CREDIT CARD FRAUD DETECTION

Bradley Beck

DECEMBER 20, 2022

MDA 620

Table of Contents

Business Issue / Background

Credit card fraud is a financial crime that millions of people fall victim to every year. It is known to be the most common form of identity theft and takes place every day in our society. This action is the unauthorized use of a person's credit card or credit card information to make purchases, obtain cash advances, as well as other forms of financial transactions. It is necessary for credit card companies to observe the differences between an actual transaction and a fraudulent one. This is important to guarantee customers are not losing money due to actions taken by someone else.

In the dataset credit card fraud detection, I observed European cardholders' transactions that occurred in September of 2013. This was an extremely imbalanced dataset that accounted for 492 instances of fraud out of 284,807 total transactions. To protect the individual's information each transaction has been transformed into numeric variables using the principal component analysis (PCA) method. This has been done to columns V1 through V28 to help prevent these people from falling victim to credit card fraud. The remaining columns in the dataset have not been transformed and consist of time, amount, and class. The time column shows the amount of seconds between each transaction, while the amount column provides the cost of the transaction, and the class column distinguished if the transaction was fraud or not. If a zero is shown it means no fraud was detected and a 1 indicates that the transaction was fraud.

Objective

Using this dataset, I explored how well a linear regression model and a decision tree could identify the cases of fraud throughout the data. In these findings credit card companies will be able to analyze the different indicators of fraudulent purchases. If a credit card company can

acknowledge an expense that is much higher than usual, they can clarify with the purchaser on whether it was them or not. Other than this detection of time between transaction can help lead to an instance of fraud. These identifiers can help limit the amount of fraud that takes place each year. The results of these models can ultimately be used as key identifier in detecting fraud.

Data Manipulation

Preparing the data for the logistic regression and decision tree model took lots of manipulation due to the dataset being imbalanced. I started this process by checking for any non-applicable data and came to find that there was none. Following this I separated all categorical variables from numeric variables, so the datatypes would not crossover. Next, it was necessary to check and remove all duplicates within the dataset. A total of 1,081duplicates were identified and needed to be removed to help improve the accuracy of the models. Prior to removing the duplicates, the data frame had 284,807 rows and after taking them away the data was left with 283,726. Now that this is taken care of it is necessary to scale and reshape the data due to its imbalance. This is necessary so the dataset can run in a train-test-split. In doing this I scaled the time and amount columns to fit within a specific measure. Thus, allowing them to be similar sizes. At the same time, I reshaped both columns so they would align with the class column. Before completing these functions, it was not possible to create my models due to an inconsistent number of samples in the columns. Now that my data has been manipulated it is time to create my models.

Linear Regression

In this model I established the train size to be 70%, leaving the test size to be 30%. On top of this I set the data to be randomly distributed. This resulted in the X_train being (199364,

29), the X_test being (85443, 29), the y_train being (199364,), and the y_test being (85443,). After accomplishing this I moved on to compare the actual and the predicted outcomes.

	Actual	Predicted
169876	0	0.000251
127467	0	-0.001557
137900	0	0.002484
21513	0	0.000044
134700	0	0.007742

With the actual being zero it means that no fraud was detected although our predicted values are seen to be slightly above or below zero. After analyzing this I was expecting the accuracy of the linear regression model to be very high although was surprised to see otherwise after evaluating the rest of the model. Next, I went ahead and retrieved the mean absolute error, mean squared error, root mean squared error, as well as the r squared.

MAE: 0.0033080484340579424

MSE: 0.0008195920816987147

RMSE: 0.028628518678037024

R2: 0.4804501758271654

While the MAE sits as low as .003, identifying that it is a good model the r squared value is only 48% and our main accuracy identifier. An ideal r squared results in 70% or better making this model not extremely accurate.

Decision Tree

When building my decision tree model, I started by putting together a test-train-split. In this I set the data so it would be randomly filtered as well as making the train size 30% while the test size was 70%. After this was successfully assembled, I was able to compute a confusion matrix where I could observe the accuracy of this model.

```
Confusion Matrix (Accuracy 1.0000)
```

```
       Prediction
Actual     0     1
     0 85301     0
     1     0   141
Confusion Matrix (Accuracy 0.9991)

       Prediction
Actual      0      1
     0 198922     92
     1     94    257
```

The confusion matrix on the top show the results of the training data which came out to be 100% accurate while the bottom demonstrates the test data which is 99% accurate. Ultimately this classification model was able to detect up to 99% of the fraud cases in this data frame.

Conclusion

Credit card fraud is a terrible crime that is constantly done to take advantage of banks and individuals. These instances of fraud can be detected through the cost of a transaction as well as the time taken between transactions. With these identifiers banks have been able to detect cases of fraud. In my linear regression and decision tree models I was able to test the accuracy of these identifiers in determining which transactions were fraud or not. My findings were that through a linear regression model the outcome of determining the difference between fraud or not was only 48%. Although when analyzed through a decision tree model the accuracy of this test was nearly 100%.

Works Cited

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019

Yann-Aël Le Borgne, Gianluca Bontempi Reproducible machine Learning for Credit Card Fraud Detection - Practical Handbook

Bertrand Lebichot, Gianmarco Paldino, Wissam Siblini, Liyun He, Frederic Oblé, Gianluca Bontempi Incremental learning strategies for credit cards fraud detection, IInternational Journal of Data Science and Analytics