

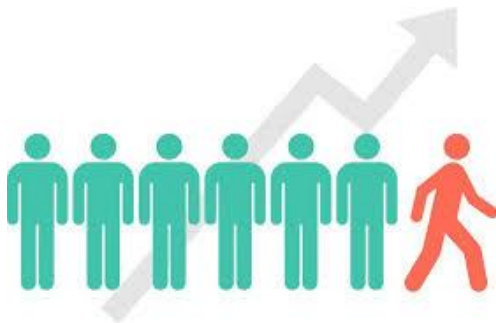
Springboard Data Science Capstone Project - 1
Predicting the Likelihood of **Customer Churn**

Capstone Report



Carolyn Massa

March 2020



Contents

[illegible]

1 Introduction

Whatever you call it – defection, attrition, turnover – customer churn is a painful reality that all businesses must deal with. Even the largest and most successful companies suffer from customer churn therefore understanding what causes formerly loyal customers to abandon ship is crucial to lasting, sustainable business growth.

Let’s say you are the person responsible for allocating and maintaining a company’s budget for client acquisition. You may spend, say, 15,000 USD just to obtain a client after marketing and customer service costs. You notice over time over 10% of your 10,000-customer client base continue to leave after your meticulous efforts made to acquire the customer and maintain them. This is where a close analysis is necessary to make appropriate modifications to your processes and possible your product offerings.

Most businesses are heavily affected by customer churn...from banks to online retailers, it matters and is mission critical to examine WHY the customer is leaving your business and putting their purchasing power into another organization.

You can measure your client churn rate in one or more of the following ways:

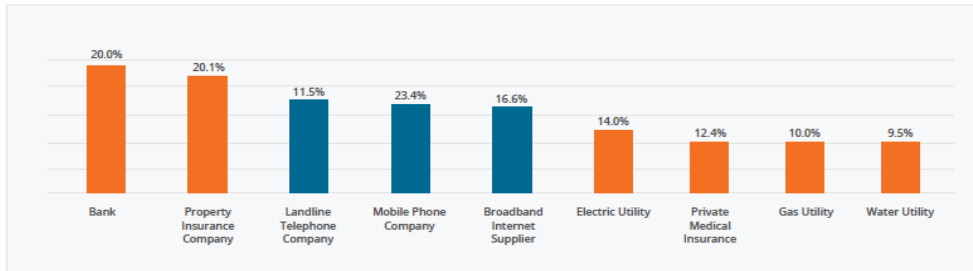
- Total number of customers lost during a specific period
- Percentage of customers lost during a specific period
- Recurring business value lost
- Percentage of recurring value lost

In my study I connect to a dataset from a Global Bank to explore their rate of churn, look for patterns, and build a model that can be put into production to serve to look for possible churn “warnings” so the Global Bank management teams can react and address and change the correct deficiencies and take appropriate measures to prevent their loss.

“Null Hypothesis”

Can consumer behavior or trends prove useful to predict whether a customer leaves a bank or not? According to a “Call Miner” Index below, banks have an average annual churn rate of 20% of their valued customer base. That being stated, my null hypothesis (H0) is that *“Gender is not a determining factor to whether a customer leaves the bank or not”*. I will test this hypothesis at the end of this Milestone Report and verify *if* this is a valid statement or not.

Chart #1 - The CallMiner Index | Switching rates per sector in the last 12 months



2 Data Acquisition and Cleaning

I acquired the “bank churn” dataset from Kaggle.com which contains a list of the bank’s 10,000 customers and how many churned in an 11-year time frame.

The data set below has 13 variables relating to 10,000 customers, which we can use to develop a predictive churn model. Below I process the data and start my exploratory data analysis. My [python code can be reviewed here](#) along with a Powerpoint slide [Presentation here](#).

The data set contains 13 variables:

| | | | | |
|------------------|------------------|--------------|--------------|-----------------|
| Customer ID | Sur Name | Credit Score | Geography | Gender |
| Age | Tenure | Balance | #of Products | Has Credit Card |
| Is Active Member | Estimated Salary | Exited | | |

Obviously, the data is incomplete and leaves a lot of unanswered questions.

- Would it be possible to obtain balances over a given time period as opposed to at a single date?
- What date did the customer exit?
- What types of products do the customers use?
- Could they have exited from a product and not the bank?

To begin I need to verify the type of data, what percent is useable and look for patterns; I notice that of the 10,000 customers of Word Bank, 2,037 have churned in the past 11 years which is a 20% Churn Rate. I also discover that the average customer has a tenure of 5 years with the bank and is 37 years old (range 18-92). The average age of the customers who churn is 44 years old.

Proportion of customer churned and retained

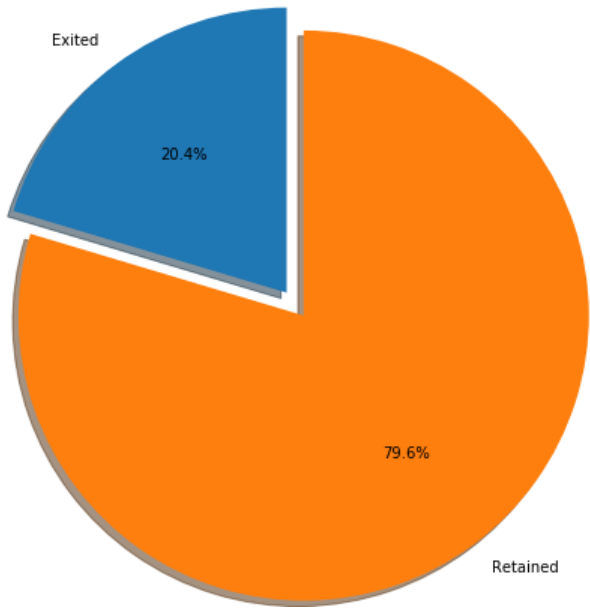


Fig 1 As you can see from the figure above 20.4% of 10,000 customers have churned over the past 11 years

