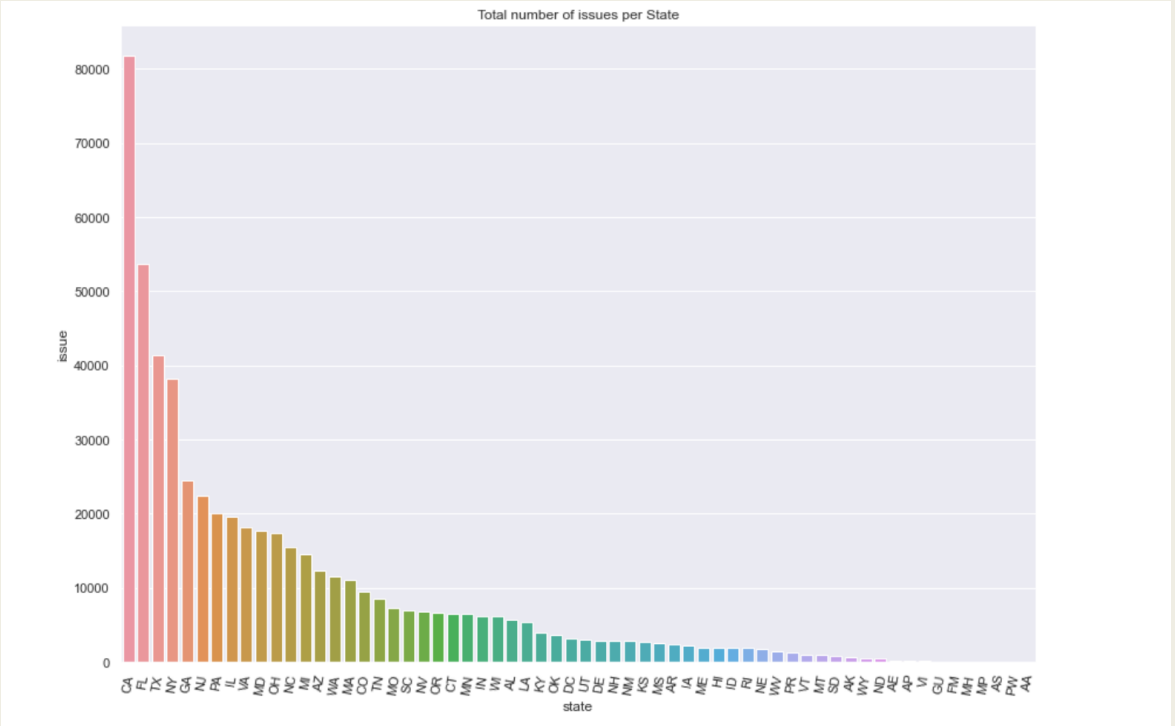
CONSUMER COMPLAINTS DATA ANALYSIS

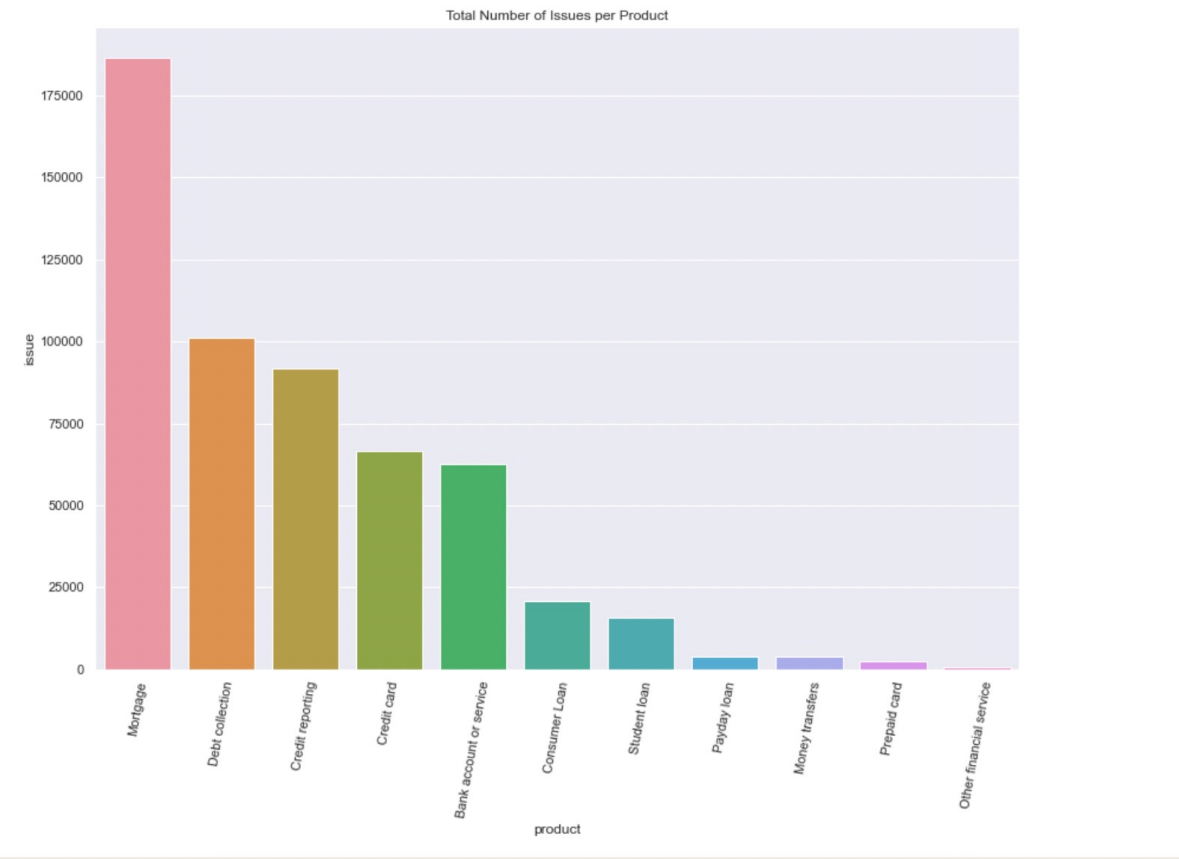
Aditya Asthana

Data insights from Consumer Complaints (States)

- Based on the visualization, we can determine that California had the highest number of complaints
- California has 33% more complaints compared to the next nearest state (Florida)
- Majority of complaints were spread across California, Florida, Texas, and New York
- Since complaints are seen to be state related, I used it as a feature in my machine learning model

