VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW
INFORMATION SYSTEMS FACULTY

REPORT SHORT-TERM INTERNSHIP

TOPIC:

PROPOSING A MODEL TO IDENTIFY CREDIT CARD FRAUD IN FINANCE

ENTERPRISE:

Vietnam International Bank (VIB)

Instructor: MSc. Nguyen Van Ho

Class code Section: 232IS2903

Group: DA_4N

Mentor: Nguyen Phat Dat, Data Analyst at VIB

Ho Chi Minh City, June 2024

GROUP MEMBER

No.	Full name	Student ID	Class	Email	Contribution
1	Nguyễn Duy Duẩn	K214061736	K21406	duannd21406@st.uel.edu.vn	25%
2	Trần Sĩ Đan	K214061258	K21406	dants21406@st.uel.edu.vn	25%
3	Nguyễn Mai Trinh	K214060418	K21406	trinhnm21406@st.uel.edu.vn	25%
4	Giả Hoàng Nam Phương	K214061744	K21406	phuongghn21406@st.uel.edu.vn	25%

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to MSc. Nguyen Van Ho and Mr. Nguyen Phat Dat, for their extremely important support and guidance throughout this internship process. Orientation of MSc. Nguyen Van Ho and Mr. Nguyen Phat Dat's expertise in finance have helped us better understand data, operations and problems in the financial sector, thereby providing a solution to the problem. Enthusiasm and interest helped us progress and complete the project smoothly.

We are grateful for the knowledge, advice and materials throughout the study process. We feel very lucky to receive the guidance from you and your teacher.

Finally, we would also like to thank each member of our team for their dedication and efforts in completing the internship project, despite the tight and demanding schedule. We believe that this internship process and internship project has brought many valuable experiences and knowledge to the whole group.

Group: DA_4N

COMMITMENT

We would like to assure that this entire project is the result of group research, carried out under the guidance of Master Nguyen Van Ho and Mr. Nguyen Phat Dat. The results from this project are completely honest.

Group: DA_4N

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

Acronyms	Full writing
IoT	Internet of Things
SQL	Structured Query Language
SAS	Statistical Analysis Software
CDO	Chief Data Officer
GLM	Generalized Linear Model
IF	Isolation Forest
MPL	Mean Path Length
AVL	Average Tree Depth
XGBoost	Extreme Gradient Boosting
LSTM	Long-short Term Memory
RNNs	Recurrent Neural Networks
DNNs	Deep neural networks
AUC-ROC	Area Under the Receiver Operating Characteristic Curve

PART 1.

DATA ANALYTICS

CHAPTER 1. OVERVIEW OF DATA ANALYTICS

1.1 What is data analytics?

Data analytics is a multidisciplinary field that employs a wide range of analysis techniques, including math, statistics, and computer science, to draw insights from data sets. Data analytics is a broad term that includes everything from simply analyzing data to theorizing ways of collecting data and creating the frameworks needed to store it.

1.2 Types of Data Analytics

Data analytics does include a wide range of methods and objectives, all aimed at extracting valuable insights from data to support decision-making processes in various industries. Descriptive analytics summarizes and lays the foundation for historical data, while advanced analytics does tools such as machine learning and deep learning play a role in order to develop predictive models and trend detection.

1.3 How is data analytics used? Data analytics examples

The development of machine learning tools, with the proliferation of big data and improvements in computing power have revolutionized data analytics and companies can now use these technologies to draw meaningful conclusions from complex data sets, ultimately resulting in informed decision making and competitive advantage.

The use of data analytics varies widely. For example, big data analytics can improve performance in a wide range of industries. Enhanced productivity enables businesses to thrive in an increasingly competitive world.

Data analytics plays an important role in banking and financial services, where it is used to forecast market trends and evaluate risk Credit scores are examples of data analytics affecting customers. This score uses multiple data points to quantify credit risk. Data analytics is also used to detect and prevent fraud to increase efficiency and reduce risk for financial institutions.

The use of data analytics goes beyond maximizing profit and ROI. Data analytics can provide important information for health information(s), crime prevention, and

environmental protection. For example, researchers use machine learning to protect wildlife.

The use of data analytics in healthcare is already widespread. Predicting patient outcomes, better allocating funds, and improving diagnostic methods are just a few examples of how data analytics is transforming healthcare. Doctors are also using machine learning. For example, drug discovery is a complex task with many variables that can be facilitated by machine learning. Pharmaceutical companies use data analytics to understand drug markets and forecast sales.

The Internet of Things (IoT) is a concept often used alongside machine learning. Together, these devices offer tremendous opportunities for data analysis. IoT devices have sensors that collect meaningful data points for their operation. Devices like the Nest thermostat control speed and temperature to regulate temperature.

1.4 Advantages and disadvantages of data analytics

1.4.1 Advantages of data analytics

Informed decision making: Data analytics help organizations make informed decisions by transforming available data into valuable insights. This reduces reliance on gut instincts and can provide a competitive advantage by reducing poor decision making that can negatively affect growth and profitability.

Increased performance: Analytics enables faster analysis of large data sets, enabling organizations to better achieve specific goals. Fosters a culture of efficiency and productivity by providing insights to employees, identifying areas for improvement and increasing overall enterprise productivity.

Consumer Behavior Insights: Analyzing large amounts of consumer data helps organizations identify changes in consumer preferences and behavior. Understanding changes in consumer behavior provides a strategic advantage, enabling companies to react more quickly to market changes and align their offerings accordingly.

Personalized products and services: Analytics enables companies to customize products and services based on individual customer preferences. By tracking customer

preferences and behaviors, organizations can tailor recommendations and offers to meet each customer's unique needs, increasing customer satisfaction and loyalty.

Quality Improvement: Data analysis plays an important role in improving the quality of products and services. It helps identify errors, identify useful tasks, and enhance the user experience through auto learning algorithms. Additionally, analytics make it easier to automatically cleanse data, improving the quality of the data, and providing value to both customers and organizations.

1.4.2 Disadvantages of data analytics

Lack of communication across teams: Data analytics insights may not have much impact on organizational design due to lack of alignment between departments or teams Siled business practices prevent effective communication and collaboration, preventing valuable insight sharing to those concerned.

Lack of commitment and patience: Implementing analytics solutions requires time, resources and commitment, the return on investment is not immediate and if immediate results are not seen, possibly labor consumption the role is exciting, leading to a loss of confidence in research processes. For the research project to be successful, it is important to provide feedback that allows problems to be identified and necessary improvements to be made.

Low-quality data: Access to high-quality data is crucial for effective data analysis. Organizations may have access to data, but it may not be sufficient or adequate to answer important business questions. Data quality issues such as incomplete or inaccurate data can lead to poor decision making and undermine effective research efforts.

Privacy Issues: Data collection practices can raise consumer concerns about privacy, especially in relation to the security and confidentiality of their personal information. Organizations must prioritize data privacy and security, ensuring that only sensitive data is collected, and anonymizing sensitive information to protect customer privacy and maintain trust.

Complexity and bias: Some assessment tools act as black box models, making it difficult to understand the reasoning behind decision-making processes. This lack of transparency can lead to hidden biases in the data and decisions these programs make, and can lead to lawsuits related to race-based discrimination.

CHAPTER 2. CAREER PROSPECT

2.1 Career opportunities

In 2023, according to the latest estimates, 328.77 million terabytes of data are created each day. In fact, it is estimated that 90% of the world's data was generated in the last two years alone. And By 2025, this figure is expected to skyrocket to a staggering 181 zettabytes.

In light of the burgeoning volume of data in contemporary times, the capacity to systematically analyze and derive meaningful insights from such datasets has become imperative. As indicated in a scholarly report authored by Research and Market.

'The global big data and business analytics market in 2022 was valued at US$294.16 billion. The market value is anticipated to grow to US$662.63 billion by 2028'

In the realm of data analysis, diverse vocational pathways await, spanning various industries, including:

- Finance and Banking: Within financial institutions, data analysts are pivotal for tasks such as risk assessment, fraud detection, and investment analysis, contributing to informed decision-making and financial stability.
- Healthcare: The analysis of patient data not only enhances healthcare outcomes but also facilitates resource allocation optimization, supports medical research endeavors, and aids in the development of pharmaceuticals.
- E-commerce and Retail: Data analysis serves as a cornerstone for understanding customer behavior, refining pricing strategies, and efficiently managing inventory, thereby bolstering competitiveness and profitability in the digital marketplace.
- Marketing and Advertising: Data-driven insights fuel targeted advertising campaigns, enable rigorous evaluation of campaign effectiveness, and empower marketers to discern and capitalize on emerging market trends.

- Technology and Software Development: Data analysts play a crucial role in product development by scrutinizing user feedback, enhancing user experience, and fine-tuning software performance to meet evolving consumer demands.
- Government and Public Policy: Data analysis is indispensable for policymakers seeking to make evidence-based decisions, comprehend societal trends, and allocate resources judiciously for the betterment of communities and nations.
- Education: Educational institutions leverage data analysis to monitor student performance, assess the efficacy of teaching methodologies, and implement strategies aimed at optimizing educational outcomes and fostering student success.
- Manufacturing and Supply Chain Management: Through data analysis, manufacturing entities streamline production processes, forecast maintenance requirements, and orchestrate efficient supply chain operations, thereby enhancing productivity and reducing costs.
- Telecommunications: Analysis of customer usage patterns and network performance data enables telecommunications companies to enhance service quality, optimize network infrastructure, and stay ahead in a rapidly evolving industry.
- Energy and Utilities: Data analysis is instrumental in optimizing energy production, forecasting demand fluctuations, and enhancing the efficiency of utility services, thus contributing to sustainability and reliability in the energy sector.

2.2 Competition in career

According to LinkedIn, as of August 2022, there are over 900,000 Data Analyst and Data Scientist jobs open in the United States alone. The supply of data experts is still catching up with demand, which means that competition for highly qualified candidates can be fierce. That's not to mention the greater demand for data skills throughout the broader labor economy, which the U.S. Bureau of Labor Statistics estimates will grow by 28% by 2026.

Meta, Apple, Amazon, Netflix, and Alphabet (formerly Google) are some of the world's leading (and largest by market capitalization) technology companies. As digital natives, their data maturity makes them an attractive employer for data scientists. Furthermore, their sponsorship gives them the ability to offer more roles at attractive salaries, attracting top data professionals. Despite this competition and the number of redundant roles available to data professionals in general, there have been tech layoffs and hiring freezes since the start of 2022. As a result, 9% of tech workers are feeling secure in their jobs right now. Such widespread pessimism contrasts with a still-hot labor market, economists say. Layoffs in June were still just under 1% of the workforce, with laid-off candidates being 'snapped up for weeks' during the hiring and quitting period near record highs.

Several key factors contribute to the high competition in the field of data analytics. First, the growing demand for skilled data analysts in businesses of all sizes has led to an increase in competition for job opportunities and opportunities. In addition, rapid advances in technology, especially in big data, machine learning, and artificial intelligence, require continuous upskilling to remain competitive. Furthermore, the rise of educational programs and courses that meet the demand for data analysis skills has led to an influx of individuals entering the field, further intensifying competition. Moreover, globalization has expanded the talent pool, allowing professionals to work remotely or travel for better opportunities, thus increasing competition for both local and international positions. As the field matures, experts increasingly specialize in niche areas within data analytics, which enhances expertise but also enhances competition. Employers prefer candidates with practical experience and a proven track record, emphasizing the importance of building a strong portfolio and gaining relevant work experience to gain a competitive advantage. Finally, networking and personal branding are essential for career advancement, allowing professionals to stand out by building connections, participating in relevant communities, and demonstrating expertise through online platforms.

2.3 Vietnam market research

Vietnam's data analytics market is growing, reflecting global trends in the field. The demand for skilled data analysts in Vietnam is on the rise, driven by the growing recognition of data-driven decision making in various industries.

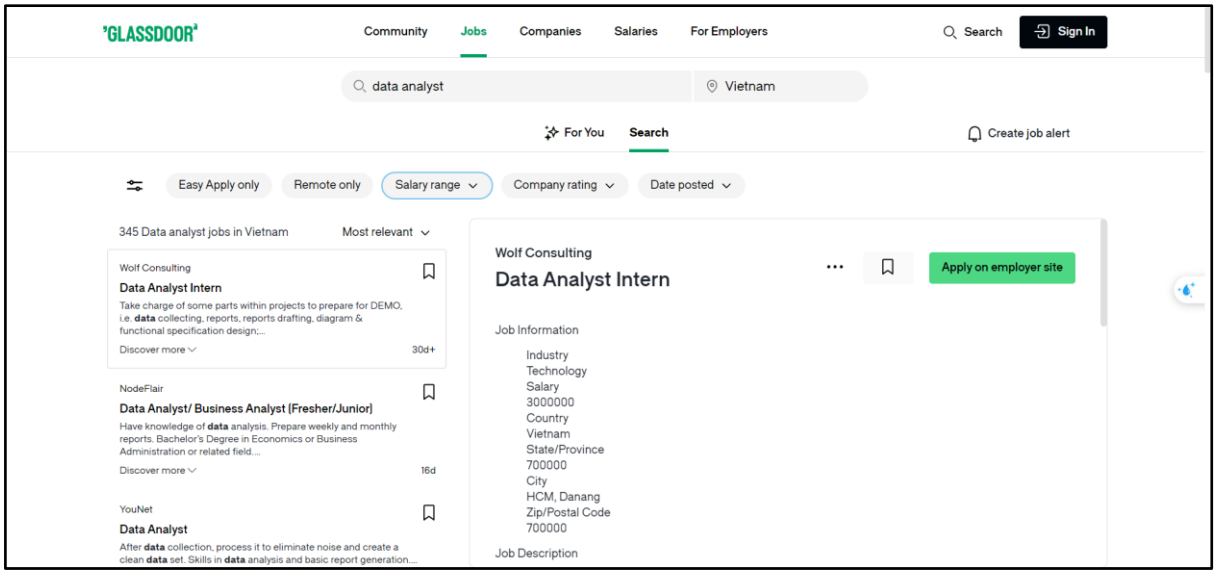

Figure 2.1. Data Analyst job example

When looking for jobs in data analysis, many positions related to the proposed are necessary and the popularity of data analytics is currently huge, the reason for this is that most businesses are oriented towards transformation and decision making based on data and AI technology application. Data analytics in Vietnam is still being assessed as a potential and beneficial market, with more opportunities for development due to the shortage of high-quality human resources, capable of adapting to the rapid change of technology.

