Implementing Our Fraud Detection Model

...

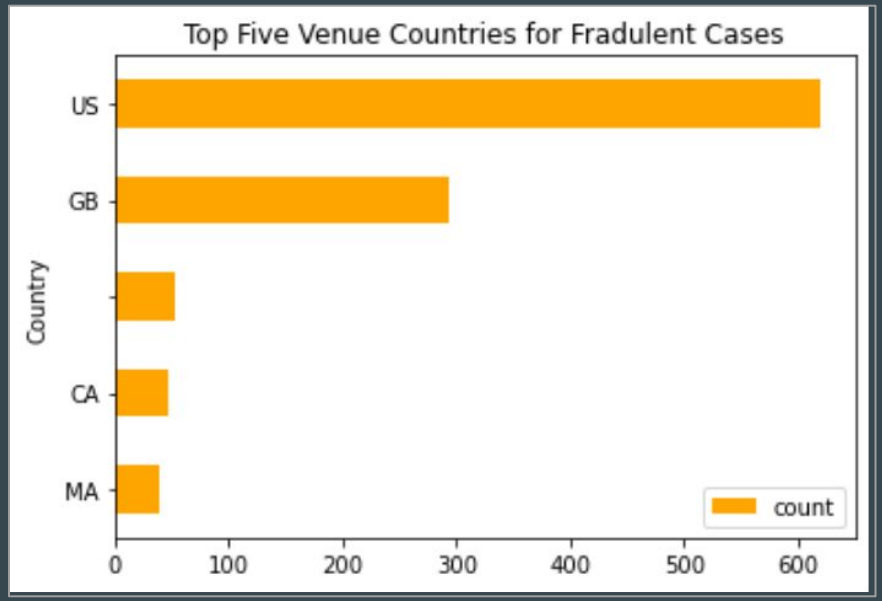
Team Members: Eric Lewis, Jeff Baksa,
Marc Mares, Vamsidhar Vurimindi

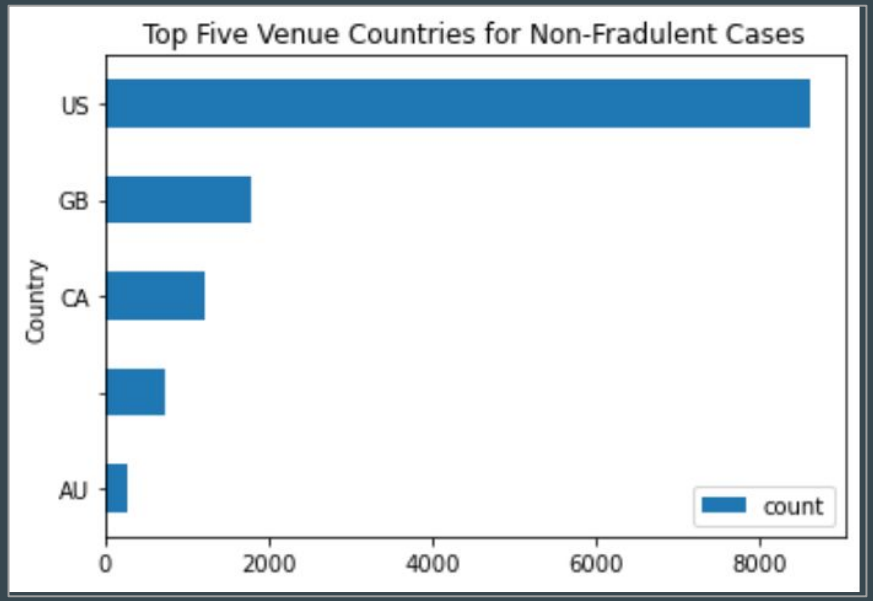
The Problem

E-commerce fraud is found more frequently that you expect!