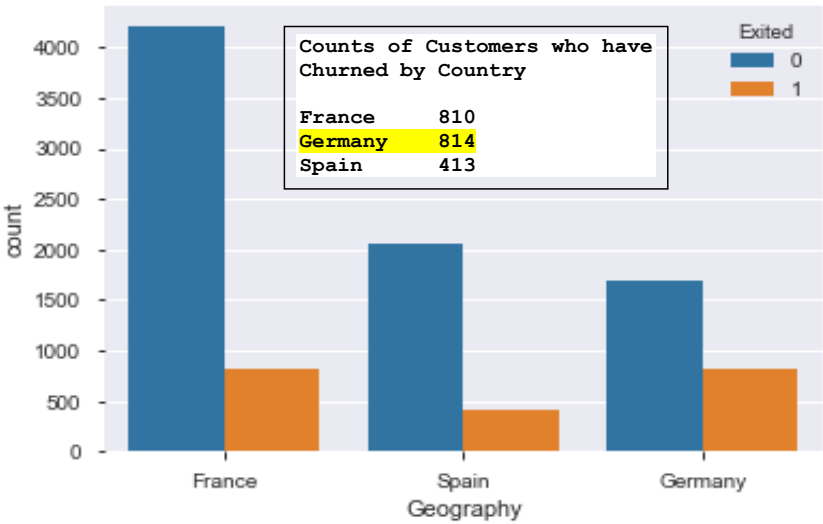


Fig 2 above as you can see from the figure above of the 2,037 who have left the bank 814 come from Germany, 810 from France, and 413 from Spain. Germany has the highest churn rate among the three countries.

I next check for missing data or outliers. I dropped 3 columns (Surname, CustomerID, and RowNumber) as they are not relevant to my research. In checking for outliers, I looked at the distributions of key variables: for example with customer age, the oldest customer is 92, the average member is 38 and the youngest is 18. The highest bank balance is \$250,898, the lowest \$0, and the average \$97,198. This data helps me put the members into prospective as I investigate further the cause of churn. Ultimately, I did not identify any outlier observations in the data set.

2 Exploratory Data Analysis

After verifying that I have no missing values and/or irrelevant data I get to work at seeking patterns and running correlations.

I first compare the Churn rates by value for categorical variables (Gender, Geography, Credit Card Holder, and Product Participation).

- Commented [RN(1)]: Is this true? Otherwise, please discuss identification and removal of outliers here.
- Commented [CK2R1]: Yes there were no outliers.
- Commented [CK3R1]:

Commented [CK4]:

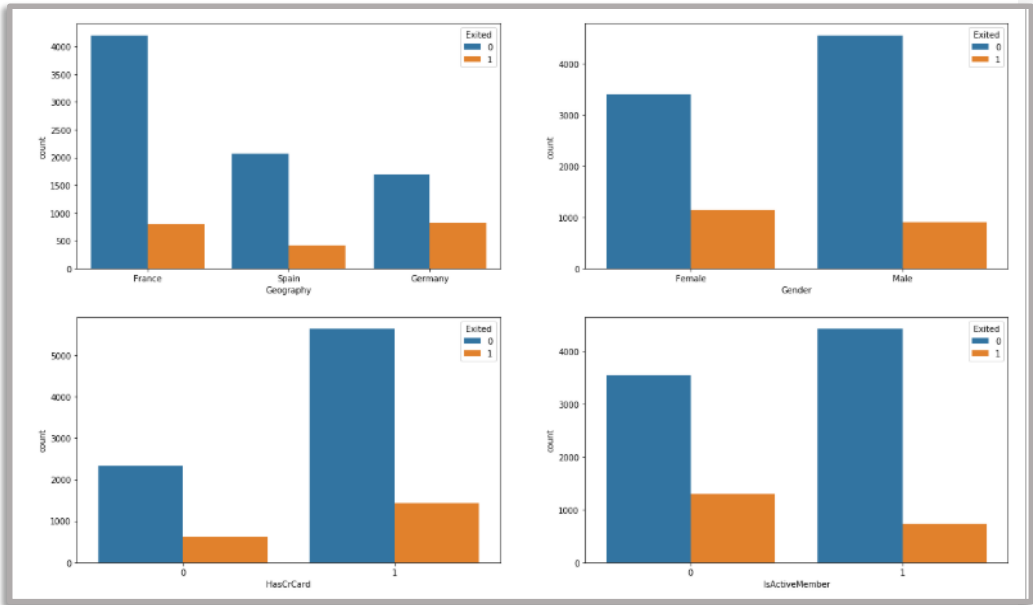
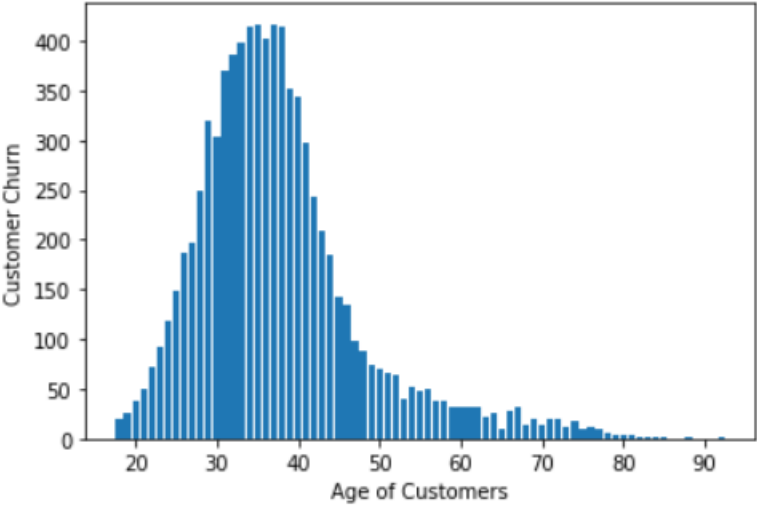
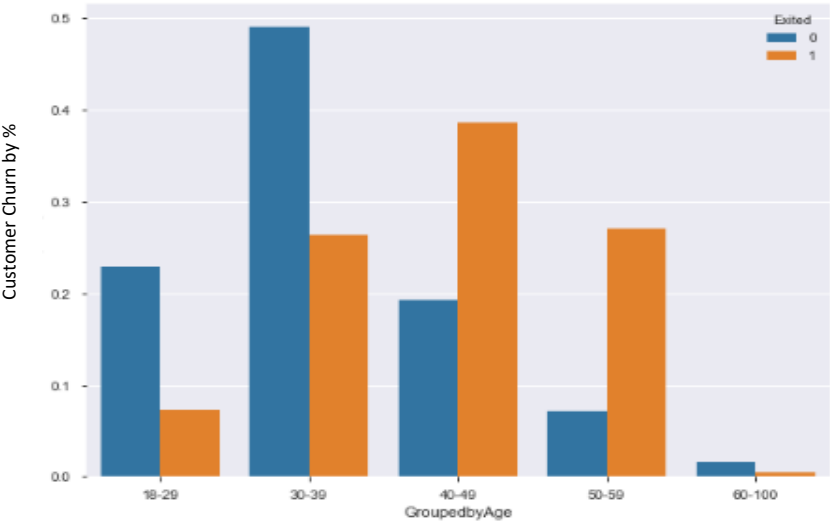


Fig. 3 above:

I discovered the following:

- As mentioned earlier, the country of France has the lowest churn rate and Germany has the highest. However, the proportion of churned customers is inversely related to the population of customers alluding to the bank possibly having a problem (maybe not enough customer service resources allocated) in the areas where it has fewer clients.
- The proportion of female customers churning is greater than that of male customers.
- Oddly, many of the customers that churned are those with credit cards. Many customers have credit cards, so this could just be a coincidence.
- The “inactive” members have a higher churn rate. The overall proportion of inactive members is quite high, suggesting that the bank may need a program to reactivate customers, which would improve customer retention.