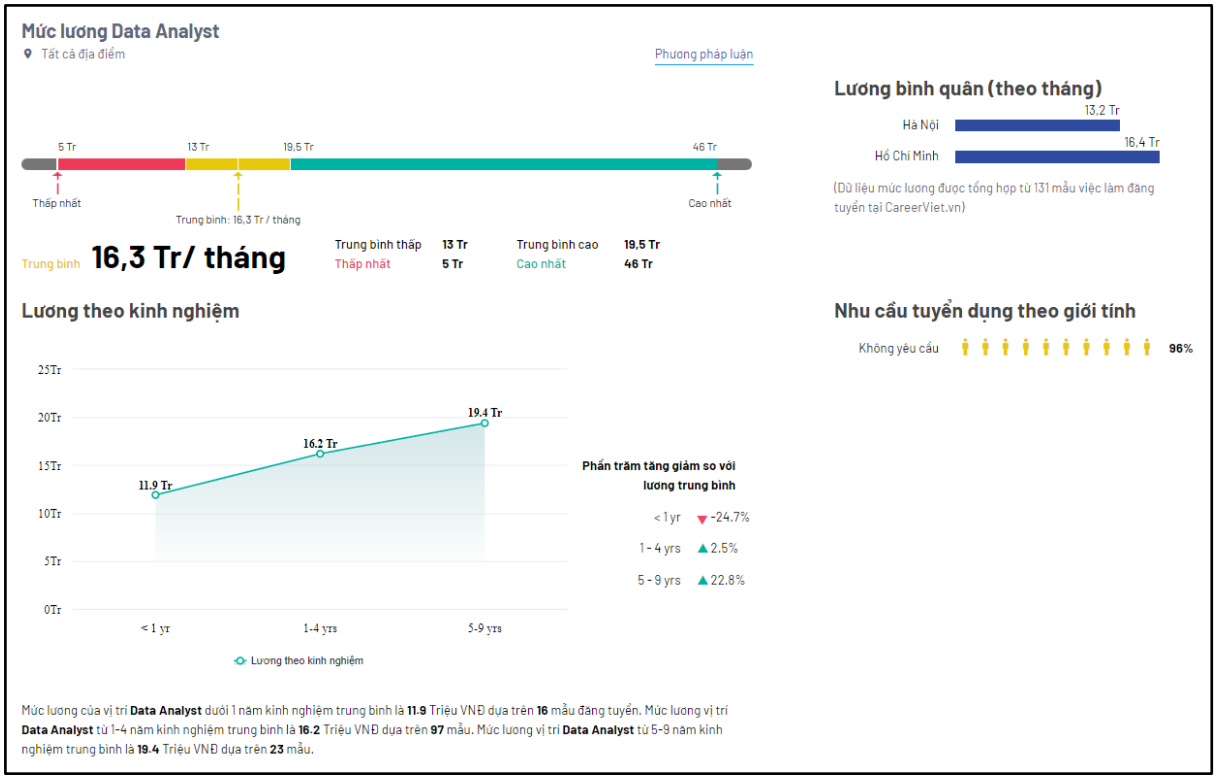

Figure 2.2. Data Analyst salary 2023 according to CareerViet

Research by VietnamSalary shows that the average salary of data analysts in Vietnam is 16.3 million VND per month. However, this figure is just an average and the actual salary of each individual in the industry can vary quite widely, ranging from 5 to 46 million VND per month. This is to say that there is a large divergence in income in the industry, and there are many factors that affect the specific salary of each person. These factors include:

- Geographical Location: Salaries may be higher in big cities like Hanoi or Ho Chi Minh due to high demand and higher cost of living. Areas that are less developed, or have lower hiring demand may have lower salaries.
- Company: Large, international companies or those with strong investment capital tend to be able to pay higher salaries. Conversely, small companies, startups or local companies may have lower salaries.
- Field of activity: Different professions have different needs and finances for data analyst positions. Some sectors such as finance, insurance, and technology may

pay higher salaries due to the need for complex data analytics and the value brought from data is large.

- Work Experience: The most important factor: More experience increases the likelihood of a high salary. Entry-level or junior data analysts tend to have lower starting salaries. People with long experience, especially those with rare and in-depth skills, can negotiate high salaries.

CHAPTER 3. CAREER PATH

3.1 Career path overview

A data analyst collects, cleans, and interprets data sets in order to answer a question or solve a problem. These professionals are employed across various sectors such as business, finance, criminal justice, science, medicine, and government. As experience is gained as a data analyst, opportunities to advance one's career may be encountered in several different directions. Depending on goals and interests, progression into data science, management, consulting, or a more specialized data role may be pursued.

3.2 Entry level

Once the necessary skills for the data analyst role have been acquired, an entry-level position as a data analyst can be begun. As a newly qualified analyst, one would typically commence in a practical role, often as a junior analyst or simply a data analyst. Primary duties would involve data extraction, ensuring data cleanliness, conducting analyses, and communicating discoveries.

3.3 Experienced level

3.3.1 Data scientist

Many data analysts are initially guided by data scientists, with this transition typically involving the enhancement of programming skills, acquisition of more advanced mathematical knowledge, and cultivation of an understanding of machine learning. Furthermore, many data scientists are often equipped with degrees in data science, computer science, or related fields, which, though not strictly necessary, can lead to increased job opportunities. On a day-to-day basis, tasks performed by a data scientist might include identifying patterns and trends in datasets to reveal insights, developing algorithms and data models to predict outcomes, applying machine learning techniques to enhance data quality or product offerings, communicating recommendations to other teams and senior staff members, utilizing data analysis tools such as Python, R, SAS, or SQL, and keeping abreast of advancements in the field of data science.

3.3.2 Data Management

Moving into management positions is another common career trajectory for data analysts. Starting out as a data analyst, one may progress to become a senior-level analyst, analytics manager, director of analytics, or even a chief data officer (CDO). Emphasis should be placed on the development of leadership skills in conjunction with data skills. In certain companies, attaining these higher-level positions may necessitate a master's degree in data analytics or business administration with a focus on data analytics.

3.3.3 Data Specialist

Data analysts can work in various industries, and their career paths may lead to specialized knowledge within those industries. This specialization requires a deeper understanding of that domain. Some commonly mentioned specialist title include:

- Business analysts utilize data to enhance the efficiency and effectiveness of an organization's IT processes, organizational structures, or staff development.
- Financial analysts employ data to steer investment opportunities, identify revenue potentials, and mitigate financial risk.
- Operations analysts are responsible for optimizing a company's performance by identifying and resolving technical, structural, and procedural issues.
- Marketing analysts, also known as market research analysts, examine market trends to aid in determining product and service offerings, price points, and target customers.
- Systems analysts utilize cost-benefit analysis to align technological solutions with company needs.
- Health care analysts leverage data from health records, cost reports, and patient surveys to assist providers in improving the quality of care.

3.3.4 Consultant

After several years of experience analyzing data for a company or multiple companies, the option to work as a data analytics consultant can be considered. Rather

than being directly employed by a company, one would work as a freelance contractor or for a consulting firm, conducting analysis for diverse clients. Working as a consultant typically entails a greater variety in the types of analysis performed and offers increased flexibility, especially for those who are self-employed.

CHAPTER 4. REQUIRED SKILLS FOR DATA ANALYST

4.1 Required skillset

According to Cousera (2024), in a world where there are more demands each year for data analysts and scientists than there are people with the appropriate skills to fill those roles. In fact, according to the US Bureau of Labor Statistics (BLS), the number of job openings for data analysts is expected to increase by 23 percent from 2022 to 2032, significantly higher than the average job growth rate nationwide. Whether it's in business, healthcare, finance, or any other industry, data analysts play a crucial role in transforming raw data into meaningful information to support decision-making. Below is the required skillset and responsibilities of a data analyst:

- SQL: Structured Query Language, or SQL, is the standard language used to communicate with databases. Knowing SQL allows you to update, sort, and query data stored in relational databases, as well as modify data structures (schema).
- Statistical programming: Statistical programming languages like R or Python enable you to perform advanced analyses that Excel cannot handle. Being able to program in these languages helps you clean, analyze, and visualize large datasets more effectively.
- Machine learning: Machine learning, a branch of artificial intelligence (AI), has become one of the most important developments in data science. This skill focuses on building algorithms designed to find patterns in large datasets and improve accuracy over time.
- Probability and statistics: Statistics is the field of mathematics and science concerned with collecting, analyzing, interpreting, and presenting data. With a solid foundation in probability and statistics, you can recognize patterns and trends in data, produce accurate results, and make reliable conclusions.
- Data management: Data management involves methods for collecting, organizing, and storing data efficiently, securely, and cost-effectively.

- Statistical visualization: Data visualization helps tell a story with data to support better business decisions. By using charts, graphs, and maps, you can present findings in a clear and compelling way.
- Econometrics: With econometrics, analysts apply statistical and mathematical models to economic data to forecast future trends based on historical data. Understanding econometrics is essential for data analysts seeking jobs in finance, particularly at investment banks and hedge funds.

4.2 Responsibilities

Responsibilities for a data analyst encompass collecting, analyzing, and communicating data to address specific challenges and inform decision-making processes. This involves various tasks:

- Understanding Business Needs: Identifying scientific objectives and comprehending business requirements to clarify the analysis plan and ensure alignment with organizational goals.
- Data Collection: Acquiring relevant data through methods like surveys, website tracking, or purchasing datasets, ensuring it meets the needs outlined in the previous phase.
- Data Cleaning: Ensuring data integrity by removing duplicates, errors, or outliers to maintain accuracy and reliability.
- Data Modeling: Designing database structures and determining data categories and relationships, laying the foundation for effective analysis.
- Data Analysis: Applying statistical techniques to identify patterns, trends, and insights within the data, which are crucial for answering the research questions at hand.
- Presentation of Findings: Communicating research outcomes effectively through visualizations such as charts and graphs, generating reports, and presenting data to stakeholders in a clear and understandable manner.

Throughout these responsibilities, data analysts leverage their skills in mathematics, statistics, economics, and computer science to extract meaningful insights from data. They also utilize their understanding of database architecture and query language to access and manipulate data efficiently. Additionally, they may work with both structured and unstructured data, such as text or audio files, to address diverse analytical challenges.

CHAPTER 5. PLAN AND DISCUSSION

5.1 Timeline

In order to carry out the project effectively, several milestones have been set. Achieving these milestones helps ensure that the project is completed on time, meeting its objectives.

(1) Exploring Career Options

Different career paths will be thoroughly explored to gain a better understanding of them. This includes researching typical tasks for each job and assessing their future demand.

(2) Problem Identification

A business case is issued by a company. It's required to address this problem within a month.

(3) Project Execution

Once the problem and its requirements are understood, the project will officially begin, guided by the lecturer. We'll focus on finding the best approach to solve the issue.

(4) Solution Development

Solutions to the problem are developed in this phase. This involves brainstorming ideas, conducting research, and testing different approaches until the most suitable one is found.

(5) Presentation

The project findings will be presented to others. This helps deliver what has been achieved and why it's significant.

(6) Reflection and Learning

Finally, this final phase involves reflecting on the project and the lessons learned. This reflection aids in improvement for future improvement and projects.

5.2 Discussion

In this discussion section, we will focus on key aspects to ensure the success of the project. Below are the main topics we will address:

- Importance of Data: We will evaluate the significance of accurate data collection, processing, and analysis for the project. A deep understanding of data will play a crucial role in making informed decisions and strategic choices.
- Challenges and Risks: We will examine potential challenges and risks that may arise during the project implementation. This could include issues related to inaccurate data, technology dependencies, or project management concerns.
- Communication and Collaboration: We will discuss the importance of effective communication and collaboration in the project. Close understanding and collaboration among team members will be essential in achieving project goals.
- Evaluation and Adjustment: We will assess the project's progress and outcomes, and explore ways to adjust the plan to ensure success in the future.

However, despite our thorough preparation, we have encountered obstacles in connecting with these companies and gaining their approval for collaboration. Moving forward, it is imperative that we shift our focus towards enhancing our approach to engage with companies and showcase the value we can bring through our expertise in data analytics.

PART 2.

