**Magic Mirror: Trend Analysis of Company Public Image**

**Hummingbird Co:**

**Larry Glerum: Analyst**

**Tyler Violillo: Team Lead, R Programmer, Analyst**

**Richie Walker: Analyst**

**Northwestern University Predict 498 SEC56, SP2018**

**Faculty Advisor/Sponsor: Nethra Sambamoorthi, PhD**

[**Team Roles and Responsibilities**](#_kepbrkphx8jt) **2**

[**Abstract**](#_tso9h2oulylw) **3**

[**Organization and Background**](#_yujzzqk5vhxf) **4**

[**Preliminary Research and Business Opportunity**](#_md3xsyxxorew) **6**

[**Data and Measures**](#_q227kidf0v6z) **7**

[**Data QA**](#_41khshwei14k) **9**

[Consumer Complaint Database](#_oukp0gfr3ex7) 9

[Evaluation and Exploratory Data Analysis](#_fqm0z263o1ek) 12

[Twitter Data](#_dlij6xmvmkx0) 23

[Consumer Sentiment Data](#_idln6nd4cwo) 25

[**Research Questions and Performance Measurements**](#_oasdusyjtoms) **26**

[KPIs](#_dac352ja8wz9) 26

[Hypotheses](#_jprkevc88hs0) 27

[**Feature Handling**](#_y2o2l85t88od) **29**

[Variable Reduction/Selection and Ranking](#_24ii7nxwe2l0) 29

[**Tools and Techniques**](#_qxt19e6juih2) **30**

[**Methodology**](#_pzkwd9enm90x) **32**

[Computational Methods](#_uas0b6w9keaf) 33

[Outputs](#_etvqzim220g9) 34

[**Dashboard Components**](#_aevi58tsfuou) **39**

[Classification Model](#_u0nfnvs06yp7) 39

[Twitter Score](#_7jpxs69743c3) 44

[Cluster Analysis](#_tnsj241hj30p) 45

[Time Series](#_ovn3wt873ax) 46

[**Project Challenges**](#_2p3pqezb5pyd) **49**

[**Opportunities**](#_k2ymqonx2bf8) **50**

[**Final Conclusion**](#_gf4s2dvn1zyl) **51**

[**Sources**](#_r2lwltw0qqxq) **52**

[**Appendix**](#_1tz2fi79ff8n) **54**

[**SDaIR and Project Plan**](#_4dpga7j36rrt) **71**

# **Team Roles and Responsibilities**

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Role** | **Responsibility** |
| Larry Glerum | Analyst | EDA, analyze data, contribute project artifacts, implementation |
| Tyler Violillo | Team Lead, R Programmer, Analyst | Ensure completion of project, clean/transform data, create models, analyze data, contribute project artifacts |
| Richie Walker | Analyst | Project Plan, analyze data, contribute project artifacts |

# 

# 

# **Abstract**

The world is getting smaller, and a positive public perception is becoming a virtue few companies can afford to destroy. Companies like Facebook, in the face of their data breach or United Airlines, affected by the broadcast video of the incident when United staff dragged a passenger off a plane, offer insight as to the importance that corporations, like financial institutions have concerning a solid plan in which to minimize or manage negative perceptions that can drive down consumer interest which can impact the bottom line. With many events happening every day under the radar, by way of customer complaints and client response surveys, companies have a huge task of optimizing their handling of customer complaints with positive marketing and strategic handling of individual complaints.

Magic Mirror is a comprehensive tool that joins consumer confidence numbers, Social Media analysis, and interpretations of analytical trends found in a large public consumer complaint dataset to create personalized dashboards. This tool should describe the information necessary to help our clientele to mobilize quicker to protect themselves from any event that would result in negative publicity or high levels of consumer complaints. A part of the implementation that will bring great value is the availability of exporting our tools to our clientele’s mobile phones and tablets.

**Introduction**

A company’s public image can be foundational to a company’s success. One way to measure how the public views a company is to look at the news. McDonalds had their hands full with hot coffee complaints and had to work to quell the debate over how hot their coffee is (McDonalds won in court, but people still are questioning their quality). The other way is to gain greater understanding toward the customer complaints against them. Companies pay thousands to learn how to deal with trends in consumer complaints. Some complaints are out of the company's hands, such as a customer having trouble with their website due to a problem with their own internet service provider, or say the loss of a job. Most other complaints about a company are valid and will hurt the business goals that a company has, and can alter resources meant for development and growth when there is a poor plan in place to confront such a phenomenon. For example, if a company has consistently poor handling of customer service calls, word will spread and the company’s image will be damaged.

Our company aims to offer solutions by using identifiable trends in customer complaints that can used to remake the way companies handle such complaints. We will focus on things such as most common complaints, regions with exceptionally high complaint numbers, and how public opinion of certain companies has changed over time. With this information, we hope to provide our customers with an actionable item that will allow them to fix these problems within their company and improve their overall customer service and public image.

# **Organization and Background**

Hummingbird Co was founded in Cleveland, OH in April of 2018. Made up of 3 members: Larry Glerum, Tyler Violillo, and Richie Walker. The company breakdown involves the three maintain co-leadership but roles are distinct. Larry is the Business Implementation analyst, Tyler is the Technical Programming analyst, and Richie is our Project manager. Each brings a dynamic that is crucial to the success of the business and the product lines Hummingbird Co. is preparing to bring to market. Our goal is to develop the best customer-focused products for the financial industry. Our vision is to design products that will impact the world. Consumers have concerns about their financial institutions, and we would like to think that we could help foster better relations between the two, impacting the world. This goal and our values are demonstrated by our mission statement:

*“Our solutions allow you to improve your company and, in doing so, better the world”*

We believe that companies will benefit from Magic Mirror, and consumers will benefit from the use of Magic Mirror. We believe that knowing ahead of time that a problem is coming is far advantageous for determining a correct response in a timely manner. For example, if one of our customers is getting complaints surrounding availability, we hope to provide an analytical picture that shows them how their current behavior can be harmful so they can take action and fix it before it becomes a crisis.

Our product is simple: a periodic report of how your company is performing based on customer complaints and public opinion trends. We went with the name “Magic Mirror” as a play on the fairytale *Snow White*. In the story, the Magic Mirror told the witch the truth, even if it was not what she wanted to see and hear. That is what we hope to do with our product. It isn’t always easy hearing your own faults, but sometimes that is the only way to fix them and grow as a company.

We will initially be offering it to financial institutions with the possibility of expansion to other markets down the road. We are flexible with our deliverables and will offer different plans to meet each company's individual needs. Our main service with a monthly (12) plan of $250 per month ($3,000 yearly subscription) we will also offer a Quarterly (4) comprehensive report at $400 per quarter ($1,600 yearly subscription), and finally a Yearly (1) report for $1000.

As a startup company, we believe we have a large potential for growth due to our solid business plan, aggressive research, cutting-edge tools, and the most dedicated team in the industry. Our use of complaint data is key to giving our customers the most up to date information.

We believe analytics can really help us paint a picture for companies to act on. Showing trends and analyses regarding complaints will allow for targeted projects to fix the issues. We believe we will have plenty of data from multiple sources to work with. Since we will be providing insights and collecting data on many different companies, we will be able to show our customers how they stack up against others in their specific industry as opposed to just an in-house project designed to deal with customer complaints.

# **Preliminary Research and Business Opportunity**

A company's primary concern is always going to be retaining and growing their customer base and a key to that is keeping those customers happy. Times have changed and it is also important to understand public sentiment, expressed through social media. One “viral” video on social media can assault a company’s image. With all the avenues for negative feedback, it is very difficult to do everything perfectly all the time. Time and resource constraints can prevent our customers from addressing every customer complaint correctly every time. We are developing a tool that will provide them with real-time figures, trends, and a predictive forecast, learning the trends of tomorrow concerning customer complaints against them and the rest of the financial industry.

We understand that companies may need us to focus on a specific need. For example, if we see their call center is causing problems but they outsource their call center to a place chosen by their parent company, they may not be able to do much there. Instead, we can focus on target areas they can improve upon. Magic Mirror will be able to identify different KPIs and, once we gather enough data, we will be able to tailor a dashboard with the most pertinent information. Selecting the right KPIs is going to be key for us from a growth standpoint because we want to provide our customer with something they find value in.

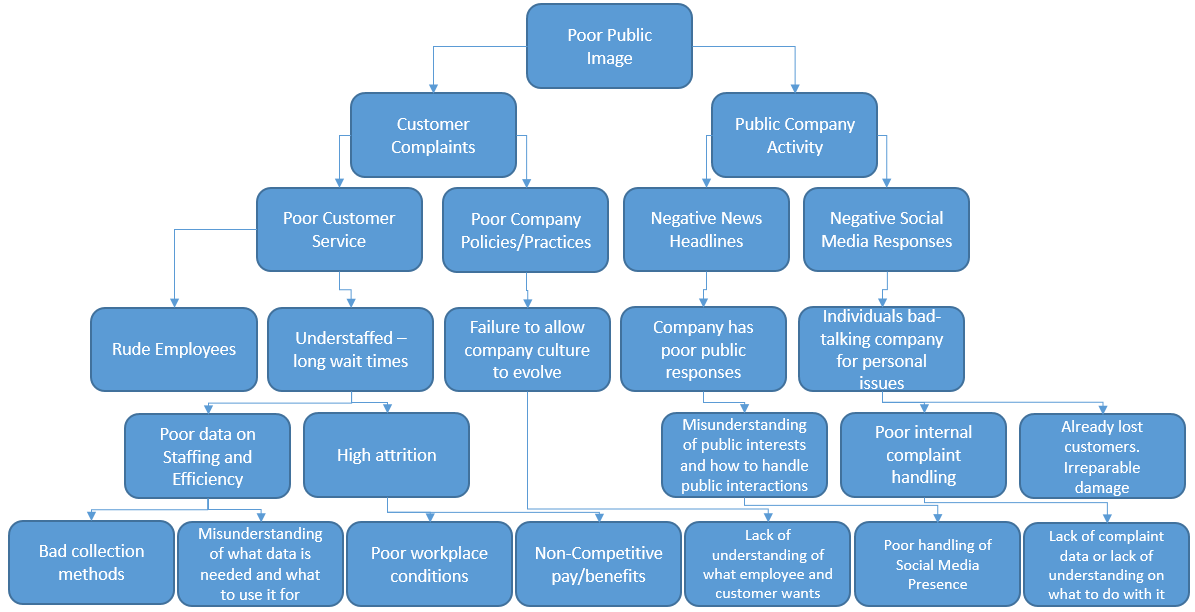
Initially, we are focusing on a few different generalized groups of KPIs: location based trends in complaints, time-series based analysis with forecasting, and a predictive value noting the likelihood of a customer concern becoming a dispute. The key for companies is that they take care of the customer concern the first time. Determining the likelihood of a dispute is a very important KPI that will help a company determine quickly if they need to act to change the trend. Location based KPIs are going to show things such as the states/branches with the highest complaint numbers for a company/ industry. This allows for improved targeted remediation with the customer. Time based KPIs will show if the number of complaints or the public image of a company is trending up or down. We can also look at seasonality data such as the time of year/month complaints happen most. Finally, the category based KPIs will allow us to tell our customers exactly what they are doing poorly and what kind of public image they have. We can show them the most common words and phrases used when people are talking about them on Social Media as well as the most common complaints they have against them.

Using these different measures will allow our customers to make anything from drastic overhauls to small tweaks of their processes that could result in major public relation gains.

# **Data and Measures**

The main source of data we are going to use initially is a public database managed by the Federal Bureau of Financial Consumer Protection (BFCP) which consists of records of consumer complaints against the financial industry that they willingly submit via a survey. We also located the web page that seems to be the source of this dataset. USA.gov is a website offering surveys to consumers about complaints of various industries, a link on the site points people to a survey concerning financial entities. We also are pursuing Social Media data such as Twitter and Facebook. We hope to use these sources in conjunction with each other to provide a broad understanding of public perception of our customers. For visualization, we also are implementing the FRED Consumer Sentiment Index found at: <https://fred.stlouisfed.org/series/UMCSENT/> , which is an index created from interest rate movements, National GDP, and unemployment numbers. These are good indicators of the overall attitude of U.S. Consumers of their financial institutions.

Our plan is to take these data and analyze it for trends such as frequency of certain words, the time of year/month most complaints are issued, etc. so we can provide a dashboard to the customer that is easy to understand. They can then set KPIs they would like us to monitor for them depending on their initial performance. We will provide updates to them regarding performance on these KPIs as well as their position relative to the industry. The Issue Tree below visualizes the beginning stages of this analysis. We know that for each company, that this issue tree will change to allow them to see where they individually are in tackling this issue:



The Issue Tree shows that a lot of the potential issues can be traced back to data - whether it be poor collection, lack of data altogether, or just a lack of understanding on what to do with the data they have. Some of these end nodes, however, are beyond what we can offer a company. We cannot help with rude employees, workplace conditions, or what they pay their employees. However, we can point out to them the issues farther up the tree that ultimately lead to these issues and contribute to a poor public image. Other end nodes are directly within our control. Any area having to do with a lack of understanding on the company’s part or a lack of data, is an opportunity for Magic Mirror to step in and provide and easy to understand public image report as well as suggestions on how to improve that image.

On top of initial KPIs, our customers will be able to tailor our product to suit their own needs. For example, looking at the Issue Tree, if a customer was primarily concerned with those factors under “Customer Complaints” but wanted to leave the rest out of the report, we could personalize the report to their liking. Essentially we could go down whatever path of the issue tree is most important to each customer while making recommendations on how to make the entire tree smaller, and easier to control.

# **Data QA**

## **Consumer Complaint Database**

**Statistician’s 10 Steps for Measuring Data**

1. We secured the data in .csv format from Data.gov in which you could upload it in a variety of formats .CSV, JSON, and XML. We are most familiar with CSV and thus our team chose to work with this in Excel to be later uploaded into R.
2. Once we uploaded our dataset in R, we used the summary command to learn about the values in each column. We then took a look at the structure of the data using the str() command ([Figure 1](#e6klbkeuaehn)). Finally, we previewed the top and bottom of the dataset ([Figure 2](#eynp1ba0guql) shows a sample of rows) to find that the data is not organized by any particular variable.
3. We confirmed the number of rows in this dataset is 1,011,100 rows or values to contend with. This is a live dataset which means that if we were to upload it again next week, the number of records would be increased. A sample this large allows us a greater understanding of the total population, however, the problem we have with this data set is that there may be a large relational differences in the variables and we have noticed large numbers of missing data. This is likely due to customer not fully responding to optional sections of the complaint survey, which is not the fault of the data provider.
4. The numeric elements consist of the Complaint ID (not used in this analysis), Zip Code (some values are blocked with x’s) , and two date variables. The index is numeric but the rest of the values we will be analyzing are all factors. A rest of the data are string variables, and are factors, as well.
5. We do not have numeric values in our dataset, but we have identified and verified the beginning and ending dates (12/01/2011 – 4/13/2018) of the two columns of Date.received, and Date.sent.to.company. In the next section we provided the [data dictionary](#emon95s5wmms) for the Consumer Complaint Database including the number of NA values for each variable.
6. We found no consistency issues with this data in terms of variables presenting contradicting information. Some of the data has stopped being collected after certain dates, however, and this is documented in the ‘Missing Data’ section of this list. This data set appears to be an accumulation of time series data about customer complaints/concerns.
7. We looked at the distributions on the main drivers of this information sample in the appendix ([Figure 3](#2dc80txda2a)): the distribution of companies (over 4000), the products and subproducts (the graph is so complex the individual relationships are blurred), and issues with customer quotations for some of the entries we viewed. This dataset is very complex, but we will work to narrow the focus to analyze it.
8. There is no need for weighting these variables. As this is a classification problem, we are going to look into cluster cores and possibly make some changes to create the best model.
9. A subset of this data will be used to create a classification model to predict whether or not a customer disputed the complaint response (Customer Disputed variable). This is to analyze the financial institutions’ ability to solve the issues, and achieve higher results of satisfied customers. We will be looking at the data as a whole, as well as on a per company basis.
10. Missing values have been identified within the dataset and will have a large effect on our analysis. Given that the original dataset contains only eighteen variables, it is the intention of the project team to mitigate potential issues around missing data and derive other “dummy” variables from the available data. Below we highlight issues found for variables containing missing data and potential mitigation:

* **Sub-Product**: Data for this variable is missing in 23% of observations. Missing values appear to be an issue up to April 2017 at which point a large proportion of the Sub-Product variable began being tagged as “Credit Reporting”. From May 2017 forward no missing values appear.
* **Sub-Issue**: Data for this variable is missing in 49% of observations. Consistently missing in observations over time. The project team believes there is potential value in this variable and that “Sub-Issue” may not be applicable to all observations. Further analysis is being conducted.
* **Consumer Complaint Narrative:** Data for this variable is missing in 73% of observations. The project team is looking into the creation of a dummy variable to identify which observations do and do not contain input into this variable. The project team believes that a dummy variable may be of benefit in potential models. On top of a dummy variable, the project team is looking at analyzing the text given in this variable. The plot in [Figure 4](#7soo4qsu9rfw) shows the count of issues as well as whether or not they included a narrative. We believe that these two variables in conjunction will be helpful in the modeling process.
* **Company Public Response:** Data for this variable is missing in 69% of observations. The project team is looking into the creation of a dummy variable to identify which observations do and do not contain input into this variable. The project team believes that a dummy variable may be of benefit in potential models.
* **State:** Data for this variable is missing in 2% of observations. Missing values for this variable directly correspond with missing values in the Zip variable. The project team is looking to impute the missing values. The assumption that a portion of the missing values may be the result of complaints coming from location where State and Zip or not applicable. Given the relatively small number of missing variables, tagging these as “other” or “not reported” may be an option. Special consideration will be given to how any mitigation might impact our effort relating to the display of information in maps.
* **Zip**: Data for this variable is missing in 2% of observations. Missing values for this variable directly correspond with missing values in the State variable. The project team is looking to impute the missing values. The assumption that a portion of the missing values may be the result of complaints coming from location where State and Zip or not applicable. Given the relatively small number of missing variables, tagging these as “other” or “not reported” may be an option. Special consideration will be given to how any mitigation might impact our effort relating to the display of information in maps. Additional investigation is being conducted regarding the presence of zip codes that are formatted in such a way that indicates potentially false information.
* **Tags**: Data for this variable is missing in 86% of observations. The project team is looking into the creation of a dummy variable to identify which observations do and do not contain input into this variable.
* **Consumer Consent Provided**: Data for this variable is missing in 52% of observations. This variable appears to have been introduced in 2015. The project team is looking into the creation of a dummy variable to identify which observations do and do not contain input into this variable.
* **Company Response to Consumer**: Data for this variable is missing in only three observations. The project team will look to see if imputing or removing from data is necessary.
* **Consumer Disputed**: Data for this variable is missing in 24% of observations. The CFPB discontinued the consumer dispute option on April 24, 2017. The project team believes this variable can be valuable in modeling efforts to determine factors that influence consumer satisfaction. The project team is exploring potential mitigation solutions.

