Final Report

Uncovering Fraud:
A Machine Learning Model to Predict Credit Card Fraud Transactions

Erin Amoueyan

4/30/2023

1. Introduction

Credit card fraud is a significant concern in the financial industry. In order to prevent fraudulent transactions, a machine learning model can be trained on historical transaction data to identify fraudulent activity in real-time. There are many factors that could indicate a transaction as fraud. Some of the most important factors include unusual activity, large transactions, geographic inconsistency, high-risk merchants, and rapid or unusual spending.

It is important to detect and prevent fraud in credit card transactions for three main reasons including:

- Protecting customers from financial losses, credit damage, and identity theft.
- Preserving financial stability to avoid financial losses.
- Maintaining trust in the financial system and ensures customers feel secure using credit cards.

There are several methods used to detect and prevent credit card fraud such as rule_based systems, machine learning, biometrics, and real_time monitoring. In this report, we will describe the process of developing a machine learning model to predict fraud transactions. With the tuned Random Forest model performed in this study, we are able to predict more than 95% of the fraud transactions correctly and with the real_time detection we will be able to prevent these fraud transactions.

2. Data Wrangling

The first step in developing the machine learning model is to obtain the data. For this project the credit dars history data was obtained from Kaggle, which includes legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants and includes information on the transaction amount, merchant, location, and time.

We have performed data wrangling to clean and prepare the data for analysis. This step included tasks such as removing duplicates, dealing with missing data, identifying and addressing outliers, converting data types, adding new useful features, and removing features that are not needed. The data included 1296675 transactions from which 7506 transactions

were fraud. That equals to %0.58 of the total transactions. The plot of the fraud transactions for each category (Figure 1) shows that most of the fraud transactions are related to shopping_net, misc_net, grocery_pos, and shopping_pos.