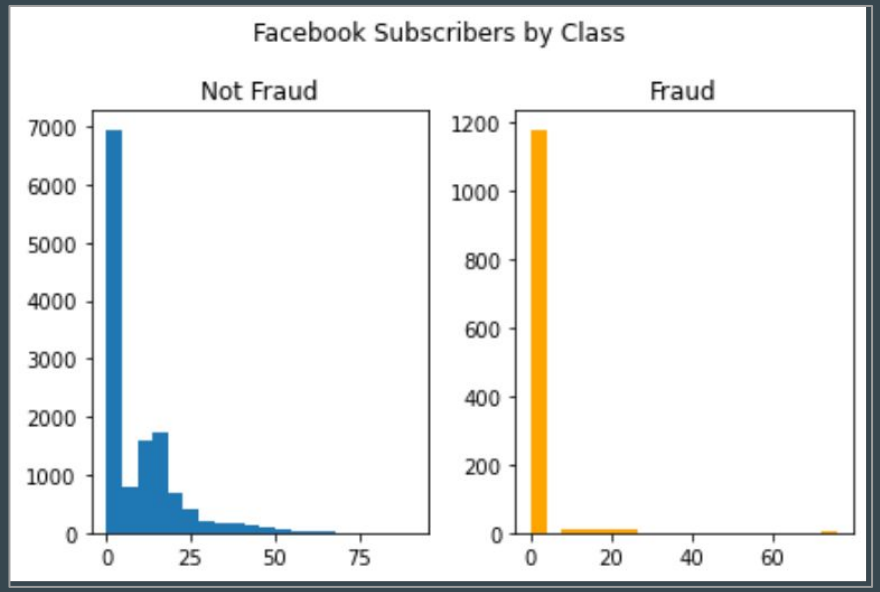
- **>10%** of the cases found in our data are flagged as **fraud** **(*imbalanced data set)***
- Want to catch fraud **before** users are able to purchase tickets to the event
    - Saving you credibility and money
- Saves people time manually checking