## **Evaluation and Exploratory Data Analysis**

The focus of our EDA (Exploratory Data Analysis) was to get us closer to understanding our data, especially the variables so that we will be able to rightly analyze the dataset and with confidence, create a model that can be employed for our product. It is important to understand the consumers that filled out the survey to create the dataset. We found some news stories that illustrate where they are at. CBS Moneywatch had a news story recently about what is keeping American’s up all night. Also in 2017, an article in Forbes discussed the financial worries of millennials, and way back in 2014, a USA Today story headlined the problems Americans are saying they are having with money. Is money a concern to Americans? Yes, and also to the American government. Through the Federal Bureau of Consumer Financial Protection (BCFP), the government collects the complaints from Americans about financial institutions they say they have umbrage with. Another interesting note is this dataset is living, meaning they are still collecting this data today and into the foreseeable future. This dataset also offers us, because of the number of rows, a good sample for us to use to find the algorithmic model in which to determine the probability of a consumer dispute from each institution’s attempt to settle the issue the first time there is contact. If we can help the financial industries improve their service, it will help people feel better about their money, making the world a better place.

**So, from these articles what are financial reasons Americans are stressed out?**

**1. Growing debt from credit cards and loans, including student loans**

**2. No retirement/ Savings**

**3. Inability to make ends meet.**

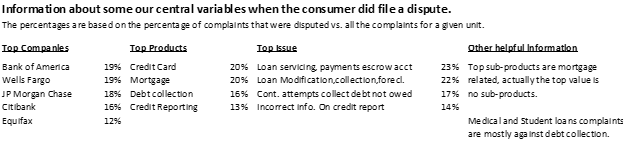
The CBS story did illustrate that health care costs are moving up fast to enter this list as a top concern. Our EDA is going to focus on who are the companies, products, and issues are that could help define variables in our model. We will also work to understand how those variables relate to our dependent variable consumer.disputed. We start by connecting with our CCD Data Dictionary for variable understanding.

**CCD** **Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field name** | **Data type**  **(levels)** | **Description** | **Missing #**  **(# of Total)** |
| ***Date Received*** | Date | The date the CFPB received the complaint | 0 (0%) |
| ***Product*** | Factor (18) | The type of product the consumer identified in the complaint | 0 (0%) |
| ***Sub-Product*** | Factor (76) | The type of sub-product the consumer identified in the complaint | 235,170 (23%) |
| ***Issue*** | Factor (166) | The issue the consumer identified in the complaint | 0 (0%) |
| ***Sub-issue*** | Factor (219) | The sub-issue the consumer identified in the complaint | 494,311 (49%) |
| ***Consumer Complaint Narrative*** | Text | Consumer-submitted description of "what happened" from the complaint | 739,706 (73%) |
| ***Company Public Response*** | Factor (11) | The company's optional, public-facing response to a consumer's complaint | 699,339 (69%) |
| ***Company*** | Factor (4756) | The complaint is about this company | 0 (0%) |
| ***State*** | Factor (64) | The state of the mailing address provided by the consumer | 17,981 (2%) |
| ***ZIP Code*** | Factor (28887) | The mailing ZIP code provided by the consumer | 22,267 (2%) |
| ***Tags*** | Factor (4) | Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers | 871,062 (86%) |
| ***Consumer Consent Provided?*** | Factor (6) | Identifies whether the consumer opted in to publish their complaint narrative | 530,270 (52%) |
| ***Submitted via*** | Factor (6) | How the complaint was submitted to the CFPB | 0 (0%) |
| ***Date Sent to Company*** | Date | The date the CFPB sent the complaint to the company | 0 (0%) |
| ***Company Response to Consumer*** | Factor (9) | This is how the company responded. For example, "Closed with explanation." | 3 (0%) |
| ***Timely Response?*** | Factor (2) | Whether the company gave a timely response | 0 (0%) |
| ***Consumer Disputed?*** | Factor (3) (Target) | Whether the consumer disputed the company’s response | 242,536 (24%) |
| ***Complaint ID*** | Int | The unique identification number for a complaint | 0 (0%) |

**Top Companies, products, and issues.**

This data, containing the details of consumer complaints, begs this question: under what circumstances are they disputing the original company response? Is it the company, product, or a nagging issue that points to the fact that consumers are going to dispute the original company response? First off, what is the likelihood of a consumer filing a dispute? The answer is at about 19.3% of the time. So how do we judge the accuracy of interpreting the variables in our data? We compare those figures with that percentage in mind. With quantitative analysis, we would compare correlation values or distributions, but here we need to keep in mind that we will be using a tool like a decision tree that looks for similarities and groups the variables. We want to look at the top companies, products, and issues when they had a consumer disputing their solution.

****

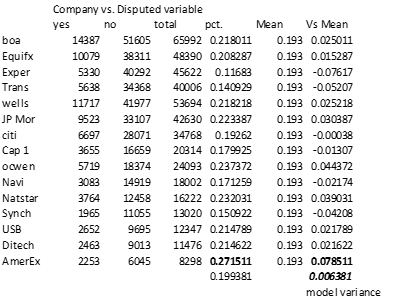
This table shows some of the main drivers for complaints and the likelihood each of these items could cause a consumer to dispute the company’s response. The most interesting note is Equifax. They had the most complaints overall, but had the lowest rate of disputes here. We will continue to examine Equifax. Overall, we see percentage values similar to the overall score of 19.3%. It appears from this that if you combined Bank of America or Wells Fargo with Credit Cards or Mortgages, products they do offer, with any of the four issues above would be strong candidates for defining the disputation rate.

Companies

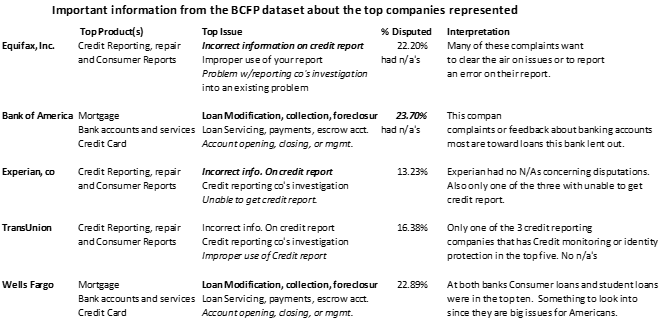
All of the companies above had levels similar to the overall percentage for the consumer.disputed variable – 19.3%. How do we judge the results of an analysis of this kind? By comparison with the that percentage when we determine values with companies, products, issues, etc…. In general, the dataset reveals complaints against over 4,000 banks, credit card companies, credit reporting agencies, loan offices, etc... all with levels of disputations for each company, product, and issue. How do we approach looking at all the possibilities with each variable? We are going to look at ones that could have an impact on our analysis of consumer disputes or describe a unique situation that tells the story of consumer issues with certain financial industries. Most on the banks on the list, like U.S. Bancorp, Bank of America, and Wells Fargo have multiple financial tools including banking accounts, credit cards, and loans they offer. That issue could become a topic of our analysis.

**Companies, products, and issues**

The companies that top the list of having the most complaints are partly described above. The question is, are there companies that have a lot higher percentage of disputes? No, as we looked at the top 15 companies, one company, the 15th top company, American Express had a value outside of a reasonable variance at 27.2%. A summation of the variances of all 15 companies landed within the mean at .5% above the dataset mean of 19.3, so if we drop American Express, then we would hit our mean, which then signifies that company would be a good variable to predict the disputation variable. NA values, which will be discussed later, of the Consumer. Disputed variable were not included in this or further analysis because of nature of many of the NA values.

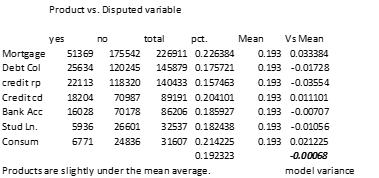
****

Some side notes: The names of the financial industry are the source of the products and issues that are on this list, so when Equifax has a security breach, Bank of America was caught selling bad mortgages to Fannie Mae, and Transunion ran into trouble with a scandal with blacklisting alerts about Americans being falsely identified as foreign nationals or as a criminal. Even though those might have led to a greater number of complaints, none of them created a bump in disputations. Student loans and healthcare were both down on the list, so if they are a concern for Americans, then they aren’t leading to complaints and/or disputations. The table below shows how each company had top products and issues.



**Product Analysis**

Products of the financial industry are household names like mortgages, checking accounts, and student loans. Just imagine, what are your issues with your bank, mortgage company, or credit card company? As we look into the data on products we find some interesting information that is helpful. We are going to look at rebranding of product titles, and how products stack up against the disputation variable. The top five products by amount of complaints in order are Mortgage, Debt collection, Credit Reporting, Credit Card, and Bank Account. This is how the products compared against the mean of the dataset’s probability with Consumer.disputes.

****

Below is the list of products that revealed that there was something happening in the dataset. Appearances are that title number one ran in the survey until a date in April of 2017, and that is when title number 2 was instituted. A study of the timeline suggests these findings. All of this was learned studying the product list.

**Duplication of product listings - All Values listed Largest number of complaints to least in each category**

**Credit reporting**

***Credit reporting \** 15.7%**

***Credit reporting, credit repair services, or other personal consumer reports \**  *100% NAs***

**Credit Cards**

***Credit card \** 20.4%**

**Credit card or prepaid card *100% NAs***

**Banking**

***Bank account or service \** 18.5%**

**Checking or savings account *~100 NAs***

**Personal Loans**

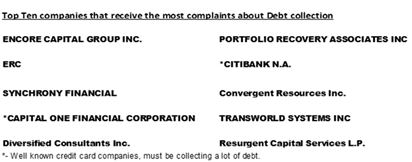
***Consumer Loan \** 21.4%**

**Payday loan, title loan, or personal loan *100% NAs***

Summation: The table reveals that a number of product branding was changed with the old brand taken out and a new one inserted into the survey, causing a split in data outcomes that should alter the method of processing of the data before building a model, where the dataset is exercised with the 1st set of product titles and then tested using the second set of titles.

When we look at the top 5 companies and then the top 5 products, there is an anomaly: debt collection. Mortgage, checking accounts, and credit cards are products that Bank of America and Wells Fargo are known for, and the credit report companies have related products we know of, but how did debt collection become number 2 in the list of those receiving the most complaints? Simply, debt collection is a wide-spread thorn in the side of consumers, and they spread their complaints across just about every company. The top ten companies receiving complaints with debt collection as the product, that most people have never heard of unless one of these companies are pursuing them. See below:

**The ascension of Debt collection to the top 5** – When we look at the top 5 companies and then the top 5 products, there is an anomaly: debt collection. Mortgage, checking accounts, and credit cards are products that Bank of America and Wells Fargo are known for, and the credit report companies have related products we know of, but how did debt collection become number 2 in the list of the receiving the most complaints? Simply, debt collection is a thorn in the side of consumers, and they spread their complaints across every company. The top ten companies receiving complaints with debt collection as the product, that most people have never heard of. See below:



What we learn is that there are 2 major credit card companies in the list, noted by the asterisks, and they get complaints in the thousands by people who don’t list credit card as the product but debt collection as the product. What this highlights is that if debt collection is a part of a companies operations, then whatever is reported concerning debt collection is technically a disputed complaint. But the way the data is laid out, that cannot be proven, because the first contact may not have been with a credit card company in this situation until the company pursued them for outstanding debt. This will need further investigation.

**Issues**

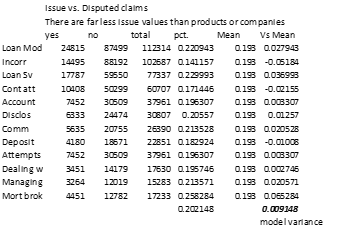
Issues would seem to be our best chance at finding keys towards finding variables for our model, but a few “issues” were discovered. Issues is the key toward understanding the complaint. Looking at companies and products, we can assume what the issues could be, but it is interesting to learn what people truly complain about, and those findings should be in the variable Issue. There are a few issues that seem to paint a picture of some of what is going on inside the data.

**Communication Tactics**

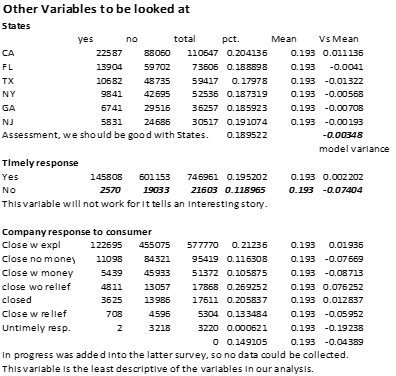
With Communication Tactics we opened up the sub-issues and the companies surrounding this Issue to uncover some interesting info. The list of Sub-issues for Communication tactics, shows us some alarming ways some companies do business:

**Frequent or repeated calls, Called after sent written cease of communication (Company) Threatened to take legal action, Used obscene/profane/abusive language You told them to stop contacting you, but they keep trying Called outside of 8am-9pm**

Some on this list are frustrating, but others are downright offensive. We decided to check, because just like debt collection is a product that would garner a lot of complaints, so would some of these issues. The table below points to how communication stacks up against the disputation variable.



Communication issues, marked by Comm, have only a slight uptick from the mean of the dataset, calculated earlier as 19.3%. Looking at the top Issues named, you can see a contrast of computed results with the consumer. Disputed variable, and the mean is still within an appropriate variance, with no sign of bias. So, with company, product, and issue being approved for the model, what other variables could we use? Because of many of our variables having a multiple of NA values, we cannot include them in this analysis. We can include state, ‘Timely response’ to see if this has an impact on disputations of past complaints, and ‘Company response to consumer’. See our table below to see how many of these three we can add.



These variables tell an interesting story but only states looks like it can help us build our model, for Timely response had surprising results with the opposite of what you would suppose, and Company response to consumer is all over the place with the mean of its values way below the range that would help us in our model. It is evident to see that people were not happy with Closed with no relief and they were good with Untimely response, whatever that determines.

**Outliers**

Outliers require a quantitative measure to be determined. We are not working with quantitative variables. Some research that in classification models, a term that doesn’t show strong relationships can be determined to be an outlier, but all resource confirm the presence of a quantitative measurement.

**Conclusion**

This dataset is broad in its capturing of consumer choices telling the stories of their complaints. All captured from one government website and released on another. There were variables we didn’t have any faith in their involvement in our model building. Variables like ‘submitted.via’ , ‘tags’, ‘Consumer narrative’, etc…. were either very inconsistent, or not useful for our investigation. We did learn some things about our dataset going forward:

1. The survey which spawned the data we are using changed in April of 2017, changing the format of the data in enough instances to be considered when analyzing this data further. The change to products and issues and taking out our dependent variable of our Consumer. disputed binary variable.

2. Although the number of complaints is higher with some companies, the numbers of disputed responses to consumers complaints was spread fairly evenly among the companies making a workable variable. Products and Issues saw the same results concerning disputations.

4. We determined in our EDA, that with Company, Product, and Issue should also be added State and Timely response. After reconsideration, having a negative correlation can actually help a model.

5. We are submitting notes of our analysis of companies and products to our marketing department.

## **Twitter Data**

To supplement data from the Consumer Complaint Database, we scraped some Tweets from Twitter using the Twitter API. Below is a list of the variables we have access to through Twitter. Due to the nature of the method we are using to obtain this data, we are unable to give a distribution of variables or missing variables. Twitter returns a sample of data and only goes back a couple weeks. Due to these limitations, our initial pitch to companies will be to monitor their Twitter data moving forward. If enough interest is shown in this aspect of the product, we will explore opportunities to purchase historical Twitter data. The top 5 rows of the data can be found in the Appendix ([Figure 5](#jsc7pa8dniv)).

**Twitter Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data Type** | **Description** |
| ***text*** | Plain Text | Text of the tweet |
| ***favorited*** | Binary | True/False. Did someone favorite the Tweet? |
| ***favoriteCount*** | Numeric | How many people favorited the Tweet |
| ***replyToSN*** | Plain Text | If Tweet is a response, Screen Name of author of original Tweet. N/A if not a response |
| ***created*** | Date | Date the Tweet was created |
| ***truncated*** | Binary | True/False. Was the returned text shortened from the original Tweet? |
| ***replyToSID*** | Factor | ID of original Tweet (if response Tweet) |
| ***id*** | Factor | ID of this Tweet |
| ***replyToUID*** | Factor | ID of author of original Tweet (if response Tweet) |
| ***statusSource*** | Plain Text | URL of source of Tweet |
| ***screenName*** | Plain Text | Screen Name of Tweet author |
| ***retweetCount*** | Numeric | How many times Tweet has been retweeted |
| ***isRetweet*** | Binary | True/False. Is this Tweet a retweet? |
| ***retweeted*** | Binary | True/False. Did someone retweet the Tweet? |
| ***longitude*** | Numeric | Longitude of posting location |
| ***latitude*** | Numeric | Latitude of posting location |

Notes on Twitter data:

* The overwhelming majority of records we have seen does not include latitude and longitude. This is because the default location setting of the Twitter app is “off” and therefore a user must manually turn it on in order for it to be collected.
* Currently, the most useful field for us is the ‘text’ field as this is what we use to define one of our KPIs (below). If a customer wants us to begin tracking Twitter presence long term, we can assume that the ‘created’, ‘favorited’, and ‘isRetweet’ fields will become more useful to us as we will have a larger collection of data.

