12-2023

IMPROVING CREDIT CARD FRAUD DETECTION USING TRANSFER LEARNING AND DATA RESAMPLING TECHNIQUES

Charmaine Eunice Mena Vinarta
California State University - San Bernardino

IMPROVING CREDIT CARD FRAUD DETECTION USING TRANSFER LEARNING AND DATA RESAMPLING TECHNIQUES

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

by

Charmaine Eunice Mena Vinarta

December 2023

IMPROVING CREDIT CARD FRAUD DETECTION USING TRANSFER LEARNING AND DATA RESAMPLING TECHNIQUES

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Charmaine Eunice Mena Vinarta

December 2023

Approved by:

Dr. Conrad Shayo, Committee Chair

Dr. Barbara Sirotnik, Committee Member

Dr. Conrad Shayo, Department Chair, Information and Decision Sciences

ABSTRACT

This Culminating Experience Project explores the use of machine learning algorithms to detect credit card fraud. The research questions are: Q1. What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud, and how do these techniques compare in terms of fraud detection accuracy and efficiency? Q2. To what extent do synthetic data generation methods effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance? Q3. To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge? The main findings are: Q1. Unconventional cross-domain methods improved fraud detection, holding promise for enhanced security. Q2. The problems caused by unbalanced datasets in credit card fraud detection were effectively addressed by the synthetic data generation techniques SMOTE and ADASYN, resulting in a more balanced dataset suitable for fraud classification. Q3. The combination of neural networks and data resampling techniques, such as SMOTE and ADASYN, significantly improved credit card fraud detection accuracy. The main conclusions are: Q1. Cross-domain methods are useful for credit card fraud detection, especially when it comes to online transactions. Q2. When used with various classifiers, neural networks show

remarkable accuracy rates: 97% for unbalanced data, 99.47% for SMOTE, and 99.11% for ADASYN Q3. A fraud recall of 0.99 is obtained by the model evaluation on imbalanced data, with 12,155 right predictions out of 12,336 and 181 incorrect ones. The identified areas for further study encompass the testing of our model on larger datasets and the optimization of hyperparameters for further enhancement.

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to Dr. Conrad Shayo, who served as the Committee Chair, for their invaluable guidance and support throughout this Culminating Experience Project. Dr. Barbara Sirotnik, as the Committee Member and Reader, who provided valuable insights and feedback that significantly contributed to the quality of this research project.

DEDICATION

To my dear Furukawa Family, Quinto Family, Vinarta Family, and cherished Friends, with a special appreciation to IA. Your constant presence in my life has been a source of unwavering strength and unending inspiration. This accomplishment is a collective victory. Thank you for pushing me through the hard times and being a constant source of encouragement. Thank you for standing by me and always believing in me.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENTS v

LIST OF FIGURES vii

CHAPTER ONE: INTRODUCTION 1

Introduction 1

Background Research 2

Problem Statement 6

Objectives 7

Research Questions 7

Organization of the Project 8

CHAPTER TWO: LITERATURE REVIEW 9

CHAPTER THREE: METHODOLOGY 19

Methods 20

Cybersecurity 20

Healthcare 20

E-commerce 20

Class Imbalance 21

Synthetic Data Generation Techniques 21

SMOTE 21

ADASYN 21

Neural Networks 22

Over-sampling and Under-sampling 23

CHAPTER FOUR: DATA COLLECTION, ANALYSIS AND FINDINGS 24
Data Collection 24
Analysis and Findings 26
CHAPTER FIVE: DISCUSSION, CONCLUSION AND FUTURE WORK 37
Discussion 37
Conclusion 37
Areas for Further Studies 41
APPENDIX A SUBSET OF THE DATASET 42
APPENDIX B CODE 44
REFERENCES 53

LIST OF FIGURES

Figure 1. Features in the Dataset (Narayanan, 2022) 25

Figure 2. Parameter Prevalence in Fraud vs Non-Fraud Transactions..... 25

Figure 3. Based on Median Purchase Price 27

Figure 4. Number of Transaction and Ratio to Median Price 28

Figure 5. Distance from Home and from Last Transaction 29

Figure 6. Distribution of Last Transaction Distance and Purchase Price 29

Figure 7. Relationship Between Home Distance and Purchase Price 30

Figure 8. The Distribution of Fraud and Non-Fraud Transactions..... 31

Figure 9. Class Which Gives Fraud And Not Fraud After Applying SMOTE..... 32

Figure 10. Class which gives Fraud and Not Fraud after applying ADASYN..... 33

Figure 11. Imbalance Data Model Performance 34

Figure 12. SMOTE Data Model Performance 35

Figure 13. ADASYN Data Model Performance 36

CHAPTER ONE

INTRODUCTION

Introduction

The financial landscape of the 21st century has experienced a seismic shift due to the introduction and integration of digital technologies, particularly internet transactions and banking (Kraus et al., 2021). This digital revolution has fostered an environment where consumers now value ease and convenience in their financial transactions (Khando et al., 2022). Banks and other financial institutions have kept pace with this change, introducing tools and instruments that cater to this new-age consumer demand (Debela, 2020). One of the most significant tools within this digital revolution has been the credit card, which has transformed the way consumers digitally shop and transact (Jain et al., 2021). However, with the conveniences offered by the credit card, there comes a host of challenges, most notably the threat of fraud (Liu et al., 2010).

In this era of digital transformation, a persistent and pervasive threat looms large: fraud. Malicious actors adeptly adapt to digital channels, exploiting vulnerabilities through tactics such as identity theft and phishing scams, posing risks not only to individuals but also to financial institutions (Liu et al., 2010). Recognizing the multifaceted nature of this threat and understanding the crucial role of robust fraud detection and prevention mechanisms are imperative for safeguarding the integrity and security of users of contemporary financial

systems. As consumers increasingly embrace these digital channels, it becomes evident that safeguarding financial transactions from fraud is not merely a challenge but a necessity (Khando et al., 2022). This Culminating Experience Project focuses on the evolving landscape of fraud detection methods and explores innovative approaches to address this ever-persistent threat (Gurney & Varol, 2022).

This culminating project aims to extend and build upon the suggestions from systematic literature review conducted by Elhusseny et al., (2022) that uses several machine learning methods, such as Support Vector Machine (SVM), Random Forest, Decision Tree, Naïve Bayes, and XGBoost. Elhusseny et al., (2022) study highlighted the potential for future works such as deep learning methods on machine learning credit card fraud. Building upon their recommendations into consideration, I intend to build a model using deep learning methods to increase project efficiency with the use of a different dataset from the Kaggle repository.

Background Research

Electronic payment systems, primarily led by credit cards, have significantly transformed the consumer experience, thus offering unparalleled ease when purchasing products and services (Ngai & Wat, 2002). In the first quarter of 2023 alone, an astonishing 1.1 billion Visa and Mastercard credit cards were issued worldwide (Statista, 2023). Furthermore, the overall number of

purchases or transactions on several major credit card offering companies, such as Visa, Mastercard, Amercian, Express, Discover, UnionPay, and JCB, have consistently increased from 2014 to 2022, reaching 625 billion transactions in 2022 (Statista, 2023). However, this surge in digital transactions has magnified the challenges of credit card fraud (Hilal et al., 2022). Malicious entities and fraudsters exploit the system, often using sophisticated social engineering techniques to deceive unsuspecting users (Hilal et al., 2022).

To address these challenges, traditional fraud detection methods are giving way to advanced computational solutions, particularly transfer learning, a technique that leverages pre-existing models from one domain and applies them to another domain, thus streamlining the detection process. A notable study by Ali et al (2022) exemplifies this approach, demonstrating that models trained on general financial transactions can be effectively repurposed for credit card fraud detection. Their findings were especially promising when these models were adapted using data from closely related financial domains, underscoring the potential of transfer learning in enhancing future fraud detection mechanisms.

Researchers have explored an intriguing approach to enhance credit card fraud detection by repurposing models from different domains, showing promise in bolstering the effectiveness of fraud detection mechanisms (Al-Hashedi & Magalingam, 2021). This strategy involves adapting models originally designed for general financial transactions to suit the specific needs of credit card fraud detection, as demonstrated by Ali et al. (2022). Furthermore, Zhou et al. (2018)

highlight the value of using pre-trained models from broader applications and fine-tuning them using specialized datasets for credit card transactions. This approach optimizes the utilization of pre-existing models while enabling them to excel at identifying suspicious activities. These cross-domain adaptations align with the evolving landscape of credit card fraud detection, emphasizing the importance of tapping into knowledge from various sources (Al-Hashedi & Magalingam, 2021). By harnessing expertise from diverse domains, these repurposed models make significant contributions to the battle against the ongoing credit card fraud pandemic.

Recent studies have explored the confluence of anomaly detection and transfer learning to combat credit card fraud. A particularly compelling study by Zhou et al. (2018) emphasized the synergistic effect of these techniques. Their research demonstrated the power of using pre-trained models, originally developed for broader applications, and fine-tuning them using niche datasets tailored to credit card transactions. This approach not only capitalizes on the extensive learning of the pre-existing models but also allows for specialization using a comparatively smaller and more focused dataset, enhancing the precision in spotting suspicious activities.

Building on the momentum of integrating advanced techniques for fraud detection, Al-Hashedi & Magalingam (2021) embarked on an exhaustive review of transfer learning methodologies tailored to credit card fraud scenarios. Their analysis shed light on the potential of cross-domain adaptation, a nuanced facet

of transfer learning. Specifically, they pointed out its efficacy in scenarios where datasets are predominantly unlabeled or have scant labelled entries. This insight underscores the importance of domain adaptation in leveraging available data more effectively, bridging the gap between vast unlabeled data sources and the critical need for precise fraud detection.

This Culminating Experience Project utilizes transfer learning methods to bridge the gap between extensive data resources and the specific demands of credit card fraud detection. By harnessing knowledge from various domains and fine-tuning it for credit card transactions, the project aims to bolster the effectiveness of fraud detection models. The incorporation of pre-trained models, as demonstrated in prior research (Ali et al., 2022; Zhou et al., 2018), provides a solid foundation for enhancing the precision and efficiency of fraud detection. Moreover, the project endeavors to pioneer innovative strategies for addressing imbalanced datasets, a persistent challenge in fraud detection. By combining transfer learning with cutting-edge data resampling techniques, it strives to create a comprehensive approach that elevates the accuracy and real-time capabilities of fraud detection systems. Through these endeavors, the project seeks to make substantial contributions to the ongoing battle against credit card fraud.

In the continuously evolving landscape of credit card fraud, staying one step ahead is crucial. Recognizing this, a groundbreaking collaboration between prominent financial institutions and leading tech companies, spearheaded by Zhou et al. (2018), highlighted the indispensability of using continuous learning in

fraud detection models. As malefactors refine and reinvent their deceptive techniques, it's imperative that existing detection models remain adaptive and forward-looking.

Researchers, as demonstrated by Ali et al. (2022), have introduced innovative hybrid models that combine the robustness of transfer learning with the dynamic nature of active learning to address evolving challenges in credit card fraud detection. This approach involves repurposing models from diverse domains and fine-tuning them for credit card transactions, as highlighted by Zhou et al. (2018), resulting in enhanced precision and relevance in the fight against credit card fraud. Nevertheless, the ongoing issue of credit card fraud underscores the need for efficient solutions, as explored further in the problem statement.

Problem Statement

While credit cards offer numerous advantages to consumers, they have been plagued by pressing issues, especially in the realm of security and fraud (Hilal et al., 2022). Recent research articles by Al-Hashedi and Magalingam (2021), and Boutaher et al., (2021), emphasize the need for further study in improving credit card fraud detection mechanisms. Al-Hashedi and Magalingam (2021) comprehensive review of financial fraud detection from 2009 to 2019 points out the importance of applying data mining techniques effectively. Boutaher et al., (2021) analysis of the utilization of machine learning techniques

to detect credit card fraud highlights the challenges posed by imbalanced data and suggests exploring synthetic data generation methods (e.g. SMOTE or ADASYN). These challenges have culminated in substantial financial losses, underscoring the critical need for an efficient solution capable of accurately detecting and preventing such fraudulent activities (Boutaher et al., 2021).

