DETECTING FRAUDULENT TRANSACTIONS

SUPERVISED LEARNING CAPSTONE

CHATHURA BRAHAKMANAGE

DETECTING FRAUDULENT TRANSACTIONS

MOTIVATION

- Fraud can be defined as money or property being obtained through false pretenses
- According to Statista, in 2018, US merchants lost an estimate of $6.4 billion dollars in payment card fraud loss in 2018
- Fraud detection can:
  - ○ Save businesses and consumers millions of dollars
  - ○ Improve existing fraud detection models
  - ○ Enhance customer experience

GOAL

- Use historical Vesta's real-world e-commerce transaction and build a supervised learning model to predict whether a transaction is fraud or not

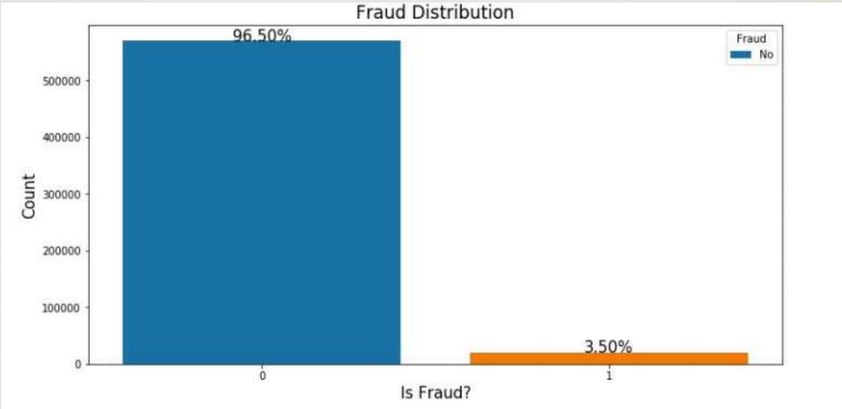
OVERVIEW

- DATASET
- CLASS IMBALANCE STRATEGY
- MODEL METRIC
- BASELINE MODELS
- RESULT
- FUTURE WORK

DATA SET

- Collected by Vesta's fraud protection system and digital security partners
- There are 590,540 online transactions
- Data types (434 attributes):
    - Transaction records
    - Identity Data

CLASS IMBALANCE



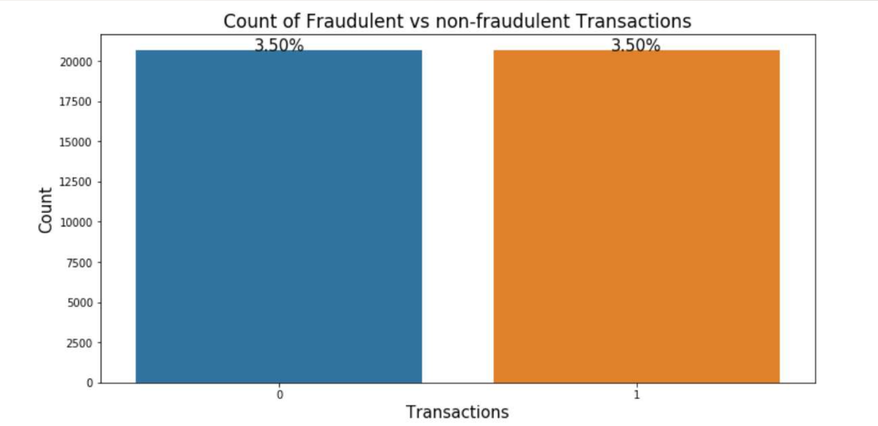
- There are 569,877 observations of normal transactions
- Only 20,663 transactions are fraud

DATA PROCESSING AND FEATURE ENGINEERING

- Drop columns missing more than 25% of data points. Then impute numerical variables by mean and categorical variables by mode.
- Set the threshold at 3 standard deviations to retain 99% of the data set and to remove outliers.
- Engineer New Features:
    - Aggregated features
    - Drop variables with high collinearity
    - Encoding categorical variables :
        - Label encoding :- Tree-based models
        - One-Hot encoding :- Any other models

