Credit Card Fraud Detection

Jade Bell

1. Introduction

Credit card fraud occurs when an unauthorized individual gains access to another person's information by making purchases and requesting cash advances without consent. An unauthorized individual can gain your information through stolen or lost cards, scam techniques, and hacking devices across a network.

The motive question with this problem is "How can we track transactions to decrease the risk of credit card fraud?" A method we can imply as a possible solution is credit card fraud detection. Credit card fraud detection is the process of identifying purchase attempts that are fraudulent. With the use of detection, the number of frauds can decrease. In general, we need to design a detection model that tracks the patterns of abnormal purchases.

2. Machine Learning Algorithm(s) Description

A dataset from Kaggle was used to fit into a detection model. The dataset provides card transactions made in September 2013 by European cardholders the occurrence of two days. The libraries used for gathering the data were numpy, pandas, and matplotlib. The main steps for card detection with the Kaggle dataset include:

1. Classify credit card transactions as legit/fraudulent.
2. Using detection methods to identify transactions from legit to fraud.
3. Split data with a training percentage of 80% and test with 20%.
4. Discover which learning model performs best.

As the data was gathered, I noticed the data was unbalanced. Unbalanced data is when one class of data is recognized more than the others. The data alone can only detect 17% of fraudulent activity as shown in the photo below. We cannot fit this data into models because it will be difficult to recognize fraudulent transactions.

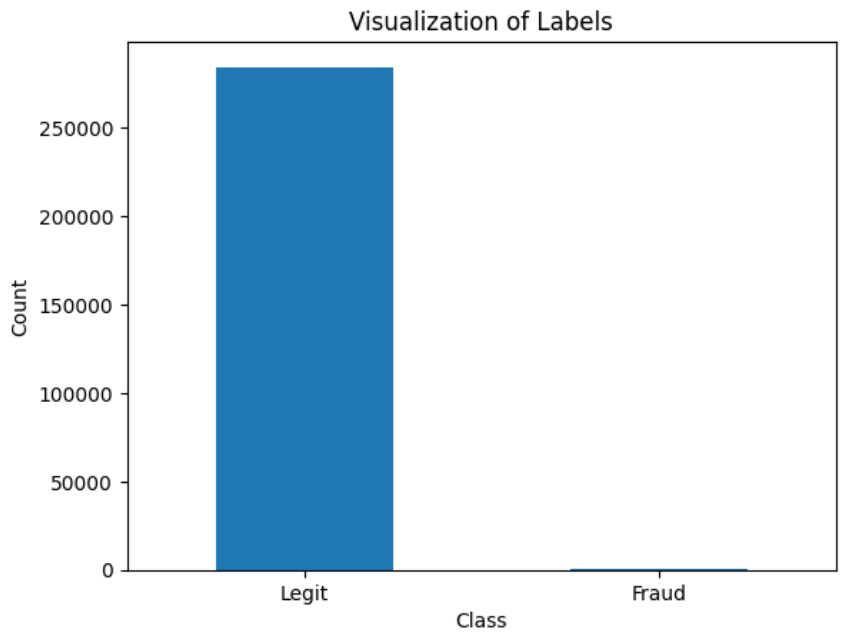

Figure 1. Unbalanced Data Chart

This highly unbalanced data will lead to machine learning algorithms performing poorly, making errors, and not detecting these fraudulent transactions. Modifications made to balance this dataset were to change the shape and number of the legit transactions to equal fraud transactions. Therefore, random samples of the legit transactions were chosen. Then, the new sample of legit transactions was concatenated with fraudulent transactions to create a new data frame.

```
▾ Change shape and number of legit transactions to equal fraud

[ ]  legit_sample = legit.sample(n=492)

Concatenating Two DataFrames

[ ]  new_ccd = pd.concat([legit_sample, fraud], axis=0)

[ ]  new_ccd.head()
```

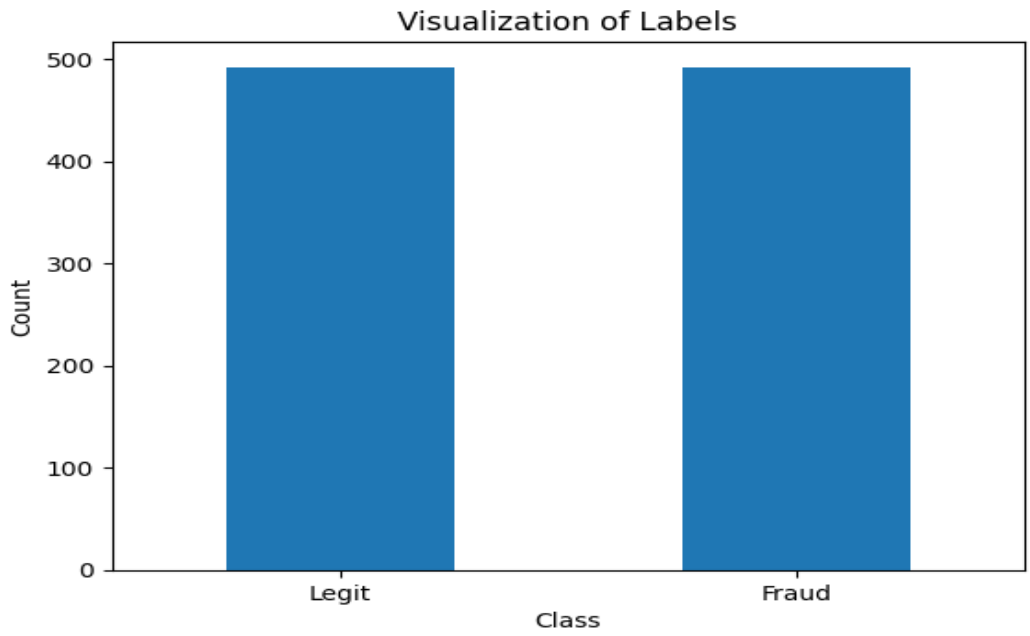

Figure 2. Balanced Chart

Machine learning algorithms used to detect fraudulent behavior: Logistic Regression, SVM, Decision Tree, RBF, and Random Forest. The logistic regression model, for example, was split into 80% train and 20% test data. Once fitted into the model, its accuracy for the training set was 93.39% and with a score of 92.38% with the test set. For a better accuracy performance, all of the models were fit into cross-validation. Because this project uses more than one model for accuracy, it is important to implement cross-validation to check their performance, which shows a higher score of prediction.