Fig(s) 4/5 Above: Age groups that has the highest “churn” rate are those between 40 and 60. Now I compare the “continuous” variables against Churn below:

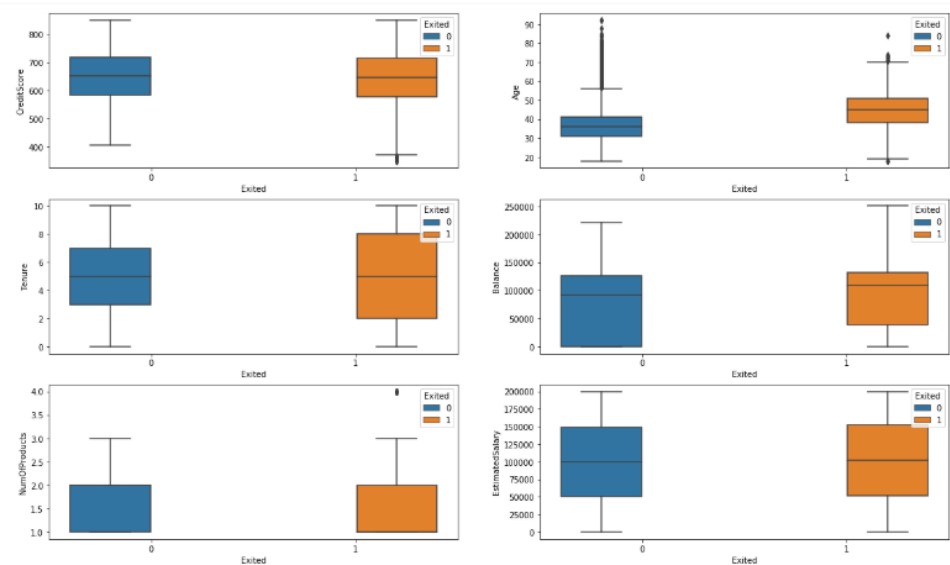


Fig. 7 above: I discovered the following:

- Interestingly, neither the number of products nor the salary has an impactful effect on the likelihood to churn. The credit score distribution is similar between retained and churned customers.
- The Average Salary of the customers that churned is 101,465 versus 99,738 who remained with the bank
- The Average Credit Score of the customers that churned is 645 versus 645 who remained with the bank
- The Average Number of Products of the customers that churned is 1.47 (of 3) versus 1.54 (of 3) who remained with the bank
- The average tenure of clients is 5 years. Customers with little tenure or very long tenure were the most likely to churn.
- Worryingly, the bank is losing customers with significant bank balances which is likely to hit their available capital for lending. The average bank balance for a retained customer is \$91,000 vs. \$101,465 among churned customers.
- One interesting find is the older customers are churning at a higher rate than the younger ones which alludes to the fact that the bank may not have adequate service standards that meet customer service expectations of older clients. The bank may gain from creating additional services plans for this client base. This is especially concerning since this age group may retirement within a 10 to 20 year window and has accumulated amounts of capital.

Commented [RN(5)]: CONFIRM

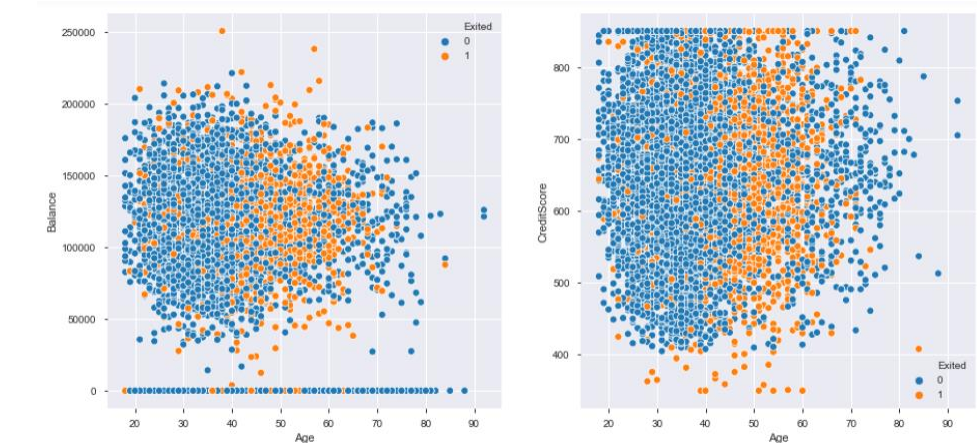


Figure 8 above: Since Age and Bank Balance presented some insights, I plotted a swarm plot to display clearer results. We see that the orange dots show clearly the Age Range between 40 and 70 had higher likelihood to churn while the bank account holder’s credit score has little relevance. There are more clusters of Churned Customers in the Bank Account Balance scatter plot between \$100,000 and \$150,000.