UNDERSAMPLING

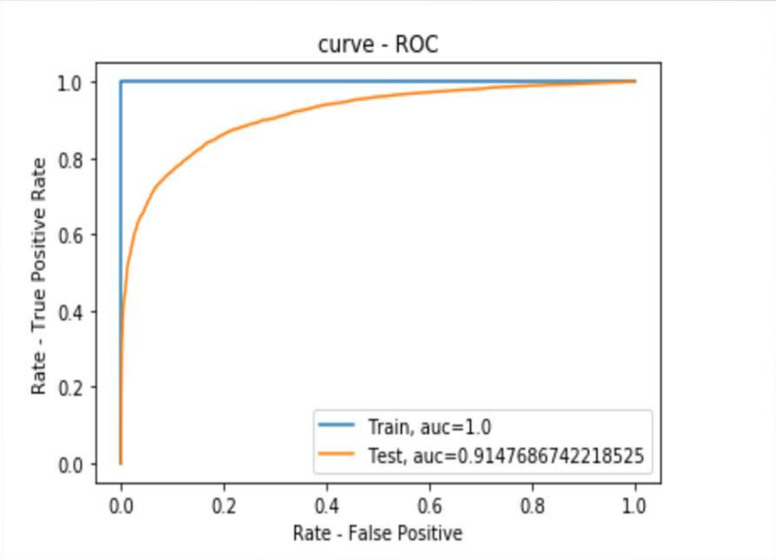
- Create class balance by randomly selecting equal amounts of normal and fraudulent observations



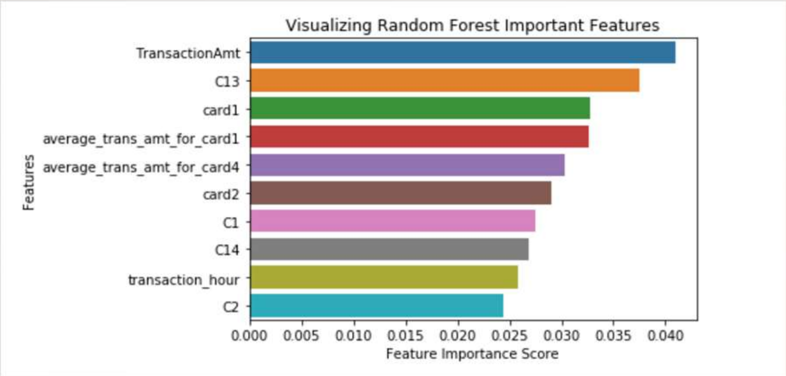
BASELINE MODELS

RANDOMFOREST

- Hyperparameters:
  ○ n_estamators = 100
  ● Performed well with a score of 0.91 and accuracy of 83%
  ● False negative rate 14%
  ● False positive rate 19%
  ● Longer computational time



FEATURE IMPORTANCE



- Transaction amount and counting matches appears to be the most important features in detecting fraud.
  Our feature engineered variables also made it in the top ten with interactions between transaction amount and card information.