Objectives

Our main goal is to improve credit card fraud detection systems' efficiency by utilizing deep neural networks capabilities. We will concentrate on using sophisticated sampling and clustering techniques to address the problem of highly imbalanced datasets, building on the findings from the study by Dang et al. (2021). In order to optimize the project's efficiency, we are also going to implement Elhusseny et al.'s (2022) recommendation—which prioritizes deep learning techniques into practice. Our aim is to make a valuable contribution to the credit card fraud detection field by combining these two methods and creating a strong and effective model that performs better than conventional machine learning techniques and offers successful fraud detection on imbalanced datasets.

Research Questions

1. What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud,

and how do these techniques compare in terms of fraud detection accuracy and efficiency?

2. To what extent do synthetic data generation methods effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance?
3. To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge?

Organization of the Project

This Culminating Experience Project is organized as follows:

- The first chapter covers the Introduction, Background Research, Problem Statement, Objectives, and Research Questions.
- The second chapter covers the Literature Review.
- The third chapter covers the Methodology.
- The fourth chapter covers the Data Collection, Analysis, and Findings.
- The fifth chapter covers the Discussion, Conclusion, and Areas for Further Studies.

CHAPTER TWO

LITERATURE REVIEW

Research Question 1. What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud, and how do these techniques compare in terms of fraud detection accuracy and efficiency?

Financial institutions face a serious threat from credit card criminals' growing expertise (Nguyen et al., 2020). Effective fraud detection requires robust classifiers that can predict fraud accurately while minimizing false positives. Features count, transaction volume, and feature correlations are examples of input data parameters that have a big influence on how well machine learning algorithms work in different instances. Deep learning approaches, such CNN and LSTM, provide better results in credit card fraud detection than traditional algorithms. These techniques were first developed for image processing and natural language processing. Notably, LSTM obtains an outstanding F1-Score of 84.85%. Sampling strategies are examined as a potential remedy for class imbalance, as they enhance performance on accessible data while significantly diminishing it on non-available data. Interestingly, efficiency improves with increasing class imbalance on unknown data. This work demonstrates the potential of deep learning in the dynamic field of credit card fraud detection.

Abakarim et al. (2018) presented a deep learning-based real-time model for credit card fraud detection. The benchmark trials demonstrated that the Deep Neural Network with Auto-encoder generated very promising results, especially in terms of the F1 score, through tests and a comparative analysis of several real-time binary classifiers and a Deep Neural Network with Auto-encoder. This result showed that deep learning performed better than logistic regression, indicating that more study in this area should concentrate on sophisticated deep learning methods for problems involving the classification of data in real time. The suggested structure might help credit card companies keep an eye out for odd activity and spot possible fraud efforts.

Research by Bolton & Hand (2002) laid the groundwork for using transfer learning in fraud detection. They emphasized how the dynamic nature of fraudsters' strategies presents problems for conventional algorithms. Their approach significantly increased the ability to detect fraud by utilizing transfer learning, even when there was a not a lot of labeled data. A study by Alruqi and Alzahrani (2023) employed transfer learning techniques to fine-tune pre-existing models, specifically sourced from e-commerce industries. Their research indicated an increase in detection rates by utilizing knowledge from how a computer program called a “chatbots” mimics human speech.

A study by Sayed et al. (2022) examined the potential difficulties of transfer learning when there are substantial domain disparities. They proposed a novel approach to weigh domain similarities, improving the model's fine-tuning

process for credit card fraud detection. Zhao et al. (2023), delved into the integration of deep neural networks with transfer learning for fraud detection. Their research highlighted the potential of integrating two potent machine learning techniques, particularly in the presence of sizable unlabeled datasets.

Maschler et al. (2021), presented a method that brought transfer learning and ongoing learning together. Their methodology allowed models to adjust over time, enhancing detection rates and lowering false positives because they were aware of the fraudsters' constantly changing strategies. The difficulties in using transfer learning for fraud detection were highlighted by Li et al. (2022). They emphasized the value of domain knowledge in the selection of appropriate source models and expressed worries about data protection during transfer.

A meta-analysis conducted by Lucas and Jurgovsky (2020) evaluated the effectiveness of transfer learning models in detecting credit card fraud, highlighting the importance of using suitable assessment criteria. They investigated these models using a range of criteria, tackling the problems associated with unequal class distributions. Accurately detecting fraud while reducing false alarms was greatly aided by the metrics chosen, especially recall and precision. The study emphasizes the significance of transfer learning models and careful measure selection in raising the efficacy and robustness of credit card fraud detection.

Given the work by Bolton & Hand (2002) and by Alruqi and Alzahrani (2023) the application of knowledge or techniques from one field or industry in

this case, e-commerce and chatbots to another domain, which is financial security and fraud detection. Cross-domain techniques involve transferring insights, methods, or models from one area of expertise to another to enhance the effectiveness of fraud detection., additionally, replicating the study conducted by Abakarim et al. (2018) with the objective of enhancing accuracy through parameter fine-tuning, this culminating Experience Project will adopt cross-domain techniques, particularly harnessing the prowess of neural network models and pattern recognition methods derived from fields such as healthcare and e-commerce. Drawing inspiration from the anomaly detection used in cybersecurity, we intend to repurpose these strategies to pinpoint unusual credit card transactions. Similarly, the pattern recognition techniques that have been pivotal in identifying abnormalities in medical images will be reconfigured to recognize suspicious transaction patterns, rather than medical anomalies.

Furthermore, behavioral analytics, which have been paramount in e-commerce to trace and analyze user interactions, will be co-opted into our methodology. This will aid in establishing a normative spending pattern for each cardholder, subsequently flagging transactions that diverge from this established pattern.

Research Question 2. To what extent do synthetic data generation methods effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance?

Data imbalance remains a pivotal challenge in the realm of fraud detection. It imposes significant impediments in distinguishing fraudulent transactions from their legitimate counterparts, thereby necessitating advanced techniques to rectify this imbalance.

The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Dang et al. (2021), emerged as a commendable countermeasure to this pervasive issue. Primarily, SMOTE aims to bolster the performance of classifiers by generating artificial minority class samples. While the underlying ambition is to balance data discrepancies, these synthetic samples intriguingly differ from genuine data. It's pivotal to note the research of Kim & Mustapoevich (2023) in this context. They underscored SMOTE's prowess in fortifying classifier generalization and curbing overfitting problems. Additionally, a comparative study by Nishat et al. (2022) with numerous classifiers illuminated both the advantages and pitfalls of SMOTE. The synthesis of SMOTE with transfer learning, as propounded by Liu et al. (2022), further augments model accuracy, especially in scenarios plagued by pronounced class imbalances.

In addition, the Adaptive Synthetic Sampling (ADASYN) technique has garnered accolades for its inherent flexibility and superior recall rates. Anowar & Sadaoui's (2020) exposition illuminates its adaptive nature, which hinges on the density distribution of minority class samples. Further reinforcing its credibility, Belle et al. (2022) unveiled significant findings from an exhaustive study, pointing towards ADASYN's pronounced efficacy in mitigating false negatives. However,

the research by Moloth et al. (2023) drew attention to its processing demands, which could potentially hamstring its deployment in real-time scenarios. An intriguing combination of ADASYN with transfer learning models was presented by Du et al. (2023), which accentuated the synergistic benefits of balanced synthetic data coupled with extraneous domain knowledge.

Venturing into a comparative domain, Thammasiri et al. (2014) embarked on an exploration of both SMOTE and ADASYN. Their revelations were twofold: while both methodologies conspicuously augmented recollection rates, they weren't devoid of precision trade-offs. Sung & Kim's (2023) innovative approach combined transfer learning with cluster-based sampling, thereby fostering a more balanced and representative dataset. This initiative, in turn, led to a marked escalation in classification accuracy.

On the frontier of dynamic resampling systems, Zhang et al. (2019) unfurled an innovative framework. Their system dynamically recalibrates the resampling rate, adapting in tandem with real-time transactional data. This elasticity ensures that the model remains perpetually attuned to evolving fraud tendencies, thereby catalyzing enhanced accuracy in classification.

In conclusion, the landscape of fraud detection is perpetually evolving, with the gravitas of data imbalance demanding continual technological interventions. Techniques like SMOTE and ADASYN, especially when seamlessly integrated with avant-garde strategies like transfer learning, signify a beacon of promise in escalating the efficacy of fraud detection mechanisms.

Research Question 3. To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge?

Hu et al. (2022) conducted a study on the Ischemic Heart Disease (IHD), a predominant global health concern, stands as the foremost cause of death worldwide. In the quest for improved diagnostic tools, the Magnetocardiogram (MCG) has emerged as a prominent non-invasive technique to detect heart anomalies. Despite its significance, the adoption of the MCG technique in regular clinical settings remains limited. A prominent challenge is the scarcity of comprehensive MCG data and the lack of professionals adept at interpreting it, particularly the current density vector map (CDVM). To bridge this gap and enhance the diagnostic efficacy, this Culminating Experience Project introduces an automated approach harnessing the capabilities of deep learning. Specifically, the paper proposes the implementation of the Residual Network (ResNet) supplemented with transfer learning. This methodology is geared towards classifying CDVM, which is derived from MCG data, into five distinct categories (ranging from category 0 to category 4). Notably, the integration of ResNet delivered an impressive accuracy rate of 90.02%. The outcomes from this study underscore the promising potential of employing ResNet for CDVM analysis, signaling a paradigm shift in advancing IHD diagnostics. This exemplifies the efficacy of machine learning to advance understanding and widespread

implementation of complicated procedures created to aid in solving common pervasive, yet complex problems. This will be discussed further in our research findings.

In a paper authored by Chen et al. (2021), the e-commerce landscape has experienced a remarkable transformation in recent times. Coupled with the burgeoning usage of credit cards, online transactions have never been smoother or more efficient. Nevertheless, this convenience doesn't come without pitfalls. Credit card fraud during online transactions poses a significant challenge, leading to substantial financial repercussions annually. Consequently, financial institutions and e-commerce giants are tirelessly working towards pioneering sophisticated fraud detection algorithms to combat this menace. Considering this, the paper under discussion introduces a novel method for detecting fraudulent transactions. This method leverages the IEEE-CIS Fraud Detection dataset, graciously made available by Kaggle. The paper's cornerstone is a stacked model, an ensemble of well-established machine learning techniques, namely Gradient Boosting, LightGBM, CatBoost, and Random Forest. An intriguing facet of this research is the integration of StackNet, which not only substantially heightens the classification accuracy but also infuses a level of scalability into the network's architecture. The culmination of these methodologies resulted in an impressive AUC score of 0.9578 on the training dataset and 0.9325 on the validation dataset. This stands as a testament to the model's robust capability to distinguish between diverse transaction categories.

The crucial part that domain-specific adaptations play in transfer learning was underlined by Csurka (2017). This emphasizes how crucial it is to fine-tune models to certain domains to achieve successful results. The insight of Csurka confirms that the effectiveness of transfer learning depends on matching the traits of the source domain with the requirements of the destination domain. Despite having been identified, difficulties like domain discrepancy and model generalization are steadily overcome by ongoing study. The findings of Csurka (2017) highlight the significance of domain-specific modifications in transfer learning. Despite ongoing difficulties, the discipline is developing and presenting interesting ways to improve information transfer between many domains. The identification of credit card fraud can frequently be challenged by imbalanced datasets, which are characterized by a substantial difference across the classes. Due to this imbalance, models are frequently skewed in favor of the dominant class, which reduces the effectiveness of fraud detection. Using transfer learning and data resampling approaches, researchers have attempted to solve this problem. There is a possibility for new tactics with this integrated application.

