School of Computer Science and Engineering

J Component report

Programme : B.Tech

Course Title : Foundations of Data Analytics

Course Code : CSE3505

Slot : F1/F2

Title: BANKING CUSTOMER CHURN PREDICTION

Team Members:

Rahul Sandireddy - 20BCE1001

Jaladi Deepika -20BPS1099

Thota.Varshika -20BPS1158

Vikkurty Vamsi -20BCE1066

Faculty: Dr. Vergin M

Sign:

Date:

ABSTRACT:

In this project we are trying to solve the problem of banking domain, by identifying whichcustomers are at risk of churning and what are the reasons for churning with the help of supervised learning classification algorithms.

In this project we are using a source of 10,000 bank records to predict the like hood of acustomer churn, we got this data set from Kaggle.

Keywords:

Churn Prediction, Machine Learning, SVM, Logistic Regression, Random Forest, Decision Tree, EDA

1.INTRODUCTION

Bank churning:

Taking advantage of bank bonuses and signing up for new bank accounts is known as bank churning. The name churning itself because once you get into it, you will want to do it over and over again, and with many banks you will be able to after an allotted time has passed. Bank Churning is completely legal, has quite a few advantages over investing, and few if any advantages.

Problem Statement:

Aiming to predict customer churn in a financial institution

Dataset Description:

The dataset was extracted from the Kaggle. This data contains 12 features about 10000 clients of the bank.

The features are the following:

- **customer_id**, unused variable.
- **credit_score**, used as input.
- **country**, used as input.
- **gender**, used as input.
- **age**, used as input.
- **tenure**, used as input.
- **balance**, used as input.

- **products_number**, used as input.
- **credit_card**, used as input.
- **active_member**, used as input.
- **estimated_salary**, used as input.
- **churn**, used as the target. 1 if the client has left the bank during some period or 0 if he/she has not.

This project uses customer data from a bank to build a predictive model for the likely churn clients. As we know, it is much more expensive to sign in a new client than to keep an existing one. It is advantageous for banks to know what leads clients to leave the company. Churn prevention allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible.
In this project we are using Logistic Regression, Svm, Decision tree and Random Forest classification algorithms

2. LITERATURE SURVEY:

Paper 1:

Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content

Dataset Description: In Naïve Bayes, the line type of users has the highest weight while in Logistic Regression, gender has the highest weight, and in Deep Learning, the customer's age has the highest weight.

Implementation details: For solving this problem we put two main approaches: For the first approach we build a dataset through practical questionnaires and analyzing them by using machine learning algorithms. The second approach is customer churn prediction model through analysing iitheir opinions through their user-generated content.

Results: Here,we build a model to analyze the behavior of customers and predict whom customers want to churn. We used Deep Learning, Naïve Bayes, and Logistic Regression algorithms.We analyzed the UGC by using sentiment analysis to analyze and classify customers' opinions.

Paper 2:

Customers Churn Prediction in Financial Institution Using Artificial Neural Network.

Dataset Description: The dataset was extracted from the database of one of the leading financial institutions in Nigeria.

Results: The data was extracted from the bank database and divided into 3 sets: training set, test set, and validation set. 80% of the dataset was used for training,10% was used for testing and 10% was used for validating the model.

Paper 3:

Prediction model for customer churn from electronic banking services

Data Set Description: Larose described this phase as the one in which data selection and data cleaning tasks are undertaken

Results: We implemented the CRISP methodology for predicting customer churn in electronic banking services. The aim of the present study is to identify the features of churners from electronic banking services.

Paper 4:

Predicting Retail Banking Churn in the Youth Segment of Customers

Dataset Description: Dataset describes various feature like ease of banking with an ATM, allied banking service and etc. These features are relevant and significant in customers' association with bank.

Implementation details: The selected models' performances were compared. The performance matrix included accuracy, F1 score, sensitivity, specificity, AUC and precision.

Results: Based on the results the mobile banking and ease of the banking with an ATM were the deciding factors for a customer to continue.

Paper 5:

Customer Churn Analysis and Prediction in Banking Industry

Dataset Description: The dataset consisted of 57 attributes such as demography, transactions, balance etc.

Implementation details: They have used five classification methods that are Decision tree, neural network, SVM, Naïve Bayes and Logistic Regression.

Results: The significant attributes are vintage, volume of EDC, balance in one month of age. This research got SVM as modelling with best accuracy

Paper 6:

Churning of Bank Customers Using Supervised Learning

Dataset Description: Dataset used for this supervised prediction is acquired from an online source. These features include row number, customer id, surname, credit score, geography, etc. The customers of the bank are identified as churn based on the potential features like age, gender, estimated salary, etc.

Implementation details: In this paper, whole focus is using flexible technique to boost the accuracy in customer churning process. So, along with K-nearest neighbours (KNN) algorithm, XGBoost algorithm is implemented.

Results: This prediction gives useful insights to the bank officials regarding its customers and functioning of bank. The performance of the prediction model is the capability to identify customers exit status accurately. XGBoost gave the best result in terms of accuracy, sensitivity and specificity. Boosting has given the increased accuracy of 86.85 with low error, high sensitivity and specificity.

