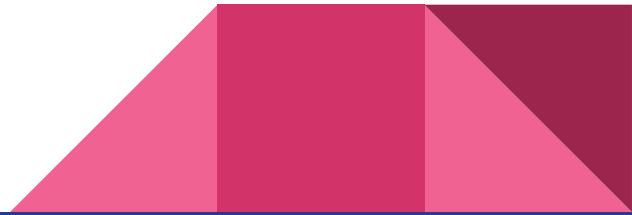
Fraud Detection

Anita Banser
Arash Shushtarian

Overview

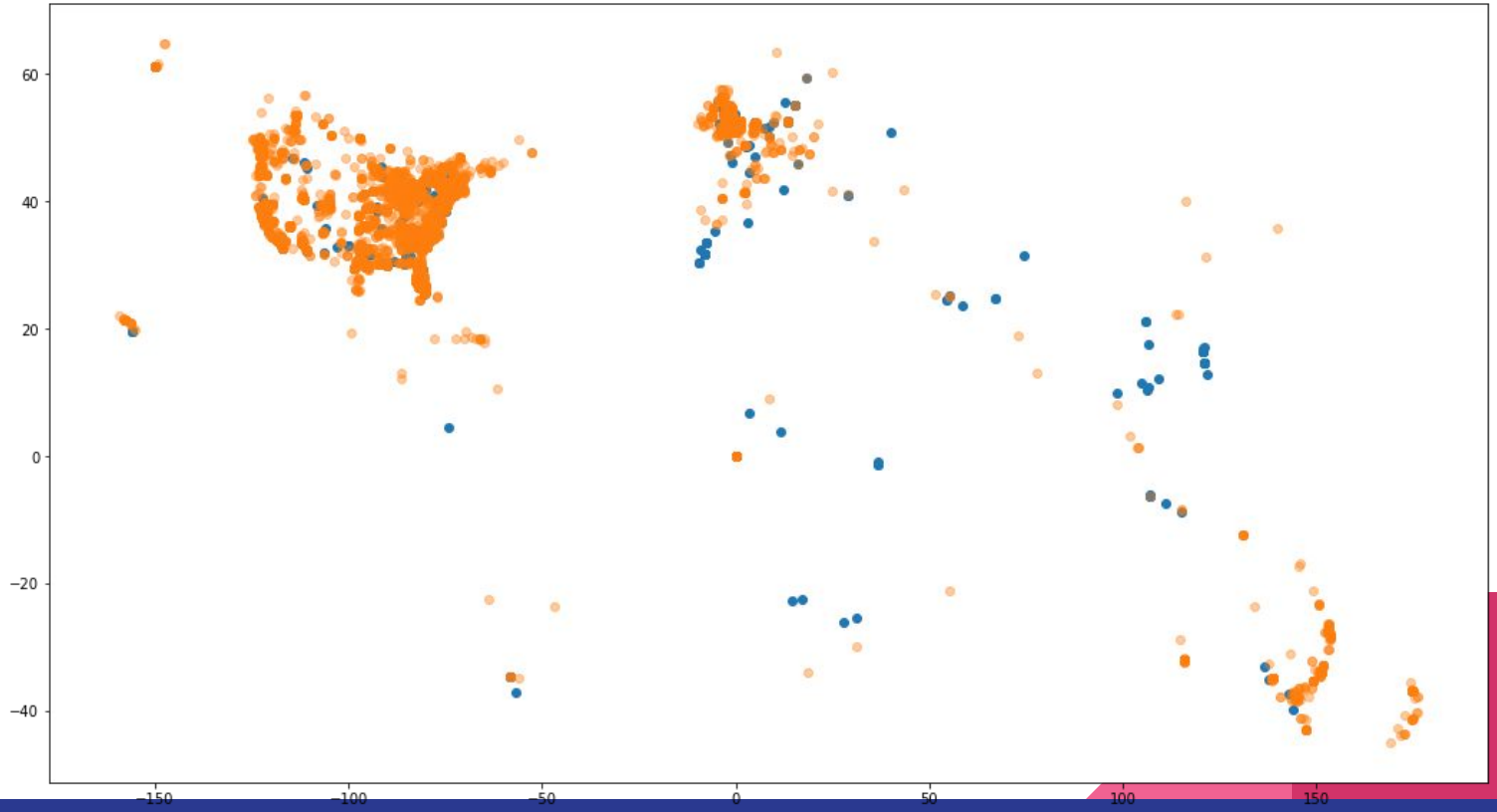
- Exploratory Data Analysis
- Methodology
- Model Selection and Advancement
- Evaluation Metrics
- Future Improvement and Implementations



Exploratory Data Analysis

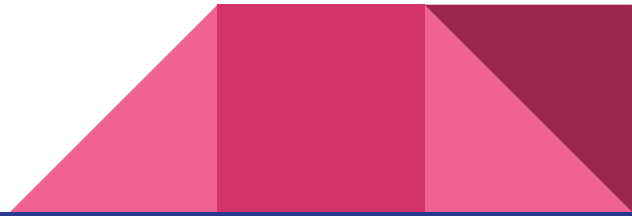
- Relationship between the following timestamps
  - Created Time
  - Published Time
  - Started Time
  - No Relationship with End Time
- Number of frauds in US and outside of USA
  - US: False - 8615 True - 619
  - Outside US: False - 4429 True - 674

Defining Fraud Zone



Exploratory Data Analysis Cont.

- Defined a fraud zone for outside of US
- Valuable Information in “Ticket Types”
  - Availability
  - Cost
  - Event ID
  - Quantity Sold
  - Quantity Total
- Natural Language Processing
  - Description
  - Not enough information gain



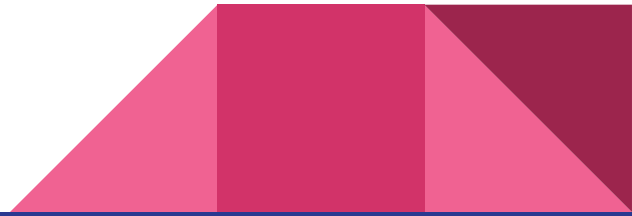
Model and Advancement

Models

- Logistic Regression
- Random Forest
- Gradient Boosting and etc.

Random Forest

- Cross-Validation
- GridSearch



Evaluation Metrics

- Accuracy of Model - Not Reliable
  - Score of 98.64%
- Confusion Matrix

```
0.986398828207
             precision    recall  f1-score   support

      False       0.99      1.00      0.99     13044
       True       0.95      0.89      0.92      1293

avg / total       0.99      0.99      0.99     14337

[[12988    56]
 [  139  1154]]
tn 12988, fp 56, fn 139, tp 1154
```

Future Improvement Implementations

- Natural Language Processing to create [zeros/ones] feature
- Synthetic Minority Oversampling Technique
- Improving missing value replacement
- Getting information about features from the Company
- Using Deep learning

Questions
Thank You!