Predicting bank mortgage customer complaint counts

Business purpose:

We wish to find some drivers of quarterly counts of complaints from bank mortgage customers. With the advent of social media, it has become more important than ever for businesses to reduce their numbers of complaints, since a large number of complaints on Yelp can have a huge negative impact on the company’s reputation and/or profitability. Although we did not find significant correlation between complaint counts and return on equity or between quarterly complaint counts and return on assets, we did find a more significant correlation (-0.212) between quarterly complaint count and the ratio of a bank’s Goodwill to its fair market value. Moreover, the relationship between complaint counts on social media and a bank’s profitability may be stronger than the relationships we found, since complaints on social media are more visible than the complaint counts in our dataset. By finding some of the banks’ characteristics that drive their numbers of mortgage complaints, we hope to enable banks to take action to prevent some of these complaints. We hypothesize that financial measures related to the quality of the banks’ customers (e.g. amounts past due and amounts of loans charged off) will strongly drive complaint counts.

Data Sources:

* The data set which contains the information pertaining to bank complaints is located at this financial products [complaint data](https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation) website.
* The data set which contains bank financial information is located at this [bank rankings](http://www.usbanklocations.com/bank-rank/) website

Description of raw data:

* The complaint data has 755,042 records and 19 fields. The set is at complaint level. The key fields for this project and next steps are the following:
  + Date complaint received
  + Product (e.g. mortgage, car loan, credit card, etc.)
  + Issue (e.g., “Incorrect information on credit report”)
  + Consumer complaint narrative: A text field with a description by the customer of his complaint
  + Complaint ID: A unique ID number assigned to each complaint
* The bank ranking data contains many tables of bank financial measures, such as equity capital, total deposits, restructured loans, etc. This data is broken up by year and quarter, with one table for every financial measure/year/quarter combination. In addition, each of these tables has the rankings of the banks by the financial measure. For example, for each of the quarters Q4 1992 through Q1 2017, there is a table containing the banks’ total assets and the banks’ ranks by total assets. Typically, the number of banks in one of these tables is around 6000.

Preparation of modeling data from raw data:

1. In order to reduce the scope of this project to a manageable level, we decided to focus on just one product. We found that the product with the most complaints was the mortgage product (see Table 1), so we chose mortgages.

**Table 1: Frequencies of complaints about bank products**

|  |  |
| --- | --- |
| **Product** | **Frequency** |
| Mortgage | 225,023 |
| Debt collection | 142,752 |
| Credit reporting | 136,672 |
| Credit card | 87,864 |
| Bank account or service | 84,971 |
| Student loan | 31,367 |
| Consumer Loan | 30,927 |
| Payday loan | 5,430 |
| Money transfers | 5,264 |
| Prepaid card | 3,736 |
| Other financial service | 1,019 |
| Virtual currency | 17 |

1. We also decided to use only about five years of bank history, from Q4 2011 to Q1 2017. This is to insure that characteristics that change over time (such as government regulations of banks) would remain somewhat constant over the data that we used in our model.
2. We needed to obtain complaint counts per bank/quarter. This seems like a simple matter of counting distinct complaint IDs for every bank/quarter combination. However, this method will not work because there are some bank/quarter combinations that have no mortgage complaints, and hence no complaint ids. Therefore, if we count distinct complaint ids per bank/quarter, we will not record the zero complaint counts. Our method for solving this problem is as follows:
   * 1. Create a set with counts of distinct complaint ids per bank/quarter. As we mentioned, this set will be missing the banks/quarters with zero complaint counts
     2. Create a set called “left set” that contains every bank/quarter combination
     3. Perform a left join of the counts set onto left set, with left set being the left set. Call the resulting set “counts2”.
     4. Replace all missing values of complaint counts in counts2 with 0.
3. We copied and pasted tables of bank names, quarters, equity capital, and bank ranking by equity capital into .csv files. This process had to be done one quarter at a time for 22 quarters.
4. We imported the ranked banks data from step 4 into R and stack the datasets into one set, called “ranked banks”
5. Upon examining the ranked banks data, we found that about 6% of the banks had more than one rank in some quarters. For example, a bank might appear to rank both #11 and #46 in Q1 2015. We eliminated these banks from our data.
6. Next, we had to merge the ranked banks onto the complaints data. This was a challenging task, because the only possible merge key is the name of the bank, and bank names often have variations in spelling. For example, in the complaints data, there is a bank called “BOKF”, whereas this same bank goes by the name “BOK FINANCIAL CORP” in the bank ranking data. At first we attempted fuzzy merging techniques to join the complaint data to the ranking data; however, this proved ineffective, since many pairs of distinct banks have very similar names; for example, “Cit Bank” and “Citibank” are two distinct entities.

So, we used the following approach: First, we selected the very large banks from the bank ranking data. We define “very large” as follows: If a bank ever achieved an equity capital rank of 100 or lower, then it would be included in our set of very large banks. Call this set of very large banks “ranked.banks.top.100”. This set contained 127 distinct banks. The next step was to manually adjust the names of the 127 banks in ranked.banks.top.100 to conform to the banks’ names in the complaint data. For example, if a bank was called “BMO Harris Bank” in ranked.banks.top.100, but that bank was called “BMO HARRIS BANK NATIONAL ASSOCIATION” in the complaint data, then the bank was given the alias of “BMO HARRIS BANK NATIONAL ASSOCIATION” in ranked.banks.top.100.

After this renaming process was complete, it was a straightforward matter to perform an inner join of the complaint data with ranked.banks.top.100, using bank names as the merge key. There were a total of 75 banks in the merged data.

1. Next, we had to normalize the complaint counts by bank size. This is because we would expect a larger bank to have more complaints, simply because it does more business. This normalization, or adjustment, is done as follows: Define a bank’s size as its amount of equity capital. Then construct a linear regression model with complaint count as the predictor and equity capital as the target variable:

Complaint count = α + β × Equity capital

Then

Adjusted complaint count = (Complaint count – α) / β

We normalized many of the other predictor variables in the same way.

1. We cut and pasted datasets containing many other bank financial measures into .csv files (one quarter at a time), imported those files into R, stacked the files to create one dataset for the financial measure, and then merged those sets onto our master dataset which was begun in step 7. The bank measures that we added are:
   1. Number of employees. These were normalized by equity capital.
   2. Return on equity (ROE). ROE measures a corporation's profitability by revealing how much profit a company generates with the money shareholders have invested. Return on Equity = Net Income/Shareholder's Equity
   3. Loan amount past due 30 – 89 days. This was normalized by equity capital.
   4. Loan amount past due 90+ days. This was normalized by equity capital.
   5. Loans restructured in troubled debt restructurings, 80 – 89 days past due, not in compliance with their modified terms. A restructured loan is a new loan that replaces the outstanding balance on an older loan, and is paid over a longer period, usually with a lower installment amount. Loans are commonly rescheduled to accommodate a borrower in financial difficulty and, thus, to avoid a default. It is also called a rescheduled loan. This was normalized by equity capital.
   6. Loans restructured in troubled debt restructurings, 90+ days past due, not in compliance with their modified terms. This was normalized by equity capital.
   7. Percent of net loans charged off. A charge-off is the declaration by a creditor that an amount of debt is unlikely to be collected. This occurs when a consumer becomes severely delinquent on a debt.
   8. Goodwill. This is defined by excess of purchase price over fair market value. We also computed the ratio of Goodwill to fair market value and called it “Goodwill percent”.
   9. Return on assets (ROA). This is an indicator of how profitable a company is relative to its total assets. ROA gives an idea as to how efficient management is at using its assets to generate earnings. It is calculated by dividing a company's annual earnings by its total assets. It is sometimes referred to as "return on investment".
   10. Nonaccrual loans. A nonaccrual loan is a nonperforming loan that is not generating its stated interest rate because of nonpayment from the borrower.
   11. Credit card loans. This was normalized by equity capital.
2. For many of the dollar amounts and percentages in the quantities mentioned in step 9, we had to remove the characters “$”, “,”, and “%” from the strings and convert those strings to numbers. For example, “$121,455” would become 121455 and “14.5%” would become 14.5.

The modeling technique

1. **Selection of model type:** Since our target variable is counts, and counts typically have a Poisson distribution, we tentatively decided to use a GLM with Poisson link to build our predictive model. However, it was necessary to test the distributions of the adjusted complaint counts for each of the 75 banks in our modeling data to confirm that the counts had a Poisson distribution. We ran goodness-of-fit tests in R (using the “vcd” package) with method Minimum Chi-Square. Table 2 gives a list of banks and the corresponding p-values from the goodness-of-fit tests. Using a significance level of 0.05, we can conclude that only 14 of the 75 banks present significant evidence that their quarterly complaint count distributions differ from a Poisson distribution. Therefore, it seems reasonable to use a GLM with Poisson link to predict claim counts.

**Table 2: p-values from goodness-of-fit tests for Poisson distribution of bank quarterly complaint counts.**

|  |  |
| --- | --- |
| **Bank Name** | **p-value** |
| THIRD FEDERAL SAVINGS & LOAN ASSOCIATION OF CLEVELAND | 1.0000 |
| RAYMOND JAMES BANK, NATIONAL ASSOCIATION | 1.0000 |
| OLD NATIONAL BANK | 1.0000 |
| IBERIABANK | 1.0000 |
| TCF NATIONAL BANK | 1.0000 |
| ARVEST BANK GROUP, INC. | 0.9999 |
| BANCO POPULAR NORTH AMERICA | 0.9997 |
| UMB BANK, NATIONAL ASSOCIATION | 0.9997 |
| ASTORIA BANK | 0.9996 |
| BANCORPSOUTH BANK | 0.9996 |
| Whitney Bank | 0.9996 |
| FIRST NATIONAL BANK OF OMAHA | 0.9995 |
| FirstBank of Puerto Rico | 0.9995 |
| PRIVATEBANK AND TRUST COMPANY, THE | 0.9990 |
| FIRST CITIZENS BANCSHARES, INC. | 0.9987 |
| SCOTTRADE BANK | 0.9987 |
| FIRST TENNESSEE BANK NATIONAL ASSOCIATION | 0.9987 |
| STATE FARM BANK, FSB | 0.9982 |
| ZIONS BANCORPORATION | 0.9974 |
| MB FINANCIAL, INC. | 0.9963 |
| HUNTINGTON NATIONAL BANK, THE | 0.9962 |
| BOK FINANCIAL CORP | 0.9955 |
| WEBSTER BANK, NATIONAL ASSOCIATION | 0.9955 |
| RABOBANK, NATIONAL ASSOCIATION | 0.9955 |
| FIRSTMERIT BANK, N.A. | 0.9954 |
| BANCO POPULAR DE PUERTO RICO | 0.9940 |
| WASHINGTON FEDERAL, NATIONAL ASSOCIATION | 0.9937 |
| EAST WEST BANK | 0.9905 |
| FIFTH THIRD FINANCIAL CORPORATION | 0.9870 |
| FIRST REPUBLIC BANK | 0.9870 |
| FIRST NATIONAL BANK OF PENNSYLVANIA | 0.9833 |
| CATHAY BANK | 0.9783 |
| ASSOCIATED BANK, NATIONAL ASSOCIATION | 0.9710 |
| Synovus Bank | 0.9627 |
| UMPQUA HOLDINGS CORPORATION | 0.9544 |
| NEW YORK COMMUNITY BANK | 0.9538 |
| FIRST NIAGARA FINANCIAL GROUP, INC. | 0.9538 |
| E\*TRADE BANK | 0.9538 |
| BankUnited, N.A. | 0.9488 |
| MORGAN STANLEY & CO. LLC | 0.9432 |
| CITIBANK, N.A. | 0.9304 |
| SUNTRUST BANKS, INC. | 0.9202 |
| FROST BANK | 0.9200 |
| PNC Bank N.A. | 0.9082 |
| FIRST HAWAIIAN, INC. | 0.8981 |
| INVESTORS BANK | 0.8925 |
| CAPITAL ONE FINANCIAL CORPORATION | 0.8779 |
| PEOPLE'S UNITED BANK, NATIONAL ASSOCIATION | 0.8597 |
| UBS BANK USA | 0.8597 |
| BARCLAYS BANK DELAWARE | 0.8597 |
| SANTANDER BANK, NATIONAL ASSOCIATION | 0.8212 |
| REGIONS BANK | 0.7858 |
| BMO HARRIS BANK NATIONAL ASSOCIATION | 0.7858 |
| CHARLES SCHWAB CORPORATION, THE | 0.7858 |
| Comerica | 0.5982 |
| AMERICAN EXPRESS CENTURION BANK | 0.5982 |
| SYNCHRONY BANK | 0.4402 |
| DISCOVER BANK | 0.4402 |
| Deutsche Bank | 0.4402 |
| NORTHERN TRUST COMPANY, THE | 0.4402 |
| BNY MELLON, NATIONAL ASSOCIATION | 0.1494 |
| WELLS FARGO BANK, NATIONAL ASSOCIATION | 0.0000 |
| TD BANK US HOLDING COMPANY | 0.0000 |
| CHEMICAL FINANCIAL CORPORATION | 0.0000 |
| BANK OF AMERICA, NATIONAL ASSOCIATION | 0.0000 |
| EVERBANK | 0.0000 |
| PACIFIC WESTERN BANK | 0.0000 |
| MUFG UNION BANK, NATIONAL ASSOCIATION | 0.0000 |
| JPMORGAN CHASE & CO. | 0.0000 |
| CIT BANK, NATIONAL ASSOCIATION | 0.0000 |
| FLAGSTAR BANK, FSB | 0.0000 |
| GOLDMAN SACHS BANK USA | 0.0000 |
| ALLY FINANCIAL INC. | 0.0000 |
| COMPASS MORTGAGE, INC. | 0.0000 |
| WESTERN ALLIANCE BANCORPORATION | 0.0000 |