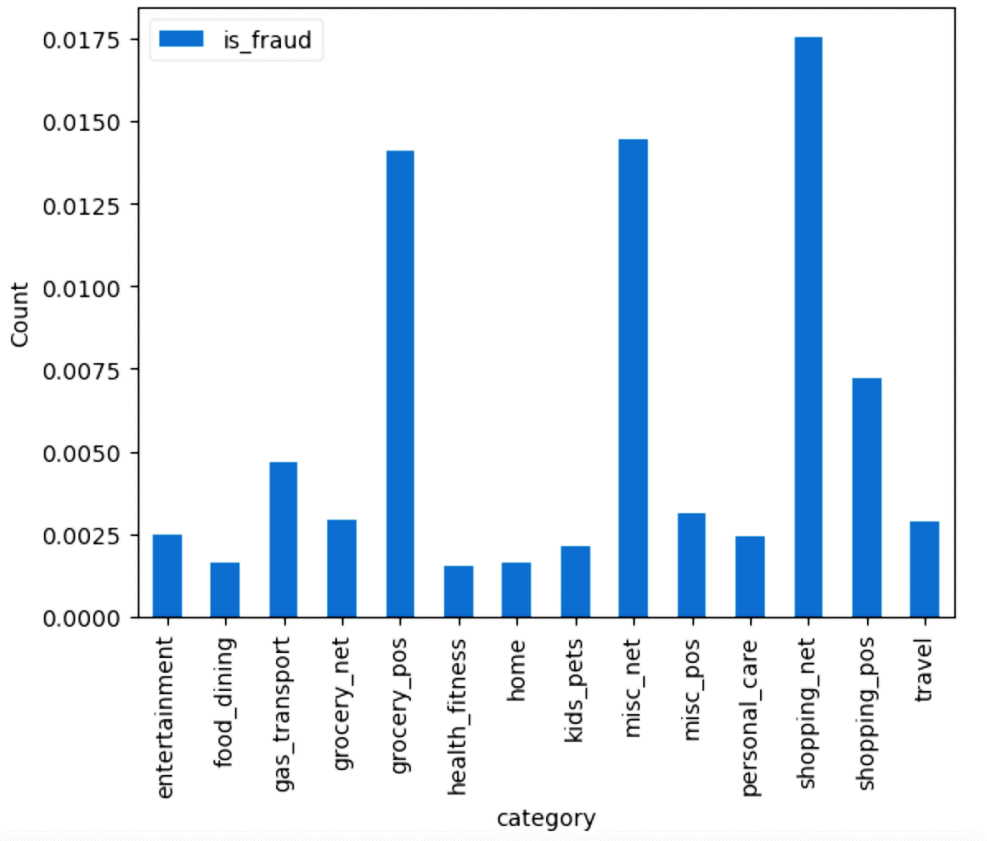

Figure 1- Plot of fraud transactions for each category

Also, the plot of fraud transactions for different transaction amounts (Figure 2) shows that for fraud transactions the mean and standard deviations are about $531 and $390, respectively. While the mean and standard deviations of the legitimate transactions are about $68 and $54, respectively. In general, we can say that the fraud transactions are mostly somewhere between $100 and $1,000 while the legitimate transactions are mostly somewhere between $10 and $100; however, they can go up to $10,000 and more.

Further analysis of the data was done in the next step, Exploratory Data Analysis (EDA).

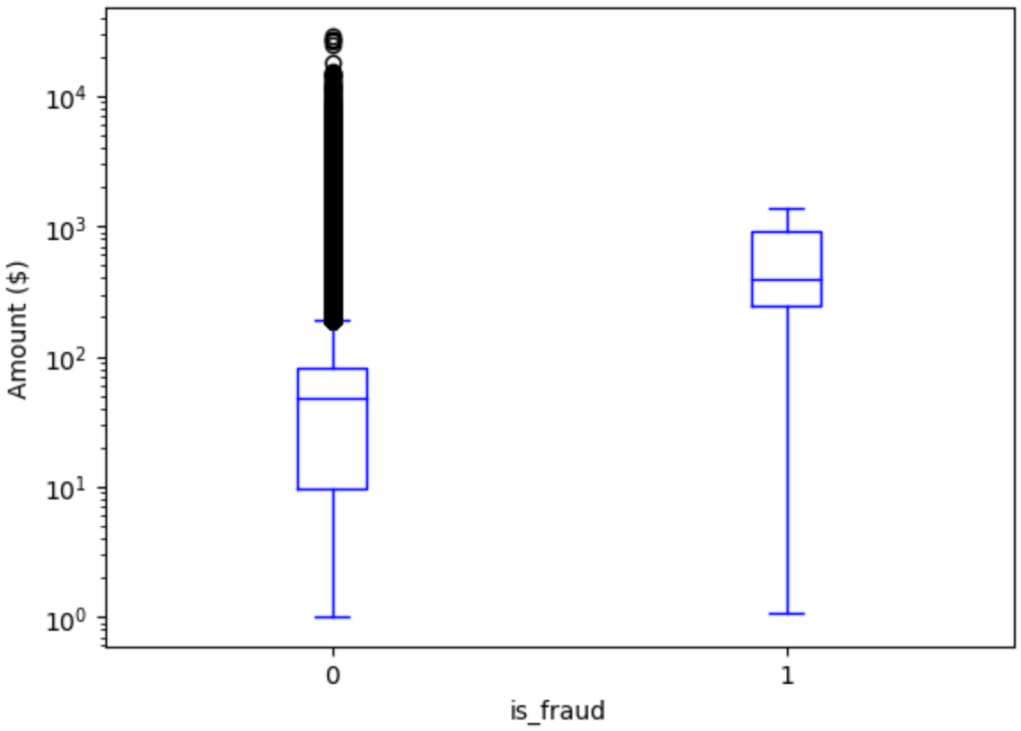

Figure 2- Boxplot of transactions distribution versus distribution amount

3. Exploratory Data Analysis

After the data has been cleaned, we have performed exploratory data analysis (EDA) to better understand the data and identify any patterns or relationships.

In this step, first we have used dummy variables to convert categorical data into numerical data that can be used in our modeling. The final shape of the data was 1296675 rows with 37 columns.

The correlation heatmap figure (Figure 3) shows the relationship between different variables in the credit card transaction dataset. The figure indicates that fraud transactions are most strongly correlated with transaction amount, with a positive correlation coefficient. This suggests that as the transaction amount increases, the likelihood of fraud also increases. There are also less strong, but positive correlations between fraud and other parameters in the dataset, such as day and time of the transactions. This could indicate that these factors are not as strong predictors as transaction amount but could still impact the model prediction. It is also

possible that are other factors that are not captured in the dataset that are more strongly associated with fraud.

