Consumer Complaint Data Analysis

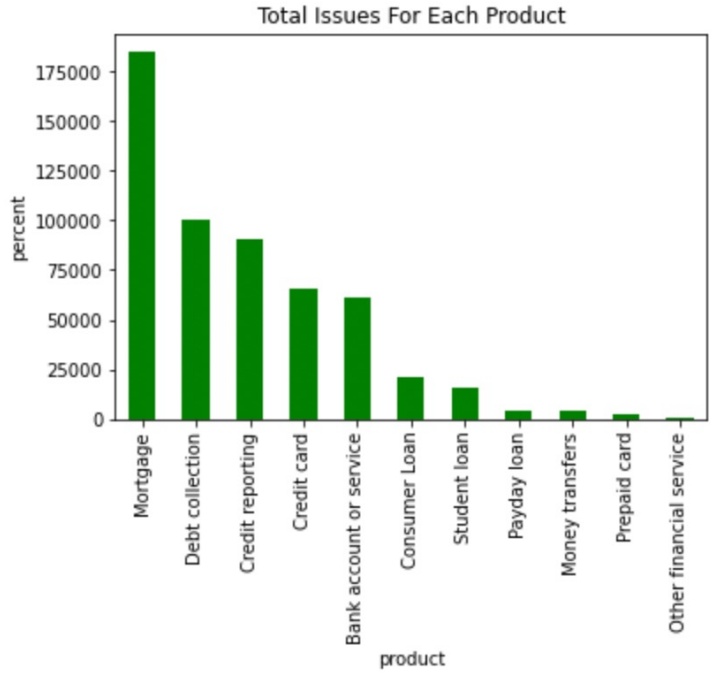
Saichetan Aitha

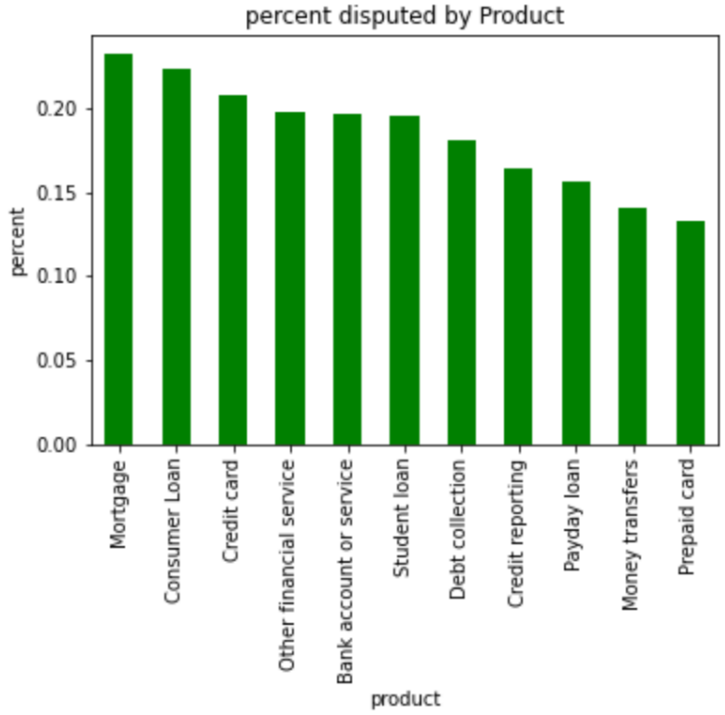
Complaints By State

- Highly Populated states such as California and Florida had the highest number of complaints
- More than 50% of complaints were from California, Florida, Texas, and New York

