Bryan Rogers
Bellevue University
DSC550: Data Mining
Term Project

Credit Card Fraud Detection

Introduction

In the modern landscape of banking, financial, and credit card transactions, the ability to detect fraudulent transactions on credit cards is becoming more and more important. The main issue that arises with credit card fraud is being able to identify fraudulent transactions from legitimate transactions. Being able to help keep financial institutions and consumers safe from significant loss due to credit card fraud. This project aims to look further into credit card fraud and aims to create, develop, and utilize an advanced model to detect fraudulent transactions in purchases using credit cards using machine learning and data mining techniques.

In solving the credit card fraud problem, the most significant aspect is protecting financial assets and consumer trust for credit card companies. The threat of credit card fraud is substantial to both individual consumers and financial organizations. Credit card fraud leads to financial loss, compromised data security, a loss of trust in financial systems, among other issues. In creating an effective system to detect and eliminate credit card fraud, the risk of credit card transactions can be lessened and ensure trust in financial transactions and institutions.

To inform stakeholders of this problem, the stakeholders would need to learn more about credit card fraud and the impact that it has on financial institutions and consumers. Highlighting key aspects such as potential financial loss, damage to institutions reputations, and the financial regulatory implications, we can show the urgent need to take proactive measures to eliminate credit card fraud. In addition to these aspects, we can utilize this model to show competitive advantages for financial organizations to provide higher levels of security, trust, and protection for consumers.

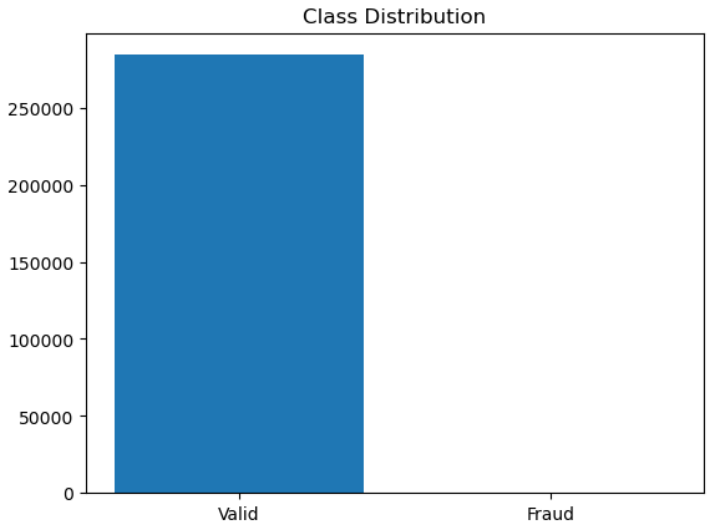
The dataset used to create the model for this project was obtained from Kaggle (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). This dataset contains anonymized credit card transactions labeled as fraudulent or legitimate, along with certain transaction features such as amount, time, and principal components obtained via PCA

transformation due to privacy issues. This dataset contains credit card transactions from European cardholders during the timeframe of two days in September of 2013. During this two-day period, 492 counts of fraud were detected out of 284,807 credit card transactions. While this is an unbalanced dataset, there does seem to show positive signs as there is only a 0.172% chance of fraud in all transactions.

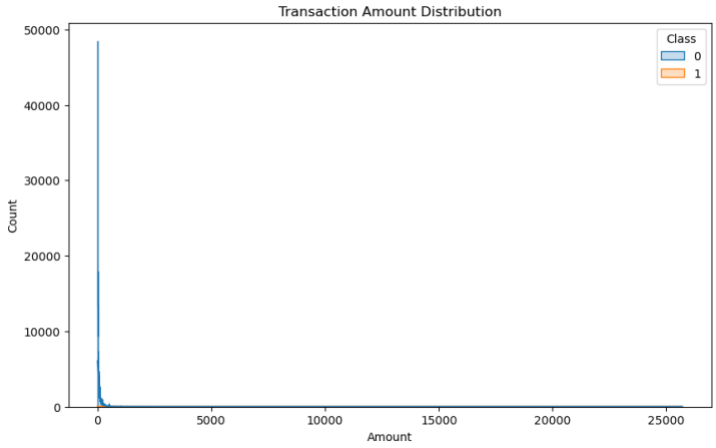
Summary of Data Analysis

Through data exploration and graphical analysis, I have been able to explore a few distributions and correlations of the dataset in order find fraudulent transactions. Through EDA, I was able to establish a distribution of transaction times and amounts. I was able to use this dataset to investigate the imbalance between fraudulent and legitimate transactions. Through graphical analysis, I was able to visualize and detect patterns in transaction amounts and times.

