Credit Card Fraud Detection

CMSC 691: Introduction to Data Science

Submitted By:

Sanjaya Gyawali sgyawal1@umbc.edu
David Ledbetter davidl2@umbc.edu
Hasin Ishraq Reefat hasinir1@umbc.edu
Mike Anoruo manoruo1@umbc.edu
Suhee Sanjana Mehjabin suheesm1@umbc.edu

Submitted To: Dr. Abhijit Dutt

Contents

Chapter 1

Introduction

The emergence of credit card fraud as a major concern has coincided with the unparalleled ease brought about by the spread of electronic transactions and online commerce in the quickly changing financial technology landscape [1]. Strengthening security measures is becoming more and more important as financial transactions move to digital platforms. In this dynamic context, credit card fraud detection emerges as a crucial and proactive solution that thoroughly analyzes transactional patterns using advanced algorithms, machine learning, and artificial intelligence [2]. This advanced approach allows for the swift identification of anomalies indicative of fraudulent activities. Beyond safeguarding the interests of consumers and financial institutions, credit card fraud detection stands as a testament to the symbiotic relationship between technology and security in our interconnected world. It not only addresses the pressing need for robust protection in digital transactions but also underscores the collaborative evolution of technological innovation and safeguarding measures. Credit card fraud detection serves as the vanguard against rising risks in this era of unrelenting technological innovation, when the frontiers of financial interactions are constantly stretched. Its significance goes beyond simple identification; it actively participates in the continuous conversation about cybersecurity resilience. This proactive solution constantly refines its arsenal as cybercriminal techniques advance, demonstrating the dynamic interaction between human inventiveness and machine precision [3]. The dedication to strengthening security not only reassures stakeholders, but also places credit card fraud detection as a critical component in the smooth integration of technical innovation and financial security.

Chapter 2

Motivation

Credit card fraud detection is crucial for multifaceted reasons, primarily centered around safeguarding both consumers and financial institutions. Foremost, it serves to protect consumers by identifying and preventing unauthorized transactions, shielding them from potential financial losses and maintaining the integrity of their financial accounts [4]. Financial loss prevention is a pivotal objective, as credit card fraud can result in substantial economic repercussions for both individuals and businesses. Every year the credit card fraudulent activity is increasing rapidly as shown in Figure 2.1. Compliance with regulations is another imperative, ensuring that financial institutions adhere to industry standards and legal requirements in safeguarding sensitive financial information. Moreover, credit card fraud detection plays a vital role in risk mitigation, preemptively identifying and addressing potential threats to the financial ecosystem [5]. It serves to prevent chargebacks, reducing disputes and enhancing the overall efficiency of financial transactions. Simultaneously, the implementation of robust fraud detection mechanisms contributes to an improved customer experience by fostering trust and security in electronic transactions, thereby fortifying the foundation of modern financial interactions [6].

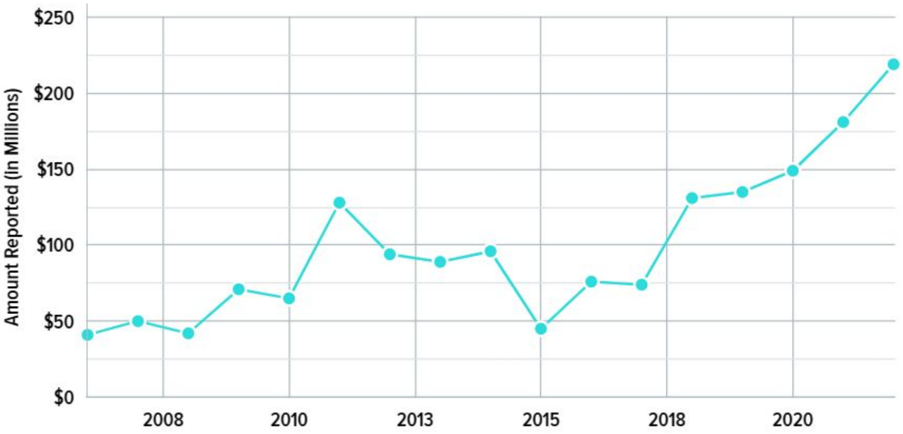


Figure 2.1: Total Value of Credit Card Fraud by Year [7]

Chapter 3

Dataset

This is a dataset of simulated credit card transactions, including both legitimate and fraudulent transactions, that occurred between January 1st, 2019, and December 31st, 2020. The dataset includes transactions from 1,000 credit cards belonging to 800 different merchants. It was generated using the Sparkov Data Generation tool created by Brandon Harris. The simulation ran for a period of two years and the data from separate files has been combined and converted into a standard format for further analysis. We collected the dataset from [8]. The dataset comprises 1,296,675 training and 555,719 test entries, including transaction details, credit card numbers, merchant information, transaction amounts, and personal details of cardholders. The dataset is explained in Table 3.1.

Table 3.1: Description of Dataset

Name	Description
index	Unique Identifier for each row
trans_date_trans_time	Transaction DateTime
cc_num	Credit Card Number of Customer
merchant	Merchant Name
category	Category of Merchant
amt	Amount of Transaction
first	First Name of Credit Card Holder
last	Last Name of Credit Card Holder
gender	Gender of Credit Card Holder
street	Street Address of Credit Card Holder
city	City of Credit Card Holder
state	State of Credit Card Holder
zip	Zip of Credit Card Holder
lat	Latitude Location of Credit Card Holder
long	Longitude Location of Credit Card Holder
city_pop	Credit Card Holder's City Population
job	Job of Credit Card Holder
dob	Date of Birth of Credit Card Holder
trans_num	Transaction Number
unix_time	UNIX Time of transaction
merch_lat	Latitude Location of Merchant
merch_long	Longitude Location of Merchant
is_fraud	Fraudulent or not (target class)

Chapter 4

Exploratory Data Analysis

Fig. 4.1 shows the correlation across all the different variables. From the figure, we see that some of the features are correlated with each other which would negatively affect the model when making predictions. The last column is where we examined to see which features are highly correlated with the dependent variable (is_fraud).