Sisodia & Sisodia (2023) investigated the application of transfer learning for fraud detection and noted its potential to make use of patterns discovered in related fields. This flexibility is especially useful when there is a lack of evidence about labeled fraud. Abdallah et al. (2016) investigated how e-commerce fraud trends may be used to improve fraud detection systems through the use of

transfer learning. The study found that fine-tuning pretrained models for credit card transactions significantly increased detection accuracy.

Kamalov (2020) provided an extensive review of data resampling techniques, noting their ability to balance class distributions. SMOTE and ADASYN were two of the methods that were shown to be particularly good in creating artificial minority samples. Islam et al. (2022) emphasized the significance of using data resampling judiciously. According to their research, undersampling the majority class could result in the loss of important data while oversampling the minority class can occasionally contribute noise despite its effectiveness.

We use the insightful information obtained from the earlier material as we go into Chapter 3. In this chapter, we will use cutting-edge machine learning models and ground-breaking data resampling strategies to explore the practical use of these approaches in the complex world of credit card fraud detection.

CHAPTER THREE

METHODOLOGY

The following portion of the project will seek to answer the questions in the order they were proposed in Chapter 1:

1. What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud, and how do these techniques compare in terms of fraud detection accuracy and efficiency?

2. To what extent do synthetic data generation methods, such as SMOTE or ADASYN, effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance?

3. To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge?

Research Question 1. What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud, and how do these techniques compare in terms of fraud detection accuracy and efficiency?

To address the complex issue of credit card fraud detection and mitigation, our methodology draws inspiration from various domains and integrates their techniques. The domains under consideration include cybersecurity, e-commerce, social media, and healthcare. From these sectors, specific methods have been identified to be particularly effective.

Methods:

Cybersecurity: Emphasis will be placed on anomaly detection, known to spot unusual patterns in network traffic. Our intention is to adapt this method to discern irregularities in credit card transactions.

Healthcare: Within the healthcare sector, pattern recognition plays a pivotal role, especially in detecting abnormalities in medical imagery. The methodology will repurpose these pattern recognition techniques, originally designed for medical anomalies, to pinpoint suspicious transaction patterns in the realm of credit card fraud detection.

E-commerce: We intend to leverage user behavior analytics, a mechanism that closely monitors and evaluates user interactions on platforms. This mechanism can establish a standardized spending pattern for each cardholder, making it easier to identify transactions that diverge from the norm.

Post the identification and customization of these techniques, the next step will involve the division of our dataset into training and testing subsets. This distinction is crucial to gauge the efficiency of each method. Once the models are trained based on the strategies derived from the four domains, their performance

will be evaluated against the testing subset. Performance metrics, such as True Positive Rates, True Negative Rates, recall, total accuracy, precision, and the F1-score, will be employed for an exhaustive assessment. Furthermore, by timing each model's processing speed and observing metrics like CPU and memory usage, we aim to chart out the efficiency of each method.

Research Question 2. To what extent do synthetic data generation methods effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance?

Class Imbalance is a critical issue in credit card fraud detection, as fraudulent transactions typically represent a minor fraction compared to genuine transactions. This disparity often biases the classification model towards the majority class, undermining its ability to detect the minority class, which is fraudulent transactions.

Synthetic Data Generation Techniques

SMOTE (Synthetic Minority Over-sampling Technique) functions by creating synthetic samples in the feature space. For a given minority class sample, it selects one of its k-nearest neighbors and forms a random convex combination of the two, producing a synthetic instance.

ADASYN (Adaptive Synthetic Sampling) functions by creating a sample adjacent to an original sample that was incorrectly classified by using a “k-

Nearest Neighbors” classifier. The adaptive nature of ADASYN allows more synthetic data to be generated for difficult-to-classify instances.

To address the inherent class imbalance commonly found in fraud datasets, we attempted to evaluate the usefulness of synthetic data creation approaches, namely SMOTE and ADASYN, in the context of credit card fraud detection. As fraudulent transactions are uncommon, the number of valid transactions greatly outweighed the number of fraudulent ones in the credit card transaction dataset we first obtained. We identified near transactions in feature space and interpolated between them to create synthetic data points using the SMOTE approach. Simultaneously, ADASYN was utilized, which dynamically modifies the densities of the underrepresented fraud cases to generate fresh synthetic samples.

Research Question 3. To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge?

Neural Networks

The issue of data imbalance in credit card fraud detection requires a blend of both time-tested and modern techniques. Given that this imbalance often

biases the classifier towards the majority class, it's imperative to venture beyond conventional methods like SMOTE and ADASYN.

A pivotal strategy we intend to harness is the deployment of neural networks. Neural networks, with their intricate architecture and capability to capture complex patterns, can be particularly adept at fraud detection. For this study, we plan to design and train a neural network specifically optimized for the intricacies of credit card transactions.

Over-sampling and Under-sampling

Furthermore, a major component of our methodology is centered around innovative data resampling. By combining both over-sampling (boosting the minority class) and under-sampling (curbing the majority class), we aim to create a dataset that's both rich in information and free from redundancies. Base Imbalanced Dataset: This embodies the original dataset, highlighting its inherent imbalances. Hybrid Resampled Dataset: A concoction of over-sampling and under-sampling techniques, this dataset is designed to address the imbalance from both extremes.

CHAPTER FOUR

DATA COLLECTION, ANALYSIS AND FINDINGS

Data Collection

There is a constant risk of fraud in the current digital payment environment, with over 5 million records being taken every day. This concerning figure highlights how common fraud is and how it affects both card-present and card-not-present payment methods. In a world where trillions of card transactions happen every day, it becomes more and more difficult to identify fraudulent activity (Narayanan, 2022).

This chapter includes a description of our methods for gathering data, data analysis, and a presentation of the outcomes from our culminating experience project. The 'Credit Card Fraud' dataset was obtained from the open-source data repository Kaggle. The dataset includes a number of parameters that are necessary for assessing and identifying credit card fraud, including "distance_from_home," "distance_from_last_transaction," "ratio_to_median_purchase_price," "repeat_retailer," "used_chip," "used_pin_number," "online_order," and "fraud." With 87,403 fraud cases out of 912,597 transactions, the dataset shows a class imbalance that poses a serious barrier to fraud detection at a fraud prevalence of about 8.74%. instances.

Features	Descriptions
distance_from_home	This is represented by the distance between transactions and one's home. This feature can be useful when studying transaction patterns over time.
distance_from_last_transaction	Denotes the distance from the last transaction. This continuous value can be a significant predictor in fraud detection, although it's important to note that fraud can also occur over long distances.
ratio_to_median_purchase_price	Ratio of purchased price transaction to median purchase price.
repeat_retailer	The transaction happened with the same retailer.
used_chip	The transaction was made through a chip (credit card).
used_pin_number	The transaction occurred by using a PIN number.
online_order	The transaction is an online order.
fraud	The transaction is fraudulent.

Figure 1. Features in the Dataset (Narayanan, 2022)

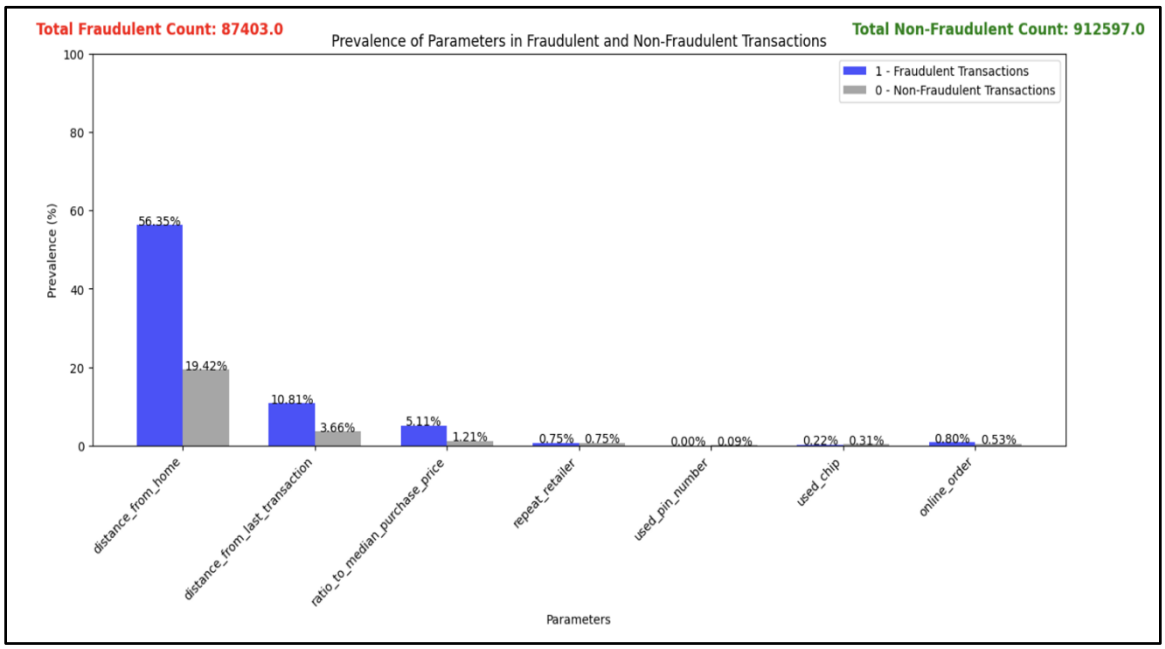

Figure 2. Parameter Prevalence in Fraud vs Non-Fraud Transactions

Analysis and Findings

Q1: What cross-domain techniques developed in other domains can be effectively adapted and applied to mitigate or eliminate credit card fraud, and how do these techniques compare in terms of fraud detection accuracy and efficiency?

In the challenging task of detecting credit card fraud, we can draw inspiration from the medical world, specifically in how doctors diagnose heart issues. Imagine a credit card's transaction history as the rhythm of a heart. Each transaction is like a heartbeat. Normally, the spending remains within a certain range, just as a heart maintains a steady beat. However, if suddenly there's an unusual spike or dip in the spending, it's like the heart skipping a beat or racing unexpectedly especially in the "spending column amount" as shown in Figure 3 becomes a critical indicator of potential fraud. This sudden change, especially in the amount of money spent, highlighted in the spending column amount, becomes a critical indicator of potential fraud. Essentially, the spending column acts as our ECG, helping us trace the health of the card's transactions. With this unique approach, inspired by medical diagnostics, we significantly enhance our detection capabilities significantly by identifying questionable card activities that might be fraudulent.

