Title: Predicting Churn for Bank Customers

Bhargava Gaggainpali  
DSC680 - Applied Data Science  
Project 3 - Milestone 2

[bgaggainpali/bgaggainpali\_DSC680 (github.com)](https://github.com/bgaggainpali/bgaggainpali_DSC680)

**Introduction**

In financial industry, banks are playing important role in challenging times like now, with COVID pandemic across the globe. People are losing jobs and financial institutions are facing more Customer churn in Bank Accounts. The increase in Customer Churn rate will result in significant financial loss to commercial banks. It is very critical for lending institutions like banks to have a prediction model to be able to predict customers churn to better serve the customers and reduce the churn.

**Any surprises from your domain from these data?**

Bank Customer Churn is part of Financial industry. I have selected the topic, as I was interested in knowing the factors which influence Bank Customer Churn and the key factors in it. As I explore more about the domain, I understand that its not same set of rules which is being followed across domain. And each category like Bank Customer Churn, Telecom Customer Churn and such classification have different factors which are influencing people choose a different company. Bank should not lose business by having more churn customers and should not hurt the good customer, who can maintain accounts and business. Also, it should be able to identify the customers who are at risk of falling Churn and provide them other options like reducing the minimum balance amount and other options. We can say that Banks are playing important role in challenging times like now, with COVID pandemic across the globe.

**The dataset is what you thought it was?**

Initially when I looked at the dataset, I had questions about the variables as they are more of general kind and was guessing that if would serve my purpose of analysis. when closely observed the stats, it surprised me as the amount of value we can retrieve from such data. I am satisfied with the dataset which I have selected.

I have identified Churn\_Modelling.csv as source for my work, below is the Kaggle link. There are 10,000 observations in the dataset, each row in the dataset represents a Bank Customer Account. Given is the list of variables in the dataset.

Source File: <https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers?select=Churn_Modelling.csv>

**Variable Description**

RowNumber Sequence Number

CustomerId Customer Account Number

Surname Customer Name

CreditScore Credit Score

Geography Location Country

Gender Male / Female

Age Customer Age

Tenure Period of time in Years as Customer to the Bank

Balance Balance amount in the Bank

NumOfProducts Number of Products availed by the Customer

HasCrCard Customer has Credit card

IsActiveMember Customer Active Member

EstimatedSalary Customer Salary

Exited Customer Churn value

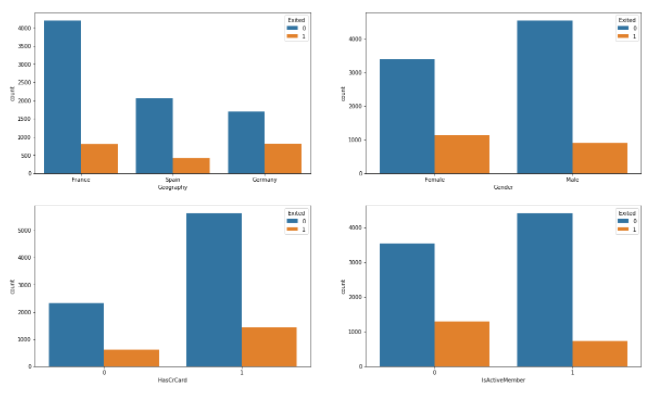
Below are the initial observations made from stats and plots:

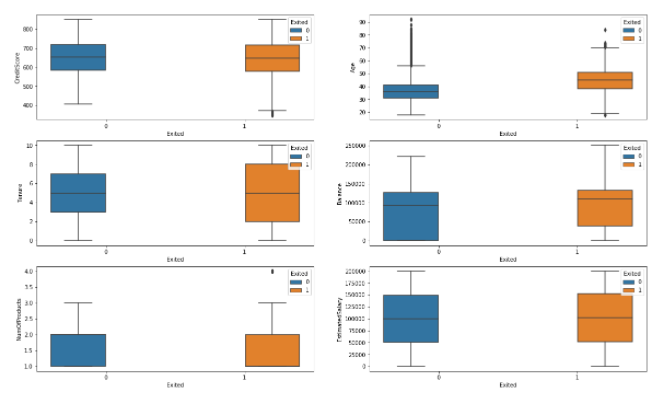
1. Majority of the data is from persons from France. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank.
2. The proportion of female customers churning is also greater than that of male customers
3. Majority of the customers that churned are those with credit cards.
4. Inactive members have a greater churn.
5. There is no significant difference in the credit score distribution between retained and churned customers.
6. The older customers are churning at more than the younger ones alluding to a difference in service preference in the age categories.
7. With regard to the tenure, the clients on either extreme end (spent little time with the bank or a lot of time with the bank) are more likely to churn compared to those that are of average tenure.
8. Bank is losing customers with significant bank balances which is likely to hit their available capital for lending.
9. Neither the product nor the salary has a significant effect on the likelihood to churn.

**Have you had to adjust your approach or research questions?**

After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression to consider variables like Balance, Age, Gender, Tenure, CreditScore, HasCrCard and EstimatedSalary. I saw surprising stats when I used visualizations to give clear idea on how each factor has its effect on the Customer Churn.

Yes, I had to increase my research questions to explore and include more variables, than initially prepared. Its based on the initial analysis using visualization.





**Is your method working?**

Till now, I am comfortable and confident that my methods would work as initially planned. The steps which I am following are giving me good results as expected.

After completing the data acquisition and initial analysis, by running the stats and using visualization, I had to increase in the number of variables to explore and analyze and also consider for model building. If the number of variables are significantly more and if need, I am planning to apply dimensionality reduction methods like PCA to reduce the number of variables for model building.

**What challenges are you having?**

So far, I have not faced any major challenges. The process and the steps I am following are giving me good results as expected. Small challenges I have faced is to convert all the data in to common format using Python functions and syntax, which I expect get better on practice. I am now focusing on building the model and might challenge on implementing it. Based on the accuracy of the results, will have to restructure the code if its not as per expected and also planning to build multiple models to choose a better one which suits the dataset values.