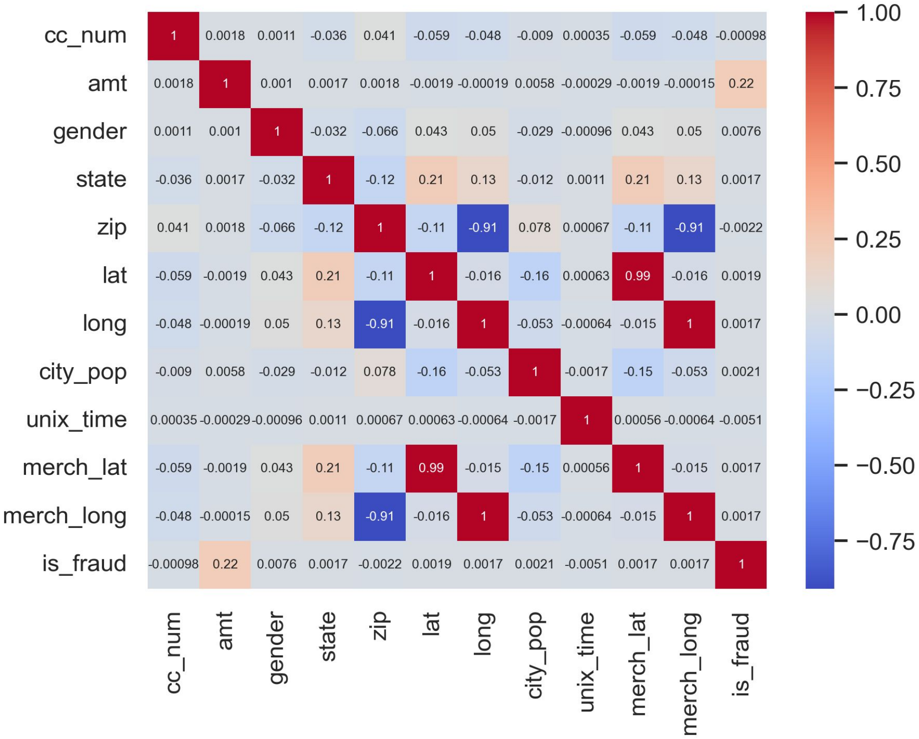


Figure 4.1: Correlation Heatmap

From this, we can confirm that the feature amt best predicts whether or not a transaction is fraudulent, as the correlation between this variable and is_fraud is .22. Later in the paper we will discuss a PCA Neural Network that filters out some of the unwanted correlations in the data set and focuses on only statistically significant features. Overall, this correlation heatmap helped us determine important features to consider when building our models.

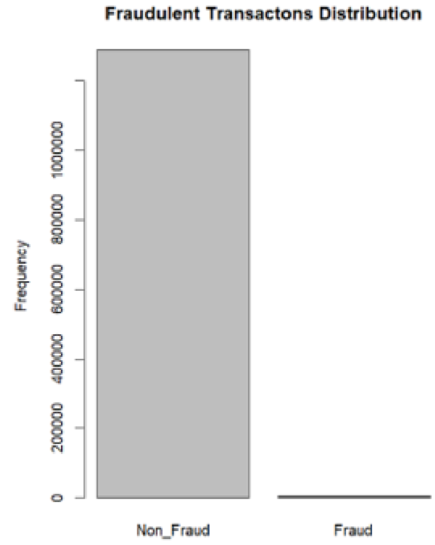


Figure 4.2: Fraudulent Distribution

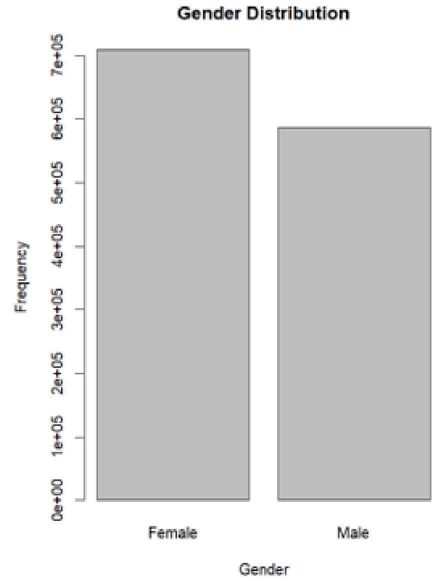


Figure 4.3: Gender Distribution

The fig. 4.2 shows a distribution of fraudulent transactions, with the x-axis representing the number of transactions and the y-axis representing the percentage of transactions. The distribution is skewed to the left, with a majority of transactions being non-fraudulent (99.48%). Fraudulent transactions account for a small minority (0.52%) of the total transactions.

The fig. 4.3 exhibits the gender distribution of credit card fraudulent activity, with women accounting for a slightly higher percentage of fraudulent transactions than men.

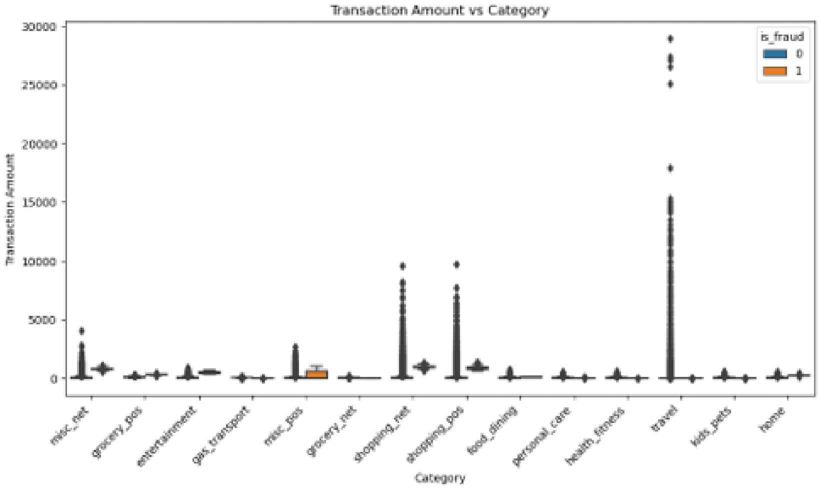


Figure 4.4: Transaction Amount vs Category

The analysis of fig. 4.4 reveals that the majority of fraudulent transactions are for lower amounts (57% less than $500) and occur mainly in shopping, net, and grocery categories (30%, 25%, and 15% respectively). This suggests that fraudsters target both low-value transactions and specific categories, highlighting the need for businesses and consumers to implement targeted strategies to prevent fraud.