3. Proposed Work

3.1 Data Source:

By taking a glimpse on our dataset, we took total of 10,000 and 14 columns. Three non-useful variables are identified in the dataset: **RowNumber**, **CustomerID**, and **Surname**. Two categorical variables: **Geography** and **Gender** need to be encoded into numbers. Because machine learning models can only work with numerical input.

3.1.1 Data Processing

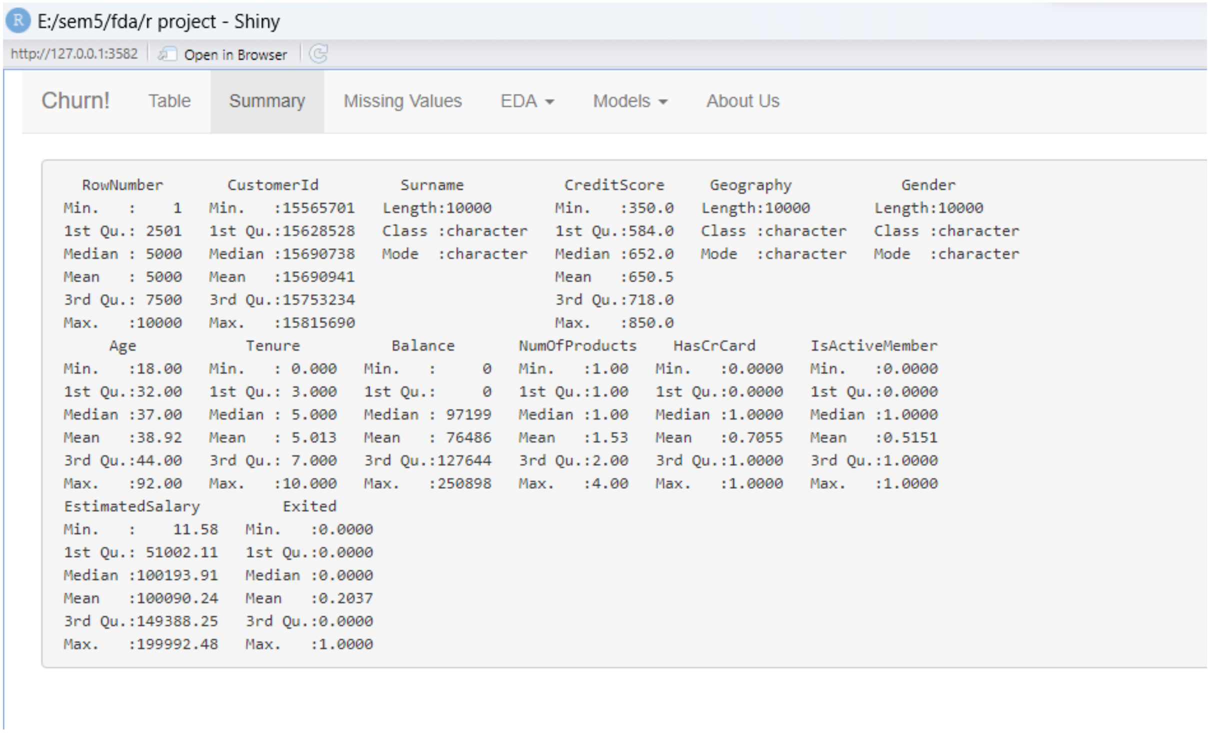
The data processing that need to be done include:
1) Drop RowNumber, CustomerID, and Surname.
2) Encode Geography and Gender.
3) Log Transform Age, CreditScore, and Balance.
4) Scale range of Age, CreditScore, Balance, EstimatedSalary from 0 to 1.

3.1.2 libraries used

library(shiny)
library(shinythemes)
library(ISLR)
library(DataExplorer)
library(ggplot2)
require(dplyr)
library(ggcorrplot)
library(tidyr)
library(purrr)
library(printr)
library(pROC)
library (ROCR)
library(caret)
library(car)
library(rpart)
library(rpart.plot)
library(e1071)
library(markdown)
library(randomForest)

3.1.3 Summary of the data:

This is a summary of the variables' statistics. When we look at the Min and Max of the continuous variables, we can see that their scales differ greatly, for example, Age and EstimatedSalary. Because the variables with larger scales would overshadow the variables with smaller scales, scaling is required to scale these variables to the same 0 - 1 range.

