B.P.H.E Society's

AHMEDNAGAR (M.S)

# DEPARTMENT OF STATISTICS

## T.Y.B.Sc.

## Certificate

**Date:**

**This is to certify that partial fulfillment of curriculum T.Y.B.Sc students, Jayshree Khomane, Geet Bidwai, Sarika Ghule, Shreya Garsund, and Snehal Lagad, have completed the project work in statistics entitled "Customer Churn" prescribed by Savitribai Phule Pune University during the academic year 2023-24.**

**Project Guide: Yogesh. R. Yewale.**

**Examiner:**

# A PROJECT ON

# “Analysing Customer Churn Patterns: Statistical Modelling and Predictive Strategies”

*******Submitted by*******

Bidwai Geet Vivek

Khomane Jayshree Arjun

Ghule Sarika Arjun

Garsund Shreya Shailesh

Lagad Snehal Parasram

*******************

UNDER THE GUIDANCE OF:

H.O.D. MALATI. C. YEOLA

PROF. YOGESH. R. YEWALE

Mrs. RESHMA WAGH

## TABLE OF CONTENT:

## ACKNOWLEDGEMENT:

In fact if not for the grace and guidance of the almighty God it would have been very difficult if not impossible to produce this project document. Although the project is part of the syllabus, it allowed us to apply the knowledge of statistics to real-life problems. This project enabled us to know various statistical tools and their application.

We are thankful to B.P.H.E. Society's Ahmednagar College, Ahmednagar for bestowing us with such an opportunity. We also wish to express our profound gratitude to

**Dr. Malati Yeola**, the Head of the Department of Statistics, and **Mr. Yogesh Yewale and Mrs. Reshma Wagh** for their guidance and invaluable contribution towards the success of this work. We also extend our appreciation to our families, friends, relatives, and all those people who have contributed to this project to get completed.

## DECLARATION:

We declare that this project work was carried out by us, in the Department of Statistics, B.P.H.E. Society's Ahmednagar College, Ahmednagar under the supervision of **Dr. Malati Yeola, Mr. Yogesh Yewale and Mrs. Reshma Wagh,** and that no previous submission for a degree of this college or elsewhere has been made. Related work by others which served as a source of knowledge has been duly acknowledged or referenced

Place: Ahmednagar

Date:  /04/2024

| SEAT NUMBER | NAME |
|---|---|
| 18533 | Khomane Jayshree Arjun |
| 18525 | Bidwai Geet Vivek |
| 18531 | Ghule Sarika Arjun |
| 18529 | Garsund Shreya Shailesh |
| 18535 | Lagad Snehal Parasram |

## OBJECTIVES:

1. Analyzing data.

2. Identify key patterns.

3. Develop effective solutions to reduce churn rates.

4. Factors affecting the Customer churn.

5. Valuable insights to improve customer satisfaction and drive long-term business success.

## INTRODUCTION:

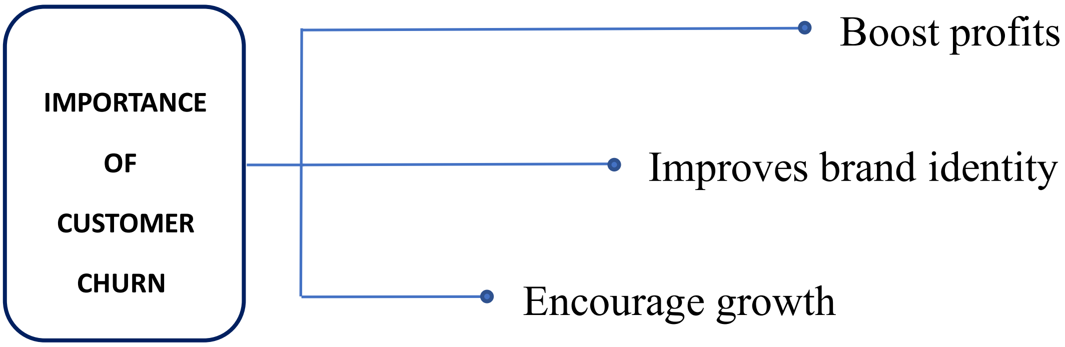
### ●What is customer churn?

Customer churn rate indicates how many of your existing customers are not likely to make another purchase from your business.

A high churn rate is something that all businesses want to avoid. A high churn rate indicates that your customers are not satisfied with the products or services you’re offering.

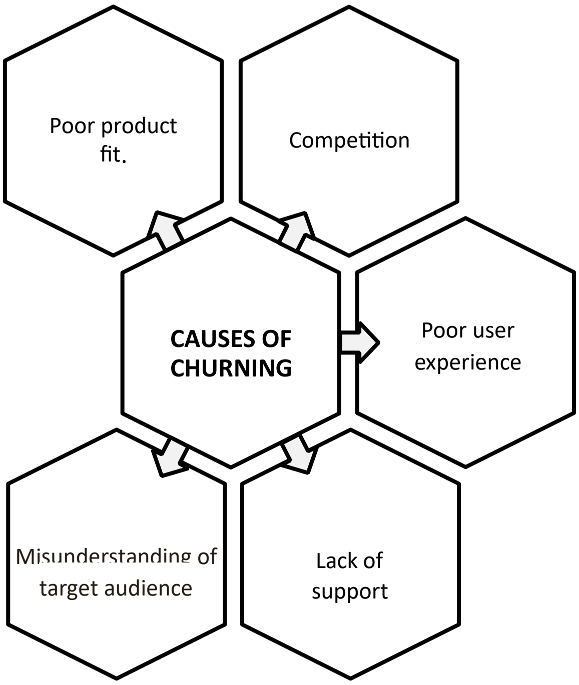
### ●Why is analyzing customer churn important?

