**Improving Credit Card Services**

**Natural Language Processing of Consumer Complaints**



**Jun Tian**

**February 2021**

1. Problem Statement

Handling consumer complaints is a common practice as well as an important business operation for many financial companies. The approaches and procedures financial companies took regarding consumer complaints would impact not only the quality of their services, consumer loyalty, but also satisfaction of governmental regulations. This project aims at identifying appropriate ways to improving credit card services through analyzing consumer complaints data in the past and providing insights for financial companies.

The Consumer Financial Protection Bureau (CFPB) manages a Consumer Complaint Database that collects complaints about consumer financial products and services. In total, there are more than 3 million records/entries in the database, ranging from December 1, 2011 to current date. More than 8% of the complaints are relevant to general purpose credit card. In 2020, there were 23,940 consumer complaints on credit card, in the CFPB database.

Complaints can provide insights into problems people are experiencing in the marketplace and help financial companies regulate consumer financial products and services under existing federal consumer financial laws, enforce those laws judiciously, and educate and empower consumers to make informed financial decisions. This project will dive into consumer complaints on credit card, study the patterns and trends of complaints over the years, investigate on what and how financial companies respond to complaints, and derive recommendations for financial companies to provide better credit card services.

Through data mining, analysis and machine learning applications, it is expected to identify an approach that could streamline consumer complaints resolutions. The overall goal is to shorten processing time on responses, saving human resources of financial companies, as well as improving consumer satisfactions. **The research problem is, given the complaint narratives, how could financial companies respond to make consumers satisfied?**

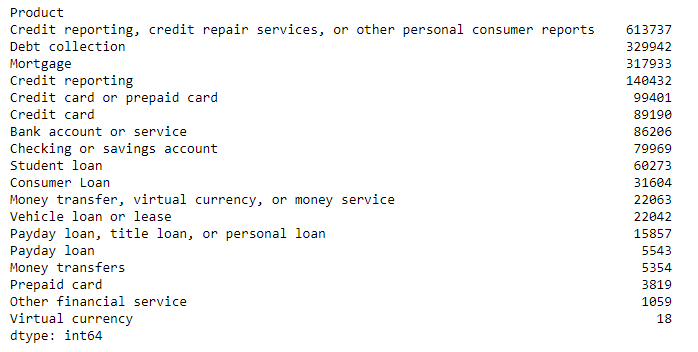
To achieve the research goal, this project will investigate and analyzing available credit card relevant complaints and responses, and will address the following questions:

* What are the patterns of credit card complaints?
* What are emerging trends about complaints?
* How did the financial companies respond to the complaints?
* Are the responses effective?
* How could financial companies improve on consumer complaints resolutions?

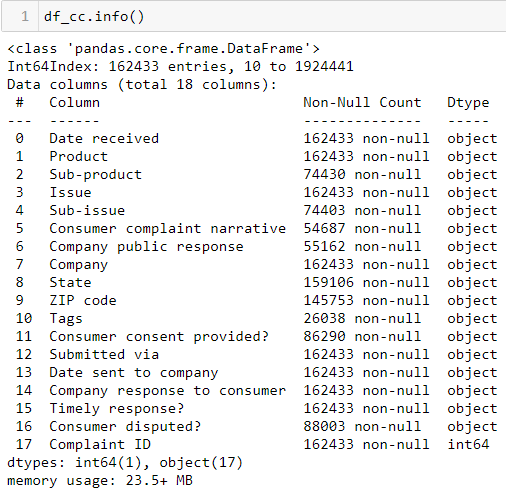
1. Data

Consumer Complaint Database is a federal managed data collection of complaints about consumer financial products and services. The Database is a rich resource for data analysts and data scientists searching for historical as well as emerging trends about consumer complaints with relevant to financial services products, including reasons for those complaints and actions financial institutions are taking to resolve them. The Database generally updates daily and is open for public access.

CFPB complaint data is organized by products. In total there are 18 products in the dataset, covering various financial products and services. The following table overviews the consumer complaints during the period of December 1, 2011 to current date. As can be seen, there are two 'Product's relevant to credit card, which needs to be reconciled.



For each of the data points, there are 18 variables and they are shown as below. A closer look into each of these variables is needed to check the integrity and consistence of each variable. Machine learning applications will be focused on two variables: Consumer complaint narrative, and Company response to consumer.



1. Methodologies

Data wrangling is the very critical part of any data science project. The quality of data determines the outcome. In this process, inconsistences of data definition, missing values, outliers, and duplicates will be cleaned up for later process. A clean dataset will be prepared in this procedure for exploratory data analysis and machine learning modelling.