GRADIENT BOOSTING

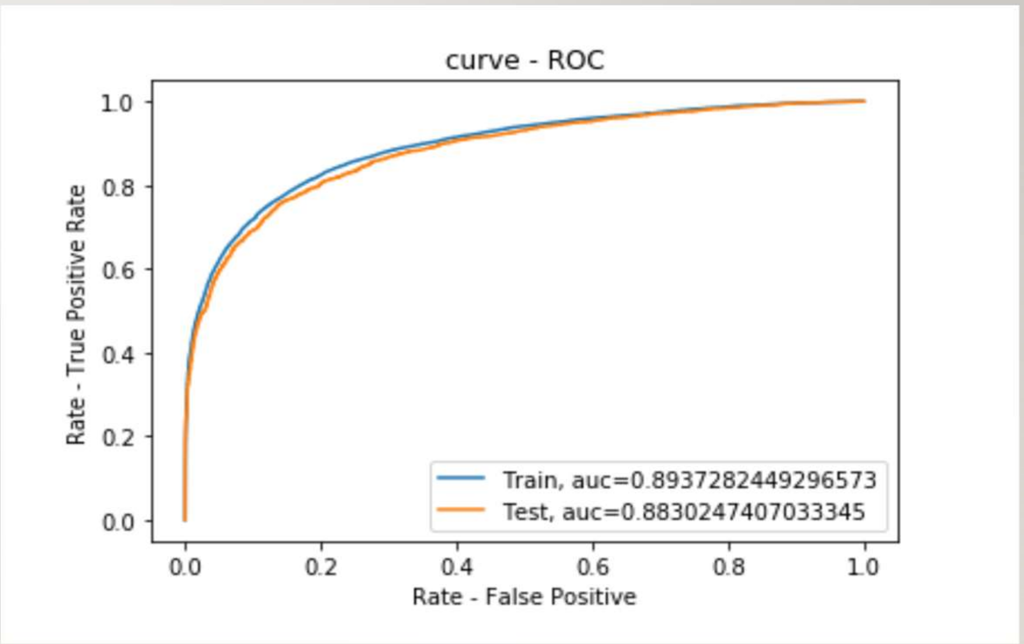
- Hyperparameters:
  ○ Random_state = 42

  ● The initial gradient boosting model has a score of 0.88 and accuracy score of 81%

