A STUDY ON FINANCIAL PRODUCT CONSUMER COMPLAINTS

Shuojia Shi
Springboard Data Science Career Track
Capstone Project 2

Key Takeaway

- Credit reporting and other consumer reports, mortgage, and debt collection are the most complained products.

- The surge in number of complaints on credit reporting in 2017 aligns with the timing when Equifax had its major data breach.

- Counter-intuitively, less concentrated complaint durations tend to result in higher dispute rate.

- Companies could proactively prepare or modify their reply to the complaint to improve consumer experience.

- The top contributions were found using feature importance analysis after classifying consumer dispute.

Outlines

- Motivation. Potential Clients.
- Data Source and Wrangling
- Explorative Data Analysis
  - Financial product analysis
  - Time effect on the complaints
  - State comparison
  - Companies and their responses
- Classification
  - Feature engineering and selection
  - Natural language processing
  - Predict consumer dispute
- Summary

Motivation & Potential Clients

- Consumer satisfaction is a key factor to a successful and profitable financial product. In this study, we take the perspective of consumer complaints to unravel apparent issues and understand trends in customers' needs. It could be a supporting evidence when analyzing customers' satisfaction level toward a product or service. It can also give insight in ways to improve their products and customer satisfaction.
- **Questions** I would like to answer:
    - How does month of the year or day of the week affect the consumer complaints?
    - Do consumers from different states have different complaint behaviors?
    - What are the most complained products?
    - What are the most complained companies?
    - How to reply to consumer complaints to maximize the satisfaction?
    - Can we predict whether a complaint will be disputed?

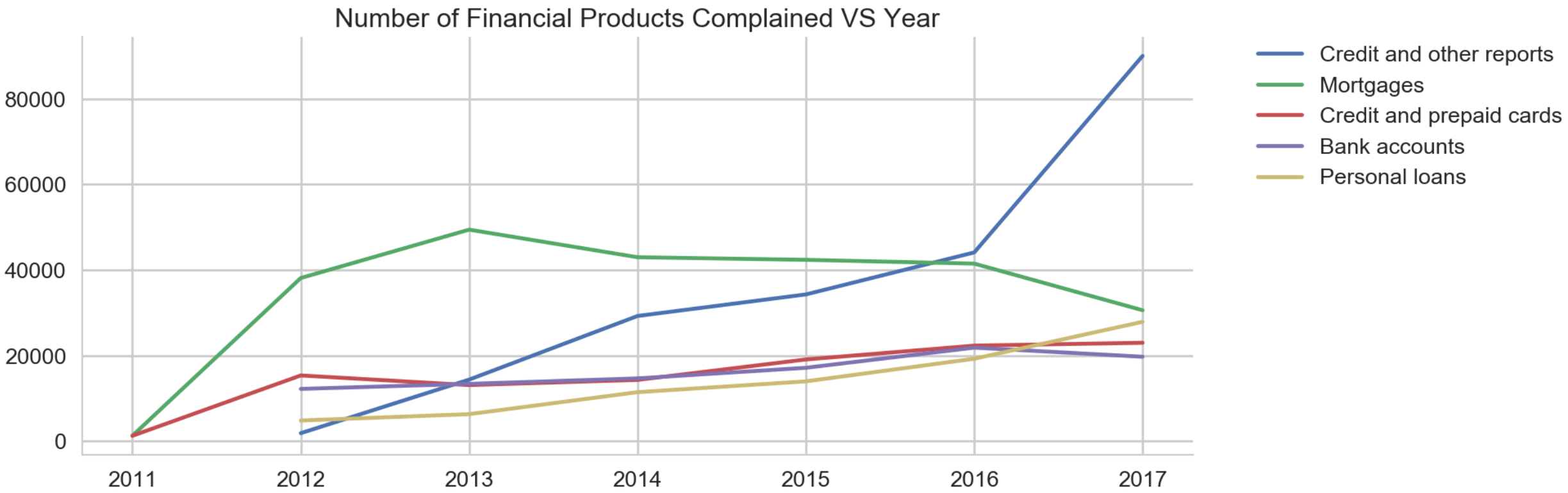
- **Potential clients**:
    - Financial companies
    - Federal and state agencies
    - Customers of financial products
    - Media and Journalists

Data Source and Wrangling

- Survey data from the Consumer Financial Protection Bureau (CFPB) https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data
- Data from 2011 to May 2018 (up to date with this study)
- Its total size is 560.9 MB.
- The table explaining all the features of the data can be found here: https://cfpb.github.io/api/ccdb/fields.html
- Data wrangling:
  - Financial products column changed over the years
  - Synchronize the column by unify the product items