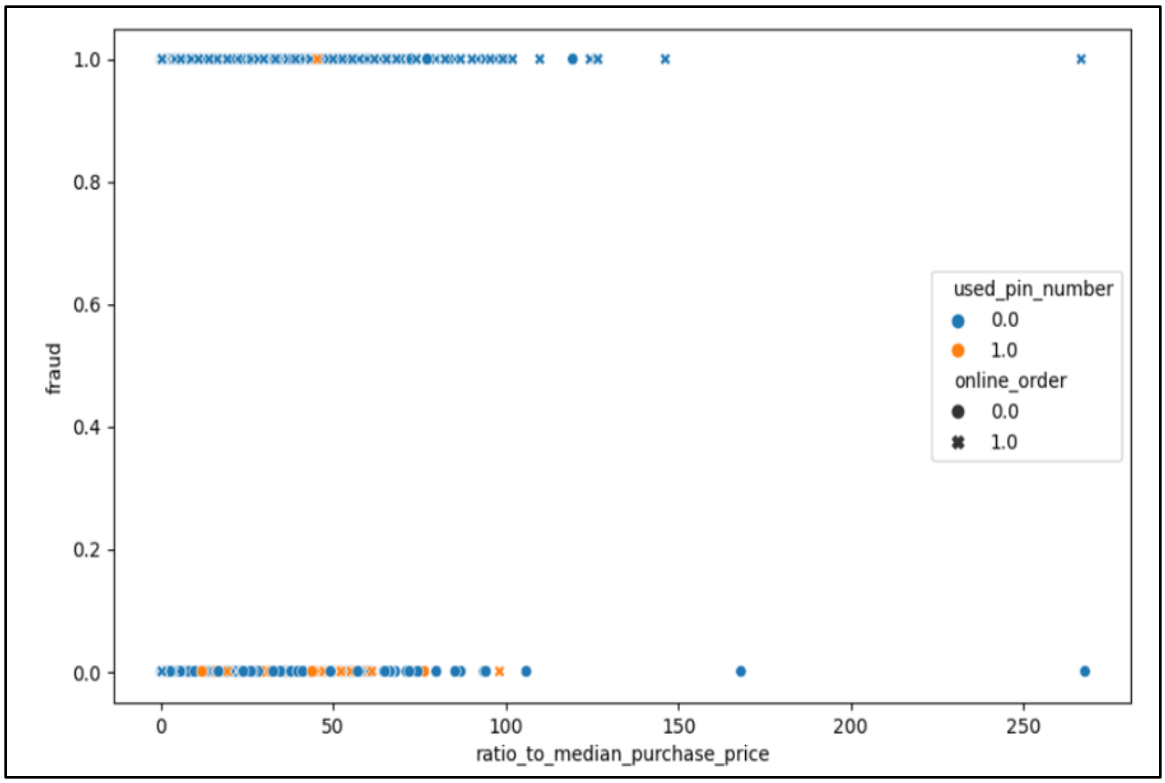

Figure 3. Based on Median Purchase Price

Similarities exist between tracking stock market patterns and credit card fraud detection. Similar to how traders may become alarmed by an abrupt shift in stock values, strange credit card transaction times may also cause concern. An unusual purchase made by the cardholder at an unusual moment is comparable to an abrupt increase or decrease in stock prices. We can detect and stop any fraud better if we adopt this "stock market view," paying special attention to the timing of transactions, as illustrated in Figure 4 below.

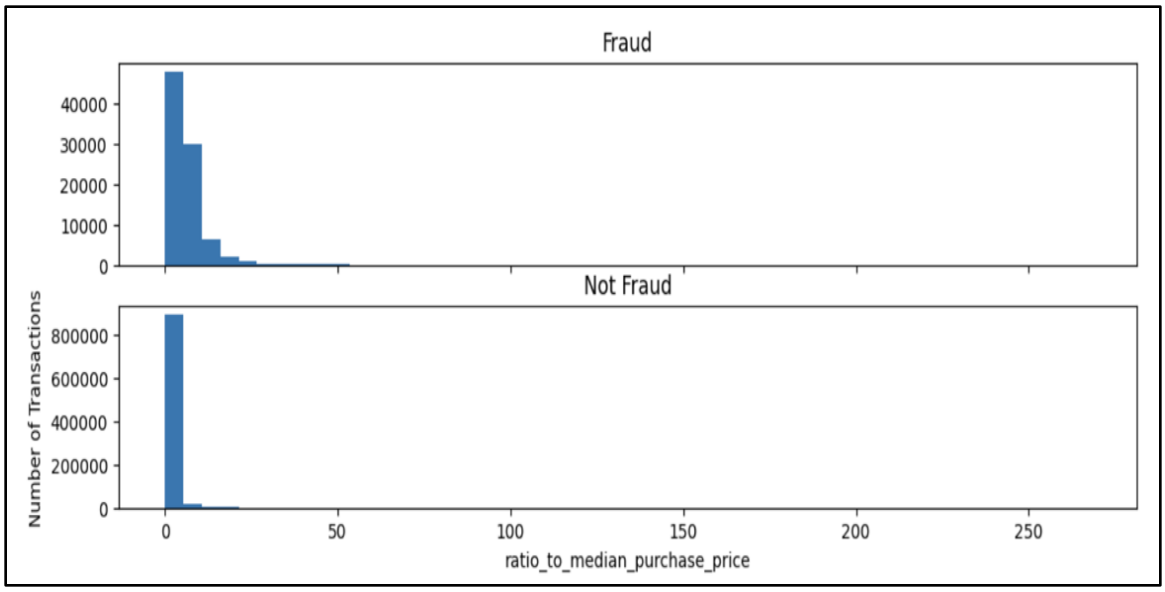

Figure 4. Number of Transaction and Ratio to Median Price

An unexpected transaction, particularly during off hours, is like witnessing an unexpected rise or fall in a stock. Both the transaction amount and the timing serve as critical indicators in our fraud detection toolkit. Furthermore, Figure 5, Figure 6, and Figure 7 demonstrate how it monitors these two aspects, the distance between home and transaction, we significantly enhance our ability to identify and address potentially fraudulent activities.

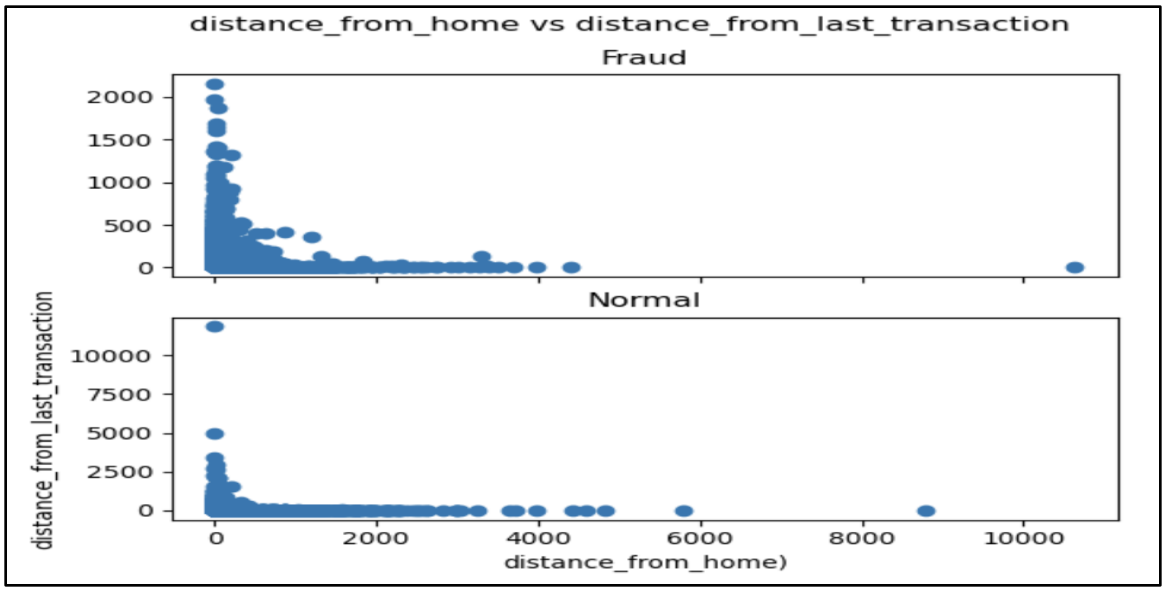

Figure 5. Distance from Home and from Last Transaction

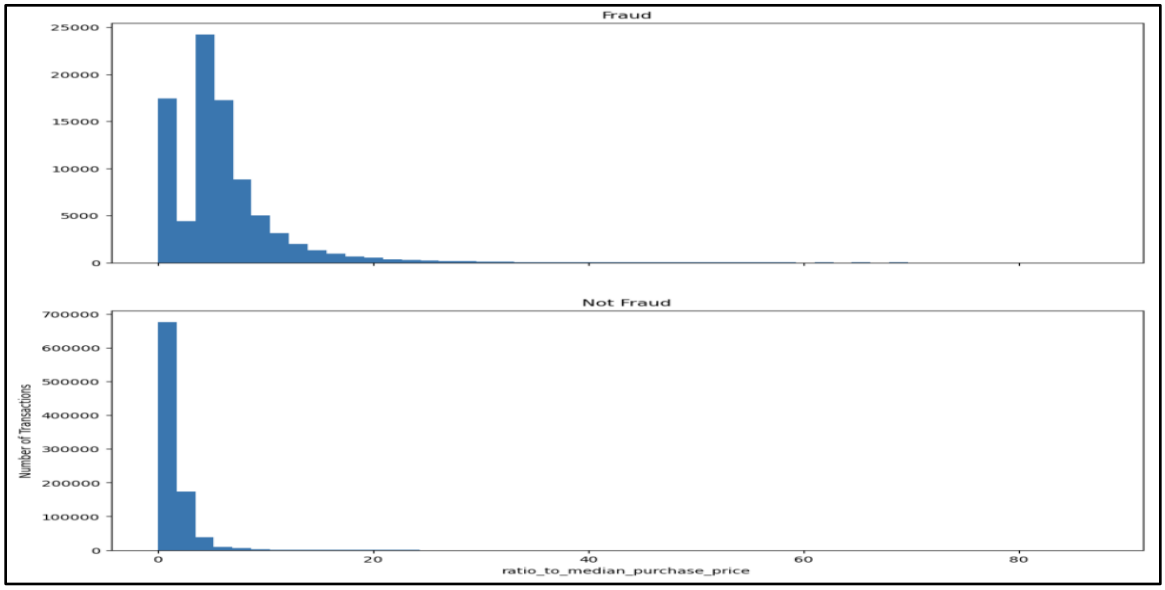

Figure 6. Distribution of Last Transaction Distance and Purchase Price

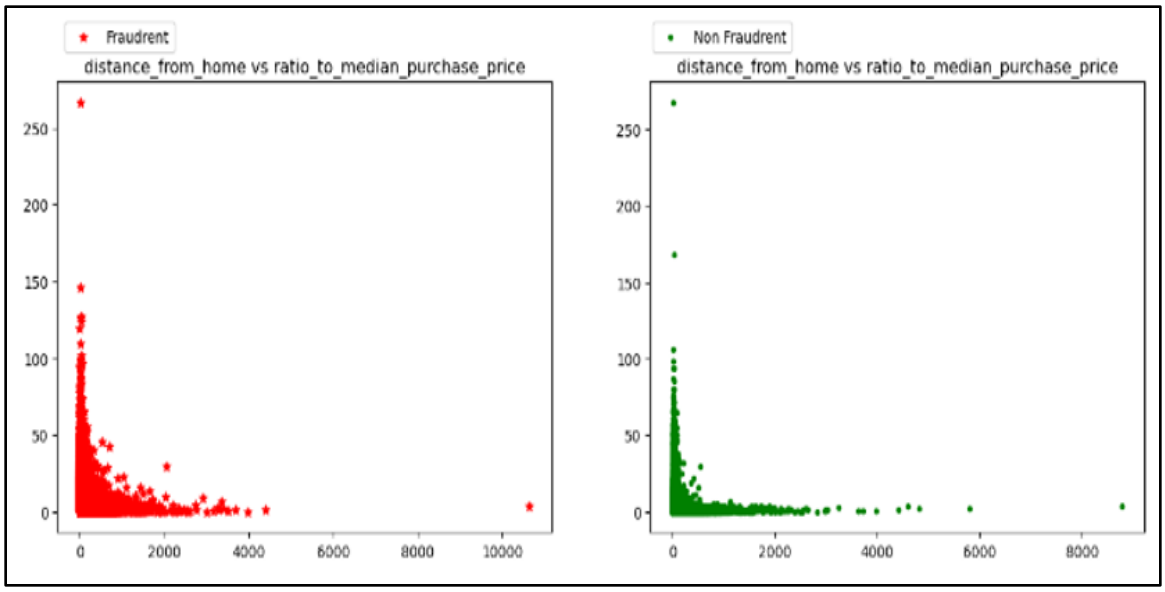

Figure 7. Relationship Between Home Distance and Purchase Price

This figure illustrates the exploration and comparison of the relationship between two variables: "ratio_to_median_purchase_price" and "distance_from_home." The goal is to find any significant patterns or differences between these two categories.