## **Consumer Sentiment Data**

|  |  |  |
| --- | --- | --- |
| **Field name** | **Data Type** | **Description** |
| ***Date*** | Date | One-month intervals |
| ***UMCSENT*** | Numeric | This quantity is a calculation based on economic factors, which right now is trending up, easy for time series graph |

# **Research Questions and Performance Measurements**

As we build our tool, we want to be able to accurately display certain metrics to our customers. This will primarily be in the form of KPIs that we agree, in collaboration with the customer, are important to the customer’s public image. Below is a list of our primary KPIs we will use in an initial pitch to a customer:

## **KPIs**

* **Time**
  + Month over Month (Year over Year) Complaint Counts/Percentages: We will provide different time frame scopes of complaint counts and percentages to allow companies to track the progress of efforts to reduce complaints. These measures can be broken out into Location, Product, Issue, and Submit Type to allow for greater levels of granulation.
* **Location**
  + High Complaint Location-Based Clusters: This metric will provide a ranking, as well as a measure for what clusters are responsible for the most complaints.
* **Product**
  + Product/Sub-Product Health: Although our primary customers are financial institutions, the range of products they offer is still pretty wide. For example, a bank may be more concerned with Credit Card or Loan products where as a credit reporting company may only be concerned with the Credit Reporting product. Since some companies overlap in these areas, it would be hard to segment everything out. The treemap in [Figure 6](#i6i7vb7b4v90) shows a breakdown of each Product and Sub-Product with over 5,000 complaints. We would be able to provide counts and percentages for each of our customers and highlight the ones important to them.
* **Public Sentiment Analysis**
  + Tweet Score: After collecting an appropriate amount of Twitter data for a company, we can provide a summary of their overall Public Image. We will apply a -1 score for each negative word found in a Tweet and a +1 score to each positive score to come up with an overall score for that Tweet. [Figure 7](#2s7uxh9tfl2l) shows an example on a week’s worth of Equifax Tweets. From this figure, we can get several scores which we can track for our customers: Average Points per Tweet (Average of all tweets collected overtimeframe), Average Points Per Reaction (Average of all non-0 Tweets per timeframe), and Average Positve/Negative (Average Positive/Negative Score per Positive/Negative Tweet).
* **Consumer Dispute Accuracy**
  + We plan on building a classification model to identify when a complaint will be disputed by a consumer and when it will not. Once the model is built, we will be able to identify how often our customers are following the “best practices” as described by the model. We can provide a percentage that will allow our customers to understand what needs to change in the way they handle their consumer complaints in order to get a more favorable outcome.
* **Flexible KPIs**
  + As the name “Magic Mirror” reflects, we want to show companies an image of themselves, through the lens of customer feedback. Sometimes this will mean creating custom KPIs to fit the products and needs of an individual customer. With the scope of data we are collecting, we believe we will be able to offer services to fit a wide range of needs.

## **Hypotheses**

Hypothesis: If we can identify what makes a customer dispute a claim response (or we can think of this as “What makes the customer unhappy?”), we can provide valuable insight into improving our customers’ public image.

[Figure 8](#gr206wdxmtmi) in the Appendix shows a table of the variables we plan to use in the modeling process for our Consumer Disputed Classification Model. The table contains the variable, whether or not it was original or derived (and the original variable it was derived from), and the expected impact on the dependent variable (Consumer Disputed) as well as some notes explaining our reasoning. Since we are working with many text variables as well has high-level factors, several variables had to be left out of this table:

* Consumer Disputed: Dependent Variable. A positive (+) direction means the complaint was more likely to be disputed (consumer way unhappy) while a negative (-) direction means the customer was more likely not to dispute the complaint (consumer satisfied)
* Company: This variable will only be used to subset the data for our different customers
* Variables needing level-reduction or clustering:
  + Product: Too many levels, possible clustering
  + Issue: Too many levels, possible clustering
  + Consumer Complaint Narrative: Text variable, cluster analysis needed
  + State: Too many levels, possible clustering
  + Zip Code: Too many levels
  + Submitted Via: Levels may be able to be combined. More analysis needed
* Variables that won’t be used in modeling for various reasons:
  + Complaint ID: Arbitrary value
  + Date Sent to Company: Derived numerical variable created
  + Date received: Derived numerical variable created
  + Sub-Product: Too many levels/missing values
  + Sub-Issue: Too many levels/missing values
  + Company Public Response: Derived binary variable created
  + Tags: Derived binary variable created with largest number of missing values

# **Feature Handling**

## **Variable Reduction/Selection and Ranking**

Since our end product is a combination of many different plots and models, we have used different variables for different components. Below is a breakdown of the manual selections we have made as well as our plan for future variable selection.

* **Manual Selection**
  + Time Series: The variables for our time series charts and forecasts are Date Received, Company, and Count (where each observation was a count of 1) from the Consumer Complaint Database.
  + Location Heat Map: The basic location heatmap is made up of State, Zip, and Count from the Consumer Complaint Database. More in-depth cluster analysis is discussed below.
  + Twitter Score: In order to calculate the score for each tweet, we only need the text of the tweet itself. The other details of the tweet can be removed for this purpose.
* **Statistical Selection**
  + Customer Disputed Model: For this model we plan on using a combination of Random Forest Models and PCA/Factor Analysis. Each of these methods provides its own way of handling feature selection. Random Forest will determine a feature ranking based on the many decision trees it creates. This will tell us which features had the most important splits when expanded over the entire dataset. PCA and Factor Analysis create their own factors/components that describe the variation over multiple variables. Each of these methods will be explored in depth during the model building phase of the project.
  + Cluster Analysis: For out cluster analysis, we will explore different methods of text mining. Currently we have worked on translating some of the fields into “Bag of Words” and we hope to use this as a foundation to explore text mining techniques such as “Word2vec” in R as well as some distance calculations between words and phrases.

# **Tools and Techniques**

For this project, we used three sources of data. Our main dataset was the Consumer Complaint Database and we supplemented that with data gathered from Twitter and FRED Consumer Sentiment. Below we will touch on the data munging for each source.

* **Consumer Complaint Database**
  + Data Retrieval
    - The data was downloaded directly from Data.gov ([Full Source in ‘Sources’ Section](#xkez93q7pfnp)) in .csv format. Since this is a live dataset, it is important to note that we downloaded the data used for this analysis on April 13, 2018
  + Data Integration
    - The majority of our analysis was done in R which required the use of the ‘read.csv()’ function to import the data. The data was also uploaded into Tableau for additional data exploration and analysis.
  + Data Transformation and Manipulation
    - *Classification Model:* Most of the data was text data. For the features with a low number of levels, we converted these to dummy variables for use in the classification model. We also created some numerical features from the date columns we had. Once everything was either a dummy variable or numerical, we were able to build our model. When building our model, we split our data into a training and test set to test the accuracy of different models. From here, we will move to using cross validation to try to get a better understanding of how our model would perform on new data. Our prior experience has shown us that just because a model performs well on the test set, that does not mean it will perform as well on production data. We are using cross validation here because it essentially gives us many different “test sets” to test the accuracy of our model on. Looking at the results of the cross validation should give us a good idea how our model will perform in production.
    - *Time Series:* We needed to convert the ‘Date Received’ variable to a date format that R would recognize. We did this using the ‘as.Date()’ function in R. We are also planning on using a train and test set for time series. To test the accuracy, we will take out the most recent year (or possibly less depending on the company) of data and build a model on the remaining data. Then we can see how far off our predictions are from the actuals.
* **Twitter**
  + Data Retrieval
    - The Twitter data was retrieved from the Twitter API. In order to access this data, we needed to have a Twitter account which we used to create a project that gave us a ‘Consumer Key’ and ‘Consumer Secret Key’ which we then used to access the data through R.
  + Data Integration
    - In R, we used the ‘searchTwitter()’ function from the twitteR package to query data. The keywords in our queries were the Company names we were interested in.
  + Data Transformation and Manipulation
    - *Rawdata:* Once the data was in R, we converted it to a dataframe using the ‘twListToDF()’ function. We then eliminated any Tweets that were actually Tweeted by the company in order to remove as much bias as possible from the data we wanted to work with.
    - *Score:* In order to come up with the Twitter Score we are using as a gauge for a company’s public image in social media, we used a function we found on the website “R-Bloggers.com” ([Source in ‘Sources’ section](#35k2e02i3wno)) which compared the text in the Tweet to two different datasets. One dataset contained “bad” words and another contained “good” words. We were then able to use these scores to create visuals such as the Equifax chart in [Figure 7](#2s7uxh9tfl2l).
* **FRED Consumer Sentiment**
  + Data Processing
    - Found from Federal Reserve Bank of St. Louis website ([Source in ‘Sources’ section](#eb9jd7wky250)). This site offers an example of the data by a predetermined calculation involving movement of the interest rate, the unemployment rate, and GDP. This reveals a certain consumer sentiment that can determine whether attitudes towards financial institutions are up or down. Complete with downloads, the data will be helpful for illustration purposes, no manipulation necessary.

# **Methodology**

Magic Mirror is made up of several key components that we plan to put together in a Dashboard as a summary for our customers on their public image. We have talked about many components of this dashboard that we believe will be useful to our customers including a Time Series Forecast of their complaint volume, a Twitter Score that tells them the number of ‘Good’ vs ‘Bad’ Tweets that are out there about them, and a cluster analysis on customer narratives that can tell them a bit more specifically what customer are complaining about. However, what we believe is our biggest component, is the Classification Model that will predict whether or not a customer disputed the company’s response to their complaint. Ideally we would like to eliminate all complaints, however, this is a nearly impossible task. A more feasible task, is to begin to eliminate customers being unhappy even after a resolution to their complaint has been found. We believe the classification model can do this by showing us what variables contribute to the customer disputing the complaint and from there we can advise our customers to take action.

Our approach to the Classification Model was to first look at the model to see what we were working with. Although the data technically came from surveys, we did not believe that a Hierarchically Structured Classification Model would work. This is because the questions asked on these surveys not not seem to have been presented in a specific way. There were no mandatory questions besides the product and issue associated with the complaint (as these were two fields that had no missing values) and there was really only one open response question.

After eliminating the possibility of a Hierarchically Structured Classification Model, we knew that we were dealing with a **Binomial Classification Model** because this was a **Supervised** analysis due to the target variable of ‘Customer Disputed’ which had only two possible responses: “Yes” or “No”.

Once deciding on our methodology, we began converting our data to a format suitable for a binary classification model. [Figure 8](#gr206wdxmtmi) contains a table of our list of variables and how we converted the many text variables into dummy or binary variables that can be used in the model. We also hope to do some text analysis on the ‘Issue’ and ‘Consumer Complaint Narrative’ variable to try and gain some understanding from them. From there, we will explore logistic regression, random forests, neural networks, and boosting models to find the best solution for our model. These methods, as well as the models we build with them, will be expanded upon in the Model Building section.

## **Computational Methods**

* **Software**
  + *R*
    - R was used for much of our initial EDA as well as the main modeling components of the project because that is what the project team felt most comfortable using. We used a variety of packages such as t-series, ggplot2, and forecast. The full list of used packages can be found in our attached code.
  + *Excel*
    - Excel was used to view the data before, and as a reference during, our work in R. It gave us a way to look at counts and do simple filtering so we had a general idea of what we were working with.
  + *Tableau*
    - Tableau was used in a similar way to Excel in the beginning of the project. As we move towards the end of the project and creating the final Dashboard, Tableau will be relied upon to create visuals on other elements of the dashboard.
* **Hardware Platform**
  + The machine where much of the coding was performed was a Windows Machine with a Core i7 Processor and 8GB of RAM. Unfortunately, we did not have direct access to a more powerful server. This has been a problem in a few areas such as running a model with too many predictor variables. To combat this, we are relying on feature selection methods to decrease the numbers of variables in the model.

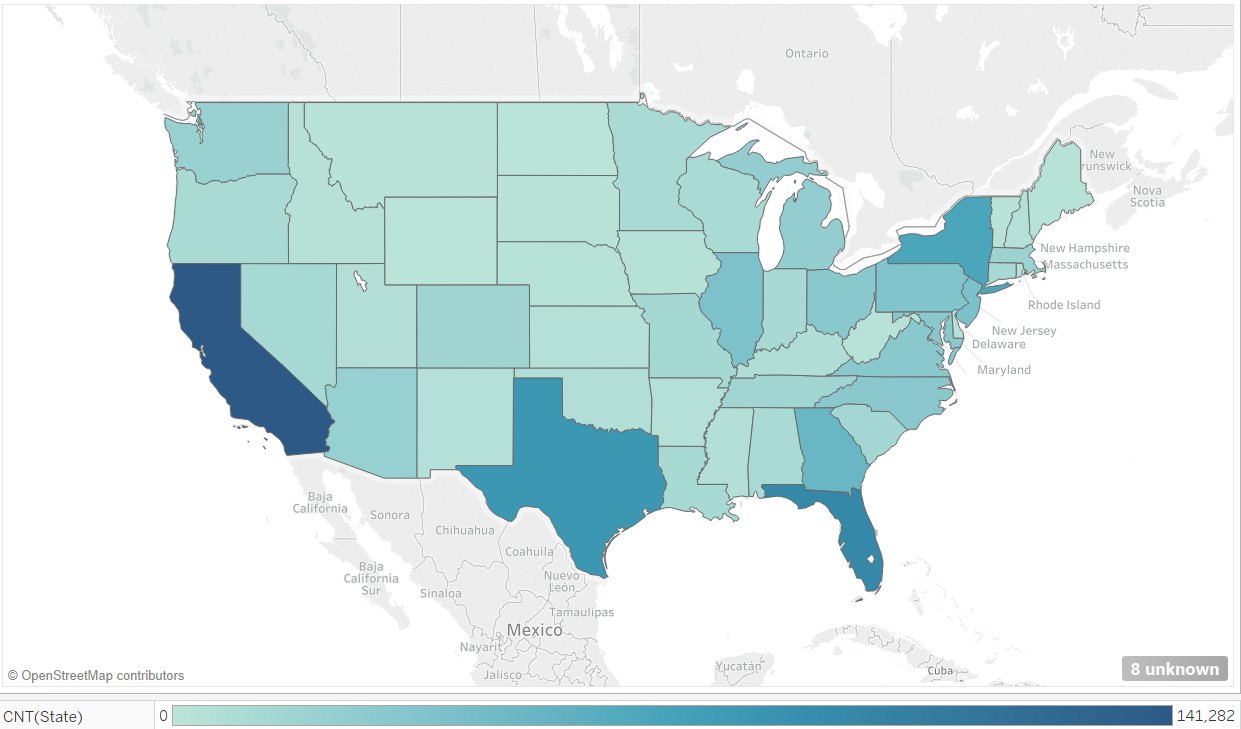
## **Outputs**

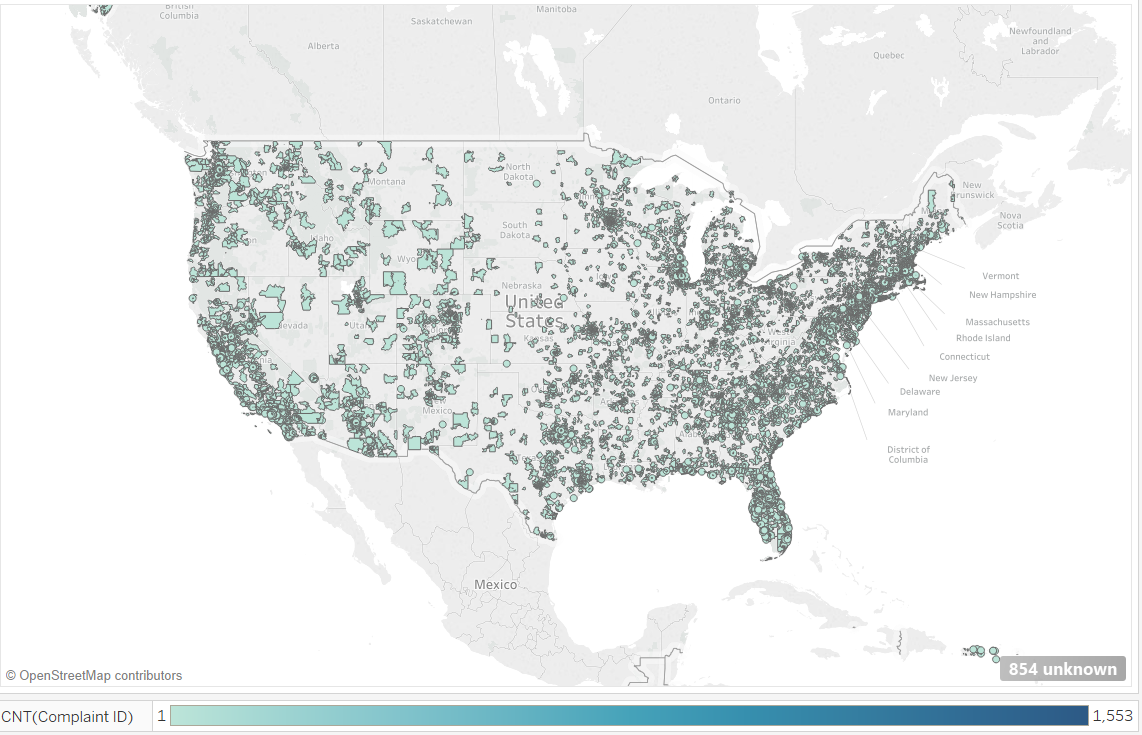
**Computational Outputs**

* *Time Series Forecast:* For our time series, we will get an output that predicts the complaint volume for a specific company over the next 12 months. Additionally, we will provide 80% and 95% confidence intervals for our forecast. While 12 months is the default value for predictions, this and other parameters can be adjusted as appropriate to better reflect the needs and situation of specific companies.
* *Classification Model:* For our classification model, we will end up with a formula or method that predicts whether or not a company’s response to a complaint will be disputed by the customer. Depending on what kind of model we choose for our final model, this output could look very different. Currently we are considering logistic regression, random forests, neural networks, and boosting models as well as anything else we find along the way that provides us with a good accuracy. From this output, we will also get an idea of what variables are important to whether or not the response is disputed. We can give this information to the customer with the hopes that they can change current processes to make the number of disputed claims go down.

