Customer Transactions Fraud Detection

- Background Information
- Question
- Key Trends
- Modelling Approaches
- Important Features
- Future Improvement

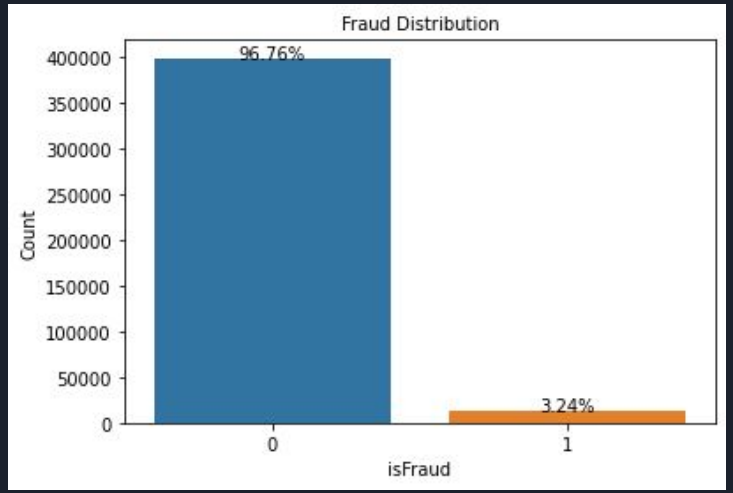
Background Information

- Vesta Corporation is seeking the best solutions for fraud prevention. They build fraud prevention systems which save consumers millions of dollars per year.
- IEEE Computational Intelligence Society (IEEE-CIS) works across a variety of AI and machine learning areas.
- They’re partnering together and seeking the best solutions for fraud detection.
- Researchers from IEEE-CIS would like to improve fraud detection accuracy and improve customer expereinces

Question

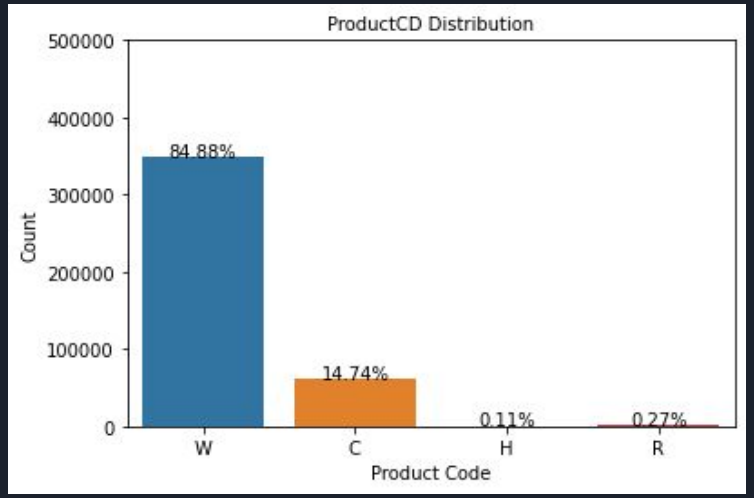
What model can be used to predict the probability that an online transaction is fraudulent?