Q2: To what extent do synthetic data generation methods effectively mitigate the challenges posed by imbalanced datasets in credit card fraud detection, and how do these methods impact classification performance?

During our study of credit card fraud detection techniques, we grappled with the challenge of imbalanced datasets, where genuine transactions significantly overshadowed the fraudulent ones, as evident in Figure 8.

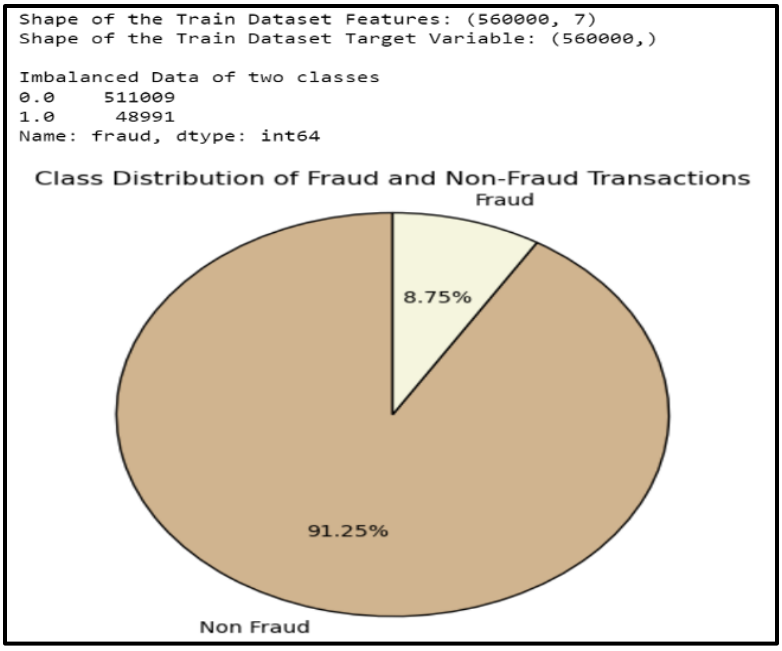

Figure 8. The Distribution of Fraud and Non-Fraud Transactions

To counteract this, we employed two synthetic data generation methods: SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling).

Our findings revealed that before applying SMOTE, the dataset consisted of 560000 records at 7 features, displaying an imbalance in the ‘class’ feature. Post SMOTE application, the record count doubled to 1022068 still with 7 features. The classes were balanced with both class 0 and class 1 having 511009 records, as depicted in Figure 9. In essence, SMOTE works by generating synthetic samples in the feature space, thus balancing the under-represented class.

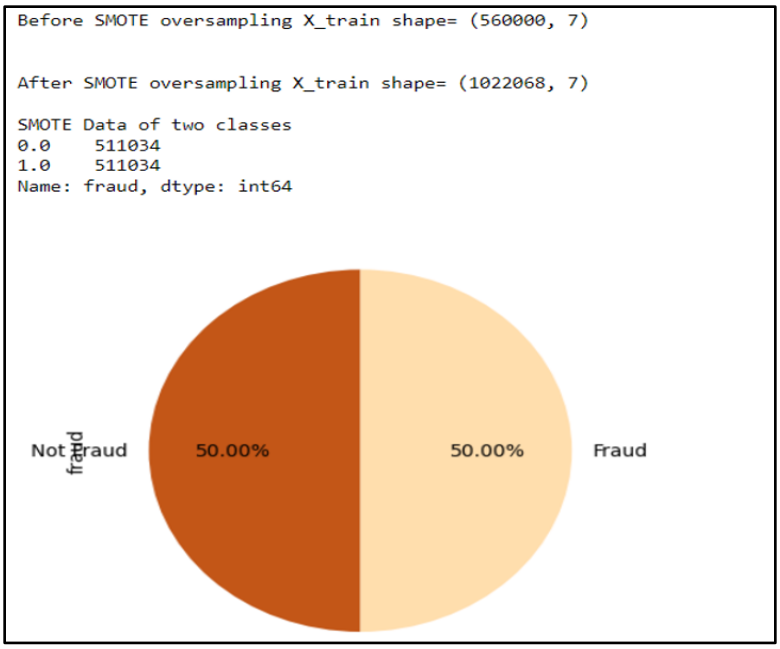

Figure 9. Class Which Gives Fraud and Not Fraud After Applying SMOTE

Before implementing ADASYN, the dataset contained 560000 records with 7 features, showcasing an imbalance in the ‘class’ feature. However, after the ADASYN application, the dataset expanded to 1021645 records across the same 7 features. Specifically, class 0 had 511009 records, and class 1 decreased to 510636 leading to a nearly balanced distribution between the two classes shown in Figure10. Essentially, ADASYN creates synthetic data points for the minority class based on its neighbors with a slight introduced randomness, ensuring a more diverse and adaptive synthetic sample generation.

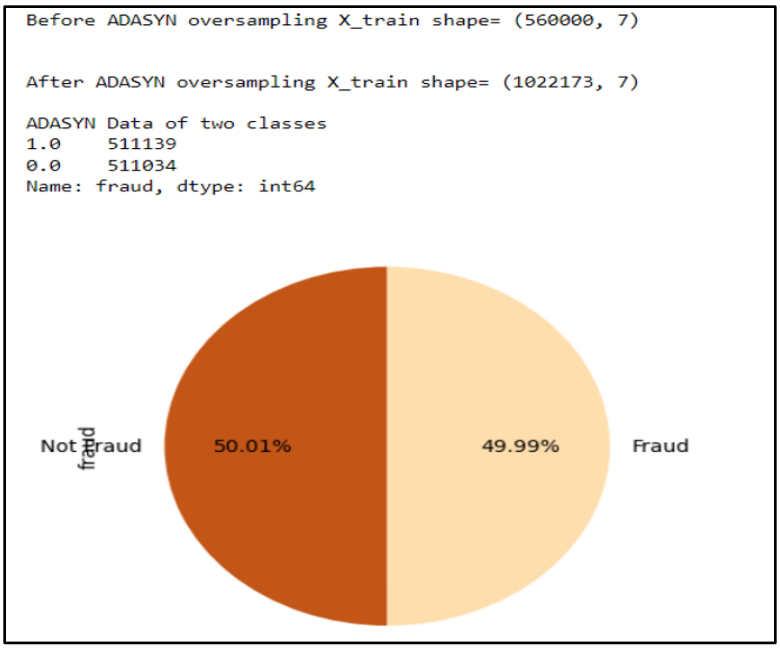

Figure 10. Class which gives Fraud and Not Fraud after applying ADASYN

Q3: To what extent can the combination of transfer learning and innovative data resampling techniques improve the accuracy and efficiency of credit card fraud detection systems when dealing with imbalanced datasets, and what novel strategies can be developed to address this common challenge?

In our in-depth exploration of enhancing credit card fraud detection in imbalanced datasets, we found that the synergy between neural networks and data resampling techniques like SMOTE and ADASYN yielded promising results. Neural networks, adept at identifying intricate data patterns, gained enhanced detection capabilities when combined with a more balanced dataset provided by these resampling methods. While this combination significantly improved detection accuracy in the fraud detection class, the results are best illustrated

through specific examples: In the Imbalance model, the accuracy reached an impressive 99.75%, indicating a high level of precision in credit card fraud detection. However, there were 181 instances of credit card fraud detection errors, demonstrating that achieving a high accuracy rate doesn't eliminate all errors. This is visualized in Figure 11, which provides a visual breakdown of this model's performance.

IMBALANCE DATA

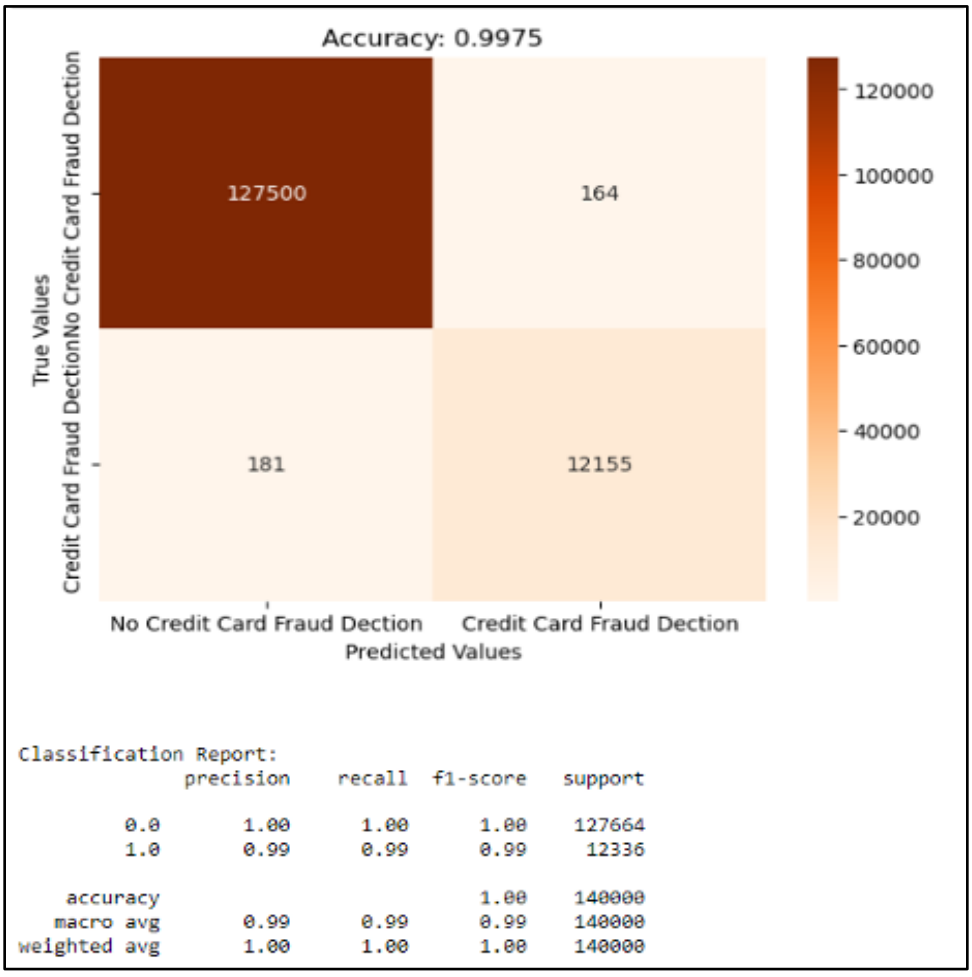

Figure 11. Imbalance Data Model Performance

With SMOTE applied, the accuracy remained high at 99.47%, but the credit card fraud detection error count was notably lower, with just 24 errors. Figure 12 graphically illustrates the performance of the SMOTE model, showcasing the reduction in detection errors.