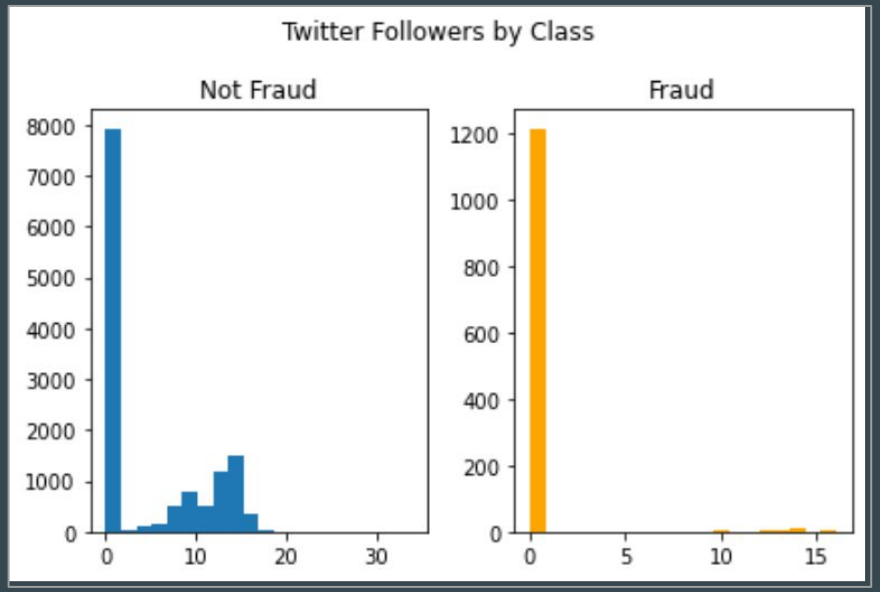
EDA - Fraud vs Non-Fraud Events





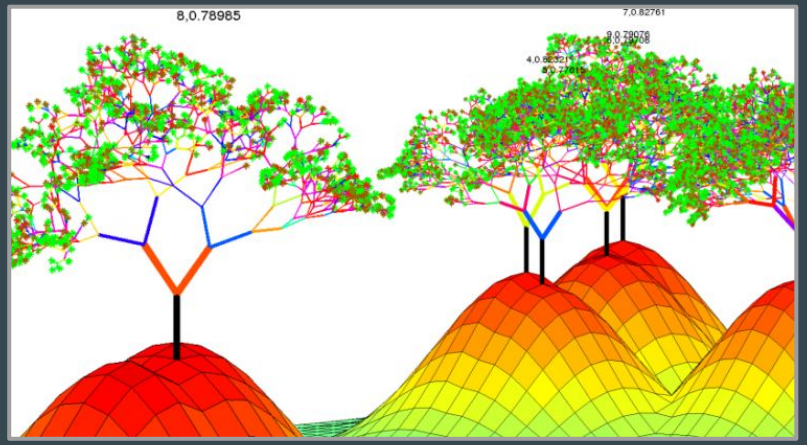
EDA - Fraud vs Non-Fraud Events





Model: Random Forest

- Others attempted:
  - Dummy Classifier
  - Logistic Regression
  - Decision Tree
  - Stochastic Gradient Descent

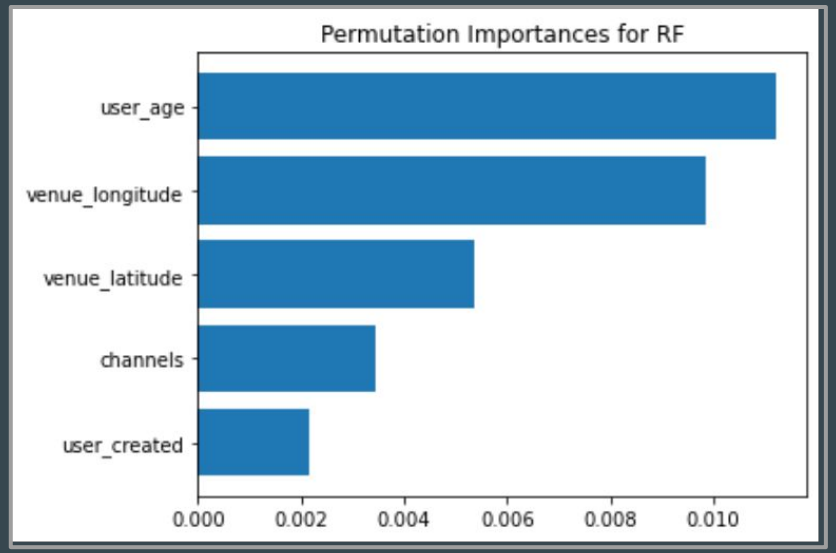


- Hypertuned RF - randomized search
  - Improved even further

- Compared precision, accuracy, and recall on all models
  - RF best combination of the three metrics

Factors/Features Influencing Prediction Model

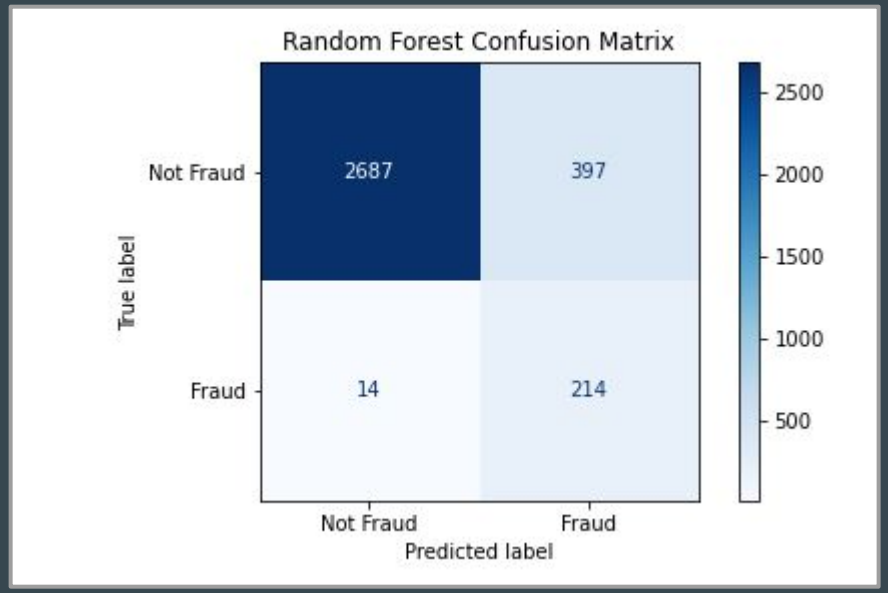
1. Body Length - how long the page/description of the event is
2. Channels - different platforms the event is able to get in contact with users
3. Delivery Method - different ways of receiving the invitation/advertisement
4. User Age - length of account life (days)
5. User Created - unix time of account
6. User Type - different type of platforms
7. Venue Latitude - location
8. Venue Longitude - location



