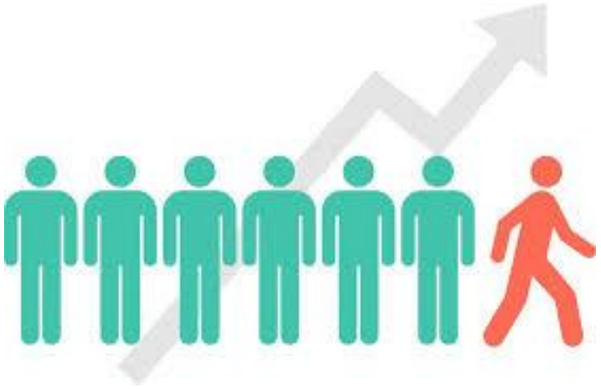
Springboard Data Science Capstone Project - 1

Predicting the Likelihood of **Customer Churn**

Milestone Report

Carolyn Massa

January 22, 2020



Contents

1 Introduction/Null Hypothesis 1

2 Data Acquisition and Cleaning 2,3

3 Data Exploration 4,5,6,7,8

4 Data PreProcessing/Feature Engineering 9

5 Model Fitting 10

6 Classification Reports 13

7 ROC Chart 13

8 Feature Importance/Hypothesis Conclusion 14,15

9 Final Thoughts 16

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 and 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.

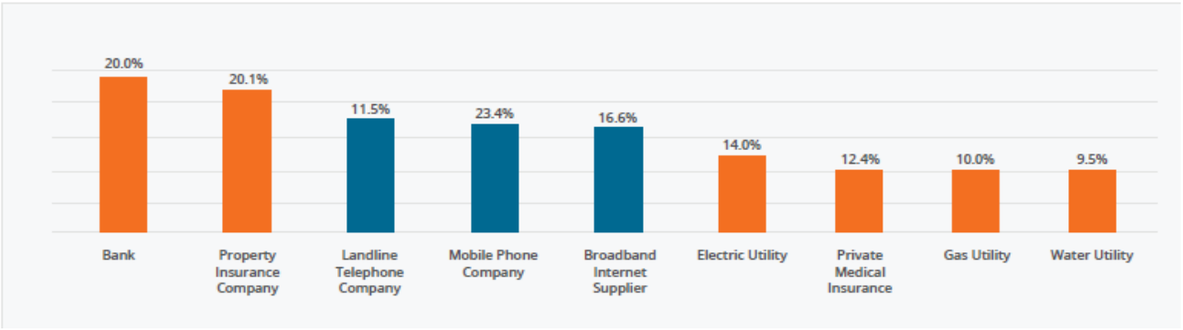
Most businesses are heavily affected by customer churn…from banks to online retailers, it matters and is mission critical to examine WHY the customer is leaving your business and putting their purchasing power into another organization.
You can measure your client churn rate in one or more of the following 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.

My "**Null Hypothesis"** is that it makes little difference in consumer behavior or trends as, according to "Call Miner" Index below banks have an average annual churn rate of 20% of their valued customer base. That being stated, I believe that the Bank "Balance to Salary" ratio of each customer has little impact on whether the bank customer will leave or not. I will test this hypothesis 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 their 10,000 customers and their # of who churned in a 10-year time frame.

The Data set below has 13 variables relating to 10,0000 customers from which to develop a predictive churn model from. Below I process the data and start my Exploratory process. My python code can be reviewed here along with both animated presentation and Powerpoint slide Presentation here.

The 13 variables are here as referenced by the INDEX:

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 period as opposed to a single date?
- What date did the customer exit?
- What types of products are the customers in?
- Could they have exited from a product and not the bank?

To begin I need to verify the type of data, what % is useable and look for patterns; I notice that of the 10,000 customers of Word Bank, 2,037 have churned in the past 10 years which is a 20% Churn Rate; relatively high by business standards. I also discover that the average tenure if 5 years with the bank and the average age who leaves is 44 years old. The average age of the Global Bank's customers is 37 years old with the youngest being 18 and the oldest being 92.

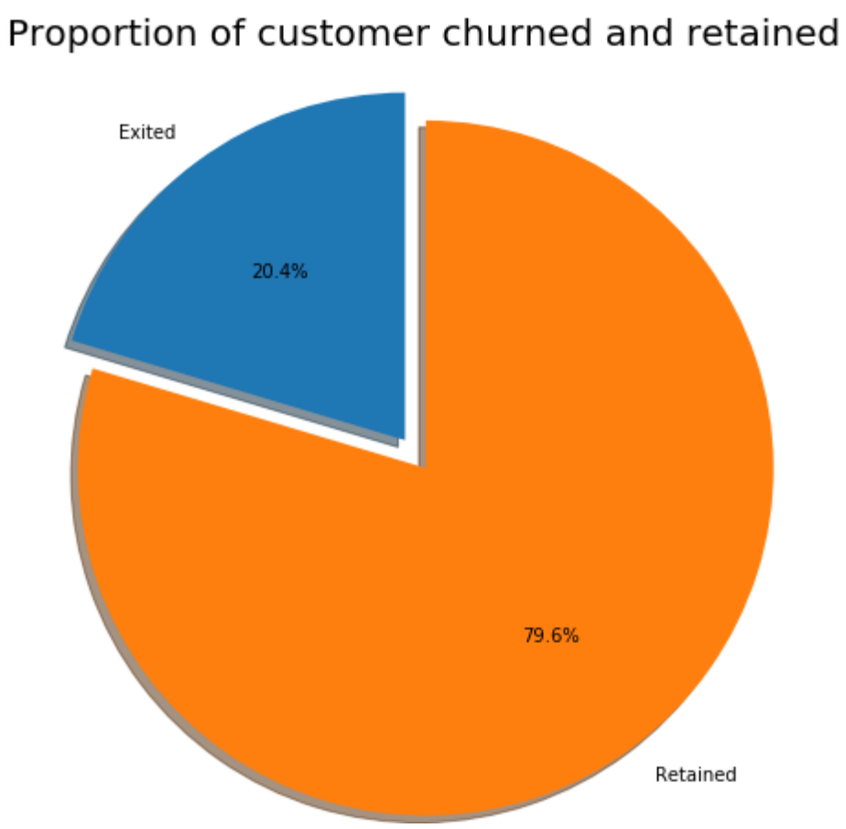


Fig 1 As you can see from the figure above 20.4% of 10,000 customers have churned over the past 10 years which is 2, 037 former bank customers.

