Credit Card Fraud

[1]Lili Balazs, [2]Brigitte Reyes, [3]Alejandro Gomez
CS 577
San Diego State University
[1]lbalazs2956@sdsu.edu, [2]breyes4059@sdsu.edu, [3]ajgomez@sdsu.edu

***Abstract*—Since credit cards continue to be one of the most common forms of payment, credit card users can be highly susceptible to becoming victims of fraud. Highly valuable and sensitive information is within credit card data and when exposed, one's credit score can greatly suffer if not caught and reported within a timely manner. In this report, we will use different machine learning models and algorithms to detect credit card fraud. These techniques will be implemented on a dataset that contains information about credit card transactions. We will use the Logistic Regression, Random Forest, and K-Nearest Neighbor models to predict credit card fraud. Our findings conclude that the Random Forest was the most effective and accurate in the prediction of credit card fraud.**

I. Introduction

Nowadays it is hard to imagine a life without the comfort of using credit cards as a method of payment. They are a far more convenient and faster way to make payments for both in-person and online settings.

However, we need to consider the security aspect of using such a form of payment and we need to evaluate the possibility of being victims of credit card fraud.

Credit card data is highly sensitive information, when exposed it can cause one's credit score to suffer if they fail to catch and report it within a timely manner.

47% of Americans have been victims of credit card fraud in the past 5 years. To resolve this issue, we can use different machine learning models and algorithms to detect fraud.

In this research project, our goal was to investigate and compare different machine-learning models that can be used to detect credit card fraud. Additionally, we would implement one of those techniques on a dataset that contains information about credit card transactions.

Using the dataset we used three different models to predict credit card fraud and compared them to see which would be superior in accuracy and efficacy. The models we used were the Logistic Regression, Random Forest, and K-Nearest Neighbor model.

After evaluating our data, we were able to conclude that the Random Forest model showed to be the most accurate and efficient model.

II. Approach

To solve this problem, we imported a dataset from kaggle that contains information about previous credit card transactions made by European credit card users in 2013. This dataset contains only numerical values and has a total of 284,807 transactions, where 492 of them are fraud.

We then examined the content of the data and took into account that this dataset contains an imbalance. Only 0.17% of the dataset are fraudulent transactions and 99.83% of the data comes from non-fraudulent transactions.

The next step to our approach was the Exploratory Data Analysis portion where we checked the distribution of our data. We compared the amount of both fraudulent transactions and non-fraudulent transactions to different parameters such as factors of time. Here we checked for correlation between the two variables.

Our last step to our approach was Data Analytics where we used the Logistic Regression, Random Forest, and K-Nearest Neighbor models to predict credit card fraud. We did many calculations to discover which was the superior model in this prediction.

III. Evaluation

Our goal was to compare different models and evaluate which would be the best model to predict credit card fraud in our dataset.

The three main metrics we would be tracking are Fraud Rate, Incoming Pressure, and Precision.

Fraud Rate: the number of cases of known fraud relative to overall sales.

Incoming Pressure: The fraud KPI incoming pressure is expressed as a percentage and refers to the number of transactions attempted which were later proven to be fraudulent.

Precision: Expressed as a percentage, precision reflects the number of fraudulent transactions proportional to the total number of declined transactions.

We applied this to our three chosen models which were the Logistic Regression, Random Forest, and K-Nearest Neighbor. We used the Confusion Matrix for each of these

algorithms to define the performance of each model. Here is how it works:

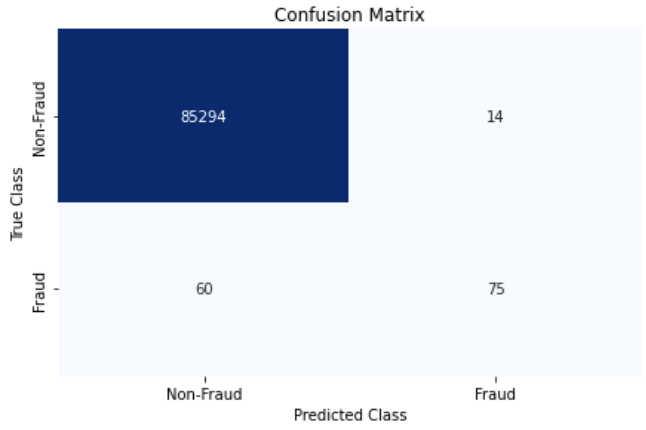
Upper Left Square: Amount correctly classified by the model of non-fraud transactions.

Upper Right Square: Amount incorrectly classified transactions as fraud, but is actually non-fraud.

Lower Left Square: Amount of incorrectly classified transactions as non-fraud, but is actually fraud.

Lower Right Square: Amount of correctly classified by the model of fraud transactions.

The first model we will evaluate is the Logistic Regression Model. Here is the corresponding Confusion Matrix:



You can see here that a total of 74 transactions were classified incorrectly.

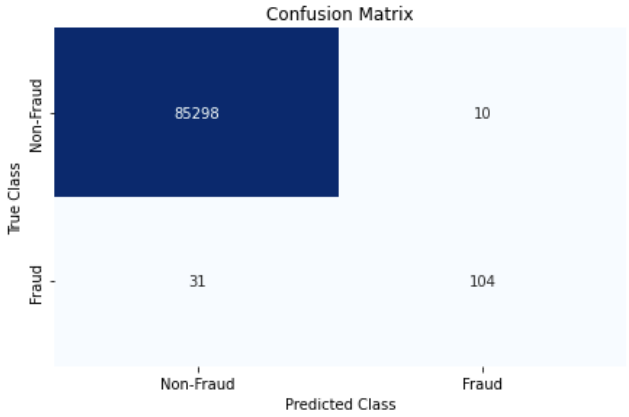
We then calculated the Mean Absolute Error, Mean Squared Error, and Root Mean Square Error:

Mean Absolute Error: 0.0008660744589960559

Mean Squared Error: 0.0008660744589960559

Root Mean Square Error: 0.02942914302177445

The next model we will evaluate is the Random Forest Model. Here is the corresponding Confusion Matrix:



You can see here that there were a total of 41 transactions classified incorrectly.

