

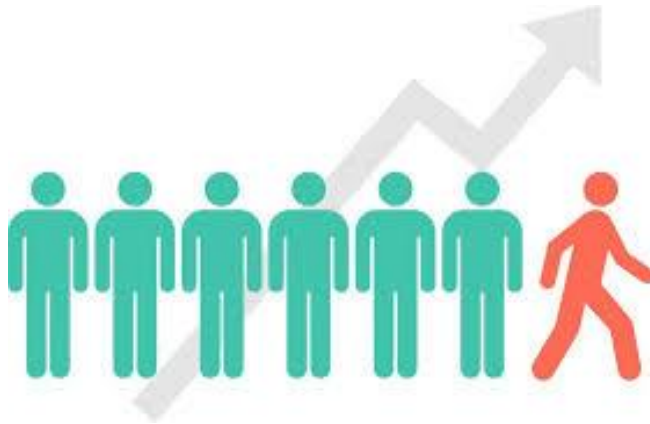
# Springboard Data Science Capstone Project - 1

## Predicting the Likelihood of **Customer Churn**



**Carolyn Massa**

November 29, 2019



# 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 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.

## 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 14 variables relating to 10,000 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](#).

**The 14 variables are here as referenced by the INDEX:**

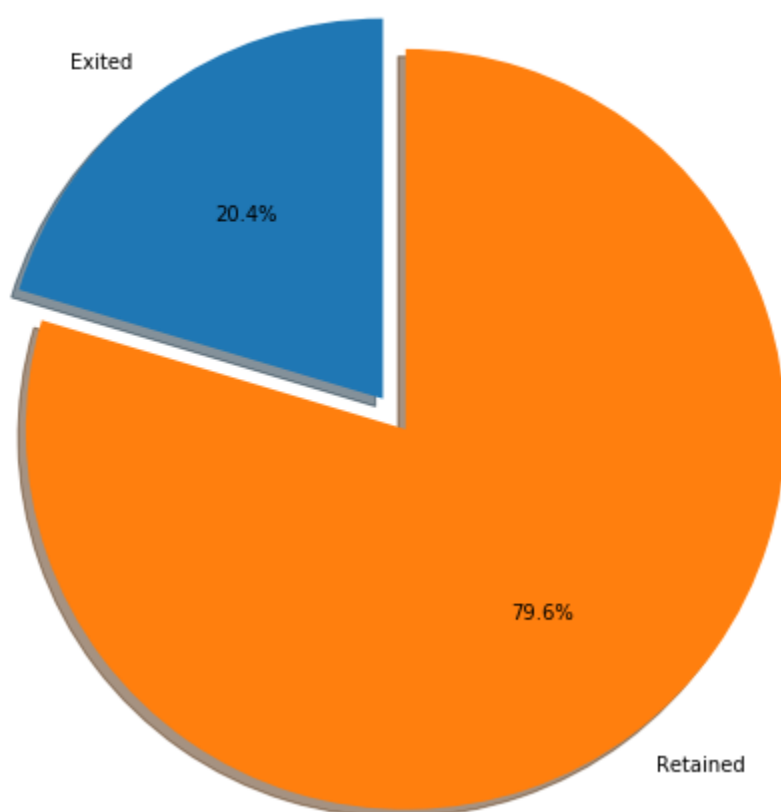
```
Index ['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
      'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts',
      'HasCreditCard',
      'IsActiveMember', 'EstimatedSalary', 'Exited'],
dtype='object')
```

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 is 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.