Customer churn analysis helps businesses understand the percentage of customers who are no longer supporting them. It’s important to analyze your customer churn for various reasons, such as:



### ●What causes customer churn?

Even the most successful of businesses will experience customer churn from time to time. However, identifying exactly what’s causing your customers to leave can give you a better idea of what you need to fix. Various factors can cause customer churn, including:



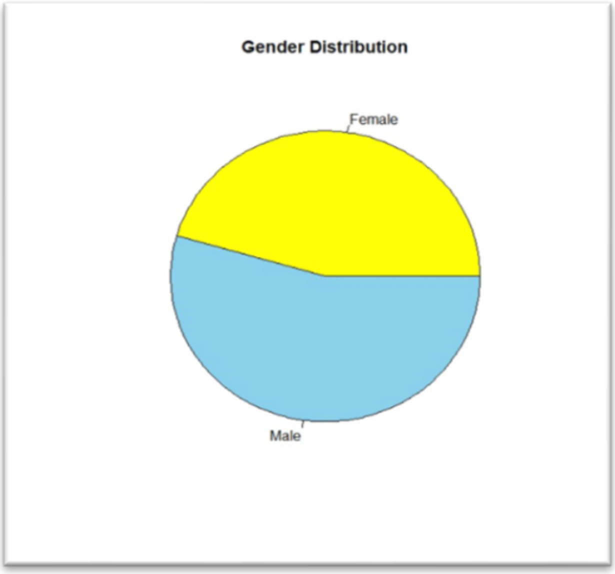
- Poor product fit**:** If your products aren't a good fit for your target audience, they're not going to keep purchasing, no matter how emotionally connected they are to your brand. A poor product fit can be detrimental to a brand and cause them to lose customers.
- Competition**:** It's very unlikely that you're the only business in your industry that sells the same products, so you always need to keep your competition in mind. If your competition offers better products or services at a lower price, your customers are most likely going to choose them over you.
- Poor user experience**:** If your website is confusing and difficult to use, your customers can get frustrated, click the back button, and visit another site. Having a poor user experience can hinder the success of your business because customers want to choose a brand that is functional and easy to use.
- Lack of support**:** Connecting is important for any business because it helps you build brand loyalty. When your customers trust you, they're going to continue to support you and will even recommend your brand to their peers
- Misunderstanding of target audience**:** If you don't understand your target audience, you don't know what they want. Misunderstanding your target audience can cause customer churn because you could be creating products or services that aren't what your customers need. The best way to understand your target audience is to track customer feedback so you can gain insight into exactly what they're looking for.

## DESCRIPTION OF VARIABLES:

- **Row Number:** Corresponds to the record number.
- **Surname:** The surname of a customer's credit score can affect customer churn since a customer with a higher credit score is less likely to leave the bank.
- **Geography:** A customer's location can affect their decision to leave the bank.
- **Gender:** It's interesting to explore whether gender plays a role in a customer leaving the bank.
- **Age:** This is certainly relevant since older customers are less likely to leave their bank than younger ones.
- **Tenure:** Refers to the number of years that the customer has been a client of the bank. Normality, older clients are more loyal and less likely to leave a bank.
- **Balance:** Also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
- **Number of products:** Refers to the number of products that a customer has purchased through the bank.
- **Has Credit Card:** Denotes whether or not a customer has a credit card. This column is also relevant since people with credit cards are less likely to leave the bank.
- **Is Active Member:** Active customers are less likely to leave the bank.
- **Estimated Salary:** As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- **Exited:** Whether or not the customer left the bank.
- **Complain:** Customer has a complaint or not.
- **Satisfaction Score:** Score provided by the customer for their complaint resolution.
- **Card Type:** The type of card held by the customer.
- **Points Earned:** The points earned by the customer for using a credit card

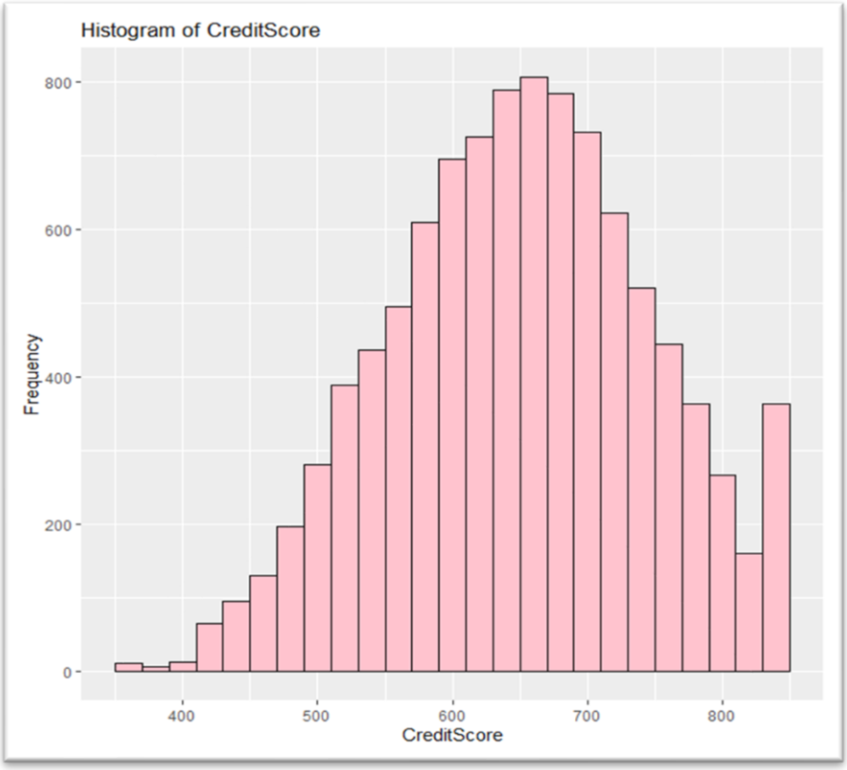
## DATA VISUALIZATION:

- GENDER DISTRIBUTION

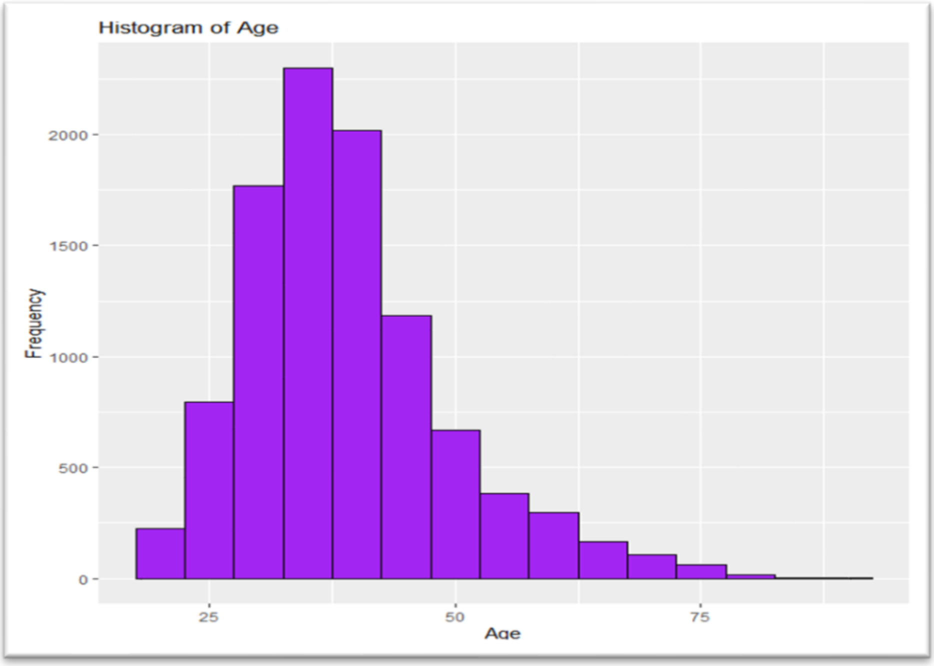


| MALE | FEMALE |
|---|---|
| 5457 | 4543 |

- DISTRIBUTION OF CREDIT SCORE

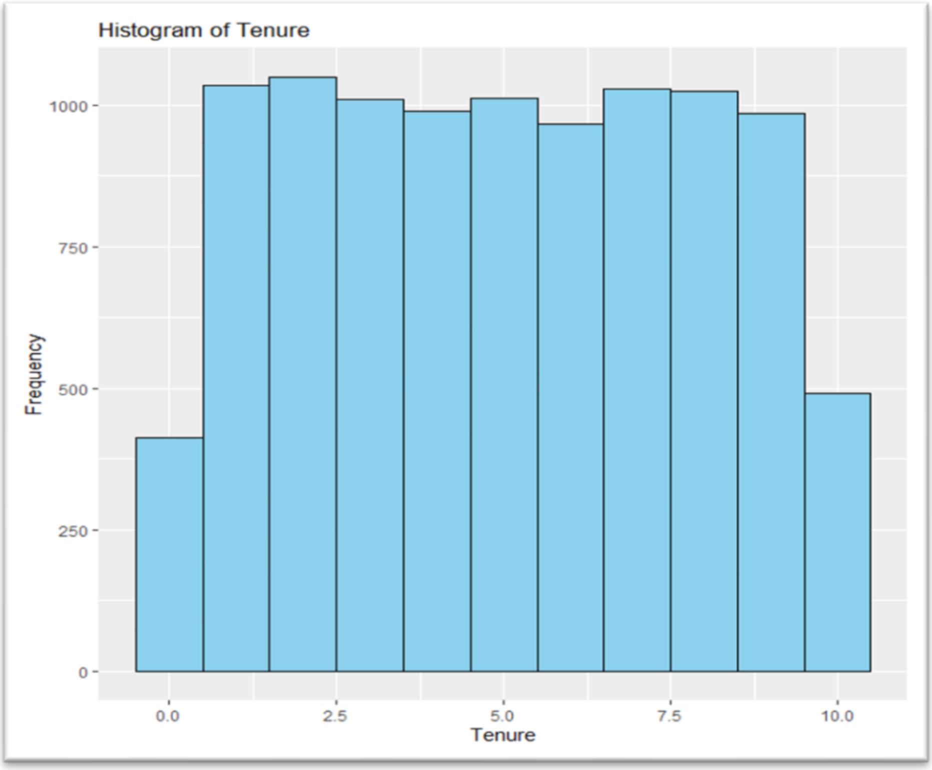


- HISTOGRAM FOR AGE



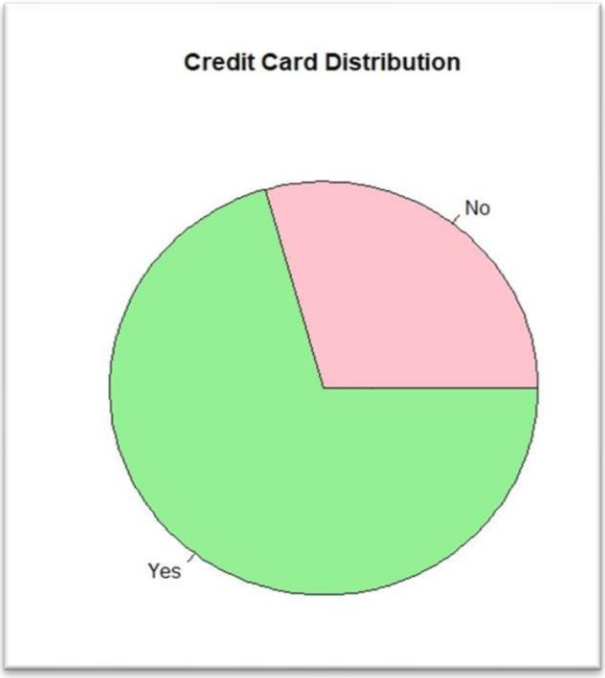
- HISTOGRAM FOR TENURE: ○ Frequency distribution

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 413 | 1035 | 1048 | 1009 | 989 | 1012 | 967 | 1028 | 1025 | 984 | 490 |



- PIE CHART FOR THE CUSTOMERS WHO HAVE POSSESSION OF CREDIT CARDS.