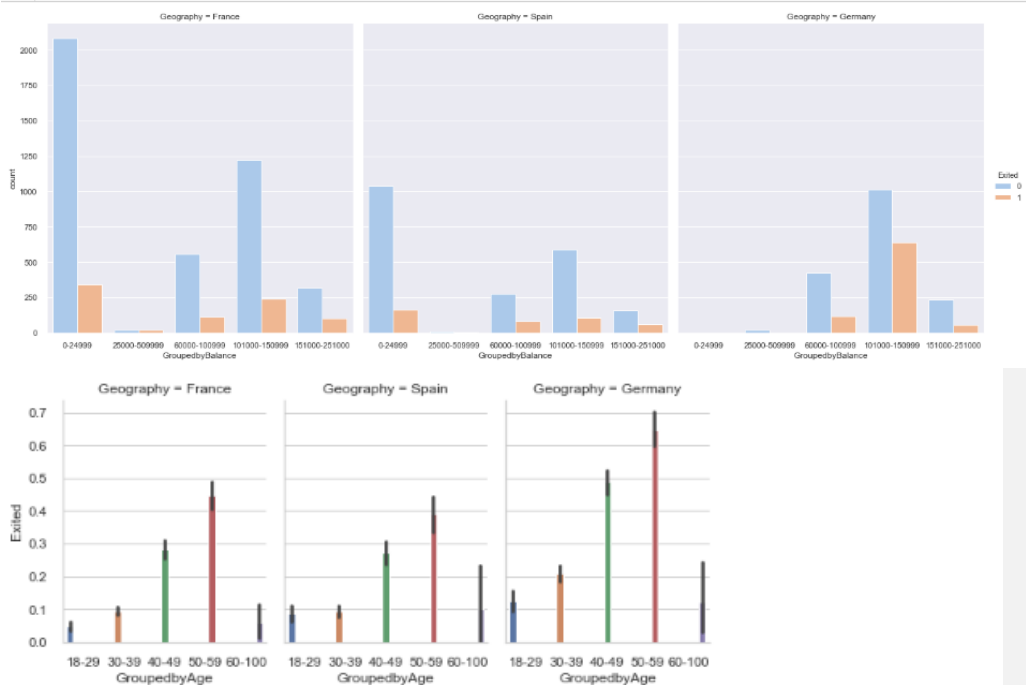


Figure 9a and 9 b above: In this bar chart I review Churn by each of the 3 countries, it is again clear that those bank members with balances between \$101,000 and \$150,000 are more likely to Churn. Germany has the highest Churn rate of the three. The age range of our bank customers is showing a clear trend also.

9b: Here I compare the Average Bank Balances in my 3 countries separated by who left and bank and who did not:

We see from the data below that of those clients who left the bank (indicated with a 1) each of the 3 countries had higher average balances.

Commented [RN(6): What is the average bank balance in each country, split out by Churned/Retained? i.e.

| Avg. Balance | Germany | France | Spain |
|--------------|---------|--------|-------|
| Churned | | | |
| Retained | | | |
| Total | | | |

Exited Geography Bank Balance Averages

| | | |
|----|---------|---------------|
| NO | France | 60339.275678 |
| | Germany | 119427.106696 |
| | Spain | 59678.070470 |

| | | |
|-----|---------|---------------|
| YES | France | 71192.795728 |
| | Germany | 120361.075590 |
| | Spain | 72513.352446 |

2.1 Correlation: Next I want to see how all the variables related to the possibility of churn using Pearson’s R:

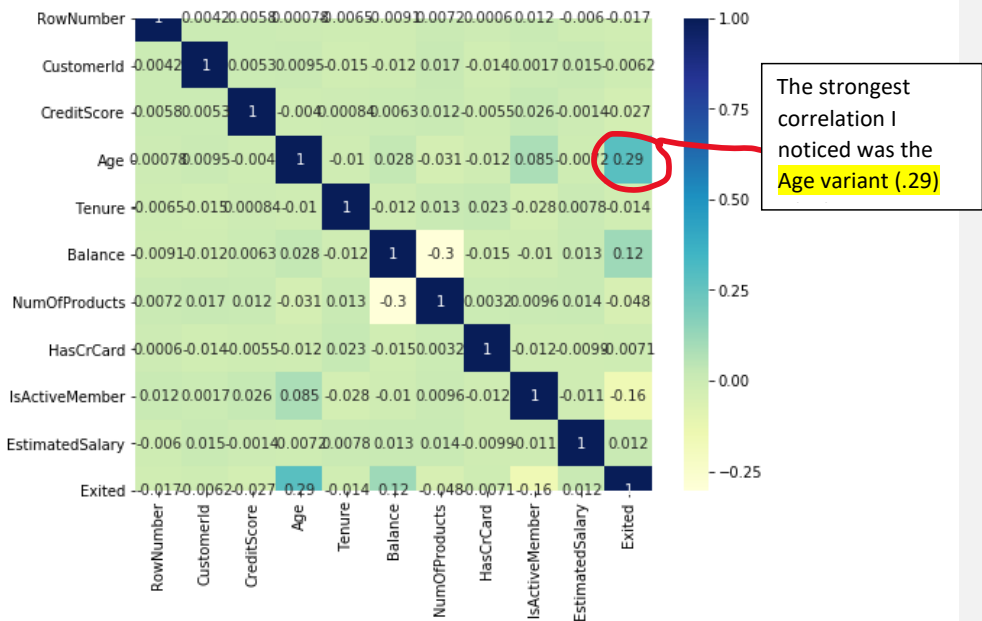


Fig: 10: In this Heatmap I used **Pearson's correlation coefficient** which is the test statistics that measures the linear relationship, or association, between two continuous variables, using covariance. Note that a score of 1.0 is perfect correlation and -1.0 is negative correlation.

2.2 Data Pre-processing/Feature Engineering

We know that since customers either churn or remain with the bank, the outcome is binary (0 = retained and 1 = churned). I used supervised machine learning algorithms to build a classification model. The models are trained using the 70% of the data and the remaining 30% is used to evaluate the performance of the models. I performed feature engineering to combine variables that have an impact on the possibility of churn. I will describe these next before moving onto my modeling where I used a Logistic Regression and Support Vector Machine Learning (SVM) with an RBF Kernel in case the data exhibits non-linear relationships.

Prior to running the algorithms, I implemented a few pre-processing steps to enhance my study and increase its accuracy. We outline these steps below. Note that some of the steps are not required (or not good for the best results) in some algorithms, but we list below all the pre-processing steps in order that they are performed that we used across all classification algorithms in this work:

- **Hot Key** encoding on categorical variables (**Gender and Geography**) to transform them to binary indicators. I also relabeled two binary variables “Has Credit Card” and “Is Active Member” to replace “0” values with “-1” to show a negative relationship more clearly.
- **New Variables:** I created 3 new variables by combining pairs of existing variables: Balance/Salary, TenurebyAge, and CreditScore Given Age

- **Data splitting:** The second step involves splitting the label encoded dataset into train and test datasets. In this project I separated the data to a 70/30 ratio. The fractions of both classes remain the same in train and test datasets.
- **Scaling:** For some algorithms, it is necessary that we scale the values of all features to lie within a fixed range. We scale features such that all features have values between 0 and 1.