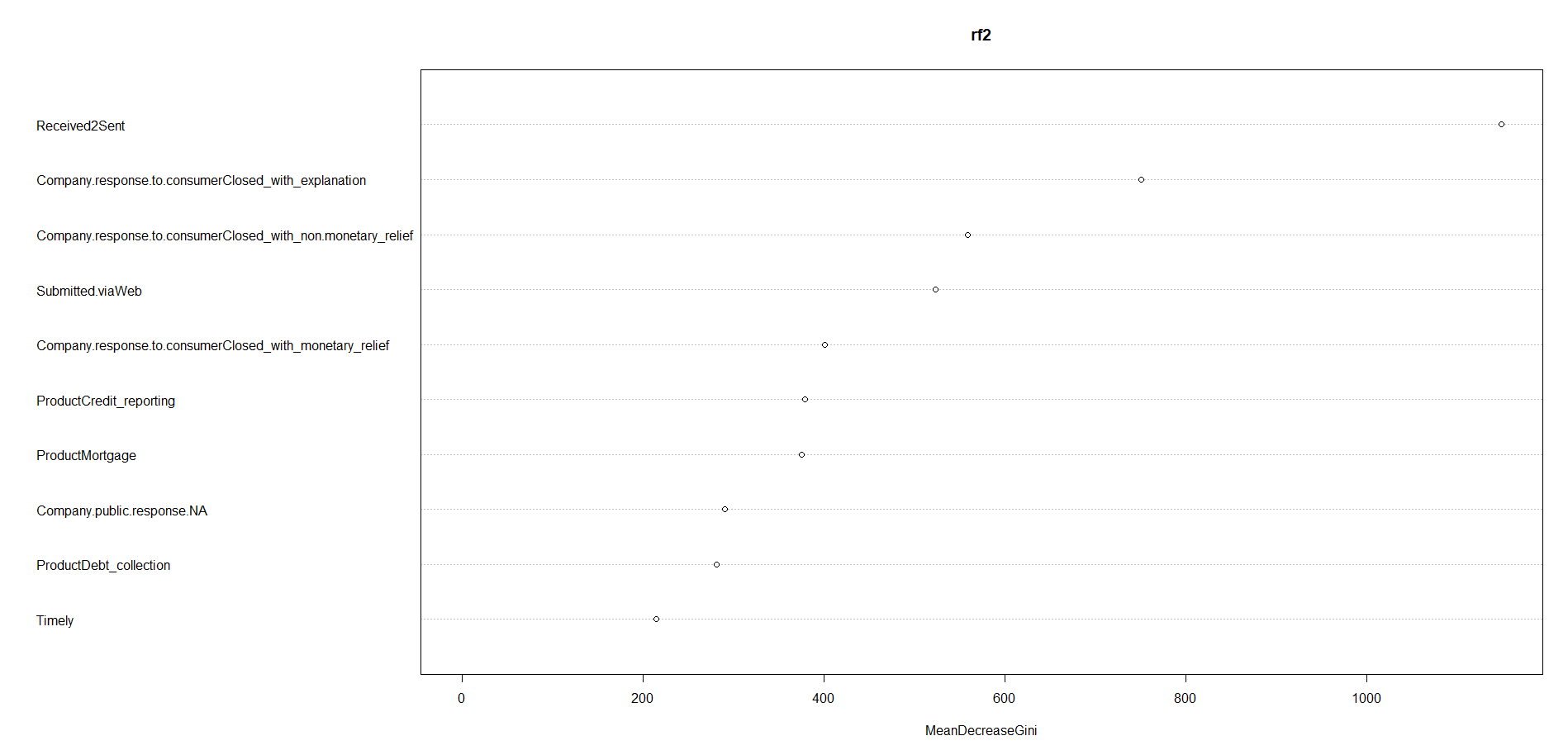
**Visuals and Reports:**

Below are some outputs from our Dashboard. We have included a mixture of exploratory data pieces as well as sample outputs from our models. These are all options our customers will be able to have on their customized dashboard.

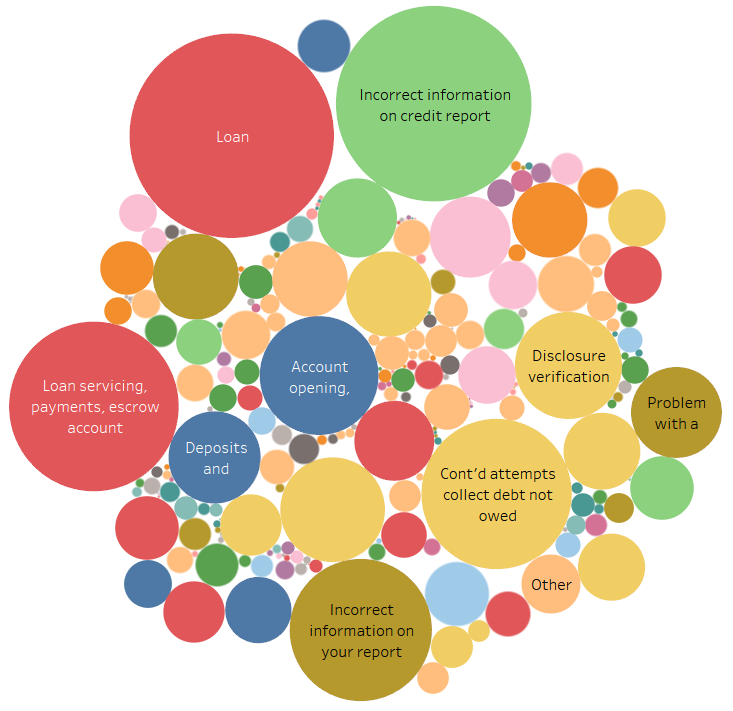




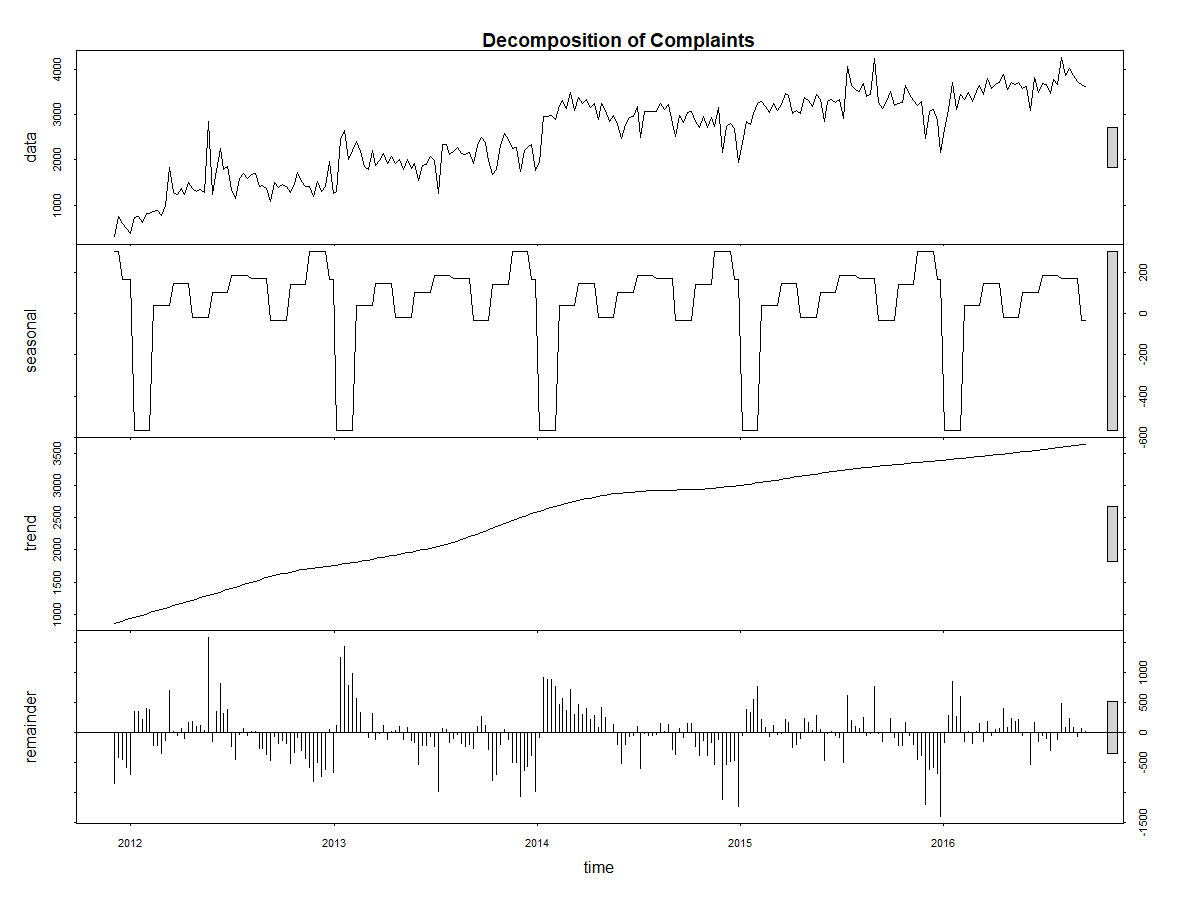
*Heat Map*: A heat map was developed that identifies the relative count of complaints by State and by Zip Code.



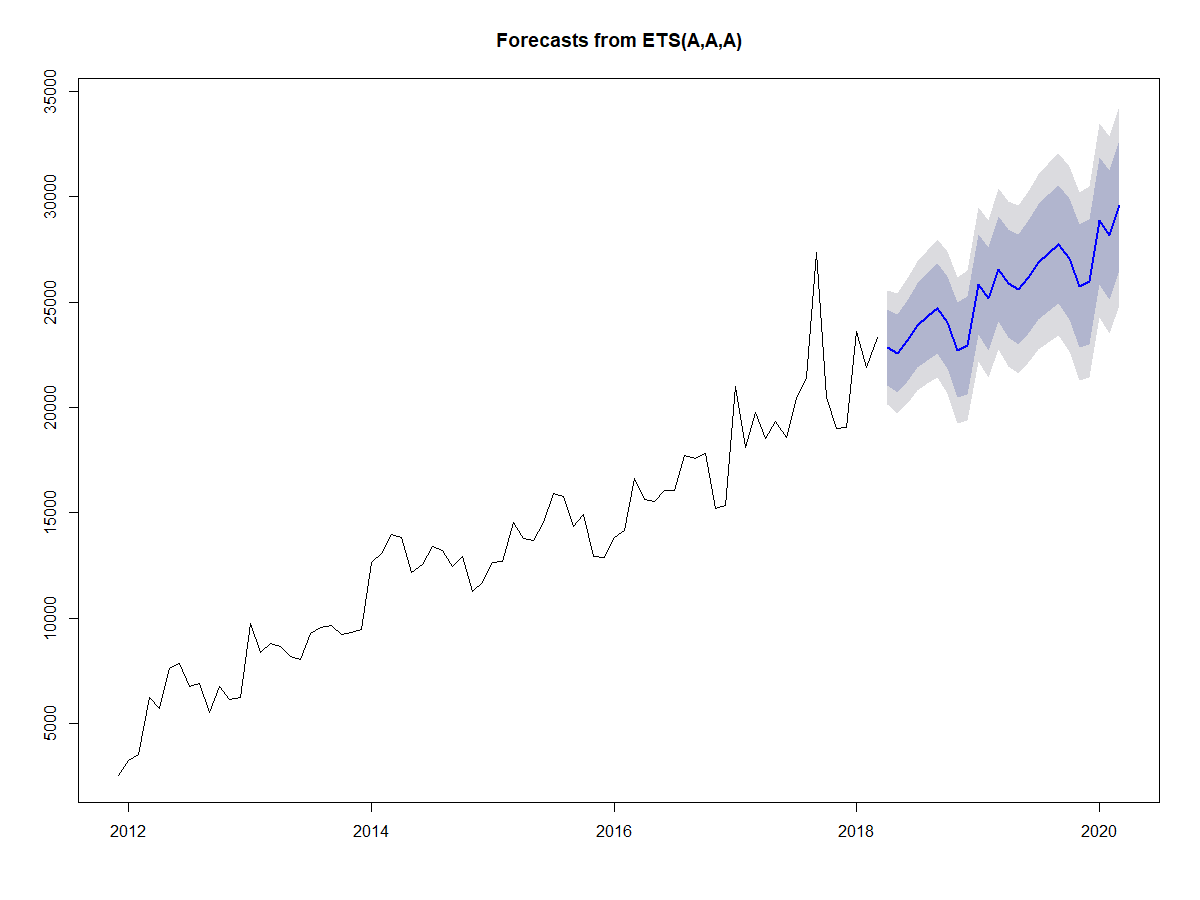
*Importance Chart*: The following importance chart is from our Random Forest model. The chart is intended to be indicative of what our final variable importance chart should look like. Chart will be adjusted accordingly based on information derived from final model.



*Packed Bubble Chart*: A packed bubble chart was developed to identify high complaint counts at the intersection of product and issue. The size of the bubble is proportional to the relative count of complaints at this intersection. Color of the bubble is used to identify the product.



*Time Series Charts:* Line charts were developed to show the decomposition of time series components for complaint counts. The chart contains four parts. (1) Actual counts are shown. (2) The assumed seasonal components are identified. (3) A calculated trend is shown. (4) The remainder (i.e., noise) is shown.



*Time Series Forecast Line Chart*: A line chart will be added showing the monthly complaint counts along with forecasted values for the next twelve months. Additionally, 80% and 95% confidence intervals are shown.

# **Dashboard Components**

## **Classification Model**

**Goal**

The goal of the classification model is to predict whether or not a consumer will dispute a company’s response to their complaint. Beyond this though, it is to see what variables are important to that outcome so we can educate our customers on what they need to improve on to reduce complaints. Since we don’t have any data points where there wasn’t a complaint, we cannot analyze what a complaint looks like compared to a non-complaint. However, we do have the disputed variable, so we can see what actions our customers can take to prevent disputes. We hypothesize that some of the same things that cause a dispute are also what cause a complaint in the first place so in reducing disputes, we hope to reduce complaints as a whole.

**Difficulties**

There were several major difficulties when building this model. The first was the format all of our data was in: text. This caused the number of features we were using to balloon when creating dummy and tokenized variables. This, combined with the number of observations in our dataset, caused several processing issues when running models. There were actually several things we had to cut out from model building because we simply did not have enough processing power to build a good model from them. For example, when tuning the XGBoost model we had trouble running a full cross validation and had to settle for a random cross validation. This should have still gotten us close to the pest hyperparameters, but it is not exact.

Another obstacle was that the Consumer Complaint Database stopped collecting the “Consumer Disputed” variable on April 24, 2017. However, we believe that this variable is still important because it gives us an idea of what makes a customer upset *beyond* the normal complaint. The insights we find from this model should be able to be applied to a normal complaint and not just a disputed one. Also, if there is enough information gathered from this variable, we could possibly petition the CCD to begin collecting information on this variable again.

The final, and possibly, biggest issue with this data was that it was imbalanced. After removing missing values, the Consumer Disputed variable was split 80/20 between Not Disputed and Disputed ([Figure 10](#3k0jkm1ojh56)). Most of our early models simply put all observations into the “Not Disputed” category in order to get an 80% accuracy rate. To combat this, we tried to create both and under and over sampled dataset based on our original data using the ROSE package in R. Due to the size of the our data and our aforementioned resource constraints, we did most of our work with the undersampled dataset as it was smaller. Our accuracy obviously decreased using this method, but we were now getting a better split between disputed and non-disputed in our predictions.

**Feature Selection and Variable Importance**

For XGBoost and Random Forest models, the feature selection was built into the model building. The different splits performed when creating each model automatically calculated which variables were important in determining the outcome. The variable importance for the Random Forest variable importance is in [Figure 11](#76m2vvpum62i). The way these importances are calculated is by looking at the distribution of the Target variable at a particular split and comparing it to the other split. For example, if we take our “Consumer Narrative NA” variable, which is a binary variable that tells us whether or not the consumer left a narrative, we see two outcomes: left a narrative and didn’t leave a narrative. This is where the split in this variable occurs. If the majority of target variables in one side of this split are 1 and a majority of the variables in the other side of the split are 0, the Gini is decreased for “Consumer Narrative NA”. At the end of the model building process, we look at the average of how much each variable decreased the Gini and order them to find the most important variables.

For the Logistic Model, we ran a forward, backward, and both stepwise selection and chose the best model. The backward, forward, and both selection gave us very similar models so we moved ahead with the forward one. For variable importance in the logistic model, we can use the varImp function in R which returns the absolute value of the t-statistic for each variable. The chart in [Figure 12](#6vatzgr3igho) shows the top 10 importances in descending order for our best Logistic Model.

**Tokenization**

Some of the variables we had to work with had too much text to break into dummy variables as they simply would have created too many features to manage. The “issues” and “Consumer Narrative” variables had a lot of text but still information we felt would be useful in the model. We used the process of Tokenization to break down each observation into the core word that formed it. This meant stripping away punctuation and stop words such as “is”, “the”, “and”, etc. as well as converting words to their stem word. From there we were able to look at some of the most common/important words for each variable. Figure contains a word cloud for the top 100 words from the “Issues” as well as one for the top 100 words in “Consumer Narrative”. We experimented with using these breakdowns of the text variables in our models by using TF-IDF which stands for “Term Frequency–Inverse Document Frequency. This basically tells you how important a word is out of a collection of words. It is used in searches and text mining and can be combined with our data in the form of a sparse matrix. Due to time constraints, we only scratched the surface of using TF-IDF scores in our models but it is definitely something we feel would be worth pursuing in the future.