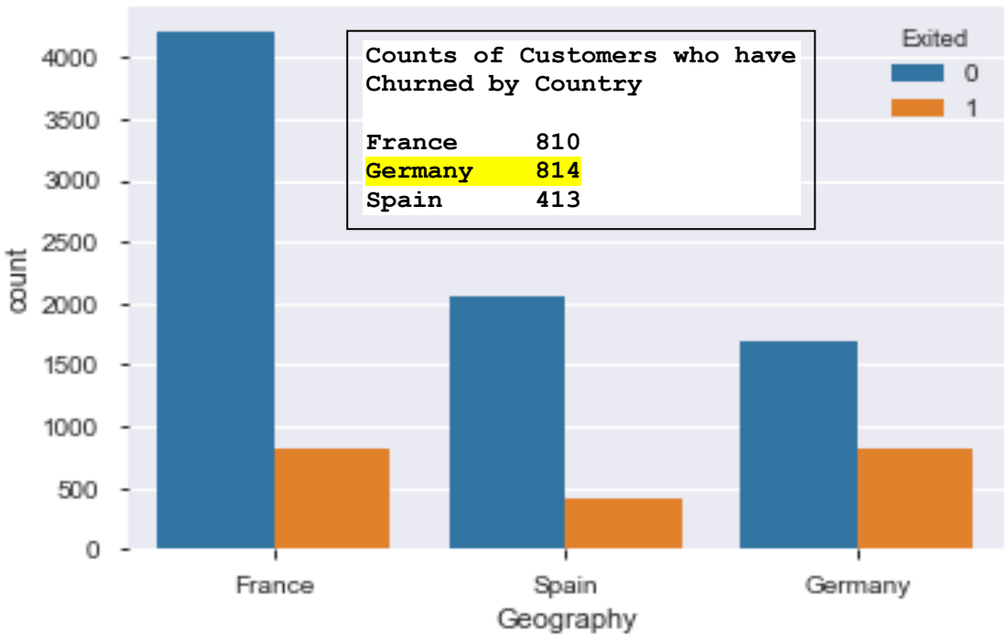


Fig 2 above as you can see from the figure above of the 2,037 who have left the bank 814 come from Germany, 810 from France, and Spain has a total of with 413 members who have left the Bank.

I next examine by data which does not missing fields or outliers. I dropped 3 columns as they played no relevance in my research. Those 3 columns were 1) Surname 2) CustomerID 3) RowNumber.

2 Exploratory Data Analysis

Introduction to the cleaned data:
After I verify that I have no missing values and/or irrelevant data; I get to work at seeking patterns and running correlations.

I compare the Churn rates by Gender, Geography, Credit Card Holder, and Product Participation which are **"Categorical Variables"** meaning they are not **CONTINUOUS** as in the Gender is either male or female and does not change.

I use the python "describe" function to cumulatively review my data set and provide the minimum and maximum values, the mean, the bottom 25% percentiles, 50% percentiles which is the same as the mean, the top 75% percentiles and the maximum values contained in my continuous variables. It also displays the standard deviation for all observations as well as my dataset datatypes.

I am curious to I check for outliers where 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.

Below I compare 4 of my Variables against their Churn Rate:

