Fraud Detection

Colin Shanahan,
Derek Jia,
Jenny Cho,
Marcia Torrico,
Noah Schumacher

Objective

Challenge: The company needs to flag potential new fraud for further review as it comes in so it can be triaged by most pressing (and costly) transactions.

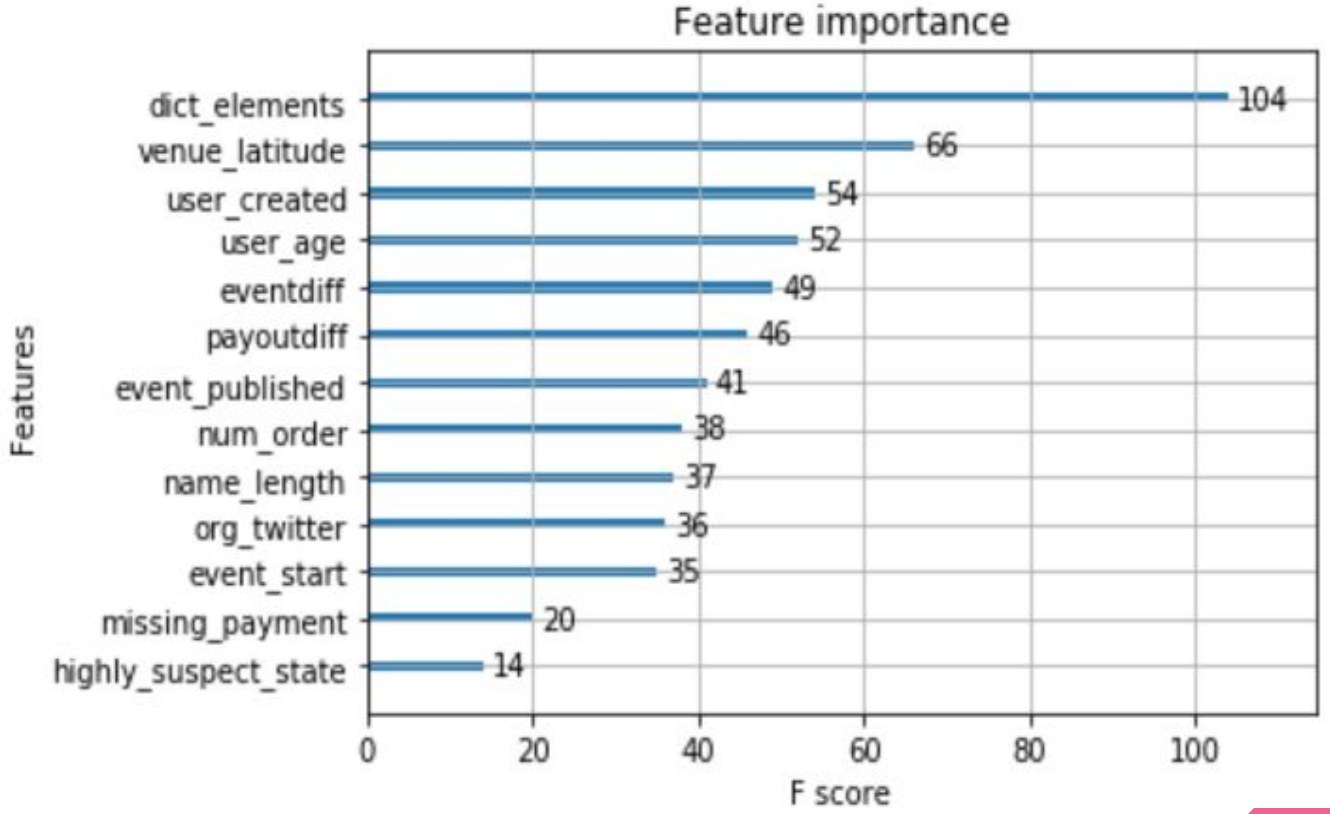
Given a JSON file with transactions that are fraud or not, and other features, create a web based front-end with machine learning back end to enable quick triage of potential new fraud.

The web front end needs to be usable by a non-technical audience for triaging, and flag each transaction as low, medium or high risk.