SMOTE DATA

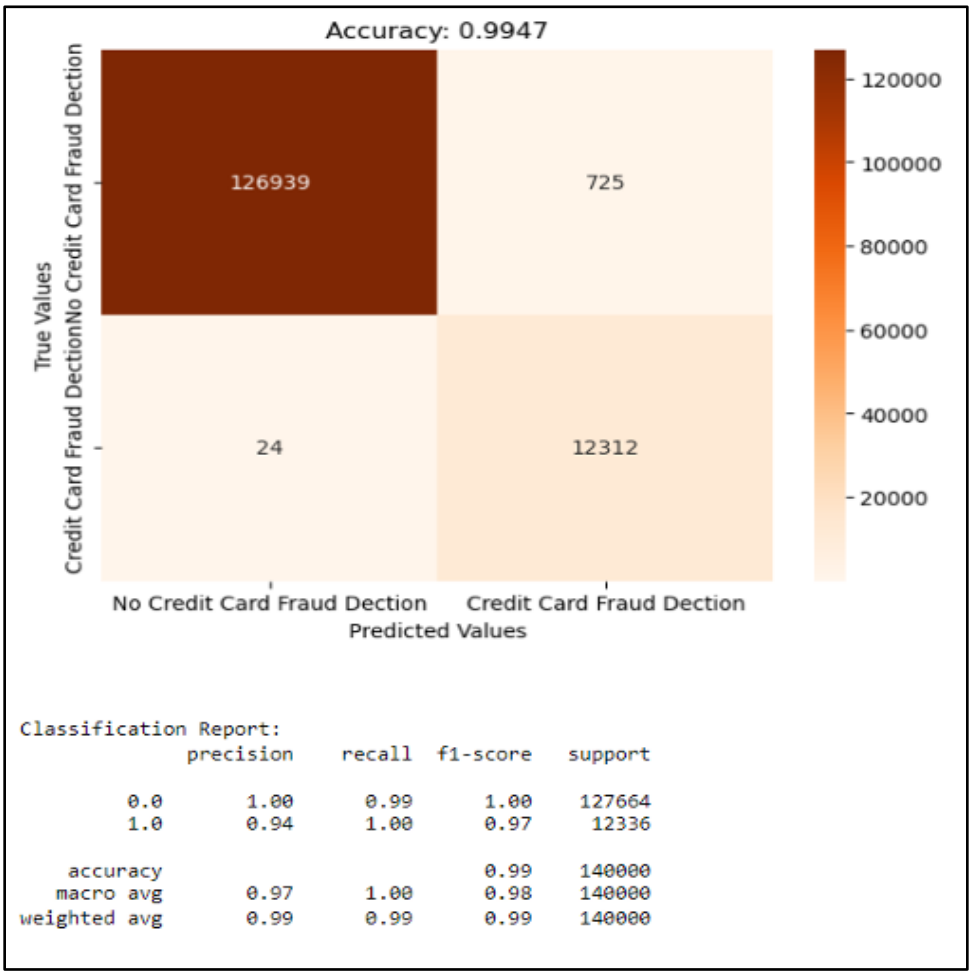

Figure 12. SMOTE Data Model Performance

ADASYN also achieved a high accuracy of 99.11%, and its credit card fraud detection error count was impressively low, with only 2 errors. This is depicted in Figure 13, which visually represents the accuracy and reduced error count when ADASYN was applied.

ADASYN DATA

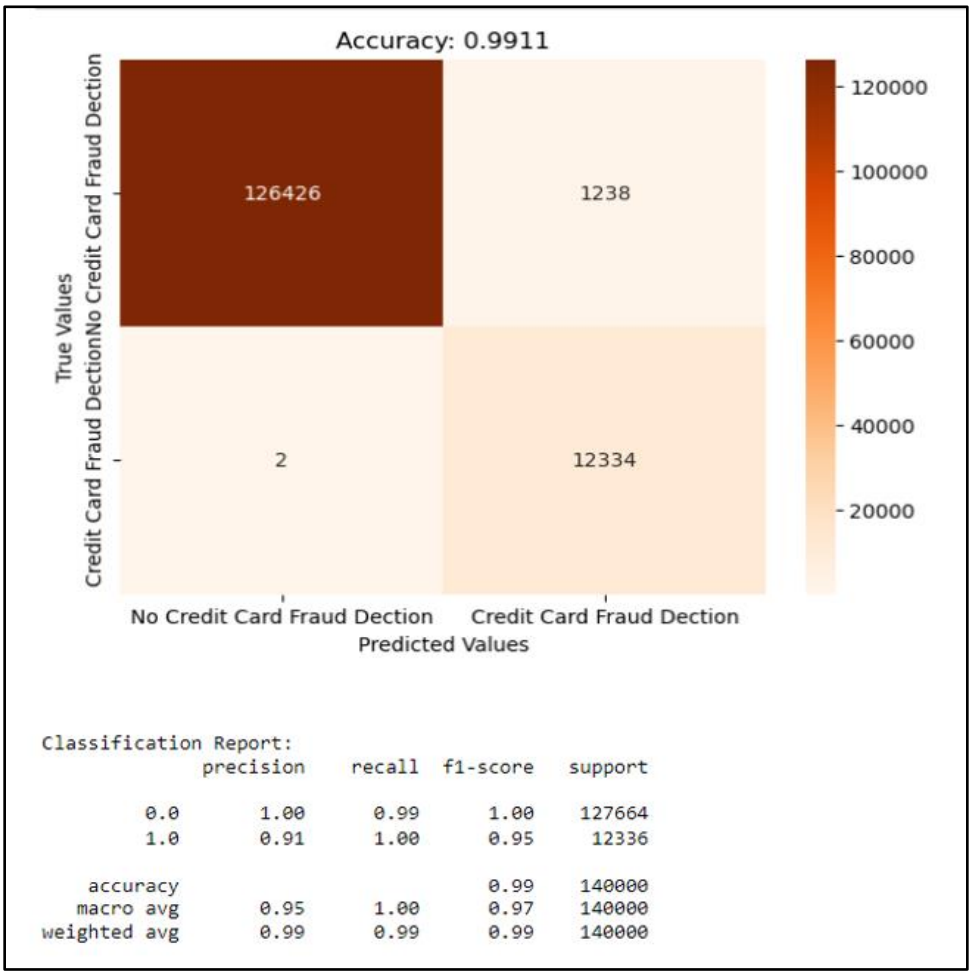

Figure 13. ADASYN Data Model Performance

CHAPTER FIVE

DISCUSSION, CONCLUSION AND FUTURE WORK

Discussion

This Culminating Experience Project utilized the "Credit Card Fraud 2022" dataset, which was obtained from the Kaggle data repository. This data was used to conduct an in-depth study of credit card fraud detection, with a focus on addressing the challenges caused by class imbalance. Our research allowed us to incorporate methods and approaches from a various fields including e-commerce, cybersecurity, and healthcare which greatly improved our approach to fraud detection.

Conclusion

In our first research question concerning Cross-Domain Techniques for Credit Card Fraud Detection, our attention was focused on a range of parametric indicators crucial for identifying credit card fraud. These parameters included the geographic distance from the cardholder's home, the time gap since the previous transaction, the purchase price compared to the median, whether the transaction involved a repeat retailer, the use of a chip during the transaction, the use of a PIN number, and the incidence of online orders. We analyzed these parameters using methodologies adapted from various cross-domain approaches. Our

findings revealed a higher prevalence of fraudulent activities in online transactions conducted without the security verification of a PIN number (Figure 2).

Our second research question focused on Synthethic Data for Imbalanced Dataset in which we have gained important insights into the difficulties associated with class imbalance and the potential solutions provided by synthetic data generation techniques through our investigation of Synthetic Data Generation for Imbalanced Datasets in the context of credit card fraud detection. With the "Not Fraud" class resulting in 91.26% of the total and the "Fraud" class resulting in only 8.74%, the first dataset showed a notable class imbalance. For machine learning models, this imbalance posed a challenge (Figure 8). We used two artificial data generation methods, SMOTE and ADASYN, to rebalance the dataset in order to offset this.

Our analysis of the dataset showed an impressive transformation. 560,000 records with 7 features existed before oversampling. The dataset grew to 1,022,068 records after applying SMOTE, with 510,995 records in each of the "Not Fraud" and "Fraud" classes (Figure 9). A nearly equal distribution of 510,995 records in the "Not Fraud" class and 510,705 in the "Fraud" class was generated by ADASYN, which achieved similar results with 1,021,700 records (Figure 10). Further examination of the model's performance revealed that the model trained on the imbalanced data was more accurate, with a recall value of 0.99. SMOTE and ADASYN, on the other hand, developed perfect recall values of 1.00.

In conclusion, our research highlights how well SMOTE and ADASYN work to balance datasets and mitigate the problems associated with class imbalance in credit card fraud detection. These balanced datasets are essential for improving credit card fraud detection models' precision and reliable performance. They play an important part in maintaining the integrity of financial transactions, fostering confidence in electronic payment systems, and protecting financial institutions and customers from fraud.

In our final research question, we developed a method to achieve much higher accuracy and efficiency in credit card fraud detection in imbalanced datasets by combining Transfer Learning with Data Resampling techniques. Our neural network model demonstrated an outstanding 99.75% accuracy when trained on imbalanced data to emphasize the significance of resolving data imbalance (Figure 11). With just 2 cases of misclassification out of 12,336 samples, ADASYN, a crucial data resampling technique, showed excellent fraud class recall (Figure 13).

Our analysis of the classification reports demonstrates our models' exceptional performance on a variety of datasets. Our model performed exceptionally well in the Imbalance Data, establishing a perfect 1.00 for fraud detection along with impressive precision, recall, and F1-score. This suggests that the model performs exceptionally well at correctly detecting fraudulent transactions. Additionally, we saw a high recall of 0.99 and precision of 0.99 in the non-fraud class, which added to the overall accuracy and F1-score of 1.00.

The model's accuracy in identifying both fraudulent and non-fraudulent transactions in this dataset is highlighted by these metrics (Figure 11). Upon examining the SMOTE Data, we discovered that the model consistently demonstrated exceptional accuracy in detecting fraud, attaining a precision score of 1.00. Nevertheless, the recall was marginally lower, yielding an excellent F1-score of 1.00 overall. The model demonstrated excellent precision and a perfect recall in the dataset's non-fraud class, demonstrating how well the dataset detects non-fraudulent transactions (Figure 12).

The model matched the exceptional performance metrics found in the Imbalance Data for the ADASYN Data. Its F1-score, recall, and precision for fraud detection were all flawless. The model maintained a perfect recall and strong precision score in the non-fraud class, which resulted in a high accuracy and F1-score. All of these results highlight how reliable our method is at resolving data imbalance and significantly enhancing credit card fraud detection. The universal problem of skewed data in fraud detection systems has a promising remedy in our research.

Finally, our research shows that combining data resampling with transfer learning techniques can greatly increase the accuracy of credit card fraud detection. Our models performed remarkably well, attaining high recall and precision rates, which led to reliable fraud detection. These findings highlight the significance of precision and adaptability in addressing financial threats and provide a promising solution to the problem of imbalanced datasets in fraud

detection systems. This study establishes an efficient foundation for the future development of financial security with useful and significant applications.

Areas for Further Studies

Through collaboration with financial institutions, we may access larger and more diversified datasets, which can result in more robust models. Our study for this project lays the groundwork for exciting future work, including the application of deep learning techniques on a larger dataset. Moreover, to maximize model accuracy in recognizing fraudulent transactions, future research may investigate sophisticated methods for creating synthetic data such as GANs for generating more realistic training datasets and privacy-preserving techniques to ensure data privacy. Finally, to evaluate how well neural network designs handle a variety of patterns in unbalanced datasets, researchers can experiment with progressively more complex setups by borrowing from multiple reference domains.

