COMP 562 Final Project: Credit Card Fraud Detection

Saurya Acharya, Jahnavi Alapati, Saketh Devareddy,
Maya Krishnamoorthy, Sheel Patel

GitHub Repository: 'https://github.com/Saurya-Acharya/COMP562-Final'

Introduction

Credit card fraud has become an increasingly significant challenge for both consumers and financial institutions. In fact, credit card fraud cost US retailers $32 billion in 2014, suggesting a pressing need for fraud prevention to keep up with today's digital age [1]. Therefore, it is important to be able to detect fraudulent transactions accurately [4].

As technology has become more advanced and the primary method of moving money, this project tries to provide a dynamic approach to recognize patterns indicative of fraudulent activity [3]. For consumers, this includes identifying unauthorized transactions that are a breach of personal security. For businesses, identifying unauthorized transactions can help prevent legal issues, damage to reputation, and even financial loss.

In summary, this project aims to address credit card fraud detection using logistic regression and random forest classification models to predict whether a transaction is fraudulent or legitimate.

Dataset Overview

In our attempt to construct a machine learning model for credit card fraud detection, we found dataset available on Kaggle. The 'Credit Card Fraud Detection' dataset encompasses transaction data from credit cards used by European cardholders in September 2013.

This dataset contains transactions over the span of two days and shows 492 instances of fraud out of a total of 284,807 transactions. Due to confidentiality issues, the numerical input variables were derived from a Principal Component Analysis (PCA) transformation. The PCA-transformed features are labeled V1 through V28 and describe the basic transaction characteristics while masking the secret information. Because of the size of the file, we removed the columns V1-V28 when uploading the file to GitHub. However, the original file can be found on Kaggle.

There are also two non-transformed features: 'Time,' which represents the seconds elapsed between each transaction and the first transaction in the dataset, and 'Amount,' which represents the transaction value. The regression response variable, 'Class,' identifies a transaction as fraudulent (assigned value 1) or legitimate (assigned value 0).

Challenges in Analysis

In typical credit card transaction datasets, fraudulent transactions are usually outnumbered by legitimate ones. In fact, real transactions outnumber fraudulent transactions in our dataset with a ratio 577:1. This imbalance makes it more difficult for machine learning models to distinguish between fraudulent and non-fraudulent data. Moreover, credit card fraudsters are constantly adapting their strategies to get past fraud detection systems. This means that the machine learning models also need to change rapidly to remain effective.

Methods Performed

A critical step in our analysis was addressing the issue of an unbalanced dataset. Because our dataset consisted of a majority of real transactions and less than 1% of fraudulent transactions, we subsampled the number of real transactions to match the number of balanced ones. This allowed us to have a more balanced training data set that made a more valuable model.

With the new balanced dataset, we proceeded to train the models using two machine learning techniques: Random Forest Classifier and Logistic Regression.

Random Forest Classifiers operate by constructing multiple decision trees. This is what helps it capture the complex interactions between features, which often happens in credit card fraud detection scenarios, where the relationship between variables can be fairly intricate. Furthermore, due to its incorporation of randomness in attribute selection and bootstrapped samples, it contributes to increased accuracy against noise [6] and is less likely to overfit the training data. The model must generalize well because fraudulent activities come in a variety of forms. In our subsample, we examine over 25 characteristics, so this model was best suited to handle this example.

Logistic regression, whether applied in a binomial or multinomial context, is a reliable statistical method adept at predicting outcomes with two or more possible values [2]. Thus, in the context of credit card fraud detection, where the target field involves binary classification (fraudulent or non-fraudulent transactions), logistic regression is particularly relevant to build a robust credit card fraud detection algorithm [5]. To serve as an efficient baseline model, we believed that logistic regression would provide a reasonable probabilistic interpretation of the results.