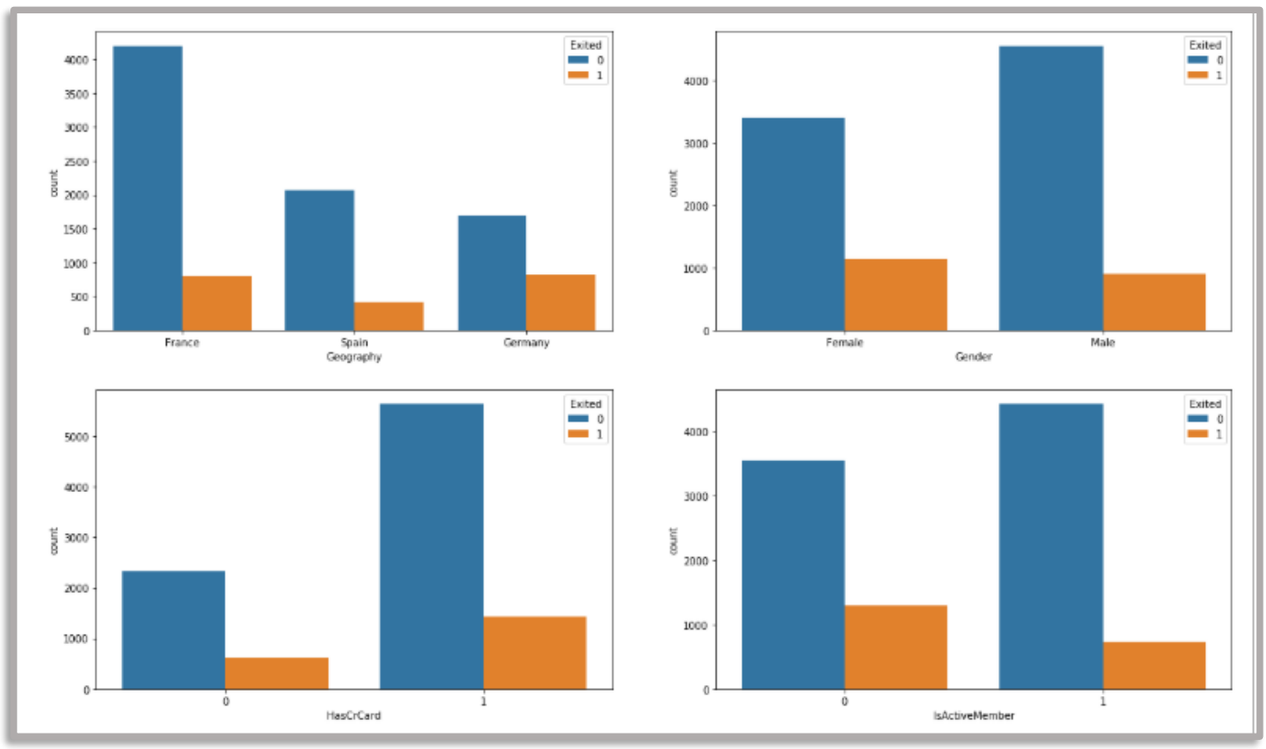
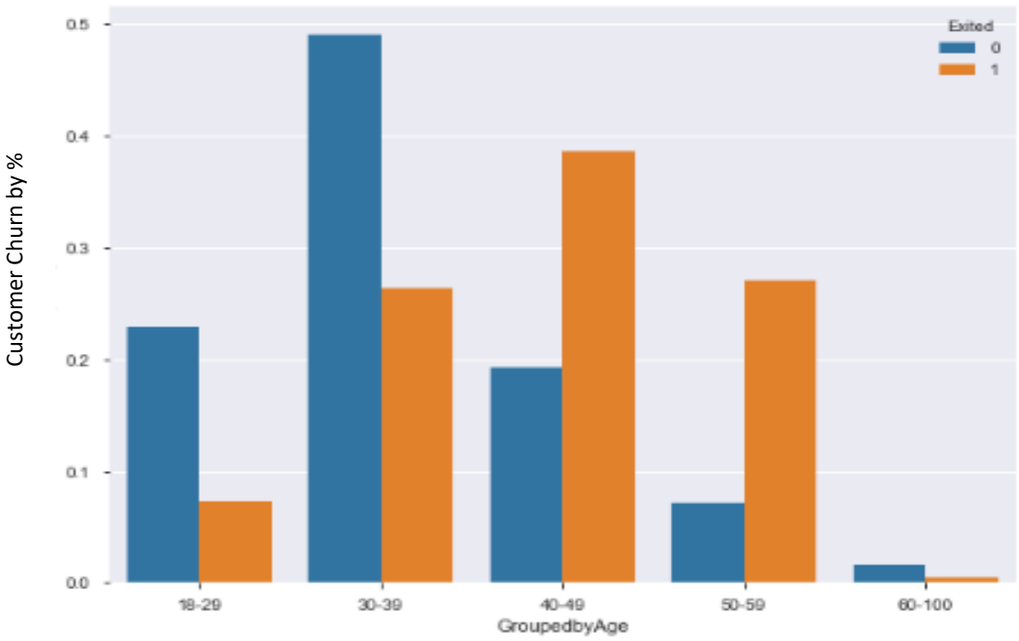
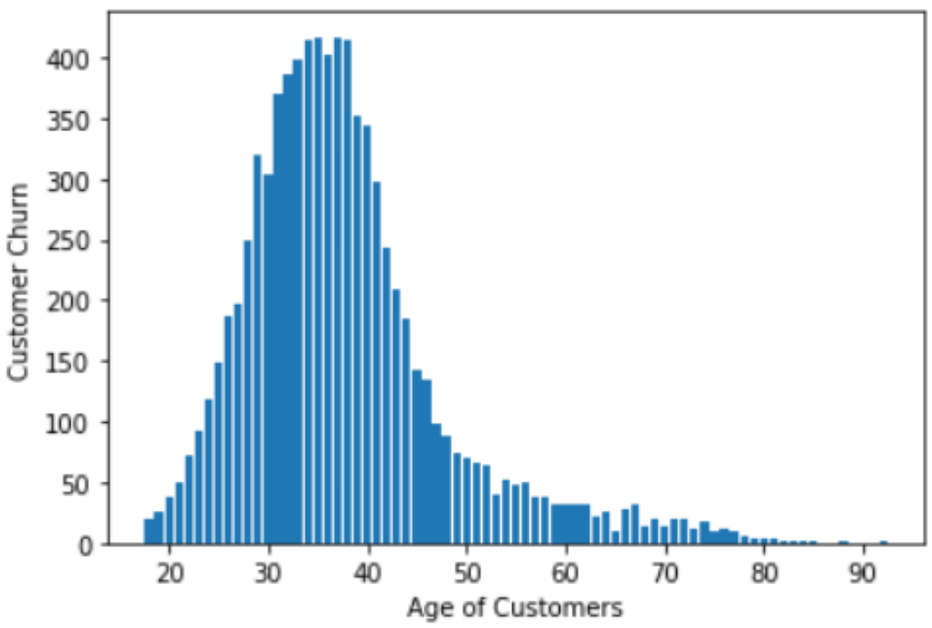


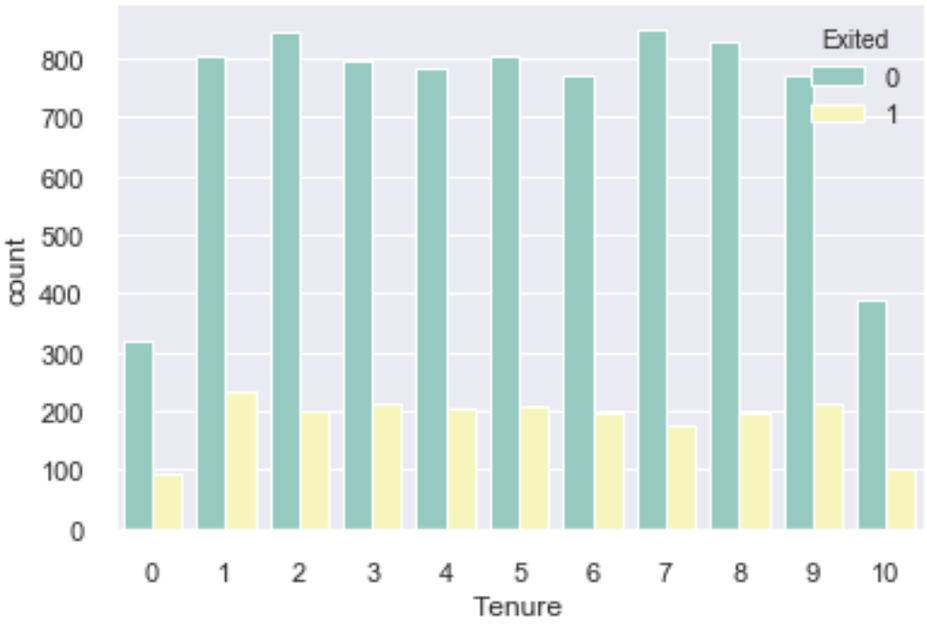
Fig. 3 above: I discovered the following:

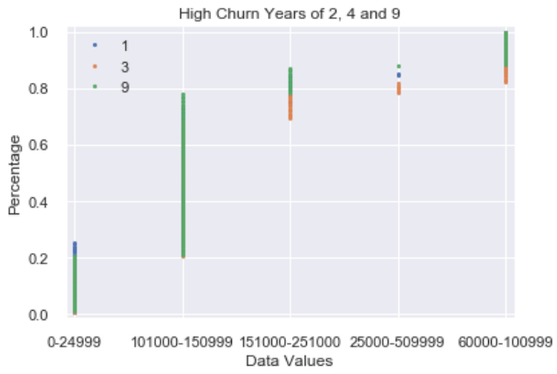
- As mentioned earlier, the country of France has the least churn rate and Germany has the highest. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank possibly having a problem (maybe not enough customer service resources allocated) in the areas where it has fewer clients.
- The proportion of female customers churning is also 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 greater churn. The overall proportion of inactive members is quite high suggesting that the bank may need a program implemented to turn this group to active customers as this can have a positive impact on the customer churn.