**Models**

We experimented with three different types of models and several different approaches within those methods. We built XGBoost, Logistic, and Random Forest Classification Models to try to predict whether or not a consumer would dispute a company’s response to a complaint based on the rest of the data provided. Most of our time was spent on the XGBoost model since we saw the most opportunity to explore the text data using some XGBoost-exclusive functions. We found the best results from the Logistic model and found the Random Forest to be a good resource for feature importance.

XGBoost

*Sparse Matrix*: Our first XGBoost attempt was focused on incorporating the TF-IDF scores from the tokenization method talked about above. In order to combined these scores with the rest of our data, we created a sparse matrix and simply used the cbind() function in R to merge the score back after they had been calculated. From there, we ran the XGBoost model with the “gbtree” booster and “binary:logistic” objective as well as several different combinations of other hyperparameters. Unfortunately, we were unable to get away from out imbalanced problem using this approach as we continually got ~80% accuracy where the model would predict almost all 0s. Moving forward, we believe that this approach has the most potential to be tweaked to a point where we get a better accuracy. [Figure 17](#bd6c4zuc1lek) shows the variables importances for this model.

*Hyperparameter Tuning*: Our next approach was to tune the hyperparameters of the model using cross validation. We set up a leanrer in R and used the tuneParams() function to take our training data and run it against a range of values for the paramaters booster, max depth, minimum child weight, subsample, and colsample bytree. Using this method, we still got the same results (~80$ accuracy) due to the imbalanced dataset. We ran the tuning function again, but this time we ran it on the under sampled dataset. This showed promising initial results with a training error of about .4. However, when we used the model to predict on the original test dataset, we saw barely any improvement with a test accuracy of only 52%. The confusion matrix and accuracy for this model can be seen in [Figure 13](#ms4s7sfw3vmr).

Logistic

*Stepwise Selection*: Our approach to Logistic Regression was to use stepwise feature section (forward, backward, and both) to selection the feature from our entire dataset (excluding TF-IDF scores). We saw similar results from all 3 methods and decided to go with the Forward selected model as our best one which had an accuracy of 55%. [Figure 14](#ugoicy8g7p6c) shows the function and variables included in this modela and [Figure 15](#jxs4ri3by2ku) shows the summary of the base logistic model as well as the confusion Matrix of the Forward model.

Random Forest

*Random Forest*:The main objective when building the Random Forest model was to get a comparison for feature importance against what we gathered from our Logistic Model. The Random Forest model we built was very simple. We used the randomForest package in R to build the model and obtain the variable importance chart in [Figure 11](#76m2vvpum62i). The dataset used was our undersampled dataset so that we could avoid an all 0 result. This model got an accuracy of 51% and the confusion matrix can be seen in [Figure 16](#rgosm085cs4d).

**Model Selection (Final Model)**

The table below summarizes the accuracies of our best model from each method:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| *XGBoost* | 52% |
| *Logistic* | 55% |
| *Random Forest* | 51% |

As stated earlier, we had models with accuracy in the 80s, however, we did not feel these models gave an accurate representation of the data. Given this caveat, our most predictive model was the Logistic Model.

**Results**

Looking past predictive accuracy, we focus on the most important variables from our models. Of our top models above, these variables were in the Top 10 Importances for all models: ProductCredit\_Reporting, ProductDebt\_collection, Timely, Company.public.response.NA, and Received2Sent.

One thing that stuck out, was the number of Product Dummy variables that were important in both models. Besides those listed, each model had several more product dummy variables in their Top 10 Importances. We believe this is due in part to the number of product variables there were and also to the number of different specific sub-industries inside the financial industry that we were dealing with.

Maybe the most important thing our customers can learn from is the importance of the Company.public.response.NA variable. Knowing when and how to respond publicly to an issue may be in key in reducing complaints and improving company image. Our group believes that exploring this variable and the relationships other variables have with it is key to unlocking the potnential of this model from an accuracy perspective.

## **Twitter Score**

**Goal**

We want to provide as wide of a picture as possible for our customers to see their public image. Since Twitter is one of, if not the most up-to-date place for public reaction, we chose it as our tool to provide insight into customers minds. When looking at the CCD, we knew that there were more people unhappy that simply didn’t submit a complaint. With Twitter, we hope to capture these negative (and positive) reactions and provide and easy to interpret score and visual for our customers so they can track their improvement in the public eye.

**Process**

We used the TwitteR package in R to get the Tweets. We first ahd to create a Twitter account and to tell Twitter you are going to mine their Tweets via their API. From there, Twitter gave us a couple of keys to enter in R which perform a handshake with Twitter so we could start pulling in these tweets. Unfortunately, we found that doing it this was only provides a sample of Tweets of over a limited timeframe so we couldn’t go back to the start date of our CCD data to get a perfect match. In the future, we hope to see what kind of value this provides for our customers and possibly explore monetary ways to gather all historical Tweets.  
 Some of the variables extracted from the tweets were: the tweet itself, the author, the date, and whether or not it was retweeted or favorited. There was location data, but very few tweets actually included it as it is set to “private” as a default and a user would actually have to go into the twitter settings and turn on that they want their location shared for us to be able to get it.

**Outcome**

As an example, we pulled a few weeks of Equifax related Tweets. We did this by searching for tweets that had the quote “Equifax” in them then deleting all those that were posted by Equifax themselves to remove any bias. We got a few thousand Tweets to work with.  
 We took the actual text of these Tweets and ran it through a function we found from Jeff Breen on R-Bloggers.com. This function compares each word in each tweet to two different repositories of words: one is of “negative” words and one is a “positive” words. They contain a few thousand words each as well as common misspellings of those words.  
 The function then gives a positive point to each Tweet for each positive words and a negative point for each negative word. We believe that when collecting and scoring this data over time, we can provide a periodic score for the company that they can use as a measure of their public image.

[Figure 7](#2s7uxh9tfl2l) shows an example chart of Equifax Tweets pulled over a one week span. As you can see, there are a large number of neutral and -1 Score Tweets. From these Tweets we give Equifax several scores:

* Average Points per Tweet (Average of all tweets collected overtimeframe): -0.42
* Average Points Per Reaction (Average of all non-0 Tweets per timeframe): -0.67
* Average Negative (Average Score per Negative Tweet): -1.25
* Average Positve/Negative (Average Score per Positive Tweet): 1.19

This tells us that most people Tweet negatively about Equifax and that those negative Tweets are “more” negative than the positive Tweets are positive according to our function.

## **Cluster Analysis**

As a supplement to the rest of our analysis, the advisor of our project team provided us with a cluster analysis of a sample of 10,000 consumer narratives to find out what groups the consumers can be broken out into. Hierarchical clustering using the tfidf scores was used in order to cluster these customer segments together.  
 This kind of analysis has a similar importance to what we tried to accomplish earlier using the word cloud and tokenization in our first XGBoost model in that it allows different companies to focus on different clusters and the specific needs of those customers. Since we are working with financial institutions that are anything from banks to financial planners, we need to be able to focus on different areas of the consumer population.Looking at the segments, we can see clear lines between them:

*Segment 1:* Deals with credit reporting. This could be because of the recent data breaches by companies like Equifax driving in a higher complaint volume.  
*Segment 2:* Looks like it is more to do with banks and smaller issues like late fees and missed payments. These could be legitimate complaints or simply that customers were frustrated with the process.  
*Segment 3:*  Has more severe complaints having to do with debt and collection agencies.  
*Segment 4:* More specific types of loans that may have similar complaints. Such as home and student loans.  
*Segment 5:* Has complaints where the customer just feels like they have been treated unfair. This can relate to many other issues where the customer just had a bad experience dealing with a company’s customer service.

This analysis will provide a lot of value to our consumers because it will allow them to hone in on how to handle different kinds of complaints. For instance, if complaints start to come in at a higher than normal rate from Segment 1, our customer can look to proven complaint handling methods in order to stop whatever is causing the complaints before it gets out of control.

## **Time Series**

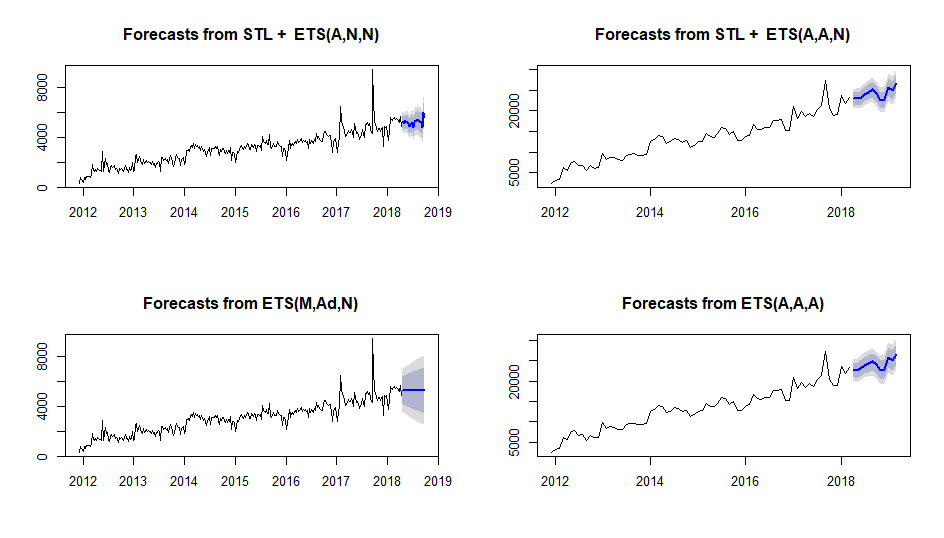
Our objective was to develop a time series component relating to the aggregation of complaint counts. This component would provide a potential client with an indication of past trend, a forecast, and associated confidence intervals for the forecast. Each record within the dataset would be considered a complaint with aggregations being made based on the date received variable. Aggregations were made by day, week, month and year. The analysis and subsequent development focused on weekly and monthly aggregations. It was felt that daily aggregations would provide too many issues and that yearly aggregations would be of little value. With regular updating of the data, these time series charts can be updated.

**Time Series Analysis**

An analysis of the data showed a clear upward trend in complaint counts over the timespan of the dataset. There are likely a number of contributing factors for this including increased usage of the internet during the duration of the dataset being used. The upward trend showed visible fluctuations in the counts. As expected, these fluctuations begin to smooth out as the aggregations move toward longer time intervals (i.e., from daily to yearly). Also, as we find a total of 4,000 companies in the data, it appears that the website started regionally, at first, but has grown over the timespan they have built this database, as well. Seasonal trend decompositions were run on the weekly and monthly data to better analyse the trend and corresponding fluctuations. Seasonality was detected in the weekly data but appeared more prevalent in the monthly aggregations. As expected, noise appeared to have a more noticeable influence in the weekly data. Once adjusted for seasonality and noise, the trends for both weekly and monthly data appeared similar. A comparison of the monthly aggregations was made for the top five companies (Bank of America, Equifax, Experian, Transunion, and Wells Fargo). For this comparison, there seemed to be a lot of similarities in the visible fluctuations. These similarities in fluctuations may partially be due to seasonality. While there are some similarities in the trends for these five companies, specifically between 2014 and 2016, the specific companies clearly have varying trends. Notable is the Equifax information breach causing the spike in September of 2017.

**Forecasting**

Forecasts were constructed for both weekly and monthly aggregations using the stlf and ets functions from the forecast package in R. Forecasts were made for six months out with 80% and 95% confidence intervals being shown. The forecasts produced for the weekly aggregations showed a more levelized forecast with the ets function producing a straight horizontal line forecast with a wide band for the confidence intervals. The monthly forecasts appeared to trend more upward as would be expected with similar forecast lines and confidence intervals being produced from both the stlf and ets functions. The graph will illustrate:



**Forecasting Challenges and Recommendations**

The challenge will be to produce a forecast suitable for a wide range of companies. Many companies may not have sufficient data to produce a desired forecast. In these cases industry data may need to be substituted. Additionally, these different companies may likely have very different needs concerning what they need out of a forecast such as weekly vs. monthly aggregations, display of confidence intervals, needed adjustments for seasonality, etc. The internal recommendation would be as a default to utilize either the stlf or ets functions to produce a forecast with either weekly or monthly aggregations being made as appropriate. In many cases, moving averages or utilization of other techniques may be appropriate. As a result of the size of the potential client base, any need to enhance or customize would likely need to be negotiated on a case by case basis.

**Implementation**

AWS Dashboard Service – These parameters are for implementation and visualization, the modeling will still be done in R, and results visualized with AWS.

1. To have workable datasets, we will be cutting down our dataset to only include the dates between April 2017\*- Present and will aggregate the data to weekly and monthly numbers for use.

2. We will use RStudio for our analysis and local storage of our very large datasets.

4. We would employ AWS, and be administrative users, who can assign author level users and read-only users. We would also have or be an author, who would make sure that we have correct dashboards for our customers. Customers would have read-only privileges to access the dashboards we assign them. The tools in each visual are usable when they upload it.

5. The way the AWS service works equips us with the opportunity for the data in our S3 bucket to be refreshed, and a schedule can also be set in Quicksight to refresh the data from S3, which will update the visuals after the refresh happens allowing our customers to have very timely information in their dashboards with Magic Mirror.

6. Availability of an AWS- Quicksight app, so they can access the most current dashboard for any approved user on any mobile device.

7. We will internalize the costs associated with AWS, as it is not costly and is billed monthly, which can be integrated into our clients fee.

\*-The date that they remade the survey to condense our variables.

# **Project Challenges**

The biggest challenge our group faced in this project was working with a completely text-based dataset. None of us had had any experience working with text mining or text variables in general beyond creating dummy variables. We spent a good portion of the project educating ourselves on Tokenization, Corpus, and TF-IDF and how we could incorporate them into our models. We learned a lot from this project and feel that if we were to do something like this again, we would have a much smoother time moving into the modeling phase and handling the text data.

Another major challenge from this project was the imbalanced dataset. We started off the modeling process thinking we had some very accurate models from the start. But once we looked at the confusion matrices, we knew that the models were simply using the imbalance in the target variable to get a high prediction accuracy. This was another new challenge for our team because we had all either worked with datasets that were more evenly distributed or developed models that were able to get a higher accuracy than even the imbalance could provide. The ROSE package and under/over-sampling our dataset was a new process that took a bit of getting used to. Besides creating the actual dataset, we learned how to use it in building a model as well as the need for switching back to the base data for prediction purposes.

# **Opportunities**

Moving forward, the next step is pitching the idea to potential customers so we can begin work from within a company. Understanding an individual company's ideals and strategies is just as important in a project such as this to the data we have collected and analyzed in this report. We want to work with the company to help them improve and not simply provide them data that is of no use to them.  
 We know there is still room for improvement on the modeling side as well as what we can offer. We want to be able to offer whatever a company needs in order to improve their processes. That will mean continuing to evolve and come up with new ideas in the future. Combining some of the methods used in the classification model is a major goal of the project team. We spent the vast majority of our coding time learning these methods and implementing them on our data, but combining them into a single model may be the key to increase our accuracy. From a social media perspective, we think that we uncovered a lot of good information through the Twitter Score, but we believe that by obtaining a larger sample of Twitter data (preferable back to the start of the CCD) we can provide a side-by-side look of negative social media response and complaint counts. If there is any sort of lag between the Tweets and the complaints, our customers could better prefer for incoming complaints after seeing a influx of negative comments on Twitter. We also hope to work with other tools and processes outside of the ones we used in this project. An idea we discussed throughout the project was providing some of these scores and importance real time via a web app. Obviously that is a bit outside of our skill sets for this project but it is something we feel would provide a lot of value moving forward. These are just some of the ideas and opportunities we would like to pursue in the future of Magic Mirror. We feel that this project and the analysis we have done has uncovered a lot of potential for future success.

# **Final Conclusion**

Although the Classification Model and it’s variable importances can be looked at as the main component of our Dashboard, we feel it is important for our customers to understand that they will only get the maximum value out of Magic Mirror if they use all the components of the dashboard in conjunction with one another. The Twitter Score and Cluster analysis can give our customers a better idea of who they are working with while the Classification Model and Time Series Forecast can give them an idea of how to fix their problems. Without this understanding of the entire picture, a company can’t hope to truly fix their public image. This all relates back to the [Issue Tree](#oau8offbj46f) and our point that if you stop before you get to the roots of the tree, you won’t really fix your problem. You will only find a temporary solution instead of ridding yourself of the issue completely.

Overall, this project was a huge learning experience for all members involved with the modeling aspect. Moving forward, we feel as though we have a solid foundation on which to continue to hone our skills. Text data and imbalanced dataset both provide a lot of obstacles, but with those obstacles are opportunities for creative solutions.

# **Sources**

Consumer Complaint Database: <https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation>

Twitter Data:

<https://twitter.com/>

Consumer Sentiment Data:

<https://fred.stlouisfed.org/series/UMCSENT/>

Twitter Score Function:

<https://www.r-bloggers.com/how-to-use-r-to-scrape-tweets-super-tuesday-2016/>

Statistician’s Ten Steps for Data Quality Management:

<http://crmportals.com/2/post/2013/06/statisticians-ten-steps-for-data-quality-management.html>

Aubin, D., (JUNE 21, 2017) *TransUnion hit with $60 million jury verdict over blacklist alerts,* Retrieved from <https://www.reuters.com/article/transunion-verdict/transunion-hit-with-60-million-jury-verdict-over-blacklist-alerts-idUSL1N1JI1ZA>.

Dilworth, K., (December 6, 2012) *Credit reports falsely tag consumers as terrorists, drug traffickers.* Retrieved from<https://www.creditcards.com/credit-card-news/credit-reports-ofac-alert-terrorists-dispute-1270.php>.

Equifax, Inc., (March 1,2018). *2017 Cybersecurity Incident & Important Consumer Information,* Retrieved from<https://www.equifaxsecurity2017.com/consumer-notice/>

Isadore, C., (January 7, 2013). *Bank of America in $10 billion settlement with Fannie Mae.* Retrieved from<http://money.cnn.com/2013/01/07/news/companies/fannie-bank-of-america/index.html>.

