Justin Tsao

Milestone Report

This project will focus on the customer service side of the finance industry. The problem I want to solve is how to increase customer service given the vast amount of data on customer complaints. From experience, I can tell you that there are a lot of complaints from the service of a firm to the products themselves. The clients for this project can be all financial firms who want to improve their responsiveness of their customer service.

The data is coming directly form the database of the Consumer Financial Protection Bureau, or CFPB. The data the project will be working can easily be downloaded and queried here https://data.world/cfpb/consumer-complaints. As of 5/20/2019, the data set is comprised of 1,284,185 individual rows of complaints, updated frequently. I plan to conduct visual and statistical analysis on the entire data set which includes 18 columns. From the data set, I only plan to use the complaints where the customers give a written response. The reason being, I want to use machine learning and natural language processing techniques on the written customer complaints as well as the categorical columns. Taking only the data with written customer complaints, leaves 383,840 rows of data.

To decide whether a complaint is handle in an ‘appropriate’ manner I will use whether the complaint was handled in a timely manner or not as a predictor. I use appropriate in quotes because the term ‘appropriate’ can be ambiguous but I assume it means that the complaint was handled as efficiently as it could. Not responding in time could mean the complaint was not resolved, difficult to handle, or maybe ignored due to lack of concern

about the customer's complaint. The goal will be a machine learning model that can correctly identify these non-appropriate response times based on the features of data and the customer complaint. The results could be extremely valuable by giving insights on which complaints are the most difficult to handle. Being able to quickly identify the difficult complaints can be leveraged to improved customer service and satisfaction which will lead to higher customer retention.

Data

As mentioned above, before any pre-processing and wrangling, the data contained 1,284,185 rows of data. The full data set was used for visual data analysis and inferential statistics. A SQL query was performed to only return 383,840 rows of data which had a consumer complaint narrative. The consumer complaint narrative contains a response about the complaint they filed. This narrative will be an important feature for machine learning using natural language processing (NLP). There were 7 columns that contained missing values. Of those seven, two of the features were removed because they comprised of data that would be given after a complaint is filed. This means the two features, 'company_public_response' and 'company_response_to_consumer', would qualify for features to us for predictions. The other columns with missing values were 'sub_product', 'sub_issue', 'state', 'zip_code', 'tags', and 'consumer_disputed'. After examining the states feature, I decided to remove the state and zip code data as I did not think it will play a relevant role as a predictor, larger state with big cities, and higher populations had larger proportion of complaints. This could throw off predictions. For the sub issues and sub categories, the missing values were replaced with the string 'None'. This made sense because not all categories of product and issues will have a

sub category. The other columns had over 100k to 300k missing values. I decided to remove those completely without filling in data because they were categorical. The cleaning for those labels was fairly simple.

Most of the cleaning was done to the customer complaint narrative. The first thing to note about the data in this column was that sensitive information like ID numbers or phone numbers were replaced by 'X's. Also, monetary values were encased in { }. The narrative data has to be pre-processed for use in NLP and machine learning. All punctuation, '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~', was removed. The 'X' values were removed as well as any extra white spaces. Since the narratives were typed in by customers there are chances of typos. One way to ease the number of types was to take care of word lengthening. An example of word lengthening would be 'realllyyyy' as way for the consumer to emphasize a point. A function was applied to the column to transform those words into a correct form, 'realllyyy' –> 'really'.