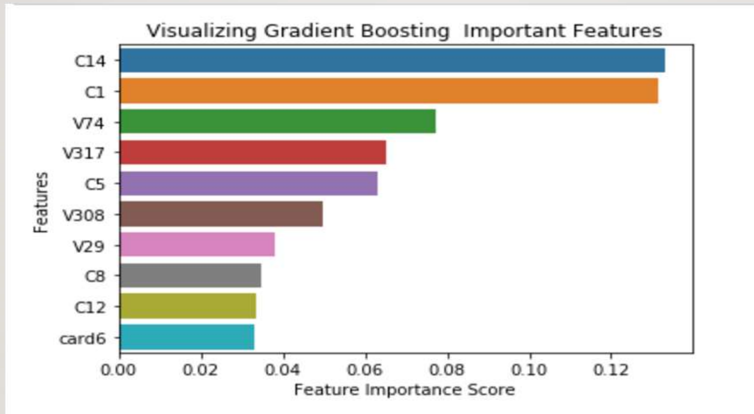
  ● False negative rate 16%

  ● False positive rate 22%

  ● Computational time was significantly longer than Random Forest models



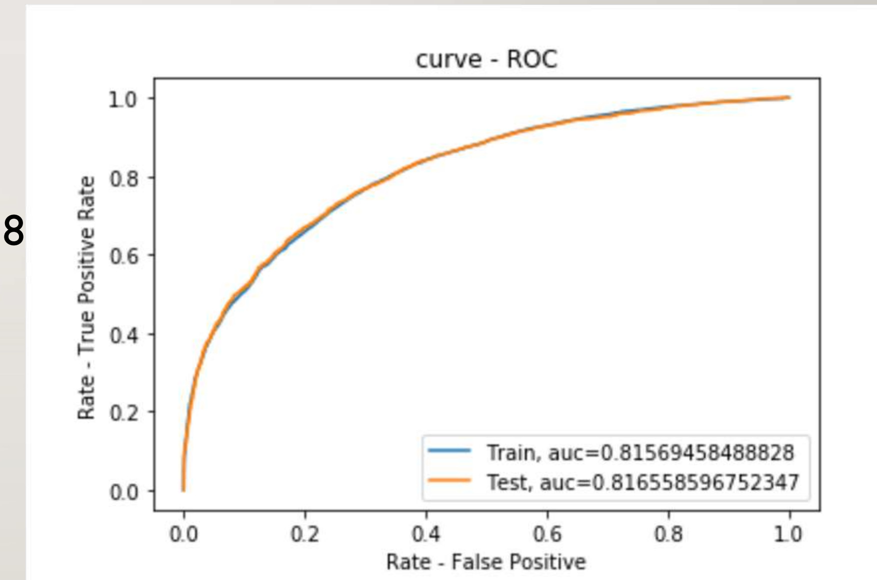
FEATURE IMPORTANCE



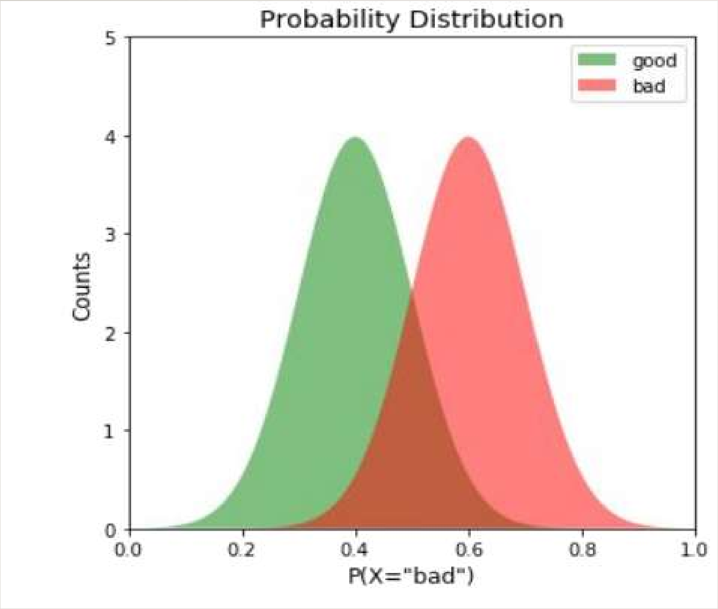
Count of card information matches, and Vesta feature engineered variables appear to the top contributing factors in detecting fraud for this model.

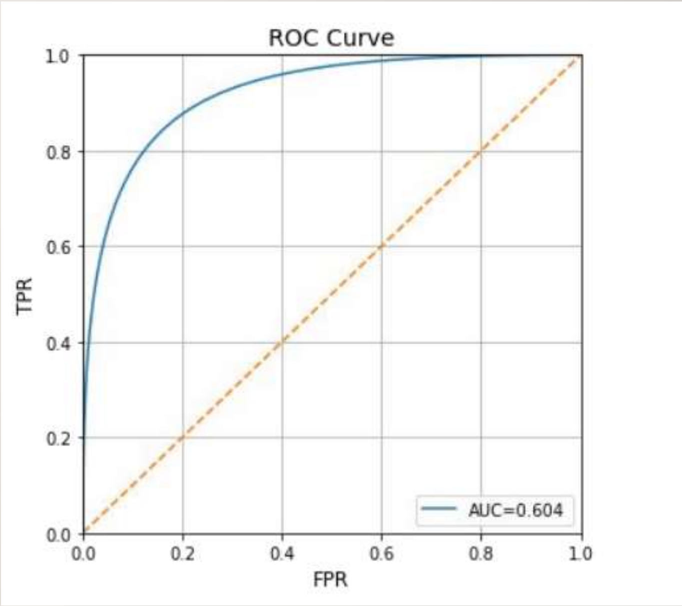
LOGISTIC REGRESSION

- Hyperparameters:
  ○ Feature selection with lasso (shrinkage method)
  ● ROC AUC score of 0.737 and accuracy of 8
  ● False negative rate 24%
  ● False positive rate 28%
  ● The computation time for logistic was relatively fast



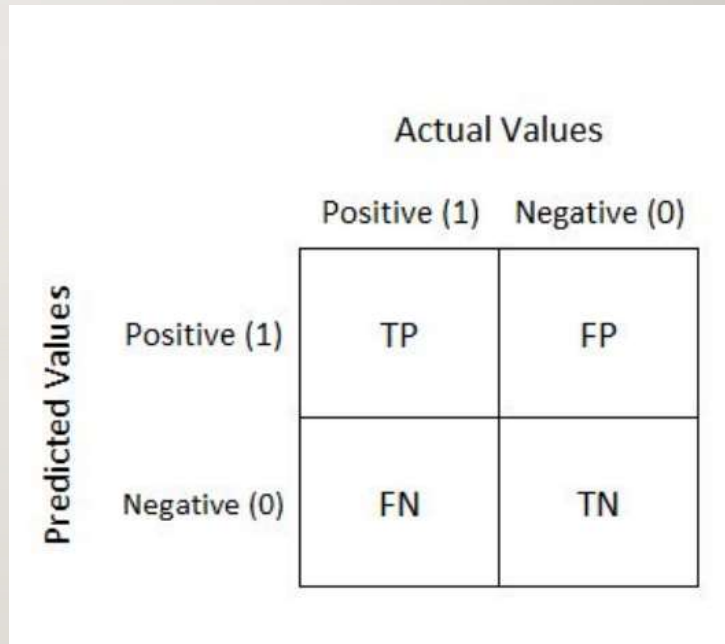
MODEL METRIC: ROC AUC





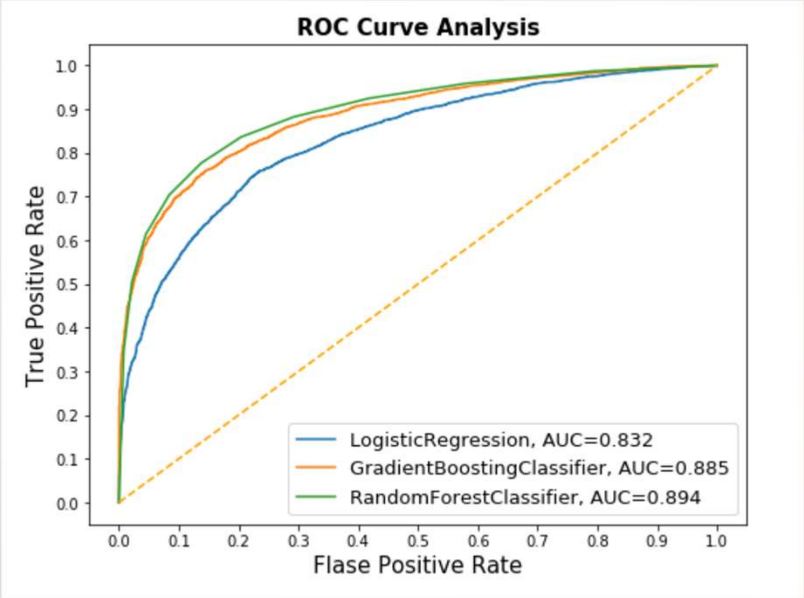
MODEL METRIC

- Other metrics to consider:
  - False negative rate
  - False positive rate
  - Accuracy



MODEL RESULT

Logistic regression model perform better than random guess with a ROC AUC score of 0.83 on the test set
○ Poor performance in classifying normal transactions
Random forest with all features performed that best with a ROC AUC score of 0.89 on the test set
○ Best model in ROC AUC score and lowest false negative rate
○ Short computational time
Gradient boosting was a close match to random forest with a ROC AUC score of 0.88 on the test set
○ Comparable to Random Forest, but slightly lower ROC AUC score and higher false rates
○ Longest computational time



FUTURE WORK

- Exploring oversampling methods and utilize different imbalanced class techniques
- More observations may improve the random forest model's performance
- Engineer more features with transaction amount, card columns, count columns and time features

THE END-