Explorative Data Analysis

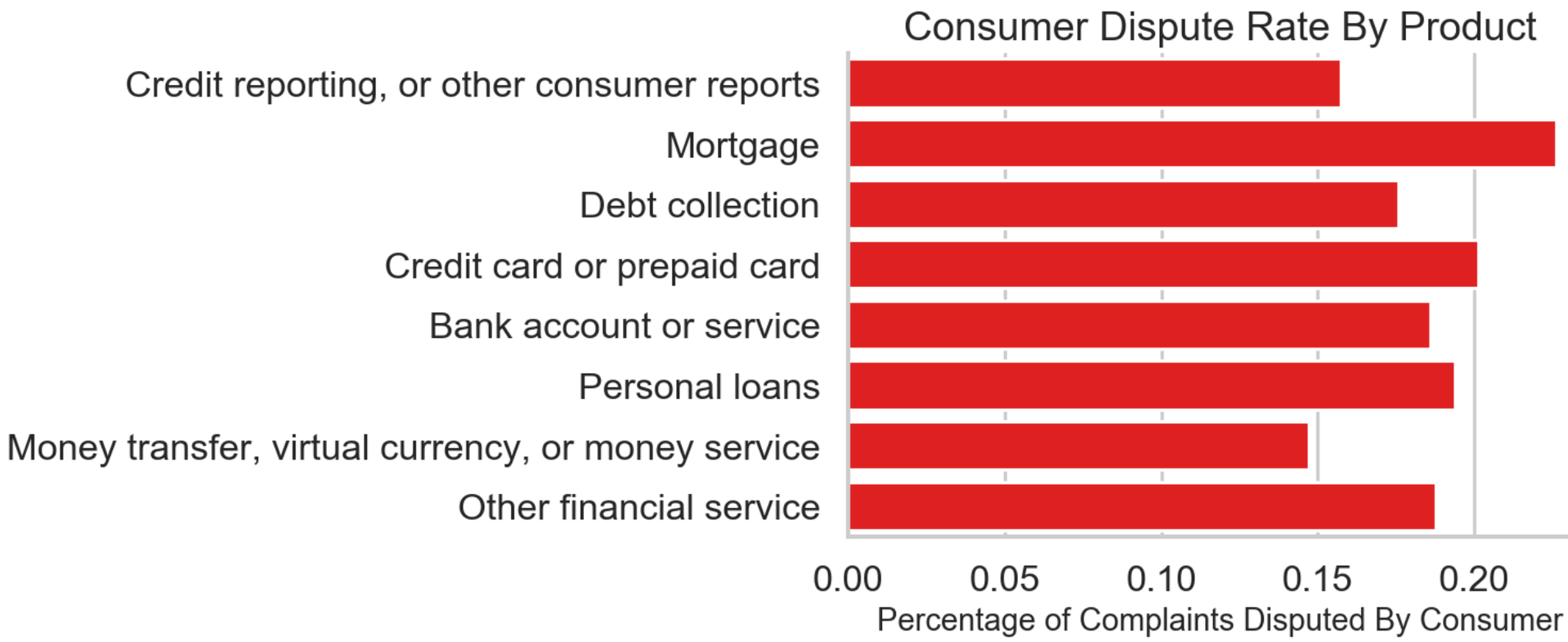
Financial product analysis



- "Credit or other consumer reports" and "mortgage" are the most complained financial products since 2011.
- The five most complained products we investigated, most held steady amount of complains year after year.
- The exception is "**Credit or other consumer reports**". We see a steady increase in complaints and a big surge from **2016 to 2017**!
- The change in complaints could be related to the huge 2017 **Equifax breach.**