- Frequency distribution



| Not having card | Having card |
|---|---|
| 2945 | 7055 |

## METHODOLOGY:

We have used the secondary data set for this project. The dependent variable exited. We coded them in numerical format, 1 for churn and 0 for no churn. For this, we use MS Excel. Since our dependent variable is binary type so we use the machine learning algorithm logistic regression for prediction.

- Logistic Regression

  Another method used in machine learning is
  Random forest & Decision tree. By comparing the accuracy of the above three models we can conclude which one is the best.

Our project includes an understanding of Machine Learning and its basic types. The classification models were used to analyze the churn status of customers. The supervised classification models namely decision tree, Random forest, and Logistic Model were fitted to our sample data of 7000 sample points.

Using the Train, Test data set in the fitting of models, first split the data into 70% Training set and 30% Testing set we fit the model train data set and check this model on the train data set.

## DATA SOURCE:

We have taken this data from Kaggle.com

https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn.

## TEST’s USED:

**TWO SAMPLE T-TEST**:

●Assumptions of the t-test are:

1)INDEPENDENCE: The observations in each sample must be independent of each other.

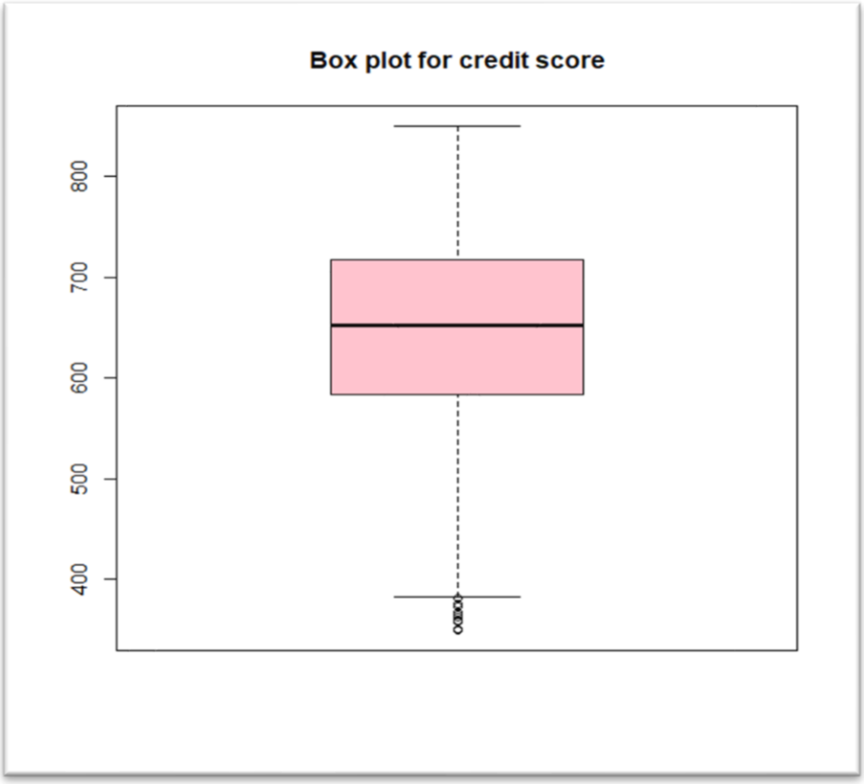
2)NORMALITY: The data in each sample should be approximately normally distributed. However, the t-test is robust to moderate deviations from normality, especially with larger sample sizes.

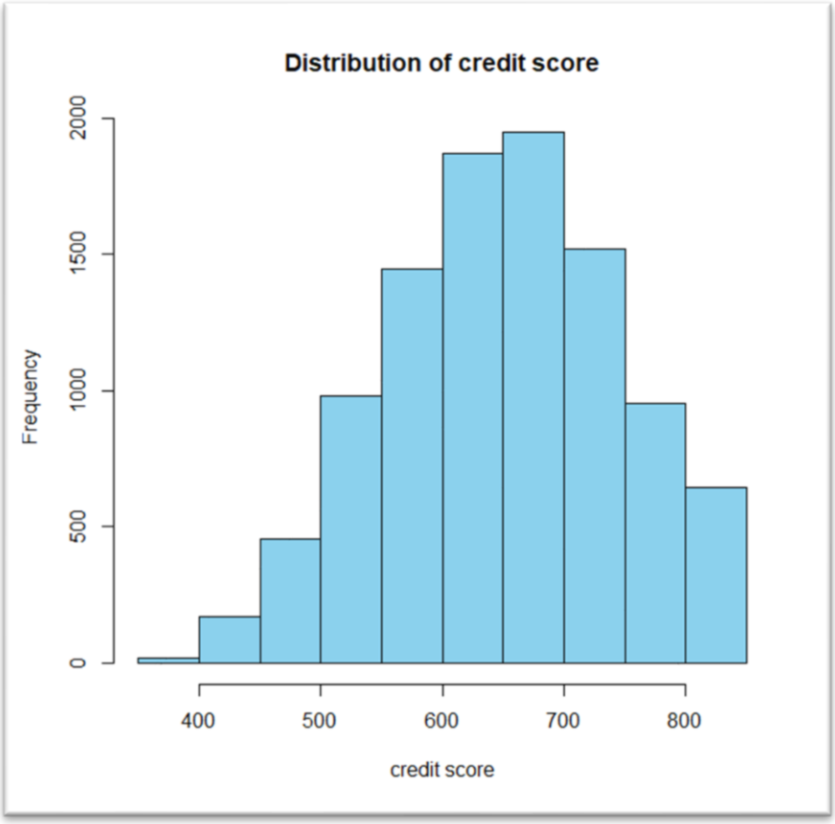
3)EQUAL VARIANCE: The variance in each group should be approximately equal.

