Project Progress Report

CSCI 4502 Project

Scott Young Michael Gilroy Corey 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. More specifically, we hope to find important patterns involving the locations where complaints took place (as in which U.S. states appear to be prone to particular types of complaints and controversial companies) so that we can gain a wide scale view of the origins of consumer issues. In the end, we feel that the most optimal outcome that we can get from this project is discovering which companies the average consumer should either avoid or consider acceptable, as many of these companies partake in important financial business such as loans, credit, and mortgages. Thus, in addition to analyzing our selected dataset for educational purposes, we also are motivated to take a practical approach to this product in that our findings may prove useful in our adult lives.

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.

A greater focus of prospective work involves creating an algorithm to quantify attributes that contain free, verbal information. Within the data there is an optional attribute known as “Consumer Complaint Narrative” that contains the consumer’s actual comments about their perceived mistreatment. We consider this attribute to be of importance because of the unique insights they provide in their respective entries, yet we have currently been unable to interpret the consumer complaint narrative of thousands of entries in an efficient manner. We have already gathered some initial training data that quantifies each entry’s narrative as positive, neutral, or negative depending on the company’s perceived hostility. This same training data will act as a model to rate the consumer narratives of other entries based on similarities in vocabulary, strong punctuation, and length. In the future, we intend to use this training data to create several algorithms that will determine the intensity of a company’s misconduct via the consumer’s narrative in relation to others. In addition, however, we also plan to combine the analysis of derived narrative data with other related attributes (such as issue, company, company response to consumer, etc.) in order to gain a greater understanding of how specific companies tend to resolve their disputes with consumers.

To find correlations between nominal attributes (company, states, issue, product), we wish to develop a more comprehensive means of representing correlation coefficients and Chi-Square values. While we currently have a reliable program that gathers these two different types of correlations, we do not have a way to highlight exceptional correlations and determine how significant other correlations are to each other. We find that determining correlations between certain elements in nominal attributes (such as the correlation between issue and company) to be important because they can help to determine if particular companies are prone to repeatedly causing the same types of misdeeds or causing issues in particular states. This information, along with other combinations of relationships, may act as proof for claims we make about the integrity of certain companies, which in turn answers questions that we have about our data.

Naturally, the prospect of simply interpreting and representing our data in more ways remains an important area. As we create code that analyzes the data set in different ways (such as determining correlation values and separating data into bins for visual representation), we constantly ask ourselves if there is another method that can yield alternate information or a type of graph that we have not derived from the data thus far. One such example includes incorporating other clustering methods. Though we have already invested resources in creating code for k-means clustering, we have also considered the possibility of including some type of density-based clustering to determine if it would reveal different trends in the data. We currently lack a specific procedure to implement density-based clustering, but we may still be able to translate some of our work done on the k-means method to this new method. Thus, we believe that implementing some density methods would be possible, yet we would need to ensure that other in-progress sub-projects are secure first.

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 shadier 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.

One of the most fundamental forms of evaluation on our project is the use of histograms to organize the entirety of the data set into a set of bins; the set of bins depends on which of the available attributes is being observed. Through this, we plan to provide a wide range of graphs, such as histograms and pie charts, that will cover a variety of attributes that the data can be separated by. Additionally, for attributes that are more numerical or abstract in nature (such as date), we may still be able to evaluate the data into a histogram by using bins that accommodate a range of values instead of accommodating only one specific and distinct value each (as would be the case in nominal attributes). Ultimately, this type of evaluation may be helpful to readers by helping them understand the general structure of the data, but the evaluation may also be helpful to us by serving as a more basic foundation for our more complex operations.

Another example of a visual data representation, or chart, will include tables that hold the correlation coefficients and possibly the Chi-Square values between strongly positively and strongly negatively correlated values in nominal attributes. This table will help to determine how closely correlated the values in one attribute are to the values in another. For example, if we were to ask the question of which company has the highest positive correlation with the issue of “billing disputes” as well as how high that correlation is compared to that of the other companies, we could create a program to produce a table that shows the highest positively correlated company and its correlation ratio along with the correlation ratios of other less positively correlated companies. This method of displaying to readers how intensely specific companies relate to an issue may show to readers how severely unethical a company may be compared to its peers..For a broader view, we also intend to create a heat map of the United States that will be color coded to help show which states contain a high number of grievances or a high occurrence rate of a certain issue. We feel that this map could help show which states may prove to more stringent or lenient when it comes to financial disputes. It could also show how some states with fewer consumer issues may have weak spots when it comes to specific issues, mainly because of how higher levels of a specific kind of issue (especially in comparison to other issues in a state) might indicate that a particular state’s laws need to be reformatted in order to make that issue less prevalent. In conjunction to our other question of determining which companies are the most troublesome, we may also use this heat map to provide further evidence of how a particularly infamous company is residing (perhaps purposefully) in states with higher occurrences of consumer issues. In the end, while this heat map may serve as a representation to a reader of the most and least tenuous states, it can also assist us in investigating deeper issues, such as paying closer attention to seemingly innocuous companies that appear to reside in states with high levels of incidents.