BUSINESS CASE SOLVING

CHAPTER 1. INTRODUCTION

1.1 Business Case

1.1.1 Overview of the Business

Vietnam International Commercial Joint Stock Bank (VIB) is one of the leading commercial joint stock banks in Vietnam, established on September 18, 1996. After nearly 28 years of development, VIB has become a top bank in retail and profitability with total assets exceeding VND 410 trillion. VIB currently has over 12,000 employees at 189 branches and transaction offices nationwide. The bank stands out in the field of bancassurance and has a strategic partnership with Prudential Vietnam, as well as signing loan agreements with the IFC. VIB is also committed to community activities, focusing on education, environment, and community development, and has received numerous prestigious awards such as the "Enterprise with Excellent Social Security and Community Responsibility" award.

1.1.2 Business Case

In the contemporary digital age, financial fraud has evolved into a sophisticated and multifaceted issue affecting individuals, businesses, and economies at large. The ramifications of fraudulent activities are far-reaching, resulting in financial losses, damage to organizational reputations, and erosion of consumer trust (Albrecht et al., 2016). Consequently, the imperative for robust fraud detection systems has never been more pronounced, driving the quest for innovative and effective methodologies to identify and mitigate fraud.

Fraud detection encompasses a wide array of techniques and processes aimed at identifying irregularities, anomalies, or patterns indicative of fraudulent behavior. Traditional methods involving manual detection are not only time consuming, expensive and inaccurate, but in the age of big data they are also impractical (West et al., 2016); besides, the advent of technologies and the proliferation of data have catalyzed the adoption of more advanced, data-driven approaches. Machine learning algorithms, owing to their ability to learn from data and improve over time, have emerged as a key part of fraud detection. Techniques such as Logistic Regression,

Random Forest, XGBoost, and LightGBM, have demonstrated substantial promise in distinguishing between legitimate and fraudulent transactions or behaviors by analyzing vast datasets (Sivanantham et al., 2021).

Despite these technological advancements, previous research in the domain of fraud detection has been hampered by several gaps. Firstly, many studies have been limited by the scope of data, often relying on outdated or narrow datasets that do not fully encapsulate the dynamic and evolving nature of fraudulent activities (Yang et al., 2021). Moreover, there has been a significant emphasis on traditional algorithms with less consideration given to newer, potentially more efficacious machine learning models. Furthermore, there has been a noticeable deficiency in comprehensive comparative studies that evaluate different algorithms side by side using modern and extensive datasets. This shortfall has led to a segmented and incomplete perspective on the efficacy of various fraud detection methods.

In light of these deficiencies, our study seeks to provide a comprehensive solution by conducting a comparative analysis of several prominent machine learning algorithms: Logistic Regression, XGBoost, Random Forest, and LightGBM. Utilizing an updated and expansive dataset, our research aims to not only bridge the gaps of previous studies but also to offer insights into the relative performance and suitability of these algorithms for fraud detection in a modern-day context. Through this endeavor, we aspire to contribute to the development of more accurate, efficient, and scalable fraud detection systems, thereby fortifying the defenses against fraud in an ever-digitizing world.

1.2 Objectives

This research aims to propose techniques and build models for identifying credit card fraud within the e-commerce domain using a range of machine learning methods. The study will focus on optimizing and assessing models employing Logistic Regression, Random Forest, XGBoost, LightGBM, and TabNet to discern the authenticity of transactions. The specific goals encompass:

- Presenting Fraud Theory: Offering a theoretical background on fraud and credit card fraud, pinpointing key factors contributing to the rise in financial fraud.
- Exploring Fraud Detection Techniques: Delving into the theoretical underpinnings of categorizing transactions as fraudulent or genuine, and examining methods and algorithms for constructing models based on historical data.
- Optimizing Fraud Detection Models: Formulating fraud detection models utilizing the provided dataset and employing diverse methods, blending algorithms to improve accuracy.
- Assessing and Comparing Models: Experimenting with different techniques, assessing, comparing, and selecting the most effective model, and refining it further to boost performance and accuracy.

1.3 Scope

Theoretical Scope:

- Explore the nature of fraud, various detection methods, and data analysis algorithms to identify fraud in transactions.
- Review existing techniques for detecting fraud and predicting future fraudulent activities.

Practical Scope:

- Analytical Methods: Utilize advanced analytical methods to detect anomalies within the dataset.
- Model Implementation: Implement and run classification and prediction models to determine the most effective method based on results and performance.
- Performance Optimization: Enhance the research by evaluating and optimizing model performance to ensure the best outcomes and efficiency.

1.4 Values and expected outcomes

1.4.1 Values

This research aims to contribute to the academic understanding of fraud detection by examining and assessing advanced machine learning methods.

- Advancement of Knowledge: Contributing to the understanding of how different machine learning models can be effectively applied to credit card fraud detection.
- Algorithmic Innovation: Investigating the potential of combining traditional models with newer architectures like deep learning model to enhance detection capabilities.
- Framework Development: Providing a framework for future studies to build upon, facilitating further advancements in the field.

Overall, the culmination of these efforts is the bolstering of trust in financial systems by mitigating the risks and impacts of fraud. The societal benefits are manifold, ranging from the protecting of individuals' assets to safeguarding the integrity of financial infrastructures, ultimately contributing to a more secure and stable economic landscape. Similarly, by reducing the costs associated with fraud, resources can be redirected towards growth and development initiatives, which can have positive ripple effects across various sectors of the economy.

1.4.2 Expected outcomes

The expected outcomes of this research include:

- Identification of the most accurate and efficient fraud detection model from among Logistic Regression, Random Forest, XGBoost, CatBoost, LSTM and DNNs.
- Enhanced ability to detect and prevent fraudulent transactions, reducing financial losses for cardholders and issuers.

- Comprehensive insights into the strengths and weaknesses of various machine learning techniques for fraud detection, contributing to the academic and professional discourse on financial fraud prevention.
- Establishing a foundation for future research and development in the field of financial fraud detection, paving the way for continuous improvement and innovation.

1.5 Structure of the project

The structure of this research project is outlined as follows:

Chapter 1: Introduction

This chapter provides an overview of the research, including the business case, objectives, scope, values, and expected outcomes.

Chapter 2: Theoretical Background

This chapter discusses the fundamental theories and methodologies related to credit card fraud detection. It covers various machine learning techniques such as Logistic Regression, Isolation Forest, Random Forest, XGBoost, CatBoost, and LSTM, DNNs, and reviews related works in the field.

Chapter 3: Proposed Model

This chapter describes the proposed methodology, including exploratory data analysis (EDA), data description, list of variables, and analysis of variables. It also details the data preprocessing steps necessary for model development.

Chapter 4: Experimental Result

This chapter presents the experimental setup, evaluation parameters, and performance results of each model (Logistic Regression, Random Forest, XGBoost, CatBoost, LSTM and DNNs). It concludes with an overall evaluation and comparison of the models.

Chapter 5: Discussion and Recommendation

This chapter provides a discussion of the experimental results, offering insights into the findings and presenting recommendations based on the analysis.

Chapter 6: Conclusion and Future Works

This chapter summarizes the research findings, highlights the contributions of the study, and suggests directions for future research. It also addresses the limitations of the current work and proposes areas for further investigation.

CHAPTER 2. THEORETICAL BACKGROUND

2.1 Credit Card Fraud

Credit card fraud refers to unauthorized or deceptive transactions made using someone else's credit card or credit card information (Bhatla et al, 2003). It involves activities such as using stolen credit card details, cloning credit cards, or making fraudulent transactions without the cardholder's knowledge or consent. Credit card fraud can have significant consequences for all parties involved, including the cardholder, merchants, banks, and financial institutions. According to (Chaudhary et al., 2012), credit card fraud is categorized into two main types: offline fraud and online fraud.

- Offline fraud occurs when a stolen physical card is used at a call center or any other physical location. Common methods include card theft, skimming, and the use of counterfeit cards. Card theft involves the physical theft of a credit card for unauthorized purchases or cash withdrawals (Bhatla et al., 2003). Skimming involves capturing card information at ATMs or POS systems to create counterfeit cards (Chaudhary et al., 2012). Counterfeit cards are made using data obtained through skimming or similar means (Sadgali et al., 2018).
- Online fraud is committed via the internet, phone, online shopping, web, or in situations where the cardholder is not present. Common methods include phishing, carding, and account takeover. Phishing uses deceptive emails or websites to trick individuals into providing their card information (Joulin et al., 2010). Carding tests stolen card details on e-commerce sites (Sadgali et al., 2018). Account takeover involves unauthorized access to a victim's online banking or e-commerce account (Delamaire et al., 2009).

Various detection techniques and prevention measures have been developed to combat credit card fraud (Sadgali et al., 2018). These techniques aim to identify suspicious transactions, patterns, or anomalies that may indicate fraudulent activity. The detection and prevention of credit card fraud are crucial for maintaining the integrity and security of electronic payment systems (Delamaire et al., 2009).

2.2 Machine Learning

2.2.1 Logistic Regression

Logistic regression, as a type of generalized linear model (GLM), has been extensively studied and applied across various fields due to its flexibility in handling outcomes not measured on a continuous scale (Cox, 1958). Unlike traditional linear regression, logistic regression does not assume a linear relationship between the independent and dependent variables, making it suitable for modeling the probability of categorical outcomes (Hosmer & Lemeshow, 2000). The logistic regression model is highly valued for its interpretability and ease of implementation, making it a staple in fields such as medicine (Hosmer & Lemeshow, 2000), social sciences (Siddiqi, 2017), and finance (Agresti, 2002).

Logistic regression is characterized by its use of the logistic function (also known as the sigmoid function) to model the probability of a binary outcome. This function maps the linear combination of predictors to a range between 0 and 1, ensuring valid probability predictions. The key formula for logistic regression is:

$$p(X)= \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$$

Where:

- $p(X)$ denotes the probability of the outcome being 1.
- e is the base of the natural logarithm.
- β_0 the intercept of the model.
- $\beta_0, \beta_{1,}, \ldots \beta_k$ are the coefficients of the predictor variables X_1, X_2,...X_k

2.2.2 Isolation Forest

Isolation Forest (IF) is an unsupervised exception detection algorithm. This algorithm attracts the attention of the scientific research community because of its high efficiency in identifying unusual data points in large data sets, especially in fields such as finance, network security, medical, etc.

The theoretical basis of IF is based on the assumption that outliers tend to be isolated in the data space. This means they are often located further away from other normal data points and have fewer neighbors. IF uses two main metrics to measure a point's isolation:

- Mean Path Length (MPL): This is the average length of the path from the root of the tree to the data point. A point with a high MPL means it is farther from the tree and therefore more likely to be an outlier.
- Average Tree Depth (AVL): This is the average depth of the tree when data points are added. A point with a high AVL means it was added to the tree at a deeper level and is therefore more likely to be an outlier.

2.2.3 Random Forest

Random Forest is an ensemble learning method widely used for both classification and regression tasks. Introduced by Leo (2001), the technique involves creating a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

A Random Forest consists of a large number of decision trees $\{h(x, \theta_k)\}_{k=1}^{K}$, where $h(x, \theta_k)$ is a single decision tree, x is the input vector, and θ_k represents the parameters of the k_{-th} tree. Each tree is trained on a bootstrap sample of the training data. For each tree, at each split in the tree, a random subset of the features mmm is selected from the total number of features MMM. The best split is then chosen from this subset. This randomness helps to ensure that the trees are de-correlated. For classification, each tree $h_k(x)$ casts a vote for the predicted class. The final prediction$H(x)$ is the class that receives the majority of votes:

$$H(x) = arg\ max \sum_{k=1}^{K} I(\ h_k(x) = y)$$

where:

$I(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise.

The accuracy of a Random Forest depends on how strong each tree is and how they are related to each other. The strength of each tree is determined by looking at how well it predicts the correct outcome compared to the incorrect ones. The generalization error is essentially a measure of how well the forest predicts outcomes it hasn't seen before. As more trees are added to the Random Forest, the predictions of the whole forest get closer to what we expect on average. This helps to make the predictions more stable and reliable. Random Forests can figure out how well they're doing by using data they haven't seen before. They do this by leaving some data out when they're training each tree and then using that left-out data to see how well the forest is doing overall.

2.2.4 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and widely-used machine learning algorithm known for its efficiency, accuracy, and scalability. Introduced by Chen and Guestrin (2016), XGBoost has become a go-to method for many data science competitions and real-world applications due to its performance and flexibility.

Grounded in the gradient boosting framework originally proposed by Jerome Friedman in 1999 (Friedman, 2001), XGBoost builds upon traditional gradient boosting through several key innovations. These include regularization techniques to mitigate overfitting, parallel processing to expedite training, tree pruning to manage model complexity, and strategies for handling missing values to enhance robustness.

In XGBoost, the final prediction is made by combining the predictions from all the individual trees. Unlike Random Forest, where each tree contributes equally to the final prediction, in XGBoost, the contribution of each tree is weighted based on its performance and the residual errors it aims to correct.

One notable aspect of XGBoost is its optimization objective, which consists of a loss function that measures the difference between the true labels and the predicted values, and a regularization term that penalizes complex models to prevent overfitting. The optimization objective is minimized during the training process to find the best parameters for the individual trees.

The prediction made by XGBoost can be represented as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \gamma \cdot Tree(x_i|\Theta^{(t)})$$

where:

- $\hat{y}_i^{(t)}$ is the predicted value for the iii-th instance at iteration t.
- $\hat{y}_i^{(t-1)}$ is the predicted value for the iii-th instance from the previous iteration.
- γ gammaγ is the learning rate, which controls the step size of each iteration.
- $Tree(x_i|\Theta^{(t)})$ represents the prediction of the decision tree model at iteration $t$for the i-th instance with parameters $\Theta^{(t)}$.