●Checking assumptions of t-test:

**CODE:**

```
data=read.csv("C:\\Users\\HP\\Downloads\\Customer-Churn-Records.csv",header=T)

head(data) x1=data$CreditScore x2=data$Exited boxplot(x1,main="Box plot for

credit score",col=”pink”) hist(x1,main="Distribution of credit score",col=”skyblue”)

shapiro.test(x1[x2==1])
```





From the above histogram and box plot we can observe that the assumption of normality is not satisfied by the data.

Alternative for checking normality condition:

```
*Shapiro wilk normality test* Shapiro-

Wilk normality test data:  x1[x2 == 1]

W = 0.99368, p-value = 1.166e-07
```

Conclusion: Since the p-value is less than 0.05, the test can state that the data will not fit the distribution normally with 95% confidence.

● As the assumption of normality is not satisfied by the given data so, we go for a nonparametric test.

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric alternative to the independent samples t-test.

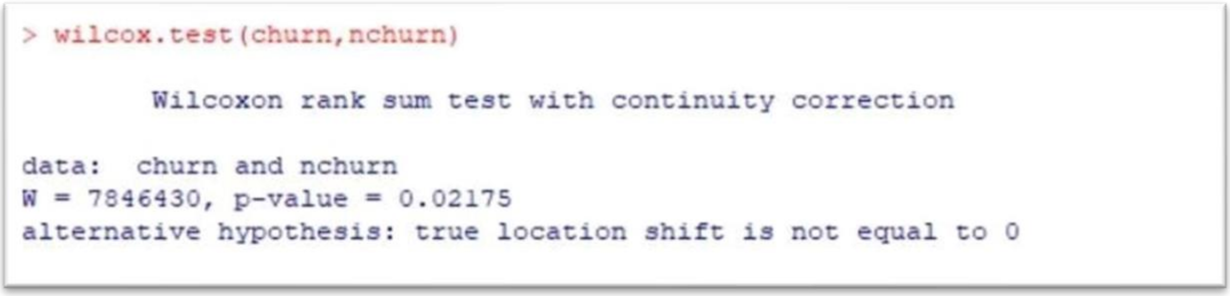
NULL HYPOTHESIS (Ho): There is no significant difference in the average credit score between customers who churned and customers who did not churned.

ALTERNATIVE HYPOTHESIS (H1): There is a significant difference in the average credit score between the customers who churned and customers who did not churn.

CODE:

```
churn_score=a[b==1] nchurn_score=a[b==0]

print ("number of customers who churned:")

print(length(churn_score))

-[1] 2038

Print ("number of customers who did not churned:") print(length(nchurn_score))

-[1] 7962 wilcox.test(churn,nchurn)
```

**OUTPUT:**



```
> wilcox.test(churn,nchurn)

        Wilcoxon rank sum test with continuity correction

data:  churn and nchurn
W = 7846430, p-value = 0.02175
alternative hypothesis: true location shift is not equal to 0
```

**INTERPRETATION:**

Since the p-value is less than 0.05 this implies that there is enough evidence to reject Ho and we can conclude that there is a significant difference between the two groups.

## CHI-SQUARE TEST:

The chi-square test is a statistical test used for categorical data analysis to examine the association or independence between two categorical variables. It assumes whether there is a significant relationship between the variability or if they are independent of each other.

●Analyzing the association between the categorical variable ‘Gender’ & ‘Customer churn’

HYPOTHESIS:

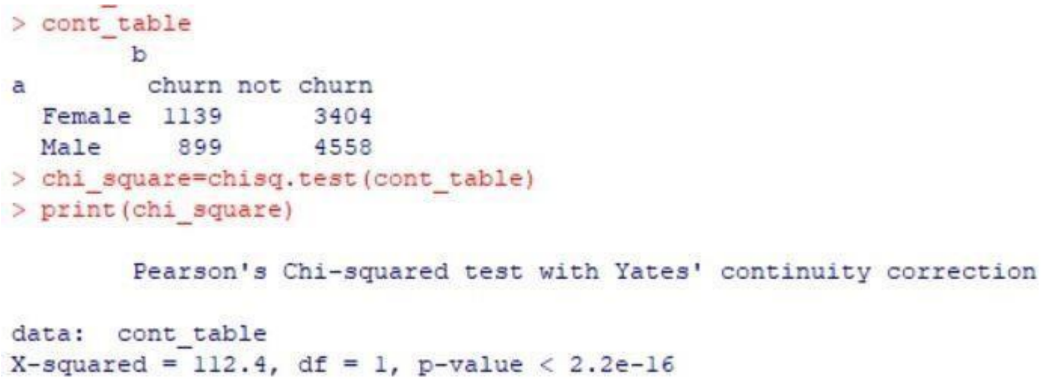
NULL HYPOTHESIS(HO): There is no association between gender and churn.

ALTERNATIVE HYPOTHESIS(H1): Gender & Churn are associated.

