Analysis of Customer Complaints on US Financial Products

Ty Andrews Dhruvi Nishar Luke Yang

Table of contents

1 Summary

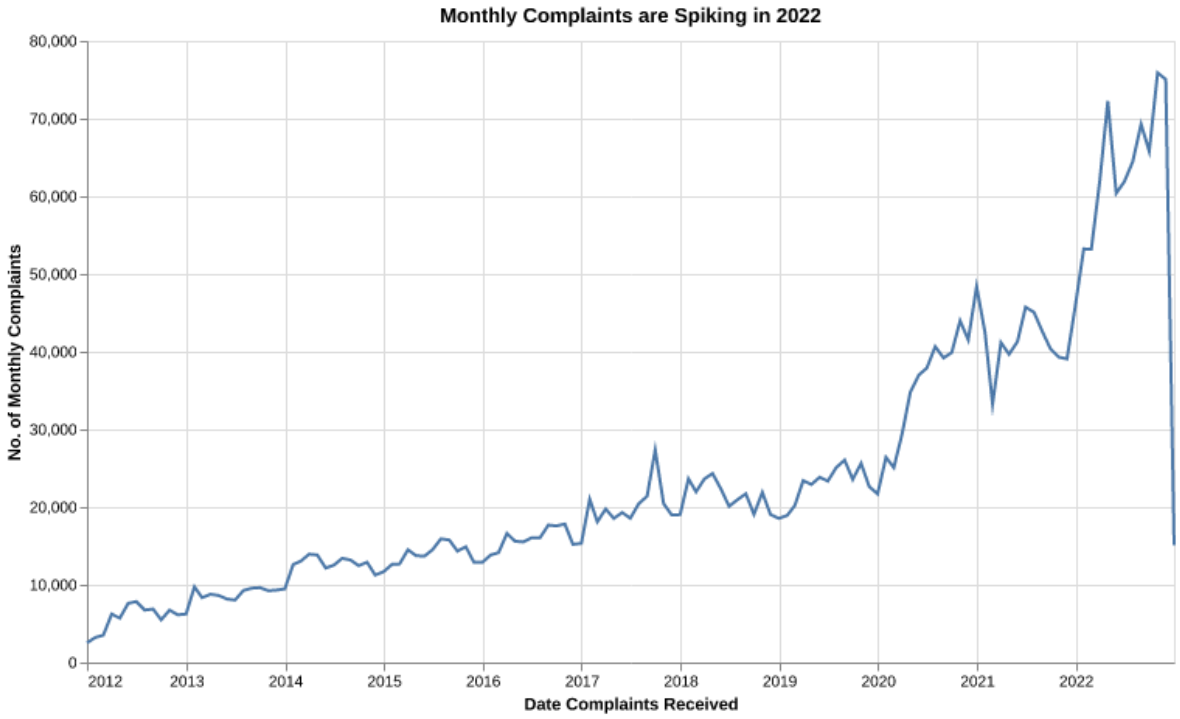
Here we used multiple classification algorithms to predict whether a financial product consumer will dispute a complaint made to the Consumer Financial Protection Bureaus' (CFPB) Consumer Complaints Database(*Consumer Complaints Database* 2022).

2 Introduction

As of 2022, the CFPB receives over 60,000 consumer complaints a month related to companies financial products. Between December 2011 and November 2022 over 140,000 complaints were disputed by consumers costing both the companies and CFPB time and money.

Complaints can be responded to by the company in multiple ways but each consumer has the opportunity to dispute the provided response. These disputes are likely costly to both the CFPB and the companies for which they are raised so being able to anticipate whether a complaint will be disputed has the potential to save both time and money for companies and the CFPB alike. Especially, as we observed that the number of complaints spikes near 2022 (Figure 1), it is essential to understand the tactics of handling these complaints.