In fig. 4.5, the relation between gender and age is shown. There was not a noticeable difference in the genders' ages, and the overall mean age of the entire dataset was found to be about 46 years with a standard deviation of about 17.4 years.

In figs. 4.6 and 4.7, we see the distribution of transactions by time and by day. Although the spread amongst days is fairly similar, Sunday and Monday transactions were more prevalent. Transactions between the hours of 12:00 and 24:00 were much more prevalent than transactions between the hours of 0:00 and 11:59.

Fig. 4.8 shows us the overall count of transactions by state, and fig. 4.9 shows us a plot of transaction amount separated by state. States with high population, like Texas and New York, generally have the higher counts, but the spread of transaction amount by state is much smoother.

The distance between the credit card holder and merchant was found using the Haversine formula with the given x and y coordinates. In fig. 4.10, we see the distribution of distance amongst all cases, fraudulent cases, and non fraudulent cases, respectively. After performing correlation analysis on these subsections, there was no correlation found.

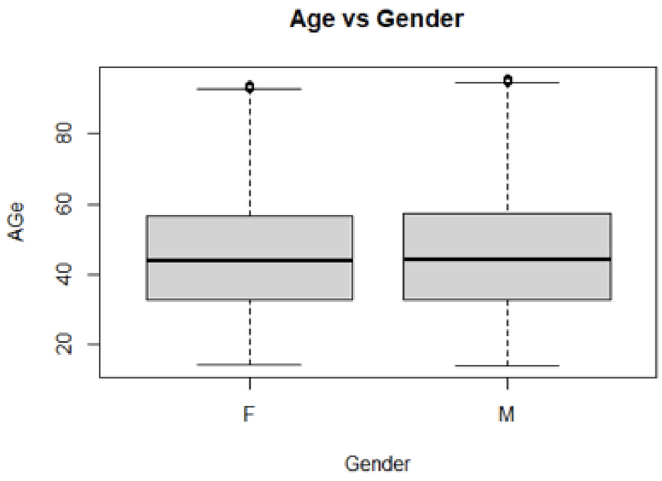


Figure 4.5: Age vs Gender

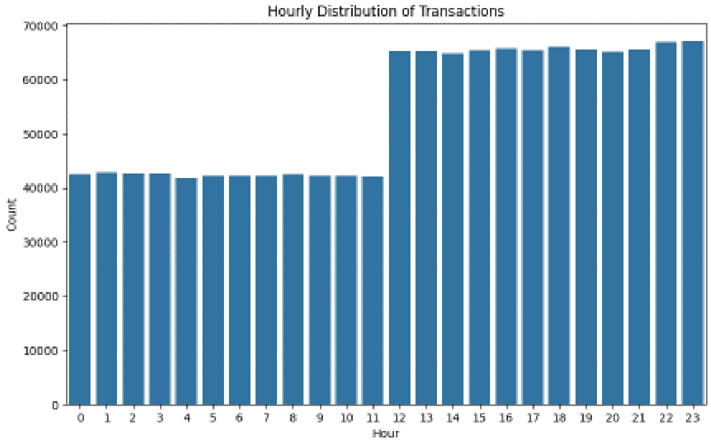


Figure 4.6: Hourly Distribution of Transactions

Fig 4.11 showcases a 2D plot of all of the transactions in the dataset. Notice that there is a clear outline of the map of the United States, with some outlying points in Hawaii and Alaska.

After careful analysis, we selected Transaction Amount, Gender, and State as the most relevant features for predicting fraudulent activity. These features demonstrated the strongest correlation with our target variable, ”is_fraud,” making them the best predictors of fraudulent transactions. While features like latitude and longitude also showed good correlation, we opted for State as it encompasses a broader geographical area and provides a more holistic view of regional trends.

Additionally, Gender was included due to its potential role in identifying fraudulent