Key Variables/Variable Transformation

- Transformed Variables with High Feature Importance:
  - Dict_eleprevious transactions
  - States with high fraud rates
  - Difference between when the event was published/ended
  - Payoutdiff: Difference between when the payout date/created
- Interesting Default Variables with High Feature Importance:
  - Venue_Latitude
  - Payee organization has a twitter account
- Low Feature Importance:
  - NLP tasks on the description variables
  - difference in country/state
  - amount between previous / current transactions

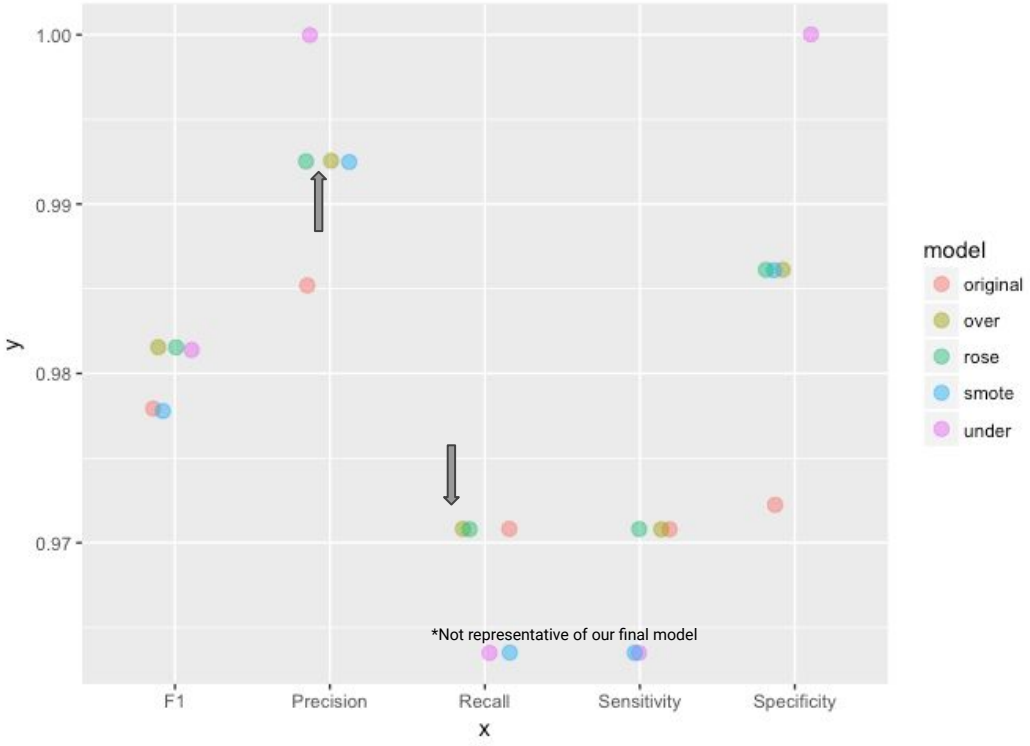
Feature Importance



Imbalanced Dataset

"However, one of the biggest stumbling blocks is the humongous data and its distribution. Fraudulent transactions are significantly lower than normal healthy transactions i.e. accounting it to around 1-2 % of the total number of observations. The ask is to improve identification of the rare minority class as opposed to achieving higher overall accuracy."

How Does Various Over/Under Sampling Techniques Affect Model Score?

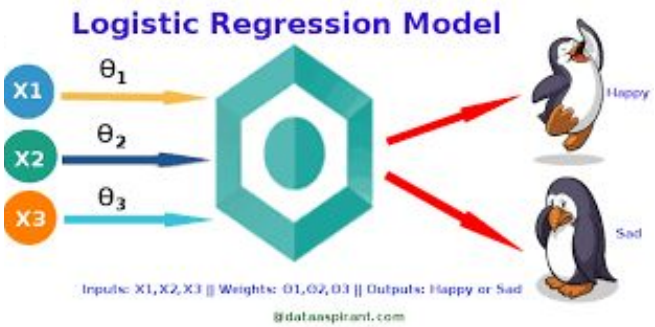


The most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique.