Similar to Random Forest, XGBoost also benefits from the addition of more trees, which helps improve the stability and reliability of the predictions. Additionally, XGBoost employs techniques like feature subsampling and column subsampling to introduce randomness and reduce overfitting, similar to Random Forest. These features contribute to the overall robustness and generalization performance of XGBoost.

2.2.5 CatBoost

CatBoost is a decision tree gradient boosting algorithm specifically optimized for data containing many categorical features. Boosting is a technique that combines multiple weak learners, often shallow decision trees, to create a stronger model. One of the key algorithmic advances in CatBoost is the implementation of ordered boosting, which enables the algorithm to handle categorical features effectively (Prokhorenkova et al., 2018). Traditional gradient boosting algorithms treat categorical features as numerical, which may lead to suboptimal performance. However, CatBoost incorporates an ordered boosting algorithm that considers the natural ordering of categorical features, thereby capturing the inherent information in these features more accurately.

CatBoost works by building many shallow decision trees, each trying to correct the errors of the previous tree. The special feature of CatBoost compared to other boosting algorithms is its ability to process categorical features without first converting them to numeric values. This is done through target-based encoding and random permutations to minimize bias and overfitting.

CatBoost also uses "ordered boosting" techniques, where the order of the data is carefully handled to minimize errors during training. Instead of using the entire dataset to estimate the value of a categorical feature, it divides the dataset into several parts and uses different parts for estimation, which helps minimize bias. This process can be described mathematically as follows:

$$\hat{y}_i = \sum_{t=1}^{T} \gamma \cdot Tree(x_i|\Theta^{(t)})$$

where:

- T is the total number of iterations (trees)
- The summation aggregates the contributions of all trees to the final prediction

Additionally, CatBoost employs a technique called "symmetric trees," where the structure of trees is constrained to be symmetric. This means that each tree splits nodes in the same way, which improves both the training speed and the consistency of the predictions.

2.3 Deep Learning

2.3.1 Deep Neural Networks (DNNs)

Deep neural networks (DNNs) are a powerful type of artificial neural network architecture that have revolutionized various fields of machine learning and artificial intelligence. Unlike traditional neural networks with few layers, DNNs contain multiple hidden layers between the input and output layers. This increased complexity allows them to learn intricate patterns and relationships within data, making them suitable for a broader range of tasks. The structure of DNNs usually includes:

- Layers: DNNs consist of interconnected layers of artificial neurons, inspired by the structure of the human brain. Each layer transforms the data it receives from the previous layer using a mathematical function.
- Neurons: Neurons within a layer are not directly connected to each other. They only receive input from the previous layer, process it, and send the output to the next layer.

- Hidden Layers: The hidden layers, stacked between the input and output layers, are responsible for extracting complex features from the data. The more hidden layers a DNN has, the "deeper" it becomes.

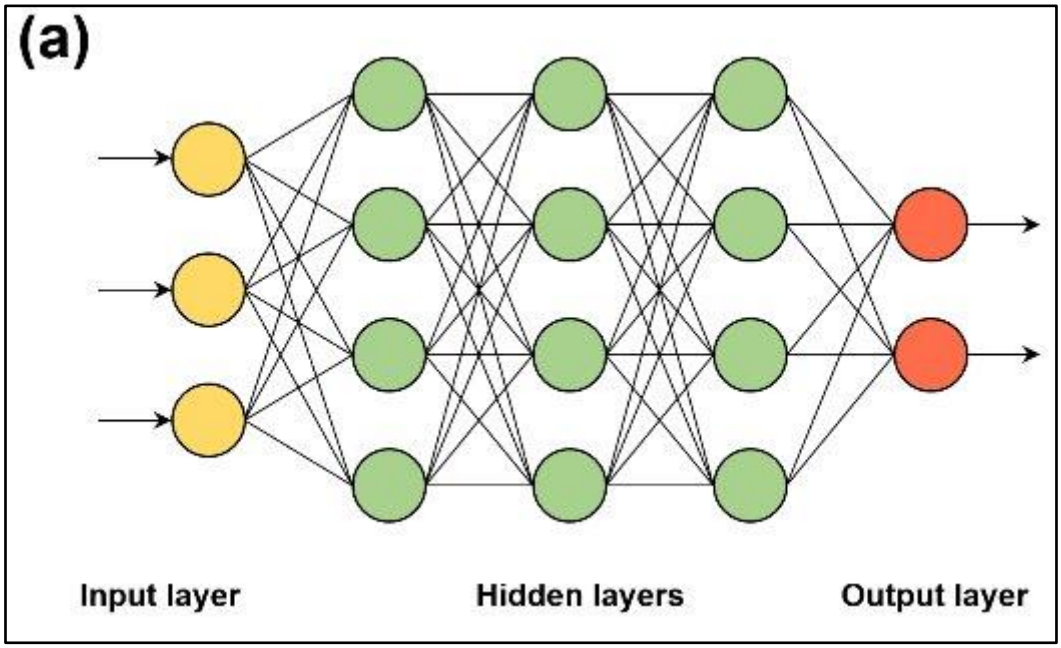

Figure 2.1. Typical structure of DNN model

2.3.2 LSTM

In the fields of Machine Learning and Deep Learning, Recurrent Neural Networks (RNNs) have proven effective in handling sequential data such as text, audio, and time series. However, a limitation of traditional RNNs is their ability to handle long and complex sequences due to the vanishing gradient problem when training models on long sequences.

To address this issue, the Long Short-Term Memory (LSTM) algorithm was introduced in 1997 by Hochreiter and Schmidhuber. LSTM is a variant of the recurrent neural network specifically designed to process and remember information from distant parts of long sequences without being affected by the vanishing gradient problem.

The LSTM algorithm is a type of recurrent neural network (RNN) designed to address the issue of vanishing or exploding gradients when training traditional RNN models. LSTM uses special gates to control the flow of information and decide which information should be retained and which should be discarded.

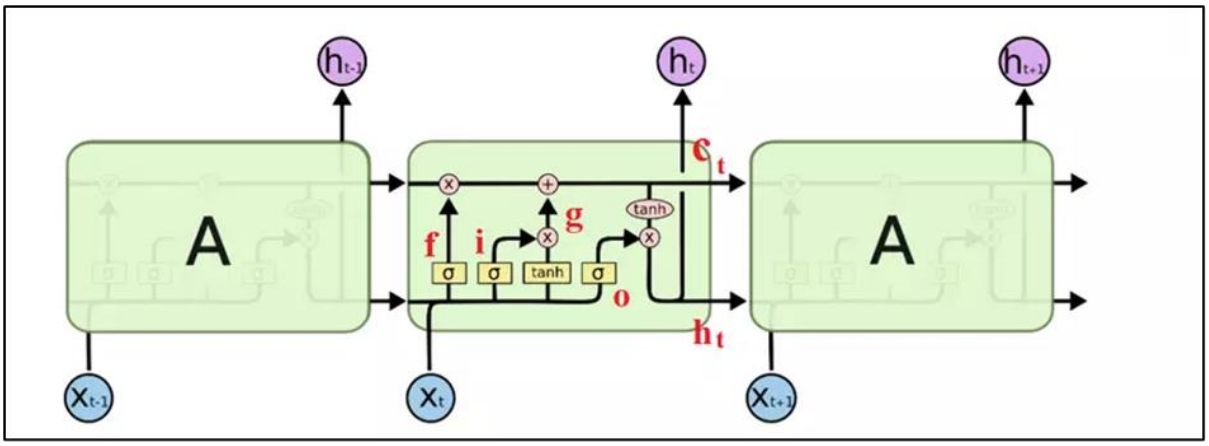

Figure 2.2. LSTM Model (source: nttuan8.com)

The LSTM operates through a series of steps to effectively manage and retain information over long sequences. The key steps involved in the LSTM's operation include:

- Forgetting and Storing Information: The forget gate in LSTM determines which parts of the old information in the cell should be retained and which parts should be discarded.
- Selecting New Information: The input gate in LSTM decides which new information should be added to the cell.
- Updating Information: The cell integrates new information and generates a predicted output based on the cell's content and the controlling gates.
- Repeating for Each Time Step: This process is repeated for each time step in the sequence, enabling the LSTM to learn and retain distant information.
- Storing and Transmitting Information in the Cell: Each cell in an LSTM maintains its state over time steps, preserving important information from previous steps and passing it on to subsequent steps.
- Handling Long Sequences and Distant Connections: LSTM's capability to store distant information in the cell allows it to manage long sequences without being affected by the vanishing gradient problem. This makes LSTM a powerful tool for modeling complex relationships in sequential data.

Training and Weight Updates: During training, the LSTM is optimized by adjusting the weights of the gates and memory units to enhance the network's

performance. Techniques such as backpropagation are used to modify these weights based on the discrepancy between predictions and actual outcomes.

Due to its ability to handle distant information and retain it for extended periods, LSTM is widely used in various applications. In natural language processing, LSTM is used for tasks like machine translation and automatic text generation. In computer vision, LSTM is utilized for object recognition and image classification. Additionally, LSTM is applied in time series prediction tasks, including weather forecasting and financial data management.

In this project, we will explore the workings of the LSTM algorithm in detail and apply it to detect fraud in credit cards.

2.4 Related Works

Previous works on fraud detection mostly provide comparisons between state-of-the-art methods, including both machine learning and deep learning. For example, Awoyemi et al. (2017) presents a comparative analysis of credit card fraud detection using naive Bayes, k-nearest neighbor, and logistic regression techniques. The study focuses on highly skewed data and explores the effectiveness of different machine learning algorithms in detecting fraud. Through experiment, this study shows that traditional machine learning techniques such as logistic regression still can handle fraud detection effectively. Another work which provide comparison between traditional machine learning techniques and deep learning include Raghavan and El Gayar (2019), which benchmarks multiple machine learning algorithms for fraud detection, including support vector machines (SVM), while also exploring the role of deep learning methods such as autoencoders and convolutional neural networks. The experimental result indicates that SVM performs slightly worse, however, still a viable choice in comparison to deep learning techniques. Thennakoon and Bhagyani (2019) discuss the evolution of fraud patterns and the introduction of new forms of fraud, focusing on four main fraud occasions in real-world transactions. The data used experiments come from a financial institution according to a confidential disclosure agreement. They highlight the use of machine learning algorithms, models, and fraud detection systems in real-

time credit card fraud detection. The machine learning models that captured the four fraud patterns (Risky MCC, Unknown web address, ISOResponse Code, Transaction above 100$) with the highest accuracy rates are LR, NB, LR and SVM. Further the models indicated 74%, 83%, 72% and 91% accuracy rates respectively. Dornadula and Geetha (2019) investigate the challenges associated with credit card fraud detection using machine learning algorithms with the aim to overcome challenges by proposing supervised and semi-supervised techniques for fraud detection, such as Isolation Forest or Local Outlier Factor. Among the models, the Random Forest model, with Precision, Recall, and F1-Score of 0.9998, 0.9996, and 0.9996 respectively, demonstrates the best performance.

However, the rapid evolution of technology and the emergence of new techniques pose a challenge to comparative research on fraud detection. Some studies may become outdated quickly due to advancements in technology and the introduction of novel algorithms. Thus, it is crucial to continuously update their comparative analyses to reflect the latest developments in fraud detection. Additionally, the comparability of datasets used in previous studies can be a limitation. Variations in data collection methods, sample sizes, and data quality can impact the validity and generalizability of the findings. Therefore, in this paper, the objectives are extended to incorporate the latest advancements in machine learning and AI by regularly updating the comparative analysis of fraud detection models. It addresses dataset comparability issues by standardizing data collection methods and ensuring high-quality data. Benchmarking new algorithms against traditional models is implemented to evaluate their effectiveness and real-world applicability. The developed models are designed to be adaptable and scalable, capable of integrating future technological advancements and efficiently managing large-scale transaction data.