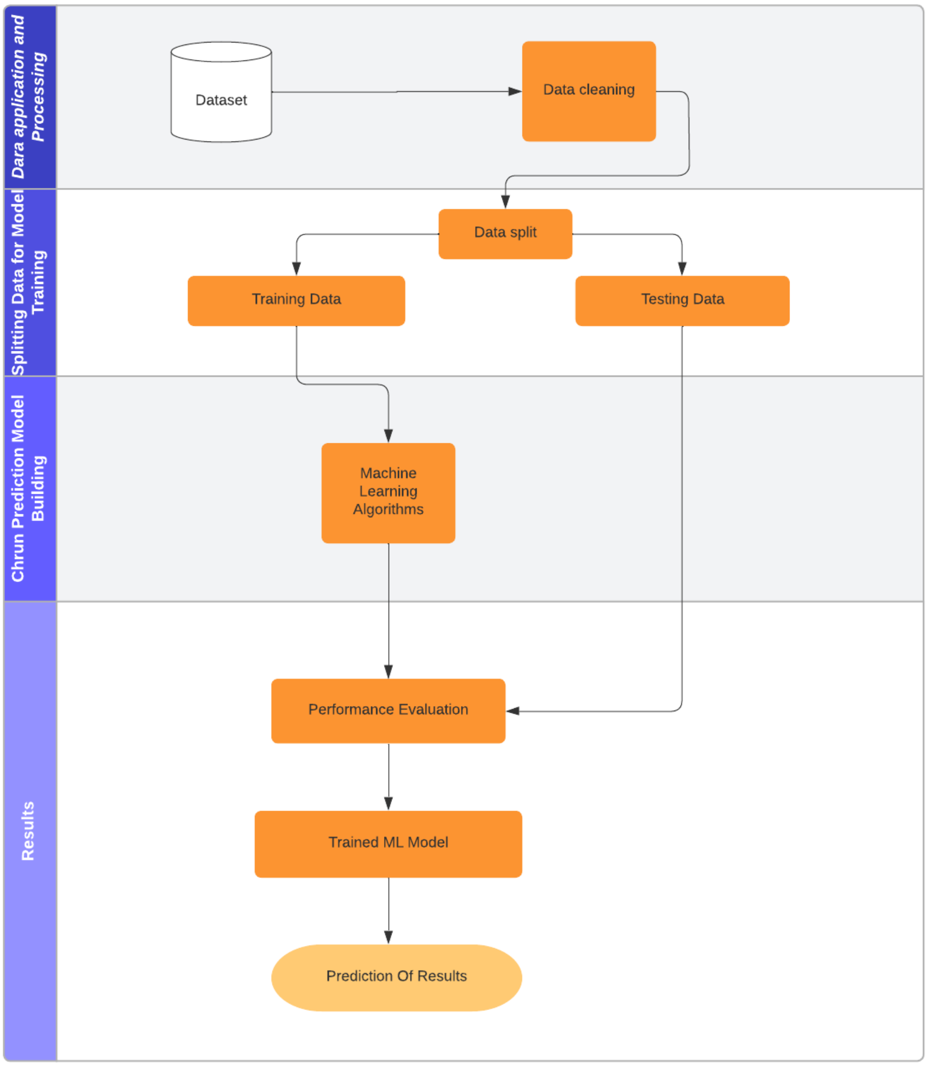


3.2. Data Analytics Models

To train a classification model, there is mainly three steps:

1. Splitting Data into Training and Testing Set
2. Model Training/ Tuning
3. Model Testing

The Exited variable will be used as the target variable to predict whether a bank customer will churn or not.



3.2.1 Method Approach

The confusion matrix represents the predicted and actual values. The output "TN" stands for True Negative; "TP" stands for True Positive, "FN" stands for False Negative, and "FP" stands for False Positive. The parameters we are using are Precision, Recall and Overall Accuracy for the Machine Learning Algorithm.

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.2.2 RANDOM FOREST CLASSIFIER

Random forest classifier is a classification technique that uses algorithms consisting of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

3.2.3 DECISION TREE

A Decision tree is a tree-like structure with attributes assigned as a node. Based on the values of the attribute's algorithm will traverse through the tree finally ending with the leaves of the tree which contain classification output. In the Decision tree we are using the Gini index (It is a measure of the impurity of the values in the attributes and split the tree into many branches).

3.2.4 SVM (SUPPORT VECTOR MACHINE)

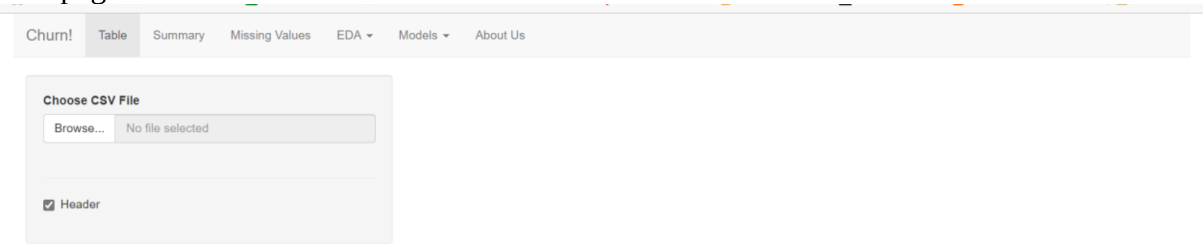
Support vector machines (SVM) are used to classify both direct and non-linear data. In short, when the algorithm receives the original training data, it uses non-linear mapping to transfigure it into an advanced dimension. In this dimension, a direct optimal hyperactive airplane is sought to separate the data of any two classes. SVM can also be used for bracket and numerical validation. The simplest form of SVM is a two-class problem where the classes are linearly divisible. For a 2- D problem, a straight line can be drawn to separate the classes, in fact, multiple lines can be drawn. In the SVM algorithm, the kernels used for the classification of websites are Default, linear, RBF, and polynomial.