1. **Train and test data:** We randomly selected 70% of our modeling data as training data and 30% as test data.
2. **Predictor variable selection:** We iteratively constructed models, using statistics in the model summaries to eliminate certain predictor variables.
   1. **Model 1:** We fit the following variables into the model:

* adj.Amt.Past.Due.30.89.Days2 : Amount past due 30-89 days, normalized by equity capital
* adj.Amt.Past.Due.90.Or.More.Days2 : Amount past due 90+ days, normalized by equity capital
* adj.Restruct.Loans.80.89.Days.P.D2 : Restructured loans past due 30-89 days, normalized by equity capital
* adj.Restruct.Loans.90.Or.More.Days.P.D : Restructured loans past due 90+ days, normalized by equity capital
* Pct.Loans.Charged.Off : Percentage of loans charged off
* adj.Nonaccr.Restr.Loans : Nonaccrual restructured loans, normalized by equity capital
* adj.CC.Loans : Credit card loans, normalized by equity capital

The model summary is :

Call:

glm(formula = adj.complaint.count ~ adj.Amt.Past.Due.30.89.Days2 +

adj.Amt.Past.Due.90.Or.More.Days2 + adj.Restruct.Loans.80.89.Days.P.D2 +

adj.Restruct.Loans.90.Or.More.Days.P.D + Pct.Loans.Charged.Off +

adj.Nonaccr.Restr.Loans + adj.CC.Loans, family = poisson(),

data = Merged.Banks.15[Merged.Banks.15$train, ])

Deviance Residuals:

Min 1Q Median 3Q Max

-9.6146 -1.3966 -0.3037 0.5700 10.6820

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

(Intercept) 1.640e+00 1.570e-02 104.471 < 2e-16 \*\*\*

adj.Amt.Past.Due.30.89.Days2 1.393e-08 6.871e-10 20.266 < 2e-16 \*\*\*

adj.Amt.Past.Due.90.Or.More.Days2 **-1.104e-09** 1.063e-10 -10.390 < 2e-16 \*\*\*

adj.Restruct.Loans.80.89.Days.P.D2 -2.093e-09 3.247e-09 -0.644 **0.519260**

adj.Restruct.Loans.90.Or.More.Days.P.D -6.940e-09 9.796e-10 -7.084 1.40e-12 \*\*\*

Pct.Loans.Charged.Off -6.223e-02 1.321e-02 -4.712 2.45e-06 \*\*\*

adj.Nonaccr.Restr.Loans 2.461e-09 7.421e-10 3.316 0.000914 \*\*\*

adj.CC.Loans 7.076e-11 9.326e-12 7.588 3.26e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 7193.8 on 1044 degrees of freedom

Residual deviance: 3455.0 on 1037 degrees of freedom

AIC: Inf

Number of Fisher Scoring iterations: 5

We will omit adj.Restruct.Loans.80.89.Days.P.D2 from the model because its p-value of 0.519260 is greater than 0.05, meaning that we have insufficient evidence that the coefficient of adj.Restruct.Loans.80.89.Days.P.D2 is nonzero.

We will also omit adj.Amt.Past.Due.90.Or.More.Days2, because the negative sign of its coefficient doesn’t make sense; we would expect consumer complaint counts to increase as past due amounts increase, rather than the reverse.

* 1. **Model 2:** After making the omissions above, we obtain Model 2, whose summary is below:

Call:

glm(formula = adj.complaint.count ~ adj.Amt.Past.Due.30.89.Days2 +

adj.Restruct.Loans.90.Or.More.Days.P.D + Pct.Loans.Charged.Off +

adj.Nonaccr.Restr.Loans + adj.CC.Loans, family = poisson(),

data = Merged.Banks.15[Merged.Banks.15$train, ])

Deviance Residuals:

Min 1Q Median 3Q Max

-9.7558 -1.3921 -0.2857 0.5443 12.7315

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

(Intercept) 1.668e+00 1.548e-02 107.725 < 2e-16 \*\*\*

adj.Amt.Past.Due.30.89.Days2 7.290e-09 2.036e-10 35.805 < 2e-16 \*\*\*

adj.Restruct.Loans.90.Or.More.Days.P.D -7.645e-09 6.878e-10 -11.115 < 2e-16 \*\*\*

Pct.Loans.Charged.Off -3.993e-02 1.244e-02 -3.209 0.00133 \*\*

adj.Nonaccr.Restr.Loans 6.454e-09 6.129e-10 10.530 < 2e-16 \*\*\*

adj.CC.Loans 6.689e-11 9.278e-12 7.210 5.61e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 7193.8 on 1044 degrees of freedom

Residual deviance: 3593.1 on 1039 degrees of freedom

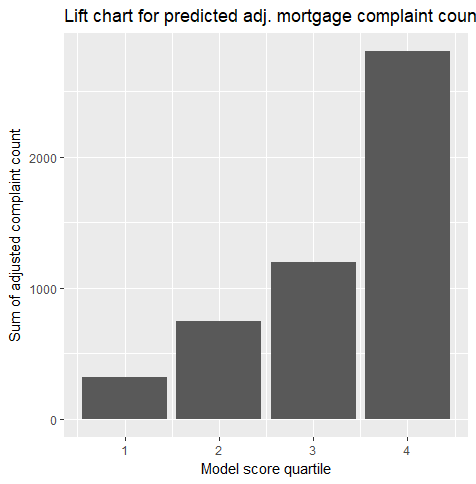
AIC: Inf

Number of Fisher Scoring iterations: 5

We selected Model 2 as our final model

Evaluation of model

1. We fit the model on the training set, then used that model to score the test set
2. Next, we partitioned the test set by score quartiles. This gave us a ranking of banks in the test set from 1 to 4, where a ranking of 1 represents lowest propensity to receive complaints and 4 represents highest propensity to receive complaints
3. We then added up the actual adjusted complaint counts by score quartile, to see how accurate the complaint propensities calculated by the model were.
4. We constructed a bar chart of the numbers computed in step 3. This **lift chart** is shown below:



We also give below a table of the bar heights, as well as the ratios of the heights of successive bars:

|  |  |  |
| --- | --- | --- |
| **Score quartile** | **Sum of adjusted complaint counts** | **Ratio of successive sums of adjusted complaint counts** |
| 1 | 319.20 | N/A |
| 2 | 749.82 | 2.35 |
| 3 | 1199.61 | 1.60 |
| 4 | 2808.58 | 2.34 |

The model does a very good job at separating banks by their complaint risk.

1. We would like to evaluate the impact of each predictor variable on predicted counts. To explain how we do this, let us begin by pointing out that our predictive model is multiplicative. That is, scored observations yield positive weights that are multiplied together to obtain a predicted count. More specifically, if x1,, x2, … ,xp are observations of the predictor variables in the model, with corresponding coefficients β0, β1, …, βp, then our predicted count is given by the product

exp(β0)exp(β1 x1)exp(β2 x2)… exp(βp xp)

To evaluate the size of the impact of these factors, we computed the five-number summary of each of the weights exp(βi xi), where the xi range over the entire test set. The five-number summaries are shown in the table below:

**Table 3: Summaries of weights in model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q25** | **Median** | **Q75** | **Max** |
| **Adjusted past due 30-89 days amount** | 1.037 | 1.192 | 1.479 | 1.843 | 80.32 |
| **Restr. Loans 90+ days past due amount** | 0.03052 | 0.6779 | 0.7752 | 0.9108 | 0.9954 |
| **Percent of loans charged off** | 0.7471 | 0.978 | 0.9948 | 0.9996 | 1.031 |
| **Adjusted nonaccrual restructr. loan amount** | 1.025 | 1.129 | 1.291 | 1.529 | 25.35 |
| **Adjusted credit card loans** | 0.9463 | 0.9953 | 0.9986 | 1.003 | 1.513 |

* We decided that adjusted past due 30 – 89 days amount is significant, because half of the time the corresponding weight is greater than 1.479, which is a significant impact upon predicted complaint count
* Likewise, we decided that restructured loans 90+ days past due is a significant predictor, because half of the time the corresponding weight is less than 0.7752.
* Percent of loans charged off is not practically significant, because most of the time the corresponding weight is close to 1.
* Adjusted nonaccrual restructured loan amount is significant, because half of the time the corresponding weight is greater than 1.291
* Adjusted credit card loans is not significant, because the corresponding weight is almost always close to 1.

Recommendations

1. Since complaint counts tend to increase with total amount past due 30 - 89 days, banks with high complaint counts should slightly increase their riskiness estimates of potential mortgage borrowers. Then, they should update their customer selection and/or loan structuring algorithms accordingly. Finally, they should do some testing to see whether this tweaking has had a significant impact on complaint counts.
2. Since complaint counts tend to decrease as restructured loans 90+ days past due increase, banks with high complaint counts should experiment with restructuring more loans in a way that is agreeable to the borrower.
3. This study could be done again with a larger number of banks and a larger number of predictors. The number of banks and predictors in this model was rather small due to difficulties in obtaining and merging data.
4. This study could be repeated for other bank products such as credit cards, car loans, etc.
5. It could be useful to repeat this study using complaint counts obtained from social media, since these complaint counts are likely to have a higher impact upon a bank’s bottom line than the complaint counts in the data we used here. A complaint in Yelp could be defined as a low rating.
6. Another future step could be digging into the complaining customers’ written comments to get more specific causes of the complaints.
7. There are many other modeling techniques besides GLMs with Poisson links that could be tried, e.g. random forests.