Consumer Complaint Analysis and Classification using Hypothesis Testing and Machine Learning

Kashmira Chawan
21261109
Dublin City University
kashmira.chawan2@mail.dcu.ie

Bhargav Anant Athavale
20210278
Dublin City University
bhargav.athavale2@mail.dcu.ie

***Abstract*—Consumer Financial Protection Bureau (CFPB) is an agency of the United States of America that is tasked with protecting the consumers in the financial sector. It has jurisdiction over various banks, debt collectors and other financial institutions. It is also tasked on behalf of the consumer with sending their complaints to various institutions within the country. Whenever the concerned company responds or within a time of 15 days these are published over the CPFB's website.**

We are analysing a dataset for the US Consumer Finance Complaints acquired from consumerfinance.gov website and available on Kaggle. It consists of 555,957 records from January 2012 to December 2015. Chi – Square Test will be carried out to test our hypothesis. We will also build a machine learning model to sort the upcoming complaints according to the product type. For this purpose, we make use of Logistic Regression. The purpose of this paper is gain insight into the data to help improve the service in the financial market.

Keywords — Machine Learning, Logistic Regression, Chi Square Test, TF-IDF

I. Introduction

In the aftermath of the 2007-08 financial crisis and the subsequent recession of 2009, the United States government passed the Dodd–Frank Wall Street Reform and Consumer Protection Act to regulate the financial industry [7]. The Consumer Financial Protection Bureau was formed as part of the process to monitor various financial institutions such as banks as well as non-bank financial institutions such as debt-collectors and mortgage collectors.

The CFPB opened its website in 2011 for consumers to lodge their complaints. By 2016, hundreds of complaints were registered with the agency with regards to consumer loans, student loans, credit reporting information, credit cards along with mortgage debt. Once an institution has assets worth $10 billion it comes under the rules and regulations of the CFPB [10]. The institute has been able to help around 29 million customers to the tune of $12 billion in refunds and debts after curtailing various malpractices with respect to various debt collection and mortgage collection [8]. It has also helped customers to understand various new virtual currencies such as Bitcoin.

By analysing the various consumer complaints to understand their relations with each other, it would be useful to predict the product type for the complaint for its speedy resolution. Also, this would improve the services offered by CFPB and build future complaint resolution models which would be time efficient and accurate.

II. Related Work

A study conducted by Yunita, Annisa, Catur and Muljono with chi-square for feature selection and k-means algorithm for classification process claims that they obtained greatest results with accuracy value of 65% and defines that dataset with feature selection has the advantage of being more efficient in the calculation process [1].

Another study by S. Rosidin, Muljono, G. Fajar Shidik, used SVM and Naive bayes algorithm for text classification. According to study the SVM gave 85.56% and Naive bayes gave 85.19% of accuracy. When the chi-square selection feature was added there was decrease in accuracy of SVM whereas, accuracy increased for naïve bayes algorithm. However, compared to other studies, using the chi-square test for feature selection has increased the accuracy for both the algorithms [2].

In a study conducted by Yao Chen, Jinfei Wang and Zhengyu Cai concluded that by standardized processing flow for complaints, we can improve quality of services and their efficiency. The government can grasp industry issues quickly, effectively, and accurately. They used SVM algorithm and Naive Bayes algorithm to solve to classify the complaints [3].

Eka Rahayu Setyaningsih and Indah Listiowarni conducted a study on categorization of 600 high school biological questions into Bloom's Taxonomy of cognitive domain. They concluded that using Naive Bayes classifier with Chi-Square as its feature selection, accompanied by Laplace Smoothing, improves its accuracy [4].

III. Data Description

The 'US Consumer Finance Complaint'[5] data consists of around 555,957 complaints and 18 columns. Among them, some of the important columns for us are as follows:

- date_received: The date the complaint was received.
- product: The type of product for which the complaint was received. For example - loans
- sub-product: The type of sub-product for which the complaint was received. For example – student loan, consumer loan
- issue: The primary issue with respect to the complaint. For example – Account management
- sub_issue: The sub-issue associated with the complaint. For example – debt payment
- company: The name of the company the complaint was issued against
- state: home state of the consumer
- submitted_via: The medium of communication for sending the complaint
- company_response_to_consumer: How the company responded
- timely_response: Was the company response within the time limits

A. Data Cleaning and Processing

For our analysis, we need to clean the dataset and only keep the necessary columns as well as rows. We observed that 'consumer_complaint_narrative' contained the greatest number of null values and decided to drop such rows to get a more focused dataset. Further we also check how much data is missing in each column and drop those columns which are necessary to our research questions, thereby reducing our dataset to 66,806 rows and 12 columns. We perform exploratory analysis on it to visualise and understand our dataset further.