Fig(s) 4/5 Above: Age groups that has the highest “churn” rate are those between 35 and 50.





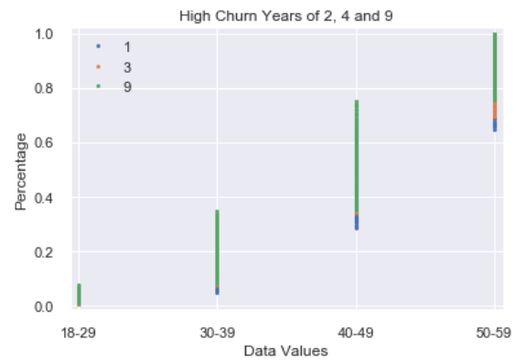


Fig (s) 6 Above: I see that the Bank has its highest customer count leave the bank in the **years 2, 4 & 10** and its lowest in its **11th year.** The bottom plots show the customers Age and Bank Balances that left during these years. **Note that the plot's legend shows '1, 3 & 9' as Python will 0 index the data. I see that that during these 3 years the customers with higher balances left the bank and the age groups between 40 and 49 were more significant.

Now I compare the **"Continuous Variables"** (which means are **non-categorical**) against Churn below:

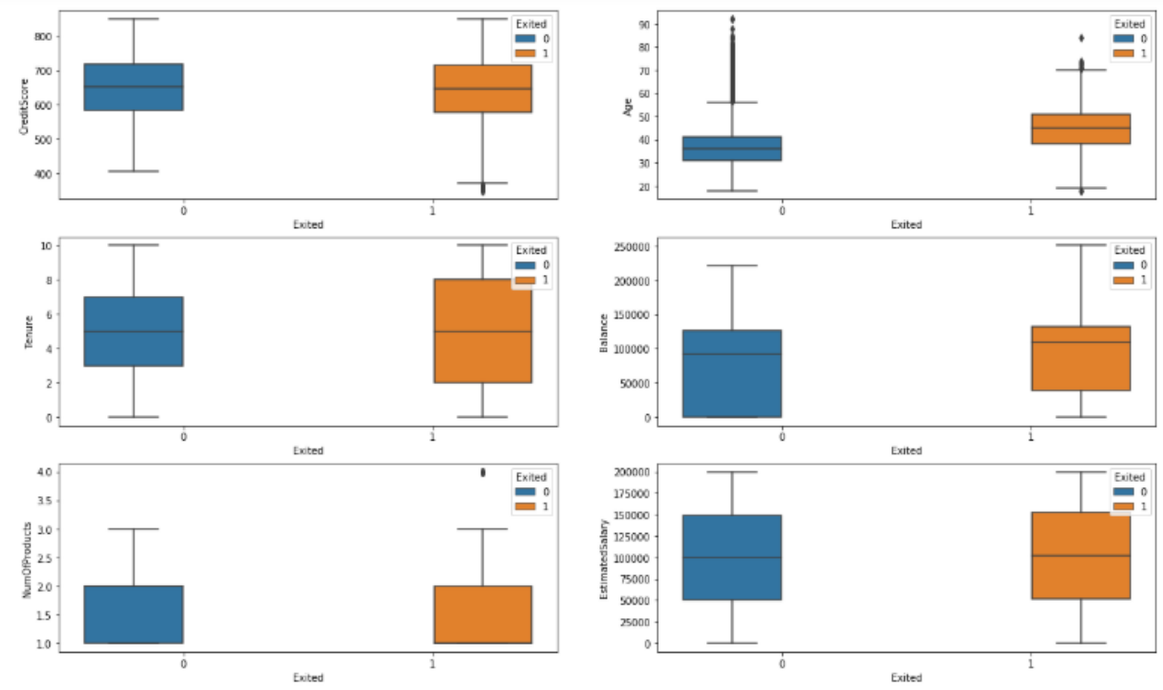


Fig. 7 above: I discovered the following:

Interestingly, neither the product nor the salary has an impactful effect on the likelihood to churn. There is no significant difference in the credit score distribution between retained and churned customers.
Regarding the tenure, the average tenure of client is 5 years and this type of client had a lesser likelihood to churn where those customers on opposing spectrums (spent little time with the bank to a lot of time with the bank) were **more 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 churned customer is $91,000 with an average bank balance of $101,465.
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.