Josuweit, A., (June 8, 2017). *3 Money Issues Millennials Stress About The Most,* Retrieved from<https://www.forbes.com/sites/andrewjosuweit/2017/06/08/3-money-issues-millennials-stress-about-the-most/#7a7ff64d2c2b>.

Rasiel, E. M., & Friga, P. N. (2017). *The McKinsey mind: Understanding and implementing the problem-solving tools and management techniques of the worlds top strategic consulting firm*. Beijing: Ji xie gong ye chu ban she.

Rawes, E., (Sept. 20, 2014). *Top 5 money problems Americans face.* Retrieved from<https://www.usatoday.com/story/money/personalfinance/2014/09/20/wall-st-cheat-sheet-money-problems/15832929/>.

Robaton, A., (April 20, 2017). *Biggest money worries keeping Americans up at night.* Retrieved by<https://www.cbsnews.com/news/biggest-money-worries-keeping-americans-up-at-night/>

Negative and Positive Word Repositiories:

Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."

Proceedings of the ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,

Washington, USA,

# **Appendix**

Figure 1

'data.frame': 1011100 obs. of 18 variables:

$ Date.received : Factor w/ 2325 levels "1/1/2012","1/1/2013",..: 1022 222 270 1764 2172 429 1639 1633 431 298 ...

$ Product : Factor w/ 18 levels "Bank account or service",..: 11 6 3 4 8 11 8 6 11 11 ...

$ Sub.product : Factor w/ 76 levels "","(CD) Certificate of deposit",..: 53 NA 75 NA 13 10 41 NA 53 11 ...

$ Issue : Factor w/ 166 levels "Account opening, closing, or management",..: 84 76 92 18 37 86 72 50 84 84 ...

$ Sub.issue : Factor w/ 219 levels "","Account information incorrect",..: NA 5 NA NA 83 NA 45 90 NA NA ...

$ Consumer.complaint.narrative: Factor w/ 262439 levels "","'Collection Notices from XXXX for Equipment that I do not have. ' XX/XX/XXXX - Contacted XXXX and cancelled TV "| \_\_truncated\_\_,..: NA 100036 116740 NA NA NA NA 9963 NA NA ...

$ Company.public.response : Factor w/ 11 levels "","Company believes complaint caused principally by actions of third party outside the control or direction of the company",..: NA 11 NA NA NA NA 6 9 6 NA ...

$ Company : Factor w/ 4756 levels "(Former)Shapiro, Swertfeger & Hasty, LLP",..: 2657 4389 950 307 949 4426 759 1603 205 3205 ...

$ State : Factor w/ 64 levels "","AA","AE","AK",..: 31 5 48 21 58 32 10 58 10 15 ...

$ ZIP.code : Factor w/ 28886 levels "","-1631","-2914",..: 13851 9587 2860 23940 4722 13810 25893 4477 26355 8318 ...

$ Tags : Factor w/ 4 levels "","Older American",..: NA NA 2 2 NA NA NA NA NA 2 ...

$ Consumer.consent.provided. : Factor w/ 6 levels "","Consent not provided",..: NA 3 3 NA NA NA 2 3 NA NA ...

$ Submitted.via : Factor w/ 6 levels "Email","Fax",..: 5 6 6 6 6 3 6 6 5 6 ...

$ Date.sent.to.company : Factor w/ 2274 levels "1/1/2013","1/1/2014",..: 1034 369 285 1566 2129 465 1621 1597 583 289 ...

$ Company.response.to.consumer: Factor w/ 9 levels "","Closed","Closed with explanation",..: 3 3 3 3 3 4 3 3 3 3 ...

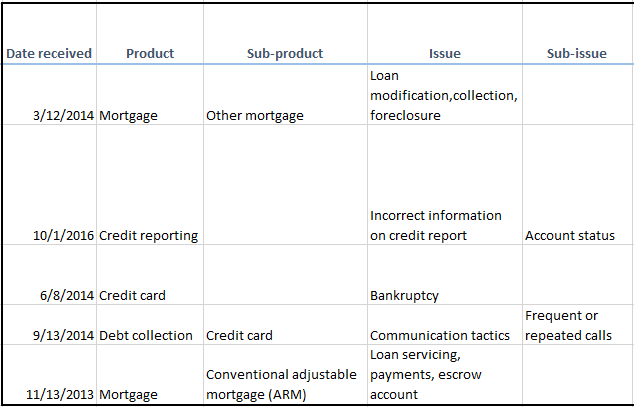
$ Timely.response. : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

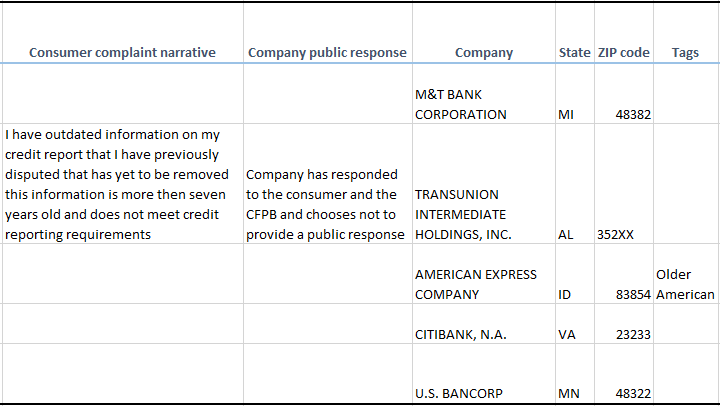
$ Consumer.disputed. : Factor w/ 3 levels "N/A","No","Yes": 2 2 2 3 3 2 2 2 3 2 ...

$ Complaint.ID : int 759217 2141773 2163100 885638 1027760 596562 1422680 1420702 1654890 1079567 ...

FIgure 2:

\* Note: The second record was left out of the sample due to a long description. Data split into 3 segments for easier viewing





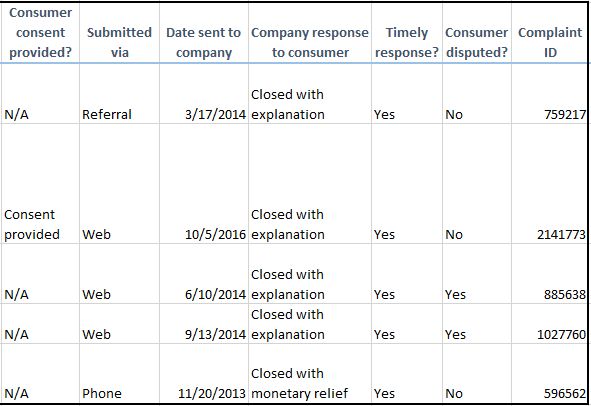
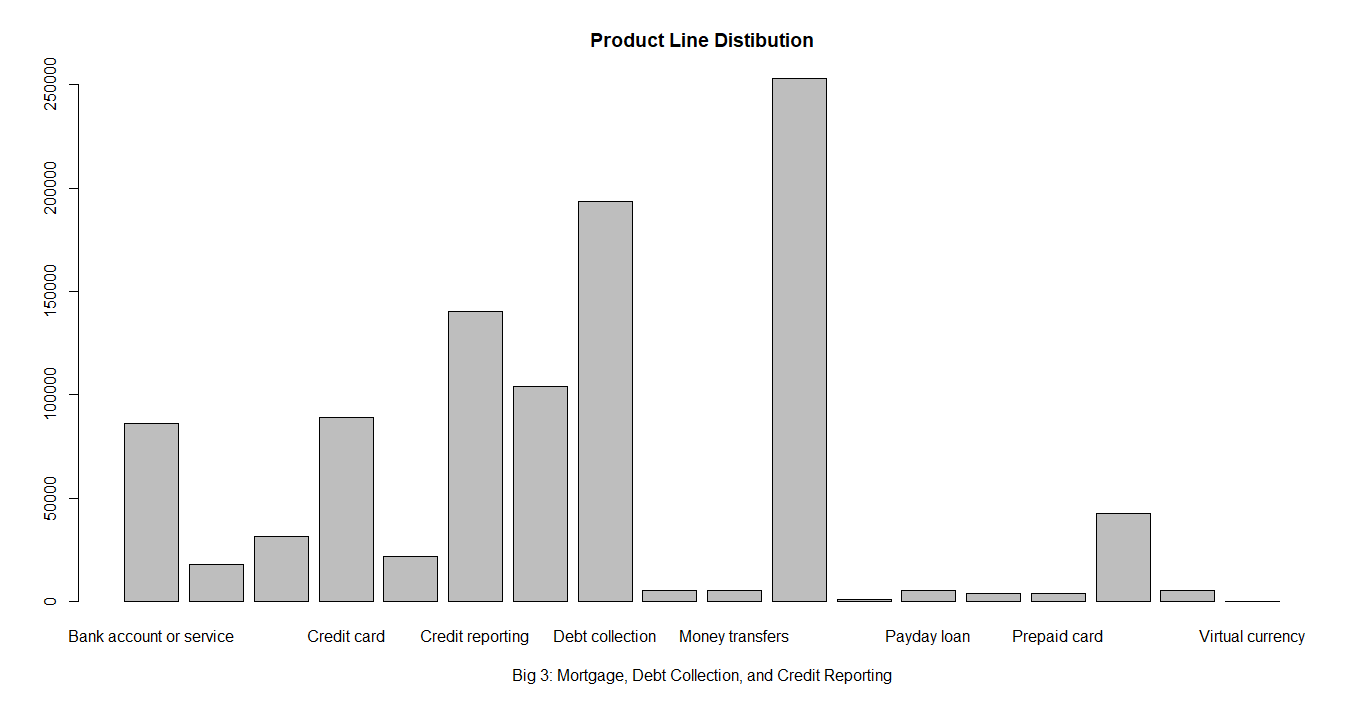
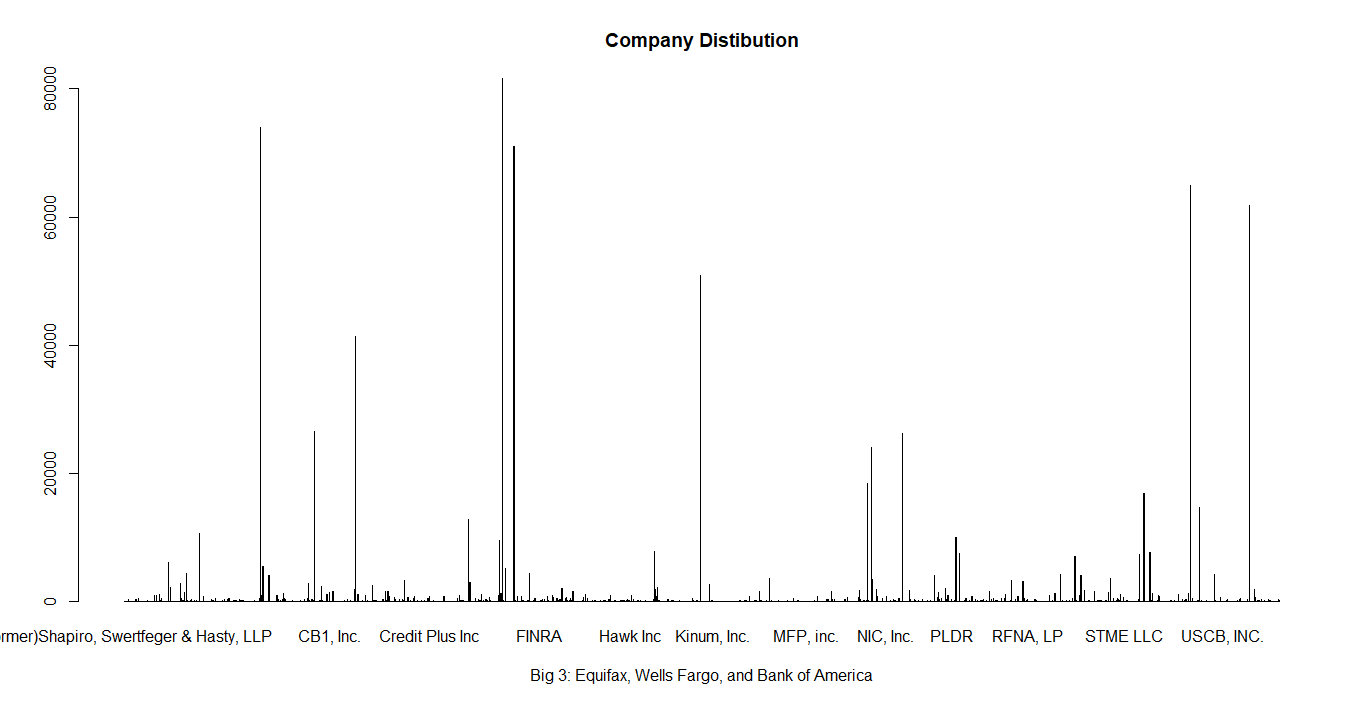


Figure 3

Distributions of key variables:





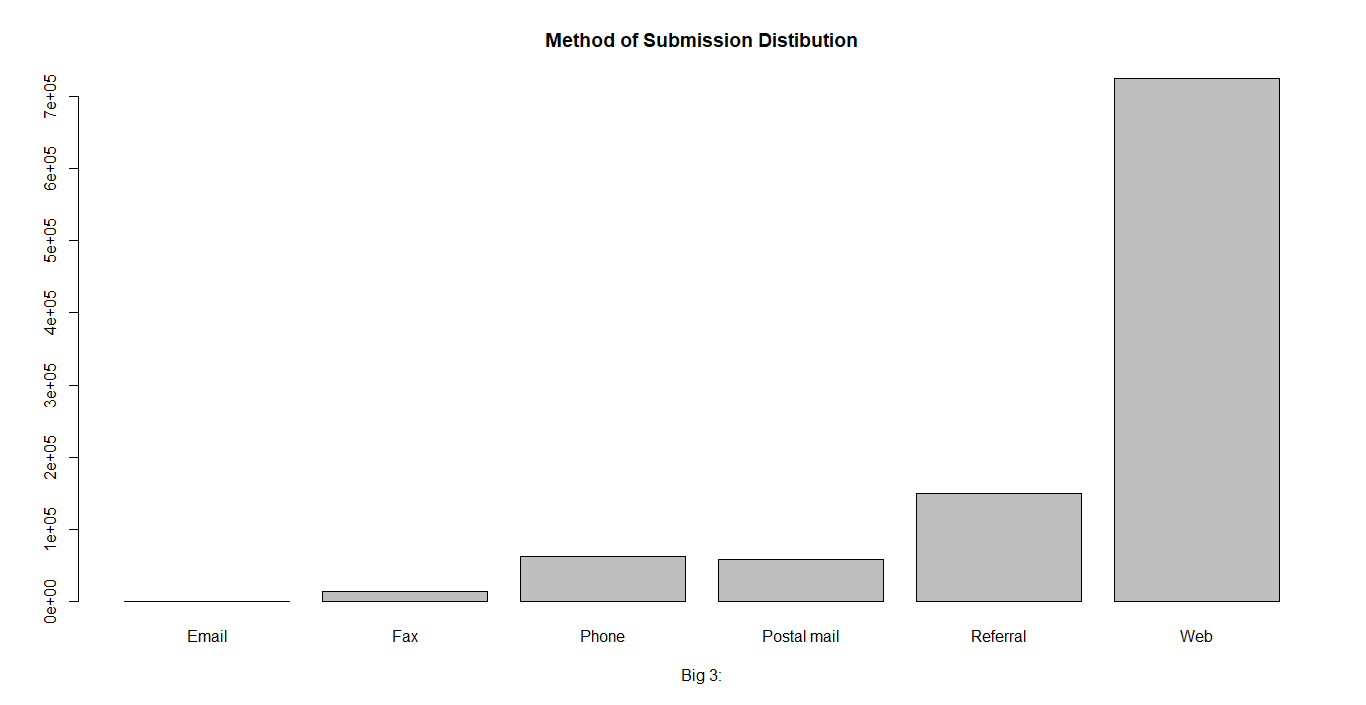
****

Figure 4

This graph shows a breakdown of the narrative variable, the dark notes where there was not a narrative complaint, the white is that there is a narrative available. It also reveals the number of distinct issues.