CHAPTER 3. PROPOSED MODEL

3.1 Proposed methodology

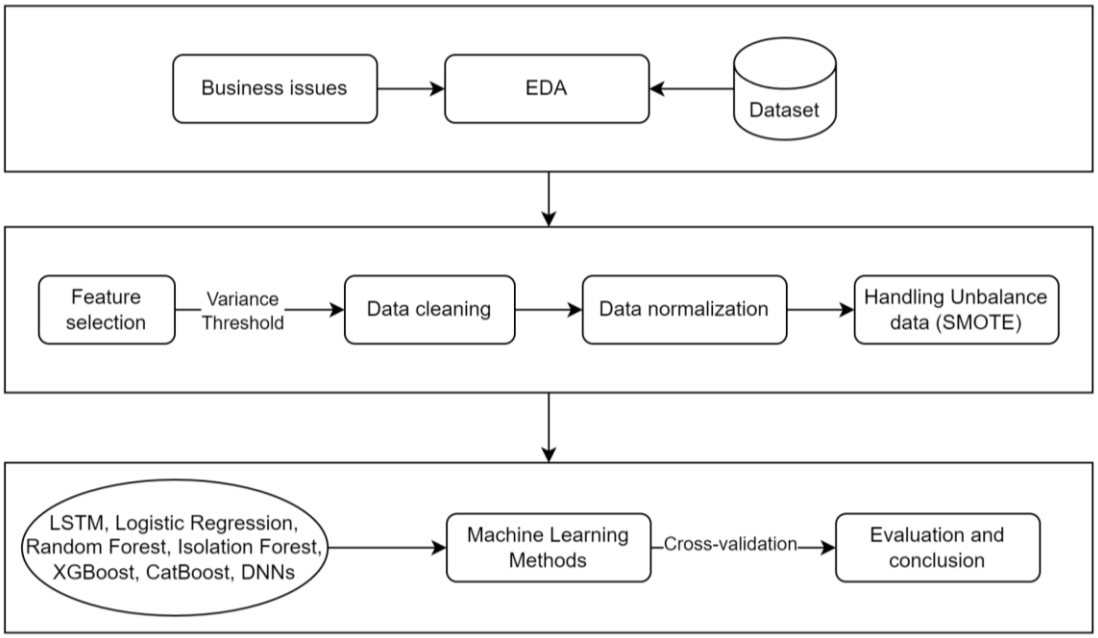

Figure 3.1. Proposed model

In order to provide a comprehensive assessment of state-of-the-art methods, a methodology is proposed to evaluate and implement a multi-step process to identify and prevent fraudulent activities. This methodology outlines the key phases to conduct the comparative experiment:

- *Business problem identification and data collection phase:*
    + Identifying credit fraud is a problem facing businesses.
    + Collect data to understand the characteristics of fraudulent and non-fraudulent transactions, identify missing values, outliers, and potential inconsistencies. Analyze statistical properties and relationships between features to better understand fraud patterns.
- *Data processing stage:*
    + Based on data exploration and understanding of fraud types, identify relevant features that will be used to predict fraud behavior.

+ Clean data by resolving missing values, outliers, and inconsistencies. Data imputation is used for missing values, while bounding and outlier detection techniques handle extreme values.

+ Normalize the data to ensure features scale similarly, which can be important for specific machine learning algorithms. In this study, z-score-based StandardScaler was used to scale the data into the same range of values.

+ Finally, handle if the data is imbalanced with the SMOTE technique.

- *Model selection and training:*

+ Choose the right machine learning algorithm for the fraud detection task.

+ Evaluate and deploy models to select the best performing model, tune and monitor its performance over time to identify fraudulent transactions.

3.2 Exploratory data analysis

3.2.1 Data description

In this study, we use the Credit Card Fraud Detection Dataset 2023 data set collected from Kaggle. Dataset contains credit card transactions made by European cardholders in the year 2023. It comprises 568,630 records, and the data has been anonymized to protect the cardholders' identities.

3.2.2 List of variables

Table 3–1. List of variables

Feature	Description of features	Data type
id	Unique identifier for each transaction	float64
V1-V28	Anonymized features representing various transaction attributes (e.g., time, location, etc.)	int64

Amount	The transaction amount	float64
Class	Binary label indicating whether the transaction is fraudulent (1) or not (0)	int64