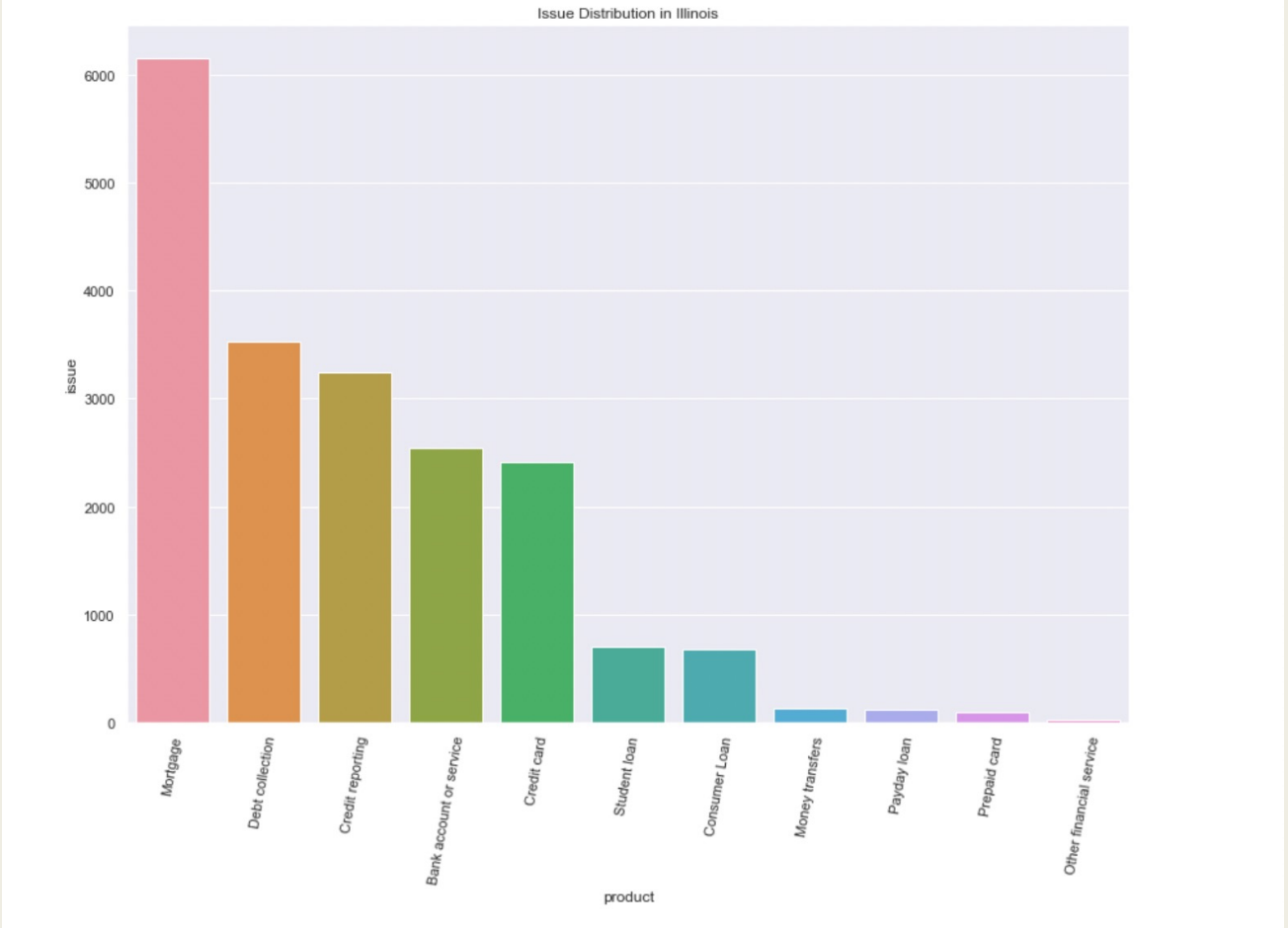


Data insights from Consumer Complaints (Products)