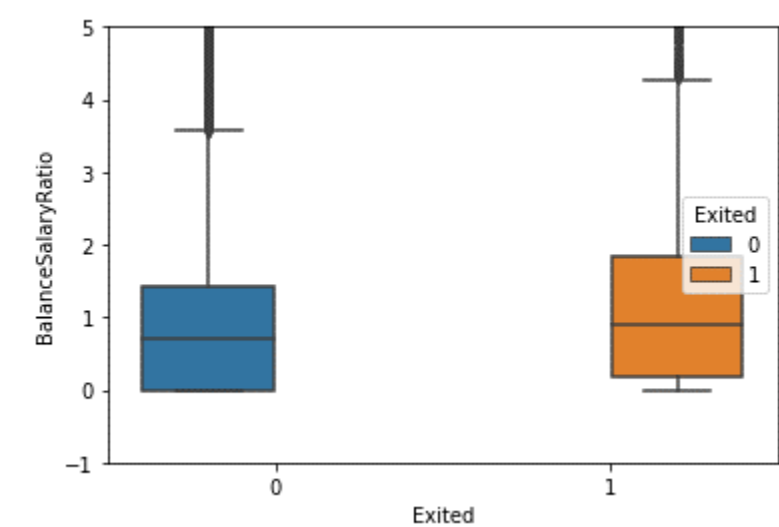


Fig 11 above: Here I combined **Bank Balance** and **Salary** to show ration to further examine correlation with Churn and enforce the evidence that higher bank balances have more Churn even after controlling for salary.

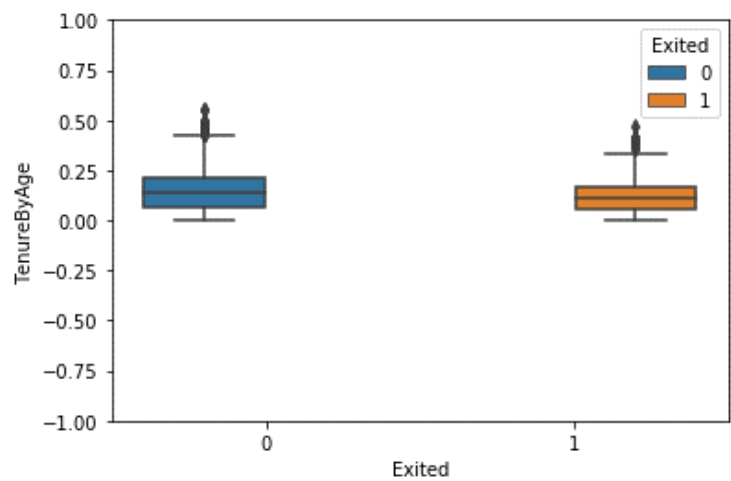


Fig 12 above: Tenure is a function of age, so I combined the two to check for trends

Commented [RN(7): I am not a fan of these charts. A better way to visualize this type of data is to use paired bar charts: put average churn rate on the y axis, and plot Age Group and Salary Group on the x axis as bars, where the height of the bar corresponds to average churn rate. You can also color the bars to differentiate between different Age groups.

Commented [CK8R7]: Please review as I have no result for "AVERAGE CHURN RATE" as I had no dates of churn to reference; just that they churned or did not churn. I added in two bar charts by color

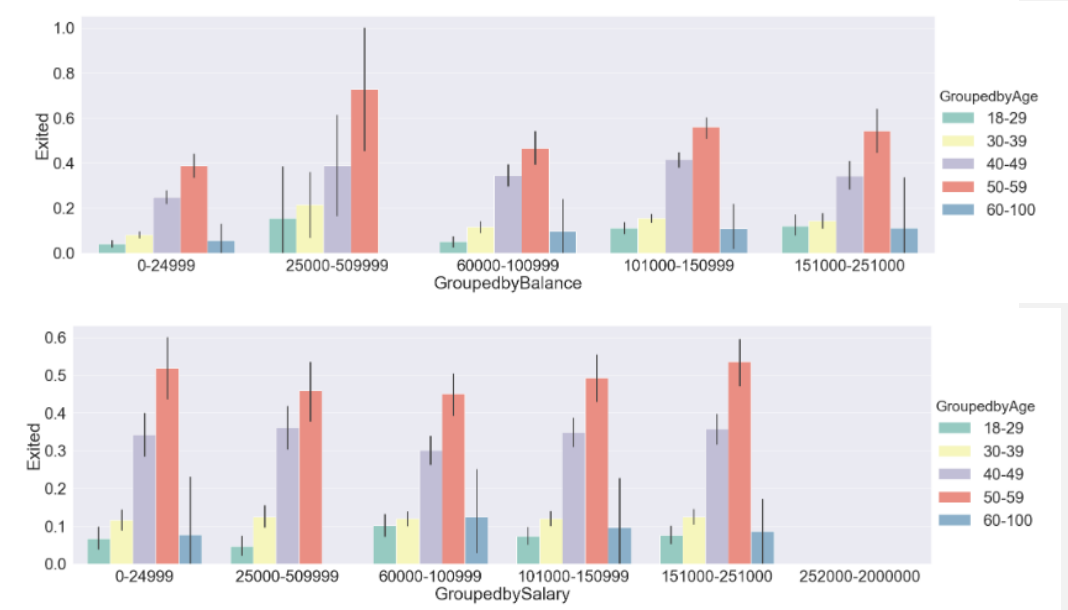




Fig 13: Again, the bar plots above shows a clear trend in Age of those bank members who left the bank with the light purple and salmon colored blocks being in the 40 to 60 age groups.

Scaling the Data Set: I scaled my continuous variables before modeling. Normalization helps in comparing values from two or more different variables while removing the effects of the differences in the scale of the data i.e. variables with large values can be easily compared with a data set of smaller values. Normalization is not necessary crucial for Logistic Regression, Random Forest, or K Nearest neighbors, but is important for SVM and Linear Regression so larger weights do not overpower smaller ones.

Normalization Formula

$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

Statistical Tests

In the beginning of the Capstone Report I mentioned making a Null Hypothesis that Gender had no impact on whether the bank customer left or not. I run a “p- value” Test via a “T-Test” to investigate further and I use the “Bank Balance” Feature as it has been shown to have impact on Customer Churn.