Explorative Data Analysis

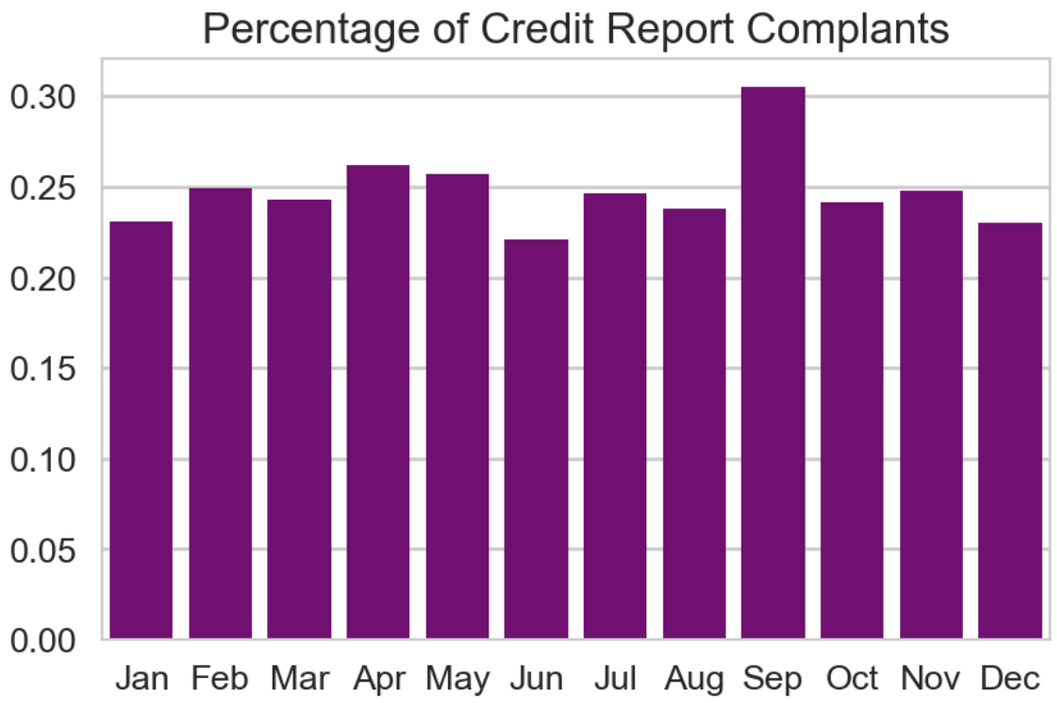
Financial product analysis

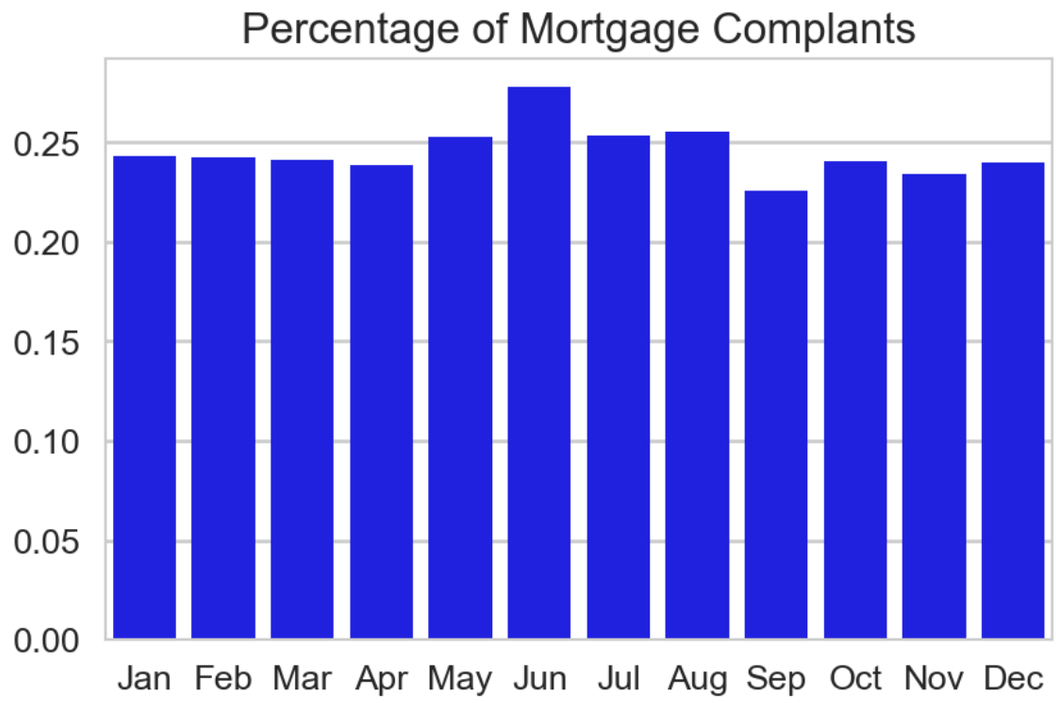


- Mortgage is not only the most complained product but also the most possible to get disputed product.
- Dispute rate ranges from 15% to 23%.

