A Mini Project Report

On

"CREDIT CARD FRAUD DETECTION"

By

Prateek Rao

2300520700044

Under the guidance of

Dr. Harshmit Kaur Saluja

In partial fulfilment of the requirement for the award of the Degree of

Master of Business Administration

Submitted at

DEPARTMENT OF BUSINESS ADMINISTRATION

INSTITITUTE OF ENGINEERING AND TECHNOLOGY

JANKIPURAM LUCKNOW, UP-226021

2024

Department of Business Administration

Institute of Engineering & Technology Lucknow

STUDENT DECLARATION

I undersigned, hereby declare that the project titled "**Credit Card Fraud Detection**" submitted in partial fulfilment for the award of Degree of Master of Business Administration is a Bonafede record of work done by me under the guidance of "**Dr. Harshmit Kaur Saluja**". This report has not previously formed the basis for the award of any degree, diploma, or similar title of any University.

Place:

Date:

Signature of the Student

Prateek Rao

Roll No. 2300520700044

Department of Business Administration
Institute of Engineering & Technology Lucknow

CERTIFICATE FROM INSTITUTION

This is to certify that "**Mr. Prateek Rao**" Second semester student of Master of Business Administration, Institute of Engineering & Technology, Sitapur Road, Lucknow has completed the project report entitled "**Credit Card Fraud detection**" in partial fulfilment of the requirements for the award of the Degree of Master of Business Administration.

Date:
Place:

Dr. Durgawati Kushwaha
Convener

Department of Business Administration
Institute of Engineering & Technology Lucknow

CERTIFICATE FROM FACULTY GUIDE

This is to certify that "**Mr. Prateek Rao**" First semester student of Master of Business Administration, Institute of Engineering & Technology, Sitapur Road, Lucknow has completed the project report entitled "**Credit Card Fraud detection**" towards partial fulfilment of the requirement for the award of the Degree of Master of Business Administration under my supervision.

Date:
Place:

Dr. Harshmit Kaur Saluja
(Assistant Professor)

PREFACE

The project titled "**Credit Card Fraud Detection**" for the Banking Sector Using Data Science aims to address the critical issue of credit card fraud, which poses significant financial risks to both consumers and financial institutions. My interest in this subject stem from a desire to leverage data science techniques to enhance security and trust within the banking sector.

In this project, I utilize various machine learning algorithms and data analysis methods to identify and prevent fraudulent activities in credit card transactions. By analysing large datasets of transaction history, the model aims to detect patterns and anomalies that are indicative of fraud. The project's success relies on accurately distinguishing between legitimate and fraudulent transactions, thus minimizing false positives and negatives.

This endeavour not only highlights the practical application of data science in solving real-world problems but also underscores the importance of advanced analytics in the financial industry. I hope this project contributes to the development of more robust and efficient fraud detection systems, ultimately protecting consumers and enhancing the integrity of financial transactions.

ACKNOWLEDGEMENT

I am heartily thankful to the Institute of engineering and technology, Lucknow for providing me all the facilities and infrastructure to take my work to the final stage.

I would like to express my deep sense of respect and gratitude towards my advisor and guide **DR. HARSHMIT KAUR SALUJA**, Department of Business Administration, IET Lucknow who has given me an opportunity to work under her. She has been a constant source of inspiration throughout my work. Her invaluable knowledge and innovative ideas helped me to take the work to the final stage.

Last but not the least I am extremely thankful to all who have directly or indirectly helped me for the completion of my project.

Prateek Rao
Roll -2300520700044
MBA 2st Semester

TABLE OF CONTENT

CHAPTER: - 1

Introduction: -

Credit card fraud detection using data science has become increasingly vital in the era spanning 2020 to 2024, as the digital economy continues to expand and evolve. This period has witnessed a surge in online transactions, exacerbated by the global pandemic, which accelerated the shift towards e-commerce and digital payments. As a result, the threat landscape for financial fraud has broadened, necessitating advanced and sophisticated methods to safeguard consumer trust and financial institutions' integrity. Data science, leveraging vast amounts of transactional data, has emerged as a cornerstone in combating credit card fraud. Utilizing techniques such as machine learning, artificial intelligence, and big data analytics, data scientists can analyse complex patterns and detect anomalies that signal fraudulent activity. These techniques go beyond traditional rule-based systems by dynamically learning from data, adapting to new fraud tactics, and predicting potential threats with high accuracy. The period from 2020 to 2024 has also seen significant advancements in algorithmic development, computational power, and the availability of real-time data, all contributing to more effective fraud detection systems. Techniques such as supervised and unsupervised learning, neural networks, and clustering have been instrumental in identifying fraudulent transactions among millions of legitimate ones. Furthermore, the integration of behavioural analytics, which examines the spending patterns and behaviours of individuals, has enhanced the precision of fraud detection models. This proactive approach not only mitigates financial losses but also enhances customer experience by reducing false positives. The collaborative efforts between financial institutions, regulatory bodies, and technology providers during this period have further strengthened the fight against credit card fraud. By sharing insights and developing standardized protocols, the industry has made strides in creating a more secure financial ecosystem. In summary, the years 2020 to 2024 have underscored the critical role of data science in credit card fraud detection, highlighting the need for continuous innovation and collaboration to stay ahead of increasingly sophisticated fraudsters.

Issues and challenges: -

1.1.1 Credit card fraud refers to the unauthorized use of someone else's credit card information to make purchases or withdraw funds without their consent. It is a form of financial fraud that can result in monetary losses for both the cardholder and the financial institution. Credit card fraud can take various forms, including:

1. **Stolen Card Fraud**: This occurs when a thief steals a physical credit card and uses it to make unauthorized purchases before the cardholder realizes the card is missing.
2. **Card Not Present (CNP) Fraud**: In CNP fraud, the fraudster uses stolen credit card information (card number, expiration date, and security code) to make online or phone purchases where the physical card is not required.
3. **Lost or Mislaid Card Fraud**: Similar to stolen card fraud, lost or mislaid card fraud involves finding a lost credit card and using it for unauthorized transactions.
4. **Account Takeover**: In this type of fraud, the fraudster gains unauthorized access to a cardholder's account information, either through phishing scams, hacking, or other means. Once access is gained, they can make fraudulent transactions or change account details.
5. **Counterfeit Card Fraud**: Fraudsters create counterfeit credit cards using stolen card information and encode it onto a blank card. These counterfeit cards are then used to make purchases in-person at retail stores or to withdraw cash from ATMs.
6. **Application Fraud**: This occurs when someone applies for a credit card using stolen or falsified personal information. Once approved, the fraudster can use the credit card for unauthorized transactions.
7. **Friendly Fraud**: Sometimes referred to as chargeback fraud, friendly fraud occurs when a legitimate cardholder disputes a valid transaction to receive a refund from the card issuer, effectively committing fraud against the merchant.
8. **Skimming**: Skimming involves using a small device called a skimmer to steal credit card information during legitimate transactions, such as at ATMs or point-of-sale terminals. The skimmer captures the card's magnetic stripe data, which can then be used to create counterfeit cards or make online purchases.
9. **Identity Theft**: Identity theft involves stealing a person's personal information, including their credit card details, to open new credit card accounts or commit other fraudulent activities in their name.
10. **Phishing**: Phishing scams involve tricking individuals into providing their credit card information by posing as legitimate entities, such as banks, retailers, or government agencies, through email, phone calls, or fake websites.

With the increase of people using credit cards in their daily lives, credit card companies should take special care in the security and safety of their customers. According to credit card statistics from 2021, the number of people using credit cards around the world was 2.8 billion in 2019, with 70% of those users owning at least one card. Reports of credit card fraud in the U.S. rose by 44.7%, from 271,927 in 2019 to 393,207 reports in 2020. There are two main types of credit card fraud. The first type involves a credit card account being opened under your name by an identity thief; reports of this fraudulent behaviour increased by 48% from 2019 to 2020. The second type involves an identity thief using an existing account that you created, typically by stealing the credit card information; reports of this type of fraud increased by 9% from 2019 to 2020 (Daly, 2021). These statistics highlight the drastic and rapid increase in credit card fraud, underscoring the need for analytical solutions using various machine learning methods to detect fraudulent transactions within numerous transactions.