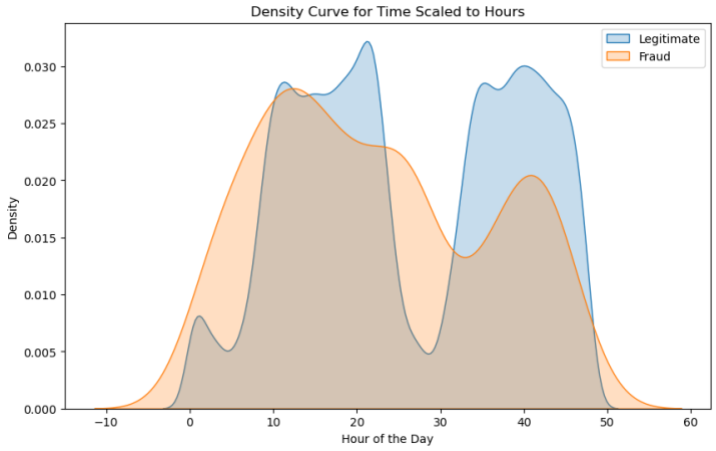
Bar Chart for Fraudulent Transaction



Histogram for Transaction Amount Distribution

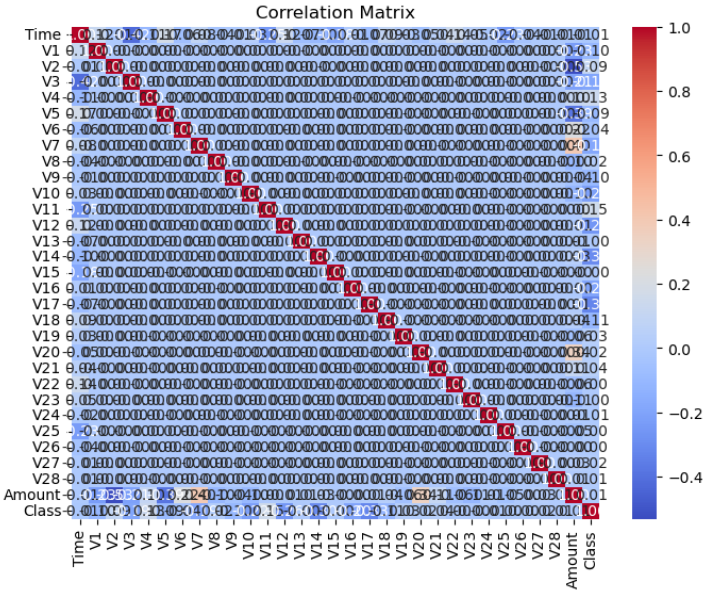


Density Curve for Time Scaled Transactions



Along with using graph to show and visualize the data, I was able to explore and examine the correlation between transaction features to better understand relationships within the dataset.

Correlation Matrix



The graphical analysis above shows valuable insights into detecting fraud in credit card transactions. The bar chart on class distribution highlights the imbalance in the dataset. The correlation matrix can show if there are any strong correlations between the features either positive or negative. Using the histogram and density curve, I can look at transaction time and amount patterns to determine fraud. The graphs provided can be built upon further to create a full, complex model to get the answers I am looking for.

This dataset was already fairly clean when it was received, and does not need much adjusting, I will be making the needed adjustments, as well as showing what if scenarios. In order to get the best understanding of the data and best way to implement a model, some features will need to be adjusted. I dropped the “Time” feature as it was converted to a feature named “Hour” to better grasp the time feature. I also considered dropping the “Amount” feature as it doesn’t have a lot of potential relevance. I also used feature engineering to engineer new features such as transaction frequency and velocity. In addition to engineering new features, I also utilized binning to categorize transactions into different categories. I added to numerical features set to the median for variables that were missing data. I considered adding dummy variable, but this was not necessary since there were no categorical variables to convert.

