Detecting Fraudulent Transactions

Supervised Learning Capstone

Thapani Sawaengsri

MOTIVATION

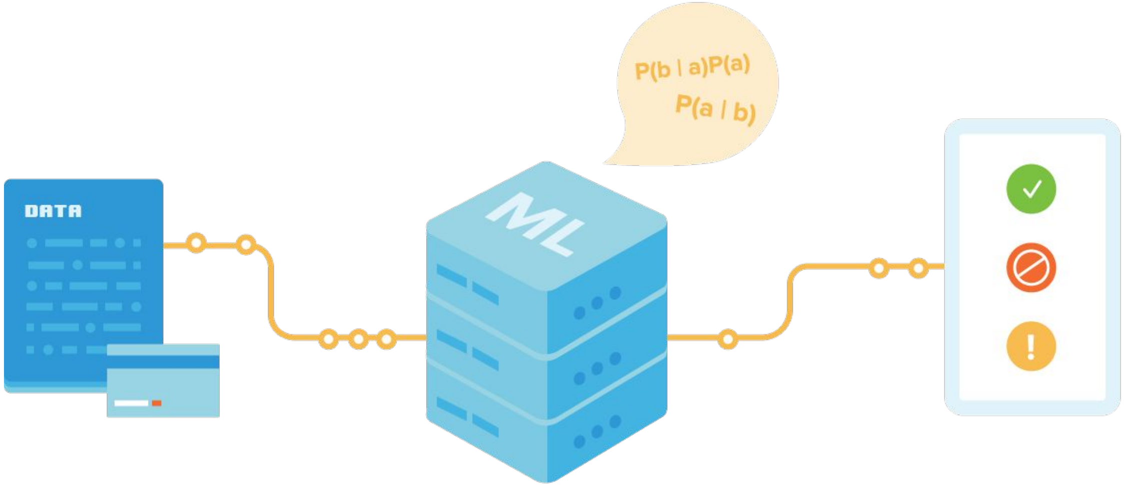
- Fraud can be defined as money or property being obtained through false pretenses
- According to Statista, in 2018, US merchants lost an estimate of $6.4 billion dollars in payment card fraud loss in 2018
- Fraud detection can:
    - Save businesses and consumers millions of dollars
    - Improve existing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

GOAL

- Participate in Kaggle's competition IEEE-CIS Fraud Detection machine dataset
- 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
- MODEL ON IMBALANCED DATA
- RESULT
- FUTURE WORK



Feed data into a machine learning algorithm to help you make a decision.

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

DATA TYPES

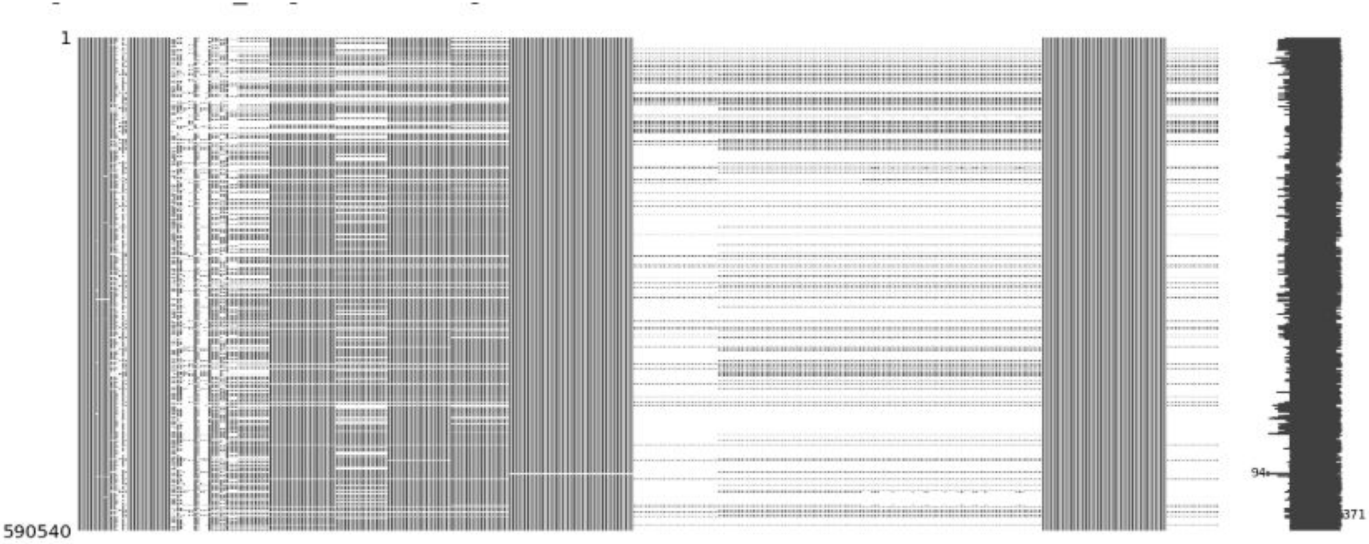
Categorical Features

- ProductCD: product code
- card4, card6: card information
- P_emaildomain: purchaser email domain
- R_emaildomain: recipient email domain
- M1 - M9: match (names on card, address, etc)
- id_12 - id_38: identity data
- DeviceInfo
- DeviceType

Numeric Features

- TransactionDT: timedelta from a given reference datetime
- TransactionAMT: transaction payment amount in USD
- card1-3, card5: card information
- dist: distance
- addr1, addr2: address
- C1-C4: counting
- D1-D15: timedelta (time between previous transaction)
- V1-V339: Vesta engineered features
- id01 to id11: numerical features for identity

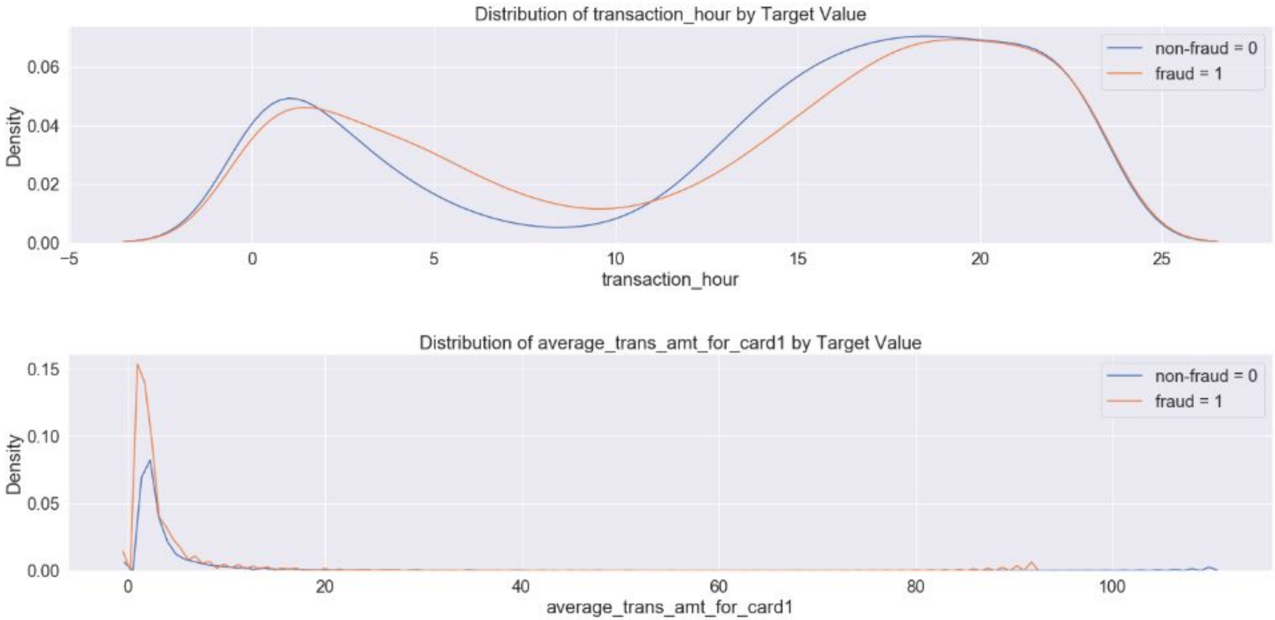
DATA CLEANING



95% of columns contained missing values

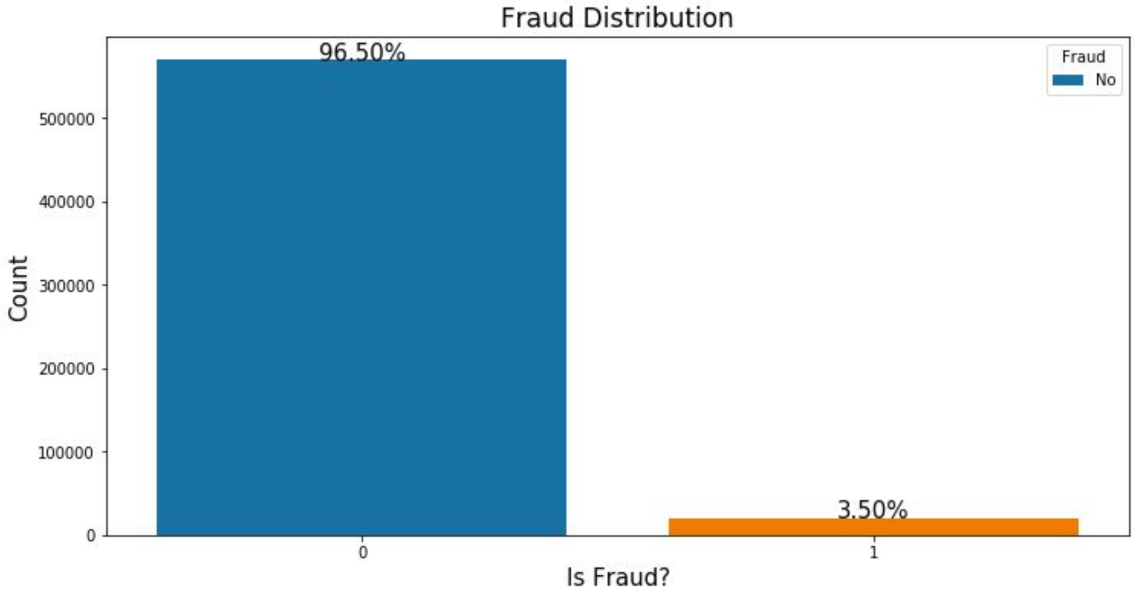
- Impute numeric variables with mean
- Impute categorical variables with mode

FEATURE ENGINEERING



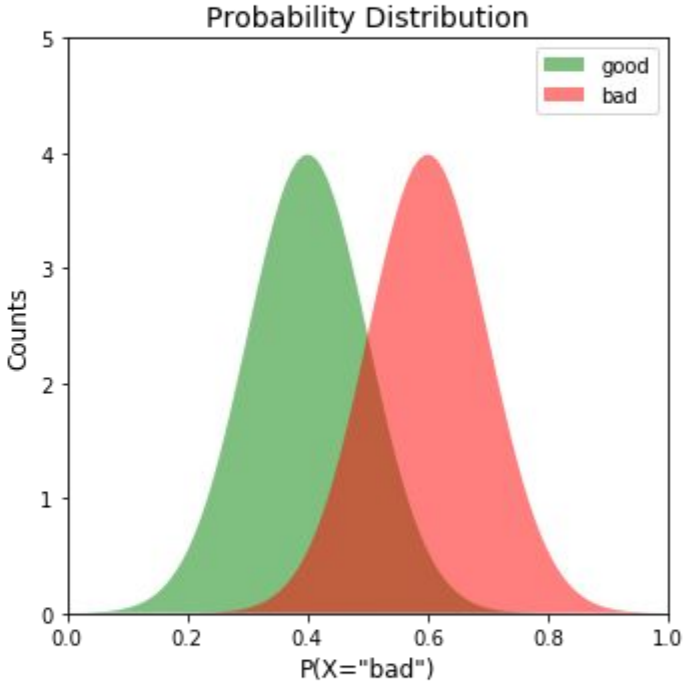
Aggregate variables to create new features

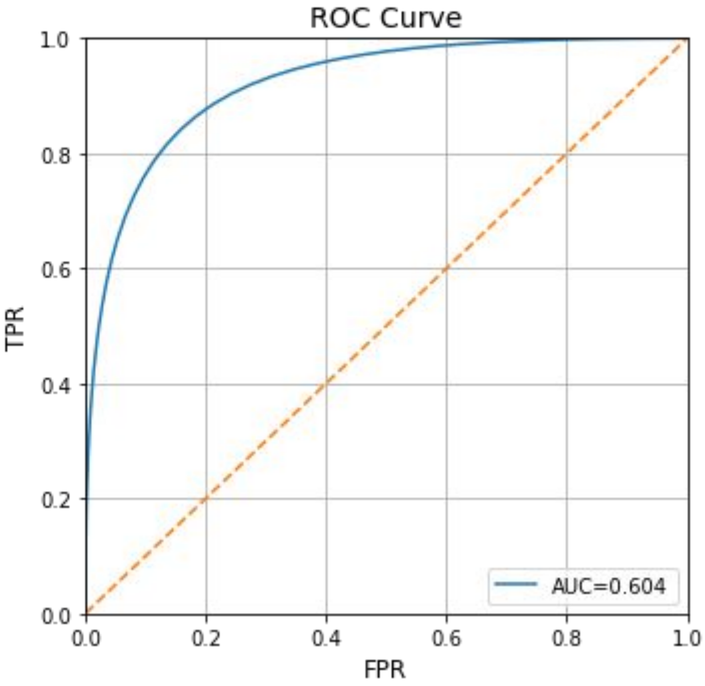
CLASS IMBALANCE



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

MODEL METRIC: ROC AUC

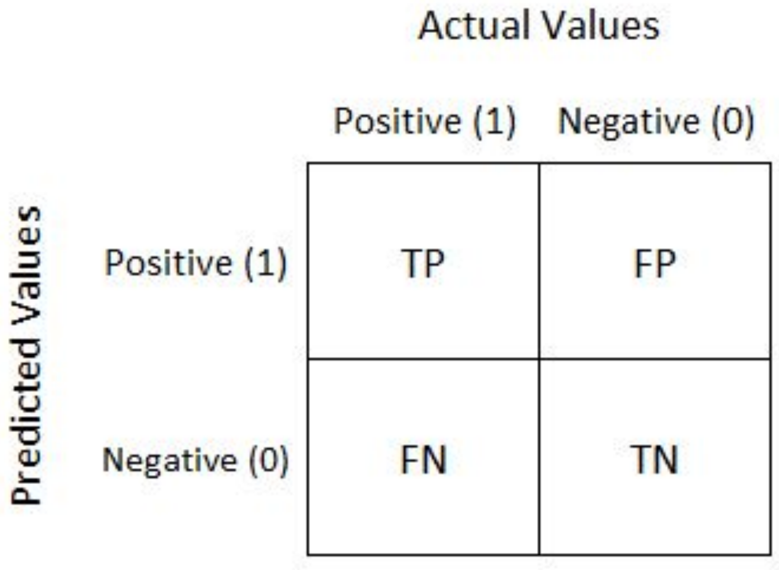




MODEL METRIC

Other metrics to consider:

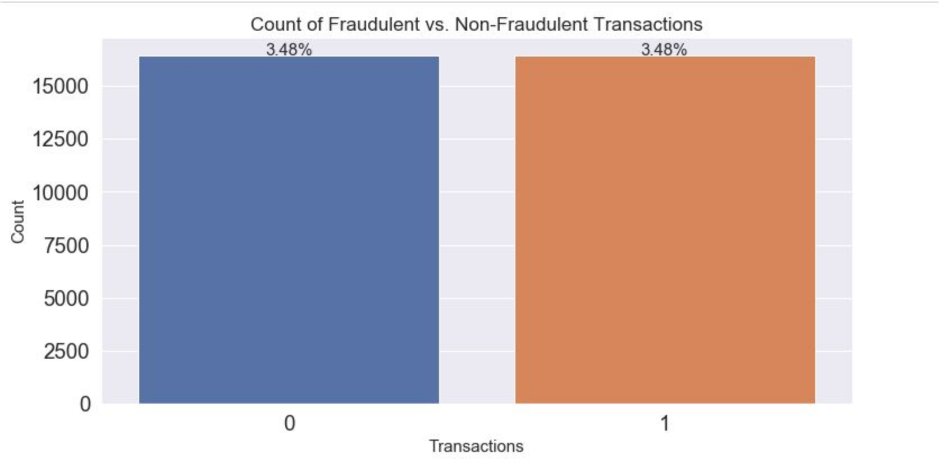
- False negative rate
- False positive rate
- Accuracy



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

UNDERSAMPLING

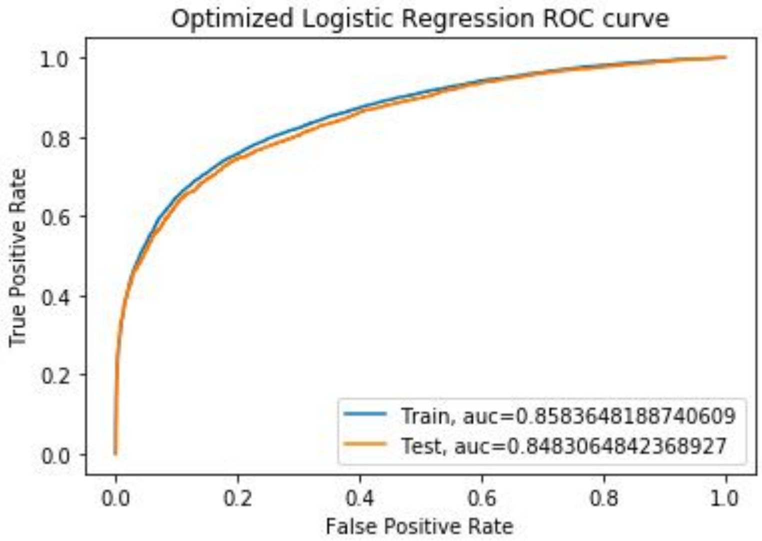
- Create class balance by randomly selecting equal amounts of normal and fraudulent observations
- Data shape:
    - 32,902 observations
    - 434 attributes



BASELINE MODELS

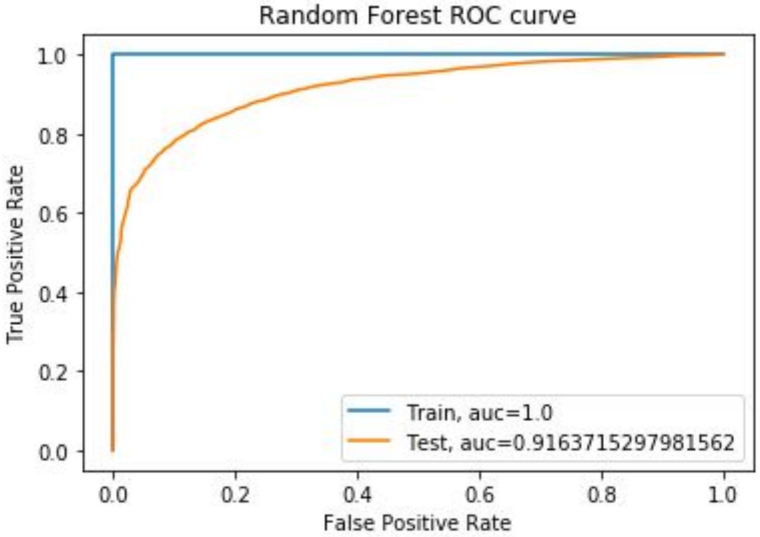
LOGISTIC REGRESSION

- Hyperparameters:
  - Feature selection with lasso (shrinkage method)
- ROC score of 0.848 and accuracy of 72%
- False negative rate 17%
- False positive rate 27%
- The computation time for logistic was relatively fast.

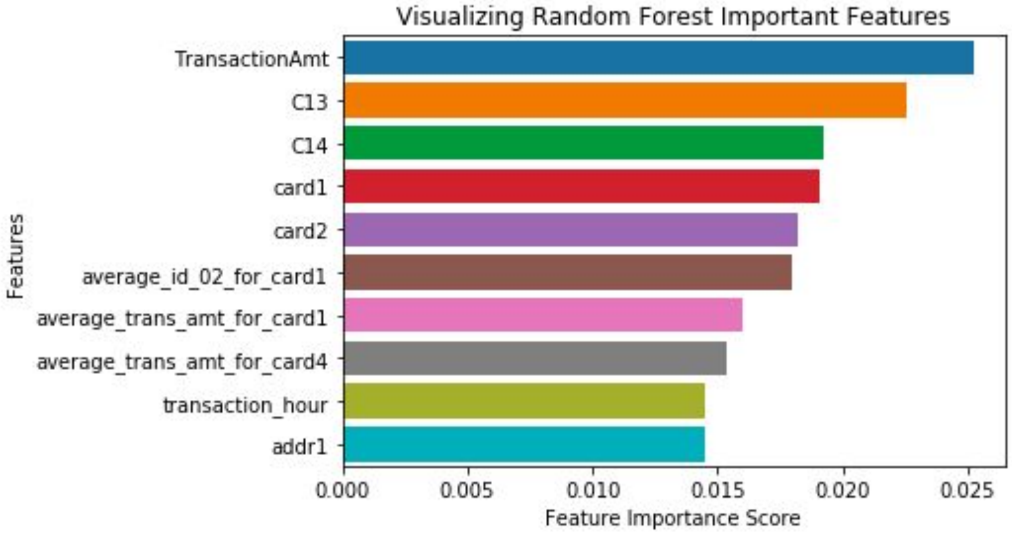


RANDOM FOREST

- Hyperparameters:
  - n_estimators = 100
- Performed well with a score of 0.916 and accuracy of 83%
- False negative rate 14%
- False positive rate 19%
- Longer computational time than logistic regression, but better performance



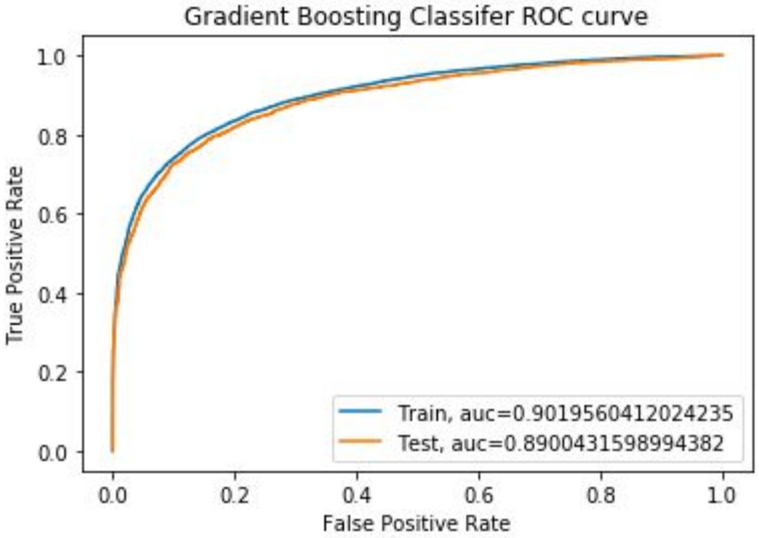
FEATURE IMPORTANCE



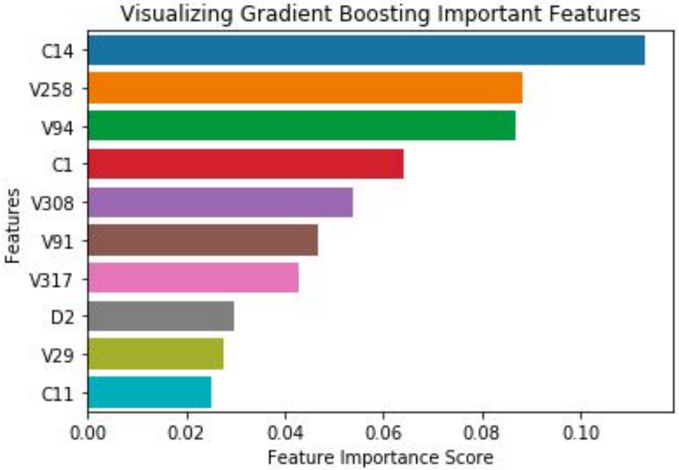
Transaction amount and counting matches appears to be the most important features in detecting fraud. Our feature engineered variables also made it in the top ten with interactions between transaction amount and card information.

GRADIENT BOOSTING

- Hyperparameters:
  - Random_state = 42
- The initial gradient boosting model has a score of 0.89 and accuracy score of 81%
- False negative rate 15%
- False positive rate 21%
- Computational time was significantly longer than other previous 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.

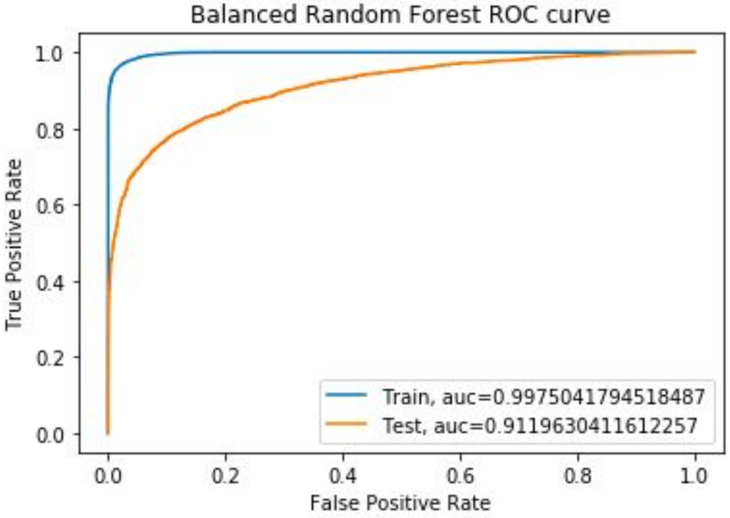
INITIAL MODEL RESULT

- Logistic regression model performly better than random guess with a ROC score of 0.84 on the test set
    - Poor performance in classifying normal transactions
- Random forest with all features performed that best with a ROC score of 0.91 on the test set
    - Best model in ROC score and lowest false negative rate
- Gradient boosting was a close match to random forest with a ROC score of 0.89 on the test set
    - Comparable to Random Forest, but slightly lower ROC score and higher false rates
    - Longest computational time

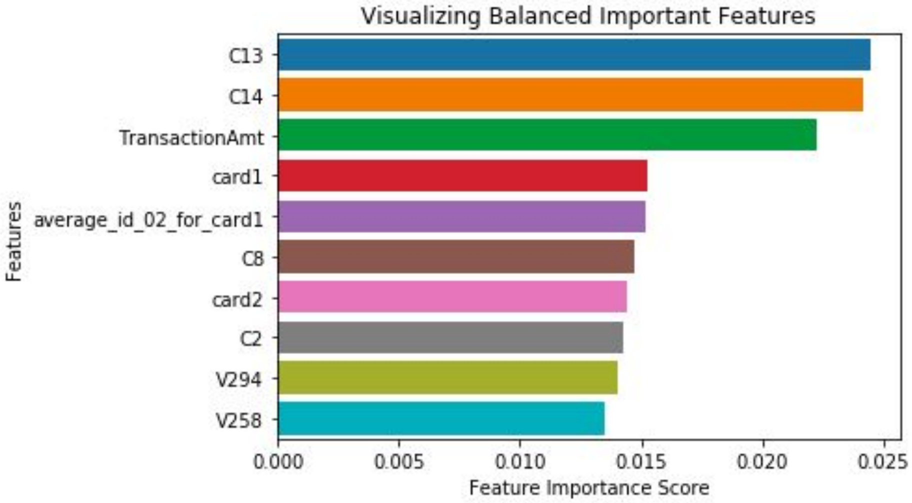
MODEL ON IMBALANCED DATA

BALANCED RANDOM FOREST

- Hyperparameters:
  - criterion: 'entropy'
  - max_depth: 100
  - n_estimators: 30
  - Class_weight = balanced
  - Random state: 42
- This model has a ROC score of 0.911 and accuracy of 81%
- False negative rate of 15%
- False positive rate 21%
- Very similar to initial Random Forest model



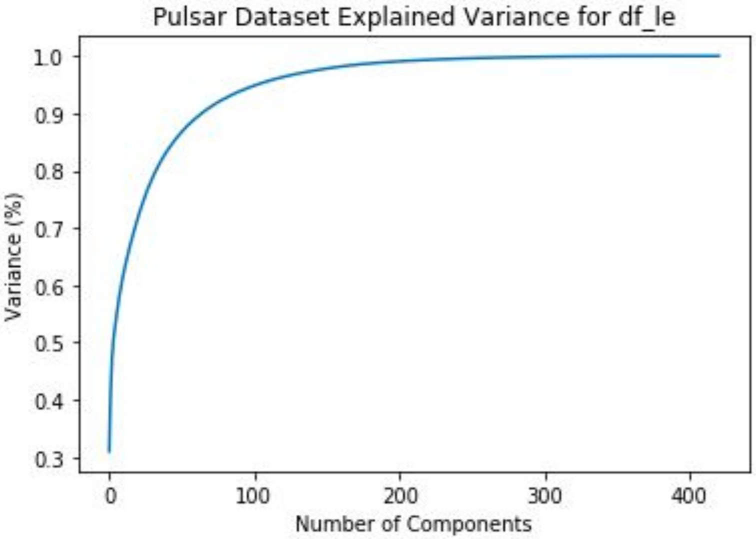
FEATURE IMPORTANCE



Similar to the previous random forest and gradient boosting models, count of matching card information and transaction amount are the most important features in detecting fraud

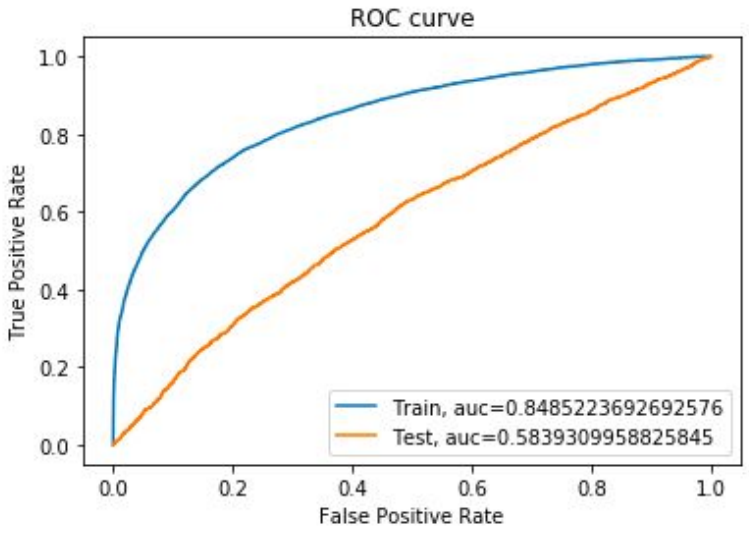
PCA FEATURES

- Try to improve performance of the models by reducing dimensions with PCA
- Around 80 components appears to capture 90% of variance



PCA FEATUREs

- All three models performed extremely poorly
- ROC AUC scores:
  - Logistic Regression: 0.58
  - Random Forest: 0.50
  - Gradient Boosting: 0.58



RESULTS

MODEL	ROC AUC TEST	FALSE NEGATIVE
Logistic Regression	0.848	0.17
Random Forest	0.916	0.14
Gradient Boosting	0.901	0.15
Balanced Random Forest	0.911	0.15

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