Customer Complaint Analysis

Executive Summary

To meet the requirements mentioned in the use case and understand the consumer's point of view, Consumer Complaints Database that has complaints and other information on the issues and feedback of the products and Finance Company's response is taken. The dataset goes through various stages of data lifecycle and insights are obtained from the complaints data.

According to the website https://data.gov Consumer Complaints Database is managed by CFPB which publishes the *generated* data related to the complaints submitted by the consumers and company's response to the complaints or after 15 days of complaint raised. It was stated that the complaints are *collected* on daily basis and stored in the database. As the dataset has various products and issues related to Banking and Finance like Mortgage, Debt, and other information, only Credit Card and other Credit information is filtered from the whole dataset as part of data *processing*. Also, R and RStudio is used to achieve this along with converting text data into categorical fields (wrangling) and selecting 16 variables from the dataset of 18 variables to study the data. These filtered data is exported into Tab-Separated Value File from R and used for further analysis. Null values are not replaced in the data as it is not mandatory that a complaint should have all the fields. Also, each ComplaintID was unique with its own data, so no rows were removed/replaced.

The exported TSV files were *stored* in Google Cloud Platform (GCP) for analysis using BigQuery and Hadoop by creating a new Project. The stored data was *managed* using Cloud Storage of GCP by creating buckets for TSV Files. Moving to *analysis*, three tools – BigQuery, Hadoop Hive and Hadoop Spark was used to execute queries and obtain required results. For BigQuery, a Dataset and Table were created to access data to find Timely Responses of Company, Dispute Rate, and other aspects. For Hadoop, Dataproc API from GCP is used to create a cluster and connect to Linux Virtual Machines. Queries were executed in Hive and Spark environments after creating a table for Complaints data. Results obtained from there were evaluated to understand the efficiency of Hadoop Hive and Spark.

As the data set was *visualized*, new insights were gained related to Credit Card Data and other Credit-related products in the dataset. The graphs were plotted using Tableau with Number of Complaints over Time in multiple aspects. Also, Map was used to plot and find the highest number of complaints state-wise which could be easily understood when compared to the table results from BigQuery. With the tables and graphs, it was *interpreted* that with change in product category and increasing its scope (For ex. considering Credit Reports along with Credit Cards), there was exponential growth in the number of complaints received. This also left blank in the complaints received as people opted for new category to fill the complaints instead of existing one. Also, on an average, 47 complaints were received based on Credit cards per day, Response of the Company was directly related to the mode through with complaints were submitted (Mail, Phone, or Web).

With these information, a Credit Card company would be able to understand the requirements of consumer, changes and improve its services and provide better service to achieve customer satisfaction.