3.2.5 LOGISTIC REGRESSION

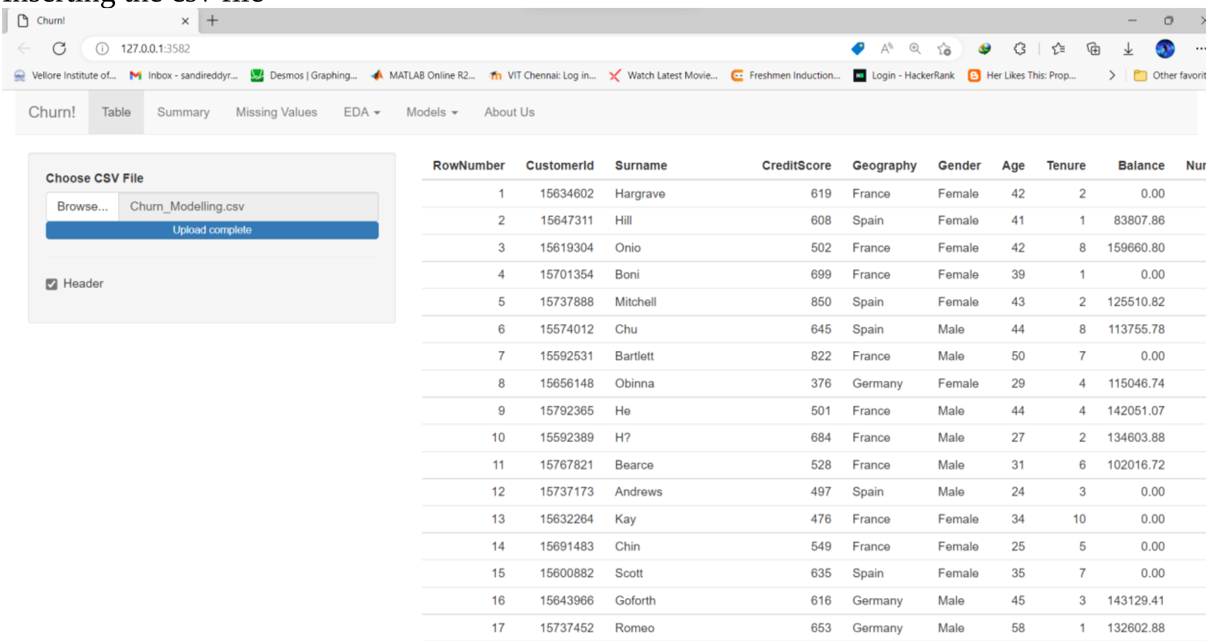
Logistic regression is a type of predictive analysis where churn prediction can be detected based on attributes. In logistic regression, the input is given as training data and test data. Based on the given input, logistic regression is calculated using a regression function called a sigmoid function, with the calculated sigmoid function, the relationship between the training data and the test data is calculated.