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.

In addition, we also will incorporate some use of HTML5 and the Python library urllib2 in order to create some types of graphs, which mainly include the heat maps of the United States.

Milestones:

What We Have Achieved So Far:

Data Cleaning:

We have finished removing entries that lack values in mandatory attributes (issue, state, company, etc.) as well as placing a constant null value in blank cells under optional attributes that are not completely necessary to evaluating an entry.

Data sampling:

Through a significant amount of effort, we have succeeded in creating an effective means of sampling data for various reasons. One of these reasons is to decrease the time needed for testing new programs with the dataset for more technical problems (like bug fixing). Also, our system of sampling data has provided us with a sample of entries that hold nonempty values for attributes with freely verbal values. Thus, we are able to use this set of samples as training data for the sake of including the freely verbal attribute of the dataset in our analysis of the perceived ethics of particular companies.

Pattern Finding:

We have made progress in finding more complicated patterns, such as lift correlation, Chi-Square correlation, and k-means clustering.

Additional Cleaning:

We have made efforts to transform the values in nominal attributes that use verbal values to a numerical form. We suspect that these changes will make the data run with a wider range of algorithms, which hopefully will yield less obvious information

What Remains So Far:

Initial Data-processing:

Though we already have a general idea of the structure of the data set, we hope to create some basic graphs (that use bins to separate the data) and charts that give us a visual direction of how the data is distributed and the nature of some basic trends. We will build off these findings later on when we want to answer highly specific questions.

Additional means of analysis:

While we have analyzed the data set in several different ways, we have still wondered if we could spend more time either refining our current data analysis method or programming new ways to view the data (such as different types of clustering). When it comes to refining our current methods, we feel that we could make them more precise or generate more useful material; in the case of our programs that generate different types of correlation ratios, we largely feel that we could generate the ratios in a more meaningful way, rather than simply producing correlation values with little context. In the case of finding new ways to interpret the data, we have wanted to find some way to effectively incorporate clustering other than k-means clustering into our project, as we feel it could be useful in gathering the data of the issues and companies into informative categories. However, it may not be feasible to do so in a competent manner before the deadline. Thus, incorporating these new types of clustering would be a significant milestone, yet we are not willing to compromise the quality of the other types of analysis currently in progress to do so.

Producing Impactful Questions:

Though we do have some idea of what questions we want to ask about the data, we do not currently have any highly specific questions pertaining to specific attributes. Among the many questions we could ask, we will have to begin formulating questions about noteworthy companies and any issues they correlate with. After that, we will be able to produce a highly detailed report that contains information about lesser known consumer complaint trends in states as well as outstanding information about a select number of individual companies.

Inclusion of different visual representations:

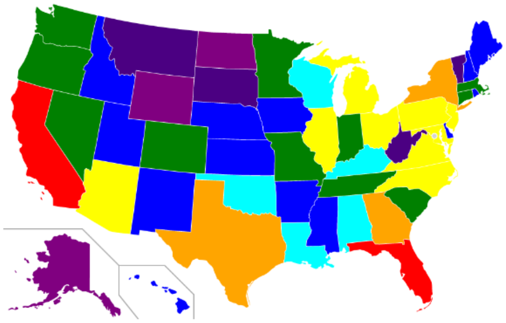
In addition to the visual graphs we are currently working on or have completed, such as bar graphs, plots, and the heat map, we also hope to include even more of these visual representations to give better insights on our data. As an example, our current plans for bar graphs include subjects such as separating into bins by their entry type or company. However, we may also provide representations using the dates of each entry. This representation, also in the form of a bar graph, may then divide data into bins that represent a slice of time. Thus, with this particular representation and many others, we believe that a collection of visual representations that contain subtler information will be a worthy endeavor for our project.