Explorative Data Analysis

Complaints by month of the year

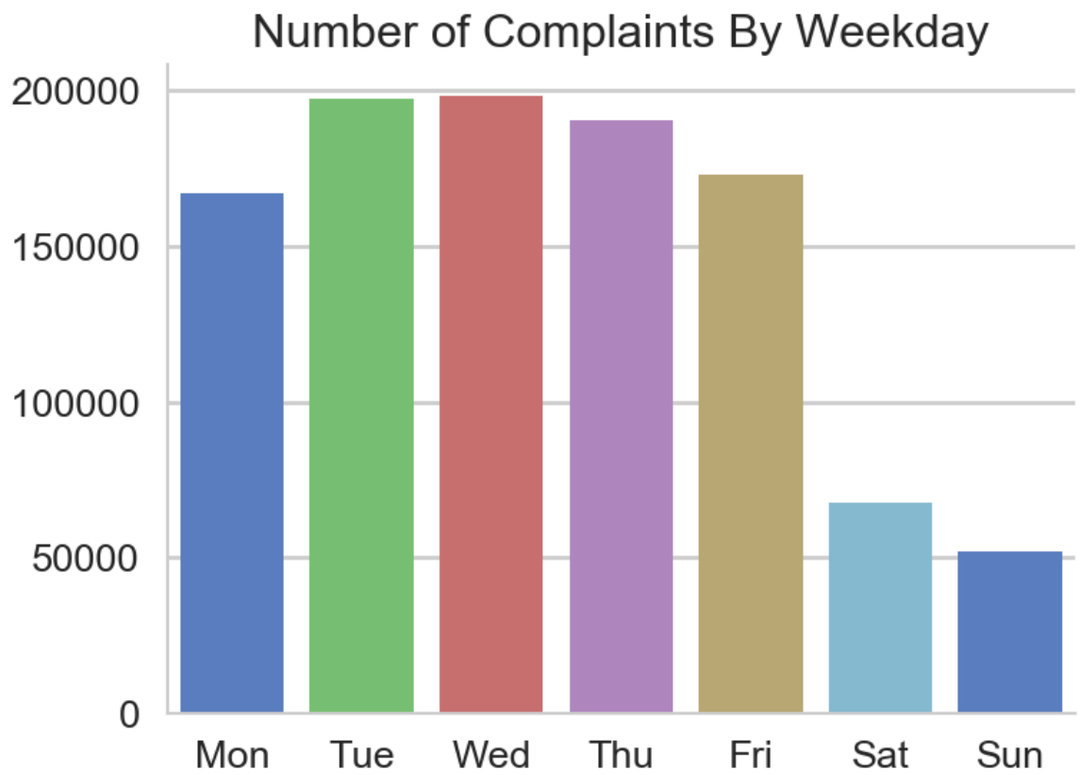


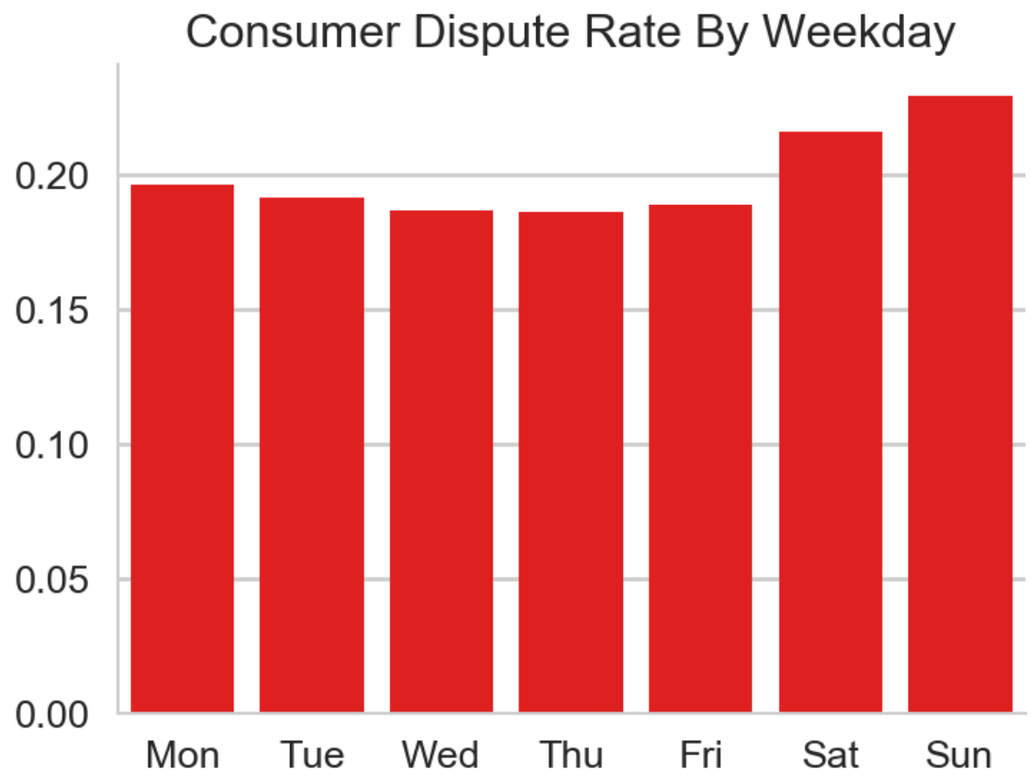


- It is worth noticing that more than 30% of the credit report complaints in September compared with about 23% in other months!
- Mortgage has the highest percentage of being complained in June.
- Both differences are statistically significant!
- This gives insight to preparing the companies to reply the consumer complaints.