### 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 10 years which is 2,037 former bank customers.

I examined the data and has 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

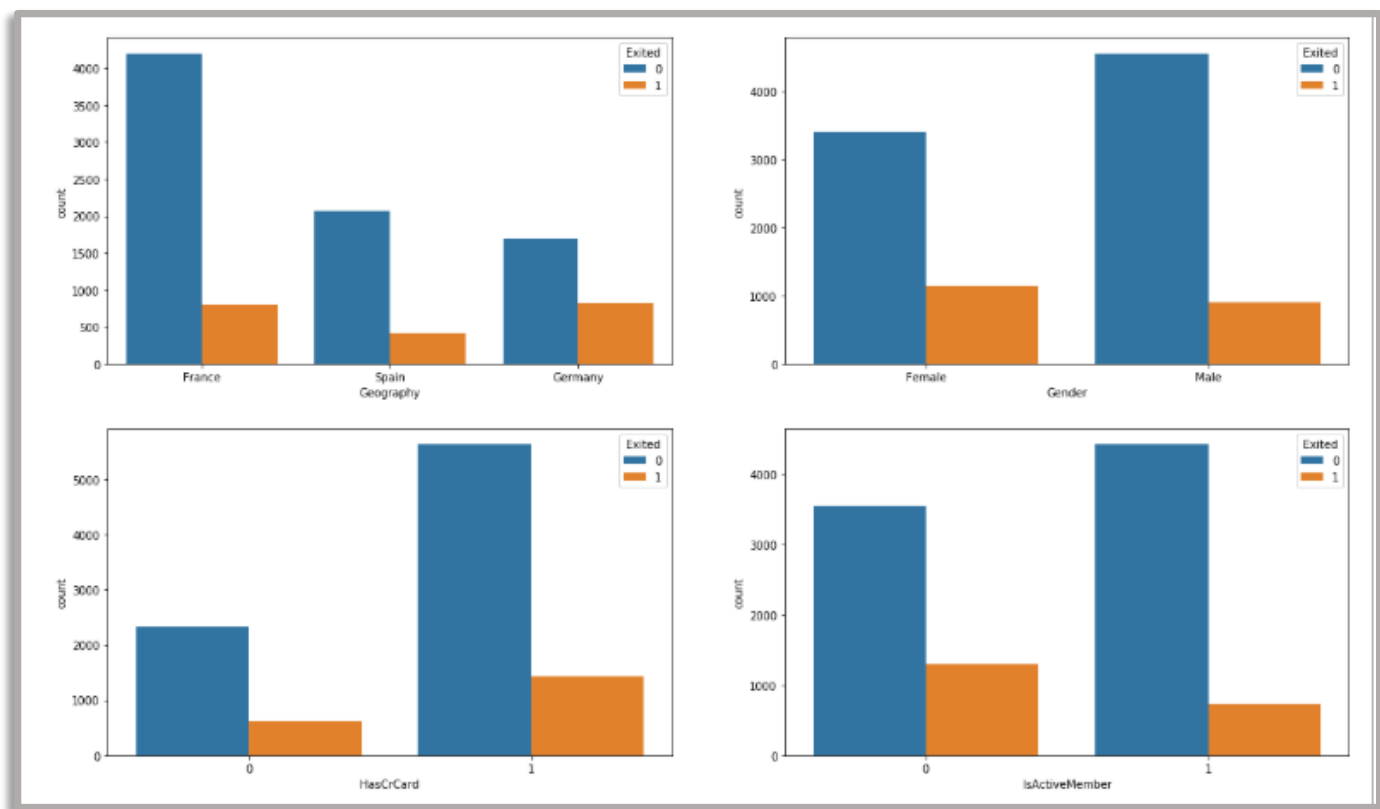
Introduction to the cleaned data:

After I verify 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 Gender is either male or female and does not change.