**CODE:**

```
 data=read.csv("C:\\Users\\HP\\Documents\\Customer-Churn-
Records.csv",header=T) a=data$Gender b=data$x cont_table=table(a,b) cont_table
```

**OUTPUT:**



```
> cont_table
        b
a        churn not churn
  Female  1139      3404
  Male     899      4558
> chi_square=chisq.test(cont_table)
> print(chi_square)

        Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 112.4, df = 1, p-value < 2.2e-16
```

**INTERPRETATION:**

Since the value of the test statistic is high this implies that there is a stronger association between gender and churn.

Also, the p-value is less than the chosen significance level(eg.0.05) which implies that there is enough evidence to reject the null hypothesis.

**CONCLUSION:**

We reject the null hypothesis and conclude that there is a significant association between the two variables.

## LOGISTIC REGRESSION:

Logistic regression is a statistical method used for binary classification, which means it predicts the probability of a binary outcome. It is called logistic because it is based on the logistic function(also known as the sigmoid function)

**ASSUMPTIONS:**

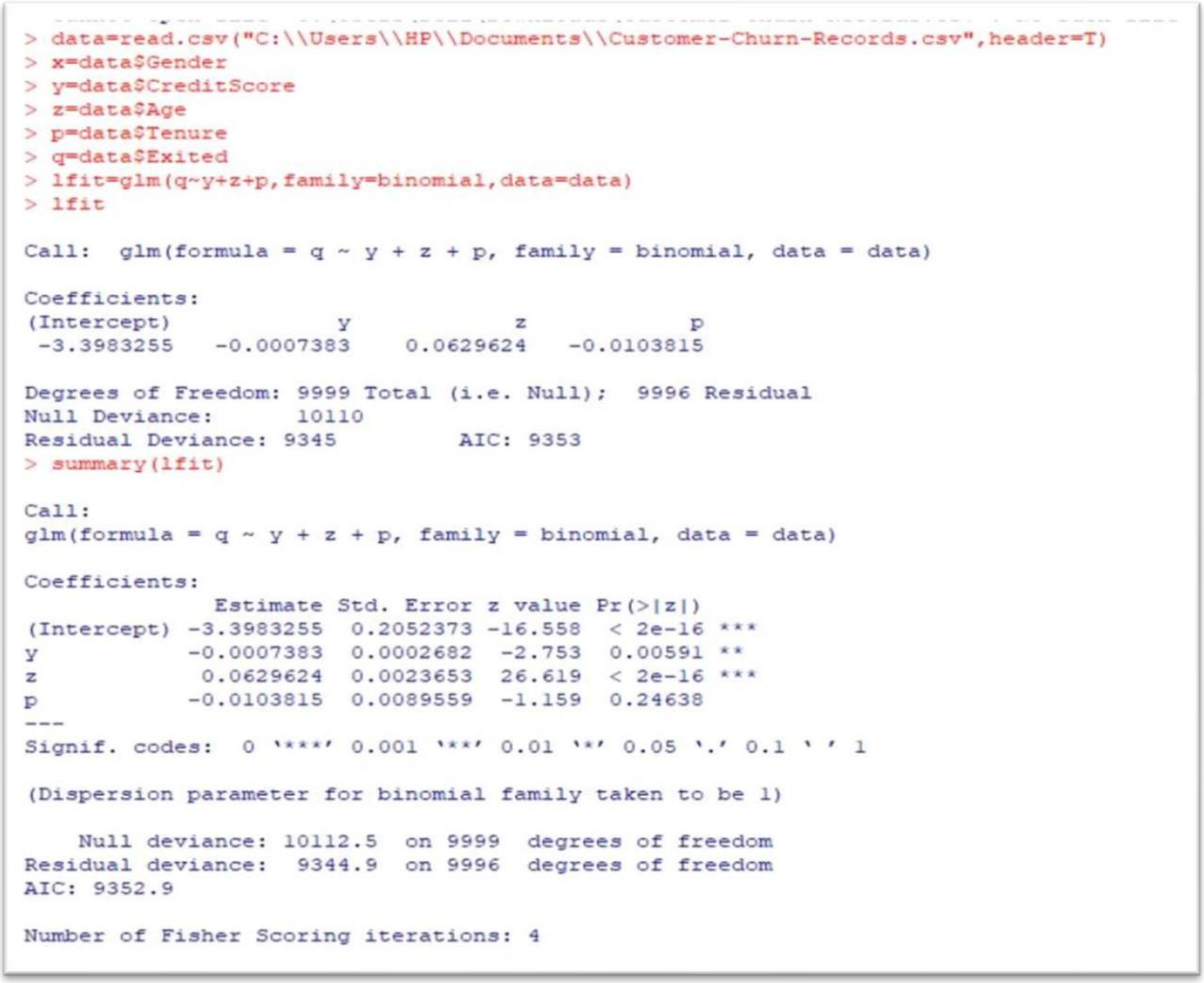
1)Binary outcome: The  dependent variable y must be binary or dichotomous. It represent the probability of the event occurring(customer churn).

2)Independent observations:Each observation is independent of the other.

3)No outliers: There should be no outliers in the dataset.

4)large sample size: The sample size is sufficiently large.

**OUTPUT:**



```
> data=read.csv("C:\\Users\\HP\\Documents\\Customer-Churn-Records.csv",header=T)
> x=data$Gender
> y=data$CreditScore
> z=data$Age
> p=data$Tenure
> q=data$Exited
> lfit=glm(q~y+z+p,family=binomial,data=data)
> lfit

Call:  glm(formula = q ~ y + z + p, family = binomial, data = data)

Coefficients:
(Intercept)            y            z            p
 -3.3983255   -0.0007383    0.0629624   -0.0103815

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:      10110
Residual Deviance: 9345        AIC: 9353
> summary(lfit)

Call:
glm(formula = q ~ y + z + p, family = binomial, data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3983255  0.2052373 -16.558  < 2e-16 ***
y           -0.0007383  0.0002682  -2.753  0.00591 **
z            0.0629624  0.0023653  26.619  < 2e-16 ***
p           -0.0103815  0.0089559  -1.159  0.24638
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10112.5  on 9999  degrees of freedom
Residual deviance:  9344.9  on 9996  degrees of freedom
AIC: 9352.9

Number of Fisher Scoring iterations: 4
```

**INTERPRETATION:**

Coefficients: it represents the estimated effect of each predictor variable on the log odds of the outcome(churn).