Key Trends



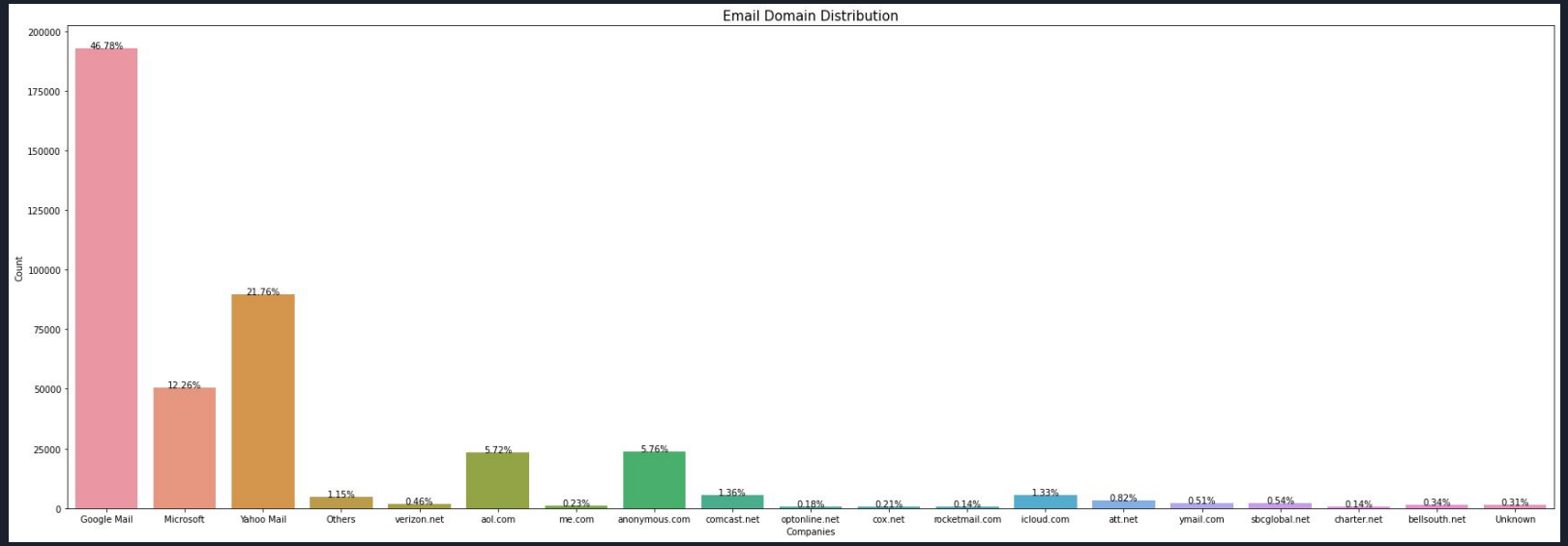
- 96.76% non-fraud transactions vs. 3.24% fraud transactions
- We have a heavily imbalanced dataset here, so we will apply some resampling technique such as SMOTE

Key Trends



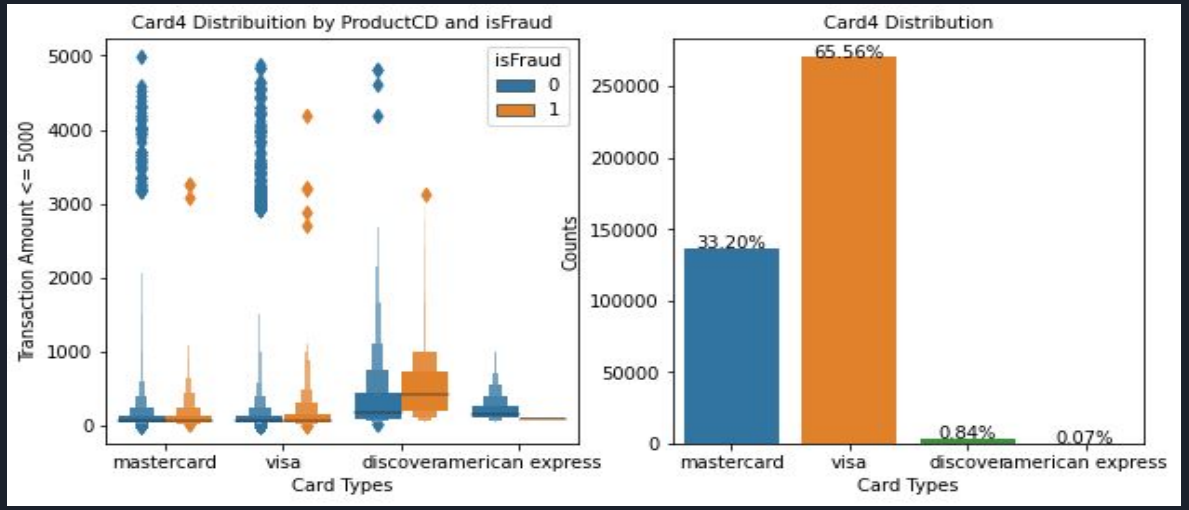
- 84.88% transactions have a product code of W, 14.74% transactions have a product ocde of C