3. Results

After the data were fit into each model, and tested with cross-validation, its results are as shown in the graph. The graph shows the training and test accuracy, cross-validation means percentage, and cross-validation's highest score(standard deviation + mean). The algorithm with the highest percentage shown in the graph is the random forest model. Even though some models obtained the same accuracy score for a test run, the model with the best performance in terms of cross-validation high score is Random Forest with 94.69%. To further confirm this accuracy score, each algorithm was shaped into a confusion matrix. The confusion matrix shows a table representation of the outcomes and the results of the classification task. The figures below show the confusion matrix of the logistic and RNF model test set.

Model Name	**Training Set Acc**	**Test Set Acc**	**CV Mean**	**CV Highest Score(m+std)**
Logistic R	93.77%	92.38%	92.75%	93.38%

SVM	95.04%	91.87%	93.77%	94.40%
RBF	95.55%	92.38%	93.64%	94.64%
Decision Tree	97.83%	91.37%	89.83%	91.65%
RNF	97.45%	92.38%	93.64%	94.69%

Figure 3. Machine Learning Al Accuracy Table

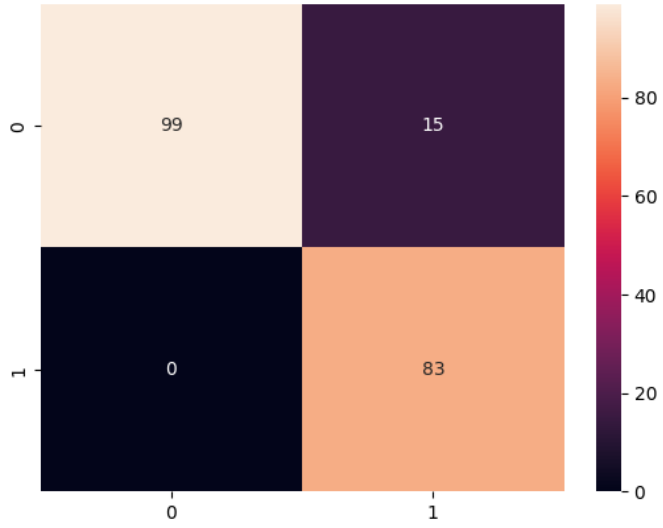

Figure 5. LR Test Confusion Matrix

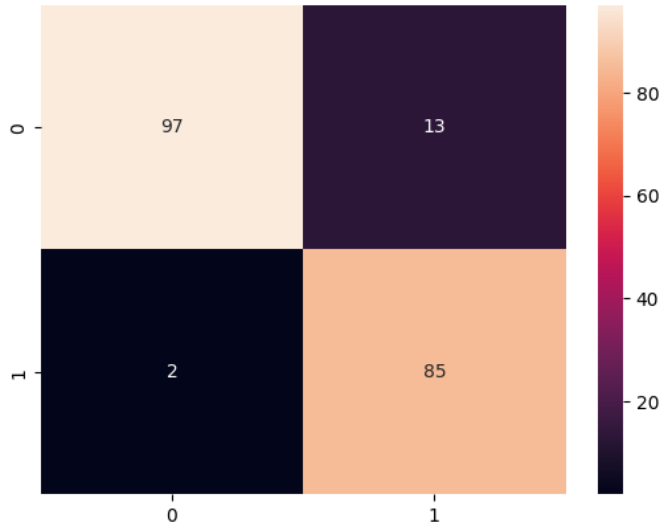

Figure 6. RNF Test Confusion Matrix

4. **Discussion and Conclusion**

It is important to decipher which model best performs well in order to people to utilize them. However, what is the importance of credit card fraud detection? Fraud detection and prevention stop scammers from stealing your personal information and making outside purchases. Another reason why implementing fraud detection is important is that companies are able to inspect fraudulent credit card transactions so that customers are not charged with items not purposely purchased. These fraud detection techniques & models help merchants identify whether a purchase is legitimate.

Reference

Inscribe. "Credit Card Fraud Detection: Everything You Need To Know." *Fraudulent Document Detection & Automation*, Inscribe, 11 Apr. 2023, https://www.inscribe.ai/fraud-detection/credit-fraud-detection#:~:text=Credit%20card%20fraud%20detection%20is,fraud%20and%20stop%20fraudulent%20transactions.

Saxena, Pranjal. "Credit Card Fraud Detection Using Machine Learning & Python." *Medium*, Towards Data Science, 13 Sept. 2022, https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc.

ULB, Machine Learning Group -. "Credit Card Fraud Detection." *Kaggle*, 23 Mar. 2018, https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

White, Alexandria. "Here's How Credit Card Fraud Happens and Tips to Protect Yourself." *CNBC*, CNBC, 27 May 2022, https://www.cnbc.com/select/credit-card-fraud/#:~:text=Credit%20card%20fraud%20occurs%20when,Hacking%20your%20computer.