$e^(\beta_1^)=0.9926$. Implies that the chances of churn decreased by 99.26% if there is one unit increament in credit score. $e^(\beta_2^)=1.0649$. Implies that the chances of

churn increased by 6.49% if there is one unit increment in age.

$e^(\beta_3^)=0.9896$ Implies that the chances of churn decreased by 98.96% if

there is one unit increament in tenure.

To test the significance of regression:

Ho: β1=β2=β3=0 Vs H1: βj ≠ 0 (for atleast one j )

Under Ho,test statistic is

G=D (model that excludes regressor)

G=Null deviance – Residual deviance

G→χ^2 with k degree of freedom. (K=3)

G=767.6 χ^2

(3,0.05)=7.815

G > χ^2 table

Decision: Reject Ho at 5% L.O.S

Accept H1 i.e βj ≠ 0 for at least 1 j.

Std error.: Variable y is more precise estimates.

## KAPLAN MEIR ESTIMATOR:

Kaplan Meir estimator is a non-parametric method used to estimate the survival function from lifetime data.

It calculates the probability that an individual survives beyond a certain time point given the observed data.

**ASSUMPTIONS:**

- The Kaplan Meir estimator does not assume any specific distribution for the survival times.
- It assumes that censoring is non-informative, meaning that the probability of being censored at any time point is unrelated to the individual's survival experience.

**INTERPRETATION:**

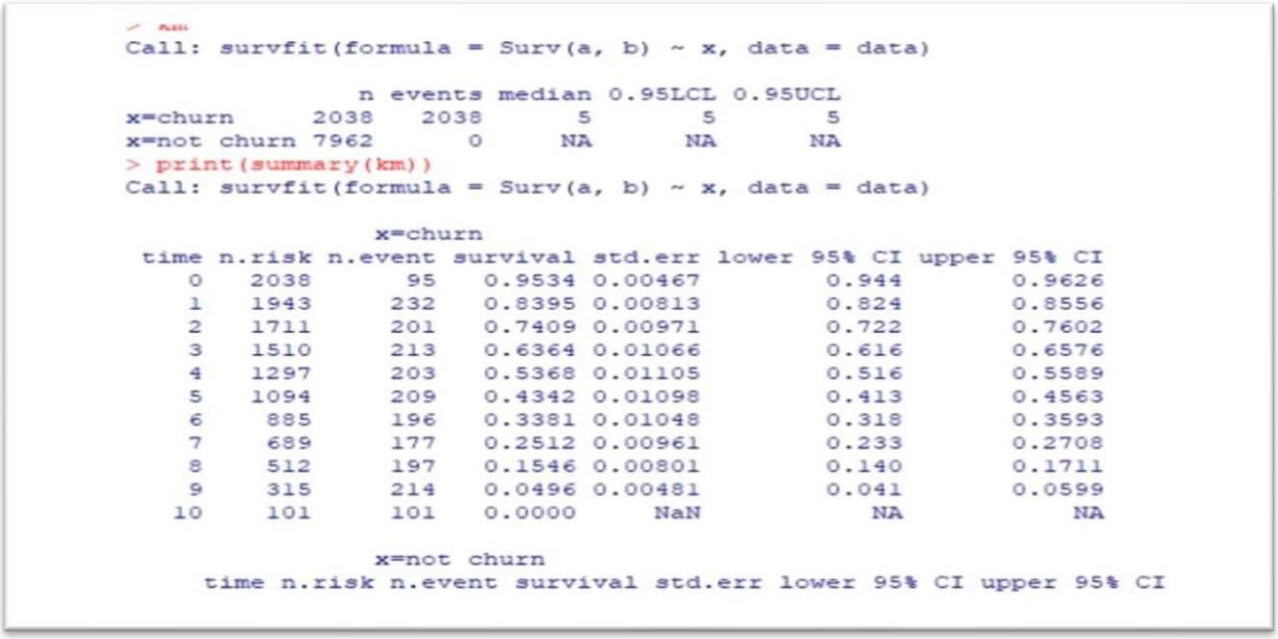
The Kaplan-Meir estimator provides a curve that represents the estimated probability of customer survival (not churning) over time.

The curve can be used to identify periods of high and low churn risk.