CreditScore	Age	Tenure	Balance
10000.000000	10000.000000	10000.000000	10000.000000
650.528800	38.921800	5.012800	76485.889288
96.653299	10.487806	2.892174	62397.405202
350.000000	18.000000	0.000000	0.000000
584.000000	32.000000	3.000000	0.000000
652.000000	37.000000	5.000000	97198.540000
718.000000	44.000000	7.000000	127644.240000
850.000000	92.000000	10.000000	250898.090000

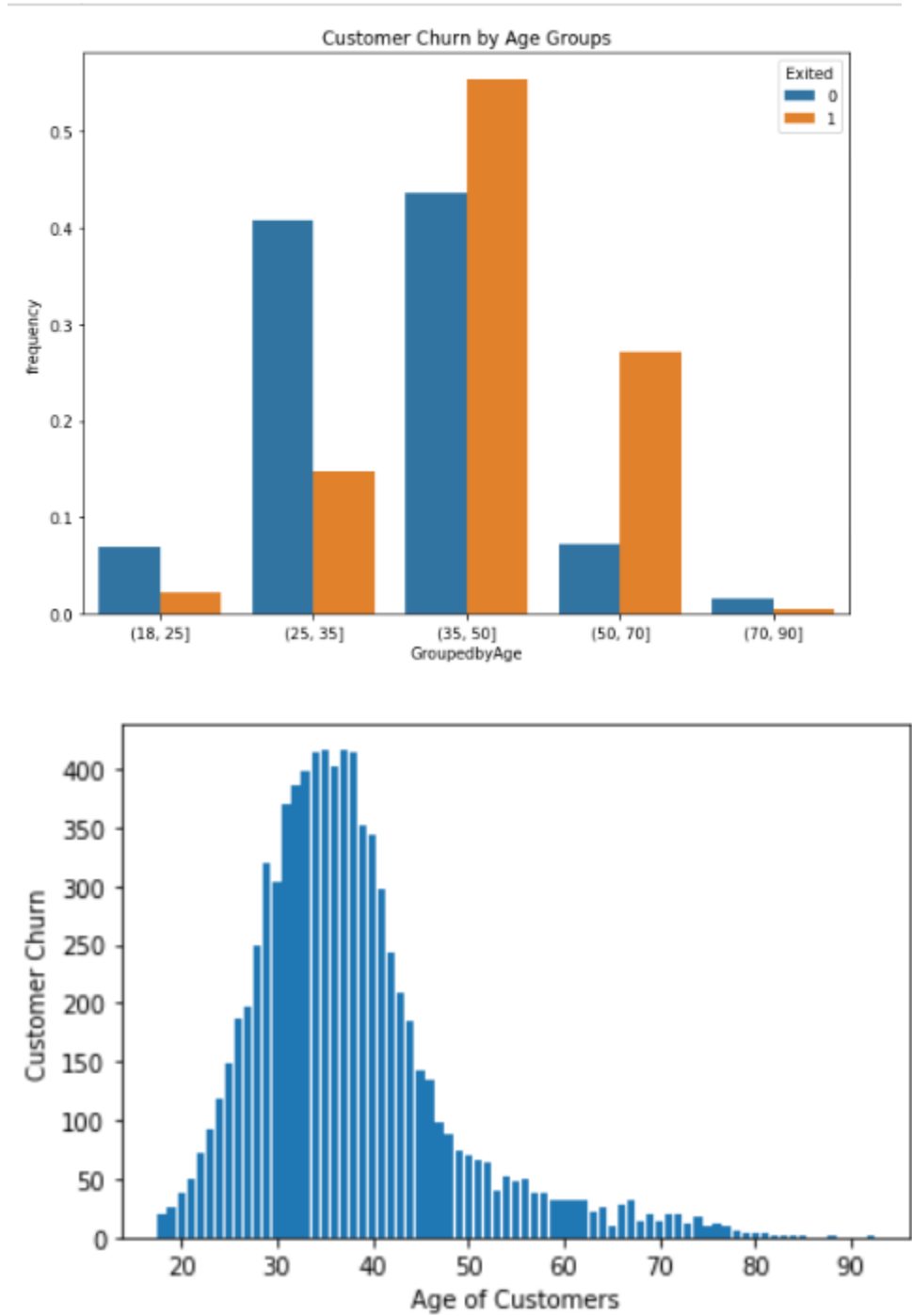
I check for outliers where the oldest customer is 92 and the highest bank balance is \$250,898.