Explorative Data Analysis

Complaints by day of the week

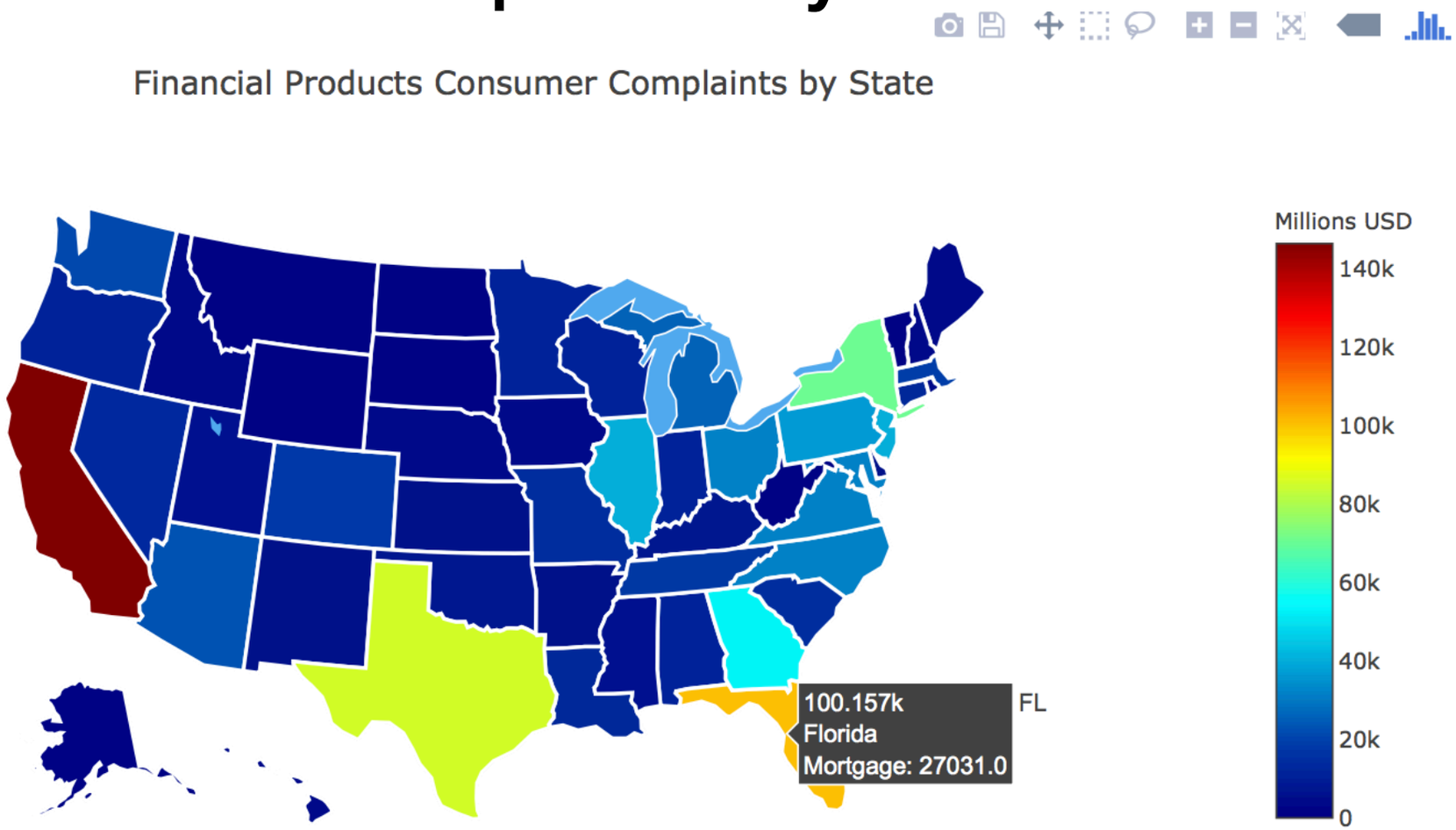




- CFPB receives most complaints on Tuesday and Wednesday.
- Only about a quarter of complaints are received on the weekend.
- There is about 15% higher chance of the complaint being disputed if it was received on the weekend.

Explorative Data Analysis

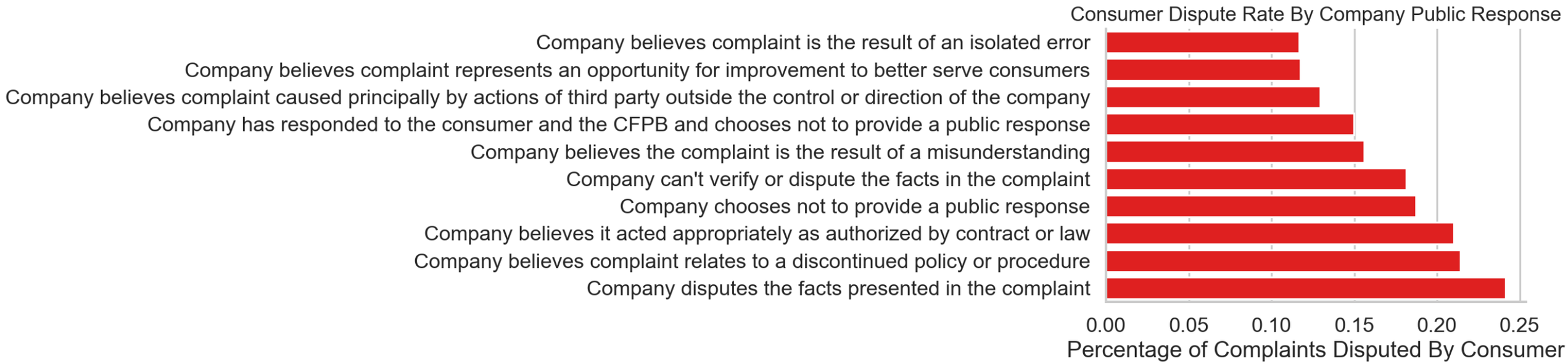
Complaints by states



- Interactive plot generated with plotly module.
- CFPB receives the most complaints from California, Florida, and Texas.
- **Mortgage** tends to be the most complaint product for states with **more expensive housings**.
- After referring to the population of the states, **Florida** stands out. The complaint number per capita in Florida is much higher than the rest of the states. What is causing this? Should financial companies and the state of Florida take a deeper look into what is happening here?