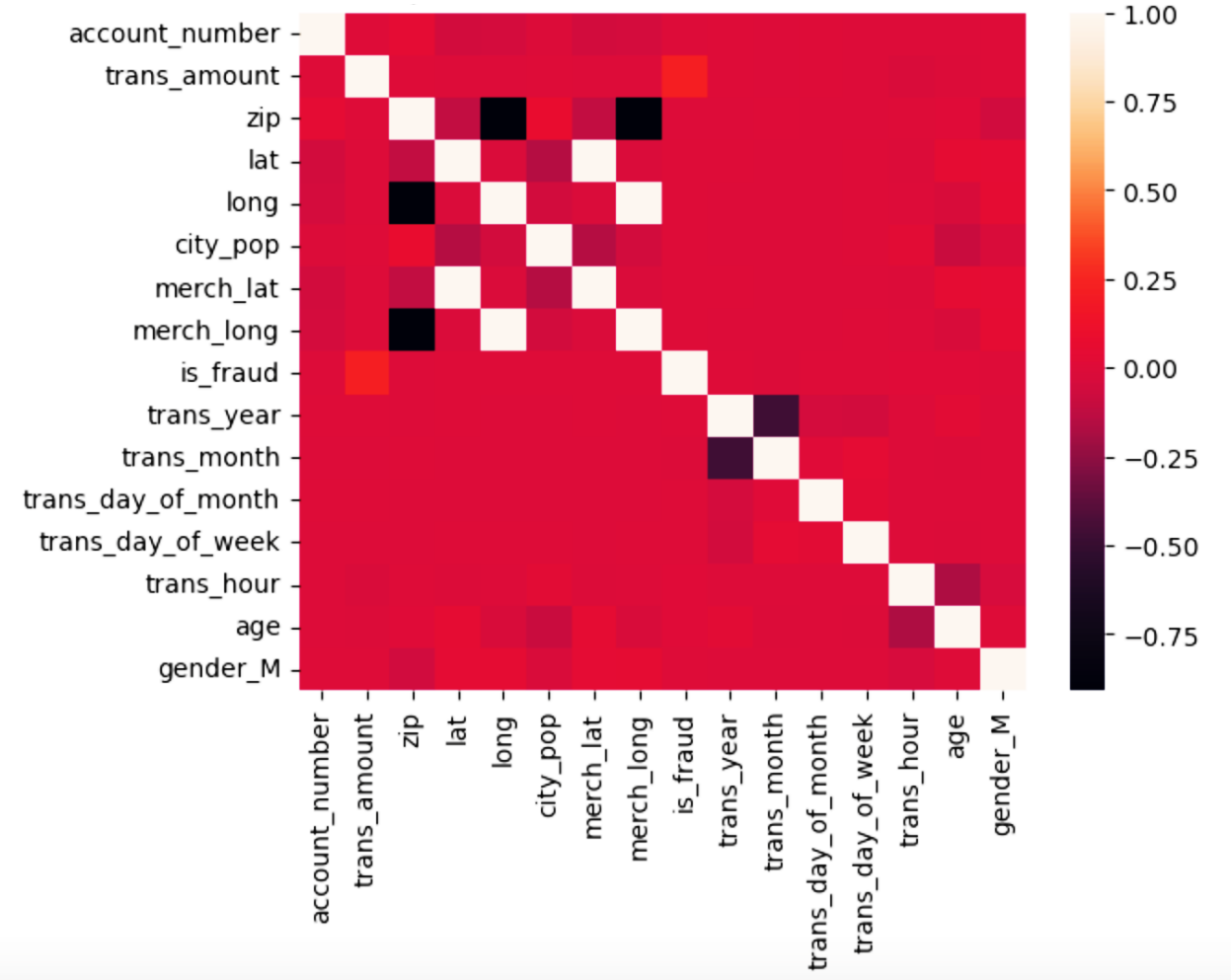

Figure 3- Heatmap of correlations in fraud detection data

Since transaction amount is the most important predictor, a violin plot has been used to better understand the amount distribution over fraud and legitimate transactions. Figure 4, shows the distribution of transaction amounts for each group, with the width of the "violin" representing the density of data points at different values. The plot indicates that in general, fraud transactions tend to have higher transaction amounts than legitimate transactions, with higher median and broader range of values around the median and also around the upper tail. This suggests that fraudulent activity may be more likely to involve high-value transactions. The plot also shows that legitimate transactions have a narrower distribution of values, however there are some values around higher transaction amounts. This could indicate that there are certain

thresholds or limits in place for legitimate transactions that are not present in fraudulent activity. Performing a t-test on the amount feature resulted in p-value of zero which suggested that this feature is significant in our prediction model.

