Ivan Calderoni
Parish Kaleiwahea
Vanadda Sermpol
ISE 5103 Intelligent Data Analytics
Peer Review

Peer Review for Casares and Olivera

Project Understanding

Casares and Olivera seem to have a very good grasp on the *Detecting Fraud on Plastic Card Transactions* project. They cite multiple sources throughout the report, which shows they have researched the topic well and use facts to back their reasoning and claims. I would like to add that most of the references used in this report so far bring something to the table and are not used as fillers. They do a great job at explaining why it is beneficial for businesses and banks to detect fraud. On top of that, it was very insightful to read about how fraud also affects "us," the customers, because it drives up interest rates, membership fees, and reduce benefits to the cardholders.

A simple suggestion: where does the dataset come from? Kaggle? Data.gov? Were there multiple data sets? How clean was the dataset from the source? Perhaps state where the dataset was found or have a link for the reader in case he or she wants to check out the source.

Data Understanding

Casares and Olivera created a table that displays and describes the different predictors. The table is very clear and concise so it is a good visual aid for the reader to use as a reference. It contains more than the nine variables that were required for the project (there is a total of 12 attributes). One can easily understand what every variable represents and which one is the target variable (is_fraud in this case).

The reader is not able to deduce if there was a Training dataset and a Testing dataset provided. It seems to be that there was only one dataset provided which they described well. They mentioned some cross validation, but wanted to know if that was a true holdout set or if the test data did not have the "fraud" field to test.

They do a great job at explaining how and why the data is imbalanced. They go on to explain what they do to overcome this “imbalance” issue. One suggestion here might be to suggest other potential resolution(s) when encountering imbalance datasets.

The graph that shows transactions/fraudulent transactions over a period of time is very insightful. It is actually very interesting and it makes the reader wonder why is that. Another suggestion for the data understanding section: provide more meaningful exploratory data visualizations like the one mentioned earlier.

Outlier analysis would be useful to visualize the dataset and how “rare” a fraudulent charge is. I would suggest color coding on a jitter graph or compare a boxplot side by side with fraudulent charges and non-fraudulent charges.

Data Preparation

This might be the strongest section in the report. Casares and Olivera seem to have done all the necessary feature engineering to at least get started. They created multiple aggregate features, which seem promising at providing some hidden information. It will be interesting to find out if these new features will improve the prediction accuracy.

They found a positive correlation between is_fraud and fraud_by_id_merchant (one of their new features) and have a nice correlation matrix to depict it. One suggestion for the correlation matrix is to add the spelt-out feature instead of just the column name. Sometimes the column names are truncated and underscores make it difficult to understand what the feature actually means.

A reader might want to know if the dataset provided is complete or not. If is not, what will be done with all those NAs? Perhaps, comment on the condition of the dataset and what method of imputation was used if applicable. With fraud charges being rare, it would be useful to know what efforts were made to keep as much of those instances as possible.

Have they considered some sort of clustering analysis on nearest neighbors of the fraudulent charges? Although it takes time to run, the nearest neighbor algorithm could provide insight on those prone to fraud and what characteristics they did not share with their neighbors who did not have fraud. Boundaries could be important for cases where the model is not confident and where a possibility of a false positive is high.

Modeling

Casares and Olivera encountered a few issues due to the size of the dataset(s). They have a nice table already set up that will display the results of the different models they have chosen to create. The models they picked were models that work well with classification type problems. However, like mentioned earlier, they were not able to run them succesfully. It will be interesting to find out the results and see which models perform well.

In addition to running the models, it would be nice to see the residual graphs and the AUC curve. Considering K-folds cross validation was used, it would be nice to know how and why you arrived at the evaluation. Were you concerned with over-fitting? If so, what did you do to consider to prevent it.

When they get to the boosted tree model, what steps were taken to tune the model will be insightful to the reader. Knowing that the boosted tree takes a considerably longer time to run than other techniques, it would be nice to know what kind of preparation work was used to assist with the run time.

Evaluation

They already know what metrics they will use to validate the models they will create. If it is a Kaggle dataset, one can also use the Kaggle score as a validation metric. Even if the competition is not live anymore, one can still submit predictions to a private leaderboard and get a Kaggle score.

Overall/Summary

Casares and Olivera are off to a great start on this project. However, some work still needs to be done to have a finished product. The data preparation and feature engineering seem to be very well developed, but the “Modeling” and “Results” sections need work. Also, a nice “Executive Summary” still needs to be added to the report and a “Conclusion” once they get the results.