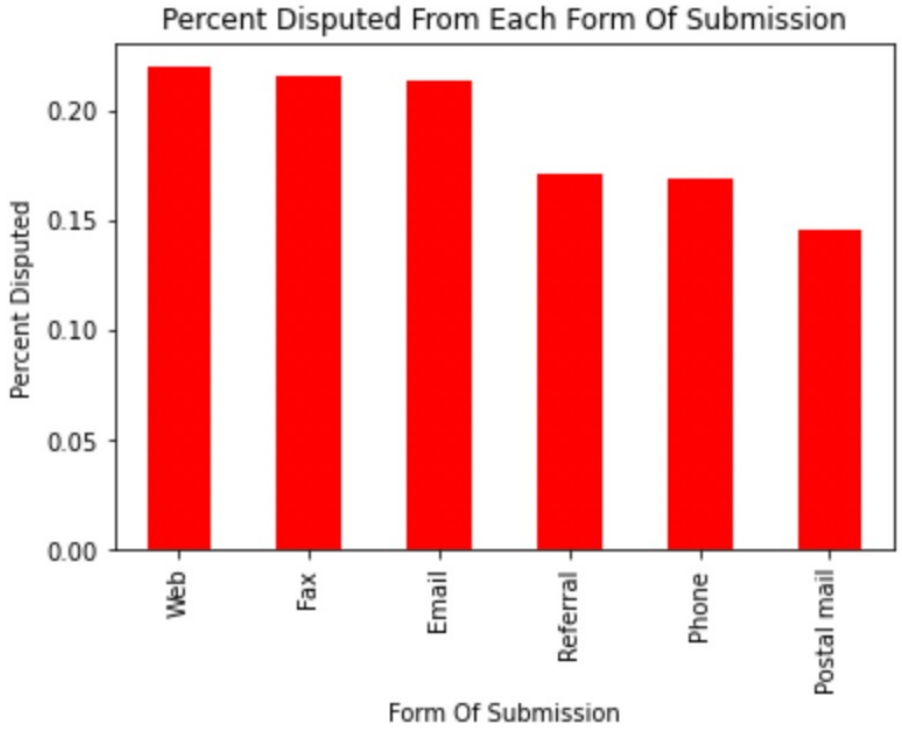
Complaints By Product

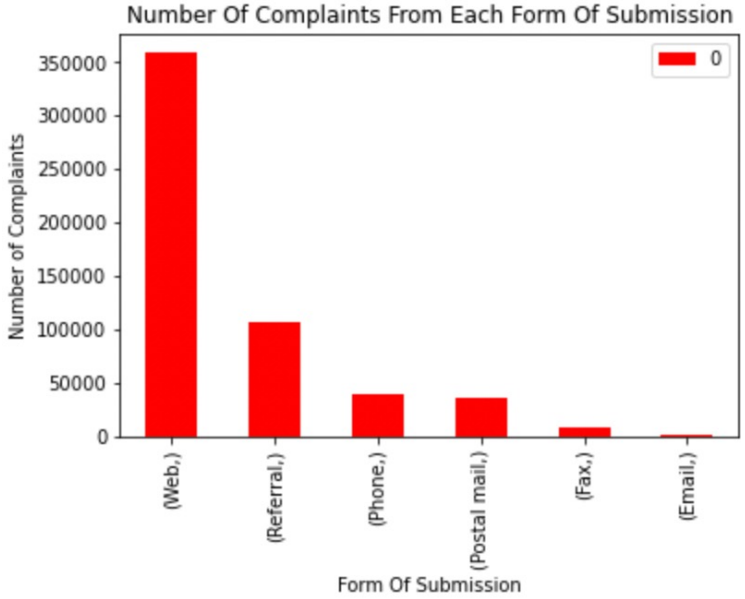
- Mortgages received the most complaints by far, 75% more than the next highest, debt collection
- Mortgages also had the highest dispute rates

