Big Data Analytics Technical Project 2021

Danfeng Wang
Big Data Analytics
Letterkeney Institute of Technology
Donegal, Ireland
L00162010@student.lyit.ie

I. PROBLEM DESCRIPTION

The dataset of this project obtained from the data.gov website is a database of consumers' complaints about financial products collected by the Consumer Financial Protection Bureau, an agency of the United States government responsible for consumer protection in the financial sector USA.

The analysis of this dataset can be used for the reference of the government to supervise and manage the financial sector by finding common problems and making precautions. By using data visualization and machine learning approach, this project is presumed to provide the regulators with give insights and understanding into consumers' perspective so that to help the authorities to improve control measures or give better regulations and decisions in the management of the financial market.

- The quality of service - responding time, where, what and which providers got the most complaints. Any improvement later? Any bigimprovement later? why?
- The motivation of complaints - top issues analysis.
- Trend - predict what is the most likely complaints in the future.

II. DATASET

The record shown in this dataset has been either confirmed by the company of the complaint or received for 15 days. The subject of complaints does not include institutions with less than 10 billion dollars in assets.

The volume of this dataset is roughly 2 million. It includes complaints dating from 2015 to the end of 2017. After giving a quick view on the dataset and investigating the domain knowledge about consumer complaints analysis, the difficulty level of this analysis is supposed to be medium based on the following reasons.

- Collected by the government which has been well organized..
- The analysis method of complaints data is relatively common. I might be easier to find the reference and the domain knowledge.
- It might brings about financial industry research work in achieving multiple dimension analysis. Some of the dimension data may require web crawler work.
- The major features of the dataset are text or date which needs more effort on multidimensional analysis or analysis with certain depth or extent.

III. GOALS

The initial goal focuses on the following points:

A. Descriptive Analysis

- The composition of the complaints
- The complaints reason rank.
- Time series analysis, such as year-on-year, month-on-month comparison.
- Complains distribution by multiple dimension.
- In-depth analysis if any outlier is found in the above.

B. Predictive Analysis

- Complaint trends prediction.
- Identify the problems within the complaints which tends to cause more complaints.

IV. DATA ANALYTIC PIPELINE OF PROPOSED SOLUTION

The analysis will be done by using Python notebook on Google Colab. The process of the analysis is listed as follows:

- data understanding - get domain knowledge by doing research according to the data.
- Data preparation – download and unzip data from data.gov . Check data integrity and outliers, process Null cells, unuseful rows or columns which have less relevant with the main purpose of this analytical work.Data type and calculated fields process and so forth.
- Data visualization – Implement the descriptive analysis.
- Data prediction – data modeling with multiple algorithms and optimization and evaluation of the models. The complaints prediction is a classification problem.