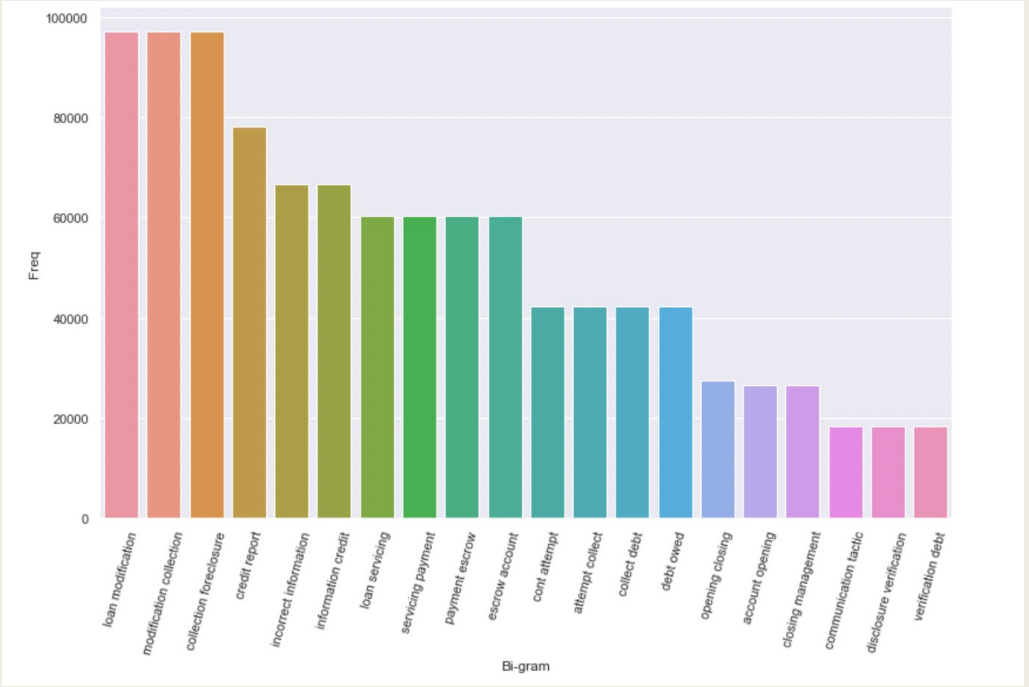
- Based on the visualization, we can determine that mortgage based products had the highest number of complaints
- Mortgage has 45% more complaints compared to the next product (debt collection)
- Majority of complaints were spread across mortgage, debt collection, credit reporting, credit card, and back accounts
- Since complaints are also seen to be product related, I used it as a feature in my machine learning model

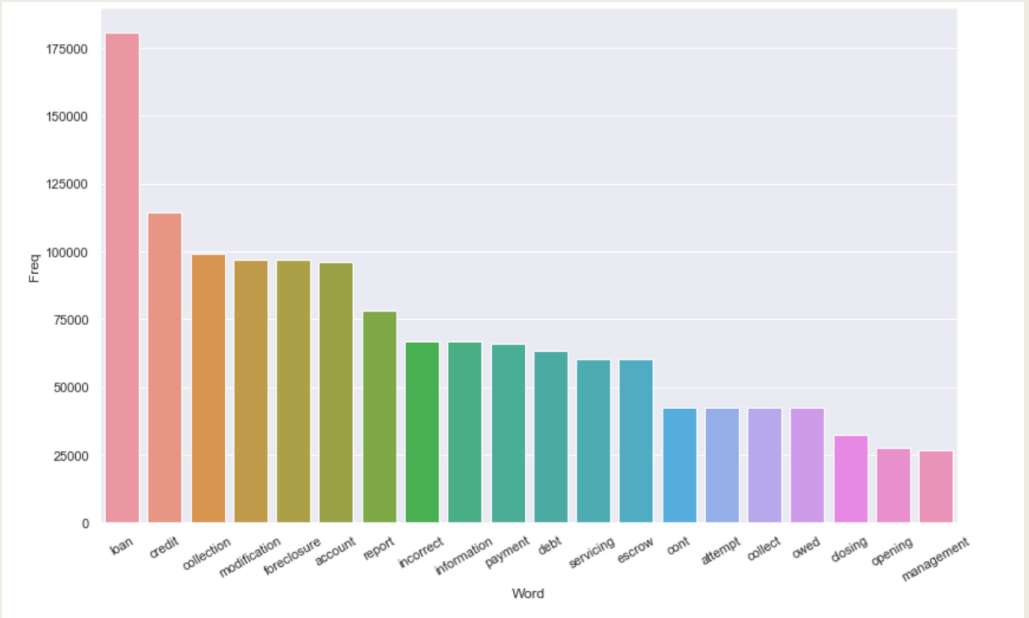
Sample Visualization for Products with the Most Issues in Illinois