Our Model

On testing data:

- Maximized recall
  - Of the actual fraud cases, we were able to identify 94%
- Minimizing false negatives
  - Predicting an event as legit when it's really fraudulent
- Retains customer satisfaction and credibility
- Slightly over-cautious when predicting non-fraud events



Predicted Probability

- Probability > 0.7
  - High risk of fraud
- Probability < 0.7 & Probability > 0.05
  - Medium risk fraud
- Probability < 0.05
  - Low risk of fraud

- Recommendations:
  - **High risk** events **should be addressed immediately** - pause events from site, their listing should be improved, reinstate if proven to be legitimate
  - Medium risk events should be sent **suggestions to improve the listing** so they won't be flagged in the future
  - Low risk events can be **left alone**

NLP Analysis

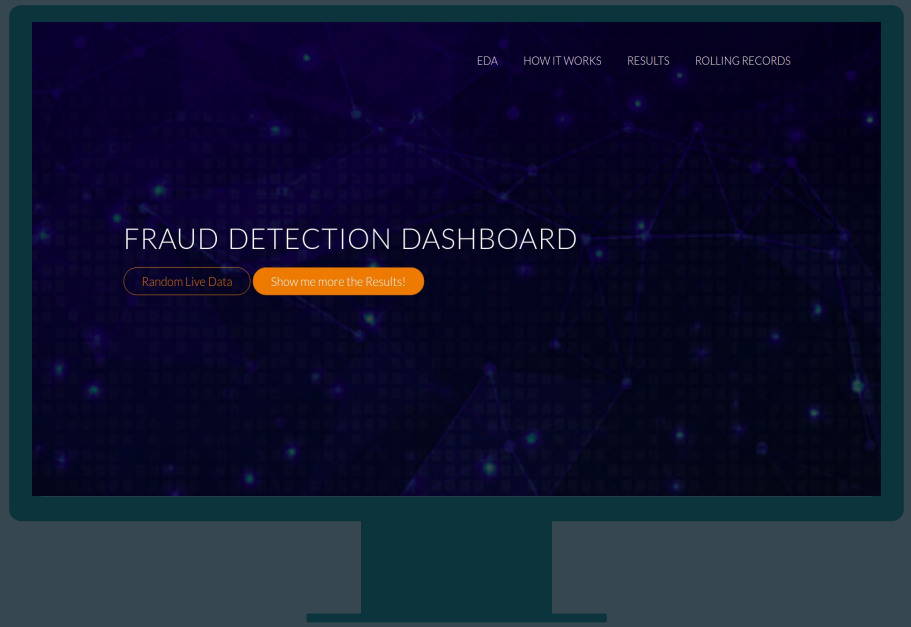
Using LDA:

- Initial analysis showed words used to exploit people's emotions

#	Top 10 Words Used for Fraud Cases									
1	event	wine	one	time	help	party	new	us	please	get
2	music	free	event	tickets	live	dj	course	communities	jardin	help
3	course	us	tickets	event	one	workshop	training	first	please	people

Display Dashboard

- Internal Company Website
  - Allow users to interact with live data
    - Manually flag events they think are fraud
  - Running log of events



Steps Going Forward

- Connect Flask with the live database
  - Live scraping updating website
- Customize the dashboard for users to interact with on the flask app
  - Phone/Email Alerts for Highly Suspicious Activity
- Continue NLP analysis of the descriptions of the events
  - Including non-fraud cases
- Profit gained analysis of correctly flagging an event

Tools/Technologies

Thank You