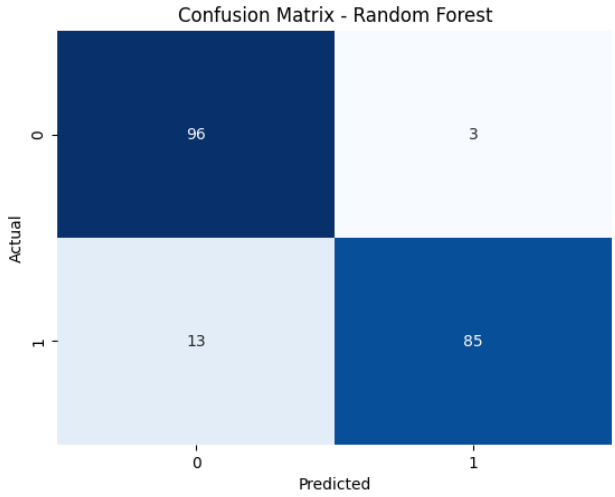

Figure 1: The confusion matrix for the random forest model implementation.

Our analysis of credit card fraud detection using our two models yielded insightful results. The Random Forest Algorithm resulted in an accuracy of 91.88%. Further, the classification report for the random forest algorithm gave us more information about how our model handles precision and accuracy. The confusion matrix revealed that the model correctly identified 96 instances of true negatives (meaning these transactions were rightfully classified as non-fraudulent). However, there were 3 instances where the model incorrectly predicted fraud, or false positives, and 13 instances of false negatives, or where fraudulent transactions were missed. Finally, there were 85 instances of true positives, or where the model successfully detected actual cases of fraud. The 13 false negatives in our model may be particularly concerning for both consumers and businesses in a fraud detection scenario, because that means 13 fraudulent transactions were not identified.

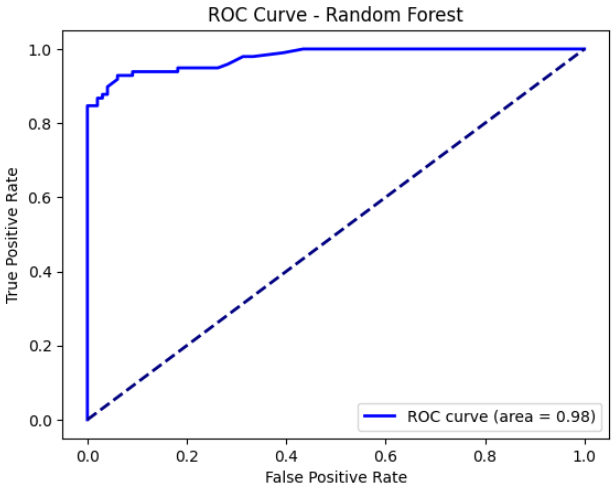

Figure 2: The confusion matrix for the random forest model implementation.

The Receiver Operation Characteristic (ROC) curve for the Random Forest model illustrates the ability of a binary classifier system. The y-axis measures the true-positive rate, or the sensitivity, and measures the proportion of actual positives that are correctly identified. For our model, that was the ability to catch fraudulent transactions. The x-axis measures the false-positive rate which measures the proportion of the negatives that were incorrectly identified as positives. In other words, the non-fraudulent transactions are wrongly classified as fraud.

The closer the ROC curve is to the left-hand border, the more accurate the test. The area under the curve (AUC) is the measure of the model's performance. The AUC for this model is 0.98. As it is very close to 1, this means that there is a very high level of distinguishability and that the model does a great job at differentiating between fraudulent and non-fraudulent transactions. The AUC tells us that there is a high true positive rate and a low false positive rate, which is the ideal scenario.

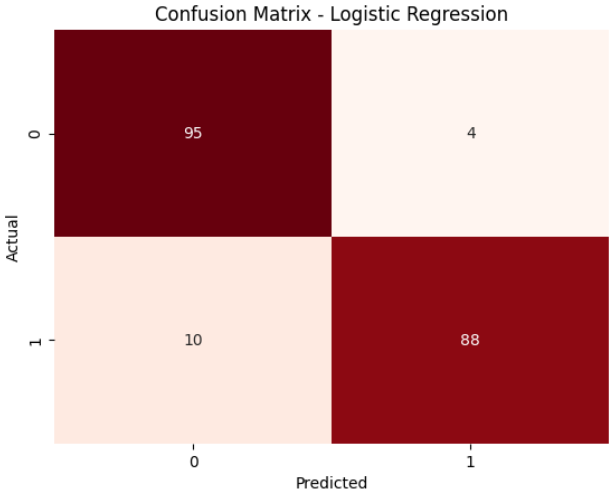

Figure 3: The confusion matrix for the logistic regression model implementation.