Explorative Data Analysis

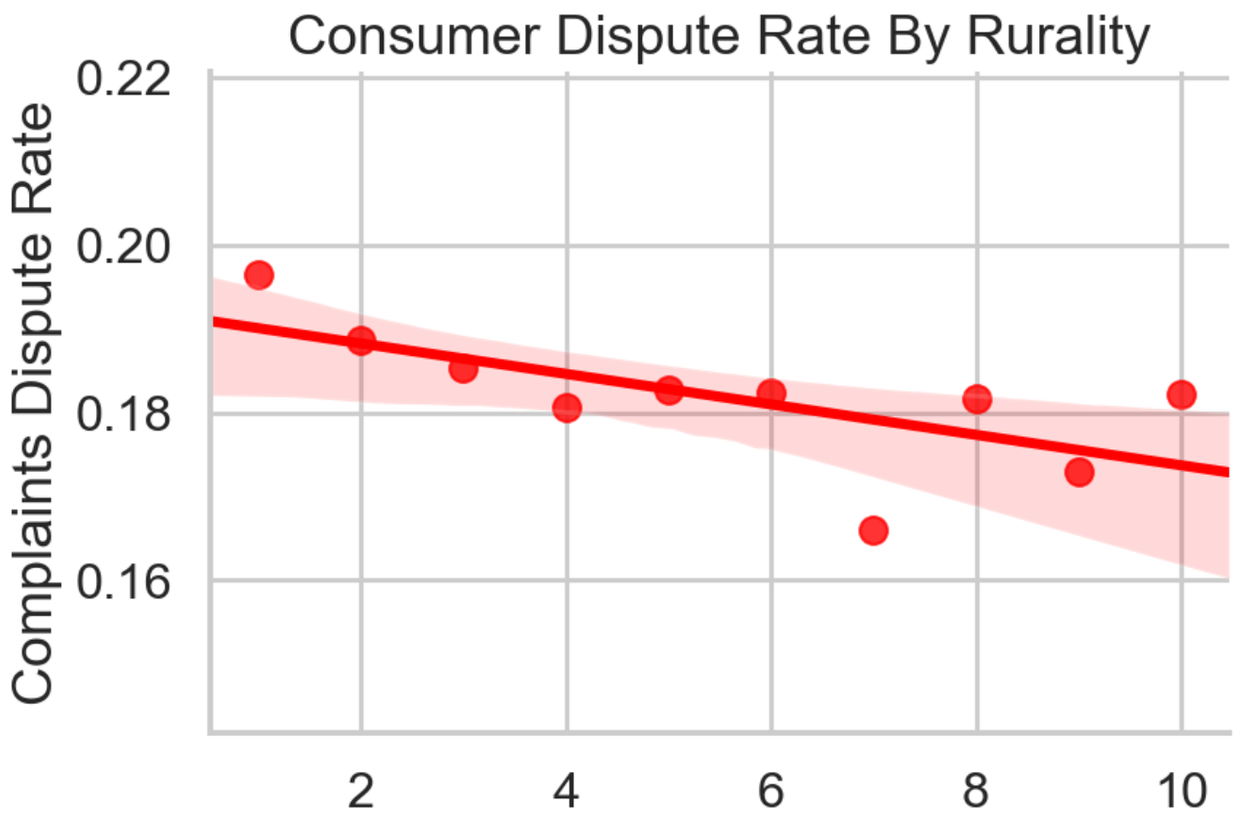
Company response and dispute rate



- It indicates how the public response by company makes a difference in the dispute rate!
- The most disputed response was when company disputes facts presented in the complaints.
- The least dispute public responses are "isolated error" and "represents opportunity for improvement".
- It is worth noting that the responses of "discontinued policy or procedure" or "acts appropriately" do not satisfy the consumers since they are among the highest disputed responses.

Classification
Feature engineering

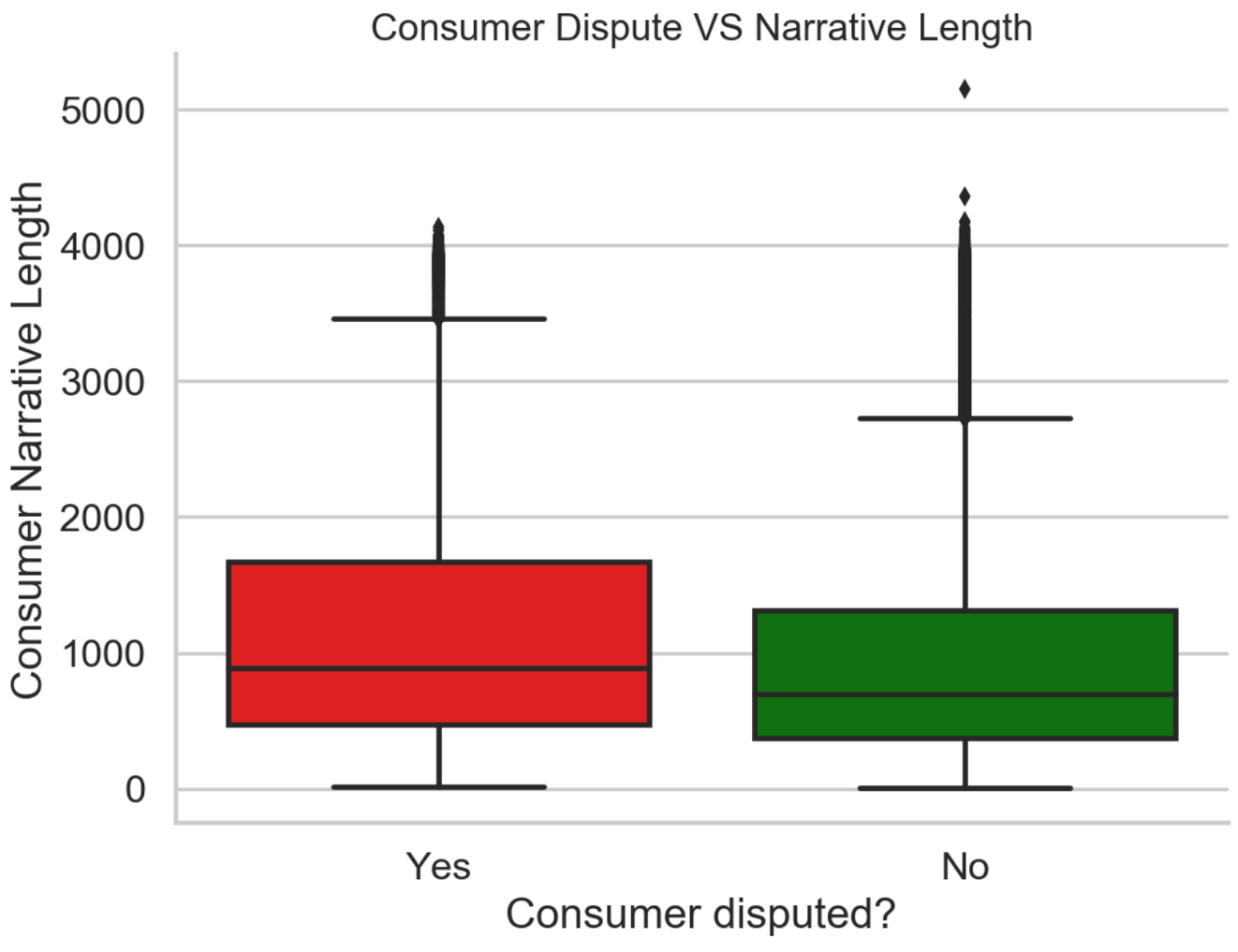
- The zip code feature adds difficulty to use for machine learning.
- Available location to rurality data.
- Zip codes could be converted to rurality of the location as an engineered feature.
- Linear trend between rurality of the location and the dispute rate of the complaint. The more rural the location is (larger x axis), the less chance the complaint will be disputed.
- Null hypothesis test shows very small p-value. The correlation is statistically significant!