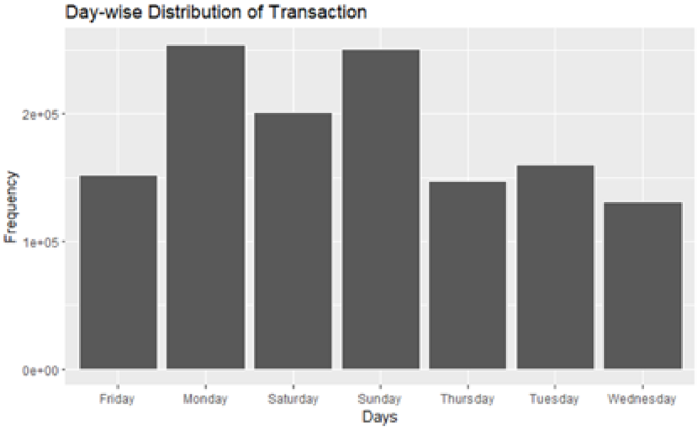


Figure 4.7: Daywise Distribution of Transactions

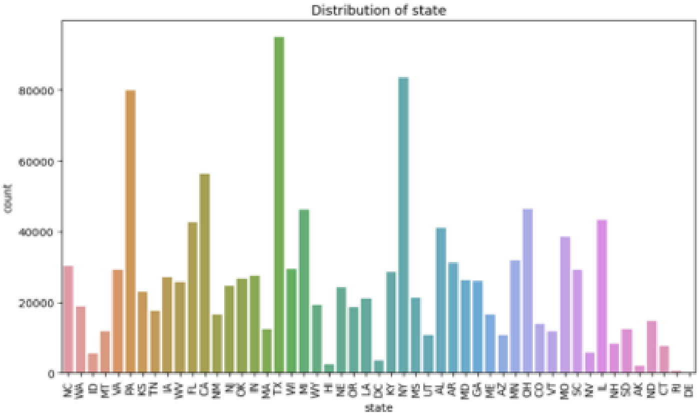


Figure 4.8: Distribution of State

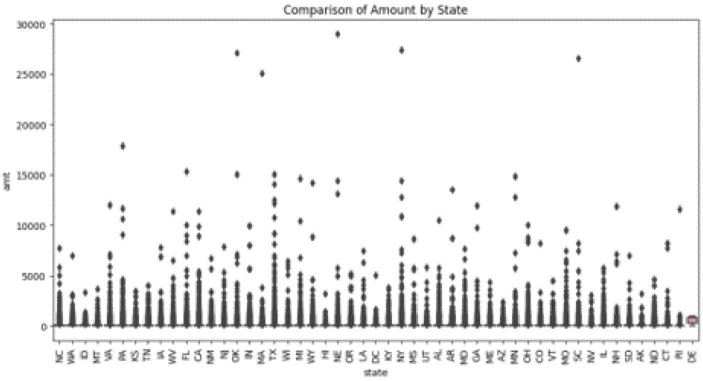


Figure 4.9: Comparison of Amount by state

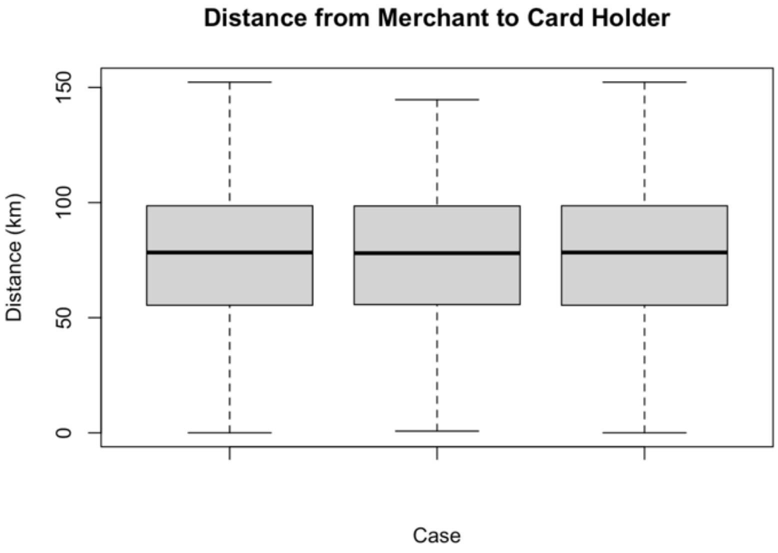


Figure 4.10: Distance from Merchant to Card Holder

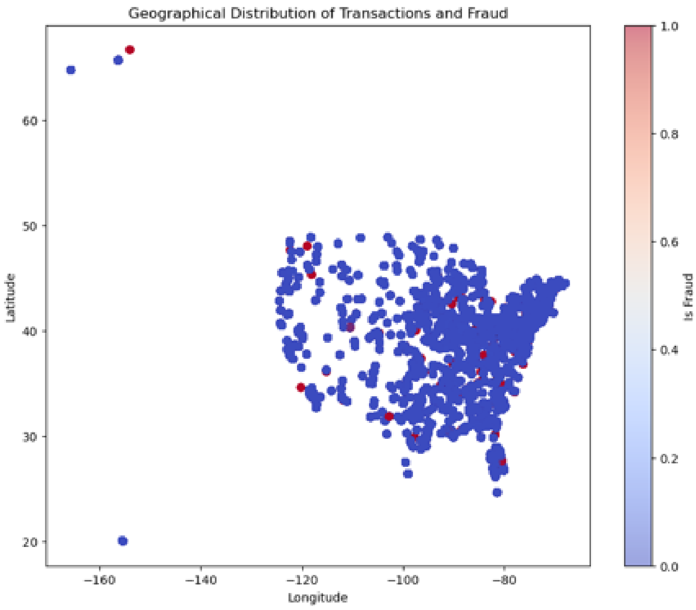


Figure 4.11: Geographical Distribution of Transactions and Fraud

behavior. This combination of features balances geographical context (State), individual characteristics (Gender), and financial context (Transaction Amount) to provide a comprehensive and accurate framework for fraud detection.

Chapter 5

Models and Result Analysis

The project utilized models such as Multiple Regression, Logistic Regression, Naïve Bayes, Neural Network, and Linear Regression. These were evaluated based on accuracy, precision, recall, and AUC scores.

5.1 Multiple Regression

After performing the exploratory data analysis, we found that the three most correlated features with the dependent variable were transaction amount, gender, and state. With this, we created a multiple regression model with the aforementioned features as the independent variables.

Table 5.1: Multiple Regression Confusion Matrix [before balancing]

preds	**0**	**1**
0	548404	5170
1	1109	1036

As you can clearly see from Table 5.1, the number of true negatives is substantially greater than the true positives, false negatives, and false positives. Because of this imbalance in the data, we balanced the test dataset by sampling an equal number of fraudulent and non fraudulent cases and tested the model again.

Table 5.2: Multiple Regression Confusion Matrix [after balancing]

preds	**0**	**1**
0	2103	42
1	781	1364

After balancing, the spread of the different cases was much more reasonable. From Table 5.1, we see that actual non fraudulent cases were predicted very well, with only

42 incorrect predictions, but the model struggled slightly with actual fraudulent cases (64% accuracy).