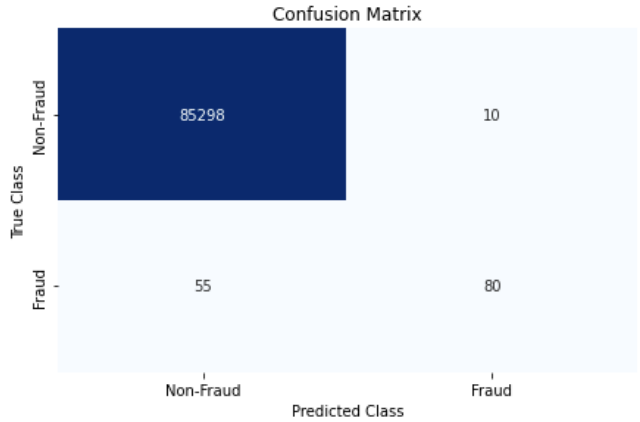
The Mean Absolute Error, Mean Squared Error, and Root Mean Square Error are as follows:

Mean Absolute Error: 0.00047985206511943633

Mean Squared Error: 0.00047985206511943633

Root Mean Square Error: 0.02190552590374028

Lastly, we will evaluate the K-Nearest Neighbor Model. Here is the corresponding Confusion Matrix:



You can see here that a total of 65 transactions were classified incorrectly.

The Mean Absolute Error, Mean Squared Error, and Root Mean Square Error are as follows:

Mean Absolute Error: 0.0007607410788478869

Mean Squared Error:0.0007607410788478869

Root Mean Square Error: 0.02758153510680446

Based on our results, we can compare the three models that were used and see that the Random Forest Model was

the most accurate and effective in predicting credit card fraud.

In this model, only 41 transactions were classified incorrectly, compared to 74 transactions in the Logistic Regression Model and 65 transactions in the K-Nearest Neighbor Model. Additionally, compared to the two other models, the Random Forest Model has the lowest error rates in all three calculations.

The K-Nearest Neighbor Model came second in accuracy with only 65 transactions incorrectly classified. As well as the second lowest error rates in all three calculations.

The Logistic Regression Model comes last in accuracy and efficacy as it had 74 incorrectly classified transactions. As well as the highest error rates in all three of the calculations.

IV. Related Work

The three models that we used are Logistic Regression, Random Forrest and K-Nearest Neighbor Models. But each of the models had its own advantages and disadvantages.

Logistic Regression

Advantages: Logistic regression is easier to implement, interpret, and very efficient to train, It makes no assumptions about distributions of classes in feature space and can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.

Disadvantages: If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting, It constructs linear boundaries and requires average or no multicollinearity between independent variables.

Random forest

Advantages: It can perform both regression and classification task, produces good predictions that can be understood easily. It can handle large datasets efficiently and provides a higher level of accuracy in predicting out comes over the decision tree algorithm

Disadvantages: The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions.

K-Nearest Neighbor Model

Advantages: KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction and because of this it is very time efficient in term of improvising for a random modeling on the available data and is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be calculated using distance formula such as- Euclidian or Manhattan.

Disadvantages: Does not work well with large dataset as calculating distances between each data instance would be very costly and does not work well with high dimensionality as this will complicate the distance calculating process to calculate distance for each dimension.Sensitive to noisy and missing data.

V. Conclusions

In this report our goal was to address the situation of credit card fraud. This is because credit cards are very commonly used as a form of payment due to their convenience and ability to build credit.

The importance of discovering when a credit card transaction is fraudulent is very high due to the consequences that can come. If credit card data is exposed someone's credit card limit can surpass with a heavy charge put on it which can leave the owner of the card in debt.

Not only this, but if a fraudulent transaction and failing to pay the balance can result in the suffering of one's credit score that can take precious time to recover. Therefore, catching such an unusual charge needs to be caught and reported in a timely manner.

Through this report, we hoped to evaluate the different machine learning models that could be used to predict credit card fraud and find the one that was the most functional.

Using the dataset with information about previous credit card transactions made by European credit card users in 2013, we evaluated three different models. These models included the Logistic Regression, Random Forest, and K-Nearest Neighbor.

After calculating the error rates and comparing the Confusion Matrix for each of the models, we were able to find the most accurate model.

We concluded that the Random Forest model was the most accurate, the K-Nearest Neighbor model was the second most accurate, and the Logistic Regression model was the least accurate model.

References

[1] Itoo, F., Meenakshi & Singh, S. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int. j. inf. tecnol. 13, 1503–1511 (2021). https://doi.org/10.1007/s41870-020-00430-y

[2] ULB, M. L. G.-. (2018, March 23). Credit Card Fraud Detection. Kaggle. Retrieved December 12, 2022, from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[3] T. Wang and Y. Zhao, "Credit Card Fraud Detection using Logistic Regression," 2022 International Conference on Big Data, Information and Computer Network (BDICN), 2022, pp. 301-305, doi: 10.1109/BDICN55575.2022.00064.

[4] Credit_Card_Fraud.ipynb