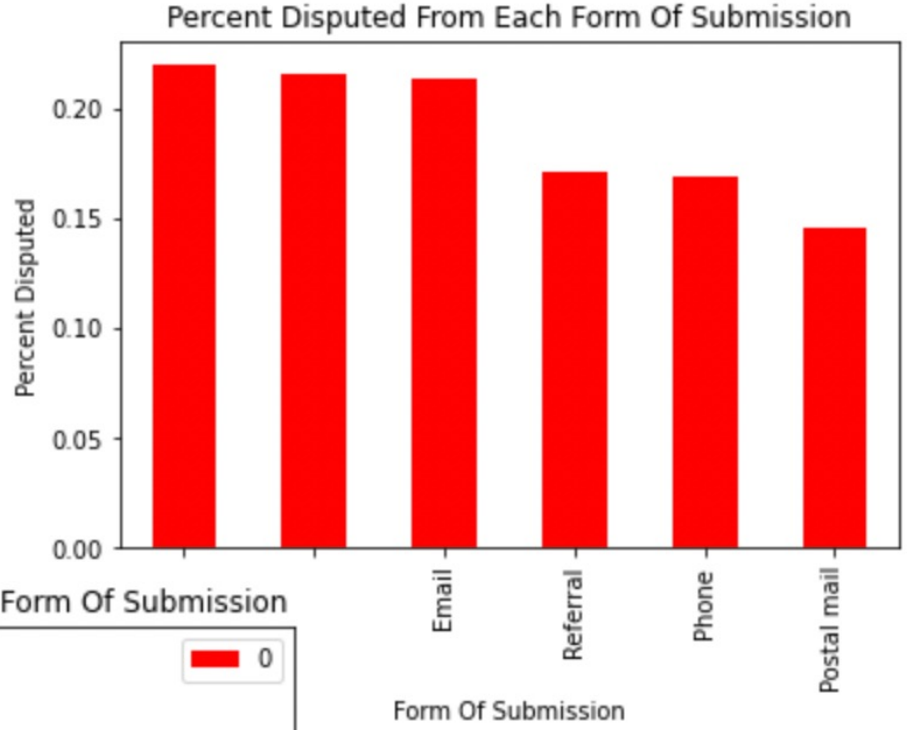


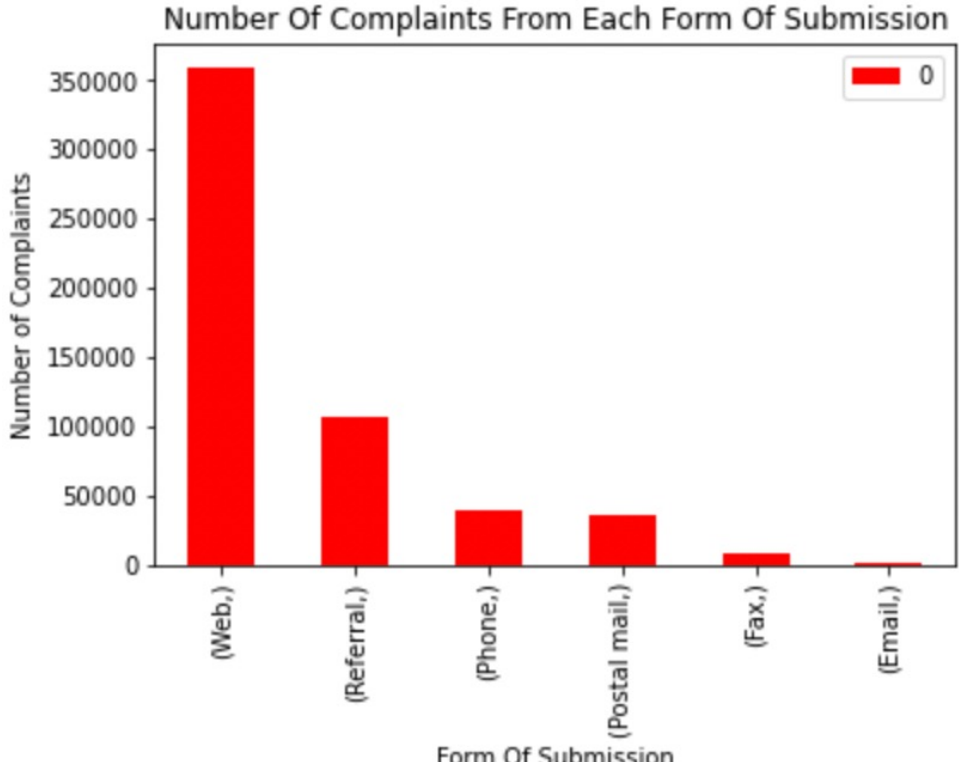


Complaints By Submission



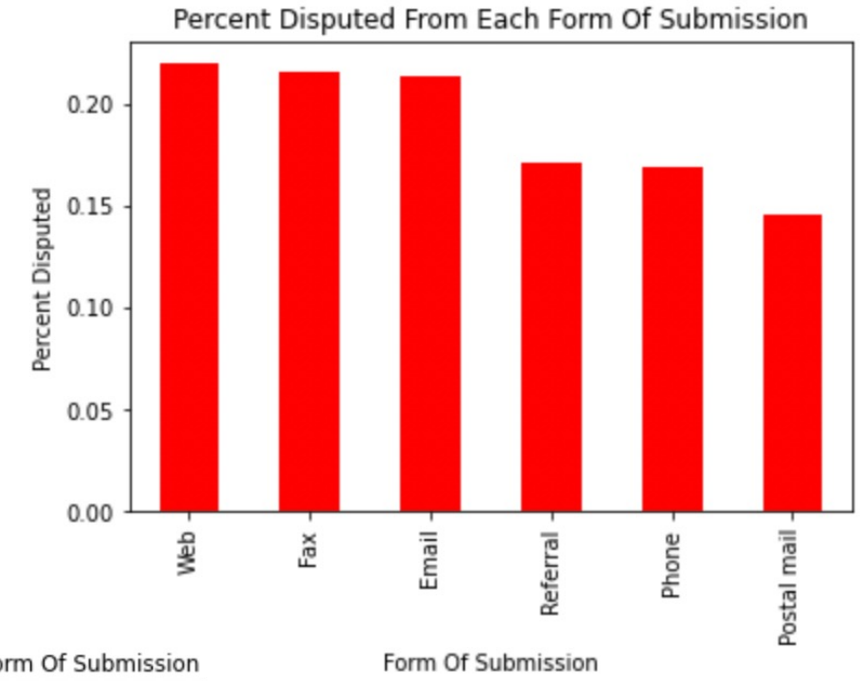


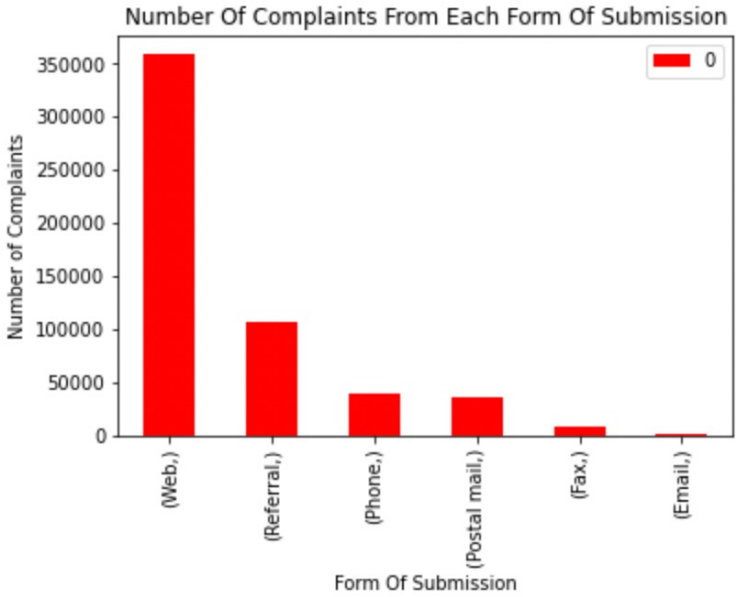

Percent Disputed From Each Form Of Submission
Percent Disputed
0.20
0.15
0.10
0.05
0.00
Email
Referral
Phone
Postal mail
Form Of Submission


Number Of Complaints From Each Form Of Submission
0
Number of Complaints
350000
300000
250000
200000
150000
100000
50000
0
(Web,)
(Referral,)
(Phone,)
(Postal mail,)
(Fax,)
(Email,)
Form Of Submission

Complaints By Submission

- Web submissions received the most complaints by far, tripling the next highest submission, which was referral
- Online forms of submission seemed to have higher dispute rates, possibly due to the lack of direct interaction





Text Classification Model

- Created a text classification model using the consumer complaint narratives
- This NLP predicts which product is being complained about based off of the complaint description.
- This Model created an accuracy of 66%

Classification Report:

	precision	recall	f1-score	support
Bank account or service	0.58	0.47	0.52	166
Consumer Loan	0.35	0.35	0.35	113
