Classifying Fraud

Group 1: Alex, Helnaz, Jon, Sherry, Trevor

The Scope

1

Business Goal:
-Correctly predicting whether or not an event is fraudulent, in order to remove that event

-Identify low,medium and high risks for risk assessment

2

Metrics:
-Total Expected Profit

3

The Data:
-14337 entries
-44 Columns

4

Unbalanced Classes:
-Not-Fraudulent: (91%), 13044
-Fraudulent: (9%), 1293

EDA

Previous Payouts
-998/1293 fraudulent events have no previous payouts

Delivery Method
-Delivery Method 0 had 14% fraud at 8K events, meanwhile Method 1 had 1% at 5K

Email Domains:
-131 email domains with above average fraud

EDA Text

Name

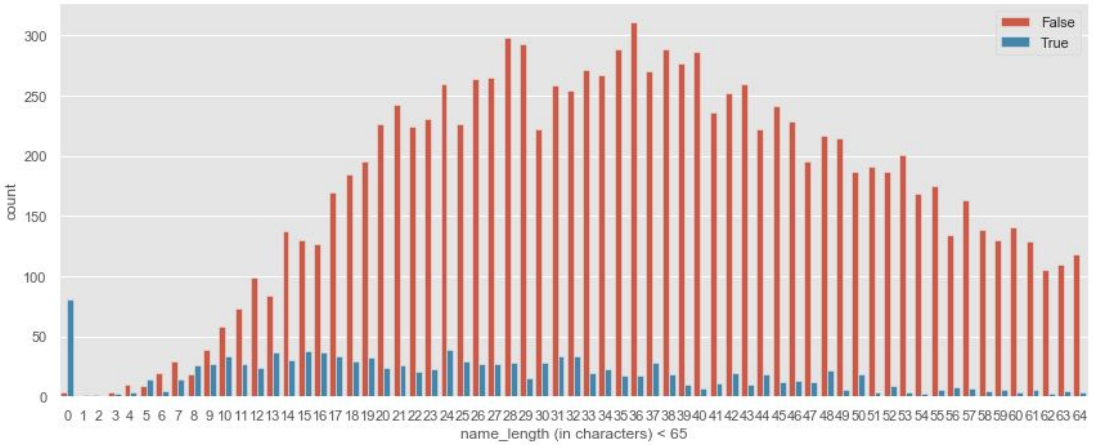
-Name length for fraudulent cases tended to be shorter (or empty) more often than non-fraudulent.

Description

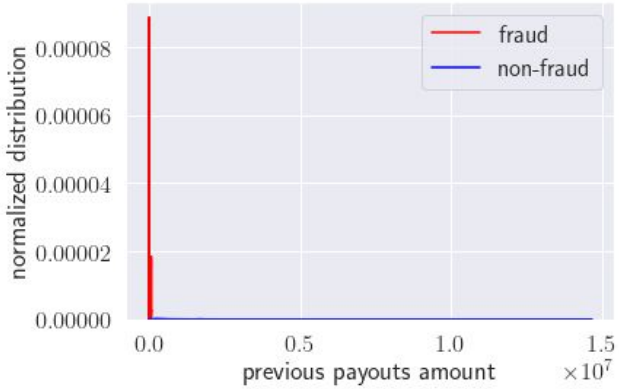
-Used Beautiful Soup & TFIDF to pull topics

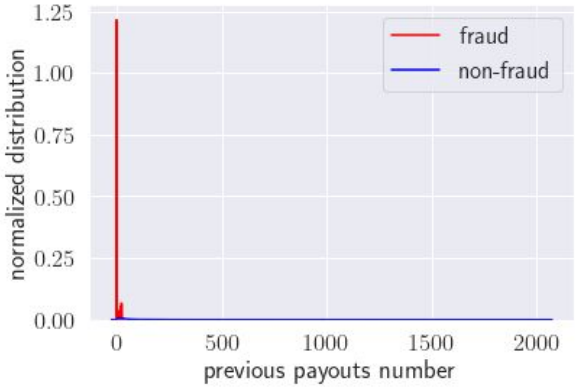
-Topics in Fraudulent Events were mainly about night-time typical events, and click-bait words like “free prizes” & “vip”:

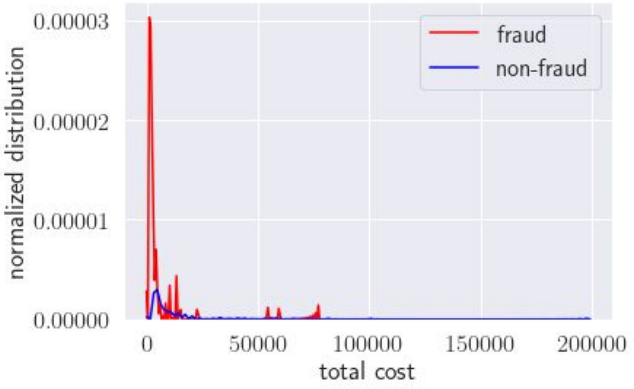
['free', 'prizes', 'rounds', 'soda', 'minutesef']
['party', 'hope', 'cheerful', 'buy', 'shall']
['event', 'night', 'music', 'vip', 'special']
['club', 'bar', 'pre', 'boot', 'clothes']
['year', 'concert', 'presents', 'end', 'kash']



Feature Engineering







Evaluation Metrics

	Predicted Fraud	Predicted Non-Fraud
True Fraud	$0	-$318
True Non- Fraud	-$355	$71

Considerations:
- FN: Average cost per fraudulent cases

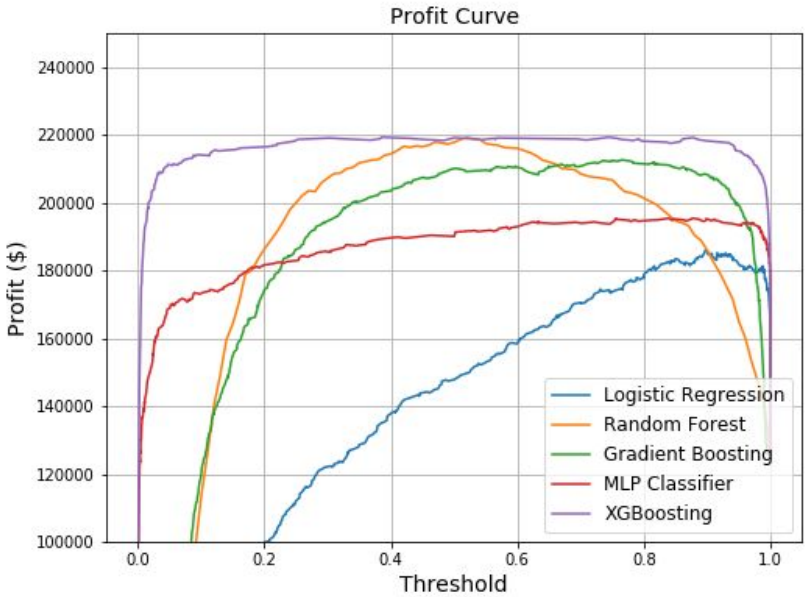
- TN: Average revenue for non-fraudulent cases (Inspired from EventBrite)

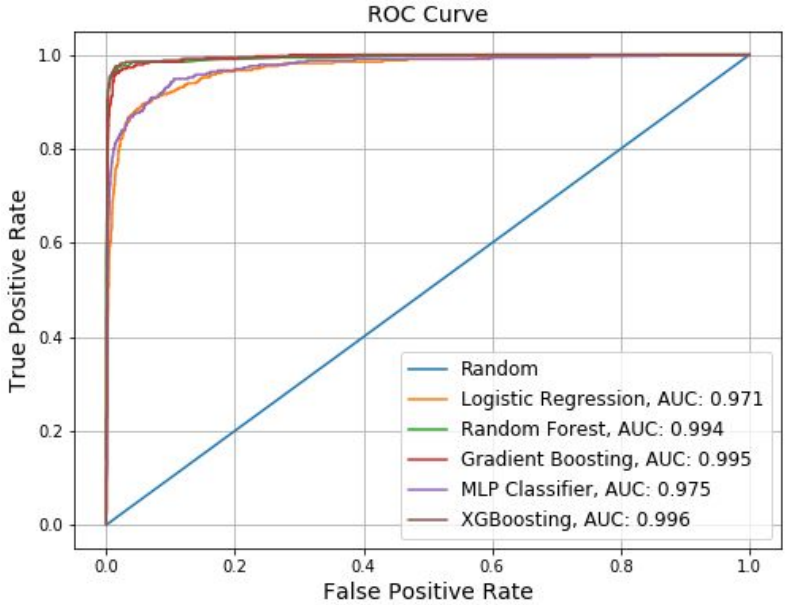
- FP: 50% of customers flagged never use the service again. Assume 10 events per customer, using average revenue for non-fraudulent cases