4. Results and Discussions

Webpage

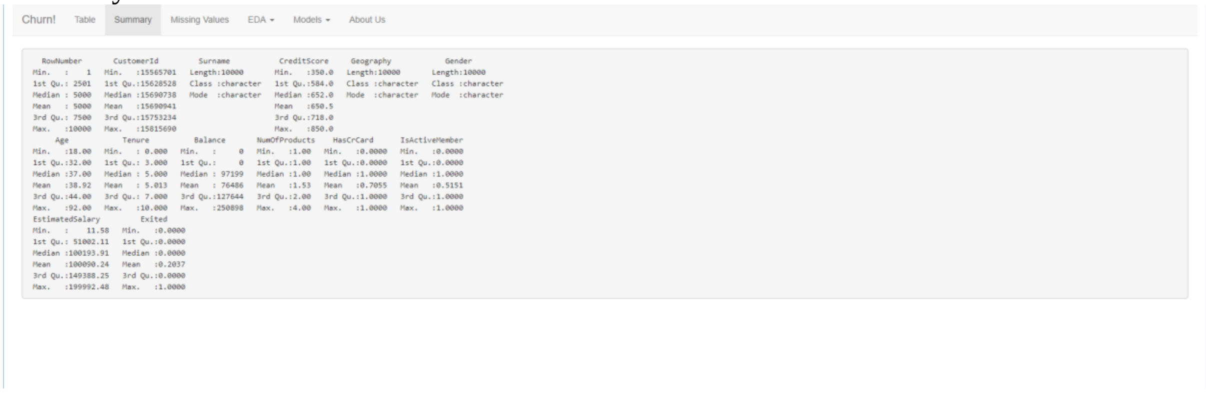


Inserting the csv file

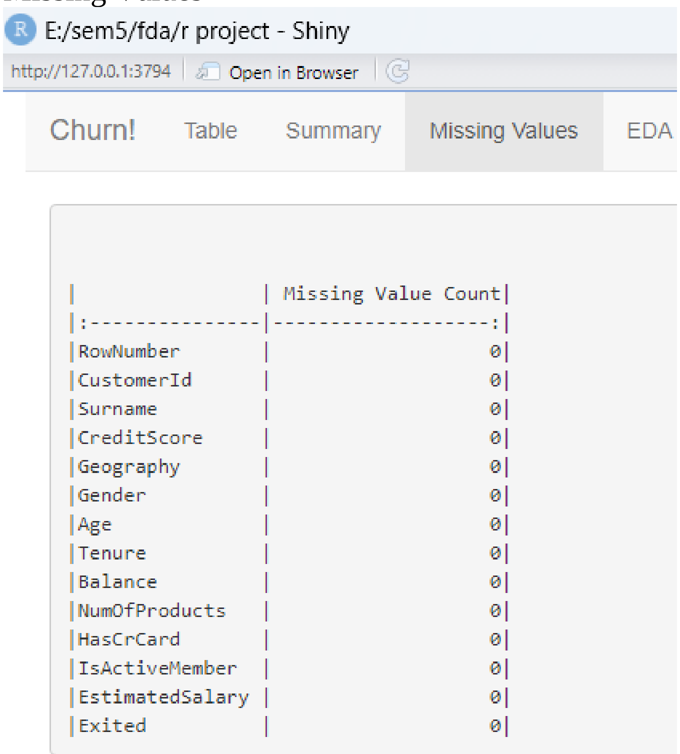


RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	Nur
1	15634602	Hargrave	619	France	Female	42	2	0.00	
2	15647311	Hill	608	Spain	Female	41	1	83807.86	
3	15619304	Onio	502	France	Female	42	8	159660.80	
4	15701354	Boni	699	France	Female	39	1	0.00	
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	
6	15574012	Chu	645	Spain	Male	44	8	113755.78	
7	15592531	Bartlett	822	France	Male	50	7	0.00	
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	
9	15792365	He	501	France	Male	44	4	142051.07	
10	15592389	H?	684	France	Male	27	2	134603.88	
11	15767821	Bearce	528	France	Male	31	6	102016.72	
12	15737173	Andrews	497	Spain	Male	24	3	0.00	
13	15632264	Kay	476	France	Female	34	10	0.00	
14	15691483	Chin	549	France	Female	25	5	0.00	
15	15600882	Scott	635	Spain	Female	35	7	0.00	
16	15643966	Goforth	616	Germany	Male	45	3	143129.41	
17	15737452	Romeo	653	Germany	Male	58	1	132602.88	