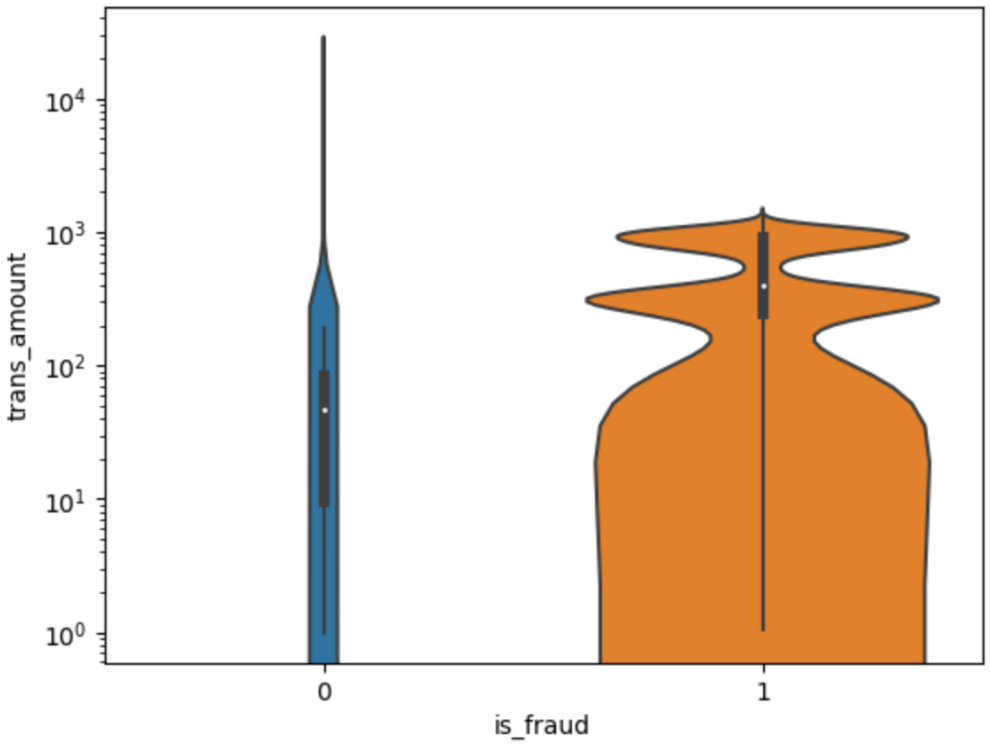

Figure 4- Distribution of transaction amounts for both fraudulent and legitimate transactions

Next, we looked into the relationship between fraud and other features from which a pattern could be found with transaction day and hour (Figure 5 and 6).