**Fig. 2** above: I discovered the following:

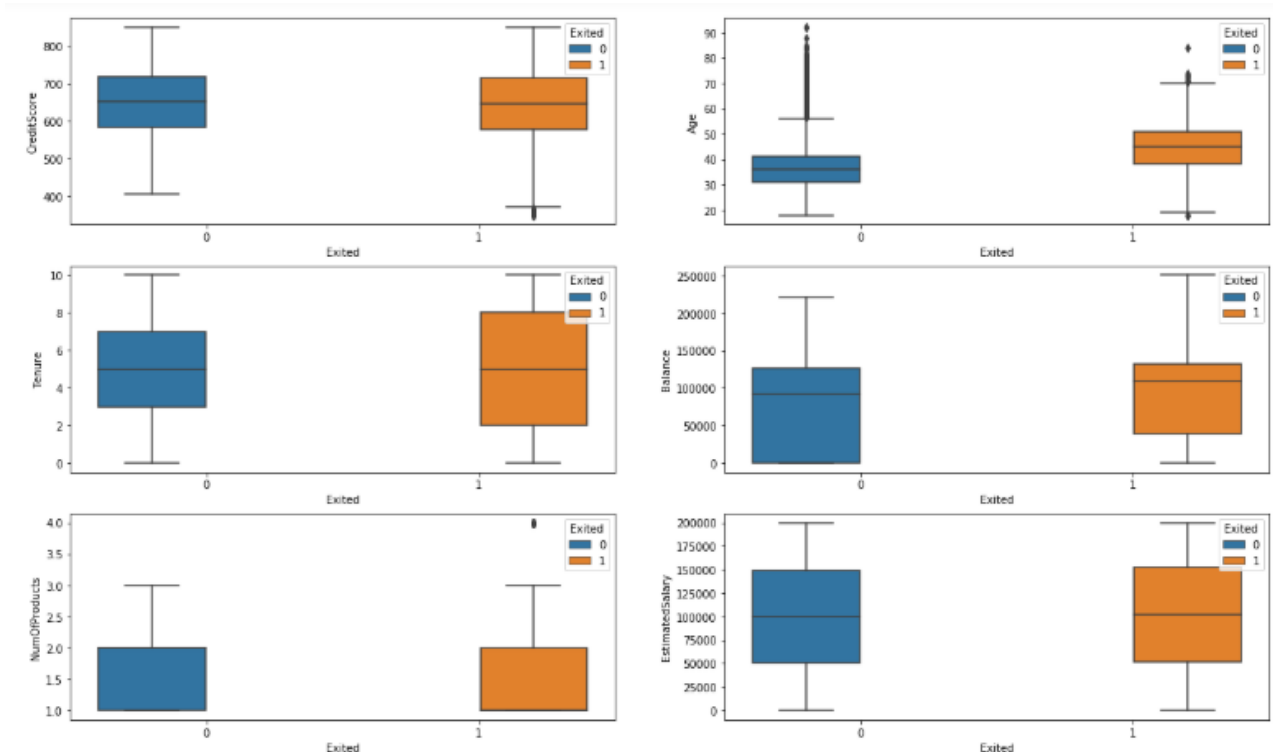
- The country of France has the least churn rate and Germany has the highest. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank possibly having a problem (maybe not enough customer service resources allocated) in the areas where 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. The majority of 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 3** Above: Age groups that has the highest “churn” rate are those between 35 and 50.

Now I compare the “**Continuous Variables**” (they are non-categorical) below:



**Fig. 4** 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 tenured client which is 5 years 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 more 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.

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:

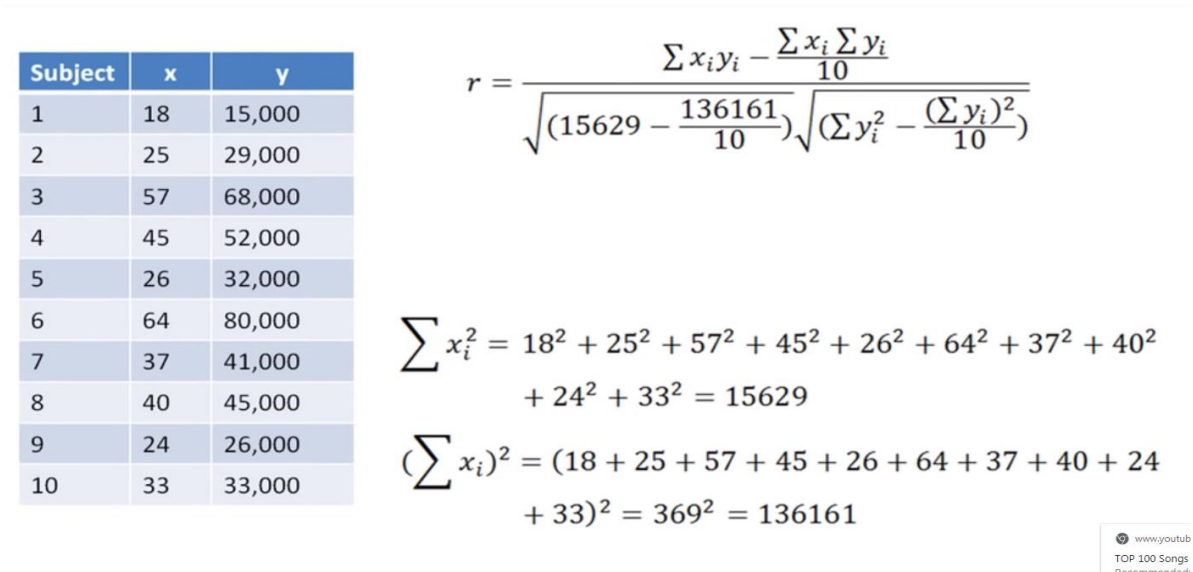


Fig 5 above: The Formula for Pearson’s R which its results are displayed below

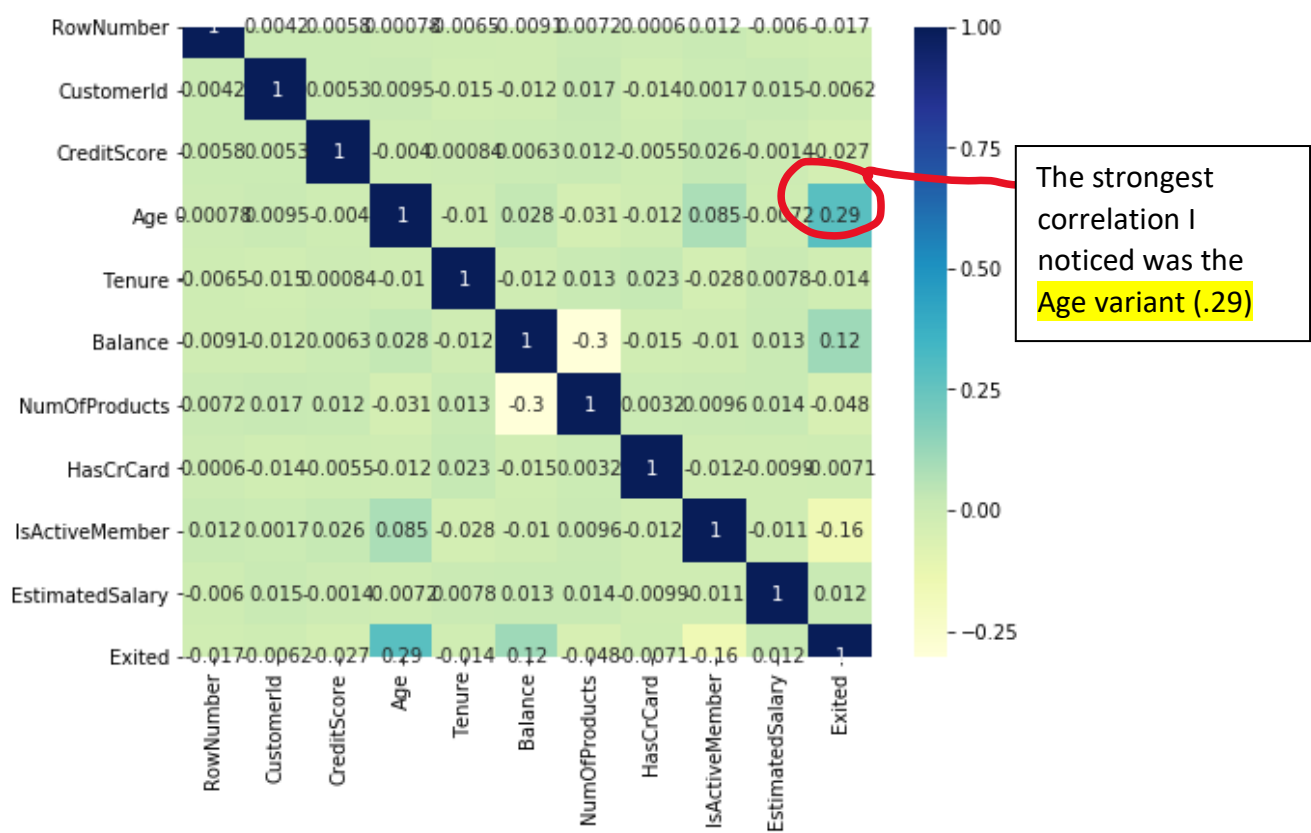


Fig: 6: In this Heatmap I used **Pearson's correlation coefficient** which is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. Note that a score of 1.0 is perfect correlation and -1.0 is negative correlation.



## 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 some “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 linear 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.

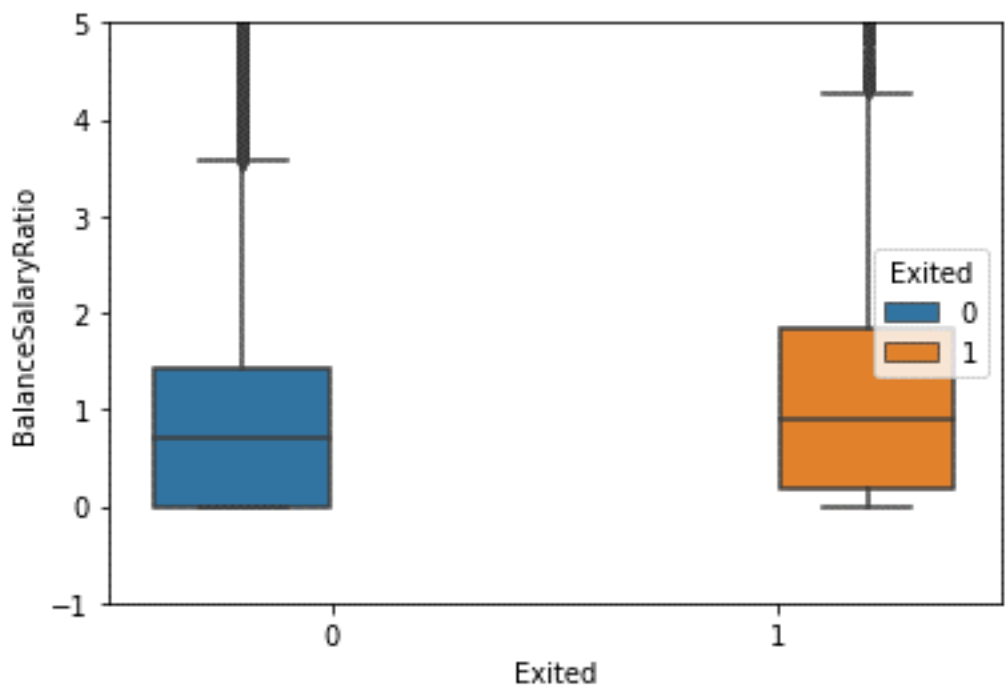
Prior to running my algorithms with my data, I needed to implement a few pre-processing steps to enhance my study and increase its accuracy. We outline these steps below. Note that some of the steps are not required (or not good for the best results) in some algorithms, but we list below all the pre-processing 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.

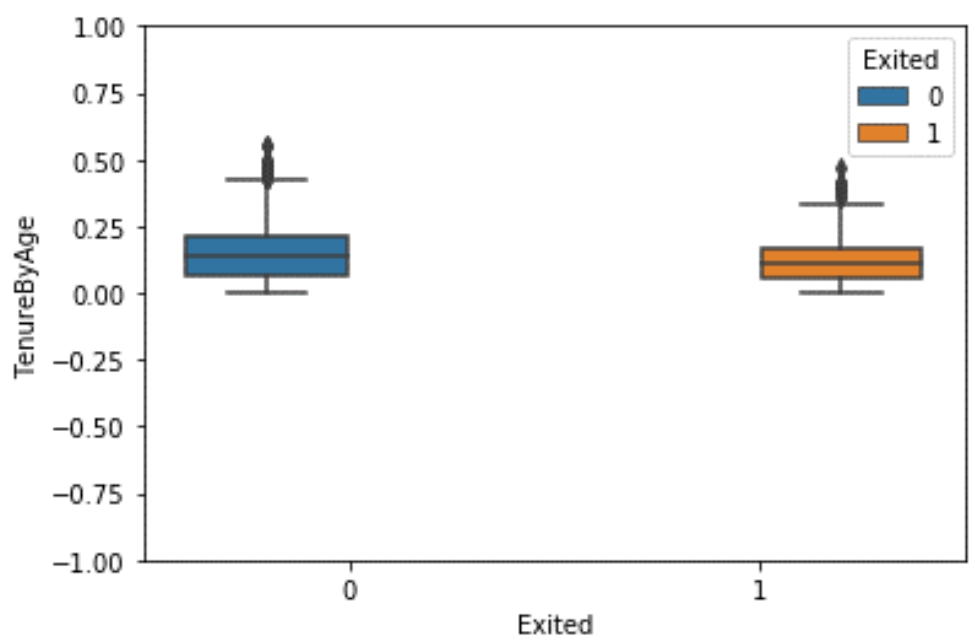
**New Variables:** I created 3 new variables that correlated to each other to better : Balance/Salary, TenurebyAge, and CreditScore Given Age

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



**Fig 7 above:** Tenure is a function of age, so I combined the two to check for trends

### Normalization Formula

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

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 both Support Vector Machine Learning

I also 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

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 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.

## 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 14 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 algorithm 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. SVM also maximizes margin, so the model is slightly more robust, but more importantly: SVM supports kernels, so you can model even non-linear relations. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier which my data requires to build the **Churn/No Churn** model as mentioned earlier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

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.

1) **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:**

**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

**RF**

	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

**Explanation of the classification reports:**

This simply tells us the percent that are correct of all instances that are classified as positive. So, for all the “1” classifications that represent churn 81% are correct in the SVM model, 91% are correct in the RF model and 81% are correct in the Logistic Regression model. The means that this % in the weighted recall score actually did churn. It is the ability of a classifier not to label an instance positive that is negative. For each class, it is defined as the ratio of true positives to the sum of true and false positives.

The  $F_1$  score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. The weighted average of  $F_1$  should be used to compare classifier models, not global accuracy. As we know a score of .5 is no better than flipping a coin and the Logistic Regression only has a score of .29 the RF model has .72 and SVM has .59.

