Final Project

By

Videesh Reddy Garlapati, Neehar Bujimalla

George Mason University

Title: Analysis of Consumer Complaint Dataset

Title of the Website or Project

Brief Introduction (about the overall project)

Briefly explain what problem you’re trying to solve through data analysis.

The main idea behind our project is to find and extract patterns or confidential information from the CFPB dataset A database of consumer complaints isn't worth important if the dataset cannot be used for the betterment of the experience of the consumer. The size of this data is enormous. It has numerous missing cells, which will decrease the accuracy score during training and testing by simplifying the data further so financial companies can use it to analyze the complaints to make a difference for their customers better. The consumer complaint narrative field within our dataset contains consumers’ complaints in detail.

• Who can benefit from your data analysis (i.e., who are stakeholders)? Specify detailed justifications on why your analysis will be beneficial to particular groups of people, researchers, or organizations.s

This data of consumer complaints is precious as we can extract intriguing information from it and conceivably give contexts to these complaints so companies can address issues in a systematized matter with the expedients to retain or gain further customers. The overall analysis can also be used by the state and federal agencies to show repetitive complaints that have not been improved over time so rules or fines can be implemented to punish companies that continue to conduct poor business practices.

This analysis could possibly be beneficial to three parties. First, Customers can better determine if a company is worth giving their business to Customer decision making could possibly improve with this dataset as consumers could easily track which company has been able to make constant improvements over time with complaints they have received. This could show initiative and therefore be a viable option for the customer to give their business over to their competitors. A reduction in the number of complaints means that a company has improved its ways of working to ensure better customer satisfaction which led to fewer complaints against them.

Second, CFPB can use this analysis to better understand the problems and possibly put an end to major issues through regulations by bringing them to the government’s attention.

Most importantly, this analysis will be vital to financial companies. It will help them to work towards decreasing the number of complaints against them in a manageable and clear way. A better way to categorize these complaints and finding relationships can help companies find a trend through the vast amount of information in these narrative fields for each complaint in easy-to-follow topics once text mining methods are implemented.

2) Nature of the Data Curation

• Who (company, agency, organization) collected the data?

The Customer Financial Protection Bureau keeps track of all consumer complaints in a database. The Consumer Financial Protection Bureau (CFPB) was established on July 21, 2011, with the goal of fighting for customers' rights and overseeing customer satisfaction and protection in the financial industry.

The Consumer Financial Protection Bureau (CFPB) provides a list of consumer complaints from its database to organizations for response. Following the company's response, the CFPB formally publishes the complaints data for the general public to view.

• Who they are, what do they do?

CFPB, will publish the data from various industries by collecting and publishing to the public and also perform data analysis. It will collect the data and send it to organizations to cross-check whether the data can be published publicly or not. Also, before publishing the data the CFPB will remove certain features (columns) which should be published from the dataset.

• Why did they collect the data (purpose)?

The dataset parameters are used to obtain some useful insights that can help us figure out what the organization needs to improve and assist customers in making informed selections.

Because our dataset is vast, data pre-processing and cleaning will be required regardless of the type of analysis we undertake.

• What is the nature of the data given the purpose of the data collection (e.g., any bias)?

The Consumer Financial Protection Bureau (CFPB) maintains a consumer complaint database, which contains complaints about consumer financial products and services provided by companies across the United States. These complaints came from their customers. It includes information such as the type of complaint, the product for which the complaint was issued, the date of the complaint, the issue, the sub-issue, and the action is taken, among other things.

• Usually, many datasets are NOT collected for data scientists themselves, but as a byproduct of the organizational process. Because of this reason, it is important to understand the nature of the data.

The data that is collected by CFPB is used by financial companies, government to analyze the consumer interest in financial products and to resolve the complaints by offering good services that can make the consumer happy. The data is also in text format and has a limited number of features which is easy to understand by the consumer as well in general.

• Given the nature of the data, how can you adjust and leverage the data (i.e., what are pros and cons of the data and how can you overcome it)?

By pre-processing, the data has one main goal: Standardization.

Standardization of dataset fields was something simple, but very useful. It’s easier to use column names that are one-word than multiple. Our main focus was the consumer complaint narratives field and any null value within that has no use, so data is cleaned by removing all rows with a null value in the consumer complaint narrative field. The field cannot be null as we are using this field to calculate topic trends and text mining of that field.