Exploratory data analysis will be followed to explore various aspects of the dataset in terms of better understanding consumer complaints and corresponding company responses. Several of the research questions will be answered through inference statistics and data visualization.

Natural Language Process (NLP) will be adopted for text analysis and data preparation for machine learning applications, as we are concerned about consumer complaints and it takes a form of language narratives recorded in English in this dataset. We will follow the steps of NLP to preprocess the text of complaint narratives.

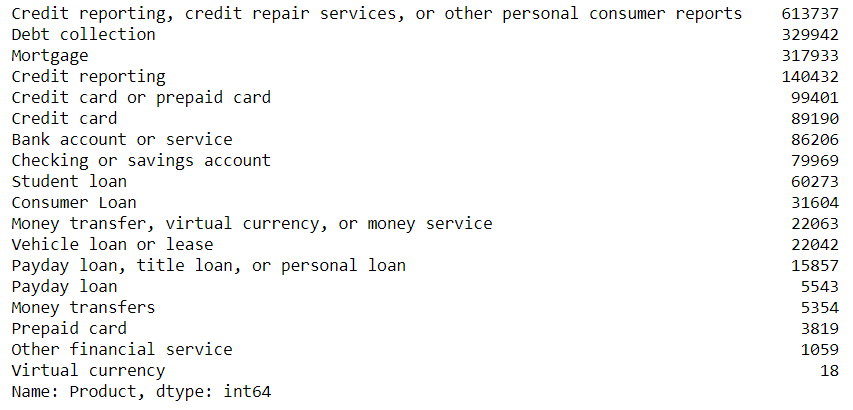
The target variable is company responses towards consumer complaints with consumer satisfaction, meaning no consumer dispute after company responses were taken. The goal of machine learning modelling is to predict company responses based on complaint narratives and consumer satisfaction. Three different algorithms will be applied for this classification problem to help financial companies automatically allocate consumer complaints into different company response categories, thus it would certainly help financial companies to streamline their consumer complaint resolutions, save on human resources, while satisfying consumers and federal regulations.

Accordingly, recommendations will be provided to financial companies for their strategies on consumer complaints resolutions.

1. Data Wrangling

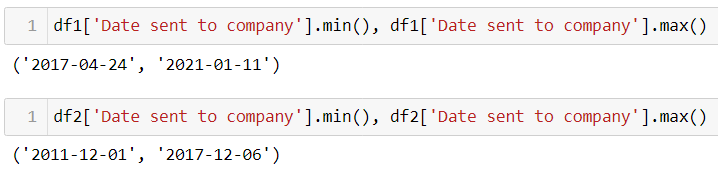
4.1 Data preparation and overview

The very first step is to access and pull original data from CFBP website. After doing so, we have more than 3 million records of consumer complaints, regarding 18 financial products and services, and ranges from December 1, 2011 to January 12, 2021. From millions of records, we would like to extract those consumer complaints that are specifically on credit card. An overview of the dataset by ‘Product’ is as follows:

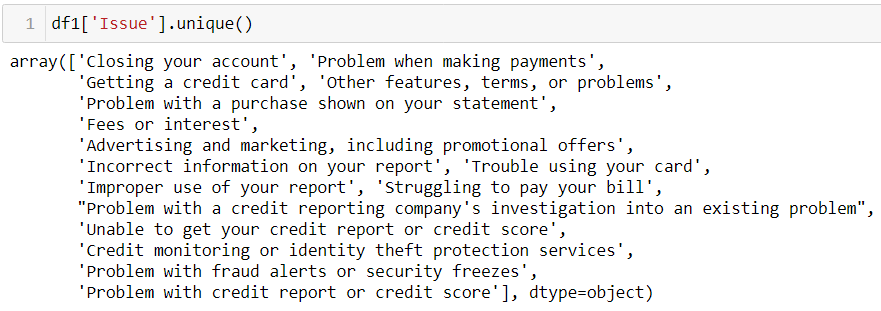


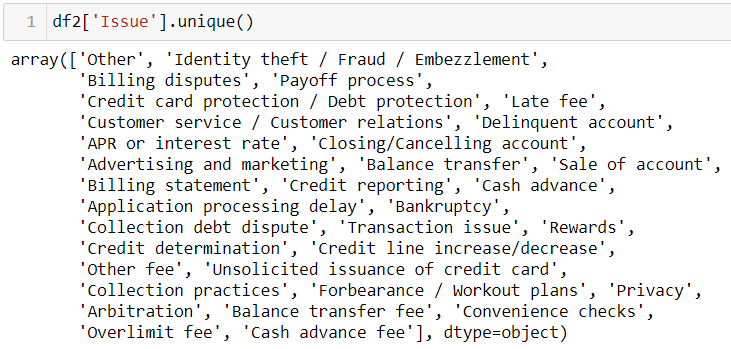
As can be seen, there are two ‘Product’s that are relevant to credit card: ‘Credit card or prepaid card’ and ‘Credit card’. A research on CFPB website found out that CFPB changed the categorization of 'Product' in 2017[[1]](#footnote-1), when CFPB combined previous 'Credit card' and 'Prepaid card' into 'Credit card and prepaid card', under which 'General-purpose credit card or charge card' refers to credit card. This discrepancy will be taken care of in the data wrangling process, in which other inconsistences will also be addressed.

In order to select the two datasets out and combine them into one file, we assign ‘General-purpose credit card or charge card’ under ‘Credit card and prepaid card’ as df1, and product ‘Credit card’ as df2, a check on the dates found out that the time range of these two datasets align with the timeline when CFPB adjusts their definition.

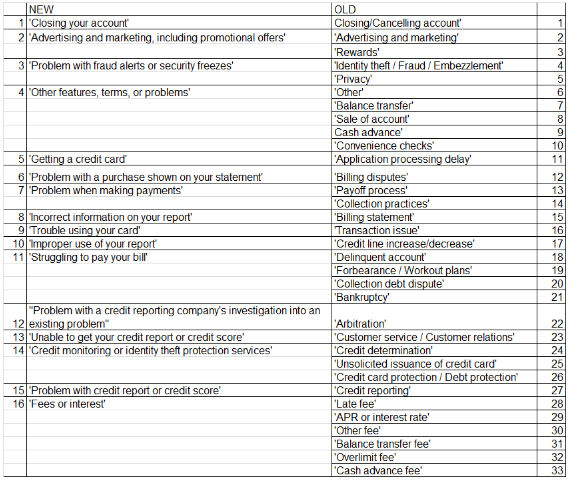


For the purpose of merge these two datasets, it is necessary to check and make sure they are compatible, meaning that other aspects of the two datasets match with each other. However, when checking on the variable ‘Issue’, it is obvious that the ‘Issue’ definition under df1 and df2 are totally different.





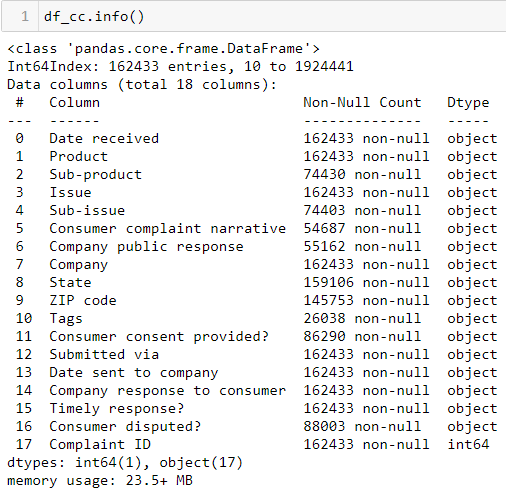
As can be seen from above, when CFPB made the change to 'Product' in 2017, they also changed 'Issue'. For old data entries, there were 33 'Issue's, while there are 16 'Issue's for newer data entries. For further analysis, we need to mapping the 33 old 'Issue's into the 16 new 'Issue's. A mapping table of 33 old ‘Issue’s to the 16 new ‘Issue’s is as follows:



After unifying the definition of ‘Issue’, the two datasets are then merged together. For the purpose of keeping data for entire year, we cut off the data points before January 1, 2012 and those after December 31, 2020. Then we have one combined dataset that is specifically on credit card.

4.2 Data Wrangling

An overview of the prepared dataset is presented as below:



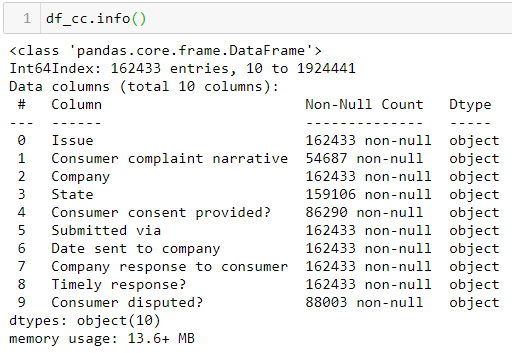
*Dropping unnecessary variables*

A close check on the variables revealed that for our purpose of exploratory data analysis and machine learning modellings, some of the variables are not necessary. After data exploration into definition of each variable, we decided to drop the following variables, and the reasons of dropping are also outlined:

**'Product' and 'Sub-product':** all complaints are relevant to credit card only  
 **'Tags':** this column does not covey important meaning.  
 **'Sub-issue':** There are already a lot of issues, sub-issue can be dropped.  
 **'ZIP code':** 'ZIP code' is highly relevant to 'State', but 'State' conveys message more directly than 'ZIP code'.  
 **'Date received':** There are two dates: 'Date received' and 'Date sent to company', in this analysis I will keep only 'Date sent to company', because 1. Most values of the two columns are identical; 2. Slight difference might bed reflected through 'Submitted via'; and 3. 'Date sent to company' is a more accurate measurement for 'Timely response?'  
 **'Company public response':** Since around half of the 'Company public response' are missing, and among 68,598 valid values 64,462 are 'Company has responded to the consumer and the CFPB and chooses not to provide a public response'. We decided to delete this column.

**‘Compliant ID’**: a unique ID that was assigned to each complaint.

After dropping above unnecessary variables, we got a dataset with the below basic information:



*Dealing with missing values*

The dataset has quite a log missing values. Out of the 162,433 records there are only 54,687 with ‘Consumer complaint narrative’. There is no way to fill the missing values of complaint narratives, but we also do not want to lose values of those data points without a narrative, therefore we decided to have two datasets for further analysis.

For Exploratory Data Analysis, we would like to keep all the data points by filling all the missing values as ‘NA’:

**Consumer complaint narrative:** only around 35% of the entries have a valid 'consumer complaint narrative', we will fill missing values with 'NA'.  
 **State:** will fill missing values with 'NA'.  
 **Consumer consent provided?:** will fill missing values with 'NA'.  
 **Consumer disputed?:** will fill missing values with 'NA'.

For Machine Learning Modelling, We will keep data only with valid consumer complaint narrative and with satisfied company response (‘Consumer disputed?’ = ‘No’).

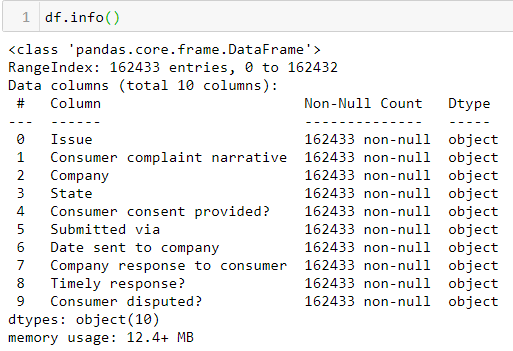
*Other Issues*

Outliers – none of the variables are numerical. A check on the variables proved that all data are within reasonable range.

Duplicates – there is no duplicates in the dataset.

1. Exploratory Data Analysis

Here is the overview of the EDA dataset:



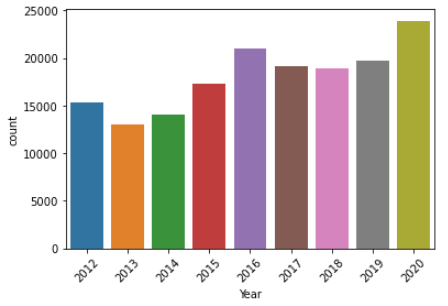
Adjusting data type

An overview of the original data types reveals that all of the data types are ‘object’’, The following adjustment need to be made:

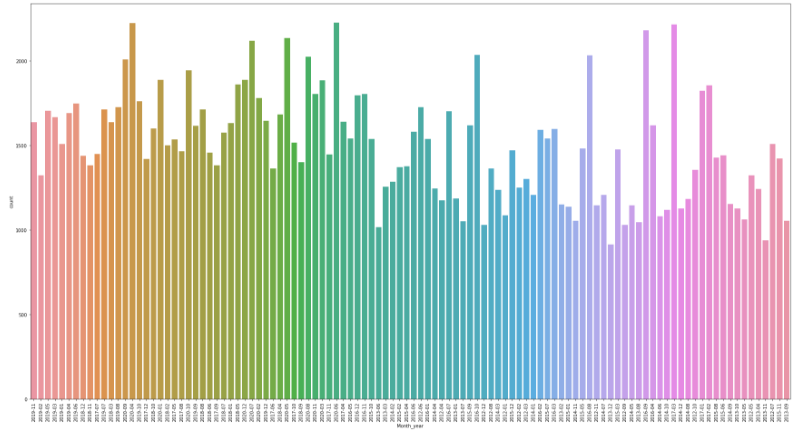
**Date time:** 'Date sent to company'  
 **Boolean:** 'Timely response?', 'Consumer disputed?'  
 **String:** 'Issue', 'Consumer complaint narrative', 'Company', 'State', 'Consumer consent provided?', 'Submitted via', 'Company response to consumer', 'Complaint ID'

5.1 Complaint trends

Through the years, we can see that consumer complaints on credit card have been changing. Overall, there is a growing trend in numbers of complaints, from22019 to 222010, there is an obvious increase.