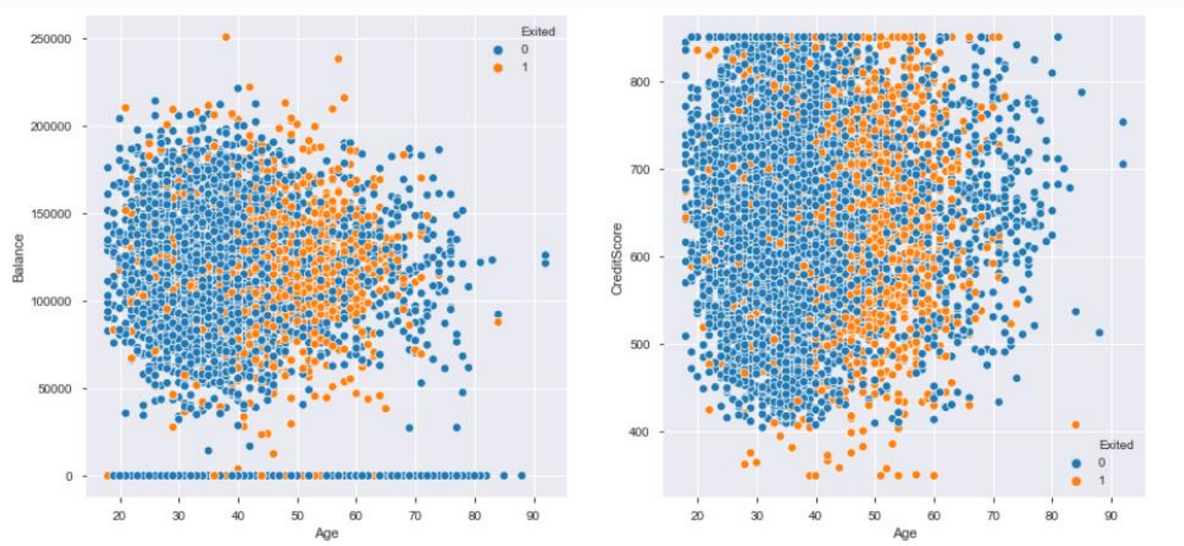


Figure 8: Since Age and Bank Balance presented some insights, I plotted a swarm plot to display clearer results. We see that the orange dots show clearly the Age Range between 35 and 70 had higher likelihood to churn while the bank account holder's credit score has little relevance. There are more clusters of Churned Customers in the Bank Account Balance scatter plot between $100,000 and $150,000.

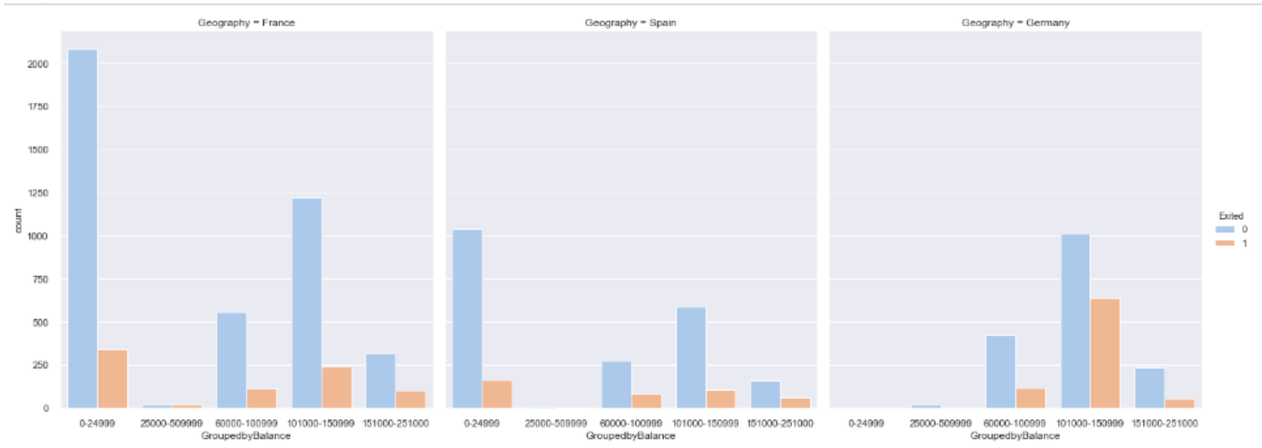


Figure 9: In this bar chart I review Churn by each of the 3 countries, it is again clear that those bank member with balances in between 101,000 and $150,000 are more likely to Churn. Germany has the highest Churn rate of the 3.

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

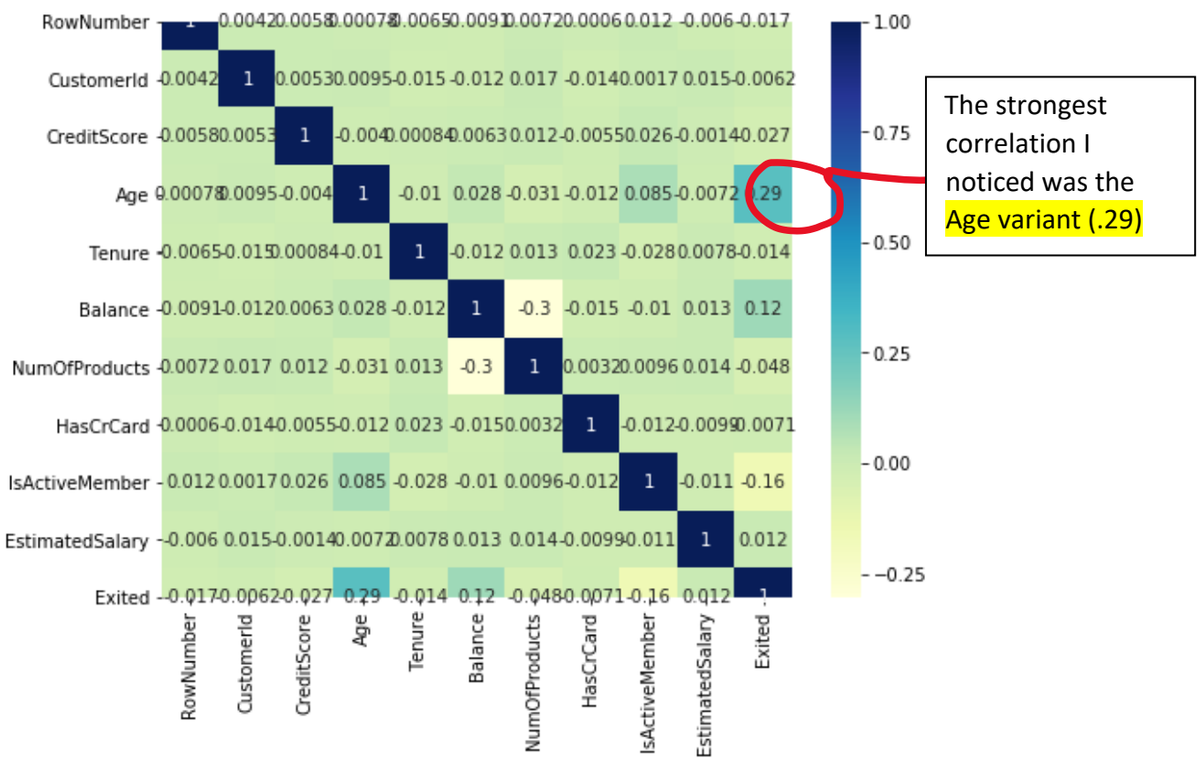


Fig: 10: In this Heatmap I used **Pearson's correlation coefficient** which is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of 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 customer "churn" is either they DO churn and leave, or they do not leave. This means that 0 = not leaving and 1 = leaving. I used supervised machine learning algorithms to build a predictive model. Furthermore, since there are only two outcomes (or classes) in the data (0 and 1), I use binary classification algorithms. The models are trained using the 70% of the data and the remaining 30% is used to evaluate the performance of the models. I had to perform "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 with a primal fit and Support Vector Machine Learning (SVM) with an RBF Kernel as my data is not liner and what the RBF kernel SVM actually does is to create non-linear combinations of features to uplift your samples onto a higher-dimensional feature space where you can use a linear decision boundary to separate your classes.