T-Test

I performed an unpaired t-test to compare the mean Bank Balance for Retained vs. Churned customers. Retained customers have an average bank balance of \$91,00, vs. \$101,465 among Churned customers. The test yielded a t-statistic of 4.10 with a p-value of 0.00067. The null hypothesis is that the two groups have the same mean, and the t-statistic and p-value tell us it is very unlikely that we would observe such large differences in the means assuming the two groups have the same mean. Thus, with a p-value < 0.05, we reject the null hypothesis in favor of the alternative hypothesis that there is a relationship between bank balance and whether the customer churned or not.

- Commented [RN(9)]: Move this section to before the modeling section. It is relevant to your analysis prior to building the models because it is generating insights that you can use in the predictive models.

Commented [RN(10)]: I thought your Null Hypothesis pertained specifically to whether Gender had any impact on churn? (See page 3)

Commented [RN(11)]: Clearly some customer traits do impact likelihood to churn (e.g., Age!). Otherwise, none of your models would have any predictive power, and would fail the F test!!

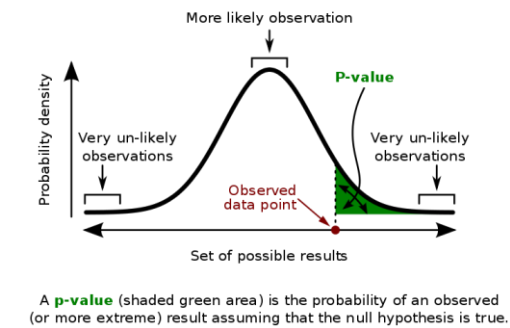
What is the “p-value”? The smaller the p-value the less likely it is that the hypothesis under consideration is true given a value as extreme or more extreme than the observation. If the p-value is less than or equal to a small, arbitrarily pre-defined threshold value, the null hypothesis is rejected in favor of the alternative hypothesis. The level of significance and is set by the researcher before examining the data and is arbitrary, typically between 0.05 and 0.001.

Important:

$\Pr(\text{observation} \mid \text{hypothesis}) \neq \Pr(\text{hypothesis} \mid \text{observation})$

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error: **the transposed conditional fallacy.**



Model Fitting

I chose the following 3 algorithms to build my model:

- 1) Logistic Regression:** Logistic regression is one of the basic and popular algorithms to solve a classification problem. The term “Logistic” is taken from the Logit function that is used in this method of classification which uses the Sigmoid function. Logistic regression is very useful for binary classification problems because it does not just output a predicted class, but also the *probability* that the observation is in a given class, which is useful for ranking exercises.
- 2)Support Vector Machine (SVM)** is a supervised machine learning algorithm which can be used for both classification and regression problems, but, is more commonly used for classification. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate.
- 3) Random Forest** works well with a mixture of numerical and categorical features which the Bank Churn data has. When features are on various scales, it works well and data does not even need to be normalized or scaled even though I did scale my data to apply the SVM model. Random Forest uses a large number of trees, works with missing values and is often considered to be a highly accurate model for both regression and classification problems.

Classification Reports:

The last line gives a weighted average of precision, recall and f1-score where the weights are the support values. The total is just for total support which is 5 here.

The f1-score gives you the harmonic mean of precision and recall. The scores corresponding to every class will tell you the accuracy of the classifier in classifying the data points in that class compared to all other classes.

The support is the number of samples of the true response that lie in that class which for this study are 2995.

Below I display BOTH my training data test results and then my test data results as I divided my 10,000 observations in a 70/30 separation.

Source: [sklearn documentation](#).

Logistic Regression – TEST DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.83 | 0.96 | 0.89 | 2411 |
| 1 | 0.57 | 0.20 | 0.29 | 584 |
| accuracy | | | 0.81 | 2995 |
| macro avg | 0.70 | 0.58 | 0.59 | 2995 |
| weighted avg | 0.78 | 0.81 | 0.78 | 2995 |

LOGISTIC REGRESSION TRAIN DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.83 | 0.96 | 0.89 | 5547 |
| 1 | 0.61 | 0.24 | 0.34 | 1453 |
| accuracy | | | 0.81 | 7000 |
| macro avg | 0.72 | 0.60 | 0.61 | 7000 |
| weighted avg | 0.78 | 0.81 | 0.78 | 7000 |

Commented [RN(12)]: Delete

Commented [CK13R12]: I was advised by a SB review contractor to add these in?

Commented [CK14R12]:

SVM TEST DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.88 | 0.98 | 0.93 | 2411 |
| 1 | 0.86 | 0.45 | 0.59 | 584 |
| accuracy | | | 0.88 | 2995 |
| macro avg | 0.87 | 0.71 | 0.76 | 2995 |
| weighted avg | 0.88 | 0.88 | 0.86 | 2995 |

SVM TRAIN DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.87 | 0.90 | 0.89 | 5547 |
| 1 | 0.57 | 0.50 | 0.53 | 1453 |
| accuracy | | | 0.82 | 7000 |
| macro avg | 0.72 | 0.70 | 0.71 | 7000 |
| weighted avg | 0.81 | 0.82 | 0.81 | 7000 |

Commented [RN(15)]: It seems strange that you would have higher accuracy on the Test data than on the Training data. But the # observations does make it appear that this is the Test results

Commented [RN(16)]: Delete

Random Forest - TEST DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.86 | 0.98 | 0.92 | 2411 |
| 1 | 0.85 | 0.36 | 0.50 | 584 |
| accuracy | | | 0.86 | 2995 |
| macro avg | 0.85 | 0.67 | 0.71 | 2995 |
| weighted avg | 0.86 | 0.86 | 0.84 | 2995 |

RF TRAIN DATA RESULTS

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.87 | 0.95 | 0.91 | 5547 |
| 1 | 0.72 | 0.47 | 0.57 | 1453 |
| accuracy | | | 0.85 | 7000 |
| macro avg | 0.80 | 0.71 | 0.74 | 7000 |
| weighted avg | 0.84 | 0.85 | 0.84 | 7000 |

Commented [RN(17): Delete

I review my scores by my Model using the F1-score as it takes both Churn and Non-Churn predictions into account and I applied it to an uneven class distribution. I will look at how well the mode did at predicting “Churn” which is represented by the “1” in my report.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you must look at other parameters to evaluate the performance of your model.
Accuracy = (TP+TN)/(number of obs.)