Monthly counts of complaints are spread more evenly. For recent years, complaints are really concentrated on a few months of the year.



5.2 Complaint breakdown

*By Issue*

A close look of consumer complaints found out that complaint issues are highly skewed. Out of the 162,433 complaints on 16 issues, complaints on ‘Problem with a purchase shown on your statement’ is the most complained issue, far more than any of other issues, accounting for more than 36,000, about 40% more than the second complained issue. Consumer complaints on ‘Other features, term, or problems’ and ‘fees or interest’ are the second and third complained issues respectively, both account for more than 20,000 complaints.

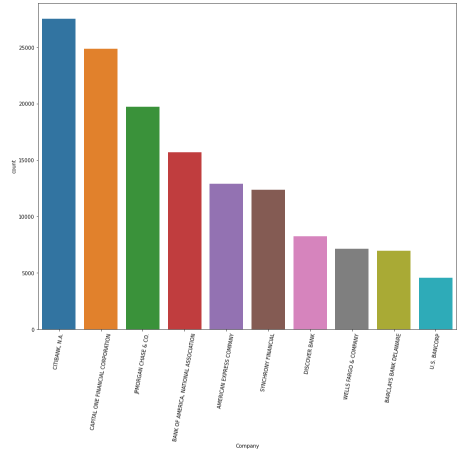
On the other hand, complaints on ‘Credit monitoring or identity theft protection services’, ‘Problem with credit report or credit score’, and ‘Problem with a credit reporting company’s investigation into an existing problem’ are the least three.



*By Company*

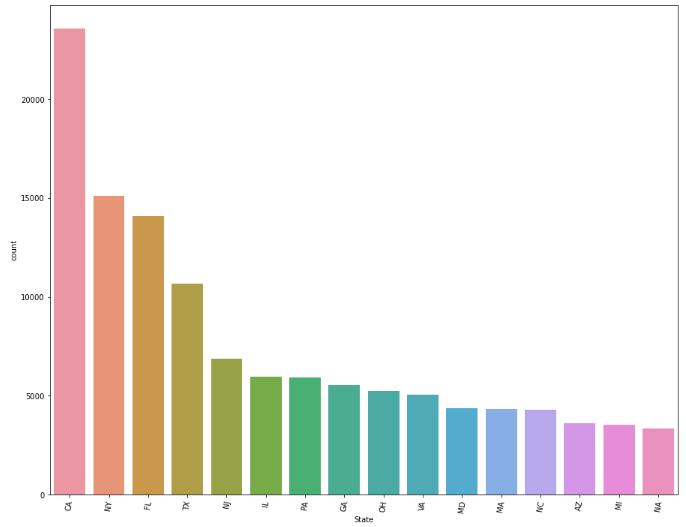
The below graph shows the top ten companies that received the most consumer complaints on credit card. Citibank, Capital One, and JPMorgan Chase are the top three companies receiving complaints, with Citibank more than 27,000 complaints, Capital one slightly less than 25,000 complaints and JPMorgan Chase slightly less than 20,000 complaints.

This complaint pattern might be proportionally corresponding to the amount of credit card issued by each financial company, or might indicate the quality of services. External data on amount of credit card issued is needed for judgement.



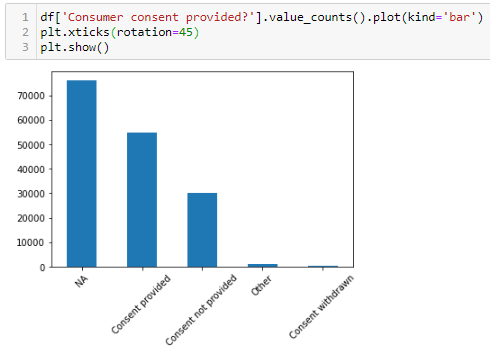
*By State*

The below graph shows the geographical distribution of consumer complaints on credit card. Not surprisingly, California is the State with most complaints, amounting to about 24,000. New York and Florida are with the second and third most consumer complaints on credit card, reaching up to slightly less than 15,000 complaints respectively. Texas is the No.4, with more than 10,000 complaints. This distribution pattern reflects the population in each state.