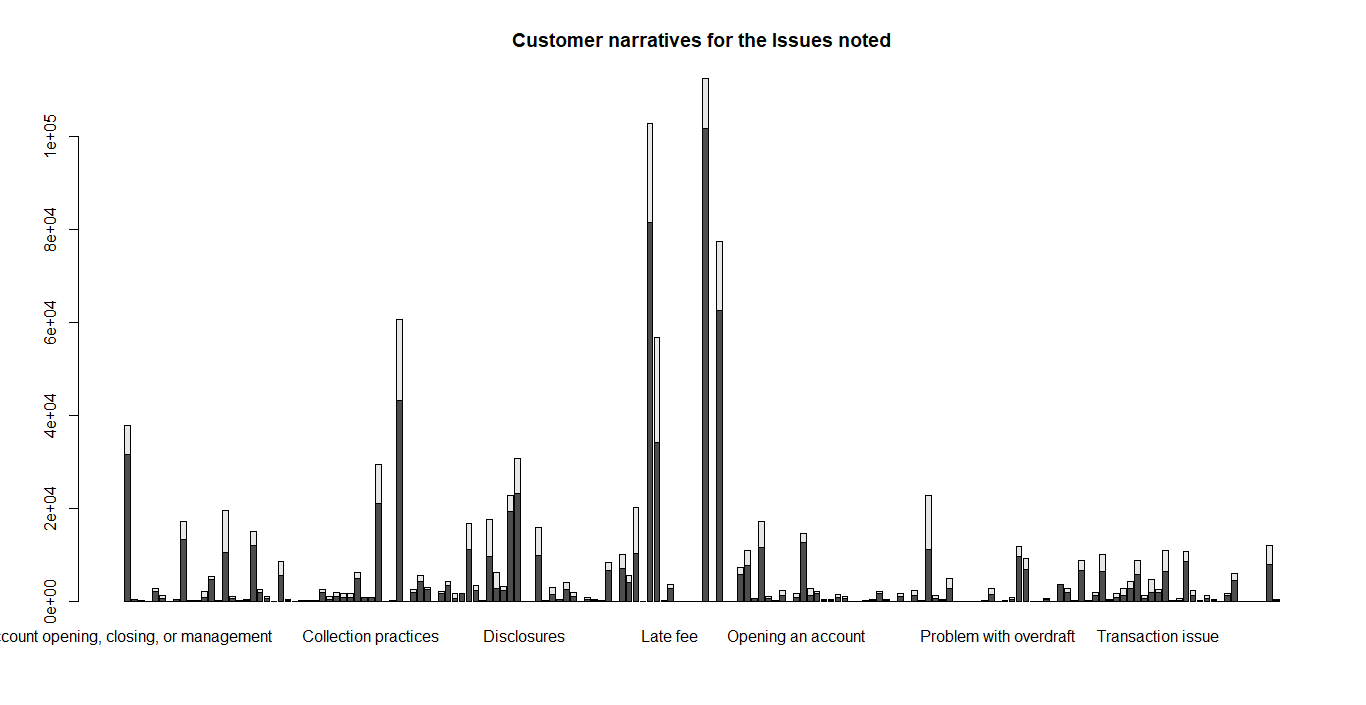


Figure 5

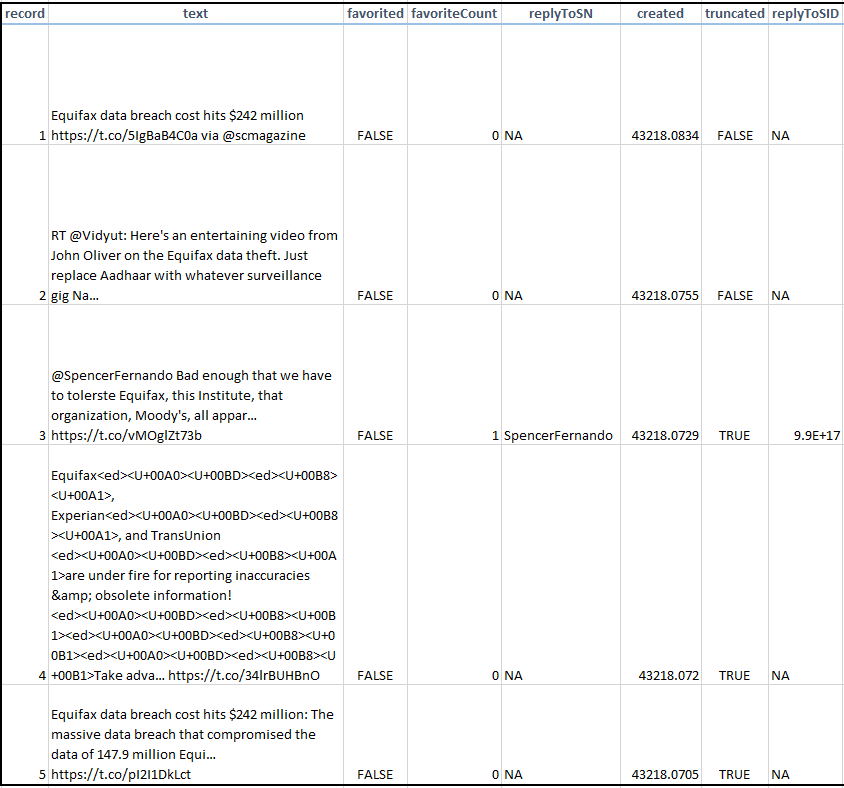




Figure 6

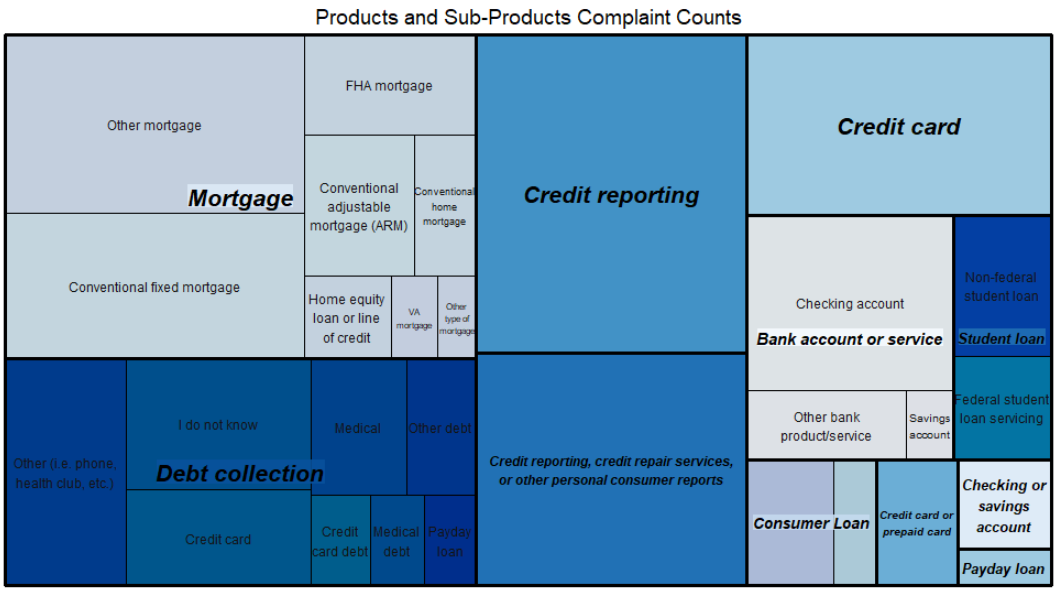


Figure 7

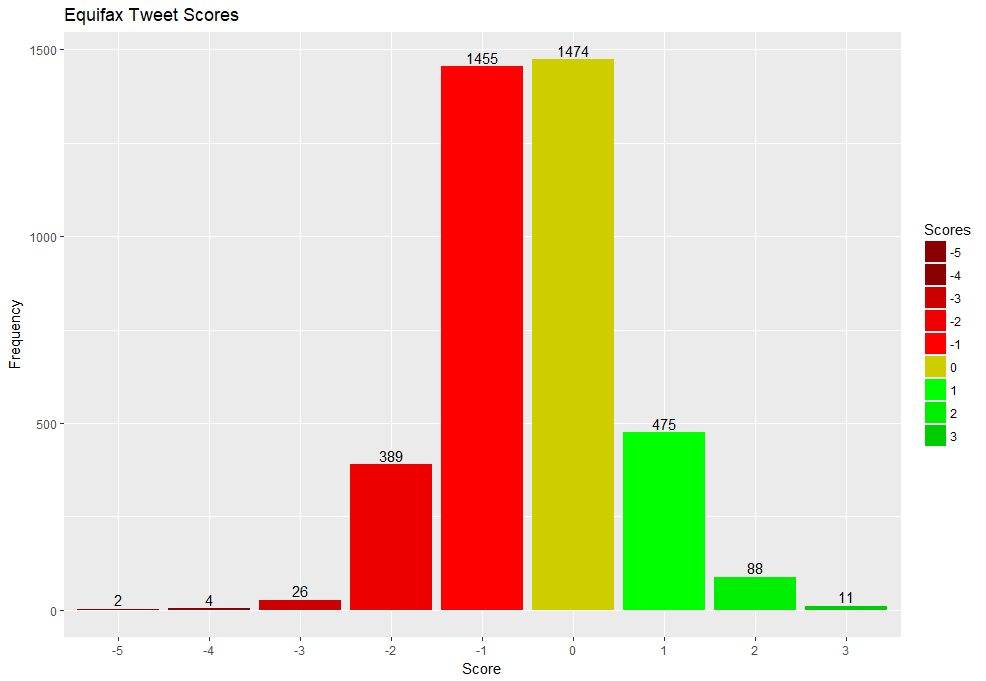


Figure 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable #** | **Definition** | **Derived?** | **Expected Direction of Impact on the Dependent Var** | **Notes** |
| ***Var 1 (Company Public Response)*** | **Did Company publicly respond?** | **Derived** | **+** | **If the company was forced to respond publicly, the customer might have still be angry with the explanation** |
| ***Var 2 (Tags)*** | **Was there a tag?** | **Derived** | **+** | **This takes into account Tag vs Blank Tag. Looking at data, if the complaint has a tag attached, it is more likely to be disputed** |
| ***Var 3*** | **Consumer consent provided?** | **Original var** |  | **Sub Variables are potential responses** |
| ***Var 3.1*** | **Provided** | **Derived** | **+** | **They want other people to see this company in a negative light** |
| ***Var 3.2*** | **Not Provided** | **Derived** | **-** | **They don't want to show this company in a negative light** |
| ***Var 3.3*** | **Withdrawn** | **Derived** | **-** | **If they provided consent then withdrew, they may now be happy with the resolution** |
| ***Var 4*** | **Company response to consumer** | **Original var** |  | **Sub Variables are potential responses** |
| ***Var 4.1*** | **Closed** | **Derived** | **+** | **Some kind of relief/explanation is better than none** |
| ***Var 4.2*** | **Closed with explanation** | **Derived** | **-** | **Some kind of relief/explanation is better than none** |
| ***Var 4.3*** | **Closed with monetary relief** | **Derived** | **-** | **Some kind of relief/explanation is better than none** |
| ***Var 4.4*** | **Closed with non-monetary relief** | **Derived** | **-** | **Some kind of relief/explanation is better than none** |
| ***Var 4.5*** | **Closed with relief** | **Derived** | **-** | **Some kind of relief/explanation is better than none** |
| ***Var 4.6*** | **Closed without relief** | **Derived** | **+** | **Some kind of relief/explanation is better than none** |
| ***Var 4.7*** | **In progress** | **Derived** | **N/A** | **There is no response for customer disputed of complaint is in progress** |
| ***Var 4.8*** | **Untimely response** | **Derived** | **+** | **If response is timely, less likely customer will dispute** |
| ***Var 5*** | **Timely response?** | **Original var** | **-** | **If response is timely, less likely customer will dispute** |
| ***Var 6 (Date received)*** | **Was the complaint received on a weekend or weekday?** | **Derived** | **Yet to be determined** | **Day of the week was extracted from the ‘Date Received’ Variable then grouped by weekend or weekday. We will see if one has a larger effect on the target than the other** |
| ***Var 7 (Date Received)*** | **Quarter** | **Derived** | **Yet to be determined** | **Quarter was extracted from the ‘Date Received’ Variable. We will see if one has a larger effect on the target than the other** |
| ***Var 8 (Date Sent to Company - Date received)*** | **Length of time between customer submitted to company received** | **Derived** | **+** | **The longer a customer has to wait to get a response, the more frustrated they may get and less likely they are to accept response** |
| ***Var 9 (State)*** | **Geographical Zone** | **Derived** | **Yet to be determined** | **The State variable was grouped into an East, Central, West, or Other Zone. We will see if one Zone has a larger effect on the target than the others** |
| ***Var 10 (Consumer Narrative)*** | **Did the Consumer provide a Narrative?** | **Derived** | **+** | **If they provided a narrative they may be more invested in the complaint and less likely to let it go** |
| ***Var 11 (Product)*** | **Dummy variable was created for each of the 18 products** | **Original Variable** | **Yet to be determined** | **We will see if one product has a larger effect on the target than the others** |
| ***Var 12 (Submitted Via)*** | **Dummy variable was created for each of the 6 methods** | **Original Variable** | **Yet to be determined** | **We will see if one method has a larger effect on the target than the others** |

