**Project Report**

**Video link:** [**https://www.youtube.com/watch?v=idr-MCXYwaw&feature=youtu.be**](https://www.youtube.com/watch?v=idr-MCXYwaw&feature=youtu.be)

**Business/Research Problem Definition**

We wanted to analyze the CFPB database and create a DSS which can be used by the financial institutions and CFPB itself. We have done this by categorizing all the complaints in four categories and doing a sentiment analysis to understand the severity of the complaint. The benefits include:

1. Financial institutions can use this to improve their response to the consumers and to identify the major concern areas which can be either in a specific product or in a specific region.
2. CFPB can use the analysis to identify the financial institutions which have a disproportionate share of the complaints and see of there is any trend in the complaint against a specific company or a specific product. Using this they can preemptively investigate the issue before too many consumers are affected by it.

The complaint database contains information about the product type, region etc. we can divide the complaint based on these parameters to identify the critical parameters. We have analyzed the text of each complaint to classify it in either of the four categories: Fraud, Fee, Policy, Unauthorized transactions. We have also done a sentiment analysis of the complaint to determine the intensity so that complaints with the most negative sentiment can be prioritized for resolution. The problem can be converted into an analytics problem by providing a descriptive statistic visualizing the multiple aspects of complaints and compare the same against their peers.

**Analytics Problem/Definition**

1. Exploratory Data analysis and Descriptive Analytics to identify various trends in the complaints.
2. Use Classification techniques to divide the complaint in either of the four categories: Fraud, Fee, Policy, Unauthorized transactions.
3. Average accuracy of classification to be more than 75%.
4. Conduct sentiment analysis using afinn library on the complaints to determine the intensity of complaints.

**Data:**

We have used 2 datasets:

1. Complaints data from CFPB website with all the fields i.e. Name of FI, Product, Complaint Text, Resolved Status etc.
2. 4 datasets – one for each complaint type, containing just the complaint text and if the complaint belongs to that category (Yes/No).

Both the datasets were complete on all the parameters of analysis. In DSS we have combined both the datasets based on complaint ID and provided them four tags and a sentiment score.

**Methodology Selection:**

To classify the complaint, we discussed various algorithms that can be used such as classification tree, logistic regression and random forest as these are the most popular classification method**.**

Sentiment Analysis: We have done sentiment analysis of each complaint into on a scale of -1 to -5 using the ‘nrc’ library.

We decided to use R as many packages are available which can be directly used for both classification and sentiment analysis. We tested all 3 methods listed above and decided to go with classification tree based on the accuracy.

**Model Building:**

Our first task was to classify the complaints into 4 categories: Fee, Fraud, Policy, Unauthorized. For that we randomly selected 4,352 complaints and got them tagged by different taggers. Then we divided those complaints in test and train data. First, we had to break the compliant into tokens using tidytext. Then for classification we have used the decision tree algorithm. We also tried running the random forest and logistic regression model, but the results were not accurate or taking too long. After running the model, we tested the accuracy on test data. Once we got a good result, we applied that to our complete dataset and assigned a category to each of them. One of the problems that we faced when we were using the data dictionary model to categorize the model was that, 340 out of 500 most frequent words were common for all categories. Hence, we tried methods like classification trees and random forest.

Result: Accuracy of Predictions

|  |  |  |
| --- | --- | --- |
| S No. | Category | Accuracy % |
| 1 | Fee | 82.7 |
| 2 | Fraud | 80.0 |
| 3 | Unauthorized | 78.5 |
| 4 | Policy | 62.5 |

**Functionality:**

The Decision Support System created allows the users to filter data based on date range, product and financial institutions that they want to analyze. Once the filters are selected, a heatmap is generated by the number of complaints. Next, any state is selected on the map to drill own further.

Clicking on any state shows the top 5 FIs for that state by complaints and display if the complaints belonged to fraud, fee, policy or unauthorized category. We have done a similar display based on the sentiment of the complaints. Here we will also display a word cloud which will show the main keywords associated with that company.

The packages used are dplyr, ggplot, plotly, stringi, lubridate, mapproj, maps, rsentiment, worldcloudtm, RColorBrewer etc

The data was filtered based on date, products and companies. The result was again bifurcated based on states. In each state we can categorize the complaint in four categories.

We will also show a variation in the number of complaints over the selected period.

The app can be further enhanced by customizing view for different type of users, automating alerts when number of complaints exceed a specific number, use the categorization to improve fraud analytics, use the sentiment analysis for better customer service

**GUI Design and Quality:**

Yes, the app work without errors for all the test cases that we tried.