Model Building

I will be implementing 3 different models to test my dataset and use those 3 to determine what is the best track moving forward. I will be looking at a Random Forest Model, an Isolation Forest Model, and a Neural Networks model. I will generate a classification report and confusion matrix from each model and use those scores to predict the accuracy of each model.

Random Forest Model

```
# Evaluating the model
print("Random Forest Model Evaluation:")
print(classification_report(y_test, y_pred_rf))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred_rf))
print("ROC AUC Score:", roc_auc_score(y_test, y_pred_rf))
```

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

           0       1.00      1.00      1.00     56864
           1       0.97      0.77      0.86        98

    accuracy                           1.00     56962
   macro avg       0.99      0.88      0.93     56962
weighted avg       1.00      1.00      1.00     56962

Confusion Matrix:
[[56862     2]
 [   23    75]]
ROC AUC Score: 0.8826354754056941
```

Isolation Forest Model

```
# Evaluating the model
print("Isolation Forest Model Evaluation:")
print(classification_report(y_test, y_pred_iso))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred_iso))
print("ROC AUC Score:", roc_auc_score(y_test, y_pred_iso))
```

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

           0       1.00      0.96      0.98     56864
           1       0.04      0.83      0.07        98

    accuracy                           0.96     56962
   macro avg       0.52      0.89      0.53     56962
weighted avg       1.00      0.96      0.98     56962

Confusion Matrix:
[[54729  2135]
 [   17    81]]
ROC AUC Score: 0.8944924445580145
```

Neural Network Model

```
print("Neural Network Model Evaluation:")
print(classification_report(y_test, y_pred_nn))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred_nn))
print("ROC AUC Score:", roc_auc_score(y_test, y_pred_nn))
```

```
Neural Network Model Evaluation:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.86      0.74      0.80        98

    accuracy                           1.00     56962
   macro avg       0.93      0.87      0.90     56962
weighted avg       1.00      1.00      1.00     56962

Confusion Matrix:
[[56852    12]
 [   25    73]]
ROC AUC Score: 0.8723434646790624
```

Model Evaluation

After running all 3 different model, I have learned some insights that should help with moving forward. Both the Random Forest and Neural Network produced high accuracy at 100% and all 3 models produced high ROC scores above 87%, showing that they would provide promising predictions for credit card fraud detection. There were certain features that proved to be significant in finding fraudulent transactions, such as transaction amount and time of transaction. Neural Network provided a more complex setup which could make the results harder to understand as opposed to the Random Forest model. There were

important metrics that will provide a more comprehensive understanding of the dataset, such as: Precision, Recall, F1, and AUC-ROC. Overall, looking at the three models that I ran, it looks like the Random Forest model will be the best model moving forward to help identify for credit card fraud.

Conclusion

The analysis and model building helped to provide valuable insights into finding credit card fraud in transactions. Through model building and evaluation, it has been determined that the Random Forest demonstrated promising performance, showing reports of high accuracy and AUC-ROC scores. The Isolation Forest model also proved to show effectiveness in identifying potential fraud in the data. Feature importance analysis showed and highlighted significant fraud predictors such as transaction amount and time.

After running analysis using these models and seeing promising results, I would say that these models are suitable for deployment in the real-world and can showcase positive results in future datasets of similar information. Based on the results of these models, I would recommend moving forward with deploying the Random Forest model for further investigation in fraud detection and model implementation, monitoring systems to ensure continued effectiveness.

This model for determining credit card fraud will face challenges in the future and will also lead to further opportunities for finding fraudulent credit card transactions. Some of these challenges that this project could face, include the ever-changing landscape of fraud patterns in credit card transactions, and facing the continuous need for model updates to keep up with the changing data. Along with these challenges, there are also opportunities in the future for further exploration with the data. These opportunities could include ensemble methods, deep learning architectures and further machine learning techniques, and advanced anomaly detection techniques to enhance the performance of the model and find better ways to detect fraudulent transactions.

In conclusion, this project shows the importance of using machine learning techniques for credit card fraud detection and can highlight the potential for proactive measures when it comes to safeguarding financial systems and protecting consumers from fraudulent transactions and help build trust between the consumer and financial institutions.

Note: Code and Content from Milestones 1-3 have been included with this writeup

References:

ULB, M. L. G.-. (2018, March 23). *Credit Card Fraud Detection*. Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud