature Review and Classification". (2020) https://doi.org/10.1186/s11782
30. Ketsetsis, A.P., Kourounis, C., Spanos, G., et al. "Deep Learning Techniques for Stock Market Prediction in the European Union: A Systematic Review". (2020)

31. Kim H, Cho H, & Ryu D. “Corporate Default Predictions Using Machine Learning: Literature Review”. (2020) https://doi.org/10.3390/su12166325

32. Kristóf T. & Virág.” EU-27 Bank Failure Prediction With C5.0 Decision Trees And Deep Learning Neural Networks.” (2022). Research in International Business and Finance, 61, 101644.DOI: 10.1016/j.ribaf.2022.101644

33. Le H.H. & Viviani J.-L. “Predicting Bank Failure: An Improvement By Implementing A Machine-Learning Approach To Classical Financial Ratios”. (2018). Research in International Business and Finance, 44, 16-25. DOI: 10.1016/j.ribaf.2017.07.104

34. Li, A.W. & Bastos, G.S. “Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review”. (2020) EEE Access, 8, 185232-185242. DOI:10.1109/ACCESS.2020.3030226.

35. Carrascosa et al. “Ecological restoration as a strategy for mitigating and adapting to climate change: lessons and challenges from Brazil” (2019)

36. Lin B. & Bai R. “Machine Learning Approaches for Explaining Determinants of The Debt Financing in Heavy-Polluting Enterprises”. (2019). Finance Research Letters, 44, 102094.DOI: 10.1016/j.frl.2021.102094

37. Marne S., Churi S., Correia D., & Gomes J. “Predicting Price of Cryptocurrency – A Deep Learning Approach.” (2021). International Journal of Engineering Research & Technology, 9(3), 387-393.

38. Mohapatra S., Mukherjee R., Roy A., Sengupta A. & Puniyani A.“Can Ensemble Machine Learning Methods Predict Stock Returns for Indian Banks Using Technical Indicators?”. (2022). Journal of Risk and Financial Management, 15(8), 350. DOI:10.3390/jrfm15080350

39. L.E. Faisal, T. Tayachi, S. Arabia, L.E. Faisal, O. “Banking, The role of internet banking in society.” (2021) 249–257.

40. A. Aditi, A. Dubey, A. Mathur, P. Garg. “ Credit Card Fraud Detection Using Advanced Machine Learning Techniques.” (2022). http://dx.doi.org/10.1109/ccict56684.2022.00022

41. A.D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi. “Credit Card Fraud Detection : A Realistic Modeling and a Novel Learning Strateg.” (2018) 3784–3797

42. R. Tyagi, R. Ranjan, S. Priya. “ Credit card fraud detection using machine learning algorithms.” (2021) 334–341.

43. B.T. Jijo, A.M. Abdulazeez. “Classification Based on Decision Tree Algorithm for Machine Learning.” (2021) 20–28. http://dx.doi.org/10.38094/jastt20165

44. Y. Liu, L. Hu, F. Yan, B. Zhang. " Information Gain with Weight based Decision Tree for the Employment Forecasting of Undergraduates." (2013) 2–5. http://dx.doi.org/10.1109/GreenCom-iThings-CPSCom.2013.417

45. J. Zhao, L. Wang, R. Cabral, F. Torre, De. " Feature and Region Selection for Visual Learning." 25(3) (2016)

46. Iwasokun GB, Omomule TG, Akinyede RO. "Encryption and tokenization-based system for credit card information security. Int J Cyber Sec Digital Forensics." (2018)

47. Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. "Prediction of pipe failures in water supply networks using logistic regression and support vector classification." (2020)

48. Liang J, Qin Z, Xiao S, Ou L, Lin X. " Efficient and secure decision tree classification for cloud-assisted online diagnosis services. IEEE Trans Dependable Secure Comput." (2019)

49. Lingjun H, Levine RA, Fan J, Beemer J, Stronach J. "Random forest as a predictive analytics alternative to regression in institutional research." (2020)

50. Campus K. et al. "Credit card fraud detection using machine learning models and collating machine learning models." (2018)

51. Guo S, Liu Y, Chen R, Sun X, Wang X. X. "Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes." (2019)

52. Hemavathi D, Srimathi H. "Effective feature selection technique in an integrated environment using enhanced principal component analysis." (2021)