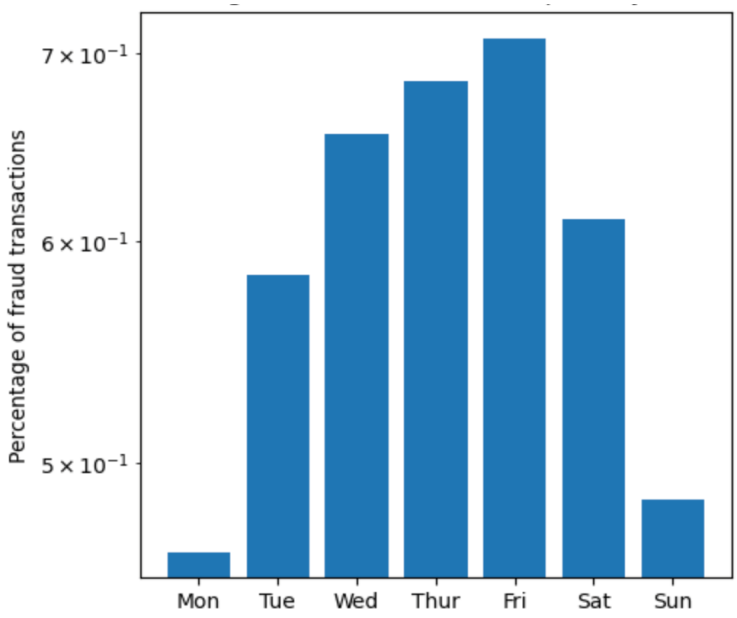

Figure 5- Percentage of fraud transactions per day of week

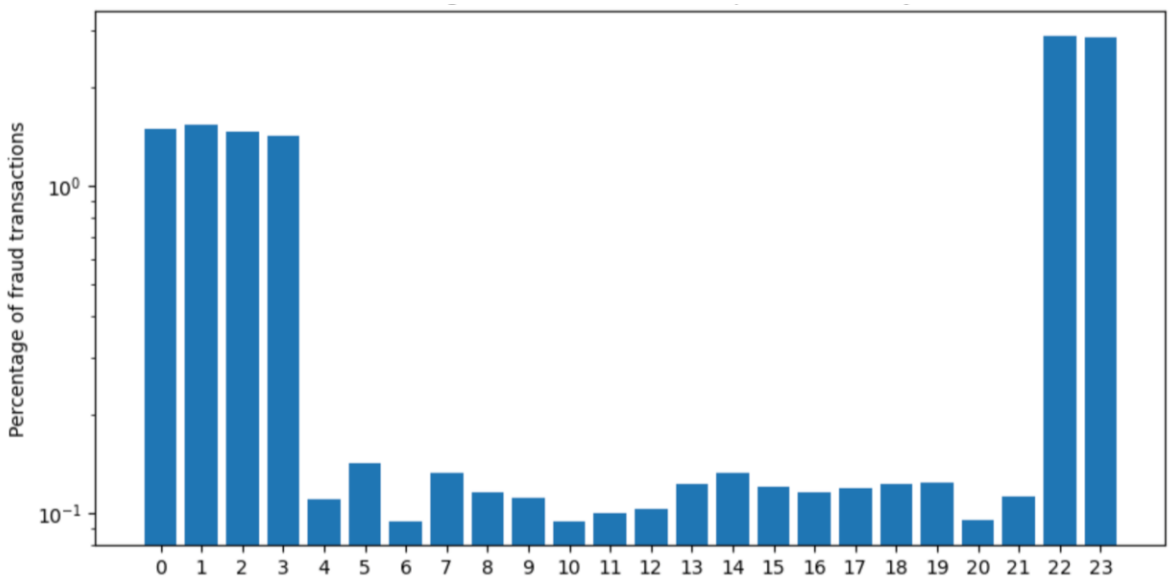

Figure 6- percentage of fraud transactions per hour of day

Figure 5 indicates that the proportion of fraud transactions are the highest on Friday, followed by Thursday and Wednesday. Monday had the least percentage of fraud transactions. Also, Figure 6 indicates that the percentage of fraud transactions are the highest at late night and early morning (from 10 pm to 3 am).

After performing appropriate statistical tests on different features, it is obvious that transaction amount (`amt`), transaction hour (`trans_hour`), customer age (`age`), and transaction month (`trans_month`) are the most significant features. Transaction day of month (`trans_day_of_month`) and day of week (`trans_day_of_week`) are also significant but at a lower level.

4. Preprocessing and Training

The next step is to preprocess the data in order to prepare it for machine learning algorithms. This will involve tasks such as scaling the data and splitting the data into training and testing sets.

Then, we trained several machine learning models such as decision tree classifier, random forest, gradient boosting, hist gradient boosting, Adaboost classifier, naive bayes, and K nearest neighbor on the training data and evaluated their performance on the testing data. Various evaluation metrics such as precision, recall, and F1 score have been used to compare the performance of different models.