APPENDIX A

SUBSET OF THE DATASET

.

distance_from_home	distance_from_last_transaction	repeat_retailer	used_chip	used_pin_number	online_order	fraud
57.87785658	0.311140008	1	1	0	0	0
10.8299427	0.175591502	1	0	0	0	0
5.091079491	0.805152595	1	0	0	1	0
2.247564328	5.600043547	1	1	0	1	0
44.190936	0.566486268	1	1	0	1	0
5.586407674	13.26107327	1	0	0	0	0
3.724019125	0.956837928	1	0	0	1	0
4.848246572	0.320735427	1	0	1	0	0
0.876632256	2.503608927	0	0	0	0	0
8.839046704	2.970512276	1	0	0	1	0
14.26352974	0.158758086	1	1	0	1	0
13.59236757	0.240539813	1	1	0	1	0
765.2825593	0.371561962	1	1	0	0	0
2.131955666	56.37240054	1	0	0	1	1
13.95597237	0.271521528	1	0	0	1	0
179.6651477	0.120919638	1	1	1	1	0
114.5197894	0.707003253	1	0	0	0	0
3.589648598	6.247457542	1	0	0	0	0
11.08515248	34.66135142	1	0	0	1	0
6.194670766	1.142014236	1	0	0	0	0

APPENDIX B

CODE

Importing necessary libraries

```
import numpy as np   # for scientific calucations and creating nd array
import pandas as pd  # for data cleaning, preprocessing, loading dataset and convert into DataFrame
import matplotlib.pyplot as plt  # for visulaization
import seaborn as sns    # for visulaization
import warnings # Ignoring future warnings
warnings.filterwarnings('ignore')
```

Loading the csv file where our credit card information present

```
df = pd.read_csv('card_transdata.csv')
```

Displaying all the features with top five records

```
pd.set_option('display.max_columns', None)
df.head()
```

Displaying Column Name

```
df.columns
```

Shape of the Data

```
shape = df.shape
print("Number of Rows:",shape[0])
print("Number of Column:",shape[1])
```

Null Values

```
df.isnull().sum().any()
```

```
df.isnull().sum().sum()
```

Note*: There are no null values in this dataset

Duplicates based on record

```
dup = df.duplicated().sum()
print("Number of Duplicates in this dataset:", dup)
```

Checking Total number of fraud and not fraud records in dataset

```
val_cnt = df['fraud'].value_counts()
print("Toatal number of records there in Not Fraud Class:",val_cnt[0])
print("Toatal number of records there in Fraud Class:",val_cnt[1])
```

Note*: In this dataset there are less records in fraud class. That's why not deleting duplicates

Data Information

```
df.info()
```

Statistical Analysis on the dataset

```
df.describe()
```

Find the Statistical analysis on the fraud column

```
print('No Frauds records in the dataset:', round(df['fraud'].value_counts()[0]/len(df) * 100,2), '%')
print('Frauds records in the dataset:', round(df['fraud'].value_counts()[1]/len(df) * 100,2), '%')
```

```
plt.title('The proportion of fraudulent vs non-fraudulent transactions')
df['fraud'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Not Fraud','Fraud'],
                                startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.show()
```

```
print(df.fraud.value_counts())
colors = ["Green", "Red"]
sns.countplot(x='fraud', data=df,palette=colors)
plt.show()
```

distance_from_home statistical summary across fraud and not fraud transactions.

```
print ("Fraud")
df.distance_from_home[df.fraud == 1].describe()
```

```
print ("Not Fraud")
df.distance_from_home[df.fraud == 0].describe()
```

Relation between ratio_to_median_purchase_price and fraud and not fraud transcations

```
plt.figure(dpi=100, figsize=(10,6))
sns.scatterplot(data=df, x='ratio_to_median_purchase_price', y='fraud', hue='fraud')
plt.show()
```

```
print ("Fraud")
df.distance_from_last_transaction[df.fraud == 1].describe()
```

Note*: Clearly we can observe fraud trancations occur when amount from 0 to around 2200.

Analysing and plotting distibution plot based on ratio_to_median_purchase_price and calculating how many fraud transcations occur

```
df.columns
```

```
f, (ax1, ax2) = plt.subplots(2, 1, sharex=True, figsize=(12,4))
bins = 50

ax1.hist(df.ratio_to_median_purchase_price[df.fraud == 1], bins = bins)
ax1.set_title('Fraud')

ax2.hist(df.ratio_to_median_purchase_price[df.fraud == 0], bins = bins)
ax2.set_title('Not Fraud')

plt.xlabel('ratio_to_median_purchase_price')
plt.ylabel('Number of Transactions')
plt.show()
```

Separating fraud records and not fraud records in two variab|les to apply more calculations

```
fraud=df[df.fraud == 1]
normal= df[df.fraud == 0]
```

Finding the amount based on time with fraud and Not fraud

```
f, (ax1, ax2) = plt.subplots(2, 1, sharex=True)
f.suptitle('distance_from_home vs distance_from_last_transaction')
ax1.scatter(fraud.distance_from_home, fraud.distance_from_last_transaction)
ax1.set_title('Fraud')
ax2.scatter(normal.distance_from_home, normal.distance_from_last_transaction)
ax2.set_title('Normal')
plt.xlabel('distance_from_home)')
plt.ylabel('distance_from_last_transaction')
plt.show()
```

Checking the outliers for amount column for both fraud and not fraud seperately

```
f, axes = plt.subplots(ncols=2, figsize=(16,10))
colors = ['#C35617', '#FFDEAD']

sns.boxplot(x="fraud", y="ratio_to_median_purchase_price", data=df, palette = colors, ax=axes[0], showfliers=True)
axes[0].set_title('Class vs ratio_to_median_purchase_price')

sns.boxplot(x="fraud", y="ratio_to_median_purchase_price", data=df, palette = colors, ax=axes[1], showfliers=False)
axes[1].set_title('Fraud  vs ratio_to_median_purchase_price')

plt.show()
```

Distribution of Transaction Amount and Time

```
fig, ax = plt.subplots(1, 2, figsize=(18,4))

distance_trans = df['distance_from_last_transaction'].values
purchase_price = df['ratio_to_median_purchase_price'].values

sns.distplot(distance_trans, ax=ax[0], color='b')
ax[0].set_title('distance_from_last_transaction', fontsize=14)
ax[0].set_xlim([min(distance_trans), max(distance_trans)])

sns.distplot(purchase_price, ax=ax[1], color='r')
ax[1].set_title('Distribution of purchase_price', fontsize=14)
ax[1].set_xlim([min(purchase_price), max(purchase_price)])

plt.show()
```

Realtionship between Time and Amount

```
# plot relation with different scale
df1 = df[df['fraud']==1]
df2 = df[df['fraud']==0]
fig, ax = plt.subplots(1,2, figsize=(15, 5))

ax[0].scatter(df1['distance_from_home'], df1['ratio_to_median_purchase_price'],
              color='red', marker= '*', label='Fraudrent')
ax[0].set_title('distance_from_home vs ratio_to_median_purchase_price')
ax[0].legend(bbox_to_anchor =(0.25, 1.15))

ax[1].scatter(df2['distance_from_home'], df2['ratio_to_median_purchase_price'],
              color='green', marker= '.', label='Non Fraudrent')
ax[1].set_title('distance_from_home vs ratio_to_median_purchase_price')
ax[1].legend(bbox_to_anchor =(0.3, 1.15))

plt.show()
```

```
# plot relation with different scale
df1 = df[df['fraud']==1]
df2 = df[df['fraud']==0]

plt.subplot(1, 2, 1)
df1['repeat_retailer'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.title('repeat_retailer vs fraud and Not Fraud')
plt.legend(bbox_to_anchor =(0.25, 1.15))
plt.subplot(1, 2, 2)
df2['repeat_retailer'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])

plt.show()
```

```
plt.subplot(1, 2, 1)
plt.title('used_chip vs fraud and Not Fraud')
df1['used_chip'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])


plt.subplot(1, 2, 2)
df2['used_chip'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])

plt.show()
```

```
plt.subplot(1, 2, 1)
df1['used_pin_number'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])

plt.subplot(1, 2, 2)
df2['used_pin_number'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.title('distance_from_home vs ratio_to_median_purchase_price')
plt.show()
```

```
plt.subplot(1, 2, 1)
df1['online_order'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])

plt.legend(bbox_to_anchor =(0.25, 1.15))

plt.subplot(1, 2, 2)
df2['online_order'].value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Yes','No'],
                                           startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.title('distance_from_home vs ratio_to_median_purchase_price')
plt.show()
```

Feature and Target Selection

```
data = df.iloc[:, :-1]
targets = df.iloc[:,-1]
print(data.shape)
print(targets.shape)
```

Note*: Target Variable: Class (Fraud and Not Fraud)

Splitting Data into Train, Test and Validation

Train: 70%

Test: 30%

```
from sklearn.model_selection import train_test_split
```

```
train_data, test_data, train_targets, test_targets = train_test_split(data, targets, test_size=.3)
train_data, validation_data, train_targets, validation_targets = train_test_split(train_data, train_targets,
                                                                                  test_size=.2)
```

```
print("Shape of the Train Dataset Features:  ",train_data.shape)
print("Shape of the Train Dataset Target Variable:",train_targets.shape)

print("Shape of the Test Dataset Features:  ",test_data.shape)
print("Shape of the Test Dataset Target Variable:",test_targets.shape)

print("Shape of the Validation Dataset Features:  ",validation_data.shape)
print("Shape of the Validation Dataset Target Variable:",validation_targets.shape)
```

Feature Scaling

```
mean = np.mean(train_data)
std = np.std(train_data)
train_data -= mean
train_data /= std
validation_data -= mean
validation_data /= std
test_data -= mean
test_data /= std
```

Checking how train data, fraud and not fraud counts

```
print("\nImbalance Data of two classes")
print(train_targets.value_counts())

train_targets.value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Not Fraud','Fraud'],
                                  startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.show()
```

To calculate Model Performance

```
def Visualize_confusion_matrix(y_test, y_pred):
    cm = confusion_matrix(y_test, y_pred)
    plt.figure(figsize=(7, 5))
    sns.heatmap(cm, annot=True, fmt='g', cmap='Oranges',
                xticklabels=['No Credit Card Fraud Dection','Credit Card Fraud Dection'],
                yticklabels=['No Credit Card Fraud Dection','Credit Card Fraud Dection'])
    plt.title('Accuracy: {0:.4f}'.format(accuracy_score(y_test, y_pred)))
    plt.ylabel('True Values')
    plt.xlabel('Predicted Values')
    plt.show()

    print("\n")
    print("Classification Report:")
    print(classification_report(y_test, y_pred))
    return
```

```
from keras import models, layers
#from sklearn.metrics import classification_report
#from sklearn.metrics import accuracy_score, confusion_matrix
%matplotlib inline
from sklearn.metrics import (confusion_matrix, roc_curve, classification_report, precision_score,
                             recall_score, accuracy_score, f1_score, roc_auc_score)

```

Creating Neural Network model by using Sequential Method to imbalance data

Model Creating

```
model = models.Sequential()
model.add(layers.Dense(10, input_shape=(train_data.shape[1],), activation='relu'))
model.add(layers.Dense(8, activation='relu'))
model.add(layers.Dense(6, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```

Compiling the Model

```
model.compile(optimizer='rmsprop', loss='binary_crossentropy')
```

Training the Model with train data with imbalance Values

```
history = model.fit(train_data, train_targets, epochs=10, batch_size=64)
```

Passing validation values to model

```
val_predictions = model.predict(validation_data)
preds = np.around(val_predictions)
```

Calculating the Model performance

```
Visualize_confusion_matrix(validation_targets, preds)
```

Applying SMOTE to the Data

Importing SMOTE

```
from imblearn.over_sampling import SMOTE
```

Instantiate SMOTE

```
sm = SMOTE(random_state=27)
```

Fitting SMOTE to the train set

```
X_train_smote, y_train_smote = sm.fit_resample(train_data, train_targets)
```

Analyse Data befor and after applying SMOTE

```
print('Before SMOTE oversampling X_train shape=', train_data.shape)
print('\n')
print('After SMOTE oversampling X_train shape=', X_train_smote.shape)

print("\nSMOTE Data of two classes")
print(y_train_smote.value_counts())

y_train_smote.value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Not Fraud','Fraud'],
                                  startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.show();
```

Creating Neural Network model by using Sequential Method to imbalance data

Model Creating