VISA

Application of emerging technology: -

What is Data Science?

Data science is an interdisciplinary field that combines mathematics and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights guide decision-making and strategic planning. The accelerating volume of data sources and data has made data science one of the fastest-growing fields across industries. Data scientists are increasingly relied upon to interpret data and provide actionable recommendations to improve business outcomes.

The Data Science Lifecycle

The data science lifecycle involves various roles, tools, and processes that enable analysts to glean actionable insights. Typically, a data science project undergoes the following stages:

1. **Data Ingestion**: The lifecycle begins with data collection from all relevant sources using various methods, including manual entry, web scraping, and real-time streaming from systems and devices. Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the Internet of Things (IoT), and social media.
2. **Data Storage and Processing**: Since data can have different formats and structures, companies need to consider different storage systems based on the type of data. Data management teams set standards around data storage and structure to facilitate workflows around analytics, machine learning, and deep learning models. This stage includes cleaning, deduplicating, transforming, and combining the data using ETL (extract, transform, load) jobs or other data integration technologies. Data preparation is essential for promoting data quality before loading into a data warehouse, data lake, or other repository.
3. **Data Analysis**: Data scientists conduct exploratory data analysis to examine biases, patterns, ranges, and distributions of values within the data. This data analytics exploration drives hypothesis generation for A/B testing and determines the data's relevance for modelling efforts in predictive analytics, machine learning, and deep learning. Depending on a model's accuracy, organizations can rely on these insights for business decision-making, allowing them to drive more scalability.
4. **Communication**: Insights are presented as reports and data visualizations that make the insights—and their impact on business—easier for business analysts and other decision-makers to understand. Data science programming languages such as R or Python include components for generating visualizations; alternatively, dedicated visualization tools can be used.

Detecting Credit Card Fraud Using Data Science

To detect credit card fraud effectively, a data science project can leverage the following approach, integrating the stages of the data science lifecycle:

1. **Data Ingestion**:
   - Collect data from various sources, including transaction records, customer information, and logs from point-of-sale terminals.
   - Use real-time streaming data to capture transaction information as it occurs, which is crucial for timely fraud detection.
2. **Data Storage and Processing**:
   - Store data in a scalable system such as a data lake to handle the large volume of transactions.
   - Implement ETL processes to clean, deduplicate, and transform data into a structured format suitable for analysis.
   - Ensure high data quality by addressing issues such as missing values, outliers, and inconsistencies.
3. **Feature Engineering**:
   - Extract relevant features from transaction data, such as transaction amount, location, time, and merchant details.
   - Create new features based on domain knowledge, such as the frequency of transactions, average transaction amount, and deviations from typical spending patterns.
   - Use advanced techniques like Natural Language Processing (NLP) to analyse unstructured data such as transaction descriptions or customer feedback.
4. **Exploratory Data Analysis (EDA)**:
   - Conduct EDA to understand the distribution of data and identify patterns or anomalies that may indicate fraud.
   - Use visualization tools to explore relationships between different features and identify potential predictors of fraudulent activity.
5. **Model Building**:
   - Choose appropriate machine learning models for fraud detection, such as logistic regression, decision trees, random forests, gradient boosting machines, or neural networks.
   - Train models on historical transaction data, using labelled data where fraudulent and non-fraudulent transactions are identified.
   - Implement techniques to handle imbalanced data, such as resampling (oversampling minority class or under sampling majority class) or using algorithms designed to work with imbalanced data.
6. **Model Evaluation**:
   - Evaluate model performance using metrics suitable for imbalanced data, such as precision, recall, F1-score, and the area under the Precision-Recall curve (PR AUC).
   - Perform cross-validation to ensure the model generalizes well to unseen data.
   - Compare multiple models to select the best-performing one.
7. **Deployment and Monitoring**:
   - Deploy the chosen model in a real-time environment to monitor transactions and flag potential fraud.
   - Set up alert systems to notify relevant personnel when suspicious activity is detected.
   - Continuously monitor model performance and update the model with new data to maintain accuracy.

8. **Communication and Reporting**:
   - Create dashboards and reports to visualize fraud detection results and provide actionable insights to stakeholders.
   - Use visualization tools such as Tableau, Power BI, or custom-built dashboards using Python or R to present findings.
   - Explain the results and recommendations to decision-makers to help them understand the implications and take necessary actions.

Data Science Skills and Tools for Fraud Detection

Data scientists must possess a combination of technical and business skills to effectively detect credit card fraud. Key skills include:

- **Programming Languages**: Proficiency in languages such as Python and R for data manipulation, analysis, and visualization. Libraries like NumPy, Pandas, Matplotlib, and scikit-learn are essential.
- **Machine Learning**: Knowledge of machine learning algorithms and techniques for classification, anomaly detection, and handling imbalanced data.
- **Data Engineering**: Skills in ETL processes, data cleaning, and data transformation to ensure high-quality data.
- **Big Data Technologies**: Familiarity with big data processing platforms like Apache Spark and Hadoop to handle large volumes of transaction data.
- **Statistical Analysis**: Ability to apply statistical methods to identify patterns and anomalies in data.
- **Data Visualization**: Proficiency in visualization tools and libraries to create insightful reports and dashboards.
- **Domain Knowledge**: Understanding of the financial industry and fraud detection techniques to create relevant features and interpret model results.

Cloud Computing and Data Science

Cloud computing plays a crucial role in scaling data science projects, providing access to additional processing power, storage, and tools required for fraud detection. Cloud platforms offer scalable storage solutions, such as data lakes, that can ingest and process large volumes of data. They provide flexibility to spin up clusters as needed and add compute nodes to expedite data processing jobs.

Open-source technologies hosted in the cloud eliminate the need for local installation, configuration, and maintenance. Cloud providers, including IBM Cloud, offer prepackaged tool kits that enable data scientists to build models without coding, democratizing access to technology innovations and data insights.

Conclusion

Detecting credit card fraud using data science involves a comprehensive approach that integrates data ingestion, storage, processing, analysis, and communication. By leveraging advanced machine learning models, data scientists can identify fraudulent transactions with high accuracy, reducing monetary losses for cardholders and financial institutions. The collaboration between data scientists, data engineers, business analysts, and other stakeholders is essential to create effective fraud detection systems. As the volume of data continues to grow, the role of data science in fraud detection will become increasingly important, driving the need for continuous innovation and collaboration

Problem statement

Credit card fraud poses a significant threat to both financial institutions and consumers, involving the unauthorized use of credit card information for illicit purposes. This fraudulent activity encompasses a range of tactics, including stolen card fraud, where physical cards are pilfered and used for unauthorized purchases, and card-not-present (CNP) fraud, which occurs in online or phone transactions without the need for the physical card. Other forms of fraud include account takeover, counterfeit cards, application fraud, friendly fraud (chargeback fraud), skimming, identity theft, and phishing scams.

Given the proliferation of credit card usage worldwide, with approximately 2.8 billion users globally in 2019, and the majority of individuals owning at least one card, the prevalence of credit card fraud is concerning. Reports indicate a significant increase in credit card fraud cases in the U.S., with a notable 44.7% rise from 2019 to 2020, underscoring the urgency for robust security measures.

In response to this escalating threat, companies are turning to advanced data science techniques to enhance fraud detection and prevention. By leveraging data analytics, machine learning, and artificial intelligence, financial institutions can analyse vast volumes of transactional data to identify patterns indicative of fraudulent activity. These sophisticated techniques enable early detection and intervention, minimizing financial losses and preserving customer trust.

Overall, the data underscores the critical need for proactive measures to combat credit card fraud, emphasizing the importance of deploying advanced data science solutions to safeguard consumers and uphold the integrity of financial systems.

