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)?

CFPB created a consumer complaint database, a collection of complaints from consumer financial products and services that companies across the United States offer. These complaints were collected from their customers. It contains information regarding the type of complaint, the product for which the complaint was issued, date of the complaint, issue, sub-issue, action is taken, etc.

• 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 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?

In this project, many software and hardware resources are used. Software resources like Excel for understanding the data, and Jupiter Notebook for data visualization. The laptop is a hardware resource, here Lenovo with Intel i7, 12 GB RAM is used for data visualization.

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

The dataset has many null values and to remove those null cells EDA-Exploratory data analysis (text data pre-processing) is required. By removing this the data shape will also reduce its size and easy to process further.

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

The limitations of the target dataset is huge in size and normal computer hardware resources will not be able to perform the tasks like Machine Learning concepts – TF-IDF, SVM, and Logistic regression fitting, Deep learning. And with this kind of data set, it is easy to train, test, and validate the data by using GPU or TPU resources, but with a regular computer, it is hard to find the accuracy of data.

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.

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

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?

Generally everyone will do some figures, plots and diagrams to tackle data of this size. It’s hard to go through such data and gather insights about it, because the dataset is large. So, here we used python to draw some insights about the data. As first step need to install and import all required python packages and import dataset. After the data should be cleaned for exploratory data concept is used and the data is cleaned by taking consumer complaint narrative as main, and we used different methods to gain several different types of insight and value.

We created a word cloud to get an idea of the common words within the consumer complaint narrative column. This would give us an idea of what the most frequent words within the consumer complaints were. A much clear idea of what kinds of issues the consumers have could be possibly derived.

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.