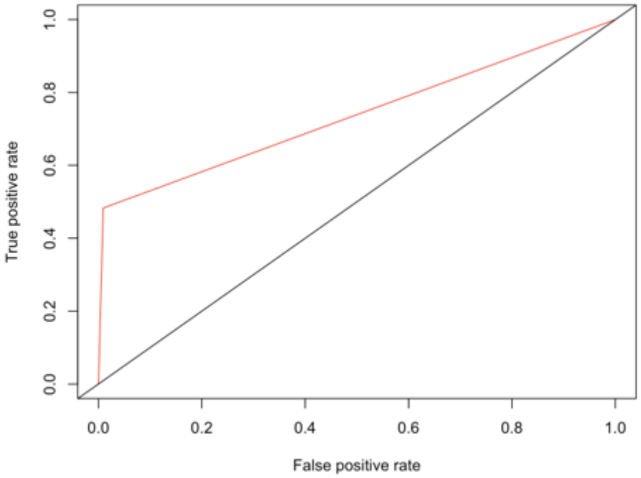


Figure 5.1: Multiple Regression ROC [before balancing]

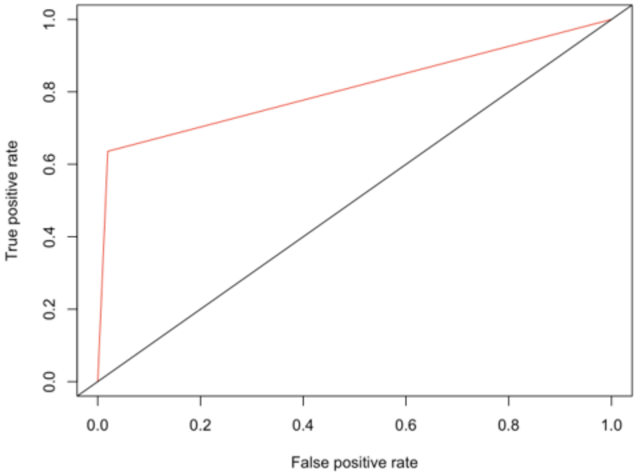


Figure 5.2: Multiple Regression ROC [after balancing]

From the two graphs in figures 5.1 and 5.1 we see that the ROC graphs are similar before and after the balancing. This makes sense, because the balancing mainly fixed the bias towards true negatives, which does not effect the ROC.

5.2 Logistic Regression

We performed a Logistic Regression model using the following three features: *amt*, *gender* and *state*. Table 5.3 represents the confusion matrix by the running the model with unbalanced data. Overall, the model seems to perform well as it correctly classified a large number of samples, although it shows a very low precision. This is because it missed a significant number of actual positive cases. This suggests that the data might have been imbalanced, meaning that one class might have significantly more samples than the other.

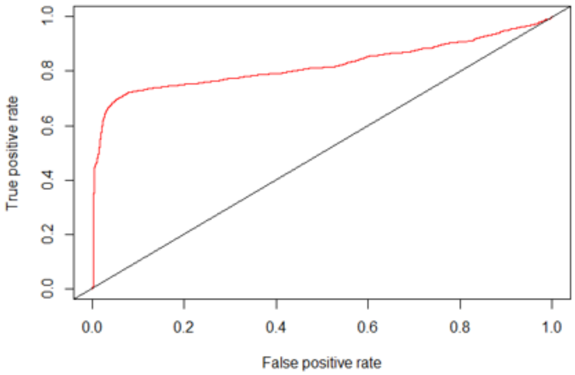


Figure 5.3: Logistic Regression ROC [before balancing]

From fig. 5.3 we can see that ROC curve shows a good relationship between the true positive rate (TPR) and the false positive rate (FPR). It has a high TPR and a low FPR at all threshold values.The AUC for this is 0.82. Because of the highly imbalanced data, we balanced the data out by identifying a sample size that is almost equally balanced between zeros and ones. From table 5.4 that balancing significantly reduced the class imbalance. The number of false positives (FP) has decreased significantly for class 1, leading to improved precision. Additionally, the number of false negatives (FN) has also decreased, leading to improved recall. After balancing we again performed the ROC and generated the AUC curve. Fig. 5.4 is very close to the perfect curve. It has a high TPR and a low FPR at all threshold values. This means that the logistic regression model is able to accurately distinguish between positive and negative cases. The AUC for this is 0.824.

Table 5.3: Logistic Regression Confusion Matrix [before balancing]

preds	**0**	**1**
0	553215	2145
1	359	0

5.3 Naïve Bayes

The Naive Bayes model is performed based on three features: *amt*, *gender* and *state* and exhibits strong performance in detecting fraudulent transactions, achieving an accuracy of 99.2%. The confusion matrix (before balancing) is shown in Table 5.5. This

Table 5.4: Logistic Regression Confusion Matrix [after balancing]

preds	**0**	**1**
0	2029	551
1	116	1594

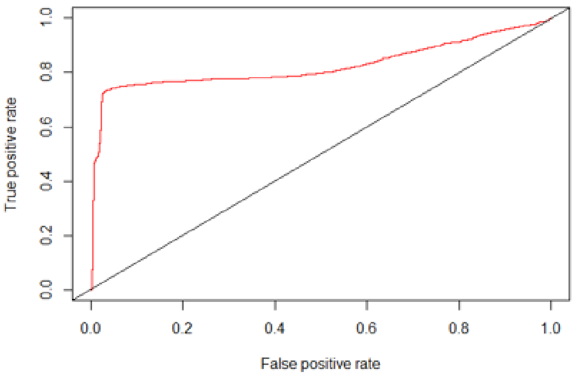


Figure 5.4: Logistic Regression ROC [after balancing]

suggests its ability to accurately classify the majority of transactions and data is highly imbalanced. Furthermore, the model demonstrates high precision (89%) and recall (96%), indicating its effectiveness in identifying genuine fraudulent transactions while minimizing false positives.