The results of the initial modeling showed that almost all of the models had a high accuracy of about 99%. However, this is not a good metric to select the best model since the data is imbalanced and the number of class 0 transactions (no fraud) is much higher than class 1 (fraud transactions). Therefore, we have a small positive class. We would like to avoid false negatives since we don't want to classify a fraud transaction as a legitimate transaction. Therefore, the metrics we should care about the most are F1 score and Recall score and also precision score. Based on these metrics, among all of the models, the random forest model was the best model, when predicting training data, with more than 99.9% accuracy, recall and f1 scores. Also, the precision score is 1 therefore, it can perfectly predict the legitimate transactions and almost all fraud transactions. Only 2 of the fraud transactions are misclassified as legitimate which is only about 0.035% of the fraud transactions. This is so far the best model that can predict the train data. The confusion matrix for this model is shown in Figure 7.

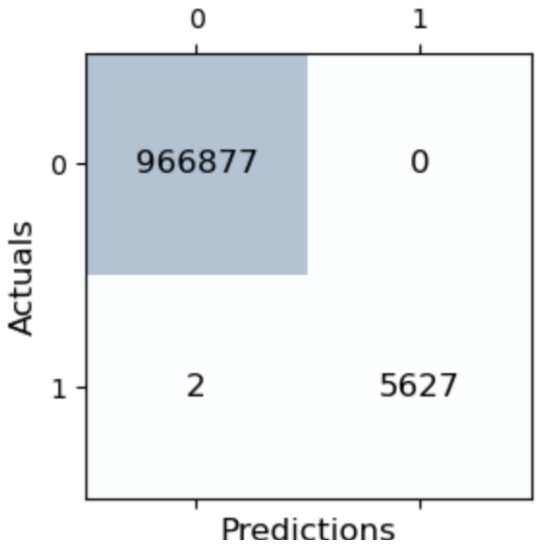

Figure 7- Confusion matrix of fitting random forest model on train data

5. Final Modeling

Once we have identified the best-performing machine learning model, we have fine-tuned it using techniques such as hyperparameter tuning and cross-validation. This helped us to improve the model's performance and reduce the risk of overfitting. The hyperparameter tuning was performed by using below ranges with cross-validation fold of 5:

- max_depth : np.arange(2, 10)
- Criterion : ['gini', 'entropy']
- class_weight : ['balanced']

The result of the hyperparameter tuning is shown in Figure 8. Recall score was used as the best scoring matrix. Az illustrated in Figure 8, both gini and entropy models perform pretty similarly with the gini model performing slightly better. Therefore, the Random Forest with max_depth of 9 and criterion gini was used as our best model.