Figure 9

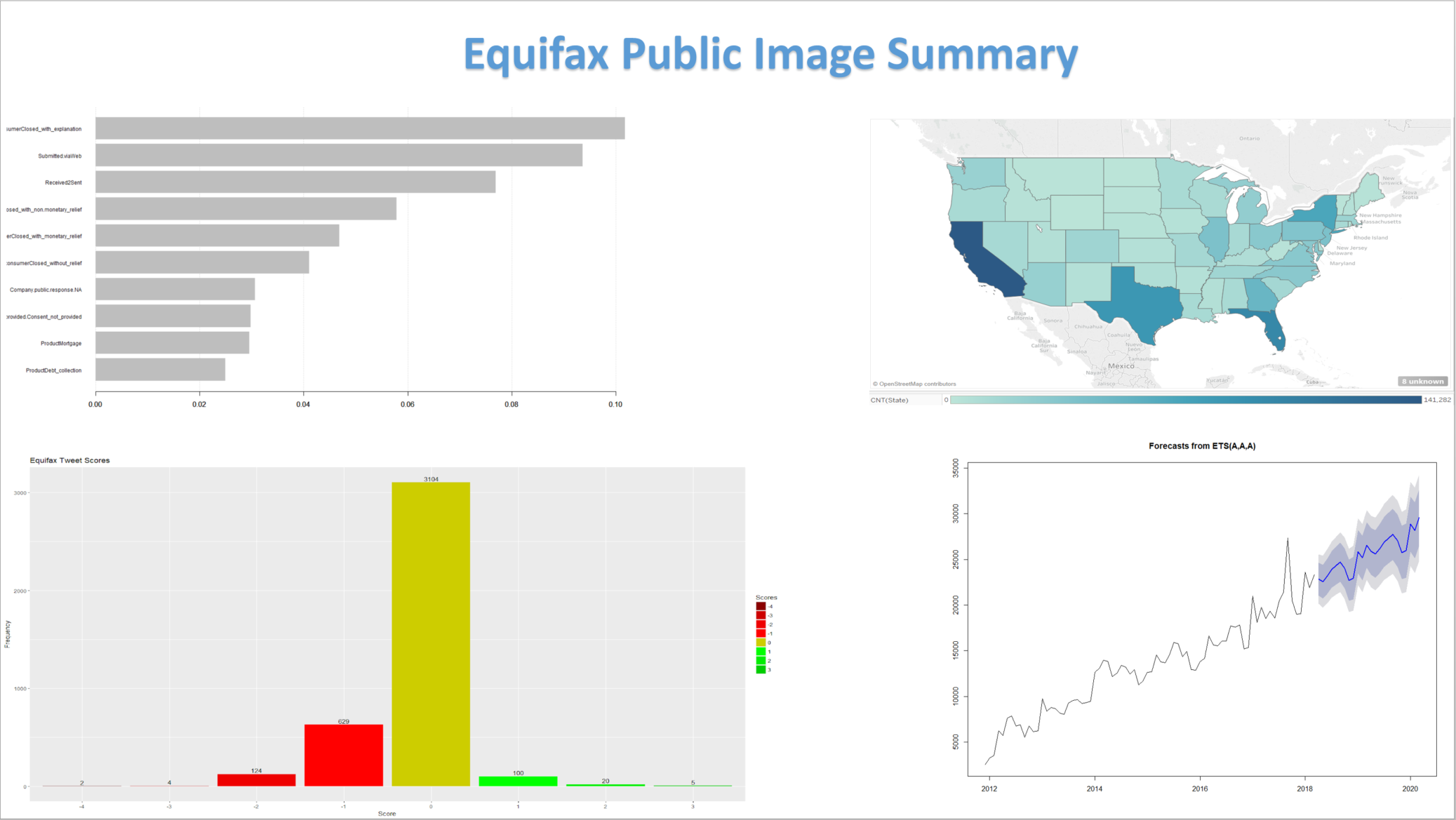


Figure 10

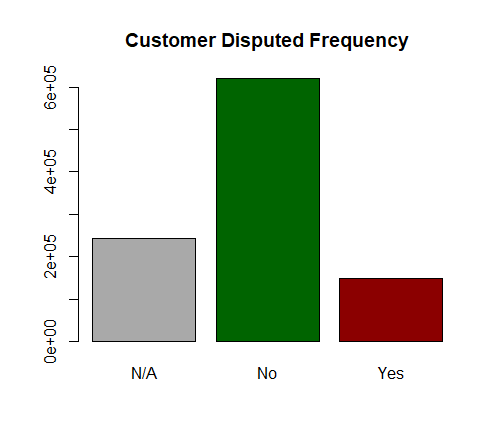


Figure 11

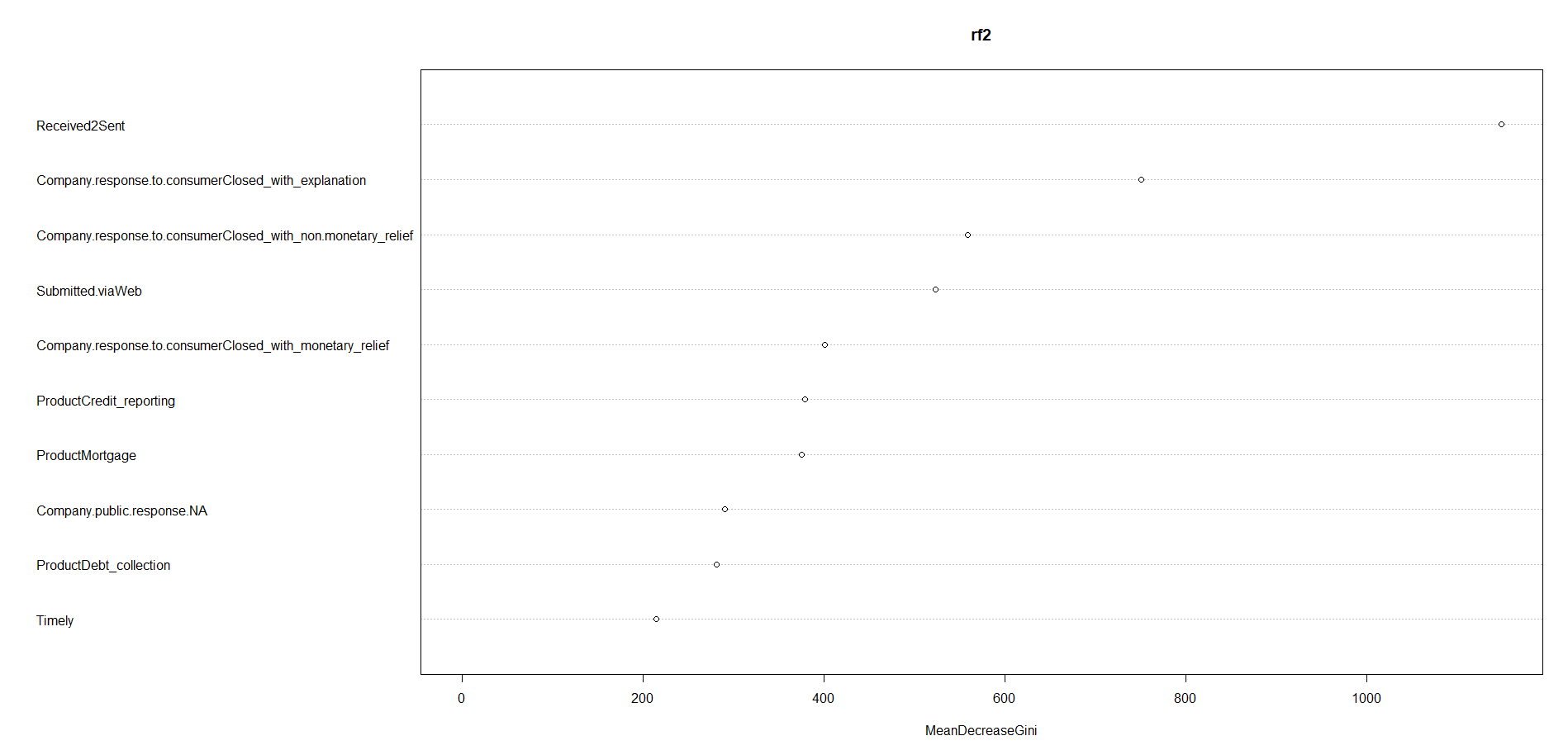


Figure 12

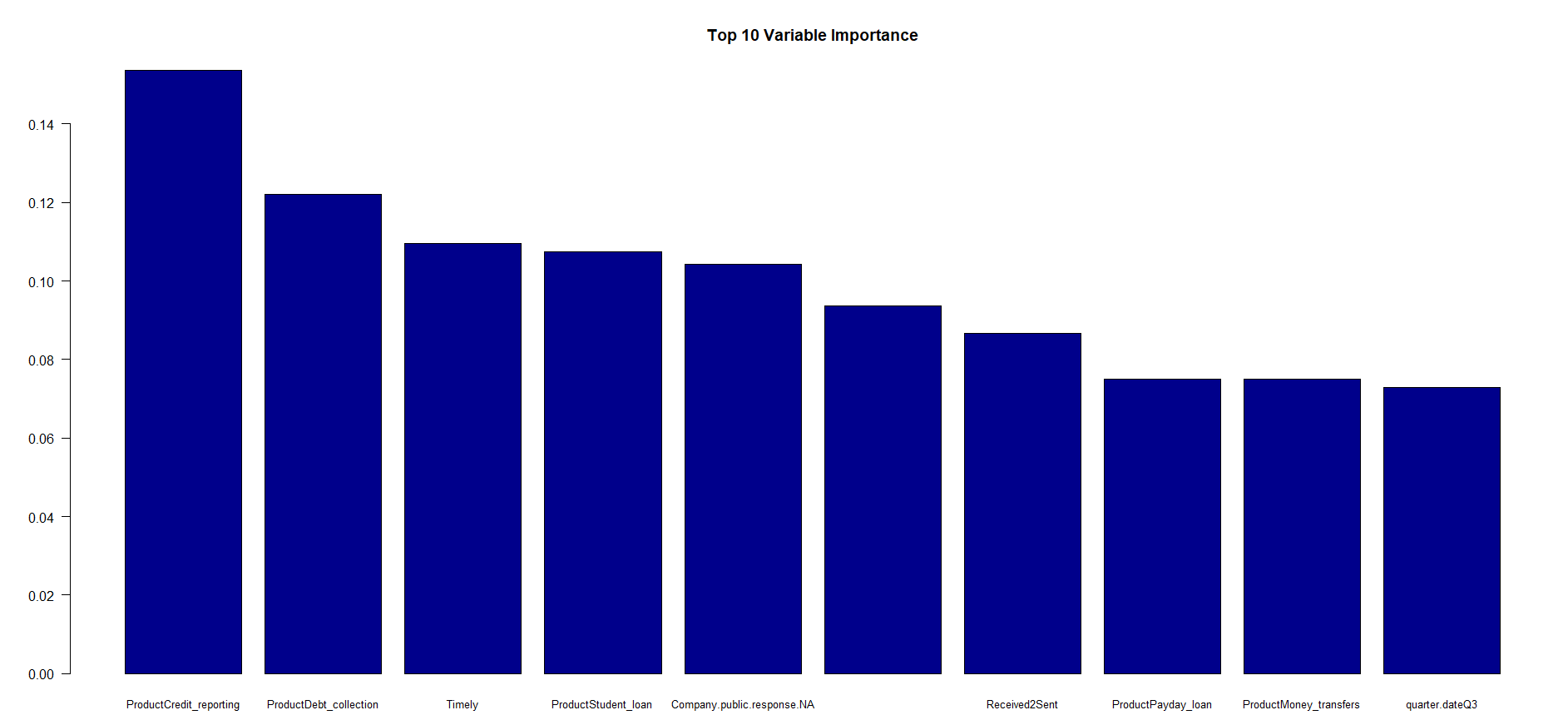


Figure 13

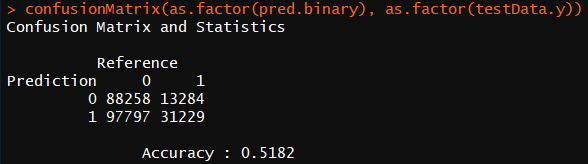


Figure 14

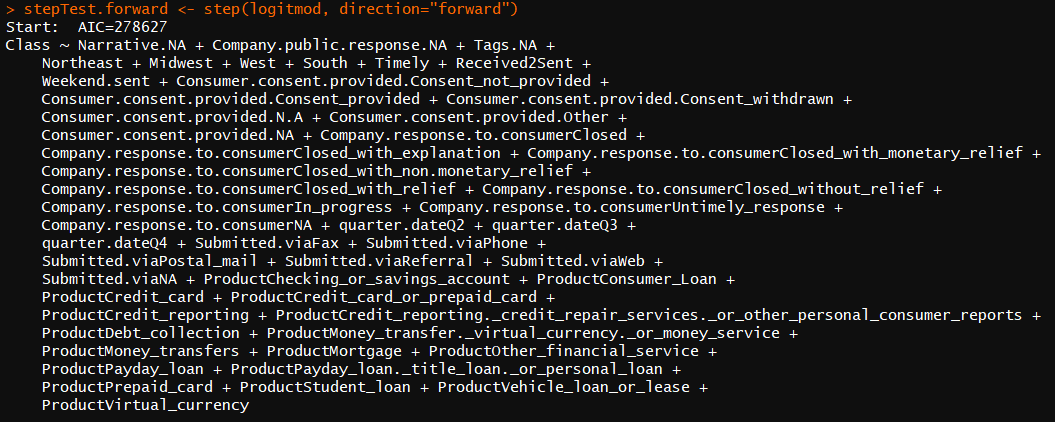
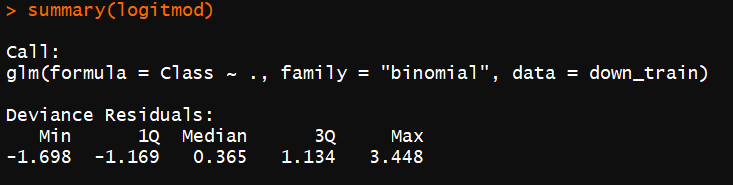


Figure 15



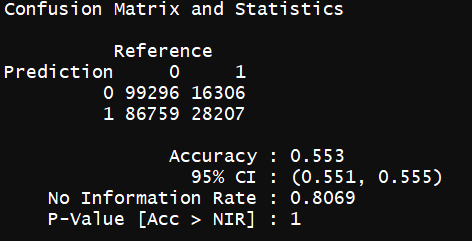


Figure 16

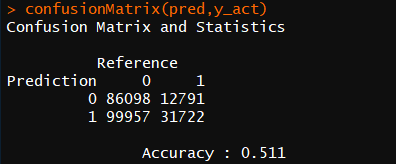
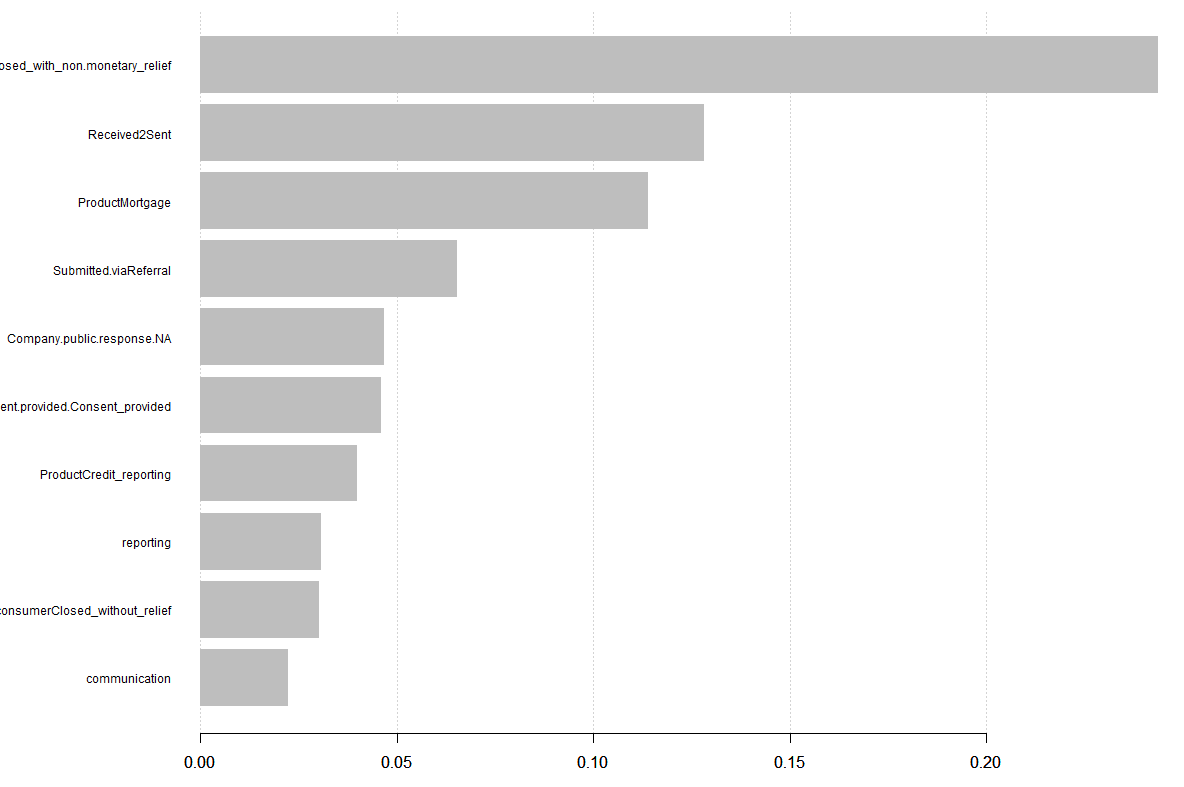


Figure 17



# **SDaIR and Project Plan**

The scope of this project is limited to the creation of a Solution Development and Implementation Report and accompanying dashboard. The period of development of the report and dashboard beginning on April 2, 2018 through formal submission of the report and dashboard on June 3, 2018 is being referred to as Phase 1. Any implementation and/or training requirements following this period will be referred to as Phase 2 and falls outside of the initial scope. The project team has chosen to implement an agile approach to complete this project. An agile approach will better allow for the needed flexibility given limited resources and a short duration for project completion.

**Project Deliverables**

Major project deliverables have been identified and include (1) Solution Development and Implementation Report, (2) Solution Development and Implementation Report presentation, (3) Native files of supporting documents, and (4) Dashboard.

1. Solution Development and Implementation Report - The Solution Development and Implementation Report will be a formal report and structured to provide needed context for project deliverables and include the following:
   1. QA report
   2. Variable list/data dictionary attachments
   3. Variable reduction and list of high value variables
   4. Methodology visuals and reports
   5. Description of how it will be implemented.
2. Solution Development and Implementation Report presentation - a video presentation and associated slide deck will be provided for the analytics project.
3. Native files of supporting documents - native files of associated code, figures, tables, etc. will be provided.
4. Dashboard - an interactive dashboard solution sample will be provided. Major components of the dashboard will include:
   1. KPI summaries
   2. Heat maps showing complaint counts by state and zip code
   3. Time Series showing historical complaint trends as well as future forecasts

**Project Requirements**

Project requires the formal submission of the Solution Development and Implementation Report and dashboard on or by June 3, 2018. Project requirements also include periodic progress reports along with a video presentation. The video presentation is required on or by June 3, 2018 and will demonstrate the dashboard, provide an explanation of the proposed implementation, and include an associated slide deck.

**Project Constraints**

Project constraints are largely driven by the location of the project team members and the time sensitivity associated with completion of the phase 1 deliverables by June 20, 2018. The project team members are located in separate geographic area. Also, all project team members have responsibilities associated with other projects. The project team will be constrained to virtual meetings and will have to coordinate interaction in such a manner that supports both this project and responsibilities outside of this project.

Another constraint we may run into is access to data. Currently we are using a free data source provided by the Consumer Financial Protection Bureau, so we are relying on their updates and accuracy for our main data source. We also may need to purchase historical social media data in order to provide a better overall picture of a company’s public image.

**Project Assumptions**

The project assumes the availability of project team members to coordinate sufficiently for submission of required deliverables. Other project assumptions include the following:

* Quality of dataset provided by the CFPB is sufficient to produce dashboard and associated models, KPIs, etc.
* No mitigation to changes in project team members during phase 1 are provided for in the project plan
* Members of the phase 1 project team are not expected to be available for any post-phase 1 activities such as testing, training, implementation, etc.
* No budget constraints beyond what is assumed in the Cost section of the project plan are assumed
* Sufficient expertise relating to applicable programming languages and platform solutions will be available for any phase 2 implementation requirements.
* A company is assumed to have sufficient and accurate information contained within the dataset to effectively populate associated metrics, graphs, etc.

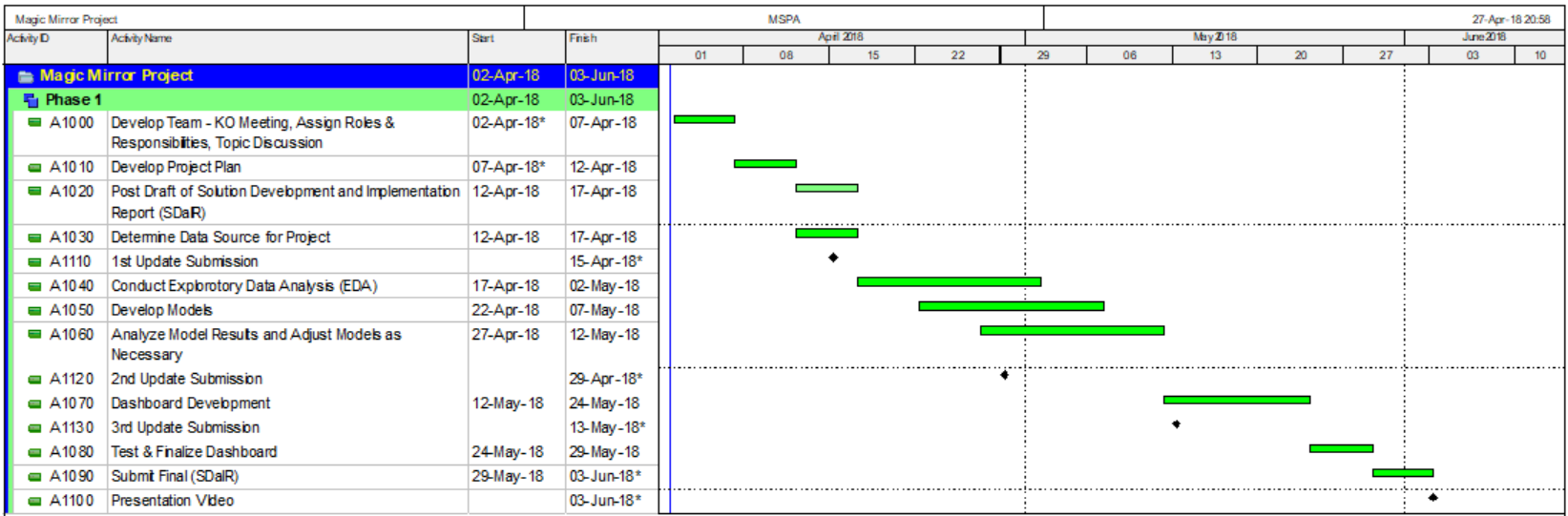
**Project Risks**

Major project for phase 1 development risks include but are not limited to the following:

* Additional analytics may be needed in the dashboard
* Limited availability of key resource
* Test facilities not available for dashboard deployment
* Potential quality issues with dataset

**Gantt Chart**

The below Gantt chart represents a preliminary Level 2 project schedule expressed in ordinal dates for phase 1 activities. The project schedule is consists of a 10 week period beginning on April 2, 2018 and completing with the required final submissions on June 3, 2018

****

**Cost**

A rough order of magnitude (ROM) estimate for the phase 1 activities of this project are projected to be approximately $70,000. This estimate reflects an accuracy range of -25% to +75% as defined by the Project Management Institute (PMI) for ROM estimates. The estimate equates to approximately 600 hours total to be proportioned among the three (3) project team members at an average billing rate of $100 per hour along with a $10,000 allowance for miscellaneous costs.

**Resources**

Project resources for phase 1 include three (3) project team members consisting of one (1) project lead and (2) supporting analysts. We each employed our own hardware and computational programs and software.