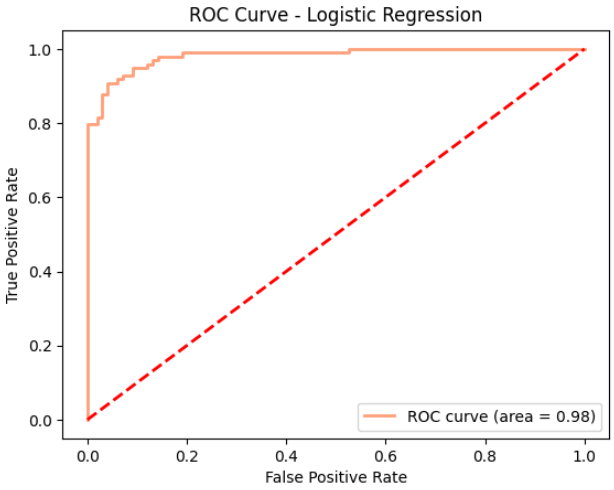

Figure 4: The ROC curve for the logistic regression model implementation.

Our Logistic Regression model yielded very similar results, with an overall accuracy on test data of 92.89%. There were 3 less false negative resulting from this model, but there was also 1 more false positive than the Random Forest model. We cannot determine if the sensitivity of this model is better than that using the random forest algorithm unless we test with more data. The ROC curve for the Logistic Regression model also yielded an AUC of 0.98, which is ideal again because it is close to 1.

Conclusion

In conclusion, both the Random Forest and Logistic Regression models show high levels of accuracy in detecting credit card fraud. Both models showed accuracy between 92% and 94%, and the AUC for both were close to 1, which is ideal.

It was particularly important that we addressed the need to balance the data set. Without this step, it would have been extremely difficult for the model to predict a fraudulent transaction.

In the future, we would try to reduce the number of false negatives. Given our goal was to accurately determine fraudulent transactions, false negatives do not help with catching those cases. Credit card fraud remains a hard-to-attack obstacle, but reducing instances of false negatives will ensure fraud detection systems are reliable and trustworthy.

Furthermore, if given more data, we would have tested our models with the additional data (especially fraudulent data) to better improve the accuracy of our model.

We believe that the work done in this project is an important step in detecting credit card fraud, and we hope to expand upon this project in the future.

Citations

1. Heggestuen, J. (2015, March 16). *THE PAYMENTS SECURITY REPORT: New security protocols aim to close the massive hole in online and in-store credit-card security.* Insider. https://www.businessinsider.com/payments-companies-close-the-massive-hole-in-payments-security-2015-3

2. Hussein, A., & Khairy, R. *Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression.* Babylon, Babylon, Iraq. https://doi.org/10.3991/ijim.v15i05.17173

3. Morgan, K. (2023, September 15). *What is credit card fraud?* Money.
   https://money.com/what-is-credit-card-fraud/

4. Pierre, S. (2023, June 14). *Getting started with credit card fraud detection.* DataDrivenInvestor. https://medium.datadriveninvestor.com/getting-started-with-credit-card-fraud-detection-9817f2fb3326

5. Sahin, Y., & Duman, E. (2011). *Detecting credit card fraud by ANN and logistic regression.* In *2011 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 315-319). Istanbul, Turkey. https://doi.org/10.1109/INISTA.2011.5946108

6. Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March 1). *Random forest for credit card fraud detection.* IEEE Xplore. https://doi.org/10.1109/ICNSC.2018.8361343