Predicting Bank Churn

Felipe Castillo

Aaron Lee

Data Mining

6/2/2022

In today’s time, we have a lot of businesses that consistently compete with one another. The truth can be said about applications and brick and mortar shops. In fact, it will be difficult coming up with an industry that doesn’t have competition. What does competition look like? If you notice, billions of dollars go towards advertising to get your attention. If you have a favorite hamburger place, someone else will also have hamburgers, and it might even be at a discount. This is also true for banks. If you notice, there is no shortage of them. Therefore, more than ever, it is crucial to find a way to retain those customers. Churn can be detrimental to any business or organization. In this case, we will be examining whether we can predict bank churn and what we can do to reduce it.

The problem with churn is that it can limit the bank’s overall lending power. The money must come from somewhere. For the bank to be able to make money, it needs money. This is a problem because the bank could lose money in the long run. Depending on the level of churn, there are ramifications for when there is less money in the bank. This will likely cause higher interest rates on loans. Even the smallest levels of churn could pose a problem. For example, if Elon Musk took his money out of the bank, it would have a larger impact than if a lot of people took out their money from the bank. With that same logic, a small group of people could potentially make most of the money in the bank. With that said, it is imperative to understand churn to either stop it or be able to prepare for it.

According to Dubes in his Forbes article, “banks today aren’t solving the problem of churn effectively (Dube, 2020).” It is safe to say that both customer and premium customer are being lost. There has been a big stride to getting new customers, but now it is not enough to just get new customers. It’s important to retain customers. This issue could potentially cost the bank thousands if not millions. Without a proper strategy, the bank will continue to lose valuable clients. How can a bank grow without both retaining and gaining new customers? As a stake holder, both categories are important. With that said, it is also important to note that in some banking areas there have been effective methods that banks have implemented thanks to AI methods. This means that focusing on bank churn should be effective.

In this project, we will work with the database Bank Turnover Dataset. This data set can be found and downloaded here: <https://www.kaggle.com/datasets/barelydedicated/bank-customer-churn-modeling>. The data set can be found on Kaggle. Its purpose is to see whether or not we can predict the bank churn with the dataset. The data set is cleaned with no null values. It has a target variable known as exits that we will be identifying. Exited refers to the people that have left the bank. This data set was last updated 4 years ago. It also supports api calls, however we will be working the csv data set.

For the final milestone submission of my code, these are the updates that I made to milestone 1, 2, and 3. First, all the cells were made to run independently from one another. For example, if you run milestone 3, it should be dependent on milestone 2. Also, I changed the PCA section to milestone 2. Instead of me using it as its own separate model, the PCA data was used in the following models. Comments were added in milestone 3 to explain the logic at every step. Lastly, I added another model, random forest classifier, to be able to compare against my model. All the graphs, values, and confusion matrix were changed to the random forest classifier data. The data in each model are scaled and I used PCA data from milestone 2. The conclusion for the milestone was updated to mimic the results.

There are three EDA visuals that are important to fully understanding the dynamic churn. First, we need to understand who the customers are. For example, are there more males leaving the bank than women? The data set holds multiple features that can be divide into categories like credit score, age, salary, etc. This information will allow us to categorize our customers into different groups, but first, we need a graph showing how many people have left.

Chart, pie chart

Description automatically generatedThe pie chart shows that 20.4% of the individuals have left the bank. Essentially, we want to know how we can retain those people and who these people are.  
 Now that we have an idea of how many people are leaving the bank, we need to know who they are. To do this it will require data preparation. Each of the features that hold weight are divided into their own classification. The two features, age and credit score, are crucial in understanding the individuals in the bank. I used Nancy a gerontologist to define age groups. The age groups are divided in three groups, young adults (18-35), middle adults (36-55), and older adults (55 and older) (Nancy, 2022). Also, I used Ulzheimer base line using Experian credit score to classify and divide credit score groups. It was divided into 5 groups: poor (less than 579), fair (580-669), good (670-739), very good (740-799), and exceptional (800 and above) (Ulzheimer, 2021). The other two values that we will be looking at are already classified into their own subgroup. The final features for our classification bar group are nationality, gender, age group, and credit score group. These features are compared with people leaving the bank.

We can gather a lot of information from the bar graph. It looks like they are more men that are staying at the bank than women. There does seem to be more women leaving the bank than men. As a base of nationality, Germany and France have similar exit rates versus Spain which has a lower rate. It also seems like middle-aged adults between the ages of 36 and 55 are disproportionately leaving the bank versus young adults and older adults. Fair and poor credit scores have the highest exit rates from the bank. It also shows that if the individual has an exceptional credit score, they are less likely to leave.

Graphical user interface, application

Description automatically generated

There are two features that have some correlation with exited. The first feature is age with 29 percent. Also, balance, which is the amount someone has in their account, is at 12 percent. Interestingly enough, the level of salary someone has is also at 12 percent. The other features had a negative correlation with exited. None of the features are highly correlated.

A picture containing graphical user interface

Description automatically generated

During the data perpetration stage, I converted some of the numeric values to meaningful values. I added a classifier to four of the features. Those features were age, credit score, bank balance, and salary. These values will help us categorize our bank customers. For example, we can define it if an individual has a great credit score or if it’s a poor credit score. All these categories will justify or reject the reason someone leaves the bank Most importantly we will be able to classify our individuals inside the bank.