Figure 1: Figure 1. Complaints Over Timeline



3 Methods

3.1 Data

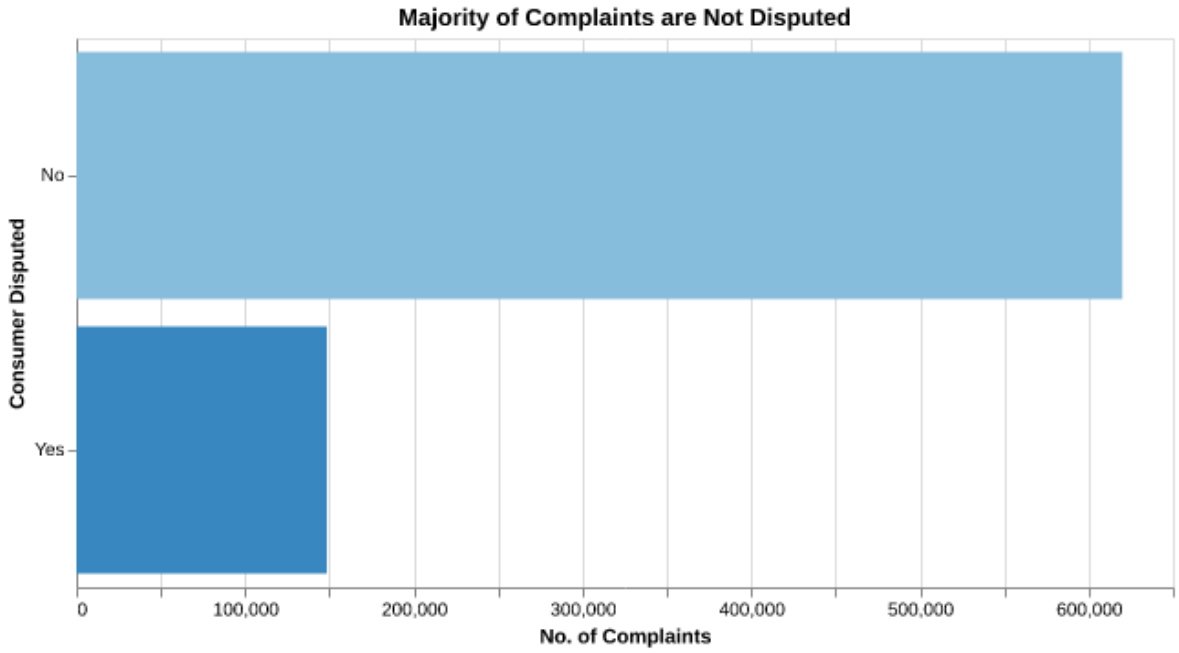
We used pandas(McKinney 2010), scikit-learn(Pedregosa et al. 2011), and TidyVerse (Wickham et al. 2019) to collect, analyze, and summarize our data. The CFPB database contains ~3 million complaints starting from December 2011 all the way up to November 2022 when this report was written. There are 18 columns included in the raw data of which the following 12 were focussed on to attempt to answer the study question. The remaining columns either had too many missing values or were id columns thus not likely to be useful.

Table 1: Table 1: Primary dataset features used for analysis.

Features	Description
consumer_complaint_narrative	Text written by the cutsomer pertaining to the complaint
product	The high level financial product to which the complaint relates
sub_product	A specific product within the high level product category
issue	The type of complaint being raised
sub_issue	The low level type of issue submitted
consumer_consent_provided	Binary of whether the consumer provided consent to be contacted by the company
timely_response	Whether the company responded to the complaint within 15 days
submitted_via	The channel by which the complaint was lodged
company_public_response	How the company publically responded to the complaint
company	The company's name against which the complaint was lodged
company_response_to_consumer	How the company responded to the customer
consumer_disputed	Did the consumer dispute the companies response/proposed resolution

Approximately 4.8% of all complaints are disputed, with the ~75.1% of complaints having an unknown dispute status. Figure 2 below shows the balance of dispute status.

Figure 2: Figure 2. Class Imbalance in the Target Column



Each field has varying amounts of missing values as can be seen in Table 1 below. Fields such as tag where there are numerous entries missing values were removed from the analysis. Individual complaints with missing information were removed from the data-set for analysis since the data set is large enough to still have a significant number of training examples for the analysis (~20,000).

Table 2: Table 2. Unique and missing value counts by data feature.

