Credit Card Fraud

Rachel Nelson

Bellevue University

Author Note

DSC530-T303 Data Exploration and Analysis (2211-1)

Statistical Hypothetical Questions

**For this paper, my statistical questions included:**

* + What are factors that correlate to fraud?
  + Can you use those factors to create predictions on which credit card transactions are fraudulent?

**My hypothesis, was that:**

* + Time between transactions will be an important factor when identifying fraud

# EDA Outcome

The outcome of my exploratory data analysis was that only two of my four Explanatory Variables correlated to fraud. Logistic Regression was used since the dependent variable was Boolean.

# What do you feel was missed during the analysis?

I feel like a solid set of statistical questions and hypothesis was missing during this analysis. Also, there was no good usage of CRISP-DM framework in the approach. Models should be compared to other models to see which ones perform better. Different methods should be used. All of that was missed during the analysis as I was just following the guidelines.

# Were there any variables you felt could have helped in the analysis?

Because I limited the data set to 5 variables, 1 dependent variable (Class) and 4 explanatory variables (Time, Amount, V4 and V9), I feel like I missed out on analyzing the other variables that were available, which would have improved the analysis. Also, the variables in the data set were transformed via PCA transformation, which prevented me from truly understanding the data set I was working with and thus resulted in a disadvantage.

# Were there any assumptions made you felt were incorrect?

I made the assumption that the data set would be imbalanced, which was the correct assumption. One of my hypotheses was that time would have a correlation to identifying fraud, which turned out to be an incorrect hypothesis. I also assumed that I should balance out the data set, and used under sampling to do so, which I am not sure was the correct approach.

# What challenges did you face, what did you not fully understand?

The biggest challenge for me was trying to interpret the results of the Probability Mass Functions. They aren’t as clear cut as the examples in our textbook, so as a result, add little meaning to the analysis. I am not sure if it has to do with just the variables that were used, or if another factor is at play. CDF also didn’t provide any meaningful information.

References

Downey, A. (2015). *Think stats: Exploratory data analysis*. Sebastopol, CA: O'Reilly Media.

ULB, M. (2018, March 23). Credit Card Fraud Detection. Retrieved November 20, 2020, from <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Tables

Table 1

Descriptive Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Mode | Variance | Skewness |
| Class | 0.00 | 0.00 | 0.00 | 23.99 |
| Amount | 88.35 | 1 | 62560.07 | 16.98 |
| Time | 94813.86 | 163152 | 2255124006.20 | -0.04 |
| V4 | 0.00 | -0.84 | 2.00 | 0.68 |
| V9 | -0.00 | 0.17 | 1.21 | 0.55 |

Figures



Figure 1. Histogram for class

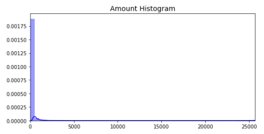


Figure 2. Histogram for Amount

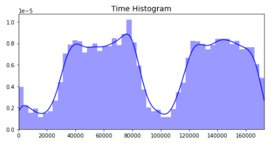
\

Figure 3. Histogram for Time

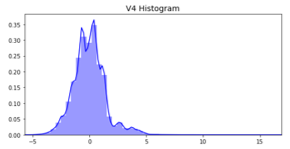


Figure 4. Histogram for V4

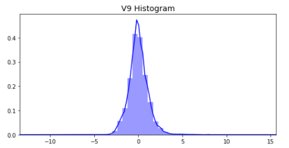


Figure 5. Histogram for V9

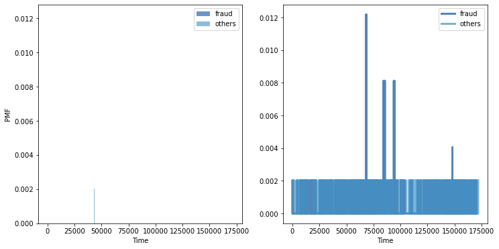


Figure 6. PMF: Time

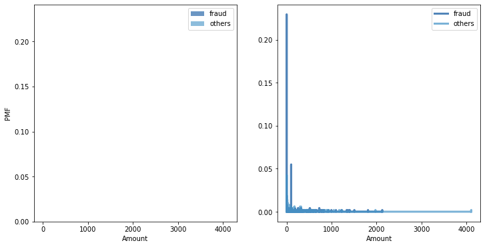


Figure 7. PMF: Amount

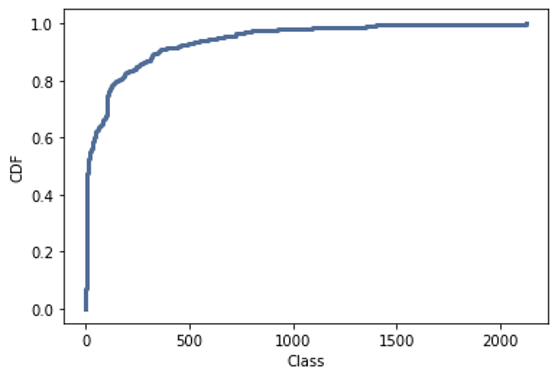


Figure 7. CDF for Class

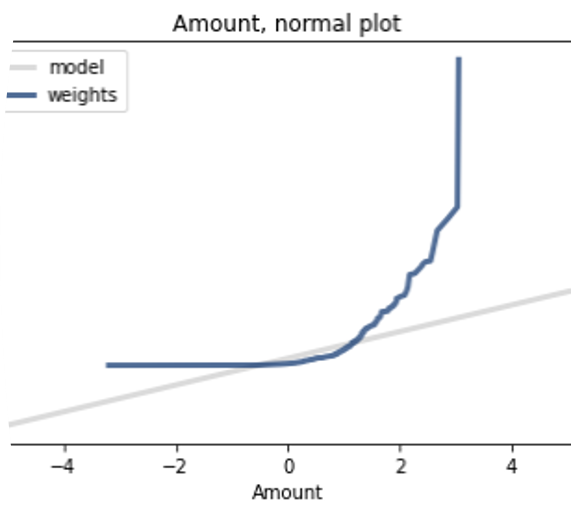


Figure 8. Normal Probability Plot for Amount

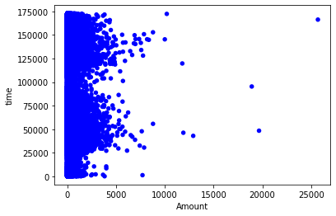


Figure 9. Scatterplot: Time and Amount

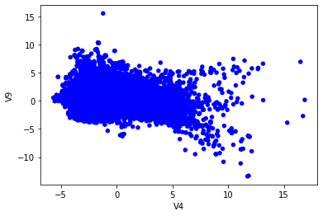


Figure 10. Scatterplot: V9 and V4



Figure 11. Correlation Matrix

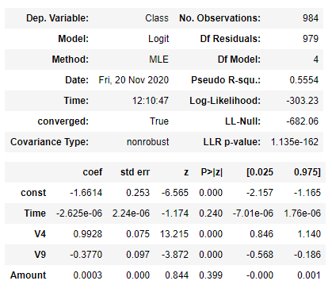


Figure 12. Logistic Regression Summary Results