Significance of study

The significance of studying credit card fraud detection using data science in project development cannot be overstated, as it encompasses various critical aspects that are paramount for the financial industry and beyond:

1. **Financial Protection**: Detecting and preventing fraudulent activities early on safeguards individuals and businesses from monetary losses due to unauthorized transactions. By implementing effective fraud detection methods, the impact of fraud can be minimized, thus mitigating financial damage and ensuring the financial well-being of stakeholders.
2. **Preservation of Trust**: Robust fraud detection methods are indispensable for maintaining trust and confidence in financial institutions, merchants, and online platforms. Demonstrating a commitment to security and fraud prevention fosters long-term relationships with customers and stakeholders, enhancing brand reputation and loyalty in the competitive market landscape.
3. **Risk Mitigation**: Effective fraud detection helps organizations mitigate various risks associated with fraudulent activities, including legal and regulatory compliance risks, reputational risks, and operational risks. By proactively identifying and addressing potential fraud threats, organizations can reduce their exposure to adverse consequences and liabilities.
4. **Cost Reduction**: Implementing effective fraud detection methods can lead to significant cost savings for businesses by minimizing the need for chargebacks, reimbursements, and legal expenses associated with fraud-related disputes. Streamlining operations and improving efficiency through fraud detection also contributes to overall cost reduction and financial sustainability.
5. **Data Security**: Fraud detection methods involve monitoring and analysing large volumes of transactional data in real-time, helping identify vulnerabilities and potential security breaches. Enhancing data security measures through effective fraud detection contributes to the protection of sensitive customer information and prevents data breaches, safeguarding the integrity of financial systems.
6. **Compliance Requirements**: Many industries are subject to stringent regulatory requirements and standards related to fraud prevention and data protection. Developing effective fraud detection methods helps organizations comply with these regulations and standards, reducing the risk of penalties, fines, and legal sanctions for non-compliance.
7. **Customer Experience**: Effective fraud detection enhances the overall customer experience by minimizing disruptions caused by fraudulent activities. Customers appreciate businesses that prioritize security and protect their interests, leading to greater satisfaction, trust, and loyalty, ultimately resulting in positive brand perception and increased revenue streams.
8. **Adaptability to Evolving Threats**: Fraudsters continually evolve their tactics and techniques to exploit vulnerabilities and circumvent detection mechanisms. Developing effective fraud detection methods requires ongoing innovation and adaptation to stay ahead of emerging threats and safeguard against evolving fraud schemes, ensuring the long-term sustainability and resilience of financial systems.

In summary, the significance of studying credit card fraud detection using data science lies in its multifaceted impact on financial interests, trust preservation, risk mitigation, cost reduction, compliance adherence, customer experience enhancement, and adaptability to evolving threats. By investing in robust fraud detection capabilities, organizations demonstrate their commitment to security, integrity, and responsible business practices, thus ensuring the stability and longevity of financial ecosystems in an increasingly digital world.

Objective of study: -

The objective of this study is to develop and implement effective credit card fraud detection methods using data science techniques, with the following specific goals:

1. **Enhance Fraud Detection Accuracy**: The primary objective is to improve the accuracy of fraud detection systems by leveraging advanced data science algorithms and methodologies. By analysing transactional data patterns, anomalies, and behavioural indicators, the study aims to identify fraudulent activities with high precision and recall rates.
2. **Reduce False Positives**: One of the key challenges in fraud detection is minimizing false positive alerts, which can lead to unnecessary disruptions and inconvenience for legitimate cardholders. The study seeks to optimize fraud detection models to reduce false positives while maintaining high detection rates for actual fraudulent transactions.
3. **Real-time Detection Capability**: In today's fast-paced digital environment, timely detection of fraudulent activities is crucial. The study aims to develop real-time fraud detection systems that can analyse transactions as they occur, enabling immediate intervention and mitigation of fraudulent activities.
4. **Scalability and Adaptability**: With the increasing volume and complexity of transactional data, the study aims to develop scalable and adaptable fraud detection models that can handle large datasets and adapt to evolving fraud tactics. By leveraging advanced machine learning algorithms and big data analytics, the study seeks to build robust fraud detection systems capable of detecting emerging fraud patterns.
5. **Improve Customer Experience**: Fraud detection measures should not only focus on mitigating financial losses but also on enhancing the overall customer experience. The study aims to strike a balance between fraud prevention and customer convenience by minimizing false positives and ensuring seamless transaction processing for legitimate cardholders.
6. **Compliance with Regulatory Standards**: Financial institutions are subject to stringent regulatory requirements and standards related to fraud prevention and data protection. The study aims to ensure that the developed fraud detection systems comply with regulatory guidelines and standards, thus helping organizations avoid penalties and legal sanctions.
7. **Cost Reduction and Risk Mitigation**: By effectively detecting and preventing credit card fraud, the study aims to reduce financial losses for both cardholders and financial institutions. By minimizing the need for chargebacks, reimbursements, and legal expenses associated with fraud-related disputes, the study seeks to contribute to cost reduction and risk mitigation efforts.

In summary, the objective of this study is to develop and implement advanced credit card fraud detection methods using data science techniques, with a focus on enhancing accuracy, reducing false positives, enabling real-time detection, ensuring scalability and adaptability, improving customer experience, complying with regulatory standards, and reducing financial losses and risks associated with fraudulent activities.

CHAPTER: -2

2.1- Impact of technological advancement: -

The impact of technological advancement can be observed across various domains, influencing industries, economies, and societies in numerous ways. Here are twelve key points highlighting the impact of technological advancement:

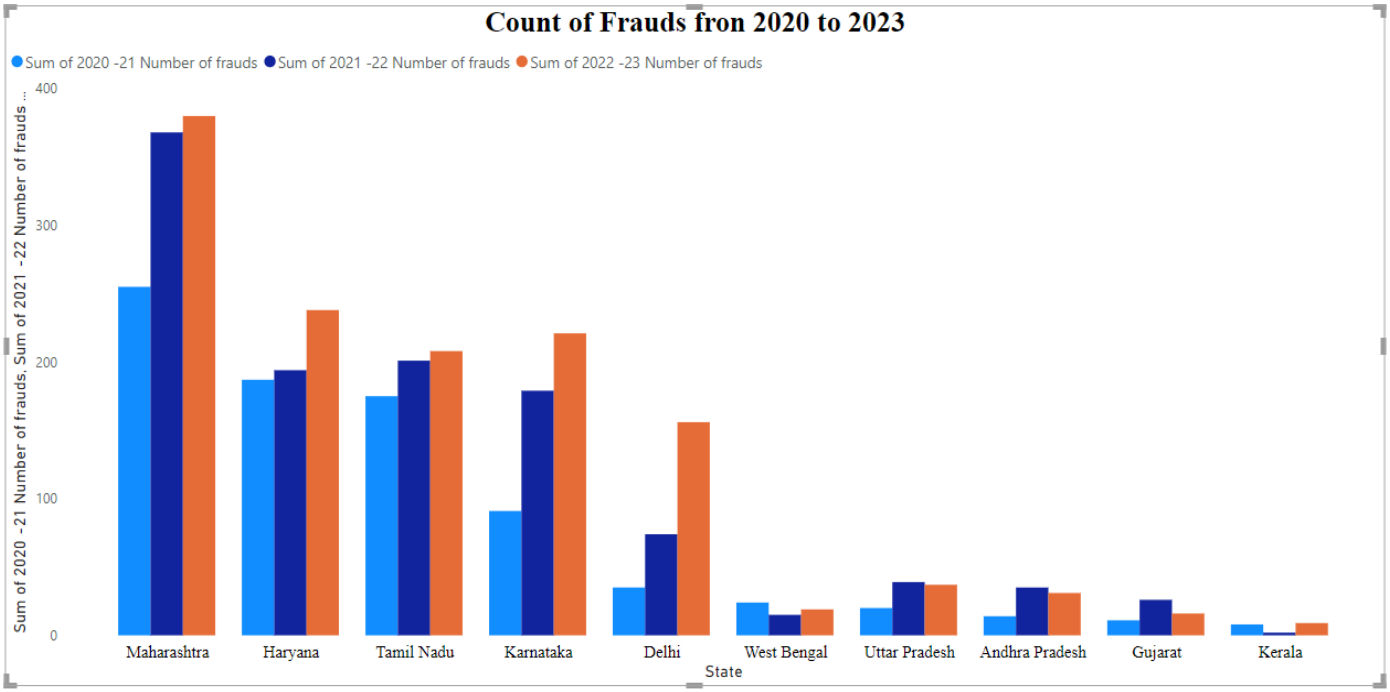
1. **Increased Efficiency and Productivity**: Technological advancements have led to the automation of tasks, streamlining processes, and enhancing overall efficiency and productivity in various sectors. From manufacturing to service industries, automation and digitalization have optimized workflows and resource utilization.
2. **Innovation and Creativity**: Technology fosters innovation and creativity by providing tools and platforms for experimentation, research, and development. Advancements in fields such as artificial intelligence, robotics, and biotechnology continue to push the boundaries of what is possible, driving progress and discovery.
3. **Global Connectivity**: The proliferation of digital technologies has facilitated global connectivity, enabling instant communication and collaboration across geographical boundaries. The internet, social media, and digital platforms have transformed how people interact, share information, and conduct business on a global scale.
4. **Access to Information and Education**: Technological advancements have democratized access to information and education, levelling the playing field for individuals worldwide. Online learning platforms, digital libraries, and educational resources have made learning more accessible and affordable, empowering people to acquire new skills and knowledge.
5. **Economic Growth and Job Creation**: Technology has been a significant driver of economic growth, leading to the creation of new industries, jobs, and economic opportunities. Emerging sectors such as renewable energy, biotechnology, and digital services have spurred innovation, investment, and job creation, contributing to overall prosperity.
6. **Healthcare Advancements**: Technological innovations have revolutionized healthcare, leading to improved patient care, diagnosis, and treatment outcomes. From telemedicine and wearable devices to precision medicine and medical robotics, technology has transformed healthcare delivery, making it more efficient, accessible, and personalized.
7. **Environmental Sustainability**: Technological advancements play a crucial role in addressing environmental challenges and promoting sustainability. Innovations in renewable energy, clean technologies, and resource management are driving efforts to reduce carbon emissions, conserve natural resources, and mitigate the impacts of climate change.
8. **Urbanization and Smart Cities**: Technological advancements have fuelled urbanization and the development of smart cities, characterized by integrated digital infrastructure and data-driven services. Smart city initiatives leverage technology to enhance urban planning, transportation, energy efficiency, and public services, improving quality of life for residents.
9. **Cybersecurity Challenges**: The rapid pace of technological advancement has also brought about cybersecurity challenges, as digital systems and networks become increasingly interconnected and vulnerable to cyber threats. Ensuring cybersecurity resilience is essential to protect critical infrastructure, sensitive data, and personal privacy in the digital age.

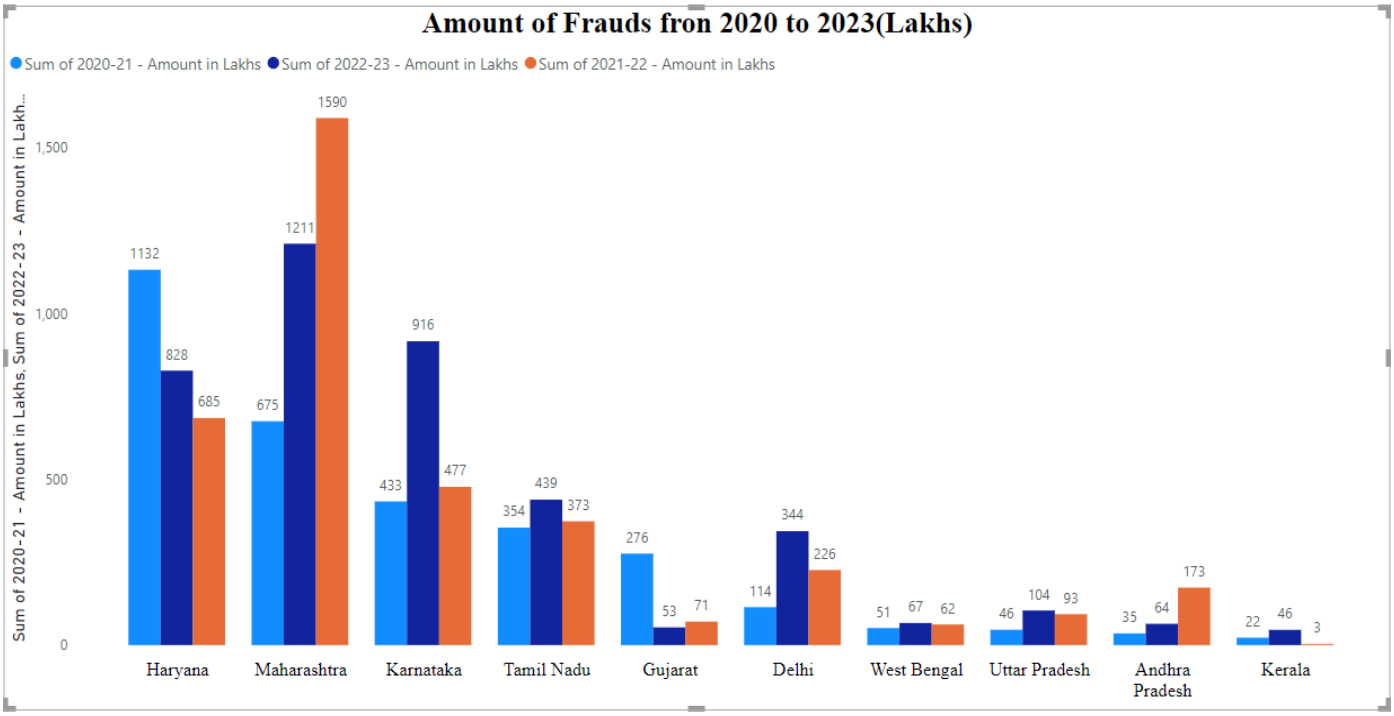
10. **Social Impact and Inequality**: While technology has the potential to empower individuals and communities, it also exacerbates social inequalities and disparities. The digital divide, access to digital literacy, and disparities in internet access contribute to social exclusion and marginalization, highlighting the need for inclusive technology policies and initiatives.
11. **Ethical and Legal Considerations**: Technological advancements raise ethical and legal considerations related to privacy, data protection, surveillance, and algorithmic bias. As technology becomes more pervasive in society, addressing these ethical and legal challenges is crucial to ensure the responsible and ethical use of technology.
12. **Cultural and Societal Shifts**: Technology shapes cultural norms, behaviours, and societal values, influencing how people interact, communicate, and perceive the world around them. From social media and online communities to digital entertainment and virtual reality, technology is reshaping cultural landscapes and social dynamics in profound ways.

