Multi-class classification of Consumer Complaints on Financial Products: An Analysis with Multinomial Naive Bayes and XGBoost

Machine Learning Engineer Nanodegree - Udacity

Capstone Project Proposal

Georgios Spyrou

14 October 2021

1. Domain Background

Complaints from customers have always been a vital component of every successful company/organization. When the complaints are well-structured and provide insightful (rather than plain judgmental) feedback regarding the consumers experience with the product or service, they can assist the organizations to understand their customers needs - which in the long run could lead to satisfied customers and thus to higher earnings. [1]

Hence, it's of high importance for the organizations to have highly responsive complaint-resolution mechanism between themselves and the customer. For that purpose, there are common Natural Language Processing (NLP) techniques and algorithms, that while working together could assist the companies to "build" such mechanisms to evaluate and respond to their customers complaints effectively.

2. Problem Statement

In this project we will focus on financial products. Specifically, we will build an automatic product categorization solution based on the customer's views regarding these products, and use the customer's complaints themselves to automatically assign them to one of the predefined categories. Predicting the correct category for a complaint can be very beneficial for the financial institutions that have to deal with responding to hundreds of complaints per day. If the customer who is filling the complaint does not submit it under the correct category, then that becomes a cost for the financial institution as -usually- they will afterwards have to:

- Identify that a complaint is under the wrong category in the first place, which is something that happens only after someone has already spent time with the case

- Re-classify and redirect the complaint to the correct category and the appropriate personnel for further examination

3. Datasets and Inputs

The dataset used in this project contains complaints that have been made by consumers regarding financial services and products (e.g. student loans, credit reports, mortgage, etc) in the United States between January 2019 and December 2020. Each of the complaints is marked to belong under one Product Category.

The seven **Product Categories** that a complaint can belong to are:

- Checking/savings account
- Credit reporting
- Credit/prepaid card
- Debt collection
- Loans
- Money services
- Mortgage

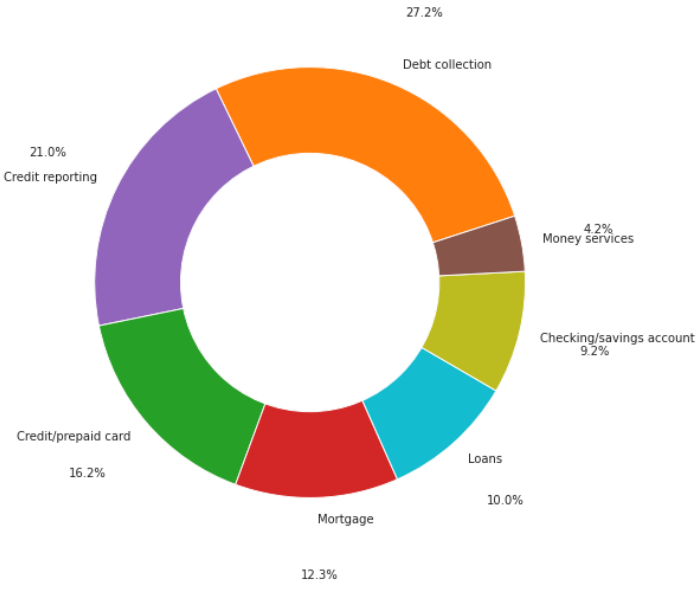
Please note that the original dataset contains data from 2011 to 2021 and it is publicly available. Furthermore, it's getting updated daily, something that makes it quite large. Hence, for the purposes of this project we will use a subset of this dataset and focus only at the aforementioned time period of two years (2019 and 2020).

Specifically, the original dataset contains about 2.3 million rows and 18 features, such as 'ZIP Code', 'Company', 'Complaint', 'Date Received' and more. As we have mentioned above, we will use a subset version of this dataset to contain complaints for two years. Hence, after subsetting the dataset we end up with approximately 400.000 rows. Please note that almost 50% of these rows do not actually contain a complaint in form of a text and therefore they are gonna be filtered out further from our analysis as they won't provide any useful input to our machine learning algorithms. Hence, the final dimension of our dataset will be close to 200.000 rows/number of unique complaints. Regarding the features of the final dataset, we will initially start with:

1. *Year*: Year that the complaint was raised
2. *Complaint*: The complaint itself as raised by the consumer in text form
3. *Product*: The product categories that the complaint can belong to (see above)
4. *Company*: Organization against which the complaint was raised
5. *State*: The US State where the complaint was submitted from

Note that Year, Company and State will helps us during our EDA exploration, but they will be later dropped as the only inputs to our algorithms will be the text/complaint from the consumer (i.e. 'Complaint') and the product category that it belongs to (i.e. 'Product').

Finally, note that after data preprocessing and data cleaning, the distribution of product categories across the dataset will not be exactly uniform, and it will follow a distribution as presented below:



We can observe that the first two major categories (Credit Reporting and Debt Collection) constitute close to 50% of the overall amount of complaints in the dataset. Therefore, we should take this imbalance between categories into consideration when we are picking the appropriate evaluation metrics for our algorithms.

More information regarding the dataset can be found in the accompanied README.md file.

4. Solution Statement

The format of the data makes them ideal for supervised learning purposes, with the text (complaint from the consumer) as the **input**, and the category that the complaint belongs to as the **target** variable.