Summary of the dataset

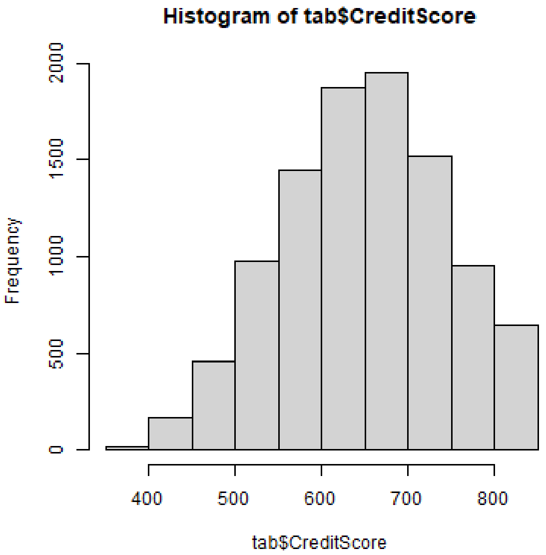


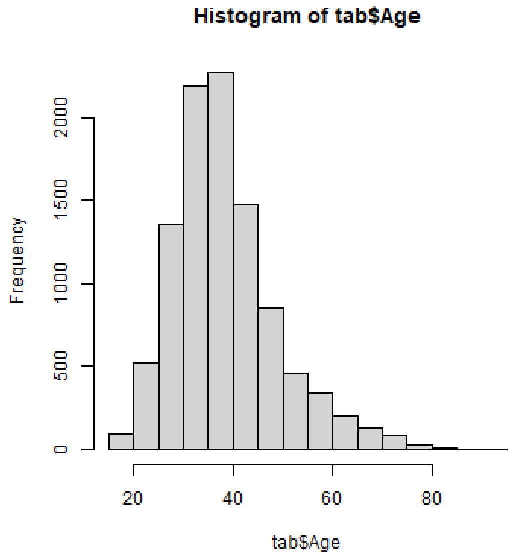
Missing Values

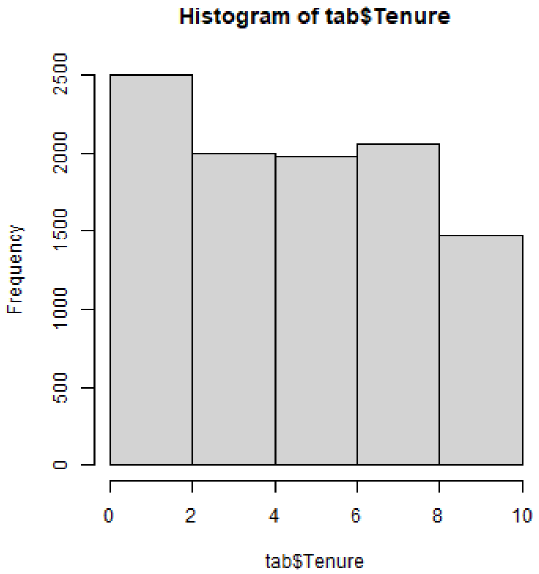


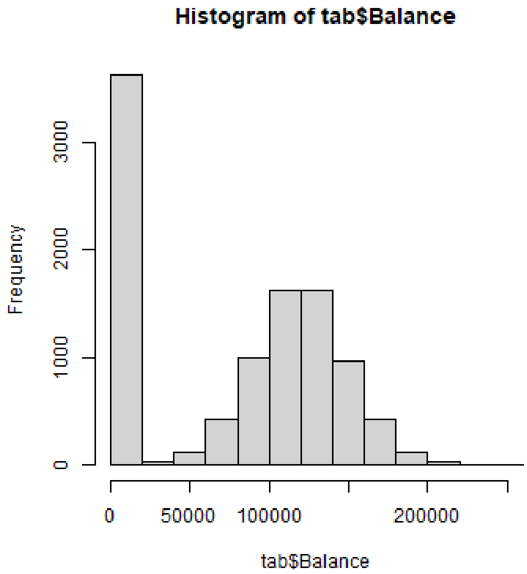
Exploratory data analysis:

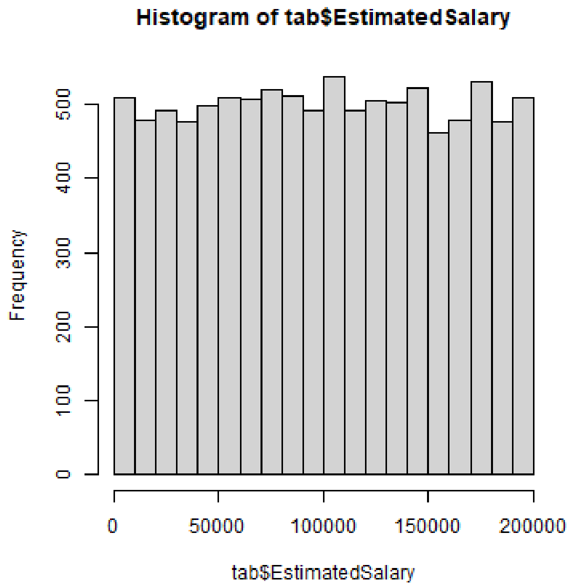
Plotting Histograms to understand the distributions