Prior to running my algorithms with my data, I needed to implement 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: In the dataset, there are some variables with numerical values, some variables with categories and some variables with binary values (0 and 1). For numerical and binary variables, we do not worry about labeling. However, we perform label encoding for the categorical variables. This step is carried out on the whole dataset. I **"**Hot Key" encoded the following variables: **Gender and Geography** to transform them to binary using a "for"/" if" statement. I performed **"Hot Label**

Encoding" where I changed the value "0" (no churn) in the two categorical variables "Has Credit Card" and "Is Active Member" to a -1 to show a negative relationship more clearly.

New Variables: I created 3 new variables that correlated to each other to better: 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 (70) and test (30) 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.

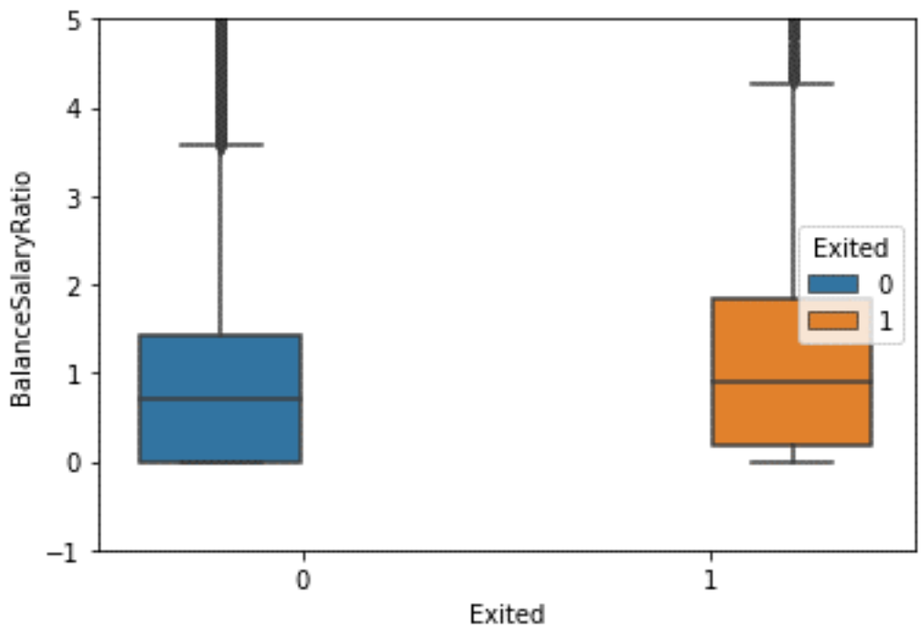


Fig 11 above: Here I combined **Bank Balance** and **Salary** to show ration to further examine correlation with Churn and enforce the evidence that higher bank balances have more Churn.

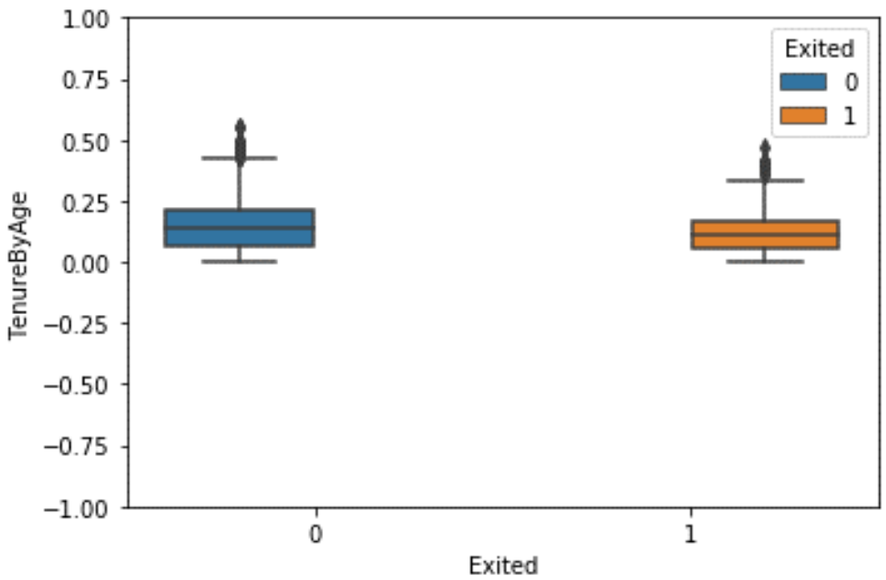


Fig 12 above: Tenure is a function of age, so I combined the two to check for trends

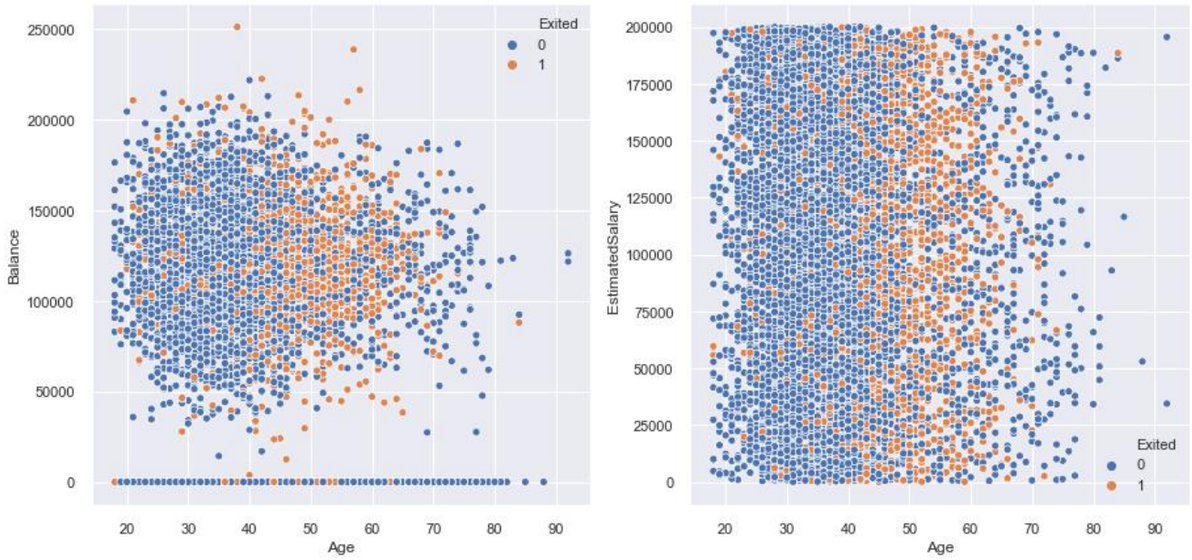


Fig 13: Again, the scatter plot above shows a clear trend in Age of those bank members who left the bank with concentrated orange dots in the 40 to 65 age groups.