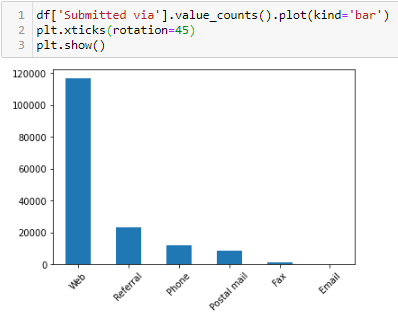


*Other Issues*

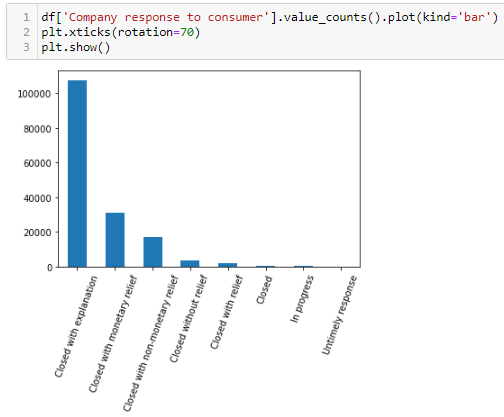
A look at ‘Consumer consent provided’, excluding the largest portion of null values, ‘Consent provided’ is much more than ‘Consent not provided’.



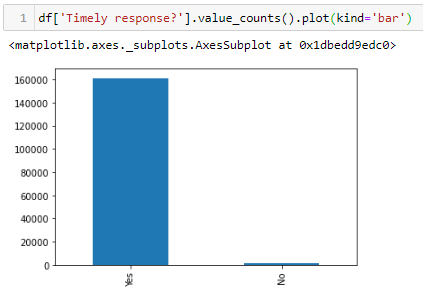
Most of the complaints are submitted via Web, while there are still significant amount of complaints submitted via Referral, Phone, Postal mail, Fax and Email, which are more likely happened in earlier years.



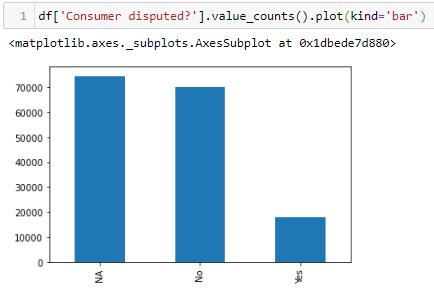
‘Company response to consumer’ are highly concentrated on ‘Closed with explanation’, which accounts for more than all other responses add up together. The second and thirdly likely company response are: ‘Closed with monetary relief’ and ‘Closed with monetary relief’.

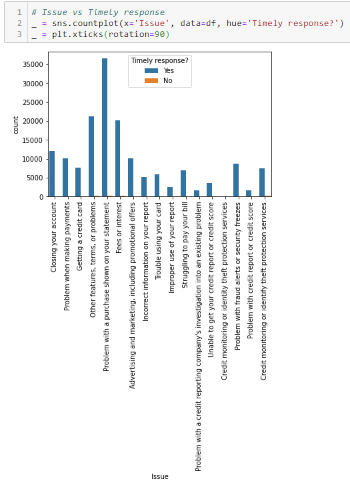


The majority of consumer complaints got timely response from companies:



Many data are not available to show whether a consumer disputed or not on company response. In the remaining data, the majority of the consumers did not dispute on company response, which might be used approximately as a measurement of consumer satisfactory.





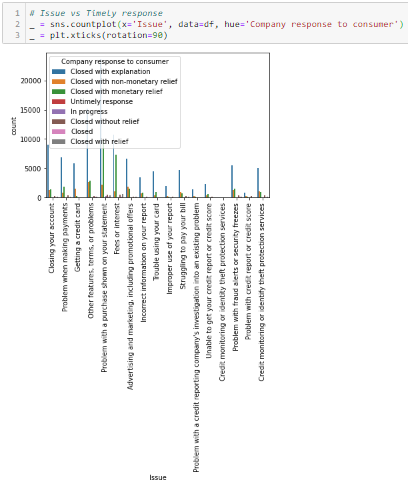
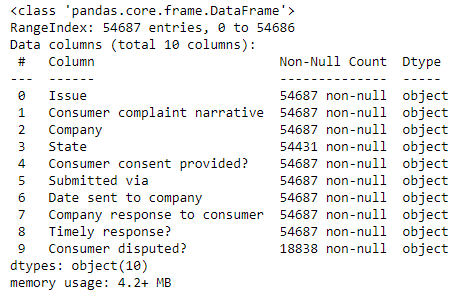