Fields	Valid Count	Unique Count
product	3142434	18
sub_product	2907141	76
issue	3142434	165
sub_issue	2457220	221
consumer_complaint_narrative	1128809	984613
company_public_response	1369029	11
company	3142434	6583
state	3102263	63
zip_code	3101735	34484
tags	354684	3
consumer_consent_provided	2313331	4
submitted_via	3142434	7
date_sent_to_company	3142434	3974
company_response_to_consumer	3142430	8
timely_response	3142434	2
consumer_disputed	768440	2
complaint_id	3142434	3142434

3.2 Data Pre-Processing for Analysis

As the dataset is quite large, >3 million complaints, it was decided to start off by dropping any complaints which were missing one of the above features. This results in a training data set of ~200,000 complaints.

Feature pre-processing approach for the modelling and rationale is as follows:

Table 3: Table 3: These features were passed into a column transformer, which was then integrated with five different estimators for prediction.

Features	Preprocessing Step	Rationale
consumer_complaint_narrative	CountVectorizer max_features = 1000	High Amount of Unique Textual Data
product	OneHotEncoder	Each field is categorical
sub_product	drop = "if_binary"	
issue		
sub_issue		
company_public_response		
company		
company_response_to_consumer		
state		
consumer_consent_provided		
submitted via		
timely_response		
consumer_consent_provided	dropped	Only one value in this column

3.3 Analysis

A predictive approach using multiple classification models was used to attempt to predict whether a consumer would dispute a complaint or not. A `DummyClassifier` was used with the a most frequent approach which predicts the most frequent class (not disputed).

The following scikit-learn models were trained and cross-validated on the training data set:

- Logistic Regression
- Support Vector Machine
- Naive Bayes Classifier
- Random Forest Classifier

Cross validation scores on both the train and validation set were recorded.

4 Results & Discussion

We applied the `DummyClassifier`, `LogisticRegression`, `Naive Bayes`, `SVC`, and `RandomForestClassifier` to predict the targetwhich is whether the customer disputed the complaint or not. The models were applied using default parameters and a five-fold cross-validation were applied using the training split. We examined and recorded the accuracy,

precision, recall, and f1 scores to be the metrics evaluating the models. The results of the cross validation were as follows:

Table 4: Table 4. Model Performance and Score.

Metric	Dummy	Logistic Regression	Naive Bayes	SVC	Random Forest
fit_time	2.477	8.422	1.879	98.956	29.277
score_time	0.484	1.028	0.503	18.946	3.297
test_accuracy	0.778	0.648	0.704	0.681	0.783
test_recall	0.000	0.448	0.347	0.437	0.056
test_precision	0.000	0.302	0.339	0.334	0.625
test_f1	0.000	0.361	0.342	0.378	0.102

Figure 3 below illustrates a visual representation of the performance of the models. We observe a high accuracy of the `DummyClassifier`. Given the imbalance of the class, the accuracy would not be an important metric in this problem. Instead, from the company's perspective, we focus more on improving the precision, the recall, and the f1 score. Noticeably the company likely wants to spot the people who are going to dispute, thus, the recall score here is more important compared to precision.

