**Financial Consumer Complaint Classification**

Project Proposal

IST 718 Big Data Analytics | Spring 2024

Group 4

Francisco Franco Arenas

Qidu Quentin Fu

Yixing (Leo) Zhu

Ximeng Deng

**Objective**

The primary objective of this project is to develop and implement a machine learning model designed to automatically classify consumer complaints about financial products based on their textual descriptions. This project aims at improving how financial institutions manage and respond to complaints, ensuring that each one is promptly directed to the appropriate support department.

Financial institutions face significant challenges in manually sorting and responding to the vast volume of consumer complaints they receive daily. This manual process is not only time-consuming and costly but also prone to errors, leading to delayed responses and, consequently, diminished customer satisfaction. There exists a critical need for an automated system that can efficiently categorize complaints, thereby streamlining the resolution process.

Our proposed solution involves the creation of a machine learning algorithm that leverages natural language processing (NLP) techniques to analyze the content of consumer complaints. By understanding the context of the complaint text, the system will accurately classify each complaint into predefined categories. This classification would enable financial institutions to automatically route complaints to the relevant department, optimizing the resolution workflow.

**Data Set Description**

**Overview/Description**

The data used for training the model is the Consumer Complaint Database from the Consumer Financial Protection Bureau. This database is a collection of consumer complaints regarding financial products and services, which have been forwarded to companies for response. Complaints are made public post-response from the company, confirming a commercial relationship with the consumer, or after 15 days, whichever occurs first.

1. **Number of rows and columns**

The raw dataset used in this project has **4,753,265 rows and 18** columns, as of February 25, 2024. Note that this dataset is directly linked to a live database (<https://catalog.data.gov/dataset/consumer-complaint-database>), so it is updated constantly. That said, the numbers of rows and columns provided will change when the dataset is accessed later. For the purposes of this project, however, we set a cutoff date of 2024-01-01; only data published before this date, which includes approximately **4.5 million** rows, has been used.

1. **Sample predictors**

As noted previously, this project aims to classify financial customers’ complaints to shorten the complaint-solving time and better the complaint-solving. Given the problem and target of our project, the following predictors are proposed to be used in the model building: **Date sent to the company, sub issue, company, narrative, and zip code, with product (complaint type) as the target variable.** Notice that the predictors and target may be altered in the future if more information is acquired to better the modeling.

1. **Link to the dataset**

<https://files.consumerfinance.gov/ccdb/complaints.csv.zip>

1. **Anything interesting or surprising about the data**

The data consists mostly of text data (13 features of 18), which require a lot of **text cleaning** before they can be fitted into machine learning models.

The data based on their categories are highly **imbalanced**, and SMOTE, various weights assigning to the different categories, and/other methods may need to be used before fitting them into the machine learning models. For more information on the imbalance, please see the section below.

The data is updated daily. For stability reasons in training the model, we set a cutoff date (2024-01-01) for training data.

**Preliminary Data Exploration**

1. **Consumer complaint narrative**

First, we dropped duplicate rows and discarded rows lacking content in the “Consumer complaint narrative” column. Upon further examination, rows containing no more than one single word were also removed, as these typically consist of irrelevant numerical data. Following these steps, our dataset was downsized to a total of 1.7 million rows.

Next, we examined the frequency distribution of word counts in the “Consumer complaint narrative” column. The majority of the word counts were concentrated in the 2-800 word range, comprising around 1.65 million entries.

|  |  |
| --- | --- |
| Figure 1 and 2: Word count distribution in Consumer Complaint Narratives, before downsizing and after downsizing. | |

1. **Complaints frequency across years**

We examined the chronological distribution of complaint narratives in the dataset, which spans from 2015 to 2023. The number of complaints has generally increased over time, with a particularly noticeable surge from 2022 to 2023.

|  |
| --- |
| Figure 3: Frequency of Complaints per Year |

1. **“Product” and “Sub-product”**

Initially, there are 21 unique values in the ‘Product’ column and 87 unique values in the ‘Sub-product’ column, with some overlap between the categories. We evaluated the values in both to determine the categories for the final ‘Product’ column.

|  | Figure 4 and 5: Product categories and count of rows, before and after cleaning. |
| --- | --- |

Moving forward, we selected 7 columns from the original dataset’s 18, specifically including ‘Product’, ‘Sub-issue’, ‘Consumer complaint narrative’, ‘Company’, ‘State’, ‘ZIP code’, and ‘Date sent to company’. We removed rows with null values in any of these 7 columns, resulting in a refined dataset of 1.47 million rows.

In this dataframe, the distribution of “Product” across different categories is illustrated below. This distribution is notably unbalanced, leading us to consider strategies such as oversampling the minority class or downsampling the majority class to correct this. Given the substantial size of our dataset, along with considerations for computational resources and processing time, downsampling (20,000 rows per category) may be the preferable approach for the next step.

|  |
| --- |
| Figure 6: Frequency of Complaints by Product. |

Another preliminary process we propose is implementing a two-stage classifier. First, we predict whether an instance belongs to the most popular class (Credit reporting, credit repair services, or other personal consumer reports), resulting in two fairly balanced classes. In the second stage, for instances classified as ‘NO’ in the first step, we predict their class using a model that excludes the most popular class, thereby achieving a more balanced classification. In this way, we plan to randomly sample 20k rows per other category, and 160k rows for the most popular class.

**Proposed Data Exploration**

We plan on examining the most frequently used words in complaint narratives for each Product category. We also aim to explore the relationships between columns, especially between Company and Product, and between the State and Product.

**A list of proposed specific data explorations and proposed predictions**

* Word frequency analysis of complaint narratives for each product category
  + Identify the most common words and phrases within complaint narratives for each product category to highlight prevalent issues or concerns using visualizations (barcharts, pie charts, word clouds, etc.)
* Part of Speech (POS) analysis of complaint narratives for each product category
  + Understand the grammatical structure of complaints in different product categories and analyze the distribution of POS tags within each category to identify patterns and trends.
* Sentiment analysis of complaint narratives for each product category
  + Calculate sentiment scores by product categories and identify which products are associated with more positive or negative feedback.
* Company and product relationship analysis
  + Explore the patterns or trends in complaints related to specific companies and their associated products using heatmap visualizations or association rule mining.
* State and product relationship analysis
  + Understand how the geographical distribution of complaints varies by product category and visualize the relationship between states and product categories using interactive maps.