-TP: No loss, no profit

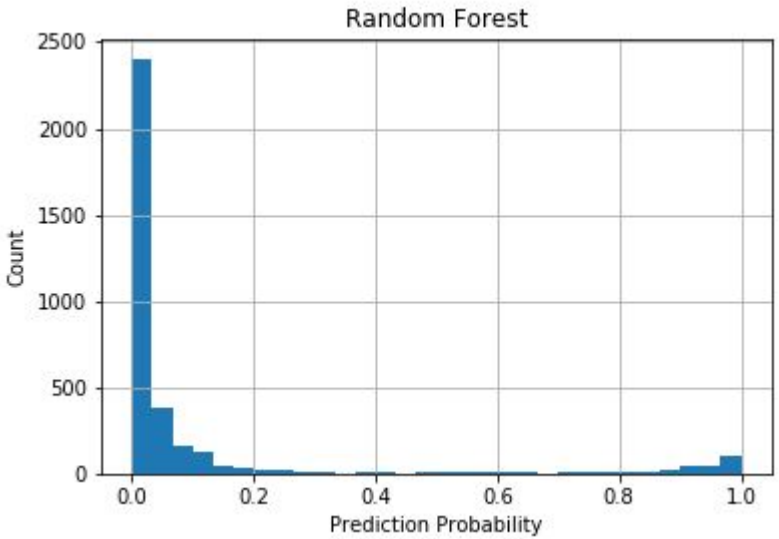
Based on the CB Matrix, we decided to optimize for the following metrics:
-Total Expected Profit

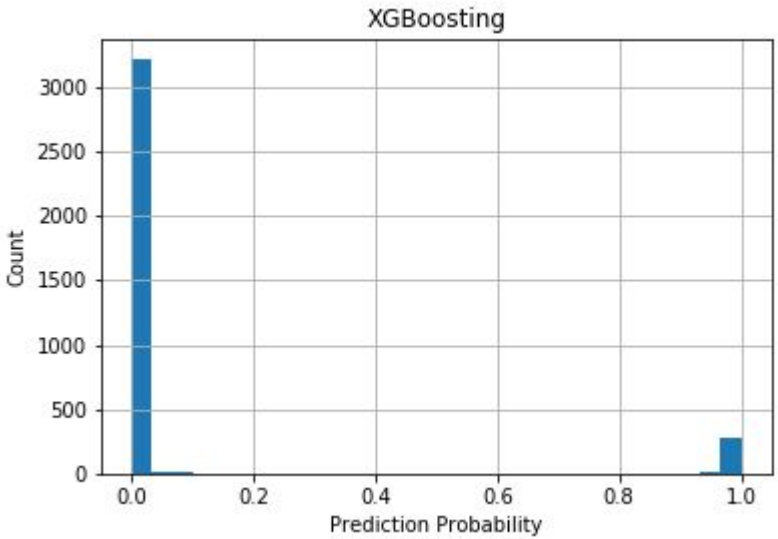
Model Selection





Choosing XGBoosting vs Random Forest





It Depends on the Business Context!

Final Models

If you want to categorize the Positive Class (low risk, medium rick, high risk)

Random Forest Final scores:

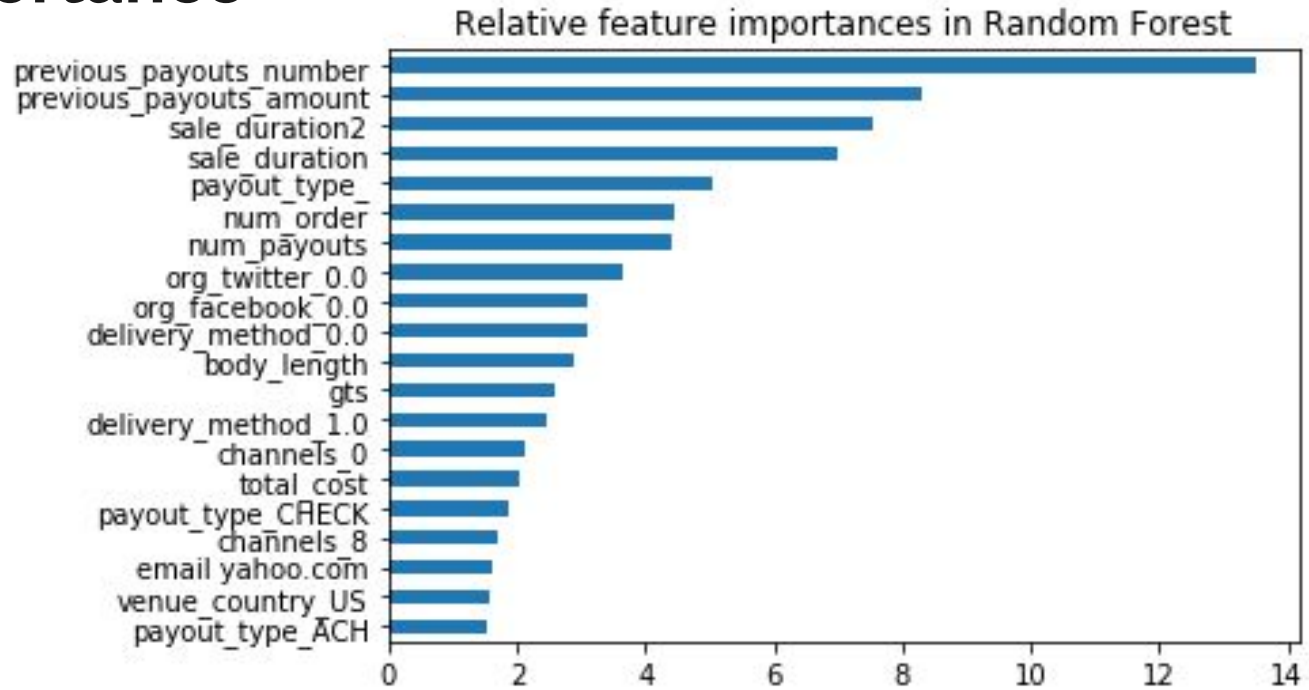
- Profit Curve Max= $219,142 at a Threshold of 0.520
- Test ROC AUC = 0.994
- Test Precision = 0.968
- Test Recall = 0.927

If you only care about Positive Class or Negative Class

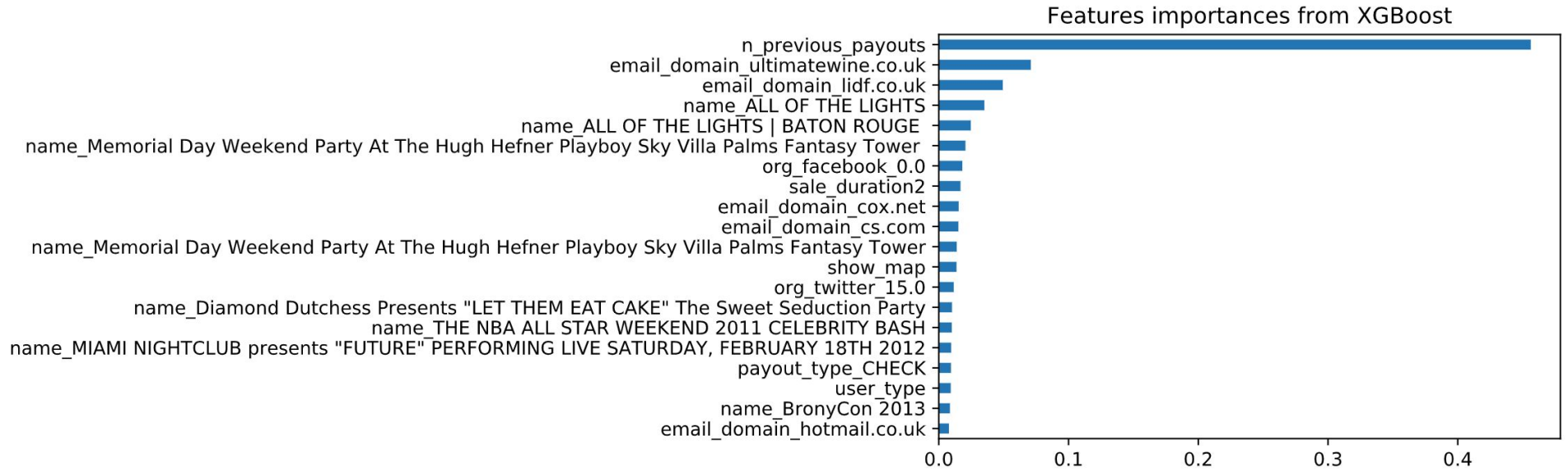
XGBoost Final scores:

- Profit Curve Max= $219,250 at a Threshold of 0.744
- Test ROC AUC = 0.996
- Test Precision = 0.971
- Test Recall = 0.924

Feature Importance



Feature Importance



Flask App

Progress made:

- Generated New Data Points
- Made a prediction via pickled model
- Store new data point with prediction in MongoDB

Not yet complete:

- Create Website
- User-interactive Dashboard for real-time fraud flagging

Thank you.