# Data Programming Project Report

## Team Braves

## Bank Churn Prediction

**Team members:**

**Keshav Balivada : kbalivada1@student.gsu.edu : Team Leader**

**Sharath Kumar Cherukuri : scherukuri1@student.gsu.edu**

**Thokala Eswar Chand : ethokala1@student.gsu.edu**

**Jasti Sai Vignesh : sjasti2@student.gsu.edu**

## Abstract

The modern way of increasing a business includes adding new customers and selling their products and services to an existing customer. In addition to that, the banks are now focusing more on how to keep their existing customers for longer periods. Our Bank Churn Prediction project solution will help the banks to predict who are the potential customers who might leave their organization in the near future. Banks can use this system to predict the tenures and throw lucrative offers to the customers who might leave soon. In this project we developed model to accurately predict the customers tenure. The developed model can be added to the banks existing environment. By entering new or existing customers profile information as input bank executives can know the predicted tenure of the customers.

## Introduction

With the current slow market, the conventional growth strategies like adding new customers and selling more to the existing customers' needs to be modified by focusing on keeping the existing customers for longer tenures. On average, the retail banks have 20-25% churning rate. It is crucial to for banks to reduce this churning rate and avoid revenue leakage by analyzing their customer behavior effectively and retain them by imposing lucrative offers. With our analysis, the banks can analyze new customer profiles and predict the tenure of their association with the bank.

## Motivation and Problem

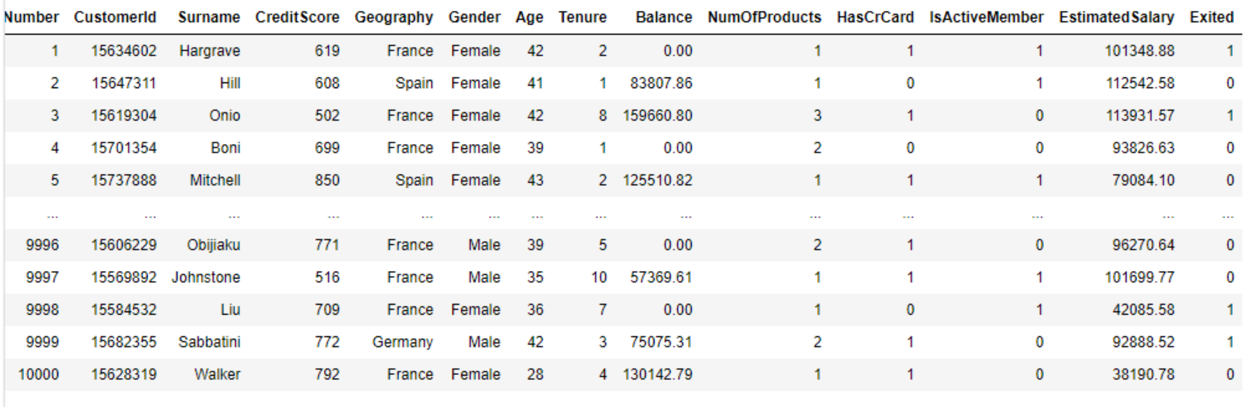
Maintaining the current customers in any bank is very important to sustain in this highly competitive world. Banks come with new innovative offers, features, services to the customers frequently. What is the use of these services or offers if there are many inactive customers. Targeting these type of inactive customers and making them active by attracting them with certain offers. According to the latest survey, on an average 20% to 25% of the customers are leaving the banks every year. This is a huge loss to the banks. In order to save banks from this problem, analyzing the customer behavior is very important. Banks must be able to know when the certain type of customers may leave the bank. If banks can able to know when customers leave the bank, then we can attract them by imposing some good offers. This not only stops them from leaving the bank but also makes them active.

## Data

The data which consists of customer details like age, gender, tenure, balance, no. of products, geography, credit score, has credit card or not, whether the customer is active or not, salary off certain bank. There are 14 features in the dataset which includes Row number, Customer Id, surname, Credit Score, Geography, Gender, Age, Tenure, Balance, No of products, Has credit card, Is active member, Estimated Salary, Exited.