3.2.3 Analysis of variables

a. Missing values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568630 entries, 0 to 568629
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   id      568630 non-null  int64
 1   V1      568630 non-null  float64
 2   V2      568630 non-null  float64
 3   V3      568630 non-null  float64
 4   V4      568630 non-null  float64
 5   V5      568630 non-null  float64
 6   V6      568630 non-null  float64
 7   V7      568630 non-null  float64
 8   V8      568630 non-null  float64
 9   V9      568630 non-null  float64
 10  V10     568630 non-null  float64
 11  V11     568630 non-null  float64
 12  V12     568630 non-null  float64
 13  V13     568630 non-null  float64
 14  V14     568630 non-null  float64
 15  V15     568630 non-null  float64
 16  V16     568630 non-null  float64
 17  V17     568630 non-null  float64
 18  V18     568630 non-null  float64
 19  V19     568630 non-null  float64
 20  V20     568630 non-null  float64
 21  V21     568630 non-null  float64
 22  V22     568630 non-null  float64
 23  V23     568630 non-null  float64
 24  V24     568630 non-null  float64
 25  V25     568630 non-null  float64
 26  V26     568630 non-null  float64
 27  V27     568630 non-null  float64
 28  V28     568630 non-null  float64
 29  Amount  568630 non-null  float64
 30  Class   568630 non-null  int64
dtypes: float64(29), int64(2)
```

Figure 3.2. Missing values of variables

Through information from the data set, it can be seen that the variables have been encoded into numeric data and there are no missing values in any cells in the data table.

b. Descriptive analysis

	id	V1	V2	V3	V4	V5	V6	V7	V8	V9
count	568630.000000	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05
mean	284314.500000	-5.638058e-17	-1.319545e-16	-3.518788e-17	-2.879008e-17	7.997245e-18	-3.958636e-17	-3.198898e-17	2.109273e-17	3.998623e-17
std	164149.486121	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00
min	0.000000	-3.495584e+00	-4.996657e+01	-3.183760e+00	-4.951222e+00	-9.952786e+00	-2.111111e+01	-4.351839e+00	-1.075634e+01	-3.751919e+00
25%	142157.250000	-5.652859e-01	-4.866777e-01	-6.492987e-01	-6.560203e-01	-2.934955e-01	-4.458712e-01	-2.835329e-01	-1.922572e-01	-5.687446e-01
50%	284314.500000	-9.363846e-02	-1.358939e-01	3.528579e-04	-7.376152e-02	8.108788e-02	7.871758e-02	2.333659e-01	-1.145242e-01	9.252647e-02
75%	426471.750000	8.326582e-01	3.435552e-01	6.285380e-01	7.070047e-01	4.397368e-01	4.977881e-01	5.259548e-01	4.729905e-02	5.592621e-01
max	568629.000000	2.229046e+00	4.361865e+00	1.412583e+01	3.201536e+00	4.271689e+01	2.616840e+01	2.178730e+02	5.958040e+00	2.027006e+01

	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
count	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05
mean	1.991314e-16	-1.183592e-16	-5.758017e-17	-5.698037e-18	-4.078595e-17	2.649087e-17	-1.719408e-17	-3.398829e-17	-5.837989e-17	2.479146e-17
std	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00
min	-3.163276e+00	-5.954723e+00	-2.020399e+00	-5.955227e+00	-2.107417e+00	-3.861813e+00	-2.214513e+00	-2.484938e+00	-2.421949e+00	-7.804988e+00
25%	-5.901008e-01	-7.014495e-01	-8.311331e-01	-6.966667e-01	-8.732057e-01	-6.212485e-01	-7.162655e-01	-6.194913e-01	-5.560458e-01	-5.653082e-01
50%	2.626145e-01	-4.104986e-02	1.620521e-01	1.760812e-02	2.305011e-01	-3.925566e-02	1.340262e-01	2.716407e-01	8.729382e-02	-2.597869e-02
75%	5.924603e-01	7.477730e-01	7.446723e-01	6.856048e-01	7.518216e-01	6.654065e-01	6.556061e-01	5.182242e-01	5.443887e-01	5.601164e-01
max	3.172271e+01	2.513573e+00	1.791356e+01	7.187486e+00	1.916954e+01	1.453220e+01	4.665291e+01	6.994124e+00	6.783716e+00	3.831672e+00

	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
count	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	5.686300e+05	568630.000000	568630.0
mean	-1.579456e-17	4.758361e-17	3.948640e-18	6.194741e-18	-2.799036e-18	-3.178905e-17	-7.497417e-18	-3.598760e-17	2.609101e-17	12041.957635	0.5
std	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	6919.644449	0.5
min	-7.814784e+01	-1.938252e+01	-7.734798e+00	-3.029545e+01	-4.067968e+00	-1.361263e+01	-8.226969e+00	-1.049863e+01	-3.903524e+01	50.010000	0.0
25%	-3.502399e-01	-1.664408e-01	-4.904892e-01	-2.376289e-01	-6.515801e-01	-5.541485e-01	-6.318948e-01	-3.049607e-01	-2.318783e-01	6054.892500	0.0
50%	-1.233776e-01	-3.743065e-02	-2.732881e-02	-5.968903e-02	1.590123e-02	-8.193162e-03	-1.189208e-02	-1.729111e-01	-1.392973e-02	12030.150000	0.5
75%	2.482164e-01	1.479787e-01	4.638817e-01	1.557153e-01	7.007374e-01	5.500147e-01	6.728879e-01	3.340230e-01	4.095903e-01	18036.330000	1.0
max	2.987281e+01	8.087080e+00	1.263251e+01	3.170763e+01	1.296564e+01	1.462151e+01	5.623285e+00	1.132311e+02	7.725594e+01	24039.930000	1.0

Figure 3.3. Descriptive analysis results

The input variables from V1 to V28 are all numeric, they are the result of some data transformation. The only features that haven't been converted are 'id' and 'Amount'. Feature ‘id’ is just a representation of each individual transaction. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise, it seems to be perfectly balanced in terms of number of cheating and non-fraud cases.

c. Distribution

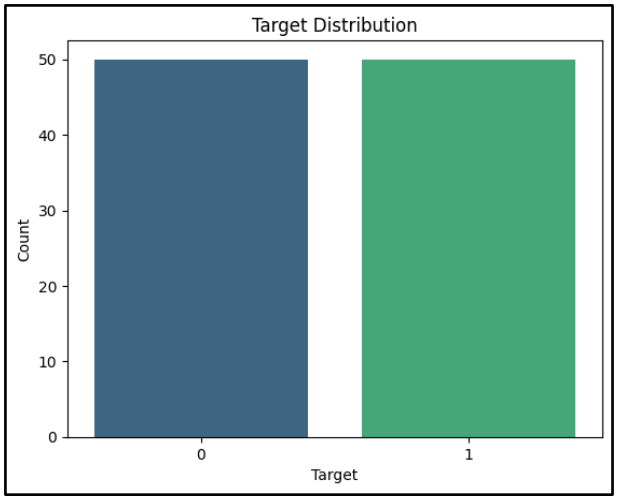

Figure 3.4. Distribution of target feature 'Class'

In this dataset, the histogram of 'Class' shows an equal number of instances for each class, indicating no class imbalance. The equal representation of fraudulent and non-fraudulent transactions in the dataset presents a unique scenario compared to typical real-world financial datasets, which often exhibit a significant class imbalance with far fewer fraudulent transactions.

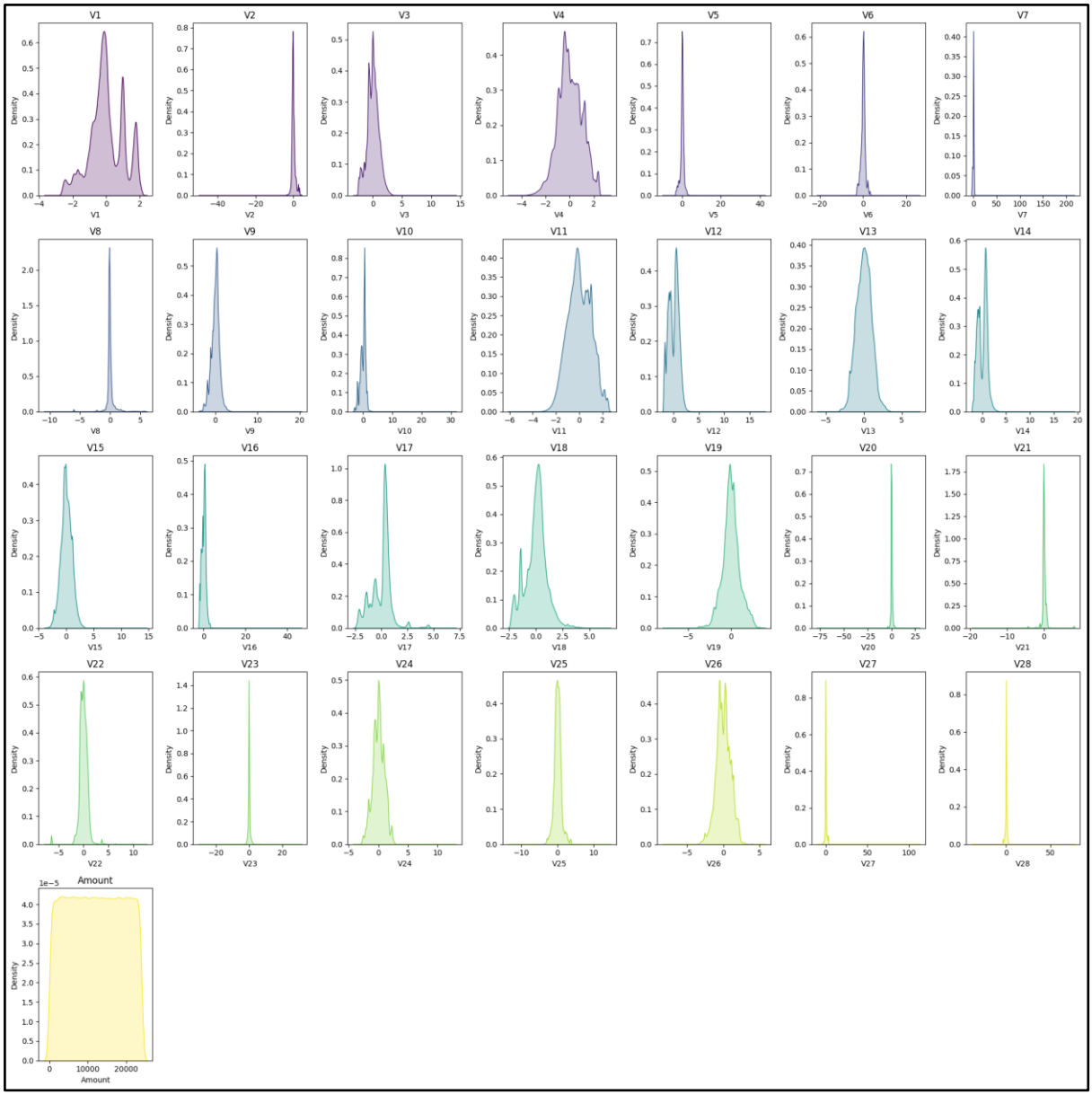

Figure 3.5. Distribution of variables in the dataset

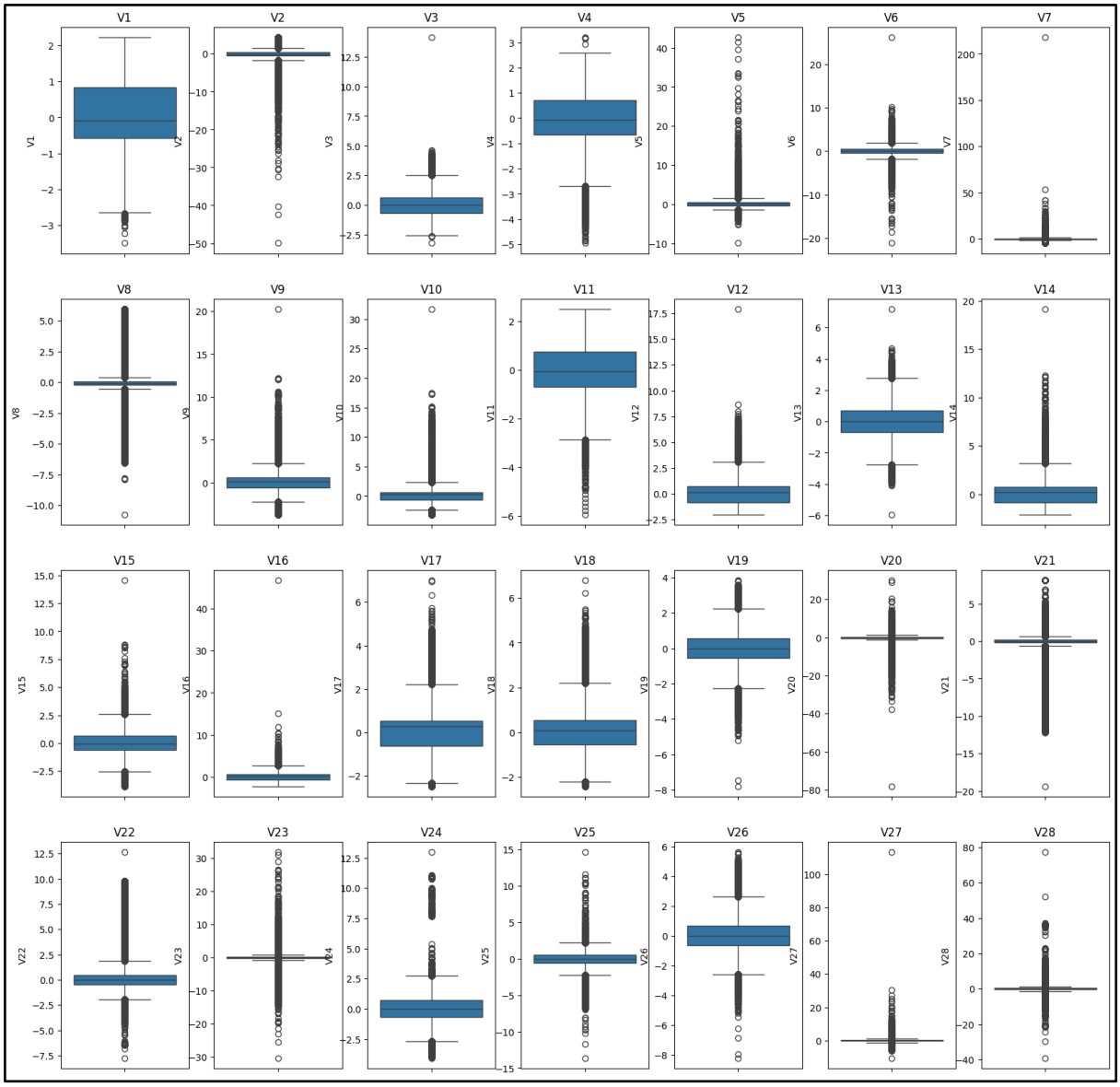

Figure 3.6. Box plot of variables in dataset

The features from V1 to V28 all have a distribution around the value 0. Many of these features, including V1, V3, V5, V13, V14, V20, V22, V23, and V25, exhibit symmetric bell-shaped distributions, characteristic of normal distributions. Other features such as V2, V4, V6, V8, V10, V12, V15, V16, V18, V19, V21, V24, V26, and V27 show skewed distributions or the presence of notable outliers. These patterns suggest these features might be derived through transformation or dimensionality reduction techniques, such as Principal Component Analysis (PCA).

The 'Amount' feature shows a right-skewed distribution, indicating a preponderance of smaller transaction amounts over smaller larger transaction amounts. This bias is typical in financial transaction data.

d. Correlation between variables

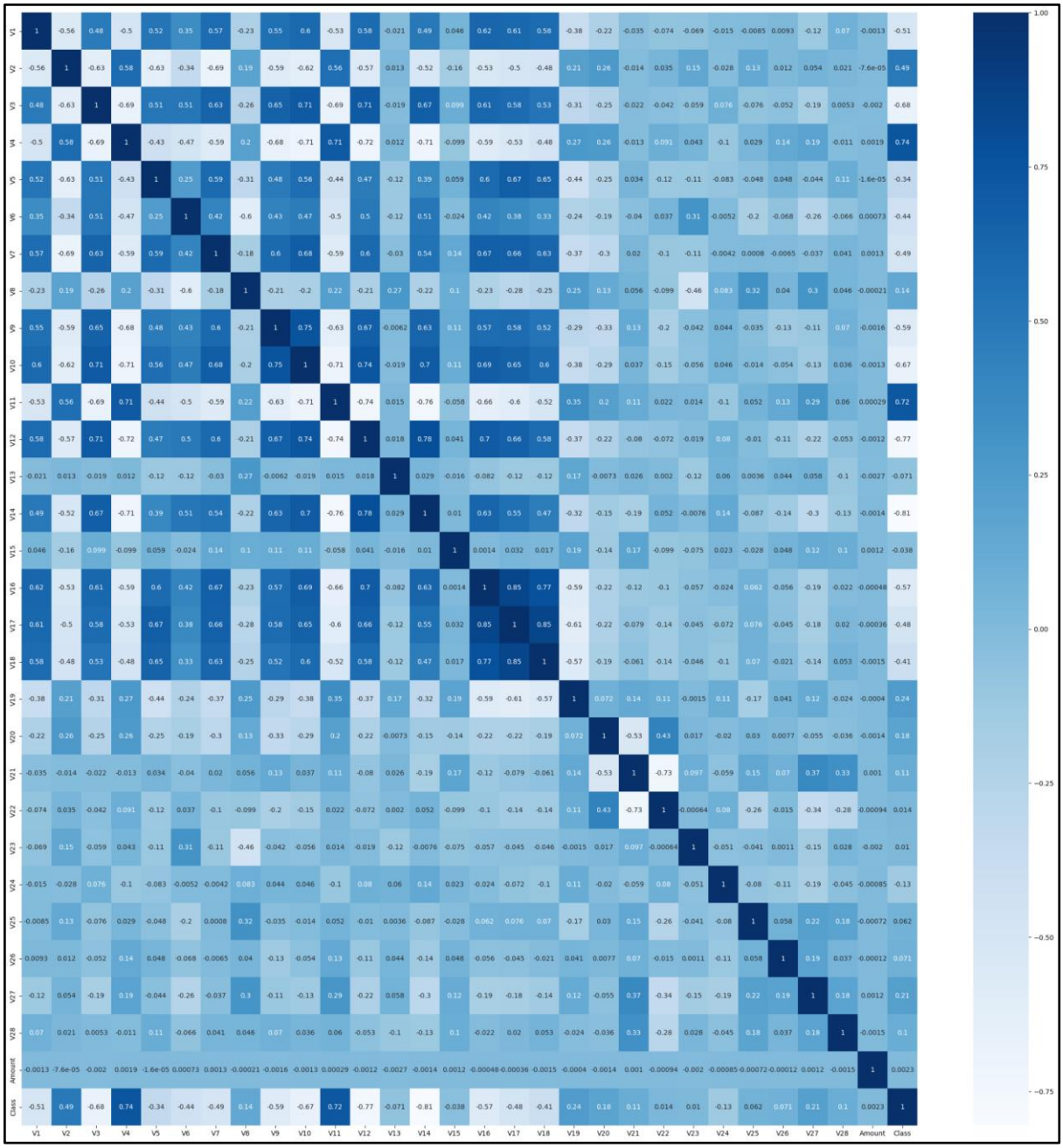

Figure 3.7. Correlation between variables

The correlation table shows the level of correlation between the variable 'Class' and other variables in the data set. From this result, some comments can be drawn as follows:

- The values on the diagonal of the matrix are all 1, indicating a perfect correlation of each variable with itself. Notably, several variables show strong positive correlations, close to 1, as shown by the brightly colored boxes. For example, pairs of variables such as V2 and V3, V3 and V4, V9 and V10, V10 and V11

have a strong positive correlation, which may be because these variables contain similar information or are generated from the same source. data.

- On the contrary, there are pairs of variables that show a strong negative correlation, close to -1, indicated by the black boxes. Pairs such as V2 and V20, V4 and V7, V12 and V28 have strong negative correlations, which may indicate that when one variable increases, the other tends to decrease, possibly due to the opposing nature of the information. which they represent.
- Many other pairs of variables have low or no clear correlation. This shows that many variables do not have a strong linear relationship with each other, possibly because they represent different aspects of the data.
- Variables V1 to V28: These are encrypted variables to protect the customer's identity and transaction details. Correlation values can reflect the degree of correlation between transaction characteristics and the likelihood of fraud. Most of the variables are not highly correlated with the target variable, which shows that no single variable has a strong linear relationship with the target variable, which can lead to the use of many variables in the model. Machine learning for more accurate predictions.Some variables such as V4, V10, V11, V12, V14, and V17 are highly correlated with the variable 'Class', which can indicate some important characteristics of fraudulent transactions.
- 'Amount': This is the amount of the transaction. The low correlation (0.0056) with the variable 'Class' may imply that the amount of the transaction does not clearly reflect the likelihood of fraud.
- V13, V15, V22, V23, V24, V25, V26, V27, V28: These variables have very low correlation with the variable 'Class' (below 0.01), suggesting that they have little or no significant influence on the fraud likelihood of a transaction.

3.3 Data preprocessing

First, we perform correlation analysis between variables to determine the relationship between them and the target class (Class), that is, determine which variables have a significant influence on the classification of transactions as cheating or

not. For this purpose, we use a correlation matrix and determine a correlation threshold of 0.1 to select variables with correlation greater than this threshold.

Next, after identifying important variables, we eliminated variables that did not have a significant correlation with the target variable. The variables that will continue to be used will be eliminated including V13, V15, V22, V23, V25, V26 and the remaining variables will be included in the prediction model. This reduces the number of variables that need to be considered during modeling, which speeds up the process and reduces computational costs.

After removing unnecessary variables, we proceed to remove duplicate rows in the data set. This ensures the uniqueness of each observation, preventing the data from being overestimated or underestimated due to duplication.

Next, we perform data normalization using standard scaling (StandardScaler). This helps bring variables to the same value range, avoiding variables with different value ratios causing unwanted effects on model performance.

After data preprocessing, we divided the dataset into two parts: training set and test set, with 80% and 20% respectively. This helps us train the model on a small portion of the data and evaluate the model's performance on the remaining portion of data on which the model has not been previously trained, thereby helping to accurately evaluate its ability to synthesize. generalization of the model.

After the data processing process, we get input for machine learning models as shown in figure:

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V16	V17	V18	V19	V20	V21	V24	V27	V28	Class
0	-0.260648	-0.469648	2.496266	-0.083724	0.129681	0.732898	0.519014	-0.130006	0.727159	0.637735	...	0.215598	0.512307	0.333644	0.124270	0.091202	-0.110552	0.165959	-0.081230	-0.151045	0
1	0.985100	-0.356045	0.558056	-0.429654	0.277140	0.428605	0.406466	-0.133118	0.347452	0.529808	...	0.789188	0.403810	0.201799	-0.340687	-0.233984	-0.194936	-0.577395	-0.248052	-0.064512	0
2	-0.260272	-0.949385	1.728538	-0.457986	0.074062	1.419481	0.743511	-0.095576	-0.261297	0.690708	...	-0.577514	0.886526	0.239442	-2.366079	0.361652	-0.005020	-1.154666	-0.300258	-0.244718	0
3	-0.152152	-0.508959	1.746840	-1.090178	0.249486	1.143312	0.518269	-0.065130	-0.205698	0.575231	...	-0.030669	0.242629	2.178616	-1.345060	-0.378223	-0.146927	-1.893131	-0.165316	0.048424	0
4	-0.206820	-0.165280	1.527053	-0.448293	0.106125	0.530549	0.658849	-0.212660	1.049921	0.968046	...	0.224538	0.366466	0.291782	0.445317	0.247237	-0.106984	0.312561	0.023712	0.419117	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
568625	-0.833437	0.061886	-0.899794	0.904227	-1.002401	0.481454	-0.370393	0.189694	-0.938153	-1.161847	...	-1.480796	-1.520928	-1.376970	1.789103	-0.751011	0.167503	-0.900861	3.308968	0.081564	1
568626	-0.670459	-0.202896	-0.068129	-0.267328	-0.133660	0.237148	-0.016935	-0.147733	0.483894	-0.210817	...	-0.545184	-0.575991	-0.664313	0.101604	-0.550260	0.031874	-0.846452	-1.528642	1.704306	1
568627	-0.311997	-0.004095	0.137526	-0.035893	-0.042291	0.121098	-0.070958	-0.019997	-0.122048	-0.144495	...	-0.370201	-0.729002	-0.251679	-0.343196	-0.076417	0.140788	-0.448909	-0.487540	-0.268741	1
568628	0.636871	-0.516970	-0.300889	-0.144480	0.131042	-0.294148	0.580568	-0.207723	0.893527	-0.080078	...	0.477402	0.848443	0.930280	-0.481058	0.288186	-0.060381	-0.554643	-0.159269	-0.076251	1
568629	-0.795144	0.433236	-0.649140	0.374732	-0.244976	-0.603493	-0.347613	-0.340814	0.253971	-0.513556	...	-0.917240	-0.936114	-0.823688	-0.330408	-0.621378	0.534853	0.931030	-1.575113	0.722936	1

568630 rows × 23 columns

Figure 3.8. Data after processing

CHAPTER 4. EXPERIMENT RESULT

4.1 Evaluation parameters

4.1.1 Confusion matrix

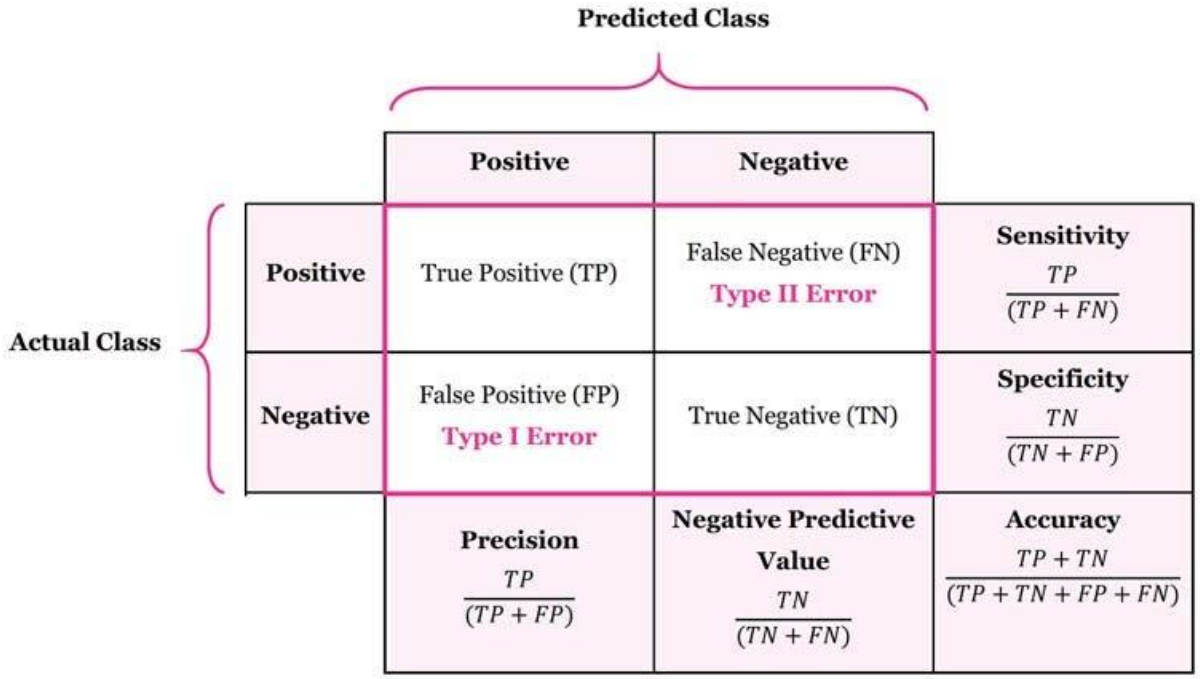

Figure 4.1. Confusion matrix

Source: How to Remember all these Classification Concepts forever

Confusion matrix is an important tool to evaluate the performance of prediction models in classification problems. It represents the number of correct and incorrect predictions for each data class, helping to accurately evaluate the model's classification ability. This matrix includes components such as True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). By considering TP, FP, TN and FN values, we can calculate measures such as accuracy, precision, recall and F1 score to evaluate and compare the performance of different models. The confusion matrix provides detailed information about the performance of the classification model and helps us accurately evaluate the classification ability of the prediction model.

4.1.2 Accuracy, Precision, Recall, F1 Score

Accuracy, precision, recall and F1 score are important metrics to evaluate the performance of prediction models in classification problems. These are measures

calculated from the confusion matrix to provide detailed information about the model's ability to accurately classify and cover each data layer.

- Accuracy is the most basic metric, representing the overall proportion of correct predictions made by the model. It is calculated as:

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ predictions}$$

- Precision focuses specifically on the positive predictions made by the model. Precision is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

- Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

- F1 score is calculated as the harmonic mean of the precision and recall scores, it is calculated as:

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

4.1.3 ROC-AUC

Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve itself is a graphical representation that plots the true positive rate (sensitivity) against the false positive rate for various threshold settings, providing insight into the trade-offs between benefits (true positives) and costs (false positives). The AUC value, ranging from 0 to 1, quantifies the overall ability of the model to discriminate between positive and negative classes, with a score of 1 indicating perfect discrimination and a score of 0.5 representing a model with no discriminative power, equivalent to random guessing. By evaluating the AUC-ROC, researchers can objectively compare the performance of different models, regardless of the chosen threshold, making it an invaluable tool in the selection and tuning of predictive algorithms.

4.2 Experience results

4.2.1 Machine Learning and Deep Learning models

The evaluation of various models for credit card fraud detection reveals distinct strengths and weaknesses among the methods tested. The models considered include Logistic Regression, Isolation Forest, Random Forest, CatBoost, Deep Neural Networks, XGBoost, and LSTM. Their performance metrics provide valuable insights into their suitability for detecting fraudulent transactions. The following results was obtained:

