Credit Card Fraud Detection With Machine Learning

Problem Statement

You are on vacation. You go shopping and when you go to pay, your credit card transaction is declined. You try your card again, but again the card is declined. You are a bit embarrassed, and as your preparing to reach for another card you get a message on your phone. It's your bank. They want to know if it is you using the card, or not. You select the option 'yes' and your card is released.

How would you feel in that situation?

Problem Statement - cont...

Vesta Corporation, the world's leading payment service company is seeking to improve the accuracy of their real time credit card fraud detection capabilities. The hope is that with increased accuracy there will be less situations like the one described previously.

This is what this project aims to accomplish.

The Challenge

- The dataset samples are each individual transactions
- A unique credit card is associated with multiple transactions
- No one transaction is indicative of fraud on its own
- To determine fraud one must look at the transaction history for the card.
- Need a way to find patterns in the transaction associated with a card

Stats

- In 2018, $24.26 Billion was lost due to payment card fraud worldwide
- The United States leads as the most credit fraud prone country with 38.6% of reported card fraud losses in 2018
- Credit card fraud increased by 18.4 percent in 2018 and is still climbing
- Credit card fraud accounted for 35.4 percent of all identity theft fraud in 2018
- Credit card fraud was ranked #1 type of Identity theft fraud

Proposed Solution

- New more accurate classification model
- Determine and select the most important features
- Conduct behavioral analysis based on transaction activity.
- Test several algorithms
- Perform parameter tuning to optimize each model
- Select the best performing model based on chosen metric

Business Value

- Flag fraudulent transactions while also avoid misclassifying legit transactions as fraudulent
- Detect fraudulent transactions immediately
- Understand what factors are more predictive to detect fraud
- Balancing convenience with security

Methodology

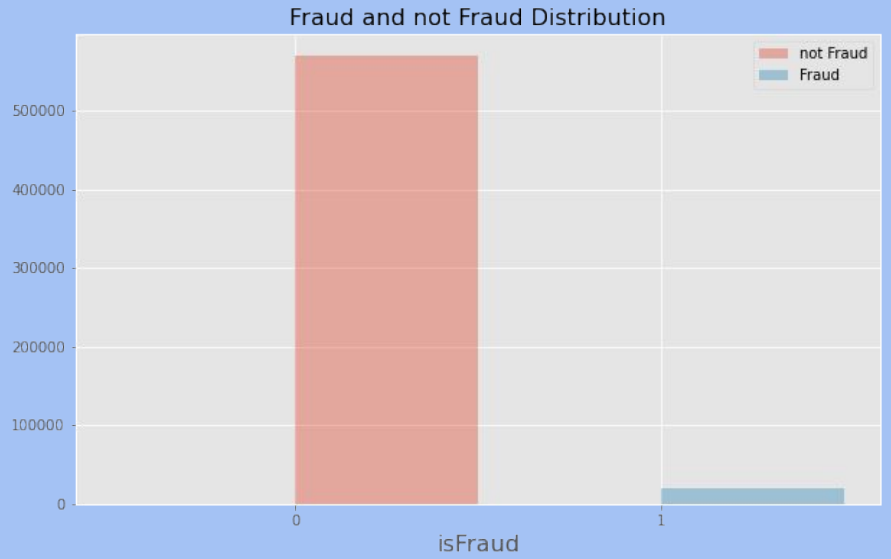
- Use the transaction dataset provided by Vesta
- Examine the data to understand relationships
- Create new features from existing features to leverage interactions
- Understand the ongoing periodic nature of the data
- Use the above knowledge to design features that contain historical information
- Use the information to conduct behavioral analysis on the transactions
- Use state-of-the-art machine learning algorithms to model the data
- Evaluate and select the best model.

The Data

- The data has been provided by Vesta Corporation.
- It is a very large dataset of over 1Gig
- The data comes from Vesta's real-world e-commerce transactions and includes 433 features.
- Many of these features have been engineered by Vesta's team.
- As is normal with this type of data, it is very unbalanced.

Fraud vs Not Fraud distributions

These distributions are typical in transactional data.



The Model

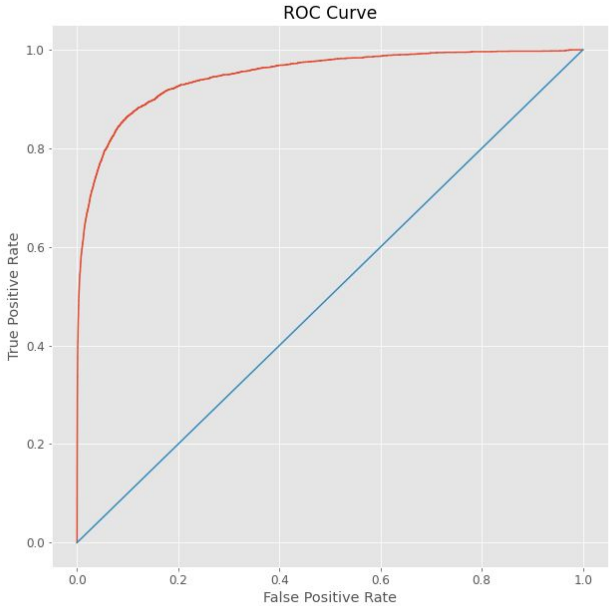
- We tried several algorithms
- Best results came from LightGBM and CatBoost
- Both are state of the art boosting algorithms
- We used AUC as the scoring metric
- AUC measures separability - How well the model can separate classes

LightGBM

AUC Score:

avg pred auc: 0.94848 avg auc: 0.94781+/-0.00016

- **A score of 1.0 would be perfect**
- **A score of 0.95 is very good**
- **The close the curve approaches the top-left corner, the closer to perfect (1).**

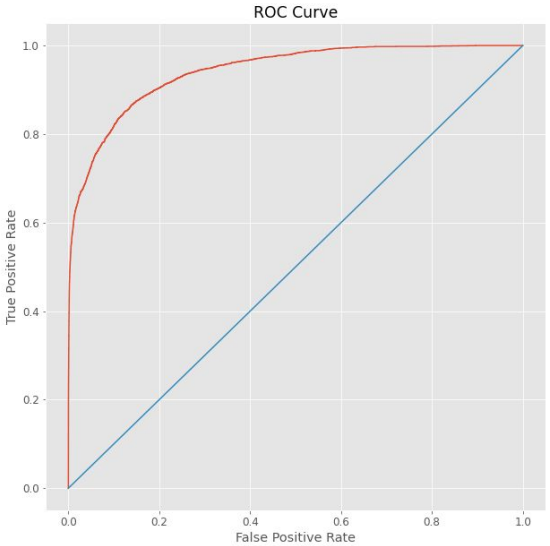


CatBoost

AUC Score:

avg preds auc: 0.94200, avg auc: 0.94161+/-0.00036

- **A score of 1.0 would be perfect**
- **A score of 0.94 is lower than LGBM**



Conclusion

- LightGBM proved to the better model for this task
- AUC score of 0.95 will allow for accurate fraud prediction
- This model can be deployed to catch fraud as it happens
- This will allow for automatic notification to client when fraud is suspected.
- Immediacy of notification means client is informed before issue escalates
- It also means that you can act on fraud before it becomes a lost

Future Work

- Continue to tune parameters
- Work on improving feature engineering
- Explore other models

Thank You

Thank you for attending. Please do not hesitate to contact me with further question.