Customer ID here is the unique id which is given to the customer. Surname is the customer name. Credit Score feature represents the customer present credit score. Geography represents the customer geographical location. Gender indicates the gender of the customer (Male or Female). Age indicates the age of customer. Tenure represents the duration of the customer’s account in the bank. No of Products represents the number of products purchased by the customer provided by the bank. Has Credit Card feature indicates that whether customer has taken credit card or not. Here, 1 represents the customer has taken credit card and 0 represents customer doesn’t have credit card. The feature Is Active Member indicates that the whether the customer is active or not. Here, 1 represents the customer is active, and 0 represents the customer is not active. Feature Exited indicates that the customer has left the bank or not.

In this project, the feature ‘Exited’ is predicted using all the remaining features. There are also some features here which are not useful in predicting like Row number, customer id, and surname. These features can be excluded for prediction.

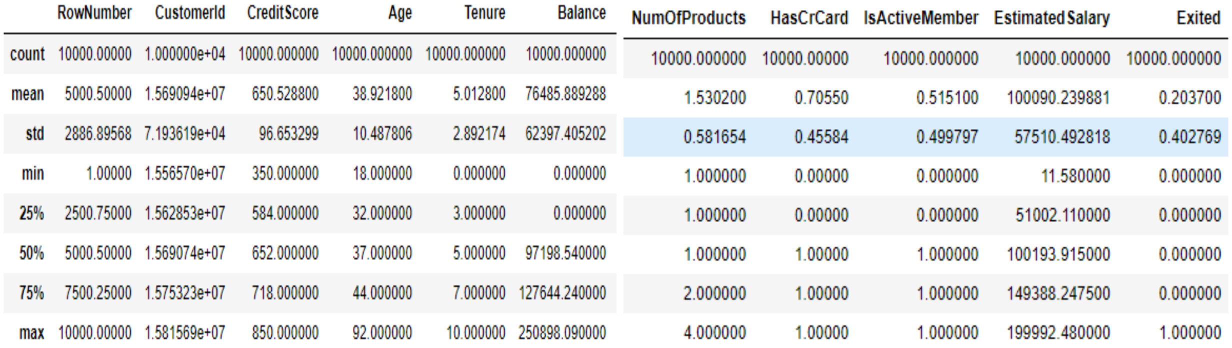


| Number | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15634602 | Hargrave | 619 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | 101348.88 | 1 |
| 2 | 15647311 | Hill | 608 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 |
| 3 | 15619304 | Onio | 502 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | 113931.57 | 1 |
| 4 | 15701354 | Boni | 699 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | 93826.63 | 0 |
| 5 | 15737888 | Mitchell | 850 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.10 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9996 | 15606229 | Obijiaku | 771 | France | Male | 39 | 5 | 0.00 | 2 | 1 | 0 | 96270.64 | 0 |
| 9997 | 15569892 | Johnstone | 516 | France | Male | 35 | 10 | 57369.61 | 1 | 1 | 1 | 101699.77 | 0 |
| 9998 | 15584532 | Liu | 709 | France | Female | 36 | 7 | 0.00 | 1 | 0 | 1 | 42085.58 | 1 |
| 9999 | 15682355 | Sabbatini | 772 | Germany | Male | 42 | 3 | 75075.31 | 2 | 1 | 0 | 92888.52 | 1 |
| 10000 | 15628319 | Walker | 792 | France | Female | 28 | 4 | 130142.79 | 1 | 1 | 0 | 38190.78 | 0 |

Fig no: 1 Data of customers from a bank.

## Statistical Analysis

Statistical Analysis is made on the collected data to understand the stats of numerical features present in the dataset. The analysis like minimum, maximum, mean, count, standard deviation, 25, 50, 75 percentiles values is made.



|  | RowNumber | CustomerId | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 10000.00000 | 1.000000e+04 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.00000 | 10000.000000 | 10000.000000 | 10000.000000 |
| mean | 5000.50000 | 1.569094e+07 | 650.528800 | 38.921800 | 5.012800 | 76485.889288 | 1.530200 | 0.70550 | 0.515100 | 100090.239881 | 0.203700 |
| std | 2886.89568 | 7.193619e+04 | 96.653299 | 10.487806 | 2.892174 | 62397.405202 | 0.581654 | 0.45584 | 0.499797 | 57510.492818 | 0.402769 |
| min | 1.00000 | 1.556570e+07 | 350.000000 | 18.000000 | 0.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 11.580000 | 0.000000 |
| 25% | 2500.75000 | 1.562853e+07 | 584.000000 | 32.000000 | 3.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 51002.110000 | 0.000000 |
| 50% | 5000.50000 | 1.569074e+07 | 652.000000 | 37.000000 | 5.000000 | 97198.540000 | 1.000000 | 1.00000 | 1.000000 | 100193.915000 | 0.000000 |
| 75% | 7500.25000 | 1.575323e+07 | 718.000000 | 44.000000 | 7.000000 | 127644.240000 | 2.000000 | 1.00000 | 1.000000 | 149388.247500 | 0.000000 |
| max | 10000.00000 | 1.581569e+07 | 850.000000 | 92.000000 | 10.000000 | 250898.090000 | 4.000000 | 1.00000 | 1.000000 | 199992.480000 | 1.000000 |