```
Logistic Regression:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     56765
           1       0.98      0.95      0.97     56961

    accuracy                           0.97    113726
   macro avg       0.97      0.97      0.97    113726
weighted avg       0.97      0.97      0.97    113726

Accuracy: 0.9660
CPU times: user 8.66 s, sys: 3.56 s, total: 12.2 s
Wall time: 3.77 s
```

Figure 4.2. Prediction results using Logistic Regression

Logistic Regression: This model achieved a balanced accuracy of 96.60%, with both precision and recall values around 0.95 to 0.98 for both classes. The F1-scores for both classes are 0.97, indicating that Logistic Regression is a strong performer for this dataset, providing a good trade-off between precision and recall.

```
Random Forest:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56765
           1       1.00      1.00      1.00     56961

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuracy: 0.9999
CPU times: user 7min 48s, sys: 6.39 ms, total: 7min 48s
Wall time: 7min 49s
```

Figure 4.3. Prediction results using Random Forest

Random Forest: Achieved near-perfect performance with an accuracy of 99.99%. Both precision and recall for each class were 1.00, indicating the model was able to perfectly classify all instances in the test set. However, this perfect performance could indicate overfitting, especially given the balanced nature of the dataset.

```
Isolation Forest:
              precision    recall  f1-score   support

           0       0.50      1.00      0.67     56765
           1       0.82      0.00      0.00     56961

    accuracy                           0.50    113726
   macro avg       0.66      0.50      0.33    113726
weighted avg       0.66      0.50      0.33    113726

Accuracy: 0.4998
CPU times: user 19.6 s, sys: 214 ms, total: 19.8 s
Wall time: 19.9 s
```

Figure 4.4. Prediction results using Isolation Forest

Isolation Forest: The Isolation Forest model performed poorly, with an overall accuracy of 49.98%. It demonstrated a high recall (1.00) for the non-fraud class but failed to identify fraudulent transactions (recall of 0.00). This suggests that the model is not suitable for this balanced dataset, likely due to its unsupervised nature.

```
XGBoost:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56765
           1       1.00      1.00      1.00     56961

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuracy: 0.9969
CPU times: user 28.8 s, sys: 97 ms, total: 28.9 s
Wall time: 7.99 s
```

Figure 4.5. Prediction results using XGBoost

XGBoost: Achieved an accuracy of 99.69%, with precision, recall, and F1-scores all at 1.00. XGBoost provides a good balance of high performance and relatively shorter training times compared to some other models.

```
CatBoost:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56765
           1       1.00      1.00      1.00     56961

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuracy: 0.9995
CPU times: user 4min 35s, sys: 5.96 s, total: 4min 41s
Wall time: 1min 14s
```

Figure 4.6. Prediction results using CatBoost

CatBoost: Similar to Random Forest, CatBoost showed an almost perfect accuracy of 99.95%, with all precision, recall, and F1-scores at 1.00. CatBoost is highly effective, but the potential for overfitting should be monitored.

```
LSTM:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56765
           1       1.00      1.00      1.00     56961

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuracy: 0.9993
CPU times: user 11min 42s, sys: 1min 4s, total: 12min 46s
Wall time: 6min 55s
```

Figure 4.7. Prediction results using Long Short Term Memory

LSTM: Exhibited near-perfect performance with an accuracy of 99.93%. Precision, recall, and F1-scores were all at 1.00. However, the training time was considerably longer, indicating higher computational costs.

```
Deep Neural Networks:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56765
           1       1.00      1.00      1.00     56961