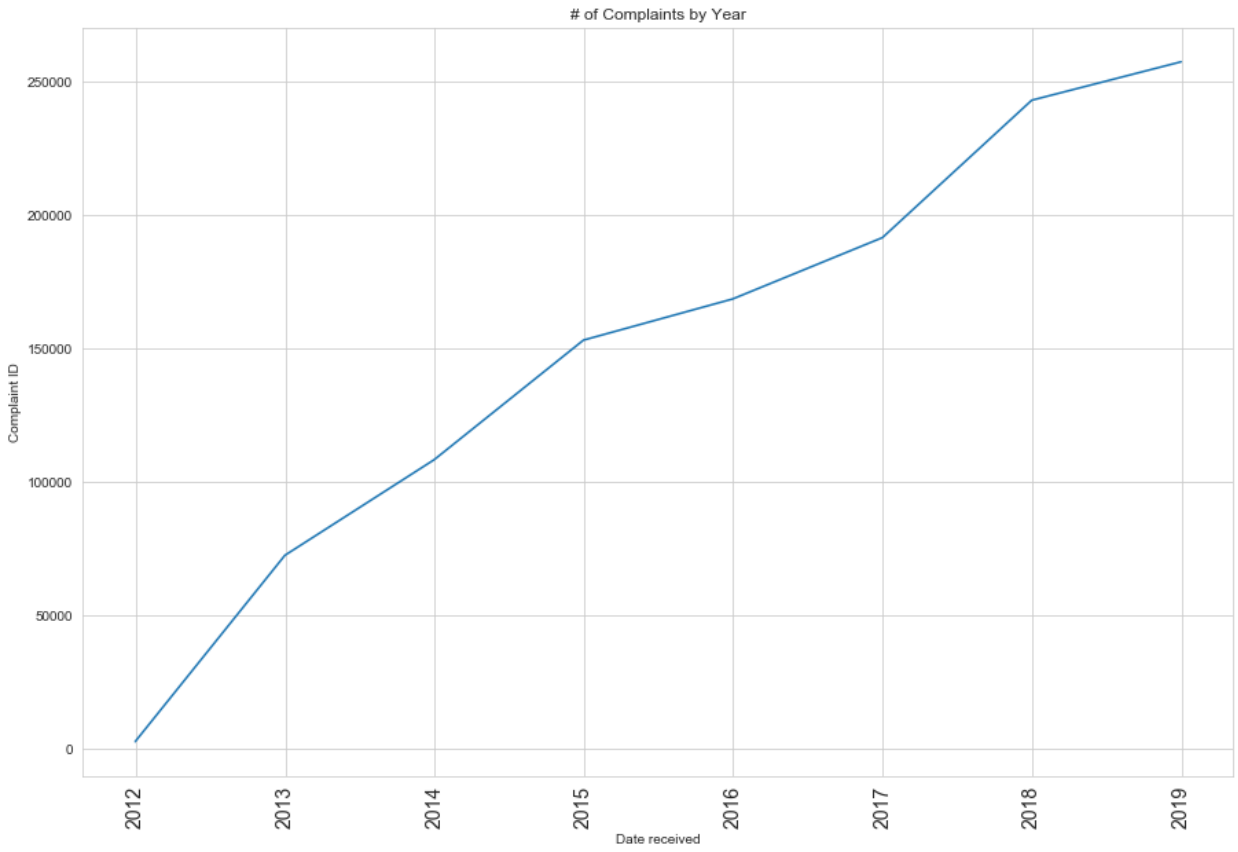
For the following data pre-processing steps on consumer narrative, I have decided that it may easier to implement in Scikit-learn. Tokenization needs to separate each word into a list of words. Next, we can remove stop words like 'is' or 'of' that give little meaning to the narrative. We can also decide to stem and lemmatize words. Stemming is essentially removing suffixes like '-ly' or '-ing' from words. Lemmatization is changing words into their root form. For instance, 'ran' -> 'run' and 'caring' -> 'care'.

As of right now, I have decided on a feature set to include [product, sub_product, issue, sub_issue, consumer_complaint_narrative, company, submitted_via]. The predictor label will be 'timely_response'.

Justin Tsao

Exploratory Data Analysis

The first thing I explored was trend of number of complaints per year. Also, keep in mind these number reflect the number of complaints submitted to the CFPB, there could easily be more complaints unsubmitted, or submitted directly to the financial firms themselves. I actuality, the number plotted below could be many times higher. Another thing to note is that most of data is categorical.