Fig no:2 Statistical Analysis on the data collected

From the fig no:2 we can see the values of minimum, maximum, standard deviation, mean, count, 25,50,75 percentile values for the numerical features we have. After performing statistical analysis, some of the interpretations obtained from the analysis are:

- Minimum Credit score is 350 where as Maximum Credit score is 850.
- Minimum Age is 18, and Maximum Age is 92. From this we can understand that, customer will be eligible to take a bank account only when he is 18 or above in age.
- Tenure feature is varying in the range 0 to 10.
- Maximum Balance noted in the dataset is $250,898 where as Minimum Balance noted is 0. From this we can understand that there are few customers who also maintaining 0 balance in their accounts.

- Number of products purchased minimum is having 1 and maximum 0. At least one of the products is to be purchased by the customers in order to become a member or to get an account at that bank.
- Maximum Estimated Salary noted here is $199,992 and minimum being $11.58. This may indicate that, every month/year this is the amount credited into their accounts.
- No. of Products, Has Credit Card, Is Active Member, Exited are categorical features. As the values present in them are numerical, the analysis is made on them by default.

## Exploratory Data Analysis

To understand the distribution and relationship among different features, Exploratory Data Analysis is made. There are some good insights which were derived from the visualizations.

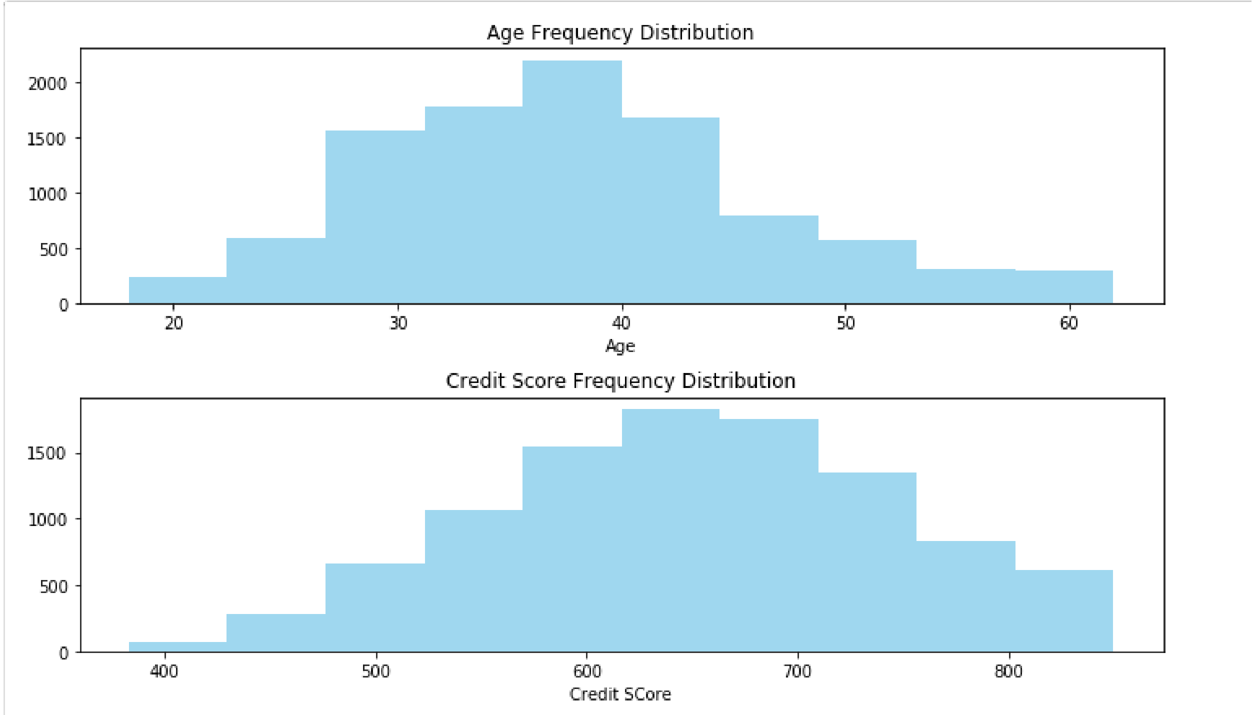


Fig no:3 Age and Credit Score Distribution

From the above figure both age and credit scores are normally distributed.