B. Exploratory Analysis

For our figure 1, we created a simple bar chart graph to understand the number of complaints against each product. We can see that most complaints were against 'debt collection' followed by 'mortgage' and 'credit reporting.'

In figure 2, we see which companies have the highest number of complaints against them. We can conclude that some of the biggest banks have the highest number of complaints against them.

In figure 3, we see whether the complaint filed by the customer was responded to by the company within the given time frame of 15 days.

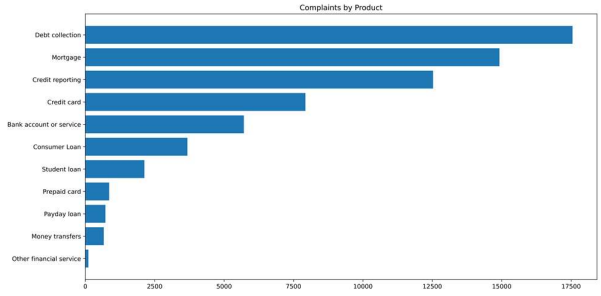


Figure 1:Complaints by Product

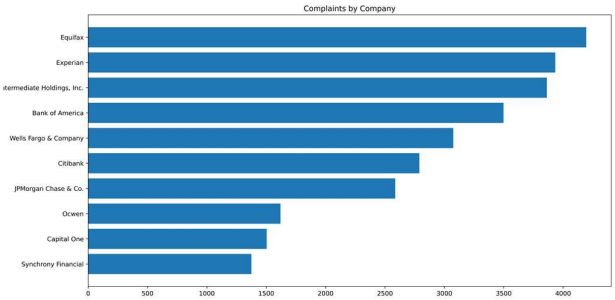


Figure 2: Complaints by Company

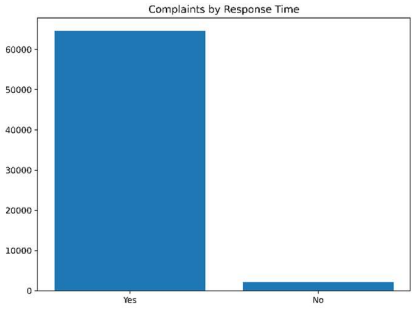


Figure 3:Complaints Timely Responded

In figure 4 , describes a map of US states which describes the number of complaints by state to visualise which state had the most complaints filed.

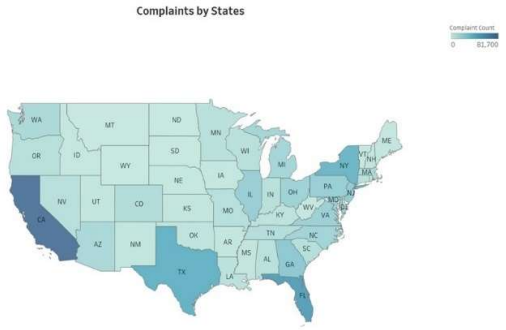


Figure 4: Complaints by State

IV. Hypothesis and Research questions

During data processing and exploratory analysis, we found the nature of the dataset and from that we came to know most of the dataset has categorical variables.
The hypothesis testing can be determined as follows:

- Null hypothesis (H0): There is no relationship between variables.
- Alternative hypothesis (H1): There is a relationship between variables.

Let's choose p value = 0.05 with 95% of confidence interval. p > 0.0.5 - It means that the test result will lie in the acceptance region, and we will accept the null hypothesis. i.e. (A ≠ B)

P < 0.05 - It means that the test result will lie in rejection (critical region), and we will reject null hypothesis and accept alternative hypothesis i.e. (A = B)

Research Question:

1. Does the company response depend upon the product for which the complaint is filed?
2. Which category do complaints belong to?

V. Methods Used

A. Chi-square Test:

We used this method to check if 'company_response_to_consumer' column, which is a categorical variable, is dependent on the 'product' column. We did this in the following way:

1. We first made a contingency table for observed data and then calculated the expected values.
2. We have chosen the p-value to be 0.05.
3. Degree of freedom was calculated to check the number of variables to be estimated to complete the dataset which was also required for chi-square statistic calculation.
4. Chi-Square depends upon observed and expected values, in our case the chi-square value came out to be 1815.74 which was the pretty good value to determine the relation between two variables.
5. Critical value is less than chi-square value then the result for this test statistically significant.

```
Significance level:  0.05
chi-square statistic: 1815.7415736890111
critical_value: 18.307038053275146
p-value: 0.0
```

Since the p-values are less than 0.05 we reject the null hypothesis and accept the alternative hypothesis, which means that the company's response is highly dependent on specific product complaints.