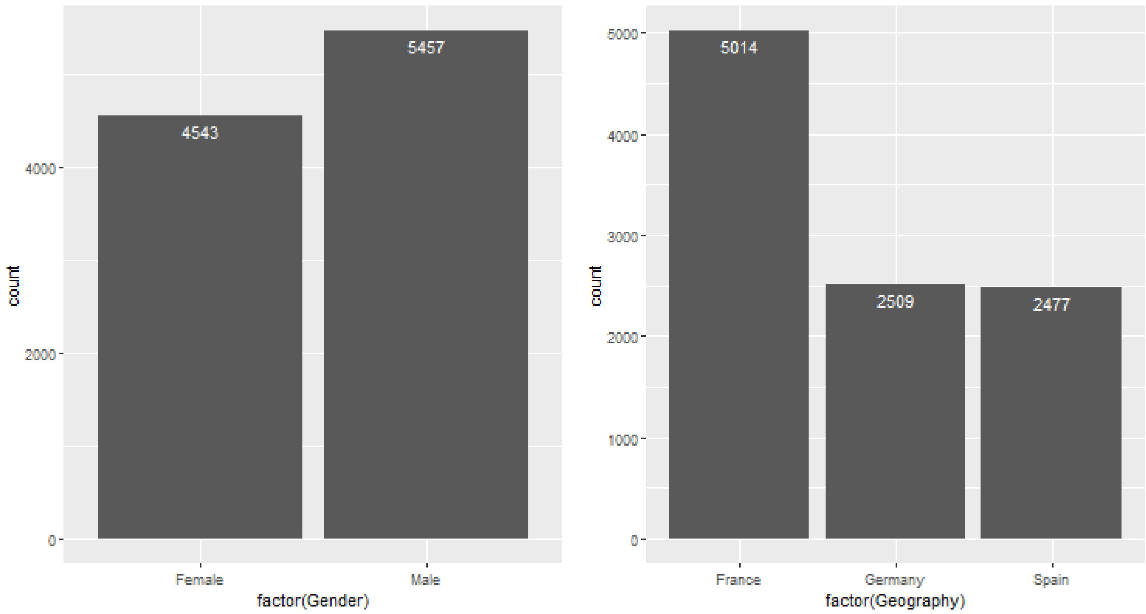


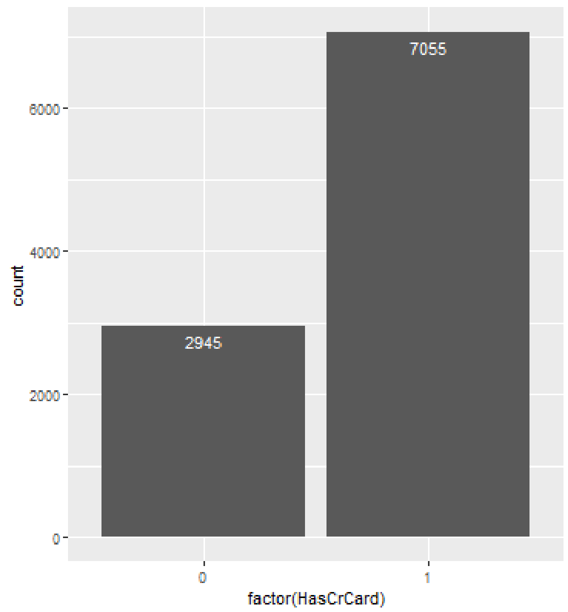


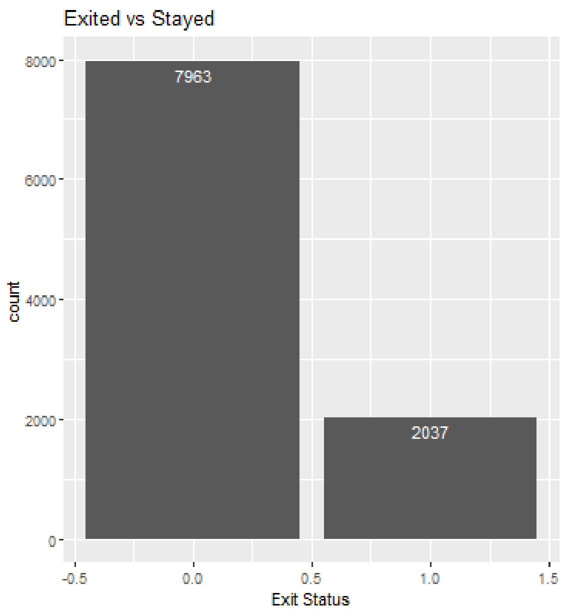




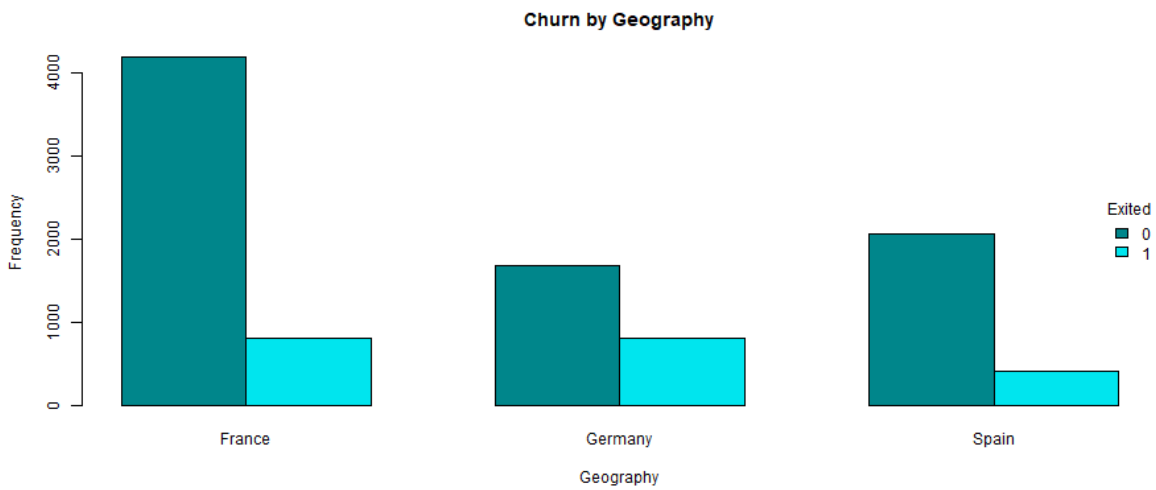
Plotting Bar Charts to understand the Categorical Variables