Key Trends



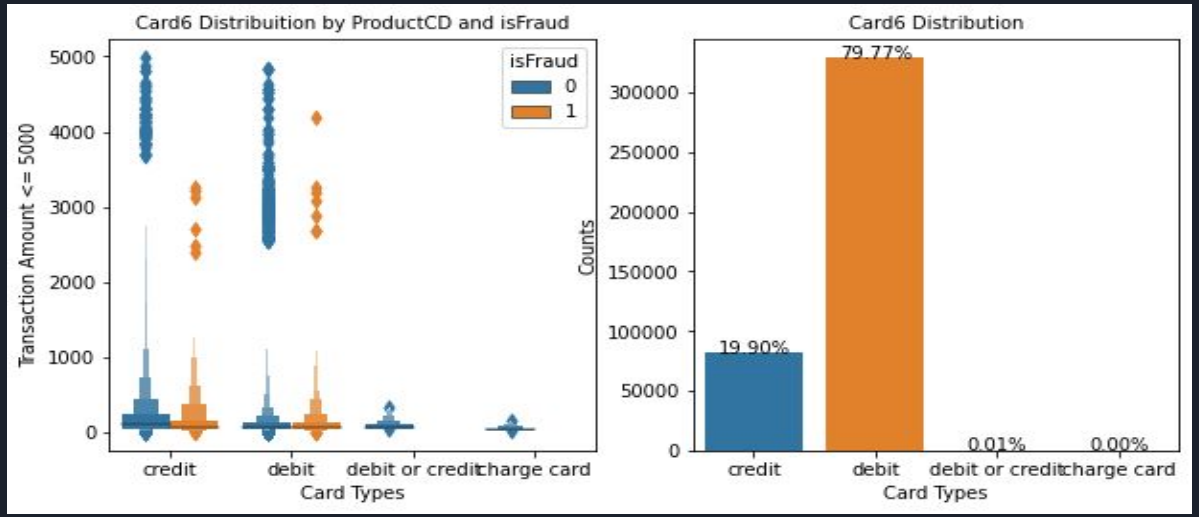
- Top three purchaser's email domain: Google, Yahoo, Microsoft

Key Trends



- Majority of purchases used visa and mastercard
- Only 0.07% of transactions made by Amex

Key Trends



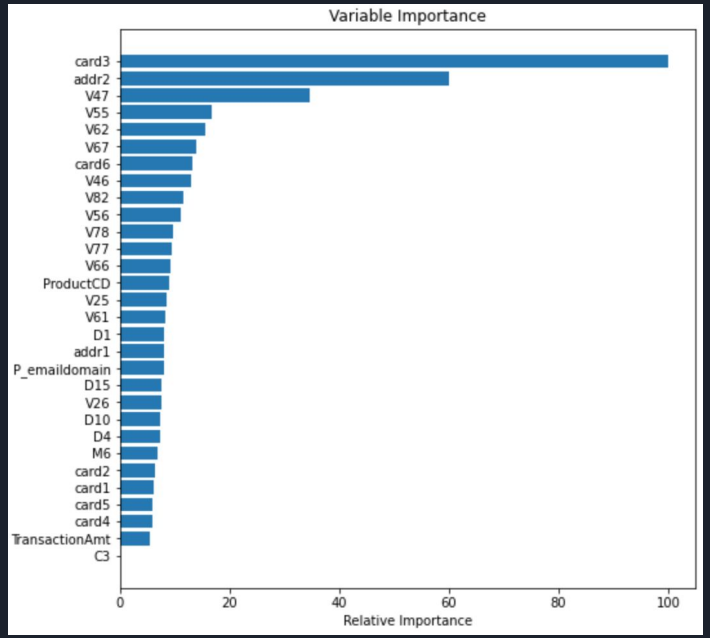
- Majority of purchase cards were debit card, followed by credit card
- Charge cards seem to be the safest card

Modelling Approaches

Model	Precision	Recall	F1	AUC ROC
Logistic Regression	0.483481	0.500000	0.491602	0.500000
Decision Tree Entropy, Max Depth 10	0.852852	0.634393	0.691286	0.634393
Random Forest	0.940277	0.702026	0.774226	0.702026
XGBoost	0.950389	0.730200	0.802431	0.730200

- Programming Package: Python (Sciki-Learn Library)
- 60% training set, 20% validation set, 20% testing set
- 4 models + parameter tuning when possible
- Best model: XGBoost, based F1 scores

Important Features



- Top three most important features: card3, addr2, V47

Future Improvement

1. Parameter tuning for several models were unsuccessful due to RAM constraints. Big data platforms such as Apache Spark may be useful in this case. Even if tuning was successful, we would not be expecting a huge jump on the performance
2. Resampling techniques other than SMOTE for dealing with imbalanced datasets could be useful.
3. More advanced data handling and modeling techniques might help to obtain a more accurate and consistent performance

Summary

- Apply machine learning algorithms to predict transaction fraudulent activities
- Actual distribution: roughly 3% fraud data - imbalanced dataset
- Most common purchase card issuer: Visa and Mastercard;
  Most common purchase card type: Debit and Credit
  Most common purchaser email: Google, Yahoo, Microsoft
- Best prediction model based on F1 score was XGBoost (80.2%) out of four models
- Some areas of improvements