Final Report: Predicting Timely Response To Consumer Complaints

Problem Statement

The Consumer Financial Protection Bureau is an independent agency of the United States government responsible for consumer protection in the financial sector. Complaints that the CFPB sends to companies for response are published in the Consumer Complaint Database after the company responds, confirming a commercial relationship with the consumer, or after 15 days, whichever comes first. Can we predict timely response (yes or no) to customer complaints made to the Consumer Financial Protection Bureau for 2022?

The complain filing process has 5 steps:

1. Complaint submitted
2. Route
3. Company Response
4. Complaint published
5. Consumer Review

Step three is where the narrative plays a role in how quickly a company responds. Many of the narratives contain legal language, specific requests, or convey a sense of urgency. NLP should be able to grasp some of this complexity in the narratives and help us evaluate how much it contributes to timely response.

Dataset

The Consumer Complaint Database is a collection of complaints about consumer financial products and services that were sent to companies for response.
https://www.consumerfinance.gov/data-research/consumer-complaints/

Field Name	Description	Type
Date received	The date the CFPB received the complaint	datetime
Product	Type of product in complaint	categorical
Sub-product	Type of sub-product	categorical
Issue	The issue the consumer identified in the complaint	categorical
Sub-issue	The sub-issue the consumer identified in the complaint	categorical
Consumer complaint narrative	Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint.	text
Company public response	The company's optional, public-facing response to a consumer's complaint.	text
Company		categorical
State		categorical
ZIP code		number
Tags		text
Consumer consent provided?	Identifies whether the consumer opted in to publish their complaint narrative.	categorical
Submitted via	How the complaint was submitted to the CFPB	categorical
Date sent to company	The date the CFPB sent the complaint to the company	datetime
Company response to consumer	This is how the company responded.	categorical
Timely response?	Whether the company gave a timely response	binary
Consumer disputed?	Whether the consumer disputed the company's response	binary
Complaint ID	The unique identification number for a complaint	number

Data Wrangling

https://github.com/UnacceptableVegetable/SpringBoard/blob/main/Capstone%20Three/EDA.ipynb

The following features do not seem useful in determining whether there was a timely response.

- 'Date received', 'Submitted via', 'Date sent to company': how and when should be irrelevant
- 'Company public response', 'Company response to consumer': this is optional on the company's end
- 'State', 'ZIP code': where is irrelevant
- 'Tags', 'Consumer consent provided?', 'Consumer disputed?', 'Complaint ID': this customer identification is irrelevant

The remaining features did not have missing except for ‘Sub-product’ (4 rows) and ‘Sub-issue’ (20k+ rows); however, the latter will be determined to not be influential.

Exploratory Data Analysis

https://github.com/UnacceptableVegetable/SpringBoard/blob/main/Capstone%20Three/EDA.ipynb

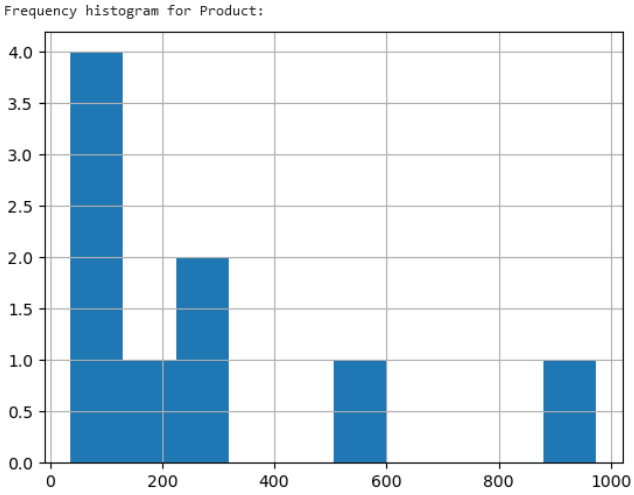
The new dataframe is now summarized:

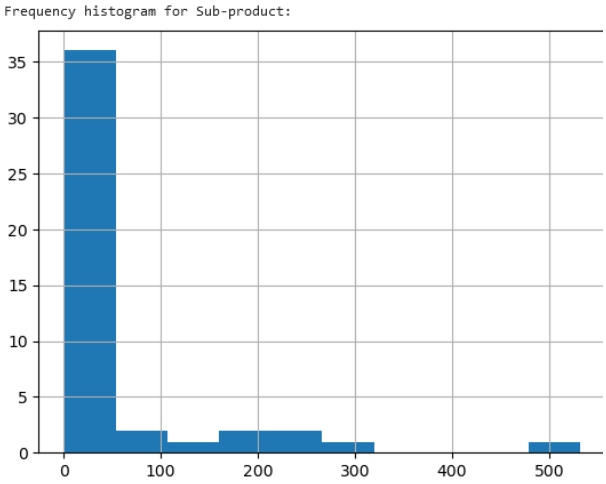
- 328213 rows
- 2732 unique companies
- 6 features, 1 label

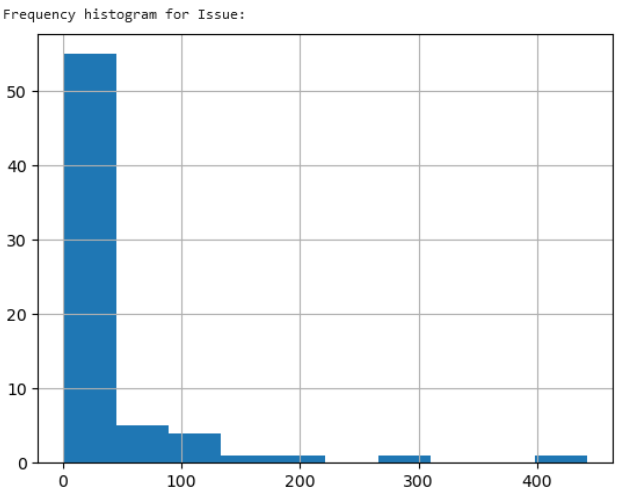
Note that the 'Timely response?' target is imbalanced. The CFPB website even says "98% of complaints sent to companies get timely responses".

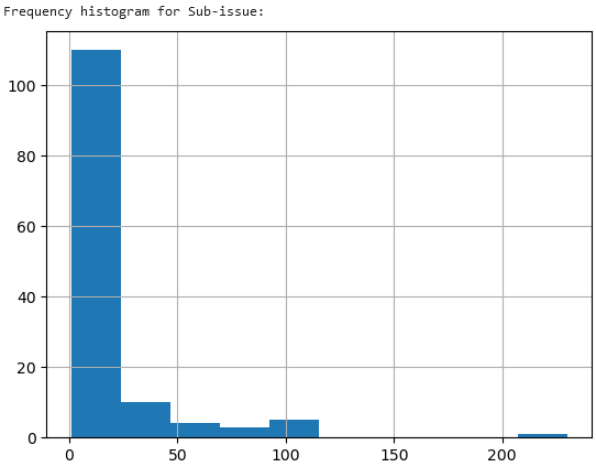
```
Timely response?
Yes    325534
No       2679
```

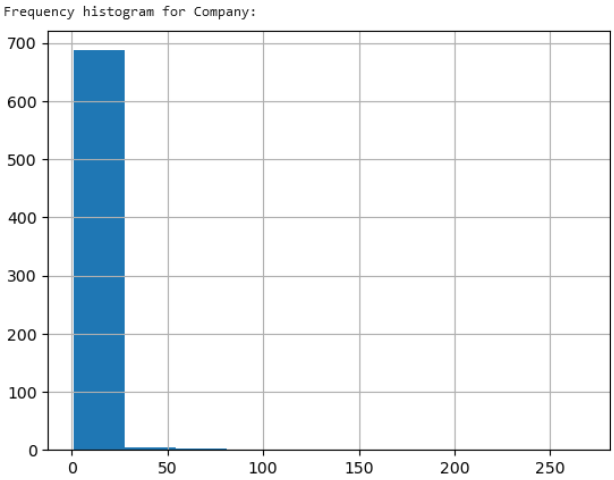
To see what features contributed to negative response, we look at frequency histograms for the value counts of each different column under the filter of “No” responses.