To construct the final model that will be able to identify and categorize consumer complaints to their relevant product category, we are proposing the implementation of a **XGBoost** classifier, which is a powerful solution for classification purposes, while also working well with high dimensional/complex datasets. [2]

5. Benchmark Model

Multinomial Naive Bayes (MNB) is an algorithm that is regularly being used for text classification tasks [3]. This model is providing good predictive performance, while keeping low the computational cost - making it a powerful tool, especially in an industry setting where time to deliver is a major factor.

In this project we are proposing a Multinomial Naive Bayes model as a benchmark solution, which will act as a "simple" but powerful baseline for categorizing the complaints. Thus, we will attempt to answer the question: *Would a more sophisticated model, like XGBoost, vastly outperform a simple prediction model like MNB* ?

6. Evaluation Metrics

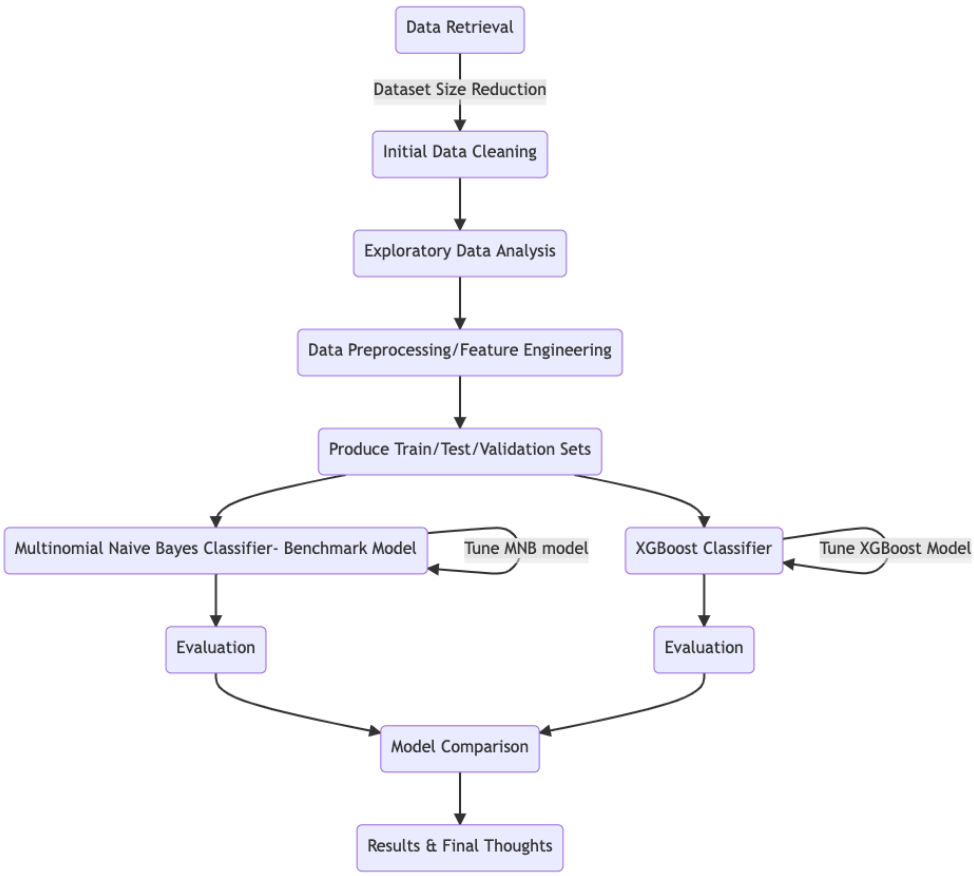
Given that we are interested in building a process where we can assign as many complaints to their correct category as possible, we are proposing the **F1 Score** as the main evaluation metric. Specifically, we will try to optimize and compare the models on the '*macro*' version of the F1 score, which does not take into consideration the specific

weights (i.e. number of instances per class) and thus assigns equal importance to each class. That way we can optimize the algorithms to favour all classes and not the majority ones. [4]

7. Project Design

The project will follow the structure below:

2. *Data Retrieval & Dataset Reduction*: Retrieve data from source and reduce dataset to the required size and the specific years of interest.
3. Initial Data Cleaning: Data cleaning tasks (e.g. removal of rows that don't contain a complaint).
4. Exploratory Data Analysis: Use summary tables and visualizations to further understand the data.
5. *Data Preprocessing/Feature Engineering*: Tokenization of complaints, further cleaning (e.g. removal of punctuation), lemmatization, text vectorization with Bag of Words, and more.
6. *Train/Test/Validation sets creation*: Split data to subsets that will be used for training, tuning and performance evaluation and generalization .
7. *Implementation of MNB and XGBoost models*: Algorithm creation, training and tuning.



8. *Model Evaluation and Comparison*: Evaluate the results, compare models in terms of performance and draw conclusions regarding the usefulness of each algorithm.

References

- https://www2.deloitte.com/content/dam/Deloitte/us/Documents/regulatory/us-aers-the-power-of-complaints-042115.pdf [1]
- https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions [2]
- https://www.cs.waikato.ac.nz/~eibe/pubs/FrankAndBouckaertPKDD06new.pdf [3]
- https://en.wikipedia.org/wiki/F-score [4]