Credit card	0.60	0.59	0.60	246
Credit reporting	0.71	0.70	0.70	376
Debt collection	0.68	0.74	0.71	538
Money transfers	0.15	0.14	0.15	14
Mortgage	0.79	0.81	0.80	435
Other financial service	0.00	0.00	0.00	10
Payday loan	0.00	0.00	0.00	21
Prepaid card	0.52	0.58	0.55	26
Student loan	0.55	0.56	0.56	55
accuracy			0.66	2000
macro avg	0.45	0.45	0.45	2000
weighted avg	0.65	0.66	0.66	2000

Multi-Class Model

- This model takes in the feature of state, issue, and submission type to predict the probability of the complaint being disputed.
- This is a quick and dirty model which needs to process more features and data in order to improve accuracy.

Classification Report:

	precision	recall	f1-score	support
0	0.80	1.00	0.89	87750
1	0.38	0.00	0.00	22062
micro avg	0.80	0.80	0.80	109812
macro avg	0.59	0.50	0.45	109812
weighted avg	0.72	0.80	0.71	109812
samples avg	0.80	0.80	0.80	109812

Future Improvements/Hypothesis

- Text-based data such as issue and narratives can be vectorized in order to be incorporated into the multi-class model for better accuracy.
- Variables that can be tested for future hypotheses include analyzing companies and company responses.