Scaling the Data Set: I use the min/max operations to scale my "continuous variables" to eliminate unnecessary variances. Min/Max is also known as **"Normalization".** This formula behind this is below: These "normalization" techniques help in comparing corresponding normalized values from two or more different data sets in a way that it eliminates the effects of the variation in the scale of the data sets i.e. a data set with large values can be easily compared with a data set of smaller values. I used it to process my data as I will apply Logistic Regression, Support Vector Machine Learning and a Random Forest Method for my algorithm and I want to eliminate variants in my data. SVM is intrinsically two-class. For multiclass problem you will need to reduce it into multiple binary classification problems. For example, if I am working with and comparing two different data sets and one has much larger values than the other, I would want them to be standardized between a range of 0 to 1. The Normalization techniques are not typically used in Random Forest or K Nearest neighbors as one is working with decision trees which are not affected by scaling your data.

Normalization Formula

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

Modeling Pipeline

Next, I build a **"Data Pipeline" In Python** which allows the me to transform data from one representation to another through a series of steps. In other words, to ensure my hot encoding, min/max normalization and both my categorical and continuous variables continue in the test and train modeling.

Model Fitting

For the Bank Churn dataset, since it is relatively small (10,000 records and 13 variables which I split into 2 sets: Training (70%) and Testing (30%) 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. It is named as **'Logistic Regression'**, because it's underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the **Logit function** that is used in this method of classification which uses the Sigmoid function.

An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a Sigmoid function, which takes any real value between zero and one. It is defined as

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

2)Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. SVM is mostly used in classification problems. 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. SVM is better suited as I need a way to separate my data into CHURN or NO CHURN and Logistic Regression uses a straight line. I used a large C to output low bias and high variance and to instill a regulation parameter and achieve a higher level of accuracy.

3) Random Forest works well with a mixture of numerical and categorical features which the Bank Churn data has. When features are on the various scales, it is also fine. Roughly speaking, with Random Forest you can use the data as it is. Random Forest uses a large # 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.
Source: sklearn documentation.

Logistic Regression

	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

SVM

	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

Random Forest

	precision	recall	f1-score	support
0	0.91	0.99	0.95	2411
1	0.95	0.58	0.72	584
accuracy			0.91	2995
macro avg	0.93	0.79	0.84	2995
weighted avg	0.92	0.91	0.90	2995

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/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all bank customers that are labeled as having remained at the bank, how many remained with the bank? High precision relates to the 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 - yes. 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 is approx. 81% accurate. We have 0.78 precision score which is also strong as well as a recall of 0.81 which is good for this model as it's above 0.5. Our F1 score is 0.90.

Support Vector Machines: The accuracy for our SVM model is 0.88 which means our model is approx. 88% accurate. We have 0.88 precision score which is also strong as well as a recall of 0.88 which is good for this model as it's above 0.5. Our F1 score is 0.86.

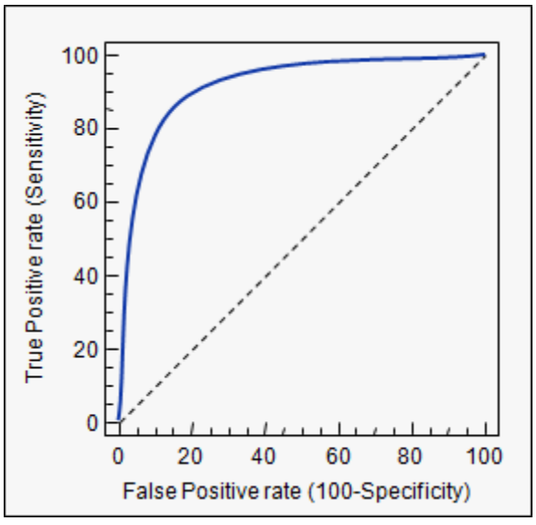
Random Forest Model: The accuracy for our RF model is 0.91 which means our model is approx. 91% accurate. We have 0.92 precision score which is also strong as well as a recall of 0.91 which is good for this model as it's above 0.5. Our F1 score is 0.90.

As we see from the Confusion Matrix Below that our RF Model resulted in the following:
From the sample of 2995 the RF model found 2387 predictions to be true positives which is a "NO CHURN" result and 329 as a True Negative which is a "Churn" result.

	CONFUSION MATRIX FOR RANDOM FOREST PREDICTED	
ACTUAL	TRUE	FALSE
FALSE	True Positive = 2387	False Negative = 24
TRUE	False Positive = 255	True Negative = 329
TOTALS	2642	353

We add the TP and the TN then divide by our sample 2995 to get .91 which is our RF weighted average score

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. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and the AUC represent degree or measure of separability. It tells how much a model is capable of distinguishing between classes. The higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

Source: Toward Data Science

To evaluate effectiveness of my model I will be using the following metrics.