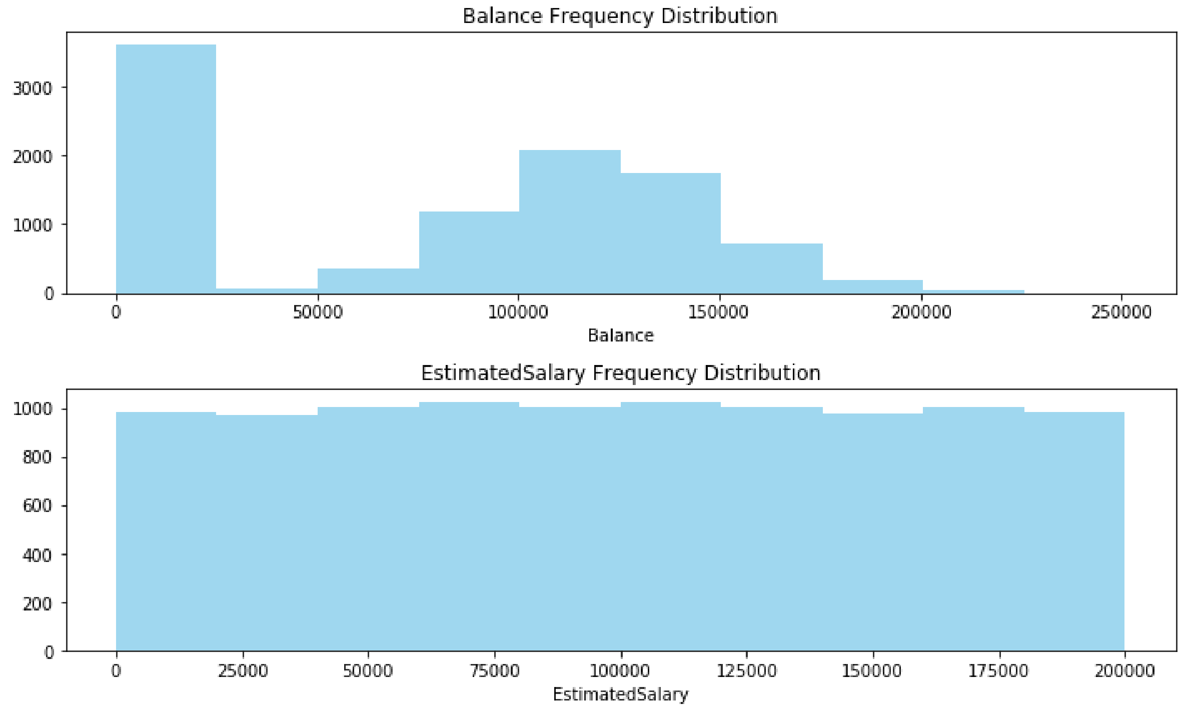


Fig no:4  Balance and Estimated Salary Distribution

From the fig no 4, Balance and Estimated are not normally distributed. We can notice that there are more number of customers having balance in the range 0 to 25000. Estimated salary is uniformly distributed.