Table 5.5: Naïve Bayes Confusion Matrix [before balancing]

preds	**0**	**1**
0	550518	1138
1	3056	1007

ROC curve in Figure 5.5 and an AUC score of 0.8366 further reinforces the model's ability to distinguish fraudulent and legitimate transactions. These results suggest that the Naive Bayes model holds promise as a valuable tool for combating credit card fraud. However, we need to balance the dataset and apply the model.

Table 5.6: Naïve Bayes Confusion Matrix [after balancing]

preds	**0**	**1**
0	2086	571
1	59	1574

The confusion matrix exhibited in Table 5.6 for the Naive Bayes model after balancing the dataset shows that the model maintains its strong performance in detecting fraudulent transactions, achieving an accuracy of 85.31%. This suggests that the model is able to generalize well to new data, even after the dataset has been balanced. Furthermore, the model demonstrates good precision (0.73) and recall (0.96), indicating that it is effective in identifying fraudulent transactions while minimizing false positives. ROC curve in Figure 5.6 and an AUC score of 0.8411 further reinforces the model's ability to distinguish fraudulent and legitimate transactions.

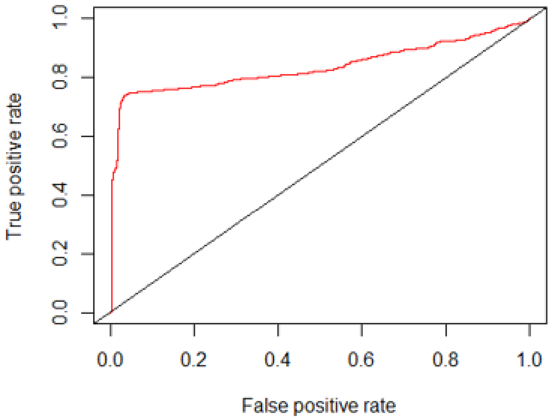


Figure 5.5: Naïve Bayes ROC [before balancing]

Overall, these results suggest that the Naive Bayes model is a promising tool for detecting fraudulent transactions, even in imbalanced datasets. The model's high accuracy, precision, recall, and AUC scores demonstrate its ability to accurately identify fraudulent transactions while minimizing false positives.

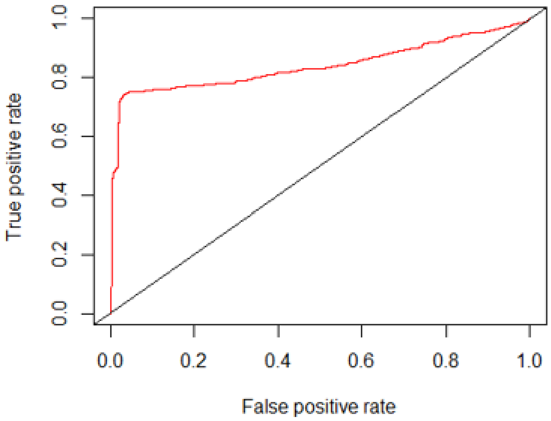


Figure 5.6: Naïve Bayes ROC [after balancing]

5.4 Neural Network

We deployed a neural network model using R's `neuralnet` library to predict credit card fraud based on various features. The training dataset is initially explored and balanced to ensure equal representation of fraud and non-fraud instances. Categorical variables were then converted to numerical form, and data normalization is applied to specific features (those that were not categorical). The neural network, with two hidden layers and linear output, is trained using a formula that includes normalized amounts, state numbers, gender numbers, and normalized city populations as input features. The evaluation of the model reveals an accuracy of 0.86, precision of 0.8, recall of 0.96, and an AUC of 0.86. The confusion matrix highlights the model's ability to correctly identify both positive and negative instances, with 2059 true positives and 1614 true negatives.

Table 5.7: Neural Network Confusion Matrix [after balancing]

preds	**0**	**1**
0	2059	531
1	86	1614

Fig. 5.7 illustrates the architecture of the neural network employed in the credit card fraud prediction model. The visual representation highlights the flow of information through various layers. The input layer receives the preprocessed features including the normalized amount (`amt_norm`), state (numerical form) (`state_num`), gender (numerical form) (`gender_num`), and normalized city population (`city_pop`).

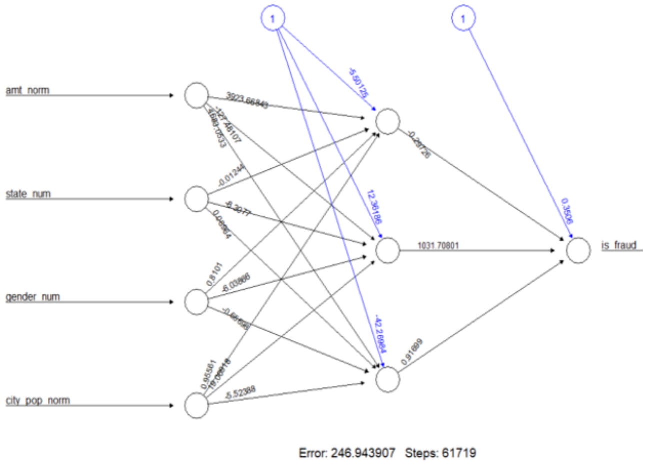


Figure 5.7: Neural Network Layers

The neural network structure incorporates 1 hidden layer, symbolizing an intermediate stage where the relationship between the inputs and output (is_fraud) is learned.

The final layer showcased is the output layer, where the prediction for the target variable `is_fraud` is executed. This layer acts as the decision point of the neural network, delivering the ultimate classification based on the learned patterns from the input features. The visualization encapsulates the sequential flow of information through the network, transforming the input features into a prediction for credit card fraud.