Goal: Increasing precision at the cost of recall

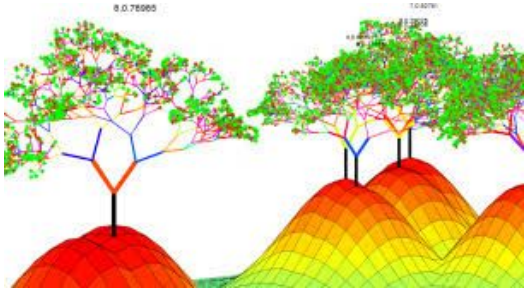
Various Models

Logistic Regression



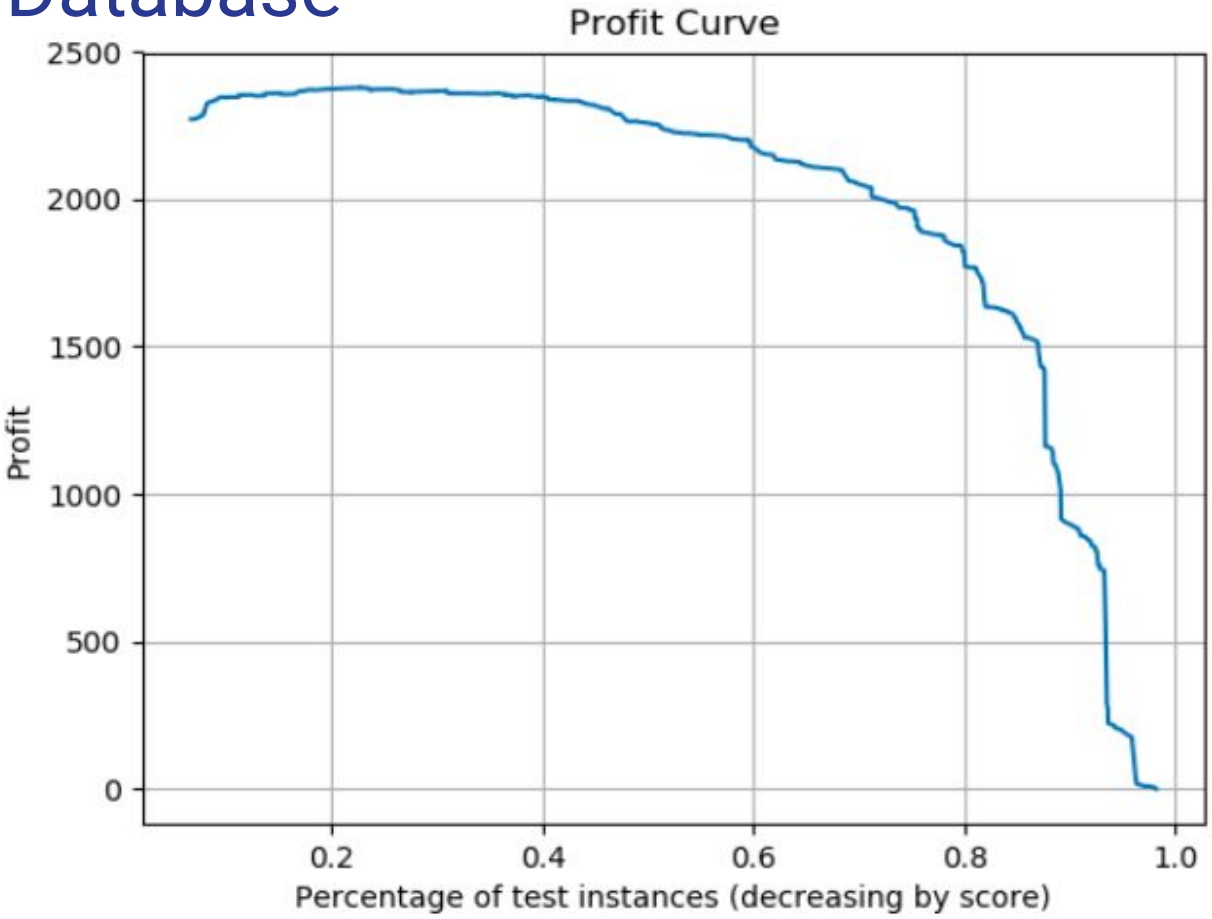
Threshold: 0.2167
TP: 2555 | TPR: 0.986
FP: 777 | FPR: 0.296
FN: 36 | FNR: 0.0139
TN: 1850 | TNR: 0.704

Random Forest



Gradient Boosting/ XGBoosting

Database



We assume it was 25 times more costly to fail to predict fraud than to inquire about legitimate events.

Web App/Getting Live Data/Website

Be right back, we are going to switch computers

Questions?