Classification

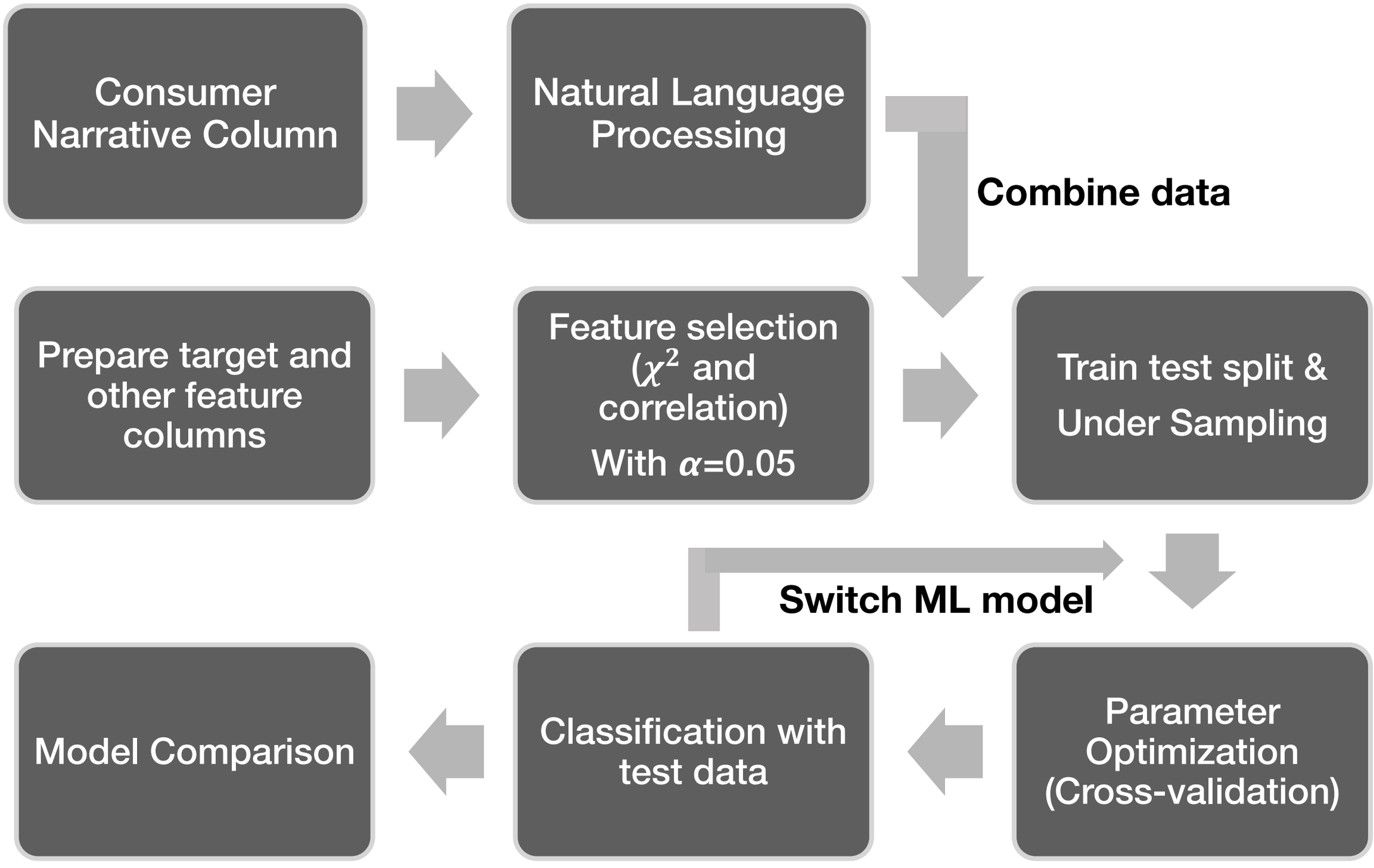
Feature engineering

- Consumer narrative records text data
- The length of the narrative is engineered into another feature for machine learning.
- Box plot shows a difference in the length difference between the two outcomes.
- Consumer narrative column will undergo natural language processing next



Classification

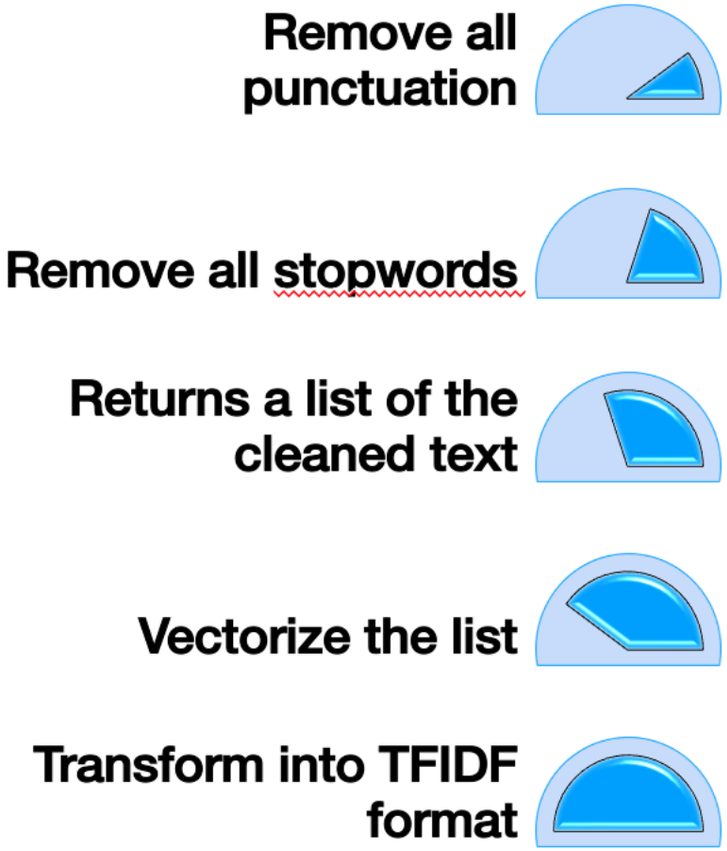
Predict whether complaint will be disputed



