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?

## 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/>

A screenshot of a computer

Description automatically generated

# 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".

A close up of numbers

Description automatically generated

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.

A graph with blue squares

Description automatically generated A graph with blue squares

Description automatically generated

A graph with blue bars

Description automatically generatedA graph with numbers and a bar

Description automatically generated

A graph with blue squares

Description automatically generated

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

A screenshot of a computer screen

Description automatically generated

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'.

A screenshot of a computer

Description automatically generated

# 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.

A screenshot of a web page

Description automatically generated

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.

# Next Steps

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.