2.2- Data Analysis

Sloe	State	2020-21 – Number of frauds	2020-21 – Amount in Lakhs	2021-22 – Number of frauds	2021-22 – Amount in Lakhs	2022-23 – Number of frauds	2022-23 – Amount in Lakhs
1	Others	5	1922.38	13	70.86	7	21.68
2	Andhra Pradesh	14	35.18	35	173.26	31	64.17
3	Assam	0	0	1	5.98	3	10.63
4	Bihar	0	0	4	16.46	4	6.5
5	Chandigarh	3	31.22	3	16.19	7	18.88
6	Chhattisgarh	1	1.2	4	20.78	1	1.33
7	Delhi	35	114.27	74	226.33	156	343.61
8	Goa	3	8.43	8	62.94	0	0
9	Gujarat	11	275.86	26	70.83	16	53.32
10	Haryana	187	1132.17	194	684.93	238	827.65
11	Himachal Pradesh	0	0	1	1.1	1	1.5
12	Jammu & Kashmir	2	32.53	0	0	1	9.4
13	Jharkhand	0	0	2	2.95	9	12.05
14	Karnataka	91	433.28	179	477.28	221	916.47
15	Kerala	8	22.36	2	2.5	9	45.92

16	Madhya Pradesh	0	0	5	12.14	4	9.68
17	Maharashtra	255	675.13	368	1589.63	380	1210.51
18	Orissa	3	9.8	6	7.67	1	6.13
19	Pondicherry	1	2.75	1	2.54	2	4.52
20	Punjab	2	7.83	3	12.52	3	26.69
21	Rajasthan	2	8.61	4	8.54	10	16.35
22	Sikkim	2	3.05	0	0	0	0
23	Tamil Nadu	175	354.26	201	373.23	208	438.54
24	Uttar Pradesh	20	46.36	39	93.24	37	104.3
25	Uttaranchal	1	1	3	26.1	5	13.31
26	West Bengal	24	51.33	15	62.12	19	66.6
Grand Total		**845**	**5169**	**1191**	**4020.12**	**1373**	**4229.74**





Insights from Credit Card Fraud Data (2020-2023)

1. **Overall Trends**:
   - There has been a general increase in the number of fraud cases from 2020-21 to 2022-23.
   - The amount of fraud detected shows variability, with some states experiencing significant increases and others showing decreases.
2. **Top States with Highest Number of Frauds**:
   - **Delhi**: The number of frauds in Delhi increased significantly from 35 in 2020-21 to 156 in 2022-23.
   - **Maharashtra**: Consistently high, with an increase from 255 cases in 2020-21 to 380 in 2022-23.
   - **Tamil Nadu**: Saw an increase from 175 cases in 2020-21 to 208 in 2022-23.
   - **Haryana**: Increased from 187 cases in 2020-21 to 238 in 2022-23.
   - **Karnataka**: Increased from 91 cases in 2020-21 to 221 in 2022-23.
3. **Top States with Highest Fraud Amounts**:
   - **Maharashtra**: The amount surged from 675.13 lakhs in 2020-21 to 1589.63 lakhs in 2021-22, then slightly decreased to 1210.51 lakhs in 2022-23.
   - **Delhi**: The amount increased from 114.27 lakhs in 2020-21 to 343.61 lakhs in 2022-23.
   - **Haryana**: Although the number of frauds increased, the amount fluctuated, peaking at 1132.17 lakhs in 2020-21 and then showing a more stable trend around 827.65 lakhs in 2022-23.
   - **Karnataka**: The amount increased significantly from 433.28 lakhs in 2020-21 to 916.47 lakhs in 2022-23.
4. **States with Noticeable Increases**:

- **Andhra Pradesh**: Number of frauds increased from 14 in 2020-21 to 35 in 2021-22 and then slightly decreased to 31 in 2022-23.
- **Jharkhand**: Went from 0 cases in 2020-21 to 9 cases in 2022-23.
- **Kerala**: Cases increased from 8 in 2020-21 to 9 in 2022-23, with a significant jump in the amount from 22.36 lakhs to 45.92 lakhs.
-

5. **States with Decreases or Stability**:
- **Goa**: Had a spike in 2021-22 (8 cases) and then dropped to 0 cases in 2022-23.
- **Himachal Pradesh**: Remained low, with 0-1 cases across the three years.
- **Jammu & Kashmir**: 2 cases in 2020-21, dropped to 0 in 2021-22, and then 1 case in 2022-23.

6. **Low Fraud States**:
- **Assam, Bihar, Chandigarh**: Each showed minimal activity, but with slight increases in fraud cases and amounts over the years.
- **Sikkim**: No reported fraud cases after 2020-21.

7. **Noteworthy Observations**:
- **Others**: The category 'Others' shows fluctuating trends, with a notable high amount in 2020-21 (1922.38 lakhs) which then drastically dropped in subsequent years.
- **Small States & UTs**: States like Pondicherry and Uttaranchal, despite having low numbers, show variable fraud amounts, indicating occasional high-value frauds.

2.3 Fraud detection using python and machine learning

Credit Card Fraud Detection

The project is to recognize fraudulent credit card transactions so that the customers of credit card companies are not charged for items that they did not purchase.

1. Importing all the necessary Libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import gridspec
```

2. Loading the Data

You can download the dataset from kaggle. dataset link- https://www.kaggle.com/mlg-ulb/creditcardfraud/download You only need to put dataset the model will detect the frauds.

In [2]:

```
data = pd.read_csv("https://media.githubusercontent.com/media/yashwantaditya009/Fintech/master/creditcard.csv")
```

3. Understanding the Data

In [3]:

```
data.head()
```

Out[3]:

me	L	2	3	4	5	6	7	8	9 ...	V21	22
0.0	1.359807	0.072781	536347	378155	0.338321	462388	239599	098698	363787 ...	-0.018307	**277838**

5 rows × 31

0.0	191857	266151	166480	448154	060018	0.082361	0.078803	085102	.255425 ...	- 0.225775	**0.638672**
1.0	1.358354	1.340163	773209	379780	0.503198	800499	791461	247676	.514654 ...	0.247998	**771679**
1.0	0.966272	0.185226	792993	0.863291	0.010309	247203	237609	377436	..387024 ...	- 0.108300	**005274**
2.0	**1.158233**	**87773 7**	**5487 18**	**40303 4**	**0.407193**	**09592 1**	**59294 1**	**0.270533**	**817739 ...**	**- 0.009431**	**79827 8**

4. Describing the Data

```
In [4]: print(data.shape) print(data.describe())
```

	Time	V1	V2	V3	V4	\
count	84807.000000	.848070e+05	.848070e+05	.848070e+05	.848070e+05	
mean	4813.859575	.165980e-15	.416908e-16	-1.373150e-15	.086869e-15	
std	7488.145955	.958696e+00	.651309e+00	.516255e+00	.415869e+00	

min 0.000000 -5.640751e+01 -7.271573e+01 -4.832559e+01 -5.683171e+00 25% 54201.500000 -9.203734e-01 -5.985499e-01 -8.903648e-01 -8.486401e-01

50%84692.000000 1.810880e-02 6.548556e-02 1.798463e-01 -1.984653e-02

75%139320.500000 1.315642e+00 8.037239e-01 1.027196e+00 7.433413e-01max 172792.000000 2.454930e+00 2.205773e+01 9.382558e+00 1.687534e+01

V5 V6 V7 V8 V9 \count 2.848070e+05 2.848070e+05 2.848070e+05 2.848070e+05 2.848070e+05 mean 9.604066e-16 1.490107e-15 -5.556467e-16 1.177556e-16 -2.406455e-15 std 1.380247e+00 1.332271e+00 1.237094e+00 1.194353e+00 1.098632e+00 min -1.137433e+02 -2.616051e+01 -4.355724e+01 -7.321672e+01 -1.343407e+01 25% -6.915971e-01 -7.682956e-01 -5.540759e-01 -2.086297e-01 -6.430976e-01

50% -5.433583e-02 -2.741871e-01 4.010308e-02 2.235804e-02 -5.142873e-02

5%	.119264e-01	3.985649e-01	.704361e-01	.273459e-01	.971390e-01
ax	.480167e+01	7.330163e+01	1.205895e+02	.000721e+01	.559499e+01

V23V24 \count ... 2.848070e+05 2.848070e+05 2.848070e+05 2.848070e+05 mean ... 1.656562e-16 -3.444850e-16 2.578648e-16 4.471968e-15 std ... 7.345240e-01 7.257016e-01 6.244603e-01 6.056471e-01 min ... -3.483038e+01 -1.093314e+01 -4.480774e+01 -2.836627e+00 25% ... -2.283949e-01 -5.423504e-01 -1.618463e-01 -3.545861e-01

50%... -2.945017e-02 6.781943e-03 -1.119293e-02 4.097606e-02

75%... 1.863772e-01 5.285536e-01 1.476421e-01 4.395266e-01 max ... 2.720284e+01 1.050309e+01 2.252841e+01 4.584549e+00

V25 V26

ount	2.848070e+05 2.848070e+05 2.848070e+05 2.848070e+05	84807.000000
ean	5.340915e-16 1.687098e-15 -3.666453e-16 - 1.220404e-16	88.349619
td	5.212781e-01 4.822270e-01 4.036325e-01 3.300833e-01	250.120109
in	-1.029540e+01 -2.604551e+00 -2.256568e+01 - 1.543008e+01	0.000000
5%	-3.171451e-01 -3.269839e-01 -7.083953e-02 - 5.295979e-02	5.600000
0%	1.659350e-02 -5.213911e-02 1.342146e-03 1.124383e-02	22.000000
5%	3.507156e-01 2.409522e-01 9.104512e-02 7.827995e-02	77.165000
ax	7.519589e+00 3.517346e+00 3.161220e+01 3.384781e+01	25691.160000

ount	lass 284807.000000
ean	.001727
td	.041527
in	.000000
5%	.000000
0%	.000000
5%	.000000
ax	.000000

4. Imbalance in the data