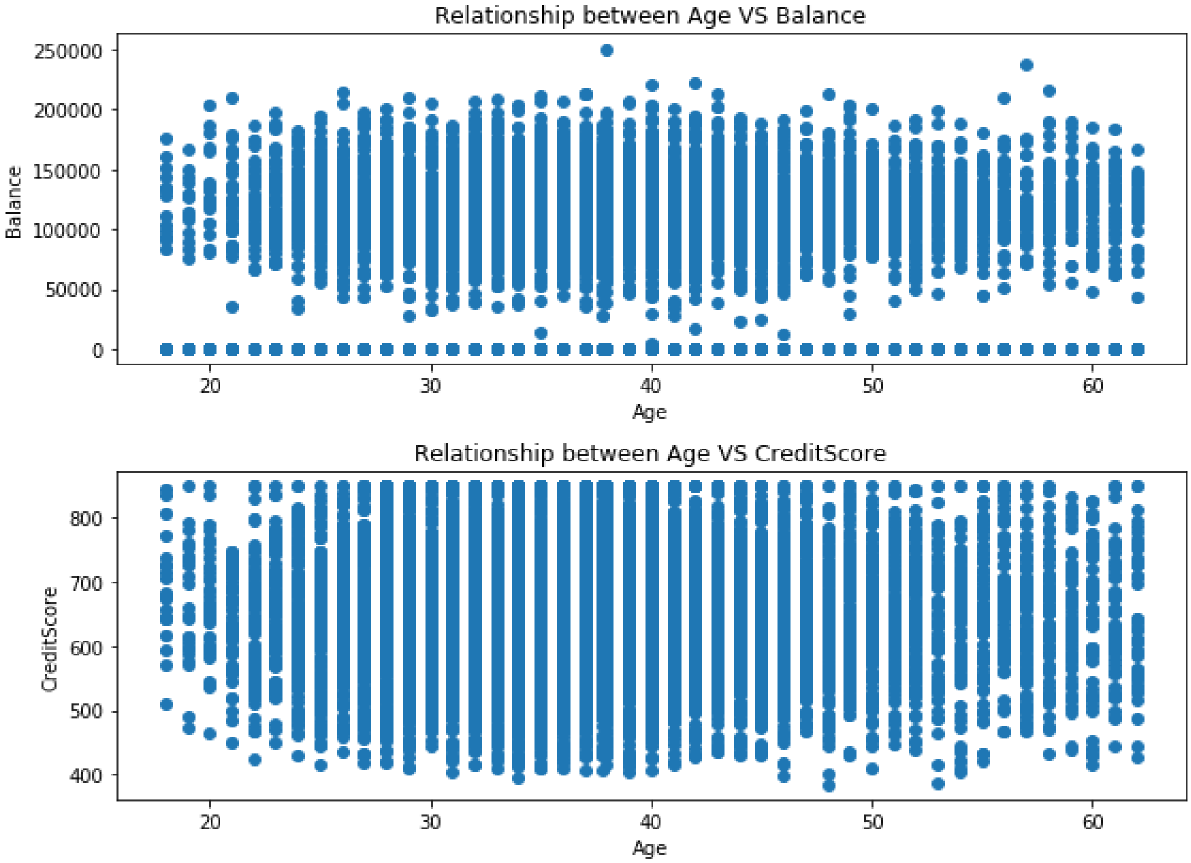


Fig no :5 Relationship between Age vs Balance, and Age vs Credit Score

From the above figure, both balance and credit score are randomly distributed with respect to age. A dense distribution is found between age 30 to 40 for the both balance and credit score.

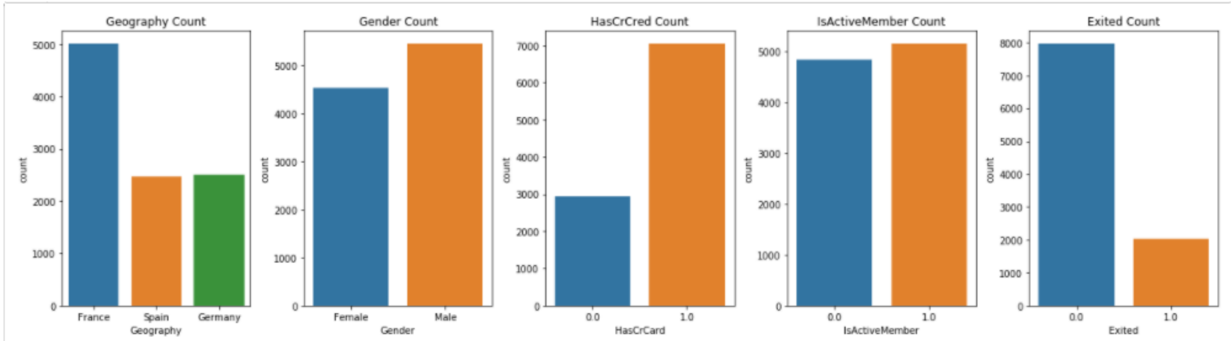


Fig no:6 Categorical Fetaures distribution

From the figure no , most of the customers data we have is from France. Customers from Spain and Germany is equal. There are more males compared to females in our data. Most of the customers who have accounts have taken credit cards. Count of active and non-active customers is equal. There is also less data for the customers who exited the bank.