The top ten companies that did not provide a timely response:

```
Company
BANK OF AMERICA, NATIONAL ASSOCIATION                268
CLGF Holdco 1, LLC                                   183
TRUIST FINANCIAL CORPORATION                         109
Self Financial Inc.                                   76
HCFS Healthcare Financial Services of TeamHealth      68
CCS Financial Services, Inc.                          67
Commonwealth Financial Systems, Inc.                  46
AmerAssist A/R Solutions, Inc.                        41
SANTANDER BANK, NATIONAL ASSOCIATION                  39
Sezzle Inc.                                           36
Name: count, dtype: int64
```

Chi-square analysis was done on the categorical features to find the most influential features on timely response. The top 3 categorical features are 'Product', 'Company', and 'Sub-product'.

	Feature	Chi2 Score	Mutual Information
0	Product	438.308877	0.035797
4	Company	63423.890990	0.033308
1	Sub-product	11637.511203	0.026384
2	Issue	116.840417	0.015173
3	Sub-issue	3821.315159	0.013437

Preprocessing

https://github.com/UnacceptableVegetable/SpringBoard/blob/main/Capstone%20Three/modeling.ipynb

The following steps were taken to preprocess the features:

- 'Consumer complaint narrative': Remove redacted text and handle dates. Then apply TF-IDF Vectorization.
- 'Company', 'Product', 'Sub-product': Replace the 4 null rows of 'Sub-product' with placeholder text. Then use ordinal encoding for these categorical variables.
- 'Timely response?': Replace "Yes" with 1 and "No" with 0.

Modeling

https://github.com/UnacceptableVegetable/SpringBoard/blob/main/Capstone%20Three/modeling.ipynb

As a binary classification problem, 3 models were considered: Decision Tree, Random Forest, and Logistic Regression. Initial testing gave AUC scores:

- DecisionTreeClassifier: 0.5478
- RandomForestClassifier: 0.7728
- LogisticRegression: 0.7997

So, for the remainder we only compared Random Forest and Logistic Regression. Because Logistic Regression was the most promising, I decided to do a grid search in order to hyperparameter tune it. I also tried undersampling to deal with the imbalanced data. The following models used balanced weight classes to address imbalanced data.

model (class_weight='balanced')	AUC
Logistic Regression with grid search and under sampling	0.8469
Logistic Regression	0.8524
Random Forest	0.8083

Recall that the Area Under the ROC curve is a value between 0 and 1 that summarizes the ROC curve (plot of the true positive rate against the false positive rate), where 0.5 AUC represents a model that is random.

Logistic Regression seems to be the best performing modeling. We would need to balance having a complex model which has the chance to overfit and using the simpler model which may not capture all important features. Here it seems that the simple Logistic Regression model with default parameters (aside from balanced class weight) is the best choice.

Recommendations

1. Prioritizing Complaint Response: The Consumer Financial Protection Bureau can use this model to automatically prioritize incoming consumer complaints based on the likelihood of a timely response.
2. Identifying Systemic Issues: By analyzing patterns in the complaints and their response times, the model can identify systemic issues within the companies or specific products that frequently result in delays.
3. Improving Consumer Experience: By providing insights into the factors that influence timely responses, organizations can develop strategies to improve overall response times and enhance the consumer experience.

Next Steps

For more advanced language models, the narratives can be broken down into more keyboards that extract phrases that indicate urgency, complexity, emotional tone, and specific requests. This requires more specific words that can be registered for their frequency.

The models can be applied to more years in the range of complaints to the Consumer Financial Protection Bureau. Complaints with narratives only account for 42% of all complaints in 2022, so we may want to consider the remaining data. This would point to particular companies or products that are more likely to have untimely responses, although this may not require machine learning as NLP is not required for 58% of the data.

Outside this dataset, we can look at whether or not the dates could be related to why some companies did not offer a timely response. Perhaps some companies were going through bankruptcy and so did not have consumer complaints as a priority. Additionally, logistics data from companies as well as other historical data can contribute to better predictions.