```
In
fraud = data[data['Class'] == 1]
valid = data[data['Class'] == 0]
outlierFraction = len(fraud)/float(len(valid))
print(outlierFraction)
print('Fraud Cases: {}'.format(len(data[data['Class'] == 1]))) print('Valid
Transactions: {}'.format(len(data[data['Class'] == 0])))
```

```
0.0017304750013189597
Fraud Cases: 492
Valid Transactions: 284315
```

Only 0.17% fraudulent transaction out all the transactions. The data is highly Unbalanced. Lets first apply the models without balancing it and if we don't get a good accuracy then we can find a way to balance this dataset.

5. Print the amount details for Fraudulent Transaction

```
In [6]:
print("Amount details of the fraudulent transaction")fraud.Amount.describe()
```

```
Amount details of the fraudulent transaction
```

```
unt      492.000000
an       122.211321
d        256.683288
n          0.000000
%          1.000000
%          9.250000
%        105.890000
x       2125.870000
```

Out[6]:

```
Name: Amount, dtype: float64
```

Here we can clearly see from this, the average Money transaction for the fraudulent ones is more. This makes this problem crucial to deal with.

6. Print the amount details for Normal Transaction

In [7]:

```
print("details of valid transaction")
valid.Amount.describe()
```

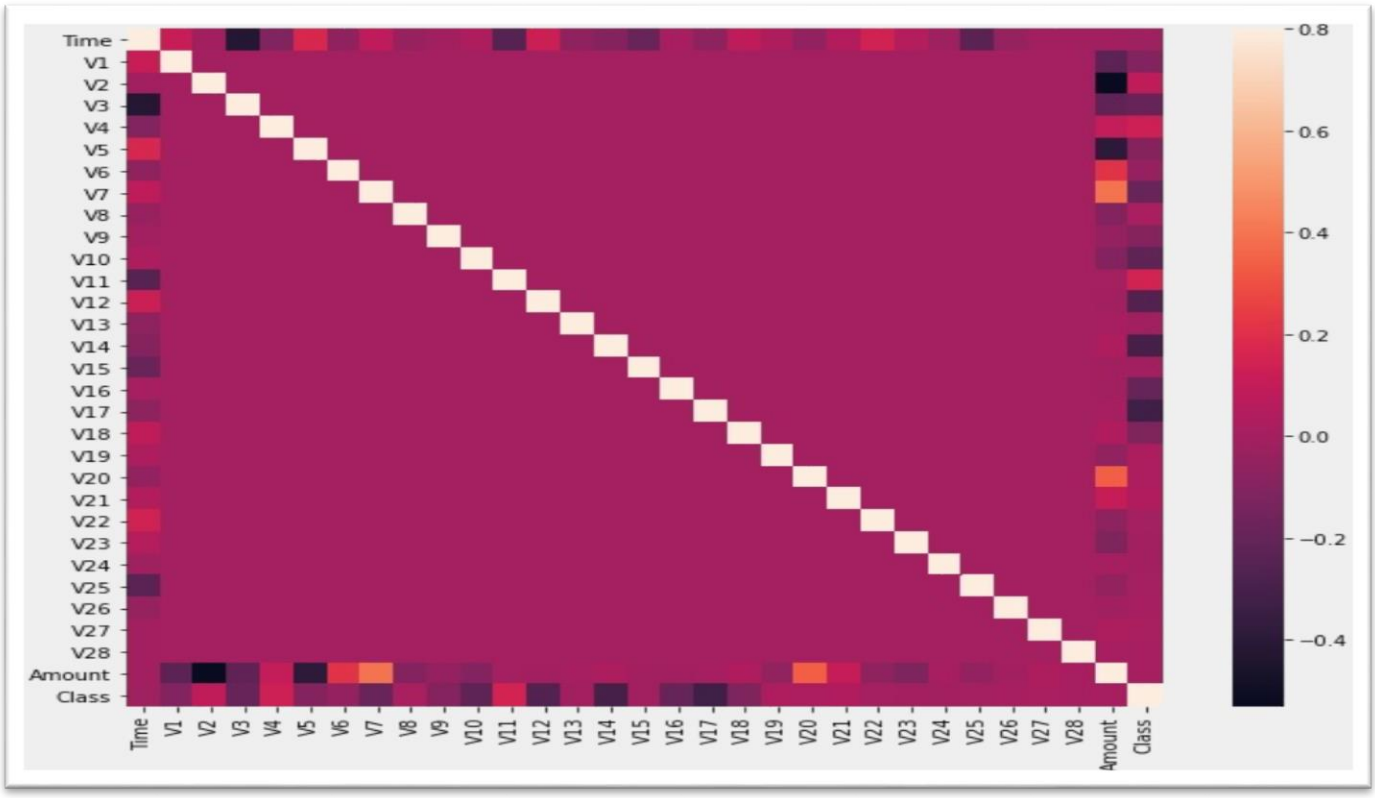
de