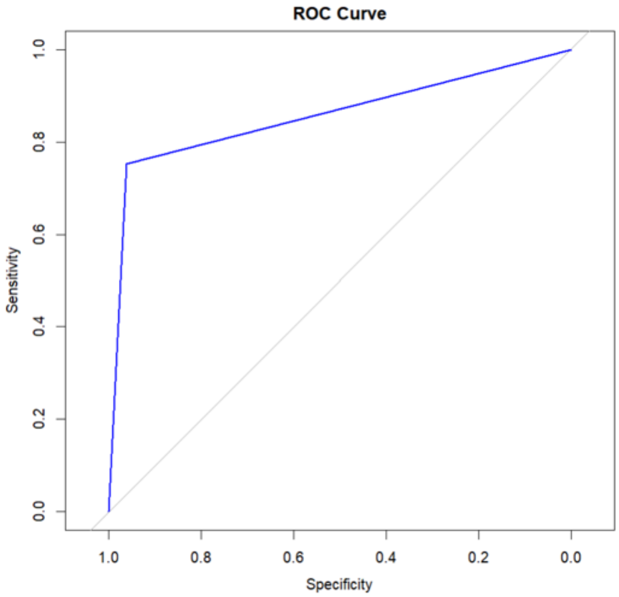


Figure 5.8: Neural Network ROC [after balancing]

Overall, the neural network demonstrates robust performance in predicting credit card fraud, as indicated by its high recall and satisfactory precision. The ROC curve further illustrates the model's discriminative capability. Which can be seen in the ROC Curve in Fig. 5.8.

5.5 Neural Network - PCA

The neural network PCA model utilizes the Principal Component Analysis (PCA) to reduce the number of features used for the credit card fraud predictions. PCA works by identifying patterns and identifying the eigenvectors and eigenvalues of the covariance matrix, allowing the model to focus on the most relevant aspects of the data. Also, as eigenvectors are calculated to be fundamental and incovariant. There are trade-offs when choosing the amount of components for the model as detailed in figure 5.9.

The 6 component vectors chosen represent just over 50 percent of the original variance. While PCA retains much of the structure and reduces the feature space, In this context, PCA likely aids in mitigating issues associated with correlated or redundant features, providing a streamlined input for the neural network.

Through the application of PCA, the model gains a condensed representation of the data, consisting of principal components that capture the primary sources of variance.

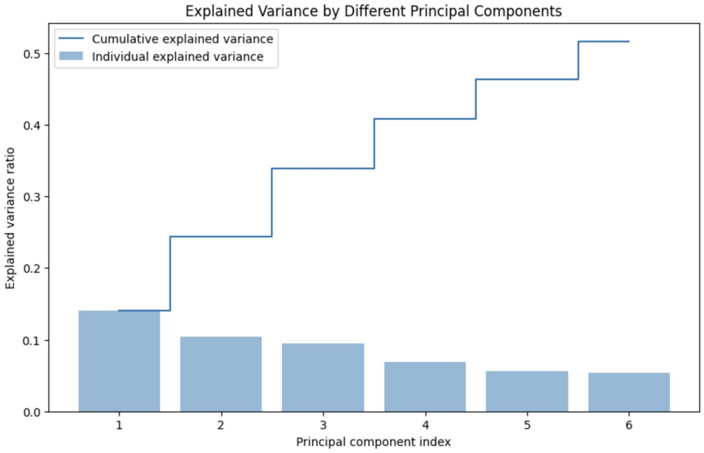
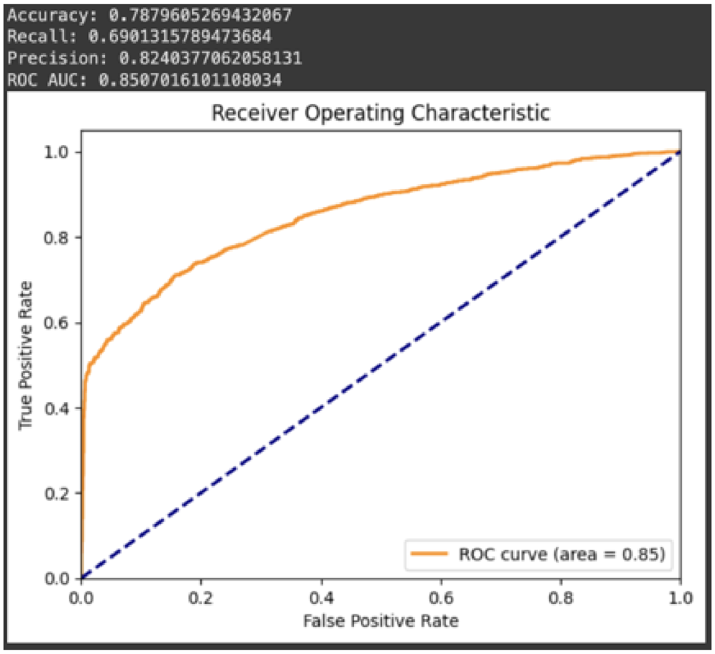


Figure 5.9: Explained Variance Ratio vs Principal Component Index

These components, which are uncorrelated and ordered by their significance, are then fed into the neural network for credit card fraud prediction. This approach not only enhances computational efficiency but also facilitates a clearer understanding of the critical features contributing to fraud detection.



```
Accuracy: 0.7879605269432067
Recall: 0.6901315789473684
Precision: 0.8240377062058131
ROC AUC: 0.8507016101108034
```

Figure 5.10: Neural Network (PCA) ROC [after balancing]

The model's performance evaluation was an accuracy of 0.78, precision of 0.82, recall of 0.69, and an AUC of 0.85. While achieving strong precision in identifying fraud, the model performed poorly when it came to recall. This suggests opportunities to better capture instances of actual fraudulent transactions. The incorporation of PCA serves as a strategic means to streamline feature input.

5.6 Linear Regression

For the dataset we used it has been explained in section 4 that the best correlation value is between *is_fraud* and *amt*. Therefore, we decided to do a linear regression with one dependent variable.