However, during the validation, surname was a value that needed to be tested to see if it holds any value. I didn’t receive any feedback from the correlation matrix on surname. However, do we need to account for family ties inside the bank? This is a good question. Can we produce a new insight if someone were to leave the bank? What if those same individuals choose a bank based on family relations. Therefore, I validated that there was indeed a high level of duplicates of the same surname. It is important that we don’t rush to conclusions as the surname could belong to a different family. It doesn’t guarantee that they are all indeed related, but it does gives us an insight whether family relations could yield a difference in our model. This was by far the hardest feature to validate.

Keeping the name was important, however, when we use one hot encoding in our data set, it increased the number of columns from 16 to 2959. One hot encoding turned our classified features into binary features. For example, if the person has an excellent credit score, it will make it as a one. If it doesn’t, it will make it as a 0. This allows our model to understand the difference between our individuals. Can we have a reduction in features without hurting our model?

I followed our professor’s advice in rearranging my code to include principal component analysis (PCA) in the data preparation section. PCA is a technique used for reducing the dimension of the data set while increasing interpretability at the same time and making sure that there is not a huge loss in information. This can be thought as rearranging the data to see different sides of it. Before the model creation, I wanted to see if it was possible to reduce the number of features. Therefore, I used PCA with a component of 50 and used a logistic regression model. The logistic regression will be our main model; therefore, it was used. To my surprise I was able to reduce 2959 of features to only 50. I was able to keep the accuracy score to 79 percent. I will use PCA for our main model building.

I used a logistic regression. I first scaled the data and ran PCA on the data. This allowed for less features to be evaluated. I wanted to compare my model with a different model. I used Random Forrest Classifier as a comparison to my logistic regression model. Both models were run with that same data and scaled with PCA. I picked random a forest classifier because it also works as a classification model. My hypothesis is that it should not be too far off the logistic regression accuracy.

Overall, our models did a good job of predicting bank churn. My logistic regression model accuracy was 77% with a recall score of .085, a precision score of .40, and a f1 score of .14. The random forest classifier had an accuracy score of 85%. It has a recall score of .37, a precision score .40, and a f1 score of .51. On average, the random forest had a higher accuracy score than the logistic regression. What’s interesting is the recall, precision, and f1 score. Recall, F1, and precision measure a model’s accuracy when making predictions. Precision measures how precise and accurate we predicted positive and how many of them are actually positive. While recall calculates how many actual positives our model captures as True positive, F1 takes both precision and recall and a balance of both (Shung,2020).Random forest has better f1 then logistic regression. The logistic regression has a low score and brings suspect to the model accuracy. The Roc plot shows that the random forest classifier is the better model. The way it works is it takes the true positive and compares it to the false positive. If you follow the line where it intersects, it should give you an idea of true positive to false positive rate. An example is if you follow the 60 percent line. This means that it correctly predicted 60 percent of the time and 20 percent of the time it will get it incorrect. The curve is closer to one for random forest classifier. I also validated the data by checking the consistency. I reran the model a few times. The logistic regression was not consistent, and the values were all over the place. Random forest was consistent with a threshold of 80%. Its accuracy would prove likely that random forest is the better model.

ndChart, line chart

Description automatically generated

At this point the model is not ready to be deployed. I picked logistic regression assuming it would be the best model for a binary classification problem. However, the compared model seems to be the better model. All scores that we went over show that even if we didn’t use random forest, logistic might not be the best choice. Even though this project points to using a logistic regression model, it might be required to evaluate random forests in detail before deployment.

My recommendations are to have an idea of multiple models that can work for a given problem and compare them. Once you find the better model, focus on this. Also, I recommend PCA. Without this method, random forests were taking about 10 minutes to compute. The models were very slow. This is due to how many records were available. Lastly, taking our professor’s advice to be careful in mixing my personal thoughts with the data. This was an issue that I had with the surname.

One of the potential challenges and additional opportunities to still be explored is gender issues. The data indicates that there is an issue with a lot of women leaving. Perhaps male and female could be compared with one another to see what is different. The issue is there, but the real challenge is to explain why it’s happening. This also opens doors for opportunity in which the stack holder could develop a plan to target these women. Also, more classification models might have to be evaluated before we give a thumbs up on the deployment. Lastly, if we were able to find an index of time, this would make the model more useful if we could predict when the customer is most likely to churn. Right now, our model doesn’t account for time.

In conclusion, with competition being so fierce in the 21st century, it’s important to hold the upper hand when it comes to gaining and retaining customers. The EDA process shows us hidden information that might not be so easy to spot. I would set up another study to focus on gender issues inside the bank. This is one of the more obvious discrepancies. After the reduction of the data, we were able to get a good accuracy score. The model I chose seems to perform worse than the alternative. Given the results, I would use the random forest over logistic regression. Overall, it is safe to say that we can predict bank churn.

**Work Cited Paper**

Dube, D. (2020, May 27). *Council post: Why retaining customers for Banks is as important as winning new ones*. Forbes. Retrieved June 2, 2022, from https://www.forbes.com/sites/forbestechcouncil/2020/05/27/why-retaining-customers-for-banks-is-as-important-as-winning-new-ones/

Ulzheimer, J. (2021, March 11). *Credit score basics*. Experian. Retrieved June 2, 2022, from https://www.experian.com/blogs/ask-experian/credit-education/score-basics/

Nancy, P. (22AD, February). *A comparison of young, middle-aged, and older adult treatment-seeking pathological gamblers*. The Gerontologist. Retrieved June 2, 2022, from https://pubmed.ncbi.nlm.nih.gov/11815703/

Shung, K. P. (2020, April 10). *Accuracy, precision, recall or F1?* Medium. Retrieved June 2, 2022, from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9