## Handling Missing Values and Outliers

After Performing Statistical Analysis, missing values were checked is there are any in the dataset. There are no missing values or null values in the dataset. Followed by, Outliers were detected for every important feature by plotting box plots. Any value which exceeds the maximum and minimum value is said to be an outlier and those are removed. This is calculated using Inter Quartile Range, Lower Quartile(Q1), and Upper Quartile (Q2). After removing the outliers, missing values will be obtained. Now, the missing values are substituted with the mean value of certain feature. This process is repeated for every numerical feature in the dataset and handled both missing values and outliers.

## Performing Normalization

Shapiro Test is performed on the numerical features to know whether the values are normally distributed or not. Based on the p value generated, the data distribution can be known. If the p value is greater than 0.5, then the data is normally distributed. In this project, normality is checked for every feature. If there is any feature without normal distribution, normalization is implemented, to make the data normally distributed.

## Performing Standardization

Every feature in the dataset, has different range. When this data is directly sent to model for training, it understood that highest value has highest weight and lowest value has lowest weight. When this comparison is made among many numerical features, our model cannot learn every minute detail. For example, if there is US$10, and Canadian $12 our model thinks that Canadian $12 is more than US $10 as the value is higher. In reality US $10 is greater than Canadian $12. Similarly, to deal with this type of problems, performing standardization is very important. Here we are using Min Max Scaler to bring all the numerical features into similar range. In Min Max Scaler, every value is subtracted by the minimum value in that feature and divides by the range.

Here standardization is implemented to only four features, as they were only the numerical features we had. Credit Score, Age, balance, estimated Salary features were scaled and changed the values to same units.

## One Hot Encoding

Till now, all the numerical features were made ready by performing normalization and standardization. There are certain categorical features like geography, number of products, tenure, gender which are to be prepared before model building. In order to prepare categorical features, one hot encoding is performed. Each and every unique value is treated as new feature and the values will be 0 or 1. 1 represents the feature is recorded and 0 being not recorded. In this way every categorical feature is changed to binary format with the help of one hot encoding. After performing one hot encoding we got 30 different features.

## Train and Test Data

The data is prepared now and next step is to build a model. The data which is pre processed is to be divided into 2 sets, train data and test data. In this project we have divided 33% of the data as test data and remaining as training data. Train data is used to train the model. Model understands the underlying patterns among features using this train data. Further, test data is used to evaluate the model we developed. While splitting the data, unimportant features like row number, customer id, Surname are dropped, as this doesn't add any value to our predictions. Number of rows got in the train data is 6689 and the number of rows in test data is 3296.

## Model Building and its Evaluation

After splitting the pre processed data into train data and test data, machine learning models is developed. In this project, we are predicting whether the customer leaves the bank with the given details. This project, started with Logistic Regression to classify the customers who will leave and not leave. Logistic Regression gave Accuracy up to 84%, Precision of 72%, and Recall 40%. Naïve Bayes model is developed further and resulted in 81% accuracy, 84% precision, and 12% recall. Decision Tree model is developed further and resulted in 77% accuracy, 47% precision, and 46% recall. As we have less data, ensemble models like random forest, xgboost works well. So as a try this project further extended hands in implementing ensemble models to see if precision and recall were also high. Random Forest model is developed further and resulted in 84% accuracy, 74% precision, and 43% recall. Even though precision increased, recall is very less. Similarly xgboost model is implemented and obtained 84% accuracy, 71% precision, 47% recall.