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuracy: 0.9988
CPU times: user 5min 23s, sys: 30.3 s, total: 5min 54s
Wall time: 3min 41s
```

Figure 4.8. Prediction results using Deep Neural Networks

Deep Neural Networks: This model also achieved high accuracy (99.88%) with perfect precision, recall, and F1-scores. Neural networks are powerful, but their training time is significantly longer compared to other models.

From the results of the above models, it can be seen that:

- Logistic regression: Effective and efficient but slightly less accurate than more complex models.

- Isolated forest: Poor performance on balanced data equal, not suitable for fraud detection.
- Random Forest, CatBoost, Deep Neural Networks, XGBoost, and LSTM: All demonstrate near-perfect accuracy and metrics. CatBoost and XGBoost stand out for their combination of high accuracy and computational efficiency.

4.2.2 Evaluation

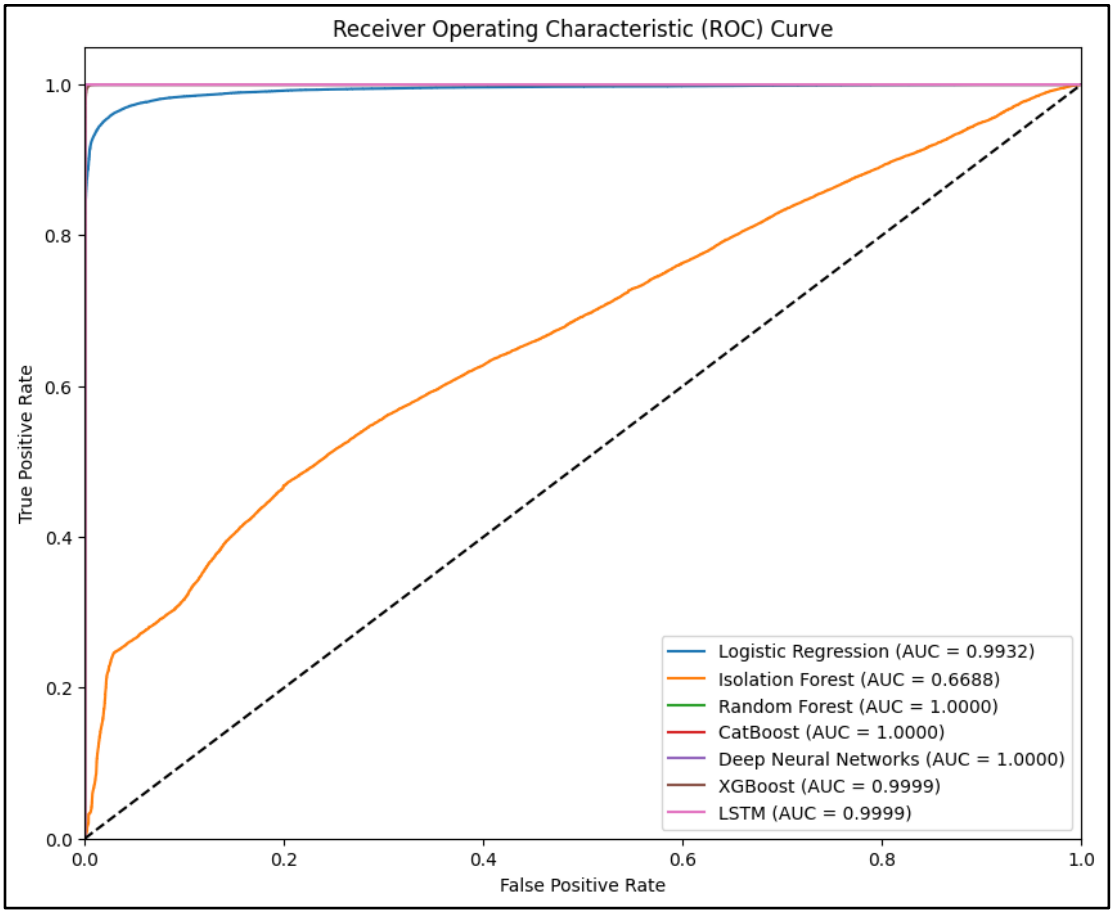

Figure 4.9. ROC Curve for Fraud Detection Models

The ROC curve analysis reaffirms the overall evaluation of the models' performance:

- Isolation Forest: Performs poorly, with a low AUC, indicating it is not suitable for this balanced dataset in fraud detection.

- Logistic Regression: Shows high effectiveness with an AUC of 0.9932, making it a strong contender for practical applications given its balance of performance and computational efficiency.
- Random Forest, CatBoost, Deep Neural Networks, XGBoost, and LSTM: All these models demonstrate near-perfect to perfect AUC scores, signifying their excellent ability to differentiate between fraudulent and non-fraudulent transactions.

Logistic Regression is noted for its effectiveness and efficiency. Despite being a simpler model, it achieves a high Area Under the Curve (AUC) score of 0.9932, making it a viable option for practical applications. Its balance between performance and computational demands makes it particularly appealing for real-time fraud detection scenarios where speed and resource efficiency are critical.

However, Isolation Forest falls short in this evaluation. Its performance on balanced datasets is poor, as reflected in its low AUC score. This indicates that Isolation Forest is not well-suited for fraud detection, particularly when dealing with datasets where fraudulent and non-fraudulent transactions are balanced. Its inability to adequately distinguish between these classes limits its utility in this context.

Meanwhile, the ensemble and more complex models—Random Forest, CatBoost, Deep Neural Networks, XGBoost, and LSTM—all show near-perfect accuracy and metrics. These models demonstrate excellent capabilities in differentiating between fraudulent and legitimate transactions, as evidenced by their near-perfect AUC scores. Among these, CatBoost and XGBoost stand out due to their combination of high accuracy and computational efficiency. These characteristics make them particularly suitable for deployment in environments where both performance and efficiency are paramount.

CHAPTER 5. CONCLUSION AND FUTURE WORKS

5.1 Discussion and Recommendation

5.1.1 Discussion

Our study indicates that CatBoost and XGBoost are highly effective for fraud detection due to their superior accuracy and computational efficiency. These models excel in distinguishing between fraudulent and non-fraudulent transactions, as corroborated by Sivanantham et al. (2021), who reported that these models outperformed others in both accuracy and computational efficiency.

Despite its simplicity, Logistic Regression also demonstrated strong performance with an AUC of 0.9932, making it a viable option for businesses with limited computational resources. Albrecht et al. (2016) emphasized its effectiveness in fraud detection, aligning with our findings.

Additionally, we have also identified several factors influencing the "y" variable in these models. Firstly, the transaction amount, one of the most crucial factors in fraud detection. Larger transaction amounts often attract closer attention due to the increased financial risk they pose. Fraudsters exploit this by frequently engaging in transactions involving substantial amounts within a short timeframe to maximize their illicit gains. Abdallah et al. (2016) found a direct correlation between higher transaction amounts and the likelihood of fraud detection, underscoring the significance of this factor in identifying suspicious activities.

Similarly, the timing of transactions can serve as a telltale sign of potential fraud. Transactions conducted during unusual hours, such as late at night or early morning, stand out as suspicious, particularly if they deviate from the user's typical behavior. Bhattacharyya et al. (2011) demonstrated this relationship, indicating that transactions occurring outside of normal business hours exhibit a heightened risk of fraudulent activity.

Another significant variable we analyzed was the location of transactions. Transactions originating from unfamiliar locations, especially those significantly distant from the user's usual area of activity, raise red flags. Ngai et al. (2011)

emphasized the significance of geographic anomalies in detecting fraud, noting that transactions from atypical locations carry a higher probability of fraudulent behavior.

5.1.2 Recommendation

Based on our findings, we recommend the following for businesses aiming to enhance their fraud detection capabilities. First, adopting advanced models like CatBoost and XGBoost can significantly improve fraud detection rates due to their ability to handle complex patterns and large datasets. These models' high accuracy and computational efficiency make them ideal for this task. For businesses with limited computational resources, Logistic Regression offers a practical and efficient alternative, as it has demonstrated strong performance in our study and in the research of Albrecht et al. (2016).

To ensure effective fraud detection, businesses should tailor their model selection to specific requirements. This involves considering factors such as dataset size, quality, computational resources, and the expertise of their analytics team. Continuous evaluation and updating of models are vital to adapt to evolving fraud tactics, ensuring long-term effectiveness. Additionally, investing in data quality is paramount for machine learning models' performance. Therefore, there should be a focus on data cleaning and preprocessing to enhance the reliability of fraud detection systems.

Furthermore, combining multiple models through ensemble methods can further enhance accuracy, providing a robust defense against fraudulent activities. A hybrid approach, utilizing both advanced models like XGBoost and simpler models like Logistic Regression, offers comprehensive fraud protection. Moreover, implementing real-time performance tracking systems enables immediate optimization and response to emerging issues, maintaining an effective and up-to-date fraud detection system.

By integrating these strategies, businesses can significantly bolster their fraud detection capabilities, safeguard financial operations, and foster a more secure and trustworthy financial ecosystem.

5.2 Conclusion

This study underscores the efficacy of advanced machine learning models, particularly CatBoost and XGBoost, in fraud detection. These models exhibit high accuracy and computational efficiency, making them excellent choices for identifying fraudulent transactions. Logistic Regression, despite being a simpler model, also showed commendable performance, highlighting its practical utility for businesses with limited computational resources.

Our findings are consistent with prior research, reinforcing the suitability of both sophisticated and traditional models based on specific business needs. The study emphasizes the importance of selecting models considering factors such as dataset characteristics, available computational power, and team expertise. By implementing the recommended models, businesses can significantly enhance their fraud detection capabilities, contributing to a more secure financial environment.

5.3 Limitations and Future works

5.3.1 Limitations

Although the study yielded promising results, some limitations must be acknowledged. First, the analysis is based on a specific data set, raising concerns about its generalizability to many different types of financial fraud and geographies, especially since the data is too complete. perfect results in no imbalance like real data. Additionally, the study used standard feature engineering techniques, potentially ignoring more advanced or domain-specific features that could enhance model performance.

Furthermore, the experiments were conducted in a controlled environment with ample computational resources, potentially diverging from real-world scenarios where resource constraints might impact model feasibility and accuracy. Moreover, the use of complex models such as CatBoost and XGBoost, while advantageous in terms of accuracy, presents challenges in interpretability compared to simpler models like Logistic Regression. This lack of interpretability could pose obstacles in regulatory compliance and decision-making processes.

Finally, the evolving nature of fraudulent techniques poses a significant challenge. Models trained on historical data may not effectively detect new fraud patterns, necessitating continual updates to maintain efficacy. Addressing these limitations is crucial for developing robust fraud detection systems capable of adapting to dynamic threats.

5.3.2 Future Works

Future research aims to address current limitations and explore new ways to enhance fraud detection systems. By using different datasets (including imbalanced data), advanced feature engineering, and adaptive models, future work can develop robust and efficient solutions. A hybrid approach combining the strengths of multiple machine learning approaches holds the promise of providing a comprehensive solution that can address the evolving challenges of financial fraud.

- Enhanced Feature Engineering: Future research should focus on refining feature engineering techniques to extract more meaningful insights from customer data. This could include looking for innovations from customer systems, temporal trends, and customers in practice. By incorporating fine-grained features, machine learning models can accurately capture the subtle nuances of fraudulent activity.
- Adaptive model architectures: There is a need to develop adaptive model architectures that can automatically learn and evolve in response to emerging fraud patterns. Future work may explore the integration of deep learning algorithms, such as recurrent neural networks (RNNs) and transformers, which can capture time dependence and complex interactions in networks. Furthermore, providing meta-learning models have been able to adapt to evolving fraud mechanisms over time -a concept that can be used.
- Defining AI for translation: Ensuring that fraud detection processes are transparent and transparent is essential to building trust and facilitating compliance. Future research could focus on incorporating interpretable AI techniques to provide translatable insights into model predictions. Stakeholders

can understand better by clarifying the rationale behind fraud detection decisions.

- Stronger Adversaries: Preventing adversary attack threats is more important than protecting fraud detection systems from sophisticated adversaries. Future works may investigate adversary training methods to enhance the robustness of the model in adversarial disturbance of the input data. In addition, research efforts can investigate anomaly detection techniques that can detect enemy gestures in real time.
- Privacy protection strategies: Maintaining the privacy and confidentiality of sensitive financial data in fraudulent applications is critical Future research could examine privacy protection strategies, such as integrated learning, privacy a separate, and equal privacy. Enabling collective fraud detection without compromising data security. Protecting data privacy enables organizations to leverage collective intelligence while protecting customer privacy.
- Real-time analysis and response: Enhancing real-time analysis and response capabilities is essential to reduce the loss of revenue due to fraudulent activities. Future work will focus on developing scalable and efficient algorithms that can handle large amounts of communication information in real time. By leveraging streaming analytics and distributed computing frameworks, organizations can quickly detect and respond to fraud incidents, minimizing potential damage
- Interagency Collaboration: Encouraging interagency cooperation and information sharing is essential to combat the increasing prevalence of fraudulent schemes. Future research efforts could facilitate the establishment of forums and collaborative associations where industry members can exchange insights, share best practices, and collaborate on fraud prevention initiatives.

REFERENCE

[1] Coursera Staff. (2024, April 19). *Data analytics*. Coursera. Retrieved 15/05/2024 from https://www.coursera.org/articles/data-analytics

[2] MastersInDataScience.org. (2023, October). *What Is Data Analytics?* Retrieved 15/05/2024 from https://www.mastersindatascience.org/learning/what-is-data-analytics/

[3] SigmaMagic. (n.d.). *Analytics: Advantages and Limitations*. Retrieved 15/05/2024 from https://www.sigmamagic.com/blogs/analytics-advantages-and-limitations/

[4] Yandex. (n.d.). *CatBoost documentation*. Retrieved 18/05/2024 from https://catboost.ai/en/docs/

[5] Albrecht, C. C., Sanders, M. L., Holland, D. V., & Albrecht, C. (2016). *The debilitating effects of fraud in organizations*. In Crime and Corruption in Organizations (pp. 163-185). Routledge.

[6] West, J., & Bhattacharya, M. (2016). *Intelligent financial fraud detection: a comprehensive review*. Computers & security, 57, 47-66.

[7] Sivanantham, S., Dhinagar, S. R., Kawin, P., & Amarnath, J. (2021). *Hybrid approach using machine learning techniques in credit card fraud detection*. In Advances in Smart System Technologies: Select Proceedings of ICFSST 2019 (pp. 243-251). Springer Singapore.

[8] Yang, D. H., Li, Z. Y., Wang, X. H., Salamatian, K., & Xie, G. G. (2021). *Exploiting the Community Structure of Fraudulent Keywords for Fraud Detection in Web Search*. Journal of Computer Science and Technology, 36, 1167-1183.

[9] Bhatla, T. P., Prabhu, V., & Dua, A. (2003). *Understanding credit card frauds*. Cards business review, 1(6), 1-15.

[10] Chaudhary, K., Yadav, J., & Mallick, B. (2012). *A review of fraud detection techniques: Credit card*. International Journal of Computer Applications, 45(1), 39-44.

[11] Sadgali, I., Sael, N., & Benabbou, F. (2018). *Detection of credit card fraud: State of art*. Int. J. Comput. Sci. Netw. Secur, 18(11), 76-83.

[12] Delamaire, L., Abdou, H. A. H., & Pointon, J. (2009). *Credit card fraud and detection techniques: a review*. Banks and Bank systems, 4(2).

[13] Li, F.-F., Yu, K., Yamanishi, M., Liu, D., Zhou, X., & Zhou, Z.-H. (2022). Isolation forest. In *Engineering Applications of Artificial Intelligence* (Vol. 112, pp. 225-238). Elsevier.

[14] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features*. Advances in neural information processing systems, 31.

[15] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). *Credit card fraud detection using machine learning techniques: A comparative analysis*. In 2017 international conference on computing networking and informatics (ICCNI) (pp. 1-9). IEEE.

[16] Raghavan, P., & El Gayar, N. (2019, December). *Fraud detection using machine learning and deep learning*. In 2019 international conference on computational intelligence and knowledge economy (ICCIKE) (pp. 334-339). IEEE.

[17] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019, January). *Real-time credit card fraud detection using machine learning*. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 488-493). IEEE.

[18] Dornadula, V. N., & Geetha, S. (2019). *Credit card fraud detection using machine learning algorithms*. Procedia computer science, 165, 631-641.

[19] Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons.

[20] Siddiqi, N. (2017). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. John Wiley & Sons.

[21] Agresti, A. (2002). Categorical Data Analysis. Wiley-Interscience.

[22] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-242.

[23] McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109-142.

[24] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[25] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[26] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

[27] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[28] Yu, H., Zhu, Z., & Yang, Q. (2019). Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689.

[29] Rasley, J., Huang, S., Huang, J., Zhang, Y., & Zhang, Z. (2020). HyperDrive: Exploring hyperparameters with POP scheduling. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.

[30] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).