From 2013 to 2014 the complaints submitted to the CFPB increased by 149.53%
From 2014 to 2015 the complaints submitted to the CFPB increased by 141.42%
From 2015 to 2016 the complaints submitted to the CFPB increased by 110.09%
From 2016 to 2017 the complaints submitted to the CFPB increased by 113.64%
From 2017 to 2018 the complaints submitted to the CFPB increased by 126.90%
From 2018 to 2019 the complaints submitted to the CFPB increased by 105.93%

As you can see The CPFB takes in complaints from consumers and sends the complaints to firm the complaint was filed against. What we can infer from this is that the number of complaints has at least double each year. This could be due to increased adoption of technology and the ease to report complaints. Worse yet, we could also infer that customer service could be falling or that consumers are increasingly unhappy with their products. Most of the complaints come from products regarding mortgages, debt, and credit reporting.

There were a total of 18 categories of products:

Mortgage: 278249

Debt collection: 245218

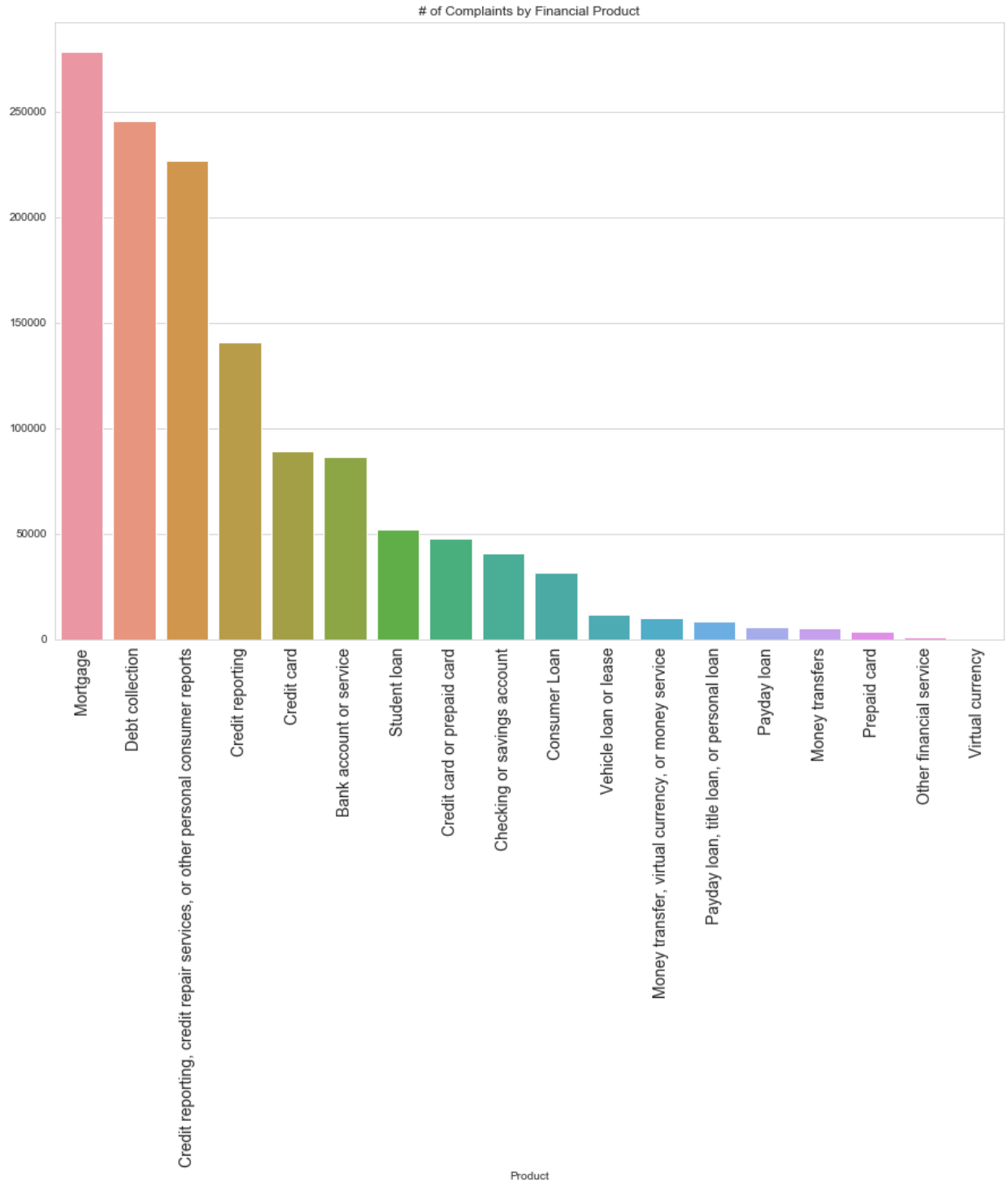
Credit reporting, credit repair services, or other personal consumer reports: 226781