Classification

Natural Language Processing (NLP)

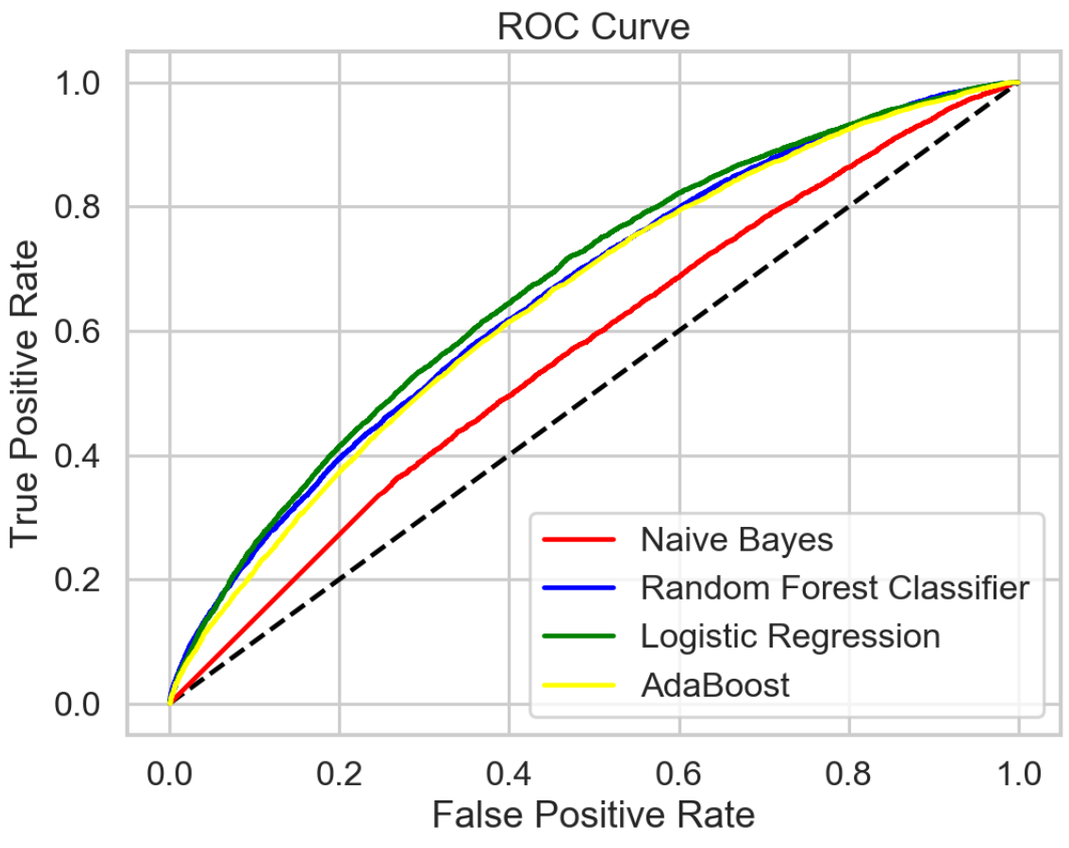
NLP flow chart:



- NLP outputs a sparse matrix
- Use it with Naïve Bayes Classifier to predict consumer dispute
- Accuracy score 0.58 on train and test data
- Next combine the TFIDF matrix with the rest of the features and classify using multiple machine learning models.
- Results shown next page.

Classification

Model	Accuracy score on train	Accuracy score on test
Naïve Bayes	0.55	0.55
Random Forest Classification	0.62	0.61
Logistic Regression	0.62	0.61
Adaboost	0.60	0.60



- **Logistic regression classifier** ranks first among the classification models. It gives 0.62 accuracy score in predicting whether the complaint will be disputed by consumer.
- Naïve Bayes gives the worst performance among models tested. In fact, naïve bayes model performed better when there was only narrative data.
- **ROC curve** indicates similar result.
- From the coefficients of attributes by logistic regression, the top 3 attributes positively contributing being disputed includes **Scottrade Bank, Equifax and VIOLATION**:.
- Logistic regression gives the **recall score** for dispute at 0.65, higher than the average accuracy score. It presents the ability for the model to detect target.
- It seems that there is generally a weak correlation between the target feature and the rest of the data.

Summary

- Credit reporting and other consumer reports and mortgage are the most complained products.
- Notably, there is surge in number of complaints on credit reporting in 2017 when Equifax had its major data breach.
- There are hot/cold months of the year and day of the week for consumer complaints. The dispute rate varies as well. Colder times tend to result in higher dispute rate.
- The analysis also shows the more rural the location is, the less chance the complaint will be disputed. Among the top complained states, Florida stands out as most complaint per capita.
- There is a strong correlation in the way the company replies and whether the consumer will dispute.
- To predict whether a complaint will be disputed, NLP and other feature preparation methods were used. The classification gave the best accuracy score of 0.62 using logistic regression classifier. Top attributes were found.
- This could help companies to proactively improve customer satisfaction by minimizing their consumer dispute.