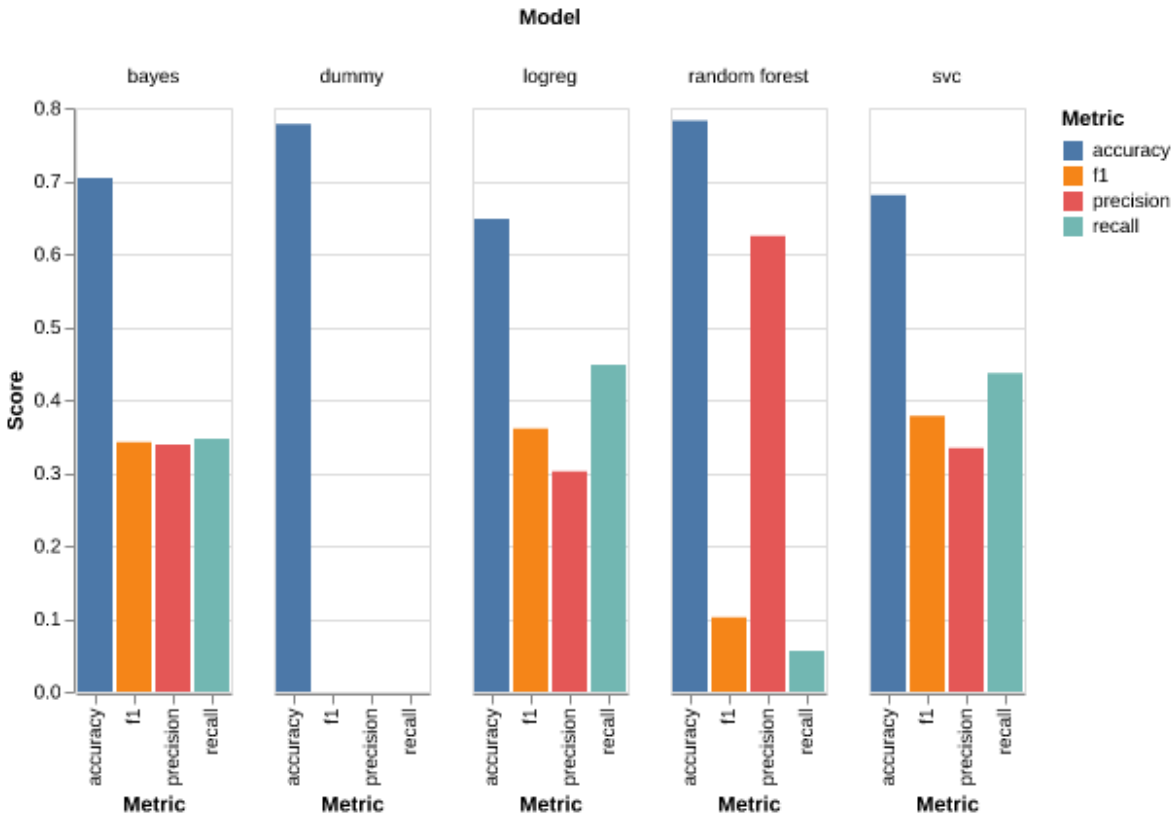
The results above motivates us to choose `LogisticRegression` as the final estimator. It has one of the overall highest f1 scores at 0.361 and among the recall is among highest scores. Though it sacrifices some accuracy, the precision, recall, and f1 scores are significantly improved uon the dummy classifier and competitive with the other models evaluated. We also see that `SVC` gives slightly higher scores than `LogisticRegression`; however, in Table 2 we see that it takes substantially more time to train and cross validate. in addition, due to the complexity of the model we also lose a degree of model incontestability. Overall, we choose `LogisticRegression` over `SVC` due to its scalability and interpretability. The following table shows the test scores of the models:

Table 5: Table 5. Model Performance and Score on Test Set.

Metric	Dummy	Logistic Regression	Naive Bayes	SVC	Random Forest
Accuracy	0.782	0.658	0.710	0.679	0.787
Recall	0.000	0.313	0.341	0.327	0.662
Precision	0.000	0.475	0.351	0.444	0.055
F1	0.000	0.377	0.346	0.376	0.102

We see that on the test set `SVR` and `LogisticRegression` performs similarly as expected based on the validation scores.

Figure 3: Figure 3: Performance of Different Models on Different Metrics



Unfortunately, an f1 score of 0.377 for `LogisticRegression` is quite low and unlikely to be particularly useful in the broader business sense. This analysis can be used as the basis to understand what target f1, precision or recall target scores should be set for further analysis.

5 Conclusion

The analysis in this reports focuses on using a machine learning approach to predict whether the consumer is going to dispute after the company's response. We processed the features such as the product, consumer's complaints, and company's responses. We trained five different models that optimized the f1 score and chose `LogisticRegression` as a suitable estimator for the first pass at attempting to predict whether consumers dispute their complaints. Next steps would be evaluating the impact of the chosen models performance and then deciding whether to try and refine further to improve the model.

References

Consumer Complaints Database. 2022. Consumer Financial Protection Bureau. https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/ 10.25080/Majora-92bf1922-00a .

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.