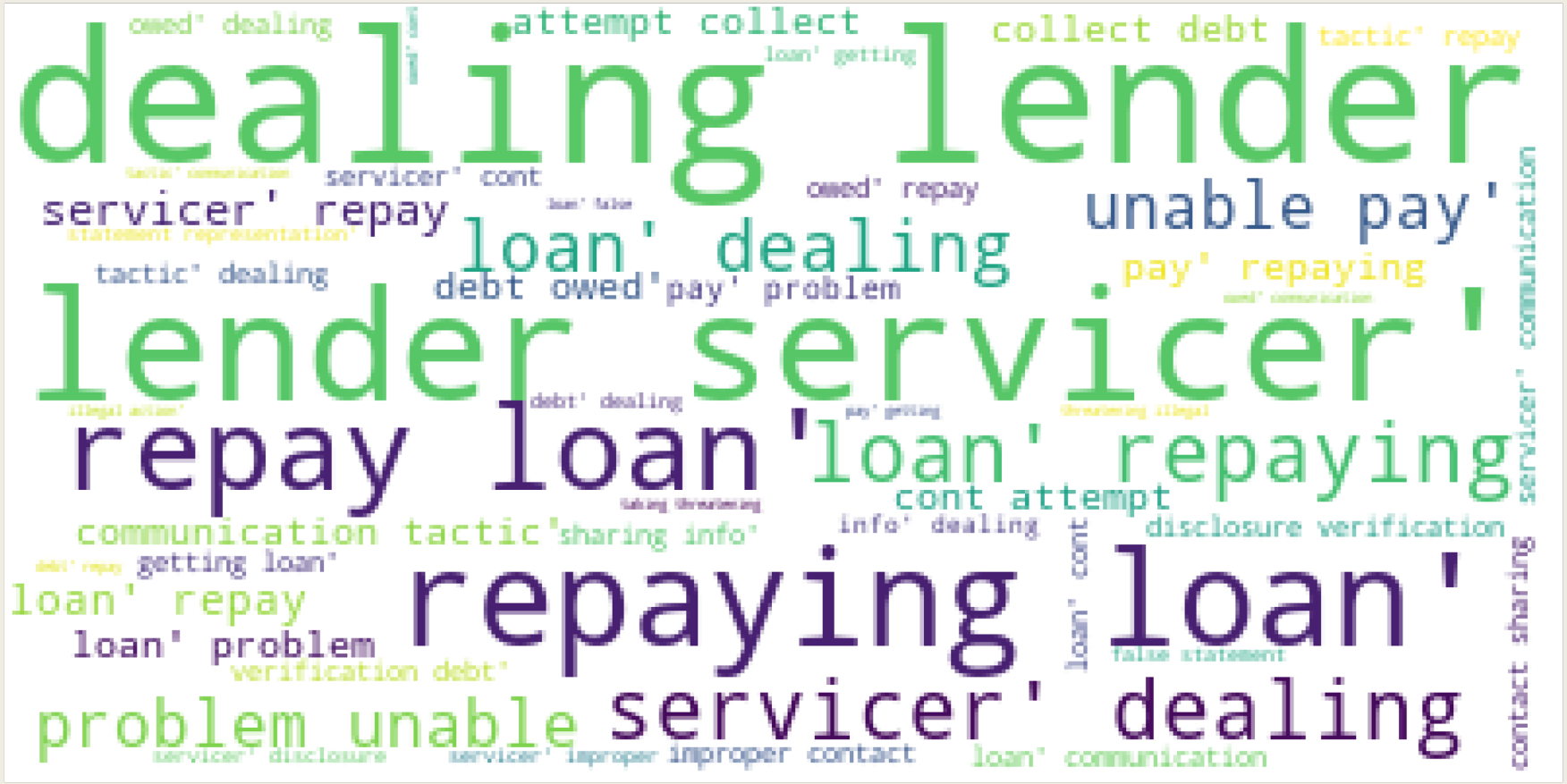
Data insights from Consumer Complaints (Top Issue Words)

- Based on the visualization, we can see the top words (unigram, bigram) for all the issues in the original data
- The top bigram word is “loan modification” and the top unigram word is “loan”
- I wanted utilize these words in my model, however I was unable to use these as features because of column explosion due to high unique word count
- Word embeddings can be incorporated to reduce dimensionality





Sample Visualization of Various Key Issues for Navient Solutions, Inc.