Final Data processing:

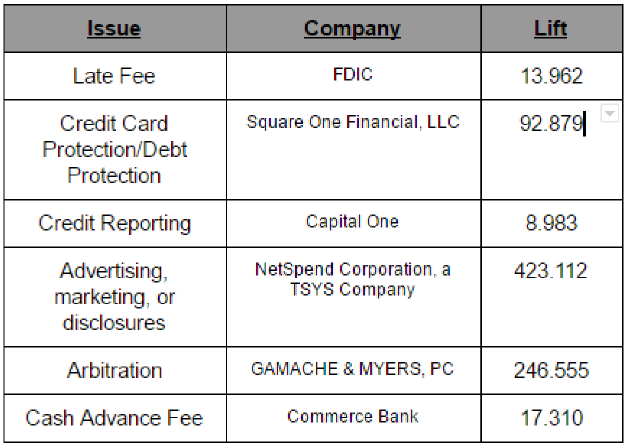
After gathering a sufficient amount of data, we must then interpret all of it and effectively describe it in either visual or verbal form. While we do have some graphs and visual representations well in progress, more work must still be done if we wish to describe the nature of other collected data and trends that are not currently analyzed.

Results:

Currently, we are in a phase of the project that involves two separate tasks: implementing additional algorithms to provide a greater range of information on the data and expressing the data and correlations we have currently gathered with a series of graphs. We have acquired a map that color codes ranges of data on the map of the United States. Each color represents a different amount of reported complaints. Purple represents the fewest amount of complaints from 0-1000. 1000-2000 complaints is represented in Indigo, 2000-4000 complaints is represented by Blue, 4000-8000 is represented by Cyan, 8000-16000 is represented by Green, 16000-32000 is represented by Yellow, 32000-64000 is represented by Orange, and more than 64000 complaints is represented by Red.

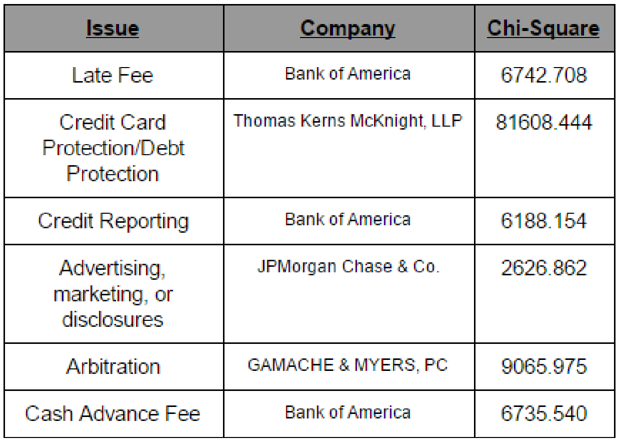


Some other concrete results at this current stage include some preliminary information on both lift value and Chi-Square correlations for the data set's nominal attributes. Provided below as an example is a table of lift values between a small sample of individual issues and the company that they are correlated the highest with.



Issue	Company	Lift
Late Fee	FDIC	13.962
Credit Card Protection/Debt Protection	Square One Financial, LLC	92.879
Credit Reporting	Capital One	8.983
Advertising, marketing, or disclosures	NetSpend Corporation, a TSYS Company	423.112
Arbitration	GAMACHE & MYERS, PC	246.555
Cash Advance Fee	Commerce Bank	17.310

In addition, the same is done in the table provided below, but using Chi-Square values instead.



Issue	Company	Chi-Square
Late Fee	Bank of America	6742.708
Credit Card Protection/Debt Protection	Thomas Kerns McKnight, LLP	81608.444
Credit Reporting	Bank of America	6188.154
Advertising, marketing, or disclosures	JPMorgan Chase & Co.	2626.862
Arbitration	GAMACHE & MYERS, PC	9065.975
Cash Advance Fee	Bank of America	6735.540

Thus, with these basic tables, one can determine the types of issues that the displayed companies tend to cause the most.

References:

http://www.uspirgedfund.org/news/usp/new-report-mortgage-problems-rank-1-cfpb-consumer-complaints

https://www2.deloitte.com/content/dam/Deloitte/se/Documents/financial-services/CFPBConsumerComplaintDatabase091913US_FSI_.pdf

https://www.consumerfinance.gov/data-research/consumer-complants/