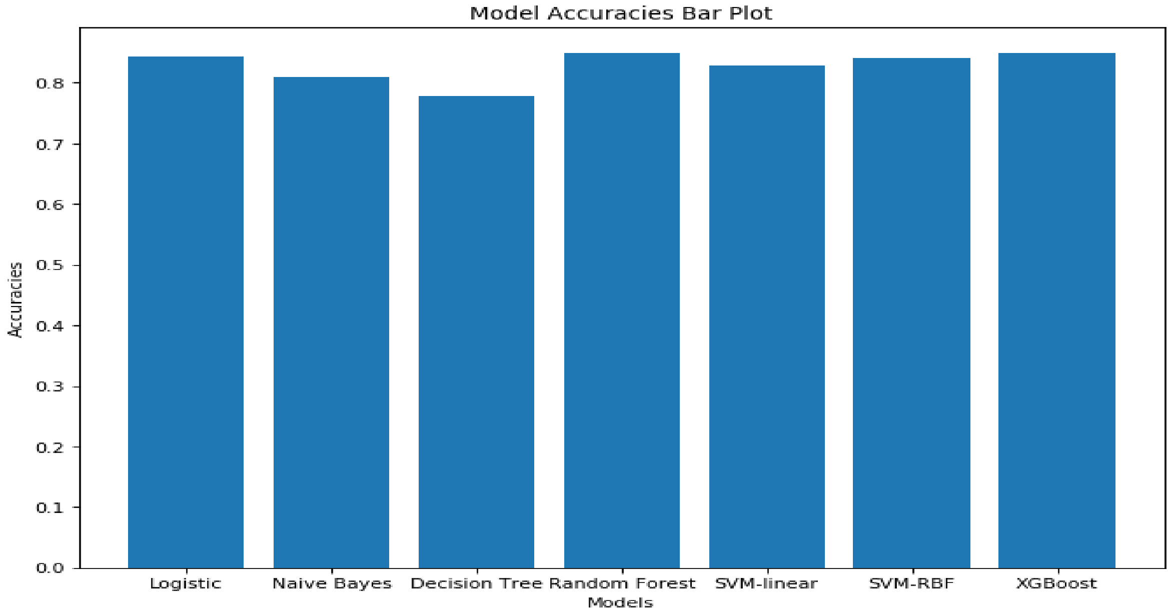


Fig no:7 Model accuracies comparison

From the Fig no:7 Random Forest has the highest accuracy recorded. Though the accuracies of all the models are uniform, precision and recall are very less. While evaluating the model, only accuracy is not important to consider, precision, recall, f1 score are other metrics which are to be considered in evaluating the model.

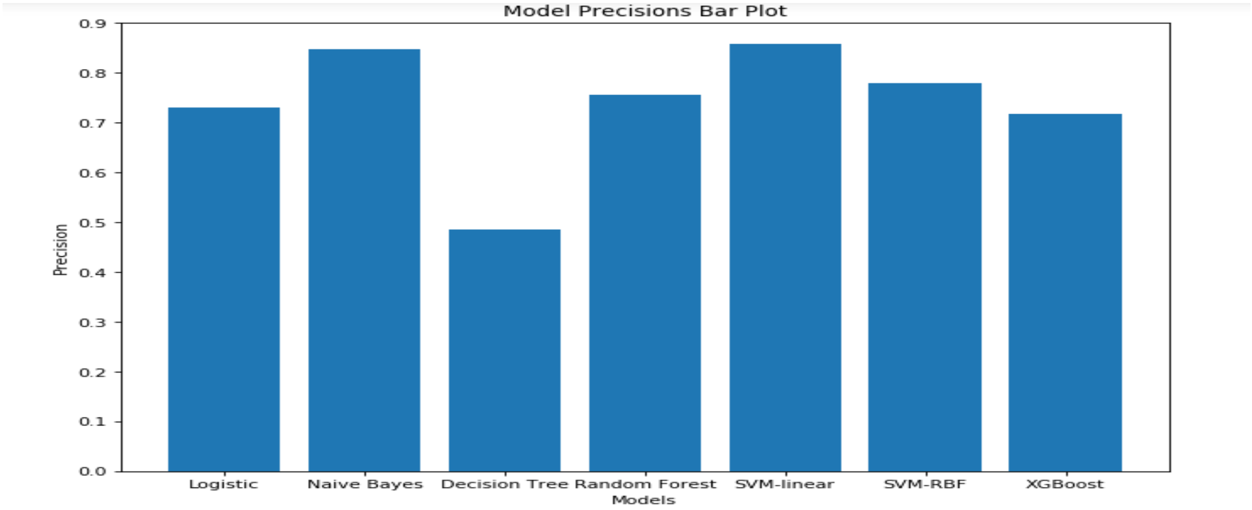


Fig no:8  Model Precisions bar graph

From the fig no:8 , though the accuracies were high we can notice that precisions were not that good. Models Naïve Bayes and SVM linear kernel has the highest precision among all other developed models.