We split the text from the consumer complaint narrative field into sentences, sentences into words, made the words lowercase, and removed the punctuation.

Words in the field with less than 3 characters were discarded, all stopwords (the, he, she, have, get, etc.) were also removed.

We lemmatized the word. This means that words in third person were changed to first person and verbs in past and future tenses were changed to present.

Words were stemmed and were changed to their root form. Words like “fraud”, “fraudster” were minimized to just “fraud”.

• Is there any privacy, quality, or other issues with this data?

4) Requirements and Resources needed

• What software and hardware resources you have used in this project?

Excel and Jupiter Notebook are the software - based resources used in this project to help you understand and visualize data. The laptop is a hardware resource, and for visualization approaches, a Lenovo laptop with an Intel i7 processor and 12 GB of RAM, as well as a Google colab GPU, TPU, was employed.

• What kinds of pre-processes were needed to make use of the data, and why?

The dataset contains numerous null values, so EDA-Exploratory data analysis (text data pre-processing) is necessary to eliminate them. By omitting this, the data shape will become smaller and easier to process.

• What are the advantages and limitations of the target dataset in answering your questions?

Adv - Stakeholders will be able to see how they may improve by correcting inaccuracies using our answers to queries for the target dataset. Banking firms will find it simple to work in departments with a significant volume of complaints.

Lim – Our target dataset restricts us to the data's properties without providing context for a complaint. This restricts those attempting to work with the dataset by requiring them to manipulate data only on the specified grounds.

The intended dataset is massive, and normal computer hardware resources will be insufficient to perform tasks like TF-IDF, SVM, Logistic regression fitting, and Deep learning. With this type of data set, employing GPU or TPU resources makes training, testing, and validating the data simple. however, finding the correctness of data on a regular computer is tough.

3) Questions:

What are the main questions of your interests that can be answered through the data that you chose? List some specific questions, and be sure to answer them in your analysis.

What type of data analysis should be done on large text data?

To deal with data of this size, most people will use graphs. Because the dataset is so huge, it's difficult to go through it and extract insights. So we utilized Python to extract some information from the data. Install and import all essential Python packages and dataset as the initial step. After that, the data should be cleaned for the exploratory data idea, and the data is cleansed by focusing on the customer complaint narrative, and we employed several approaches to acquiring various sorts of insight and value.

To obtain a sense of the most common words in the consumer complaint narrative column, we developed a word cloud. This would give us a sense of the most used words in the customer complaints narrative column. It is possible to draw a much clearer picture of the challenges that the customers face.

Which state did consumers report more complaints?

Which Product has more complaints state-wise and overall?

What are the most common words used in complaints in the overall dataset?

Provide justifications on why your question(s) are important for stakeholders.

How their services can be improved based on the analysis.

5)

Briefly describe the dataset. Prepare and describe relevant metadata (types of attributes/variables in the dataset)

The Consumer Financial Protection Bureau (CFPB) provides a list of consumer complaints from its database to corporations for action. Following the company's response, the CFPB formally discloses the complaints for the public to see.

The data was accessible via JSON, CSV, and API. The CSV file was downloaded for this project. We planned to use these dataset attributes to extract some useful insights that would help us figure out what the company needs to improve and assist customers in making informed decisions. This was a huge dataset. At the time of the final download, the dataset had a total of 2653059 complaints (March 2022). We have a date, string, and number formats in our columns. Our CSV file was about 1.60GB in size. Because our dataset is vast, data pre-processing and cleansing will be required regardless of the type of analysis we undertake.

Product object

Sub-product object

Issue object

Sub-issue object

Consumer complaint narrative object

Company public response object

Company object

State object

ZIP code object

Tags object

Consumer consent provided? object

Submitted via object

Date sent to company object

Company response to consumer object

Timely response? object

Consumer disputed? object

Complaint ID int64

dtype: object

Statistical Summary findings of the dataset.

complaints.describe(include**=**'all')