We had most of the columns in object datatype so instead of doing chi-square test on each column we tried to do it for all attributes together using a for loop considering the product as the main variable. Using the column, we can decide which column to utilize. We created a contingency table also called a crosstab they are used to summarize between several categorical values. Using *pd.crosstab* we gave the *product* as one categorical variable and *company_response_to_consumer* as another categorical variable. Output we get is the count of occurrence in each bucket i.e., value for each category in *company_response_to_consumer* column with respect to *product.*

We get the following output.

	features	Hypothesis
0	date_received	Reject the null hypothesis
1	issue	Reject the null hypothesis
2	consumer_complaint_narrative	Reject the null hypothesis
3	company	Reject the null hypothesis
4	state	Reject the null hypothesis
5	consumer_consent_provided	Accept the null hypothesis
6	submitted_via	Accept the null hypothesis
7	date_sent_to_company	Reject the null hypothesis
8	company_response_to_consumer	Reject the null hypothesis
9	timely_response	Reject the null hypothesis
10	consumer_disputed?	Reject the null hypothesis

Categorical feature selection using Chi–Square Test:

Chi-Square is best suited for categorical variables and binary targets only, and the variables should be non-negative and typically Boolean, frequencies or counts. In our data science work, we often encounter categorical features. Some people would be confused about how to handle these features, especially when we want to create prediction model where those models basically accept only number and not a category

Why did we do feature selection?

In this project we are predicting the classification of complaints posted by consumers. Feature selection can reduce the error in our prediction model giving the best accuracy result. The method through which we did feature selection is the chi-square test of independence.

The Chi-Square test of independence is used to determine if there is a significant relationship between two categorical variables. It means the chi-square test of independence is a hypothesis test with two hypotheses present: the null hypothesis and the alternative hypothesis.

Chi-square Test statistics:

We created a contingency table also called a crosstab they are used to summarize between several categorical values. Using *pd.crosstab* we gave the *product* as one categorical variable and *timely_response* as another categorical variable. Output we get is the count of occurrence in each bucket i.e., yes or no value for each category in *product* column

$$\chi^2 = \Sigma \frac{(Observed - Expected)^2}{Expected}$$

In this formula, observed is the actual count for each category and expected is the expected count based on the distribution of the population for the corresponding category.
In our case chi-squared test value came to be 1815.74 and compared it with critical value. We took 0.05 (i.e., 5%) as the significance level and degree of freedom is 10. Degrees of freedom is calculated as (no of rows-1) * (no of columns-1). Critical value is calculated by considering the degree of freedom as 10. In p-value we are getting the cumulative distribution function to cover all the area under the distribution up to our chi-squared test with degrees of freedom equal to 10. Critical value came out to 18.30 which is extremely lower than chi-squared value so it will be significant, and the p-value is 0.00 which is a small value. Since the p value is less than 0.05 and chi-squared value exceeds the critical value, we will reject the null hypothesis that the two categorical attributes are significant.

B. *Logistics Regression:*

Logistic regression models are great for analysing multinomial categorical data allowing you to do classification. It is mostly used when the target variable is categorical. In our case the target variable was the '*product*' column. Using the factorize() function each variable under product column was labelled from 1 to 7.

A logistic regression will model the chance of an outcome based on individual characteristics. Because chance is a ratio, what will be modeled is the logarithm of the chance given by:

$\log(odds)=\text{logit}(P)=\ln\left(\frac{P}{1-P}\right)=a+b_1x_1+b_2x_2+b_3x_3\ldots\ldots$

where,
p= probability that a case is of a particular category
a= constant of the equation
b=coefficient of the independent variable

C. *TF-IDF vectorizer*

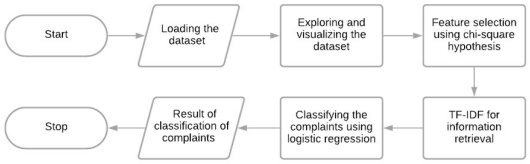
To convert the text data to vectors we used tf-idf. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents. Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is within the document. Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents[tf-idf] IDF can be calculated as follow:

$$idf_i = \log\left(\frac{n}{dfi}\right)$$

Where idf_i is the IDF score for term i, df_i is the number of documents containing term i, and n is the total number of documents.