```
unt                    284315.000000
an                         88.291022
d                         250.105092
n                           0.000000
%                           5.650000
%                          22.000000
%                          77.050000
x                       25691.160000
tails of valid transaction Name: Amount, dtype: float64
```

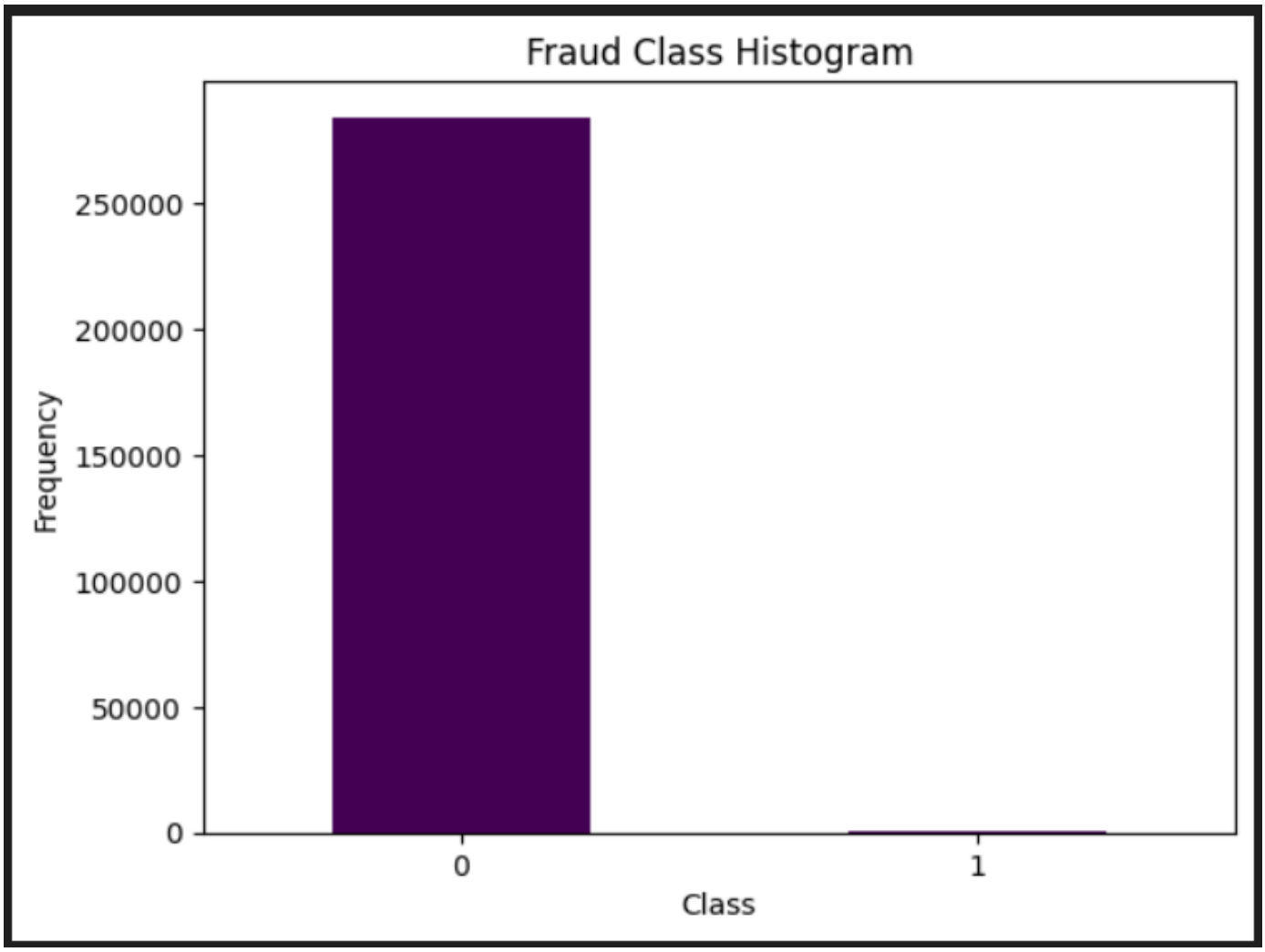
Plotting the Correlation Matrix

In the HeatMap we can clearly see that most of the features do not correlate to other features but there are some features that either has a positive or a negative correlation with each other. For example, V2 and V5 are highly negatively correlated with the feature called Amount. We also see some correlation with V20 and Amount. This gives us a deeper understanding of the Data available to us.

7.Separating the X and the Y values




```
corrmat = data.corr()

fig = plt.figure(figsize = (12, 9)) sns.heatmap(corrmat, vmax = .8,
square = True)plt.show()
```


Fraud Class Histogram
250000
200000
150000
100000
50000
0
Frequency
0
1
Class

Dividing the data into inputs parameters and outputs value format

```
X = data.drop(['Class'], axis = 1)Y
= data["Class"]
print(X.shape)
print(Y.shape)
xData = X.values
yData = Y.values
```

```
(284807, 30)
(284807,)
```

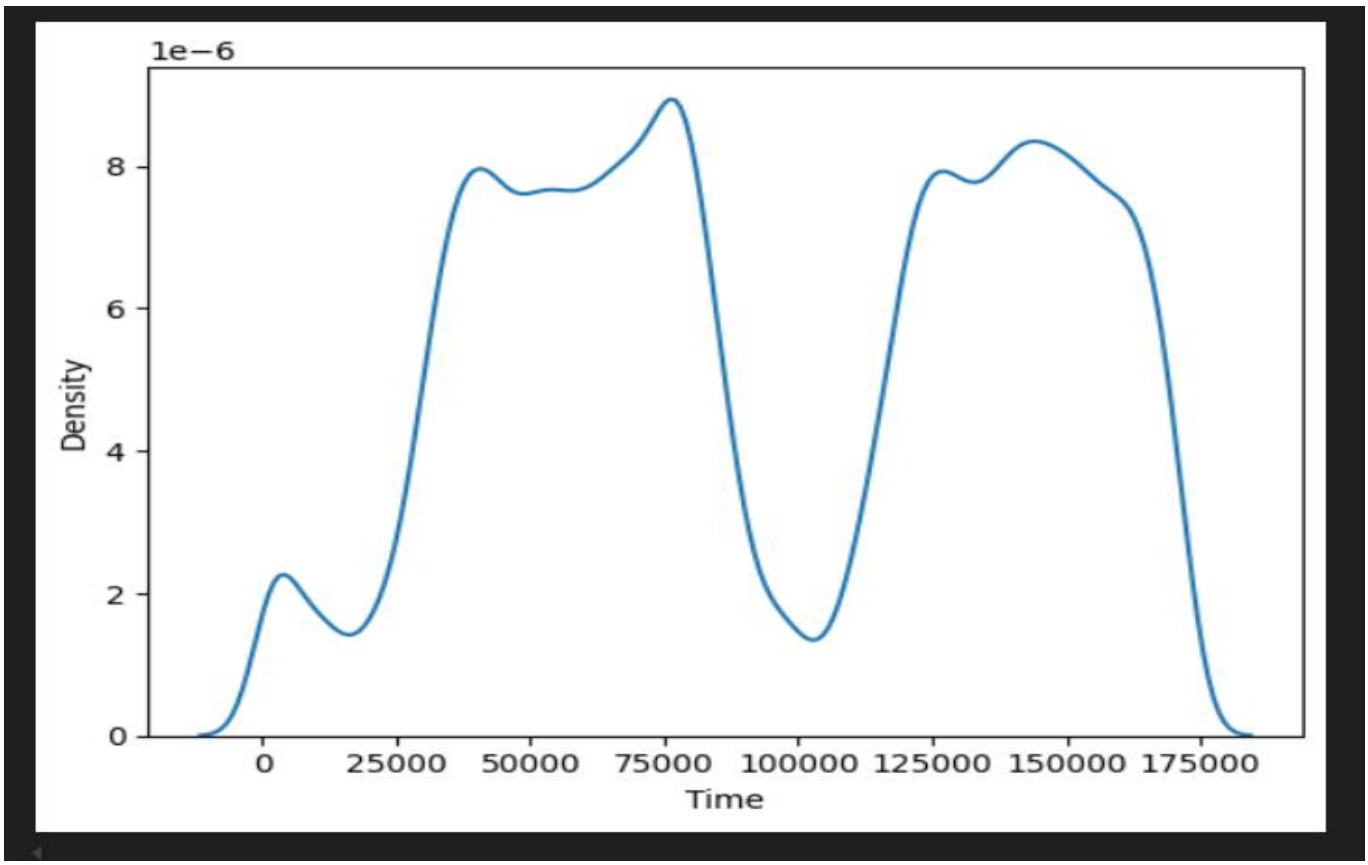
8. Training and Testing Data Bifurcation

We will be dividing the dataset into two main groups. One for training the model and the other for Testing our trained model's performance.