Out[29]:

|  | **Product** | **Sub-product** | **Issue** | **Sub-issue** | **Consumer complaint narrative** | **Company public response** | **Company** | **State** | **ZIP code** | **Tags** | **Consumer consent provided?** | **Submitted via** | **Date sent to company** |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 2653059 | 2417896 | 2653059 | 1994535 | 926071 | 1101765 | 2653059 | 2613424 | 2613170.0 | 311092 | 1872149 | 2653059 | 2653059 |  |
| **unique** | 18 | 76 | 165 | 221 | 822331 | 11 | 6368 | 63 | 59023.0 | 3 | 4 | 7 | 3757 |  |
| **top** | Credit reporting, credit repair services, or o... | Credit reporting | Incorrect information on your report | Information belongs to someone else | My credit reports are inaccurate. These inaccu... | Company has responded to the consumer and the ... | EQUIFAX, INC. | CA | 30349.0 | Servicemember | Consent provided | Web | 2022-04-19 | Closed with explanation |
| **freq** | 1088990 | 1075668 | 638340 | 407377 | 1197 | 905287 | 427629 | 333513 | 2285.0 | 167759 | 926513 | 2180651 | 3550 |  |
| **mean** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **std** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **min** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **25%** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **50%** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **75%** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |
| **max** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |  |

6) Results/Findings

• Analyze data using R or Python so that you can provide answers to your questions.

We used python to analyze the dataset of the Consumer complaints database.

Data shape: (2653059, 18)

Issues of California state:

complaints[complaints['State'] **==** 'CA']['Issue'].value\_counts().head(10)

Incorrect information on your report 74376

Problem with a credit reporting company's investigation into an existing problem 29783

Loan modification,collection,foreclosure 22282

Attempts to collect debt not owed 15395

Incorrect information on credit report 12994

Improper use of your report 12763

Loan servicing, payments, escrow account 11195

Managing an account 11042

Cont'd attempts collect debt not owed 8454 Trouble during payment process 7782 Name: Issue, dtype: int64

Classification Report with Precision, recall, F1 score.

Precision recall f1-score support

Credit reporting, credit repair services, or other personal consumer reports 0.84 0.86 0.85 96063

Credit card or prepaid card 0.60 0.79 0.68 17627

Checking or savings account 0.84 0.86 0.85 96063

Debt collection 0.84 0.86 0.85 96063

Vehicle loan or lease 0.65 0.79 0.71 11066

Money transfer, virtual currency, or money service 0.84 0.86 0.85 96063

Mortgage 0.84 0.86 0.85 96063

Payday loan, title loan, or personal loan 0.84 0.70 0.76 43874

Student loan 0.84 0.86 0.85 96063

Consumer Loan 0.84 0.86 0.85 96063

Bank account or service 0.84 0.86 0.85 96063

Credit reporting 0.36 0.77 0.50 4278

Credit card 0.84 0.86 0.85 96063

Prepaid card 0.84 0.86 0.85 96063

Money transfers 0.84 0.86 0.85 96063

Other financial service 0.60 0.79 0.68 17627

Payday loan 0.84 0.86 0.85 96063

Virtual currency 0.84 0.86 0.85 96063

accuracy 0.83 49198868371

macro avg 0.77 0.76 0.76 49198868371

weighted avg 0.83 0.83 0.83 49198868371

Provide justifications on how and why each of your analysis answers your question(s).

• Include the result for each of the analyses. You can add more than one analysis for each of them but at least one analysis on each of them is required. Your results could be represented as one or more of the followings (but not limited to these)

7) Limitations:

b) However, companies do get a very close accuracy when it comes to classification of consumer complaint narratives with their respective products. This information can be vital when sifting through large swathes of data. Companies can better tackle and prioritize issues while CFPB can better understand certain services financial companies offer that plague customers.

We examined recent data. However, if mortgage complaints rise sharply while credit reporting, credit repair services, or other personal consumer reports rise only slightly, and if both rise close to the same level, our analysis could change. These are examples of limits to our analysis.

References:

9. Explanations/ Definitions

We installed some necessary libraries which were required to preprocess the data and implement LDA:

Spacy: This library is used to facilitate the implementation of NLP.

NLTK: It is also used to work with Natural Language Processing to use stop-words library, tokenize, stem and parse documents.

Pandas: To work with dataframes.

String: It has constants, utility functions and classes which are used for string manipulation

from nltk.corpus import stopwords: Used to get a list of stopwords, this is used to find a list of stopwords from the word corpus.

Gensim: Used for topic modeling which helps in the implementation of LDA.

Seaborn: It provides a high-level interface for drawing attractive and informative statistical graphics.

Sklearn: Simple and efficient tools for predictive data analysis