VI. Results and Finding

The flow of our methodology was as follows –



A. *Chi-Square Test*

By performing the chi-square test [9], we got the output p-value < that 0.05 i.e., 0.0 which means that null hypothesis can be rejected. From the fig.3 we can see the distribution of chi-square with degrees of freedom as 10

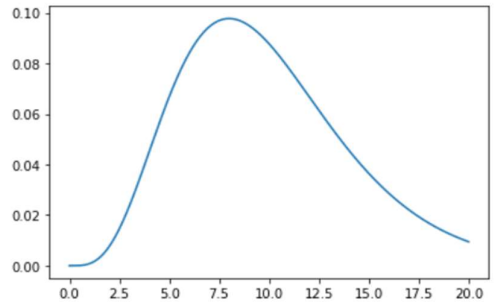


Figure 5:Chi-square distribution with 10 degrees of freedom

B. *Logistic Regression*

We used the logistics regression to analyze the relationship between predictor variable product and response variable consumer_complaint_narrative and got the accuracy of 84%.

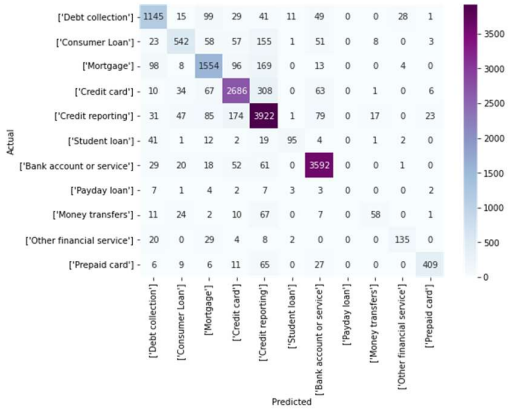


Figure 6: Actual vs. Predicted

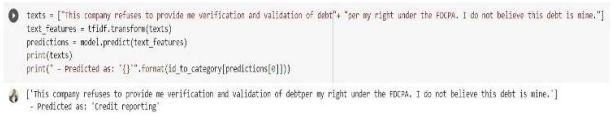


Figure 7: Model Tested

VII. Conclusion

Therefore, according to our implementation and study, the chi-square test is better for feature selection. Based on that selection we got 84% of accuracy with the logistic regression model. To conclude, the method of chi-square feature selection with logistic regression works well with classification.

VIII. References

[1] Y. D. Setiyaningrum, A. F. Herdajanti, C. Supriyanto and Muljono, "Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm," 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019, pp. 1-4, doi: 10.1109/ISEMANTIC.2019.8884290.

[2] S. Rosidin, Muljono, G. Fajar Shidik, A. Zainul Fanani, F. Al Zami and Purwanto, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data," 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), 2021, pp. 32-36, doi: 10.1109/iSemantic52711.2021.9573196.

[3] Y. Chen, J. Wang and Z. Cai, "Study on the Application of Machine Learning in Government Service: Take Consumer Protection Service as an Example," 2018 15th International Conference on Service Systems and Service Management (ICSSSM), 2018, pp. 1-5, doi: 10.1109/ICSSSM.2018.8465040.

[4] E. R. Setyaningsih and I. Listiowarni, "Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing," 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021, pp. 330-333, doi: 10.1109/EIConCIT50028.2021.9431862.

[5] "US Consumer Finance Complaints",< https://www.kaggle.com/kaggle/us-consumer-finance-complaints/>

[6] "Consumer Compliant Database" < https://www.consumerfinance.gov/data-research/consumer-complaints/>

[7] Jean Eaglesham, "Warning Shot On Financial Protection", < https://www.wsj.com/articles/SB10001424052748703507804576130370862263258?mod=googlenews_wsj>

[8] Steve Eder, Jessica Silver-Greenberg & Stacy Cowley, "Republicans Want to Side-line This Regulator. But It May Be Too Popular", < https://www.nytimes.com/2017/08/31/business/consumer-financial-protection-bureau.html>

[9] Atul Patel , "Categorical Feature selection using chi squared", 25 June 2021 , <https://www.youtube.com/watch?v=2PbfYO7fu3I&ab_channel=AtulPatel>

[10] Susan Thomas Springer," What happens when a bank hits $10 billion?", <https://independentbanker.org/2017/02/what-happens-when-a-bank-hits-10-billion//>