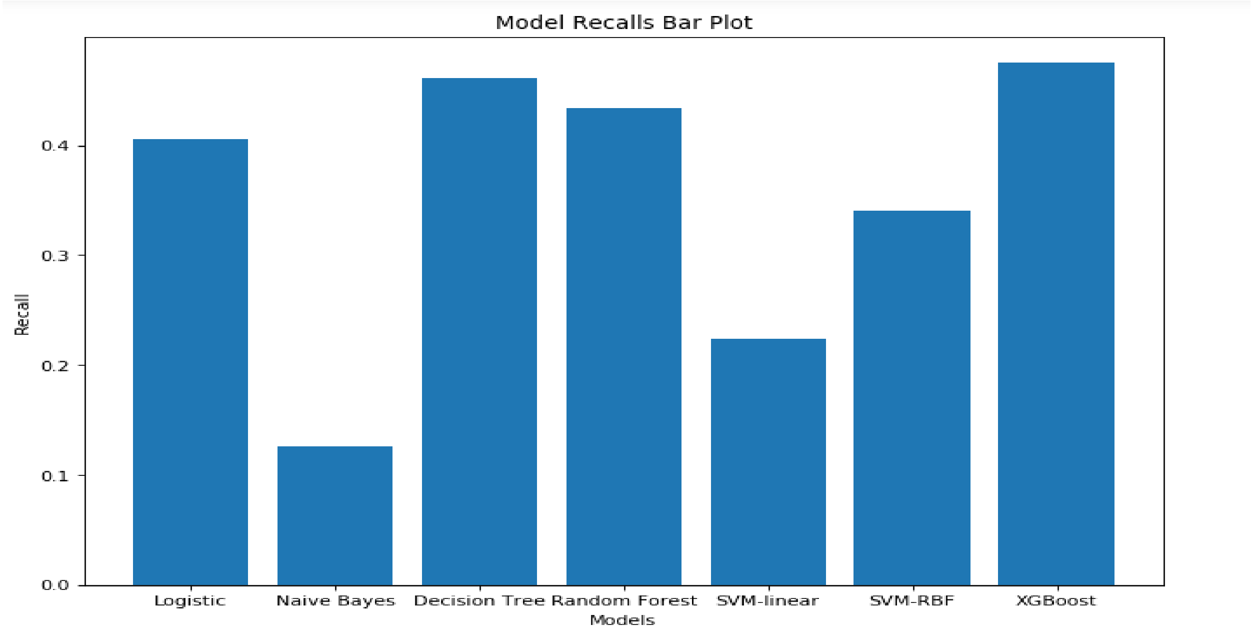


Fig no:9  Model Recall bar graph

From the fig no:9 , though the accuracies were high we can notice that recalla were not that good. Models Naïve Bayes and SVM linear kernel has the lowest recall among all other developed models. Models XGboost and Decision tree has the highest recall.

## Handling Imbalance in the dataset

The reason behind getting very low precision and recall is because of imbalance in the dataset. The ratio of the values of Target feature that is Exited are not in same ratio. There are very less number of customers who left the bank in the data. There is more data on the customers who are current members. Due to this imbalance created, the model developed is unable to identify the patterns for the customers who left the bank. This is the reason why very low precision and recall is recorded for every model. This leads to under fitting of the model.

To handle this type of situation, up sampling is performed to the values which have 1 in the recorded feature. As the values were very less compared to the values of 0, up sampling increases the count and sees that both the values in target feature have same ratio. After implementing Up Sampling both he values 0 and 1 has the same count. Building a machine learning model now, can boost the precision, recall and accuracy.

## Model Building and its Evaluation after Sampling

After sampling the data, random forest is implemented. When the data is less, ensemble models like random forest, xgboost gives better results. Random Forest model resulted in 93% accuracy, 90% precision, 95% recall. Further XGboost model is implemented and resulted in 89% accuracy, 86% precision, and 92% recall. After balancing the dataset accuracy, precision, and recall were all boosted.

## Results

Designed a model which can predict the churn rate. This model further can be deployed into bank portals. Any layman who have no knowledge in machine learning concepts can also know the customers who are about to leave. By entering just small inputs like customer age, tenure, salary, gender, etc. model can be able to generate a list of customers who are about to leave. Banks can concentrate on these customers in retaining by imposing some attractive offers or services.

## Conclusion

Our Bank Churn Prediction project solution will help the banks to predict who are the potential customers who might leave their organization in the near future. Banks can use this system to predict the tenures and throw lucrative offers to the customers who might leave soon. In this project we developed model to accurately predict the customers tenure. The developed model can be added to the banks existing environment. By entering new or existing customers profile information as input bank executives can know the predicted tenure of the customers.

## References

1. Dataset Reference : https://www.kaggle.com/santoshd3/bank-customers