Credit reporting: 140432

Credit card: 89190

Bank account or service: 86206

Justin Tsao


# of Complaints by Financial Product
250000
200000
150000
100000
50000
0
Mortgage
Debt collection
Credit reporting, credit repair services, or other personal consumer reports
Credit reporting
Credit card
Bank account or service
Student loan
Credit card or prepaid card
Checking or savings account
Consumer Loan
Vehicle loan or lease
Money transfer, virtual currency, or money service
Payday loan, title loan, or personal loan
Payday loan
Money transfers
Prepaid card
Other financial service
Virtual currency
Product

There are 76 categories of sub products. Other than credit related products (debt, loans, credit cards etc.), checking accounts stand out as the 2nd highest sub products with complaints against it.

The plot below shows the top issues resulting from those complaints.

Issue

Incorrect information on your report: 134809

Loan modification, collection, foreclosure: 112311

Incorrect information on credit report: 102686

Loan servicing, payments, escrow account: 77333

Cont'd attempts collect debt not owed: 60687

Problem with a credit reporting company's investigation into an existing problem: 51498

Attempts to collect debt not owed: 43181

Account opening, closing, or management: 37961

Communication tactics: 35449

Improper use of your report: 33441

Disclosure verification of debt: 30800

Managing an account: 25535

Written notification about debt: 23766

Trouble during payment process: 23188

Deposits and withdrawals: 22851

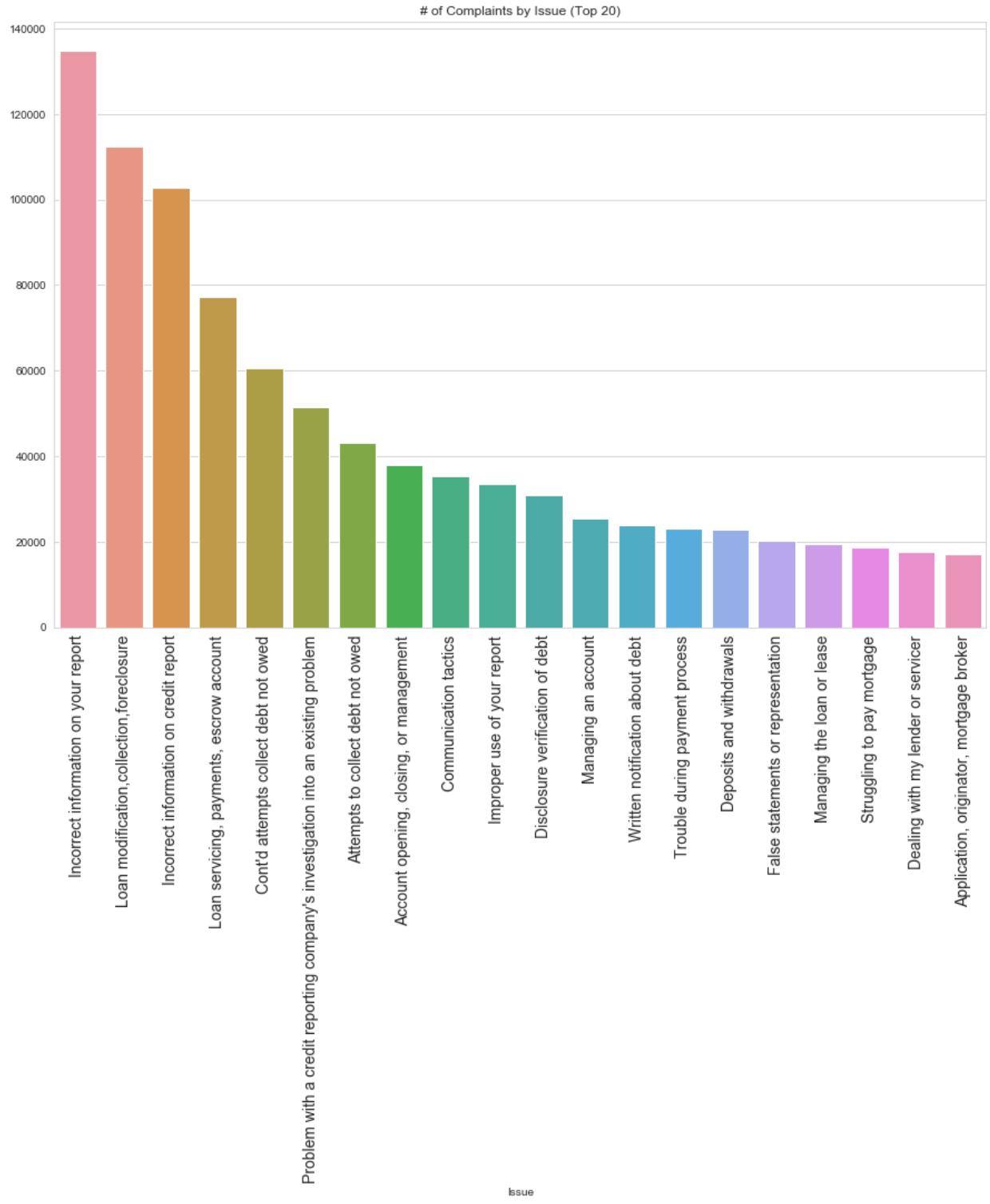
False statements or representation: 20278

Managing the loan or lease: 19438

Struggling to pay mortgage: 18682

Dealing with my lender or servicer: 17630

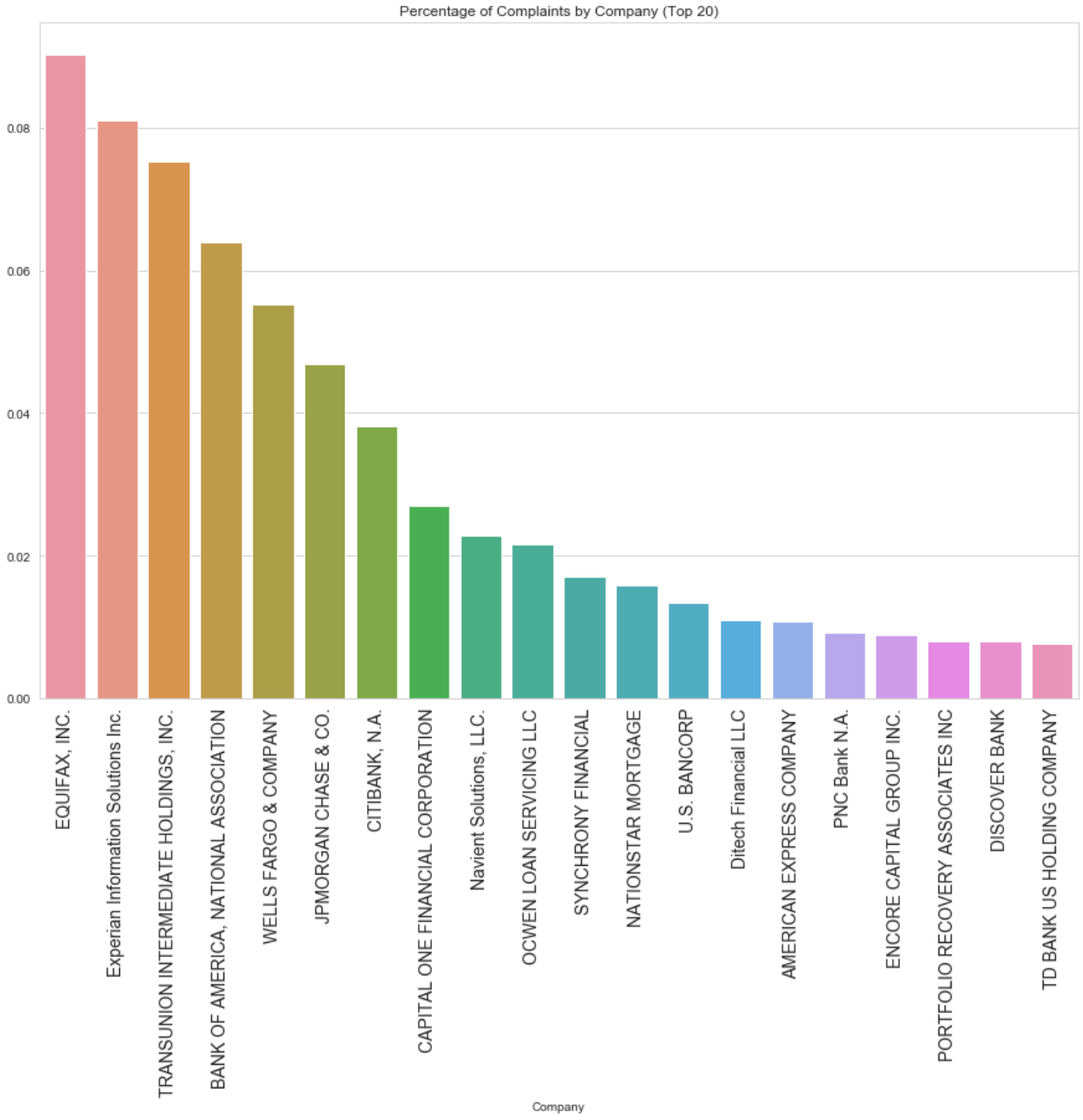
Application, originator, mortgage broker: 17229



Many of these issues seem to stem from how consumer information is handled and reported. Perhaps this is a red flag that employees need to be trained to handle

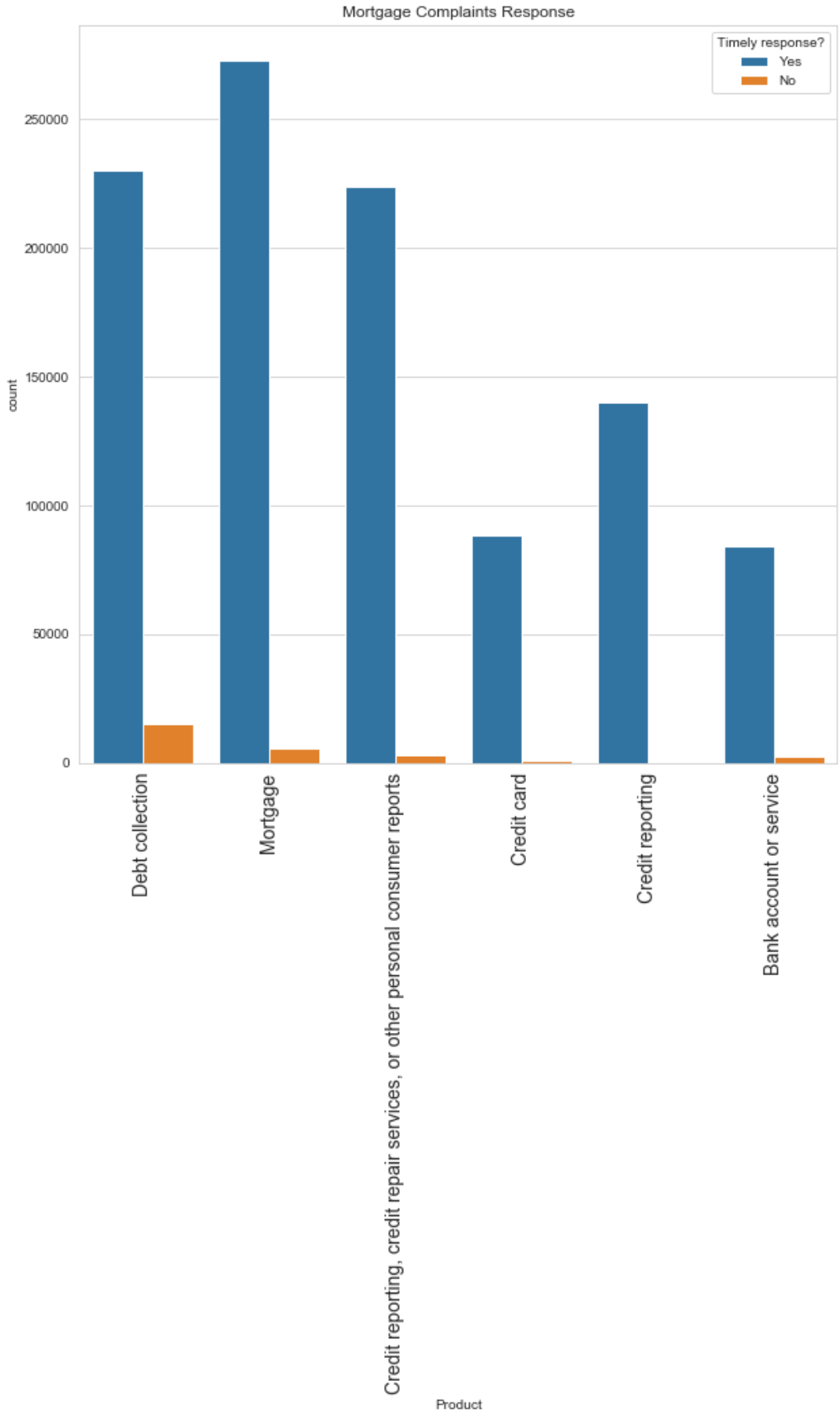
consumer information with better care both in use and when entering in data. From my experience working both at a bank and a mortgage lender, there are plenty of applications taken on a daily basis. Much of the information entered is sensitive information like ID numbers, SSN's, DOB, etc. Not properly communicating a product is also seen a lot in finance. At least in retail, there is always the rush of getting the next client in and out of the seat as fast as possible. After all, more sales means more commissions.

Below is a plot of financial firms and number of complaints. As a past financial professional, the graph above is not surprising at all. The biggest culprits are the credit reporting agencies and big banks. With the larger firms, it does make sense that there would be more over all complaints due to the size of their customer base. It is also harder to manage customer service and sales reps as a company becomes larger and the customer base increases.