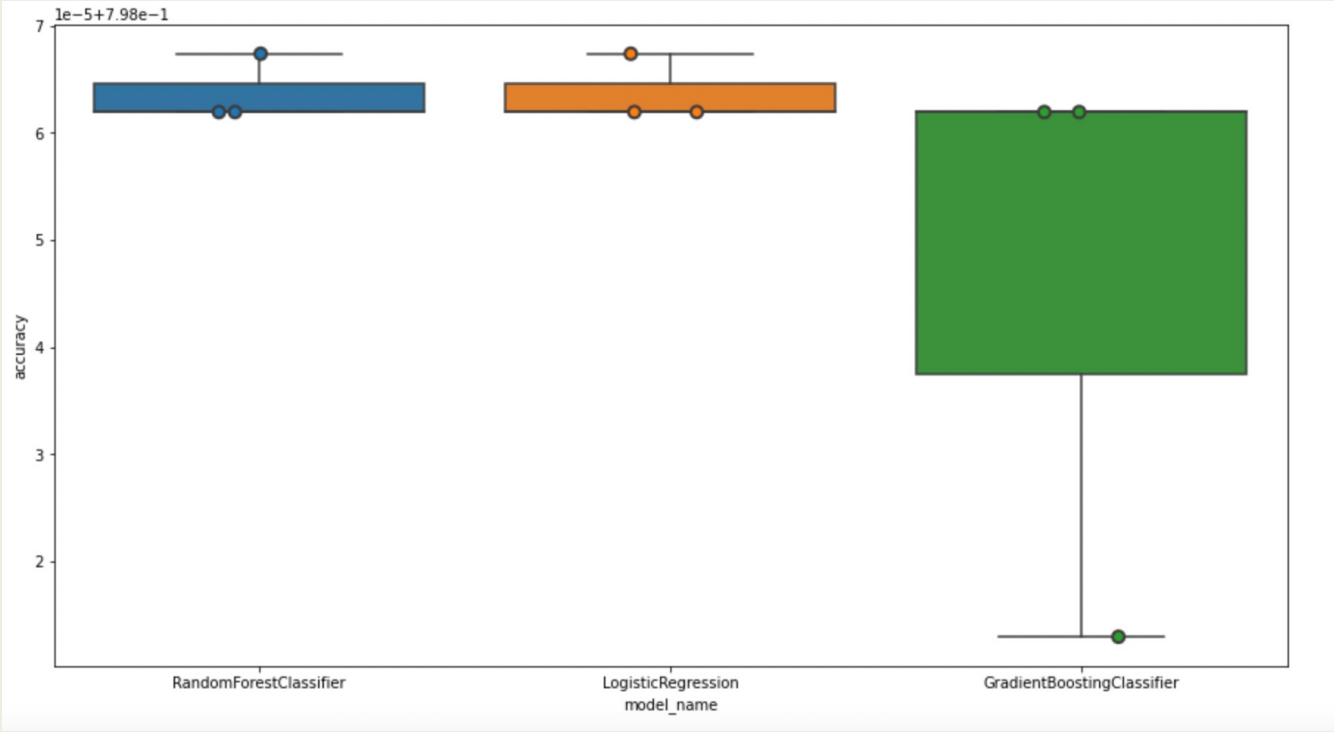


Machine Learning Model Details

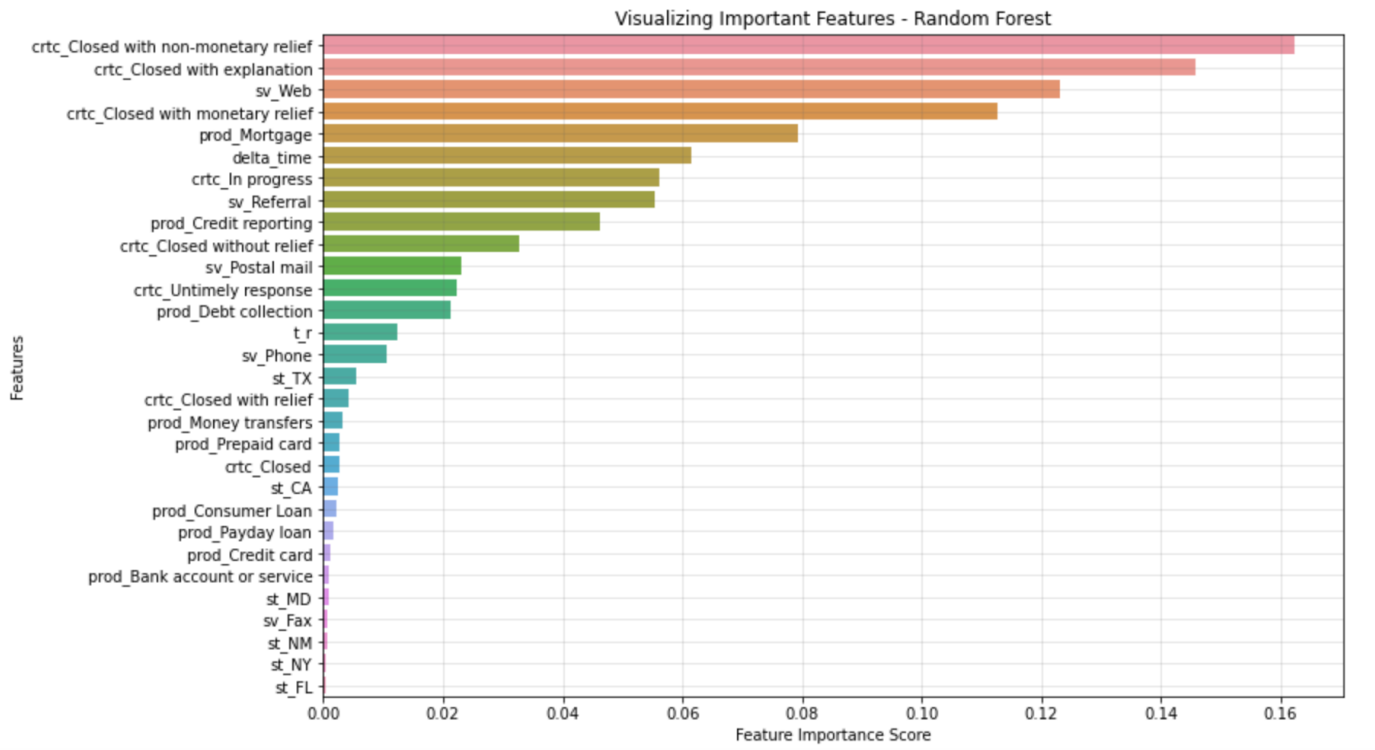
- Supervised learning model that predicts if consumer's claim is going to get disputed through classification ("yes" = 1, "no" = 0)
- Input features include the following: product, state, form of submission, companies' response to customer, and issue
- Featured engineered delta_time as the number of days between when the consumer complained and when it was sent to the company
- Converted categorical data using 1 hot encoding

Machine Learning Model Details

- Tested three different machine learning models:
  - *Random Forest Classifier*
  - *Logistic Regression*
  - *Gradient Boosting Classifier*
- Random Forest and Logistic Regression performed nearly the same (80% accurate)
- I chose random forest as my final model



Machine Learning Model Details



- Visualization determines the main features that decide the outcome (whether the consumer complaints will be disputed or not)
- The knowledge of these features will help the company to reduce issues and sell the right products