Table 5.8: Linear Regression Confusion Matrix [after balancing]

preds	**0**	**1**
0	2094	51
1	562	1583

From the creation of the previous models, we knew that balancing the data was very important for this particular dataset, so we only tested the linear model on the balanced dataset. From table 5.8, we see that the results are similar to the results from the multiple regression from 5.2. Since the independent variable *amt* was the sole variable used in this model, this may indicate that it is the most significant variable out of the three used in the multiple regression.

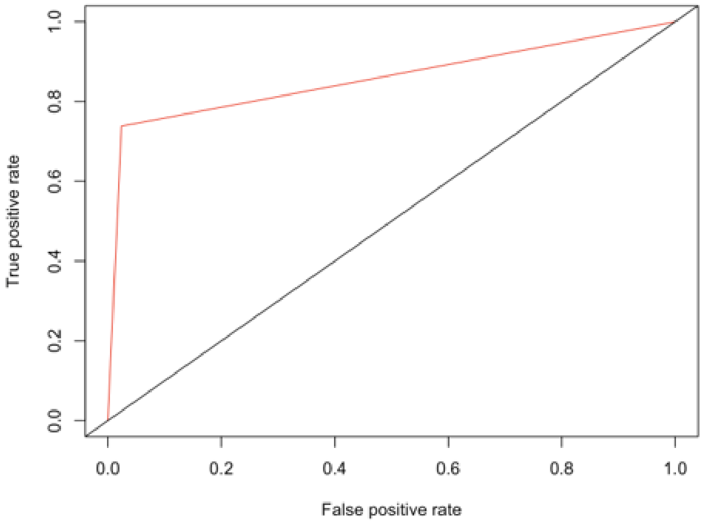


Figure 5.11: Linear Regression ROC [after balancing]

Similarly, the ROC graph of the linear regression in fig. 5.11 closely follows the ROC graph of the multiple regression from fig. 5.2 with slightly better accuracy and AUC values.

Balancing the data changed the models' accuracy and AUC scores significantly, indicating improved discrimination between fraudulent and non-fraudulent transactions.

Chapter 6

Discussion

In addressing the scarcity of fraud cases in our dataset, we opted to balance the data, aiming for a more representative training set. However, this adjustment resulted in a significant decline in overall accuracy, attributable to the substantial influence of the originally skewed data. The comparison of the models are exhibited in Table 6.1.

Table 6.1: Comparison of Models

ML Models	Accuracy	Precision	Recall	AUC
Multiple Regression	0.81	0.97	0.66	0.81
Logistic Regression	0.84	0.74	0.93	0.825
Naïve Bayes	0.85	0.73	0.96	0.841
Neural Network	0.86	0.80	0.96	0.86
Neural Network (PCA)	0.79	0.82	0.69	0.851
Linear Regression	0.86	0.97	0.74	0.857

AUC scores closely mirrored accuracy, serving as a reliable metric for evaluating the models' ability to discern between true and false positives. Notably, linear regression demonstrated the highest accuracy, leveraging the ”Transaction Amount” feature, which exhibited a stronger correlation with fraudulent activity than other variables. Among non-linear models, Naïve Bayes and Neural Networks outperformed others, prioritizing recall over precision, indicating a focus on minimizing false negatives, a crucial consideration in real-world scenarios where the Naïve Bayes model might hold a distinct advantage. Conversely, Multiple Regression and Linear Regression prioritized precision, capturing more true positives but potentially introducing false positives. These nuanced trade-offs highlight the complex decision-making process in selecting models tailored to the intricacies of credit card fraud detection.

Chapter 7

Conclusion and Future Works

The study highlights the effectiveness of various models in detecting credit card fraud, with a focus on the importance of data balancing for improved model performance. The future landscape of fraud prevention is poised for significant advancements, characterized by real-time transaction analysis that proactively identifies and prevents fraudulent activity before it materializes. This approach hinges on the integration of adaptive risk scoring, which tailors risk assessments to individual spending behaviors and the real-time context of transactions. Moreover, as financial transactions and businesses increasingly transcend borders, the complexity of fraud patterns escalates, necessitating a global perspective in fraud detection and prevention strategies. Embracing adaptive and self-learning models is imperative, with systems capable of evolving and learning from new patterns autonomously. The ability to dynamically update in response to emerging fraud techniques without constant human intervention emerges as a crucial asset in the ever-evolving landscape of financial security.

Bibliography

[1] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, “Credit card fraud detection in the era of disruptive technologies: A systematic review,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 145–174, 2023.

[2] R. Bin Sulaiman, V. Schetinin, and P. Sant, “Review of machine learning approach on credit card fraud detection,” *Human-Centric Intelligent Systems*, vol. 2, no. 1-2, pp. 55–68, 2022.

[3] Y. Jain, N. Tiwari, S. Dubey, and S. Jain, “A comparative analysis of various credit card fraud detection techniques,” *International Journal of Recent Technology and Engineering*, vol. 7, no. 5, pp. 402–407, 2019.

[4] D. Wang, B. Chen, and J. Chen, “Credit card fraud detection strategies with consumer incentives,” *Omega*, vol. 88, pp. 179–195, 2019.

[5] S. Kakati, C. Goswami *et al.*, “Factors and motivation of fraud in the corporate sector: A literature review,” *Journal of Commerce & Accounting Research*, vol. 8, no. 3, pp. 86–96, 2019.

[6] O. Ogbanufe and R. Pavur, “Going through the emotions of regret and fear: Revisiting protection motivation for identity theft protection,” *International Journal of Information Management*, vol. 62, p. 102432, 2022.

[7] J. S. Kiernan, “Credit card fraud statistics,” 2023. [Online]. Available: https://wallethub.com/edu/cc/credit-card-fraud-statistics/25725

[8] K. SHENOY, “Credit card transactions fraud detection dataset,” 2020. [Online]. Available: https://www.kaggle.com/datasets/kartik2112/fraud-detection/data