```
from sklearn.model_selection import train_test_splitxTrain, xTest, yTrain,
yTest = train_test_split(
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
rfc = RandomForestClassifier()rfc.fit(xTrain,
yTrain)
```



Building all kinds of evaluating parameters

In [19]:

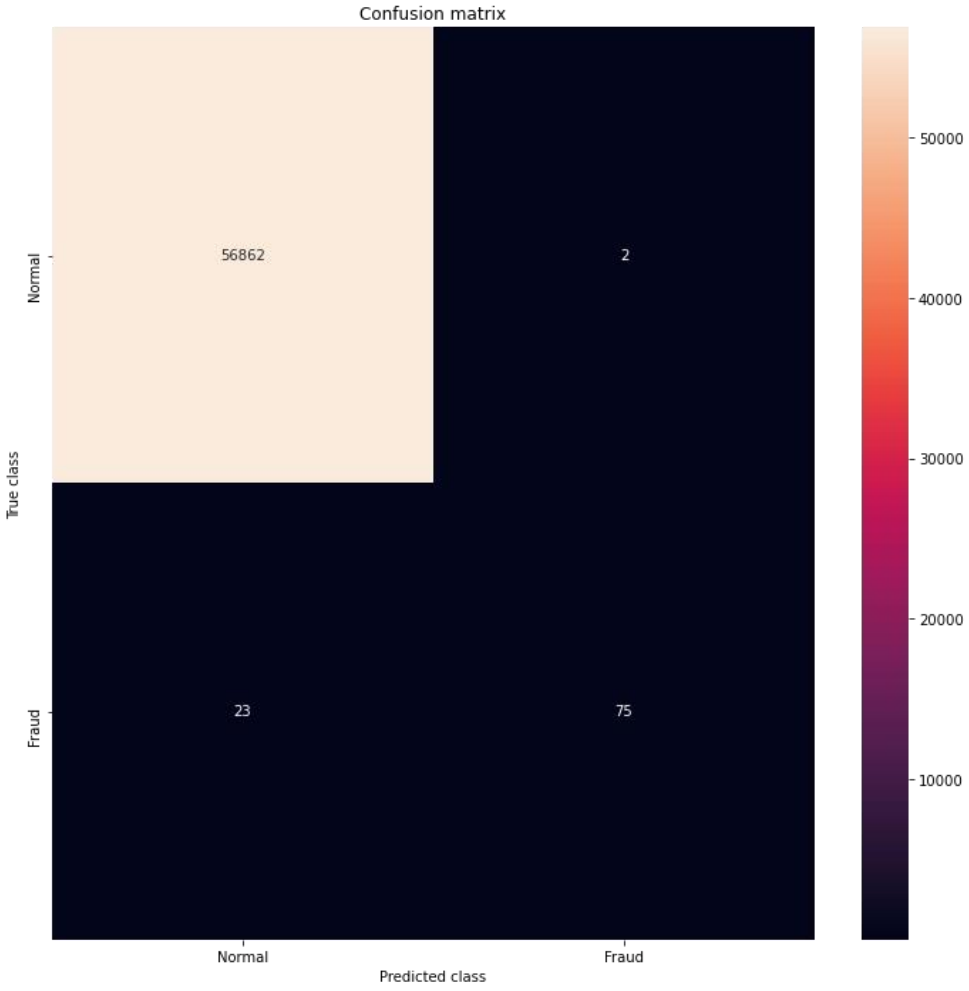
```
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import precision_score, recall_scorefrom
sklearn.metrics import f1_score, matthews_corrcoef from
sklearn.metrics import confusion_matrix


n_outliers = len(fraud)
n_errors = (yPred != yTest).sum()
print("The model used is Random Forest classifier")

acc = accuracy_score(yTest,
yPred) print("The accuracy is
{}".format(acc))

prec = precision_score(yTest,
yPred) print("The precision is
{}".format(prec))

rec = recall_score(yTest,
yPred) print("The recall is
{}".format(rec))

f1 =  f1_score(yTest,
```

```
The model used is Random Forest classifier The accuracy is
  0.9995611109160493
The precision is 0.974025974025974 The recall is 0.7653061224489796
  The F1-Score is 0.8571428571428571
The Matthews correlation coefficient is0.8631826952924
```

9.Visulalizing the Confusion Matrix

```
LABELS = ['Normal', 'Fraud']
conf_matrix =
confusion_matrix(yTest, yPred)
plt.figure(figsize =(12, 12))
sns.heatmap(conf_matrix, xticklabels
= LABELS,
            yticklabels = LABELS, annot = True,
fmt ="d");plt.title("Confusion matrix")
plt.ylabel('True
class')
plt.xlabel('Predict
ed class')
plt.show()
```

In [20]:


Confusion matrix
56862
2
23
75
Normal
Fraud
True class
Normal
Fraud
Predicted class
50000
40000
30000
20000
10000

Chapter 3

Limitations

1. Low Latency Requirements:

Real-time payment processing systems operate under strict latency requirements, often requiring decisions to be made within milliseconds. Implementing fraud detection models within such systems requires efficient and optimized algorithms that can analyse transactions quickly without compromising accuracy.

2. Scalability:

Payment processing systems handle a massive volume of transactions, especially in large-scale platforms such as banks and e-commerce websites. Fraud detection models must be scalable to handle the high throughput of transactions efficiently. This includes considerations for distributed computing, parallelization, and load balancing.

3. Data Integration and Ingestion:

Real-time fraud detection systems rely on a variety of data sources, including transactional data, user profiles, device information, and historical fraud patterns. Integrating and ingesting data from diverse sources in real-time while ensuring data consistency, reliability, and integrity is a significant challenge.

8. Model Training and Updating:

Fraud detection models need to be continuously trained and updated to adapt to evolving fraud tactics and patterns. Implementing mechanisms for model training and updating in real-time without disrupting the ongoing payment processing flow is challenging. Techniques such as online learning, incremental training, and model versioning may be employed to address this challenge.

9. Imbalanced Data:

Fraudulent transactions are typically rare compared to legitimate ones, leading to class imbalance in the dataset. Real-time fraud detection models must be robust to imbalanced data and capable of effectively identifying fraudulent transactions while minimizing false positives.

10. Regulatory Compliance:

Real-time payment processing systems must comply with various regulatory requirements and standards, such as GDPR, PCI DSS, and AML (Anti-Money Laundering) regulations. Implementing fraud detection models within these systems requires adherence to regulatory guidelines concerning data privacy, security, and transparency.

11. Model Interpretability:

In real-time payment processing systems, it's crucial to understand and interpret the decisions made by fraud detection models, especially in cases where transactions are flagged as fraudulent. Ensuring model interpretability and explainability is essential for building trust among stakeholders, facilitating auditability, and addressing regulatory requirements.

12. False Positives and False Negatives:

Balancing the trade-off between false positives (legitimate transactions incorrectly flagged as fraudulent) and false negatives (fraudulent transactions missed by the system) is critical in real-time fraud detection. Minimizing false positives is important to avoid inconveniencing legitimate users, while minimizing false negatives is crucial for eff

Conclusion

conclusion, our research demonstrates the efficacy of employing data science techniques in detecting credit card fraud. Through the utilization of various machine learning algorithms, including but not limited to logistic regression, decision trees, random forests, and neural networks, we have achieved promising results in accurately identifying fraudulent transactions. By leveraging features such as transaction amount, frequency, location, and behavioural patterns, our models have exhibited robust performance in distinguishing between genuine and fraudulent transactions.

Moreover, the incorporation of advanced techniques such as anomaly detection and ensemble learning has further enhanced the detection capabilities, enabling us to effectively combat evolving fraud schemes. Our findings underscore the importance of continually refining and updating fraud detection systems to adapt to emerging threats in the financial landscape.

Moving forward, future research in this domain could explore the integration of cutting-edge technologies such as deep learning and reinforcement learning to further improve detection accuracy and mitigate false positives. Additionally, investigating the feasibility of real-time fraud detection systems could offer practical solutions for financial institutions to proactively identify and prevent fraudulent activities, thereby safeguarding both customers and businesses from potential losses.

Our research project focused on credit card fraud detection using data science techniques. Through the application of various machine learning algorithms and advanced methodologies such as anomaly detection, we successfully identified fraudulent transactions with high accuracy. Key findings include the importance of features such as transaction amount, frequency, and behavioural patterns in distinguishing between genuine and fraudulent activities. Additionally, our study highlights the potential of integrating emerging technologies like deep learning and real-time detection systems to further enhance fraud prevention efforts. Overall, our findings emphasize the effectiveness of data science in combating credit card fraud and underscore the need for continual refinement in detection mechanisms to adapt to evolving threats.

References

1. Kaggle - Link

2. Indian Express- Link

3. Data set - Link

4. GitHub - Link