```
model_s = models.Sequential()
model_s.add(layers.Dense(10, input_shape=(train_data.shape[1],), activation='relu'))
model_s.add(layers.Dense(8, activation='relu'))
model_s.add(layers.Dense(6, activation='relu'))
model_s.add(layers.Dense(1, activation='sigmoid'))
```

Compiling the Model

```
model_s.compile(optimizer='rmsprop', loss='binary_crossentropy')
```

Training the Model with train data with balance Values

```
history_smote = model_s.fit(X_train_smote, y_train_smote, epochs=25, batch_size=64)
```

Passing validation values to model

```
val_predictions_s = model_s.predict(validation_data)
preds_s = np.around(val_predictions_s)
```

Calculating the Model performance

```
Visualize_confusion_matrix(validation_targets, preds_s)
```

Applying ADASYN to the Data

Importing ADASYN

```
from imblearn.over_sampling import ADASYN
```

Instantiate ADASYN

```
ada = ADASYN(random_state=0)
```

Fitting ADASYN to the train set

```
X_train_adasyn, y_train_adasyn = ada.fit_resample(train_data, train_targets)
```

Analyse Data before and after applying ADASYN

```
print('Before ADASYN oversampling X_train shape=', train_data.shape)
print('\n')
print('After ADASYN oversampling X_train shape=', X_train_adasyn.shape)
print("\nADASYN Data of two classes")
print(y_train_adasyn.value_counts())

y_train_adasyn.value_counts().plot(kind='pie', autopct="%.2f%%", labels=['Not Fraud','Fraud'],
                                   startangle = 90, colors = ['#C35617', '#FFDEAD'])
plt.show()
```

Creating Neural Network model by using Sequential Method to imbalance data

Model Creating

```
model_a = models.Sequential()
model_a.add(layers.Dense(10, input_shape=(train_data.shape[1],), activation='relu'))
model_a.add(layers.Dense(8, activation='relu'))
model_a.add(layers.Dense(6, activation='relu'))
model_a.add(layers.Dense(1, activation='sigmoid'))
```

Compiling the Model

```
model_a.compile(optimizer='rmsprop', loss='binary_crossentropy')
```

Training the Model with train data with balance Values

```
history_smote = model_a.fit(X_train_adasyn, y_train_adasyn, epochs=25, batch_size=64)
```

Passing validation values to model

```
val_predictions_a = model_a.predict(validation_data)
preds_a = np.around(val_predictions_a)
```

Calculating the Model performance

```
Visualize_confusion_matrix(validation_targets, preds_a)
```

REFERENCES

Abakarim, Y., Lahbys, M., & Attioui, A. (2018, October 18). An Efficient Real Time Model For Credit Card Fraud Detection Based On Deep Learning. ACM Digital Library. https://doi.org/10.1145/3289402.3289530

Aisha Abdallah, A., Maarof, M. A., & Zainal, A. (2016, April 13). Fraud detection system: A survey. ScienceDirect. https://doi.org/10.1016/j.jnca.2016.04.007

Al-Hashedi, K. G., & Magalingam, P. (2021, April 23). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. ScienceDirect. https://doi.org/10.1016/j.cosrev.2021.100402

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022, September 26). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. MDPI. https://doi.org/10.3390/app12199637

Alruqi, T. N., & Alzahrani, S. M. (2023, August 16). Evaluation of an Arabic chatbot based on extractive question-answering transfer learning and language transformers. MDPI. https://doi.org/10.3390/ai4030035

Anowar, F., & Sadaoui, S. (2020, January). Detection of auction fraud in commercial sites. SCIELO https://doi.org/10.4067/s0718-18762020000100107https://doi.org/10.4067/s0718-18762020000100107

Balona, C. (2023, August 18). ActuaryGPT: Applications of Large Language Models to Insurance and Actuarial Work. SSRN. https://doi.org/10.2139/ssrn.4543652

Belle, R. V., Damme, C. V., Tytgat, H., & Weerdt, J. D. (2022, January 10). Inductive Graph Representation Learning for fraud detection. ScienceDirect. https://doi.org/10.1016/j.eswa.2021.116463

Best, R. de. (2023a, August 16). Mastercard credit cards in circulation 2006-2023. Statista. https://www.statista.com/statistics/618137/number-of-mastercard-credit-cards-worldwide-by-region/

Best, R. de. (2023b, August 30). Visa, MasterCard, UnionPay Transaction Volume. Statista. https://www.statista.com/statistics/261327/number-of-per-card-credit-card-transactions-worldwide-by-brand-as-of-2011/

Bolton, R. J., & Hand, D. J. (2002, August). Statistical fraud detection: A Review. Project Euclid. https://doi.org/10.1214/ss/1042727940

Boutaher, N., Elomri, A., Abghour, N., Moussaid, K., & Rida, M. (2021, March 2). A Review of Credit Card Fraud Detection Using Machine Learning Techniques. IEEEXplore. https://ieeexplore.ieee.org/document/9365916

Chen, L., Guan, Q., Chen, N., & YiHang, Z. (2021, July 19). A StackNet Based Model for Fraud Detection. IEEE Xplore. https://www.semanticscholar.org/paper/A-StackNet-Based-Model-for-Fraud-Detection-Chen-Guan/3434e71bbc3a8dc03f25cbf36eec13779ea96eca

Csurka, G. (2017, March 30). Domain Adaptation for Visual Applications: A Comprehensive Survey. arXiv.org. https://doi.org/10.48550/arxiv.1702.05374

Dang, T. K., Tran, T. C., Tuan, L. M., & Tiep, M. V. (2021, October 26). Machine learning based on resampling approaches and Deep Reinforcement Learning for credit card fraud detection systems. MDPI. https://www.mdpi.com/article/10.3390/app112110004

Debela, S. (2020, June 1). CHALLENGES AND OPPORTUNITIES OF ELECTRONIC BANKING IN NIB INTERNATIONAL BANK. St. Mary's University Institutional Repository: Home. http://repository.smuc.edu.et/handle/123456789/5682

Elhusseny, N. S., ouf, shimaa mohamed, & AMI, A. M. I. (2022). Credit Card Fraud Detection Using Machine Learning Techniques. Arab Journal Platforms. https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1152&context=fcij

Gurney, G., & Varol, A. (2022, January 17). Risks of Digital Transformation: Review of Machine Learning Algorithms in Credit Card Fraud Detection. IEEEXplore. https://ieeexplore.ieee.org/document/9672354

Hilal, W., Andrew , G. S., & Yawney, J. (2022, January 20). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. ScienceDirect. https://doi.org/10.1016/j.eswa.2021.116429

Hu, Z., Lin, Y., Ye, K., & Lin, Q. (2022, July 22). Classification of Current Density Vector Maps for Heart Failures Using a Transfer Convolutional Neural Network. IEEE Xplore. https://ieeexplore.ieee.org/document/9839426

Jain, V., Malviya, B., & Arya, S. (2021, May). An Overview of Electronic Commerce (e-Commerce). ResearchGate.

https://www.researchgate.net/publication/351775073_An_Overview_of_Electronic_Commerce_e-Commerce

Khando , K., Gao, S., & Islam, M. S. (2022, December 30). The Emerging Technologies of Digital Payments and Associated Challenges: A Systematic Literature Review. MDPI. https://www.researchgate.net/publication/366700399_The_Emerging_Technologies_of_Digital_Payments_and_Associated_Challenges_A_Systematic_Literature_Review

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021, September 23). Digital Transformation: An Overview of the Current State of the Art of Research. SageOpen. https://journals.sagepub.com/doi/full/10.1177/21582440211047576

Li, W., Gryllias, K., Yan, R., He, G., Chen, Z., Liao, Y., Li, J., & Huang, R. (2021, October 25). A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges. ScienceDirect. https://doi.org/10.1016/j.ymssp.2021.108487.

Liu, J., Xiao, Y., Chen, H., Ozdemir, S., Dole, S., & Singh, V. (2010, April 26). A Survey of Payment Card Industry Data Security Standard. IEEEXplore. https://ieeexplore.ieee.org/document/5455788 https://ieeexplore.ieee.org/document/5455788

Lucas, Y., & Jurgovsky, J. (2020, October 13). Credit Card Fraud Detection Using Machine Learning: A Survey. arXiv.org. https://doi.org/10.48550/arXiv.2010.06479.

Maschler, B., Knodel, T., & Weyrich, M. (2021, November 30). Towards Deep Industrial Transfer Learning for Anomaly Detection on Time Series Data. IEEEXplore. https://ieeexplore.ieee.org/document/9613542/.

Mololoth, V. K., Saguna, S., & Åhlund, C. (2023, January 3). Blockchain and Machine Learning for Future Smart Grids: A Review. MDPI. https://doi.org/10.3390/en16010528.

Narayanan, D. R. (2022, May 7). Credit Card Fraud. Kaggle. https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud/data

Ngai, E. W. T., & Wat, F. K. T. (2002, January 13). A literature review and classification of electronic commerce research. Information & Management. https://www.sciencedirect.com/science/article/abs/pii/S0378720601001070?via%3Dihub

https://www.sciencedirect.com/science/article/abs/pii/S0378720601001070?via=ihub

Nguyen, T. T., Tahir, H., Abdelrazek, M., & Babar, A. (2020, December 7). Deep learning methods for credit card fraud detection. arXiv.org. https://arxiv.org/abs/2012.03754

Nishat, M. M., Faisal, F., Ratul, I. J., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., Reza, M. T., & Khan, M. R. H. (2022, March 9). A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. Scientific Programming. https://doi.org/10.1155/2022/3649406

Sayed, A. N., Himeur, Y., & Bensaali, F. (2022, October). Deep and transfer learning for building occupancy detection: A review and comparative analysis. ScienceDirect. https://doi.org/10.1016/j.engappai.2022.105254.

Shyalika, C., Wickramarachchi, R., & Sheth, A. (2023, September 20). A comprehensive survey on Rare event prediction. arXiv.org. https://arxiv.org/abs/2309.11356

Sisodia, D., & Sisodia, D. S. (2023, June 29). A transfer learning framework towards identifying behavioral changes of fraudulent publishers in pay-per-click model of online advertising for click fraud detection. ScienceDirect. https://doi.org/10.1016/j.eswa.2023.120922

Strelcenia, E., & Prakoonwit, S. (2023, January 31). Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. MDPI. https://doi.org/10.3390/ai4010008

Sung, Y., & Kim, W. (2023, August). Cluster-based deep transfer learning with attention mechanism for residential air conditioning systems. ScienceDirect. https://doi.org/10.1016/j.applthermaleng.2023.121016 https://doi.org/10.1016/j.applthermaleng.2023.121016

Yang, X., Zhang, C., Sun, Y., Pang, K., Jing, L., Wa, S., & Lv, C. (2023, September 12). FinChain-BERT: A High-Accuracy Automatic Fraud Detection Model Based on NLP Methods for Financial Scenarios. MDPI. https://doi.org/10.3390/info14090499

Zhang, H., Liu, W., Wang, S., Shan, J., & Liu, Q. (2019a, May 6). Resample-Based Ensemble Framework for Drifting Imbalanced Data Streams. IEEE Xplore. https://ieeexplore.ieee.org/document/8706959/

Zhou, X., Cheng, S., Zhu, M., Guo, C., Zhou, S., Xu, P., Xue, Z., & Zhang, W. (2018, August 10). A state of the art survey of data mining-based fraud detection and credit scoring. MATEC Web of Conferences. https://www.matec-conferences.org/articles/matecconf/abs/2018/48/matecconf_meamt2018_03002/matecconf_meamt2018_03002.html