Figure: Issue vs. timely response

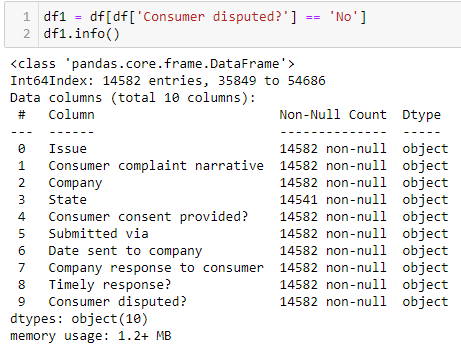
1. Pre-Processing and Machine Learning

6.1 Pre-Processing

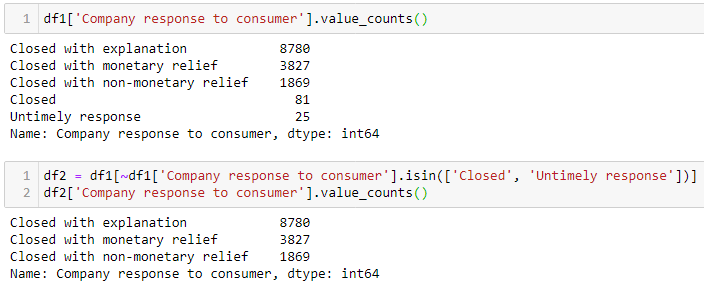
As the first step, we would like to access to the dataset. Recall that in the process of Data Wrangling, we have prepared a separate dataset for machine learning. Here is an overview of this dataset:



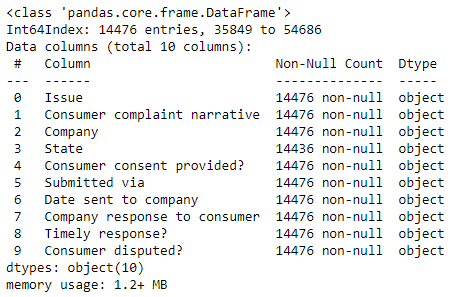
We would like to extract those data points that when company responded, they did not dispute, which we adopt as an approximate measurement as consumer satisfactory of company response.



As we also take a look at the company response to consumer, we would like to drop the values of ‘Closed’ and ‘Untimely response’.



After all these process, now we have a dataset that is ready for NLP text analysis and machine learning modelling, for which we have complete ‘Consumer complaint narrative’ and ‘Company response to consumer’ information, and with ‘Consumer disputed?’ values equal to ‘No’.



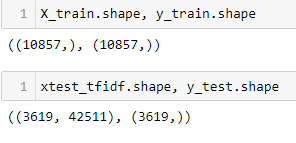
*Natural Language Process*

As consumer complaints are narratives, we adopt natural language process for text analysis. Following the standard steps of NLP, we converted all text data to lower case, removed punctuations, worked on text standardization, removed stopwords, corrected spellings, and lemmatized the text.

6.1 Machine Learning Modelling

*Train test split*

Assign dependent variable (y) as ‘Company response to consumer’ and independent variable (X) as ‘Consumer complaint narrative’. We applied label encoding on y and tf-idf vector representation of X. Then considering the skewness of values of y, we did a stratified train test split.

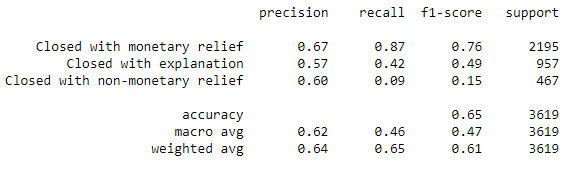


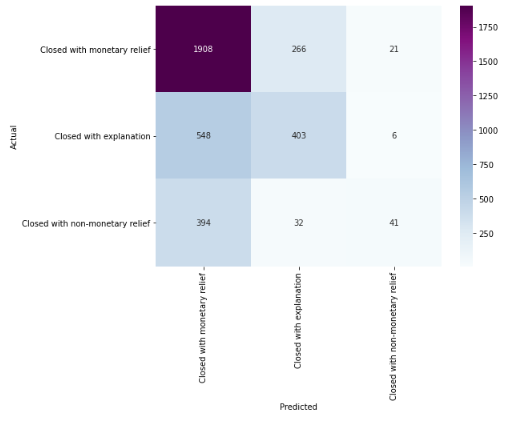
*Machine Learning Modelling*

Since this is a multi-classification problem, we are going to apply three different algorithms for machine learning and then compare the parameters to select the best model. The algorithms we are going to apply are: Logistic Regression, SVM, and XGBoost.