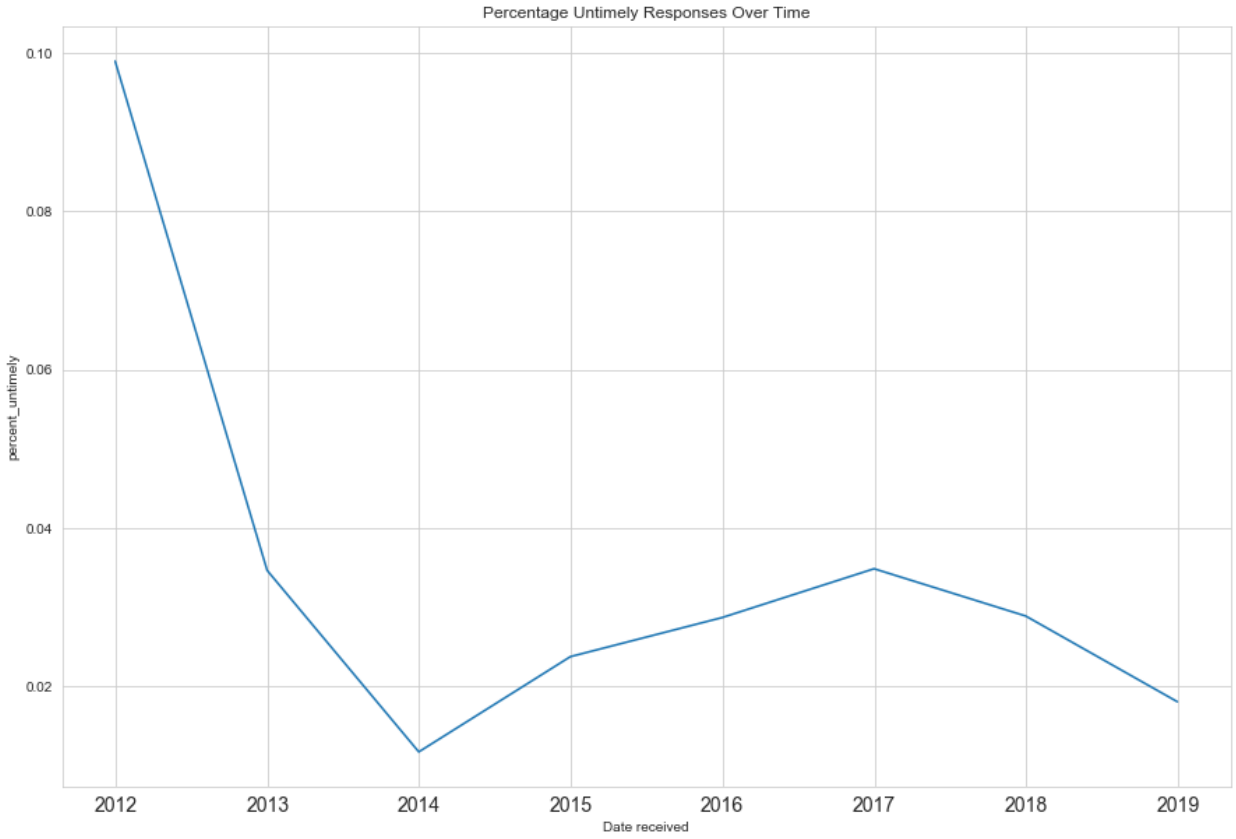
Looking at the full data, there were 1,251,987 responses which were answered in a timely manner. 32,198 were not handled appropriately. That number seems pretty low, but just because a complaint was handled in a timely manner does not mean it was handled in the best way. Also, 2.51% is reflected of complaints reported to the CPFB and is still around 32198 customers.

Justin Tsao


Mortgage Complaints Response
Timely response?
Yes
No
250000
200000
150000
100000
50000
0
count
Debt collection
Mortgage
Credit reporting, credit repair services, or other personal consumer reports
Credit card
Credit reporting
Bank account or service
Product

```
Percentage of untimely responses
Debt Collection: 6.22%
Bank Account: 2.58%
Mortgage: 1.99%
Credit Services: 1.41%
Credit Card: 1.10%
Credit Reporting: 0.21%
```

Though most of the complaints were against credit reporting/services products, the companies responded in a timely manner relative to companies responding to complaints about debt collection and bank accounts.



Based on CFPB's data, untimely responses have started to decrease over all from 2017 for financial firms over all. The last thing I want to remind readers, is that the numbers represent only complaints filed to the CFPB. Therefore, the actual numbers are higher. But the decrease in untimely responses is a good sign! Think about how many of these

claims are mortgage or credit based. Errors in customer service could be millions of dollars in combined losses for consumers and companies. Bad credit reporting could prevent customers form getting a loan from their dream house and thus a mortgage firm would lose out on this sale. Even worse, what if an error in customer service caused funds to not be released from a bank account to a consumer. This means bills go unpaid or late. Late fees rack up, and this can spiral out of control. From experience, I can tell you of a horror story about a teller who forgot to check the endorsement on the back of a check for a college student. Because the teller deposited the check without seeing if there was an endorsement, the college student was with out money to pay for books, rent, or food for weeks. They had to wait for the check to be return and re deposited.