Project Proposal

CSCI 4502 Project

Scott Young Michael Gilroy Cory Morales

Michael Min

Problem Statement/Motivation:

For this project, we are doing financial complaint analysis using the Consumer Financial Protection Bureau's publically available dataset. Founded in 2010, the Consumer Financial Protection Bureau has been a resource for consumers who feel they have been treated unfairly by a company's financial product or service. Our goal is to find trends in these complaints, locate potential flaws in this complaint-compensation system, and evaluate effectiveness of this system as a whole.

Literature Survey:

The previous work that has been done on the Consumer Financial Protection Bureau dataset can really be boiled down into two categories: visualization and statistical analysis. Some financial analysts at *Deloitte* put together a publically available document giving some insight as to things to be taken from this particular dataset, such as trends in types of mortgage complaints over time. *Deloitte* includes an abundance of visual aids to support their claims- histograms, line graphs, and stacked bar charts are all included in their analysis of the CFPB dataset. They make claims based on demographics, such as how age and income have direct correlation with the number of mortgage complaints. They also acknowledge that this dataset is comprised of only *reported* complaints, and that a healthy amount of skepticism should be exercised when evaluating the semantics of the data.

The organization uspirgedfund.org also did an analysis on the CFPB dataset, and has similar tone to the Deloitte analysis. Similar to Deloitte, uspirgedfund.org agrees that mortgages are a pivotal part of this financial complaint dataset. They divided these mortgage complaints into two large categories: customers who were unable to pay, and those who had issues paying. This kind of analysis is interesting because it details the percentage of complaints where the company is at fault.

Proposed Work:

The dataset seems to require data cleaning in order to resolve the many inconsistencies that do not appear to have conclusive solutions. Under the "Consumer disputed?" attribute, which is a binary attribute that only accepts "Yes" and "No" as values, there are many entries that have no value. In addition, there does not appear to be any way to infer the data for these missing cells, suggesting that we must either disregard the missing data or create an "Unknown" class for these cells despite the increased ambiguity of doing so. Several other attributes also have many entries that are empty; in the case of more meaningful attributes like "Company Public Response", the empty cells may prove detrimental to mining for answers, limiting what questions we will be able to ask about the data.

Another focus of work may be on data reduction, which may help with eliminating irrelevant data to make the data easier to mine. Under the "Issue" attribute, which will likely be the most important attribute in our analysis, there are entries that contain ambiguous information. Though these complaints may have some meaning in determining the variety of complaints companies have against them, they may also prove to be negligible because of how they do not mention an explicit offense. Thus, if the dataset proves too large to mine effectively, these vague entries may have to be reduced for the sake of efficiency.

For processing derived data and evaluating, we plan to search for patterns in the complaints, the companies, and the types of products/sub-products that are being offered. The “Issues” attribute appears to be nominal and reuses its data classes (i.e. “Billing disputes” and “Disclosure verification of debt”). This means that we may be able to segregate the data into clusters based on how similar issues are to each other in order to discern outliers, or issues that are particularly rare or serious. This method of detecting noteworthy complaints may prove useful in searching for more controversial companies. Additionally, we may also be able to link the data we find in the “Issues” attribute to the “Product”, “Sub-product”, and “Company” attributes in order to determine which company and product combinations tend to lead to the most conflicts.

Our work differs from the previous studies done with the dataset in that we plan to focus on how individual companies appear to handle consumer disputes. In the case of the Deloitte analysis of the dataset, we intend to differentiate our work by focusing instead on the types of offenses that specific companies appear to be prone to. To differentiate our work from the USPIRG article on mortgage complaints, we will attempt to mine information and conclusions from the database that are less apparent. Some examples include searching for which companies appear to be attempting to atone for their wrongdoings and which companies appear to be more ethical despite the complaints against them.

Data set:

Our dataset is the database of consumer complaints as provided by the Consumer Financial Protection Bureau. It contains complaint submissions dating from December 1st, 2011. On the date of download (February 21st, 2017), the dataset had entries up to February 20th, 2017. The dataset has a total of 720,000 entries and 18 attributes. Information like the text of the consumer complaint is available in an entry when release was allowed. Other information like the nature of the complaint, its resolution, what company the complaint was against and other relevant things are also included.

Evaluation:

The evaluation of the data will be composed of several things. We will include several graphs in order to create a visual aid to help people understand where the companies are at and what trends exist in complaints. We will use these graphs to see if any complaints arose when new policies were enacted, as well as seeing which companies are more shady in their practices. We can use this data to determine if any companies are adjusting their policies to accommodate new laws or if they are not.

Tools:

We will be using python as a basic framework, using the matplotlib, numpy, and pandas libraries to detect patterns in the data. This will get us some idea on what we could look for when we transition to WEKA. WEKA will grant us a more accurate and in depth look at the data. We will be able to get more succinct and accurate look at our data and can reach better conclusions

Milestones:

Data Cleaning:

Cleaning and removal of unwanted data. Convert strings to a standardized practice.

Initial Data-processing:

Identify rough patterns to later investigate further.

Pattern Finding:

Investigate the initial rough patterns to identify precise patterns in the data.

More datasets:

Find other datasets that pertain to the patterns we have found.

Additional Cleaning:

Cleaning and removal of unwanted data from the additional datasets. Convert any strings to a standardized practice.

Final data processing:

Use the additional datasets to determine the causes of the patterns found, and draw conclusions pertaining to the business practices of the companies.

Summary of peer review session:

We were informed that we did not initially phrase out question properly, as the question was rather vague. We have since modified our phrasing of our question to make it clear.