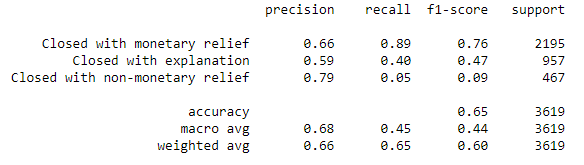
For each of the modelling process, we also conducted grid search to identify the best hyper parameters deployed in each algorithms. The following graphs demonstrate modelling outcomes of each algorithm:

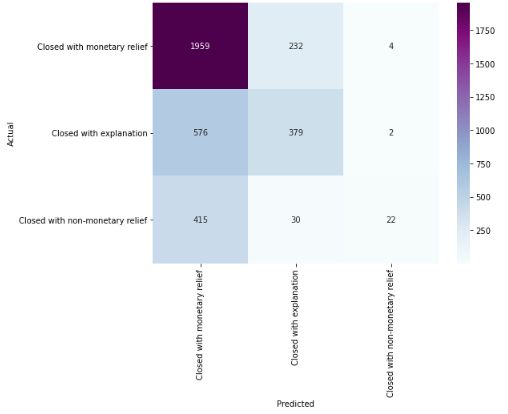
Logistic Regression:



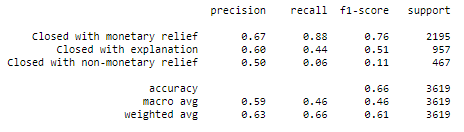


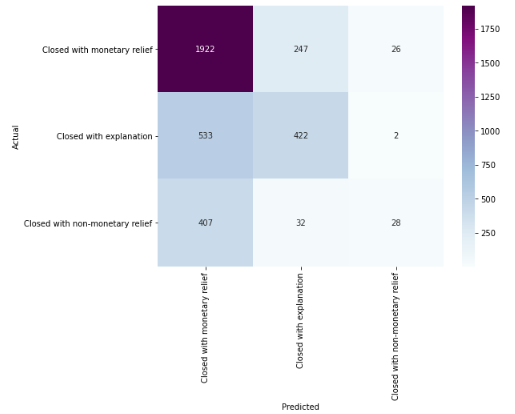
SVM:





XGBoost:





*Modelling Evaluation*

From the modelling we can see that all three different algorithms result in comparable results, with approximately 65% accuracy rate. Comparing to the proportion of each of the three company responses in the dataset, our machine learning model provides a large improvement in terms of determining what kind of company responses should be taken.

1. Findings and recommendations

Consumer complaints and corresponding company responses provide an insight for financial companies derive important decisions and strategies that could help with improving their services and products. This research dive into consumer complaints on credit card, through exploratory data analysis and machine learning modelling, we have the following findings:

Our machine learning model provided a convenient strategy for financial companies to adopt, with regard to company responses based on consumer complaint narratives. Overall, our model has an accuracy of 65% of predicting the correct company responses that make consumers satisfactory, although the precision rate for the three categories may slightly vary. Comparing to the baseline of 61% (8,780 out of 14,476) of response falling into ‘Closed with explanation’, 26% (3,827 out of 14,476) in ‘Closed with monetary relief’, and 13% in ‘Closed with non-monetary relief’ (1,869 out of 14,476), our model shows a huge improvement in correctly classify company responses towards consumer complaints. Integrating our model in the process of handling consumer complaints, financial companies could improve effectiveness, save on human resources and therefore cost.

During the research process, we found that there are potential rooms that could help improving the modeling and accuracy:

1. There are a lot of missing data on company responses and on whether or not the consumer is dispute or not. For those who disputed company responses, there is no follow up with what ultimate response did the companies took. Those missing information might provide a lot of insights.
2. From CFPB we extracted data ranging from 2012 to 2020. As financial trends may change over time, it should be beneficial to look at only data for recent years.
3. CFPB data covers a large range of companies, while there are similarities, but different companies may tend to have different strategies (systematic error), which can be corrected using the companies own data. Thus, if our model could be combined with financial company’s own database on consumer complaints, improvement of accuracy is expected.

1. <https://files.consumerfinance.gov/f/documents/201704_cfpb_Summary_of_Product_and_Sub-product_Changes.pdf> [↑](#footnote-ref-1)