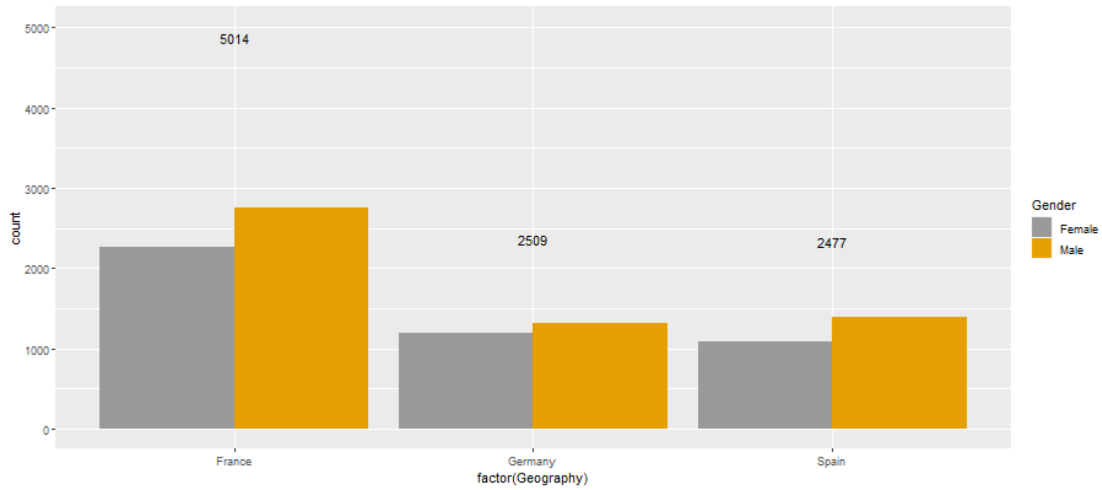






Customer Churn





Logistic Regression:

Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1528   46
         1  318  108

               Accuracy : 0.818
                 95% CI : (0.8004, 0.8347)
    No Information Rate : 0.923
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2924

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7013
            Specificity : 0.8277
         Pos Pred Value : 0.2535
         Neg Pred Value : 0.9708
              Precision : 0.2535
                 Recall : 0.7013
                     F1 : 0.3724
             Prevalence : 0.0770
         Detection Rate : 0.0540
   Detection Prevalence : 0.2130
      Balanced Accuracy : 0.7645

       'Positive' Class : 1
```

Accuracy = 81.8%

Decision Tree:

Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1538   36
         1  236  190

               Accuracy : 0.864
                 95% CI : (0.8482, 0.8787)
    No Information Rate : 0.887
    P-Value [Acc > NIR] : 0.9993

                  Kappa : 0.5105

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8407
            Specificity : 0.8670
         Pos Pred Value : 0.4460
         Neg Pred Value : 0.9771
              Precision : 0.4460
                 Recall : 0.8407
                     F1 : 0.5828
             Prevalence : 0.1130
         Detection Rate : 0.0950
   Detection Prevalence : 0.2130
      Balanced Accuracy : 0.8538

       'Positive' Class : 1
```

Accuracy = 86.4%

SVM:

Support Vector Machine

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1545   29
         1  242  184

               Accuracy : 0.8645
                 95% CI : (0.8487, 0.8792)
    No Information Rate : 0.8935
    P-Value [Acc > NIR] : 1

                  Kappa : 0.5057

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8638
            Specificity : 0.8646
         Pos Pred Value : 0.4319
         Neg Pred Value : 0.9816
              Precision : 0.4319
                 Recall : 0.8638
                     F1 : 0.5759
             Prevalence : 0.1065
         Detection Rate : 0.0920
   Detection Prevalence : 0.2130
      Balanced Accuracy : 0.8642

       'Positive' Class : 1
```

Accuracy = 86.45%

Random Forest:

Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1516   58
         1  199  227

               Accuracy : 0.8715
                 95% CI : (0.856, 0.8859)
    No Information Rate : 0.8575
    P-Value [Acc > NIR] : 0.03787

                  Kappa : 0.5641

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.7965
            Specificity : 0.8840
         Pos Pred Value : 0.5329
         Neg Pred Value : 0.9632
              Precision : 0.5329
                 Recall : 0.7965
                     F1 : 0.6385
             Prevalence : 0.1425
         Detection Rate : 0.1135
   Detection Prevalence : 0.2130
      Balanced Accuracy : 0.8402

       'Positive' Class : 1
```

Accuracy: 87.15%

5. Conclusion:

In predicting if a customer will churn or not, we employed 4 types of models: Logistics Regression, Decision Tree, Support Vector Machine and Random Forest. The performances of the models are fairly good with accuracies ranging from 81% - 87%. Other performance metrics that we considered are sensitivity, recall and f1 score

Overall, Random Forest is the best model for predicting churn among the four models.

6. References

Abou el Kassem, E., Hussein, S.A., Abdelrahman, A.M. and Alsheref, F.K., 2020. Customer churn prediction model and identifying features to increase customer retention based on user generated content. *International Journal of Advanced Computer Science and Applications*, *11*(5).

Amuda, K.A. and Adeyemo, A.B., 2019. Customers churn prediction in financial institution using artificial neural network. *arXiv preprint arXiv:1912.11346*.

Szmydt, M., 2018, July. Predicting customer churn in electronic banking. In *International Conference on Business Information Systems* (pp. 687-696). Springer, Cham.

Bharathi S, V., Pramod, D. and Raman, R., 2022. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. *Data*, *7*(5), p.61.

Kaur, I. and Kaur, J., 2020, November. Customer churn analysis and prediction in banking industry using machine learning. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 434-437). IEEE.

Dalmia, H., Nikil, C.V. and Kumar, S., 2020. Churning of Bank Customers Using Supervised Learning. In *Innovations in Electronics and Communication Engineering* (pp. 681-691). Springer, Singapore.