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question this metric asks is, of all bank customers that are labeled as having remained at the bank, how many actually remained with the bank? High precision relates to a low false positive rate.
Precision = TP/(TP+FP)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class. The question recall answers are: Of all the bank customers that truly remained with the bank, how many did we label?
Recall = TP/(TP+FN)

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)
Source = [Exsilio](#)

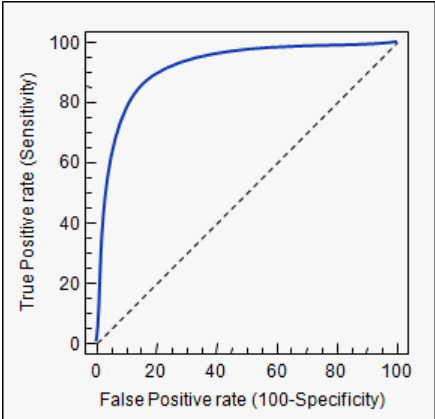
Logistic Regression results: The accuracy for our LR model is 0.81 which means our model correctly classifies customers 81% of the time. We have 0.78 precision score which is also strong as well as a recall of 0.81. Our F1 score is 0.90.

Support Vector Machines: The accuracy for our SVM model is 0.88. We have 0.88 precision score which is also strong as well as a recall of 0.88. Our F1 score is 0.86.

Random Forest Model: The accuracy for our RF model is 0.86. We have 0.86 precision score which is also strong as well as a recall of 0.86. Our F1 score is 0.84.

Commented [RN(18): Consider that 80% of your customers did not churn. Thus a naïve model could simply predict that ALL customers were retained and still be 80% accurate! 80%, not 50%, is the baseline we need to beat.

ROC CHART



A **ROC curve (receiver operating characteristic curve)** is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

Fig 14: Illustration of ROC Chart (Source = Wikipedia)

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The ROC curve allows us to evaluate classification performance at various thresholds settings. It also helps us visualize how well the model does at ranking observations based on their probability to fall into each class. (In this case, in ranking the Bank customers from most likely to least likely to churn). The Area Under the Curve (AUC) is simply the size of the area underneath the ROC curve. It tells how well a model can distinguish between classes. The higher the AUC, the better the model is at distinguishing between the classes. An AUC of 1.0 corresponds to a perfect model, and an AUC of 0.5 indicates that the model is no better at predicting than flipping a coin.

Source: [Toward Data Science](#)

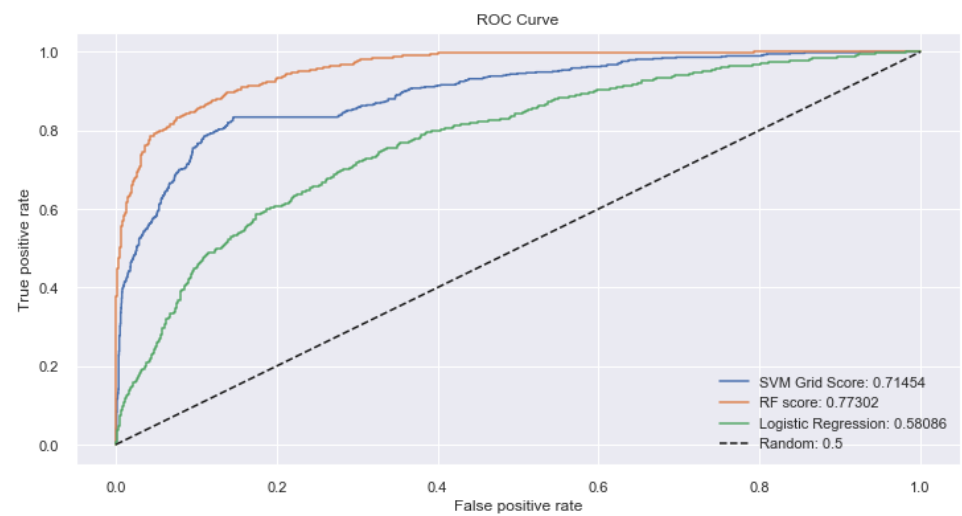


Fig 15 above: ROC/AUC compares the effectiveness of the Logistic Regression, the SVM and RF models. The RF model has an AUC of 0.77, vs. 0.71 in the SVM model, and only 0.58 in the Logistic Regression model. This metric suggests the Random Forest has the highest performance of the three.

Fig 16 below: Confusion Matrix for Random Forest “test-data” results:

| | | | | |
|---|---|-----------------|--|--|
| | Confusion Matrix for Random Forest | | | |
| | Predicted by Model | | | |
| Actual | Positive | Negative | | |
| | TP = 2373 | FN = 38 | | |
| | FP = 375 | TN = 209 | | |
| | | | | |
| TRUE | | | | |
| FALSE | | | | |
| We then add the TP and the TN then divide by our sample of 2995 | | | | |
| to get .86 which matches our weighted average in our RF | | | | |
| Classification Report. | | | | |

