IEEE-CIS FRAUD DETECTION

KAGGLE COMPETITION

Sifat Ul Alam Nabil
sifatnabil@gmail.com

Friday 11th June, 2021

Abstract

In today's environment, complex problems necessitate complicated answers. Often, traditional rule-based systems fail to tackle these complex problems. Data Science, which is a method of gaining insight from data and letting machines determine the rules accordingly, has gone a long way and is widely used in today's fast-paced technological environment. This report discusses areal-world challenge of identifying fraud in credit card transactions and presents a study on how to solve this problem using the data that was provided to us. The paper explains how the author handled the challenge and how much better of a result they were able to achieve.

1 Methodology

Both the Test and Train dataset was divided into two tables. Transaction table and Identity table. The Transaction table had more rows, so we joined the two tables to create the Train and Test set using left join. The Train set had 590540 data points where each data point had434 features where the Test set had 506691 data points and 433 features. After our dataset was ready we ran some exploratory data analysis. From our observations, the data set was highly imbalanced and a lot of the features had null values. So, our next step was to clean the data. But we also observed the names of the features were different between the Train and Test set. So, we renamed the features in the Test set according to our Train set. After that, we calculated the number of null values and their percentage for each feature. As our dataset was quite large and with a lot of features, we initially removed all features that had more than 1,00,000 missing values from the Train set and we also dropped the same features from the Train set to keep our test set consistent with our Train set. The first column TransactionID represented a unique number for each transaction which did not represent the Target variable. So we dropped the first column followed by the features that had more than 15,000 missing values. After the clean-up performed above, we were able to reduce our features from 434 to 111. As some of the features still had null values, we filled those values with the mean of the column for numerical features and with the mode of the column for categorical features. As our model expects numerical features only, we had to replace the categorical column values with numerical values to make them suitable for our model.

For model selection, we used **Light Gradient Boosting** which is a lightweight, fast, and efficient boosting algorithm with a learning rate of 0.05 and ran for 2000 boosting iterations. [1]

2 Result Analysis

With our multiple data cleaning, processing, and model selection approaches, we were able to achieve 0.901537 private and 0.926844 public scores. We observed a great score difference based on feature selection. We also tried to apply Principal Component Analysis and reduce more features but the score decreased.

3 Conclusion

All our approaches towards solving the problem were based on dataset analysis. However, with more time and proper domain knowledge, it would be possible to figure out and construct more useful features that would improve the accuracy of the model. More in-depth hyperparameter tuning for the model and advanced approaches like Neural Networks, Time Series Analysis may also help us generalize the model better and gain more accuracy in terms of score.

References

[1] Vesta Corporations. *IEEE-CIS Fraud Detection Dataset*. IEEE Computational Intelligence Society, 2019.

[2] Ahmed Khaled. *IEEE-CIS Fraud Detection Notebook*. Kaggle.

[3] Microsoft. *Ligth Gradient Boosting Official Documentation*.

[1] The code for this project can be found here sifatnabil-github