CS506 midterm report

Matthew Levine

April 1, 2024

Kaggle username - Matthew Levine/mlevine6

1 Finding Features

My methodology for finding features to add to the model was was first to plot the base features to find any noticeable correlations with a fraudulent transaction. Through this I found that certain states, cities, merchants, transaction categories, and city populations all had noticeably higher rates of credit card fraud.

Noticeable observations included:

- Fraudulent transactions seem to only happen at amounts below about $1500
- Fraudulent transactions are less likely to happen in cities with large population
- Transactions in the shopping_pos, grocery_pos, shopping_net, and misc_net categories had much higher rates of fraud than other categories
- Connecticut, Hawaii, Alaska, and Idaho all have scignificantly higher rates than average of credit card fraud

My second method for finding features was to use intuition about what types of things might indicate a transaction being fraudulent and creating a new feature to represent that. Through thsi I made features for

- MII - The first digit of the credit number that indicates the industry/card issuer. My logic for this was that people using credit cards for certain issuers might be more prone to having thier information stolen or that certain card types might be less secure.
- Hour - Fraudulent transactions might occur more at certain times of the day

- High Risk City, State, Merchant - Features indicating if this transaction was made in a city/state or with a merchant that has a higher than normal rate of fraud
- Weekday - Fraudulent trnsactions might occur more or less on certain days of the week. Ultimately the only feature that improved model performance was whether the transaction happened on a Sunday.
- Age decade, Age - Certain age groups might be more prone to having thier cc information stolen
- Distance - Distance between the card issuers address and the merchant, transactions further away from the owners address might be more likely to be fraudulent.
- Info on past/future fraudulent transactions - Multiple features indicating whether there was a fraudulent transaction on this card in the prior day, prior 3 hours, prior hour, next hour, next 3 hours. As well as the time of the most recent fraudulent transaction prior to this one if any. My logic for this variable was that fraudulent transactions would most likely happen close together and therefore if a fraudulent transaction happened in a short frame of time before or after this transaction then there is a high likelihood that this transaction is also fraudulent

Overall key findings I found about the dataset through these created features were:

- Fraud is overwhelmingly more likely to happen between 10 p.m. and 4 a.m.
- Fraud is less likely to happen on Monday/Tuesday then the rest of the week
- Airline credit card are most likely to have fraudulent transactions
- People between the ages of 50-70 are scignificantly more likely to have fraudulent transactions happen on thier credit card
- Fraudulent transaction happening in a short time before or after a given transaction results in that transaction being overwhelmingly (Near 100%) likely to be fraudulent

Plots exploring the correlation between these features and the probability of a transaction being fraudulent are located in the jupyter notebook for my model.

A that I encountered was in using the information about past/future fraudulent transactions the data used for the entire dataset could only be pulled from the training set as fraudulent data was unknown for the testing set. This could lead to situations where fraudulent transactions close to a given transaction are hidden in the submission data, it also lead to better scoring on the testing set that was split off of the training data compared to the submission set.

Something else of note is that the first time I created the feature indicating the time difference of the most recent previous transaction, there was an error and it instead indicated the time difference of the last fraudulent transaction for that card in the dataset. However when I fixed this the F1 score for the testing data improved but the F1 score for the submission data dropped significantly. So it seems that there is some correlation for this data and a transaction and the data being fraudulent and I eneded up using both the previous value and the fixed value in the final model.

2 Model Selection, Parameter Tuning, and Validation

In order to find the best possible model for this problem I surveyed several different types of models to see how they performed including KNN Classifiers, Support Vector Machines, Logistic Regression, Naive Bayes Classifiers, Decision Trees, and Random Forests. Ultimately I found that most of these models performed significantly worse than tree based models and I found that the best tree based model was Gradient Boosted Trees.

I used the package xgboost for my gradient boosted trees as it has more options for hyper parameters and trains much faster compared to the sklearn implementation. I tuned the parameters for number of estimators, max tree depth, learning rate, minimum loss reduction, evaluation metric, tree method, and grow policy to get the optimal parameters that gave the best F1 score.

Furthermore, I used a soft-voting ensemble of models with slight changes in hyper parameters between each model to prevent over fitting. I found when training the models that unweighted models (models that treat fraudulent and non fraudulent transactions as the same when training) tend to have a better false positive rate while weighted models (models that add more significance to fraudulent transactions when training to account for the dataset imbalance) tend to have a better false negative rate, therefore in order to balance out my predictions and improve F1 score I split the ensemble between weighted and unweighted models.

To validate my ensemble I split of 25% of the training data prior to training and used it as a testing set, I used F1 as the score to judge which parameters/models were the best as thats what is used for the competition.