Accuracy = TP+TN/TP+FP+FN+TN

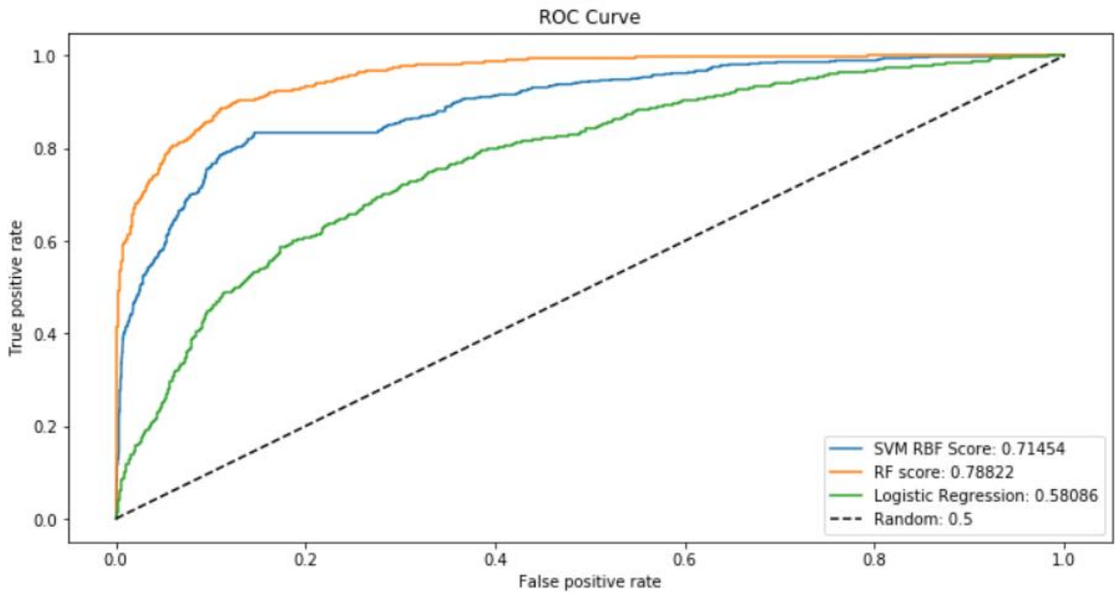


Fig 15 above: ROC/AUC compares the effectiveness of the Logistic Regression, the SVM and RF models. This accuracy is measured by the area under the ROC curve. An area of 1, for instance, is a perfect score while .5 is meaningless. So, for all the customers, .78 of the instances are correctly classified in the **RF model** .71 are correctly classified in the SVM model, and only .58 of the instances in the study are correctly classified in the Logistic Regression model.

Here is a rough guide for classifying the accuracy of the ROC/AUC using a diagnostic test is the traditional academic point system:

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

Source UNMC

Now that we now the accuracies of our 3 Models let's look at how each Data Point contributes to the level of Churn for our Bank.

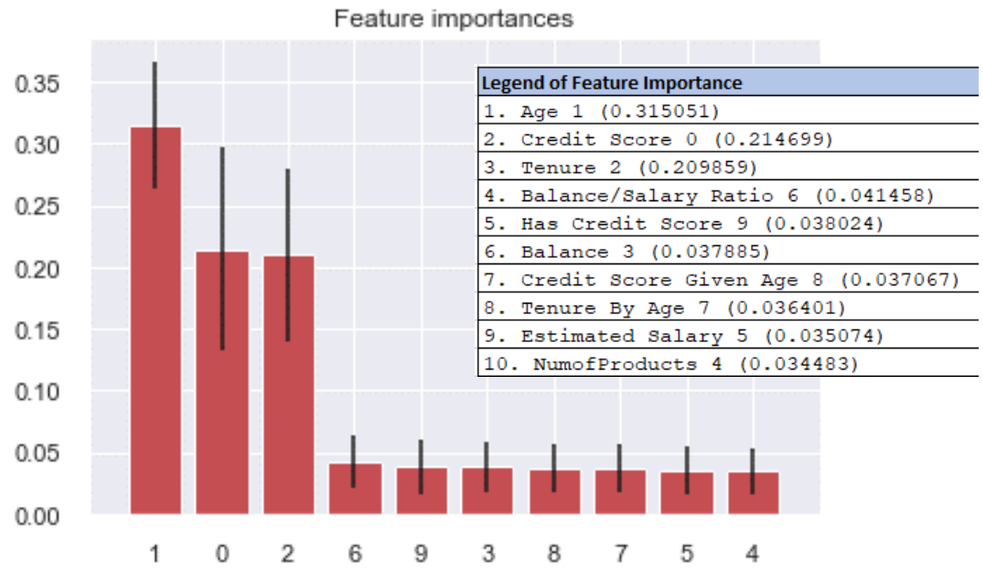


Fig 16 above: I rank each Variable by its significance

Null Hypothesis Conclusion

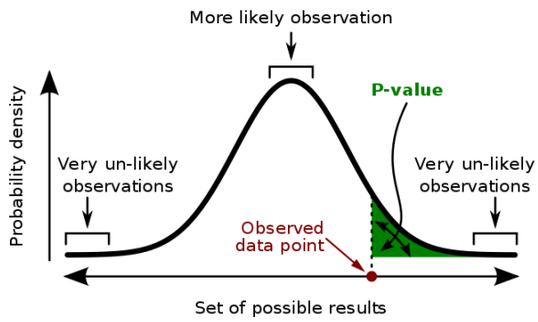
In the beginning of the Milestone Report I make a Null Hypothesis that the *"BankSalaryRatio"* variable had no impact on whether the bank customer left or not. I run a "p- value" Test to investigate further. The smaller the p-value the higher the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation. The null hypothesis is rejected. If any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value. This is referred to as the level of significance and is set by the researcher before examining the data and is arbitrary. It typically ranged from .05 to .001. Here I review the results and determine that since "Balance" and "BalanceSalaryRatio" are below .05 I will reject my Null Hypothesis.

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | 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.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Fig 17 above: p-value illustration

const	6.320921e-04
CreditScore	2.087470e-01
Age	5.318891e-50
Tenure	8.387585e-01
Balance	2.091051e-02
NumOfProducts	6.449643e-02
EstimatedSalary	9.835981e-02
BalanceSalaryRatio	3.140875e-02
TenureByAge	8.052536e-01
CreditScoreGivenAge	4.167467e-01
HasCrCard	9.617095e-01
IsActiveMember	1.444840e-26
Geography_Spain	8.212921e-02
Geography_Germany	1.106078e-09
Geography_France	4.471835e-02
Gender_Female	8.469826e-08
Gender_Male	8.469826e-08

Fig 18 left: The "p- value" results of my test data. These clearly indicate that the p-value of *"BalanceSalaryRatio"* is below .05. I will reject my "H0" and move to a "HA" where I will make a hypothesis that the "BalanceSalaryRatio" variable in my study does impact Customer Churn**.**

Final Thoughts and Summaries:

Though 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 35) 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
- 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.

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.