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. I used 30% of my data set to train so the weighted and macro averages are shown as they are calculated with 3000 instances.

## 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

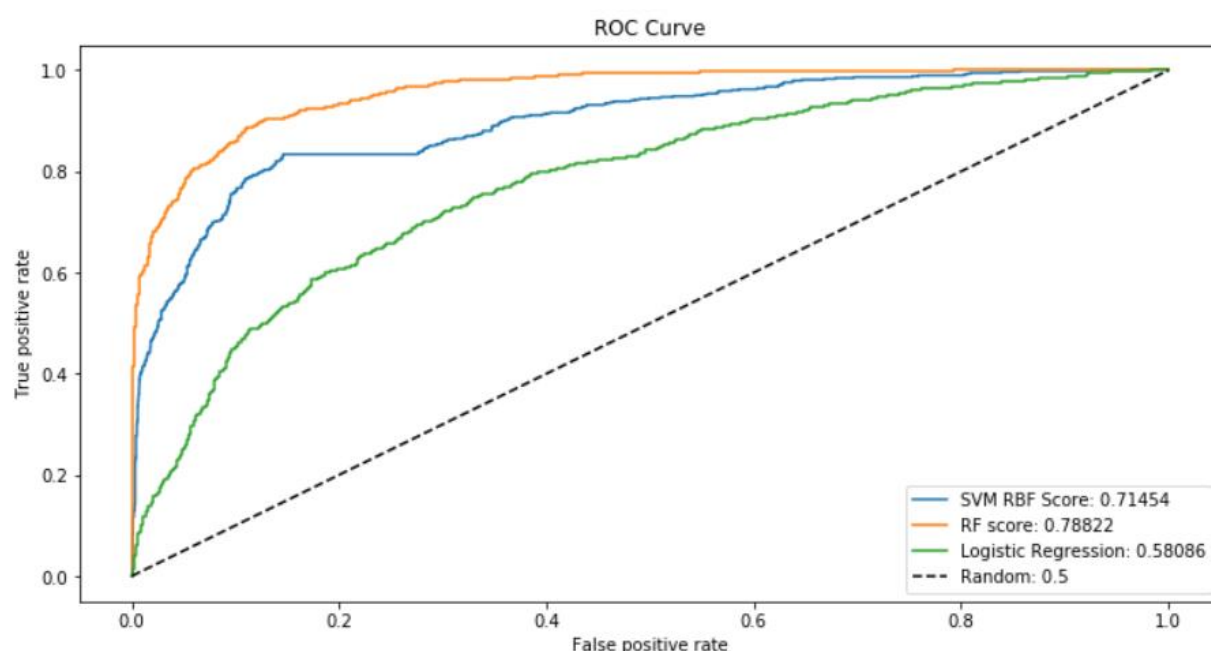
**True Positive Rate (TPR)** is a synonym for recall and is therefore defined as follows:

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

**False Positive Rate (FPR)** is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

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 following figure shows a typical ROC curve.



**Fig 8 above:** ROC/AUC compares the effectiveness of the Logistic Regression, the SVM and RF models.

## Final Thoughts and Summaries:

From the review of the fitted models above, the best model that gives a decent balance of the recall and precision is the random forest model where, according to the fit on the testing set, with a precision score on 1's of 0.91, out of all customers that the model thinks will churn, 91% do actually churn and with the recall score of 0.72 on the 1's, the model is able to highlight 72% of all those who churned.

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
- There is a small pattern that indicates that those bank members with an average tenure are less likely to churn than those on the lower or higher 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
- More credit card holders churn though most of the bank customers possess credit cards. The bank can benefit from increasing incentives in keeping credit card holders.

Furthermore, I noticed that the 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 a period of time 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?
- Does the bank have an investment division?
- Did the customer actually 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.