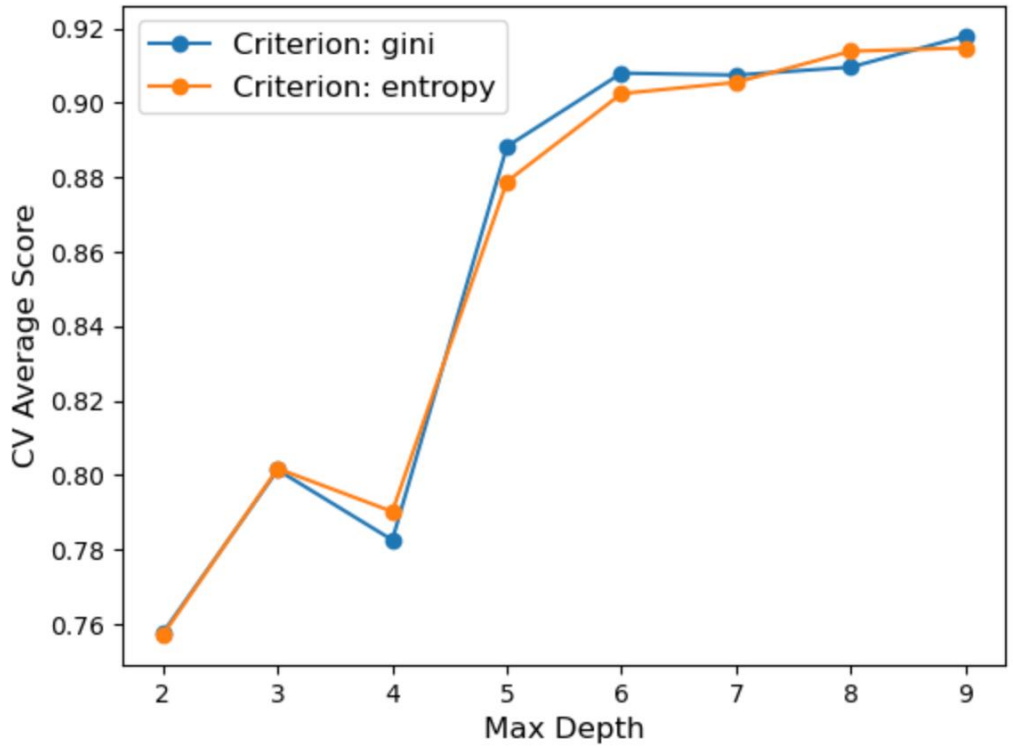

Figure 8- Results of the hyperparameter tuning on random forest model using cross-validation

Then, we evaluated the performance of this model on the test data. The results of the confusion matrix for test data is shown in Figure 9. The results showed that the model could accurately classify 90% of the fraud transactions in the unseen data. Even though some false positives could be seen in the results, the false positive rate is only about 1.9% which can be negligible since in this case we care more about the false negatives and would like to minimize the number of fraud transactions that are mistakenly classified as legitimate transactions. Here, about 10% of the fraud transactions are mistakenly classified as valid transactions. The model can be further improved by doing more feature engineering, for example adding more features to calculate the time difference between the transactions and the distance difference using latitude and longitude of the merchant.

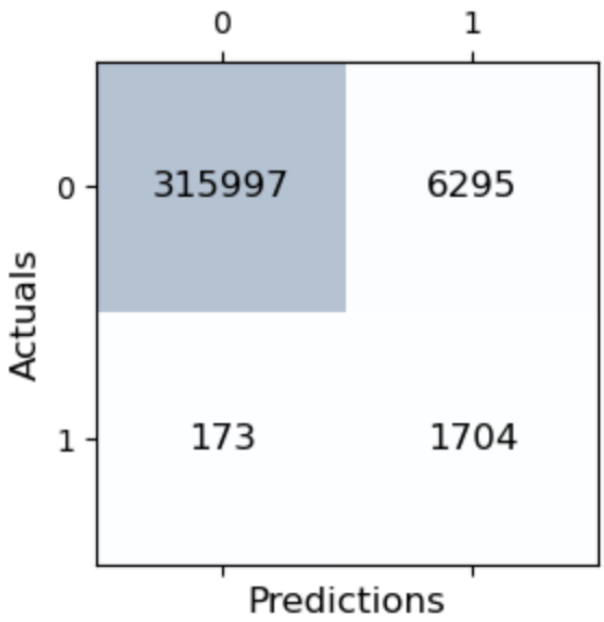

Figure 9- Confusion matrix of fitting random forest model on test data

6. Conclusion

In conclusion, we have explored different machine learning models for credit card fraud detection and found that the random forest classifier with max depth of 9 and gini criterion performed the best. This model was able to predict 90% of the fraud transactions with a relatively low false positive rate of 1.9%, meaning that only 1.9% of legitimate transactions were mistakenly classified as fraud. Overall, the model had an impressive accuracy rate of 98%. These results demonstrate the potential of machine learning algorithms in detecting and preventing credit card fraud. However, it is important to note that fraudsters may adapt their strategies over time, so ongoing monitoring and updates to the model may be necessary to maintain its effectiveness. Nonetheless, the success of our model provides a promising approach for credit card companies to proactively detect and prevent fraudulent activity, protecting both the company and its customers.

7. Future Research

While our current machine learning model for credit card fraud detection has shown promising results, there is still room for improvement. In particular, future research could focus on incorporating additional features and doing more feature engineering to better capture the complex patterns of fraudulent activity. For example, variables such as the location of the

transaction, the IP address of the device used to make the transaction, and the type of card used could be included as additional predictors. Additionally, feature engineering techniques such as principal component analysis or clustering could be used to create new variables that may improve the model's accuracy. Moreover, it would be beneficial to evaluate other machine learning models to determine if there are better-performing models than the random forest classifier we used. This could include models such as neural networks, which may be able to capture more complex interactions between variables. Overall, future research can build upon our current model to improve the accuracy and effectiveness of credit card fraud detection systems.