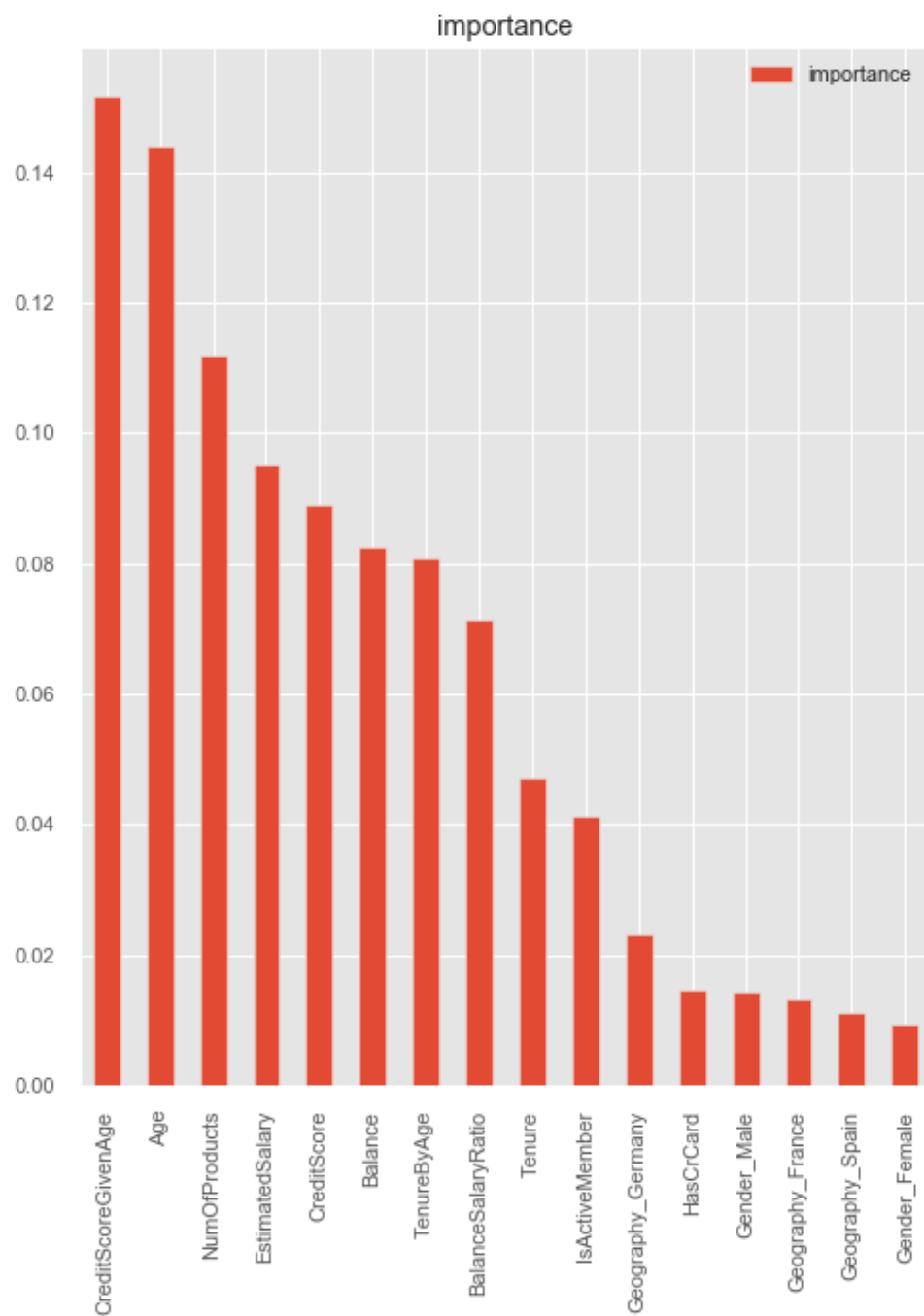


Fig 17 (above): I rank each Variable by its importance according to the “Random Forest” model as it was the best performing model. We clearly notice that Age and Bank Balance related features had more importance then, say, if the customer had a credit card or not or was from a certain country though Germany had a higher rate of Churn than France or Spain.

Chi Squared Test to prove or disprove our Null Hypothesis

The **Chi-square test** is intended to **test** how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Below I performed a Chi Squared Test to test my first Null Hypothesis that “Gender is not a determining factor to whether a customer churns.” We see from the results below the “p-value” is less than our significance of 0.05 so we **reject** the H0.

Observed Values : These are the values that were validated from our test (TRUE VALUES)

| | Stayed | Left |
|--------|--------|------|
| Female | 3404 | 1139 |
| Male | 4559 | 898 |

*3404 females stayed with the bank but 1139 left the bank
*4559 Males Stayed with the Bank but 898 left

Expected Values : Below is the expected given outcomes given our data so more females left in reality than what was predicted according to our data. More males were expected to leave though in reality fewer left the bank.

| | Stayed | Left |
|--------|-----------|-----------|
| Female | 3617.5909 | 925.4091 |
| Male | 4345.4091 | 1111.5909 |

Degree of Freedom:- 1
chi-square statistic:- 113.44910030392086
critical_value: 3.841458820694124
p-value: 0.0
Significance level: 0.05
Degree of Freedom: 1
chi-square statistic: 113.44910030392086
critical_value: 3.841458820694124
p-value: 0.0 - we see our P-value is less than our .05 Significance level so we reject our **H0**
Reject H0,There is a relationship between 2 categorical variables

Commented [CK19]: I will need help in trying to explain this.

Final Thoughts and Summaries:

Commented [RN(20): This is great.

- Through my examination of the small data set I did discover a few significant findings:
- There is no significant difference in the credit score distribution between retained and churned customers.
 - The older customers (over 40) are churning at a higher rate than the younger ones alluding to a difference in service preference in the age categories. The bank may need to review their target market or review the strategy for retention between the different age groups as losing older customers who typically have higher balances is a serious problem that needs resolution.
 - Bank members with an average tenure are slightly less likely to churn than those with either low or high number of tenure years.
 - The data shows that customers with higher balances are churning at a higher rate which is cause for concern for their lending capability. The bank could benefit from offering special programs when, say, a balance of \$75,000 and offer a higher rate of interest on a savings account or special investment privileges.
 - Neither the product nor the salary has a significant effect on the likelihood to churn.
 - More females have churned than males therefore perhaps the bank can initiate “Female Investment” education programs to address female concerns

- More credit card holders churn though most of the bank customers possess credit cards but it may benefit the bank to review their credit card benefits to stay competitive with other providers and initiate a “point system” in which points are converted to cash in your account.
- The bank can benefit from increasing incentives in keeping credit card holders.
- Salary has little effect on the chance of a customer churning.

Model Review

- My Random Forest Model performed the best as it was able to capture both the True Positives and True Negative outcomes .86 out of 100. This is a strong result and in this model an analyst is not required to scale/normalize the data.
- According to my Feature Selection results from my Random Forest model the most important determinants of Churn are Age, Gender and Bank Balance followed by Tenure and Geography.
- I think developing an accurate predictive model should be incorporated in every bank’s business plan as the closer a bank can get to modeling the characteristics of those who may leave, the more detailed a retention plan the bank can create and implement. Insights carefully extracted from data are invaluable to the health of the competitive nature of banks, especially credit cards, bank-balance specific programs and other key ancillary features a bank provides to keep their clients.

I would like to train banking data using the Naïve Bayes Classifier as this assumes independence between every pair of features and accounts for new changes relating to the features and is quick to run. I would also like to try K nearest neighbors as it is a classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point and is robust against noisy data. It will also work well for a Bank Data set with over 100,000 observations.

The study can be greatly improved with the following data since there are many unanswered questions:

- Would it be possible to obtain balances over time as opposed to a single date?
- What date did the customer exit?
- What types of products are the customers in? Do they leave as they are not happy with the products? What did competitors offer?
- Could they have exited from a product and not the bank?
- Does the bank have an investment division?
- Did the customer retire and consolidate assets elsewhere?

Of course, every business needs to perform analysis and take measures to prevent Customer Churn; considering the cost of acquiring each customer, a study should be an annual requirement.