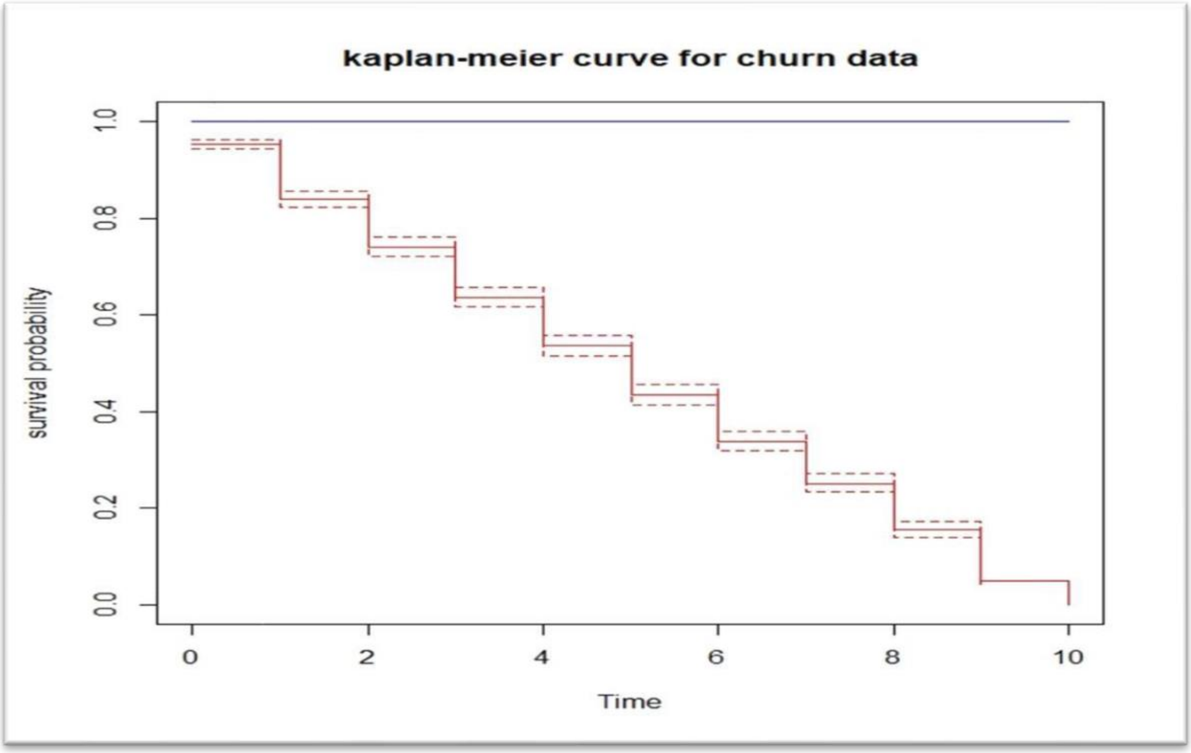
**CODES:**

```
data=read.csv("C:\\Users\\HP\\Documents\\Customer-Churn
Records.csv",header=T)  a=data$Tenure  b=data$Exited  x=data$x
library(survival)
km=survfit(Surv(a,b)~x,data=data) km  print(summary(km))
plot(km,xlab="Time",ylab="survival probability",main="kaplan-meier curve for churn
data",conf.int=T,col=c("red","blue"))
```

**OUTPUT:**



```
> km
Call: survfit(formula = Surv(a, b) ~ x, data = data)

               n events median 0.95LCL 0.95UCL
x=churn     2038   2038      5       5       5
x=not churn 7962      0     NA      NA      NA
> print(summary(km))
Call: survfit(formula = Surv(a, b) ~ x, data = data)

                x=churn 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0   2038      95   0.9534 0.00467        0.944       0.9626
    1   1943     232   0.8395 0.00813        0.824       0.8556
    2   1711     201   0.7409 0.00971        0.722       0.7602
    3   1510     213   0.6364 0.01066        0.616       0.6576
    4   1297     203   0.5368 0.01105        0.516       0.5589
    5   1094     209   0.4342 0.01098        0.413       0.4563
    6    885     196   0.3381 0.01048        0.318       0.3593
    7    689     177   0.2512 0.00961        0.233       0.2708
    8    512     197   0.1546 0.00801        0.140       0.1711
    9    315     214   0.0496 0.00481        0.041       0.0599
   10    101     101   0.0000     NaN           NA           NA

                x=not churn 
     time n.risk n.event survival std.err lower 95% CI upper 95% CI
```



**CONCLUSION:**

A higher no of events at a particular time point indicates a higher event (customer churn) rate during that period. Since at t=1 rate of event is high.

Survival: Probability of not experiencing the churn.

Std. error: A smaller standard error indicates a more precise estimate of the survival probability.

Median: the median survival time is 5 months, which means that half of the customers have churned by the end of 5 months.

## MACHINE LEARNING:

- **What is Machine Learning?**

Machine Learning (ML) is basically the study of computer algorithms that can improve automatically through experience and by the use of past data. It is seen as a part of Artificial Intelligence (AI). Machine Learning algorithms build a model based on sample data, known as training data, in order to make decisions and test its accuracy with the help of test data. Machine Learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, computer vision, etc.

Nowadays, the demand of statistics in ML is increasing day by day. In models, statistical methods are required in the preparation of train data and test data and also to check the accuracy of the models

This includes:

- Outlier detection
- Missing value imputation
- Data sampling
- Data scaling
- Variable encoding.

This all can be done in machine learning by applying the proper statistical tools.

- **Why we use it?**

The response variable of our data was in the form of classification type. So, we classify our data in two groups namely churn or not churn water as like a binary variable.

Churn=1

Not churn=0

There are also some classification models that are used in machine learning. Examples of those models are:

1. Logistic Regression

2. K-Nearest Neighbour

3. Support Vector Machines

4. Kernel SVM

5. Naïve Bayes

6. Decision Tree Classification

7. Random Forest Classification

8. ANN

9. CNN

We want to develop a model that can predict the values of churn. Our focus is on both accuracy of the predictions and interpretability of the model.

Therefore, we have chosen the models that suits our data best. We will evaluate three different models covering the complexity spectrum.

1)Decision Tree

2)Random Forest

3)Logistic Regression

To use the machine learning model the basic assumption is that there should be no multicollinearity between the regressor. So, in our data type let X1, X2, X3, X4, X5, X6, X7, X8, X9 be Balance , Age , Is active member, Num of products , credit score, Tenure , Estimated salary respectively. These are the regressor in our data which affects the value of the response variable. So to check the multicollinearity between the regressor we use the Heat Map as statistical tool.

## HEAT MAP:

- **What is heat map?**

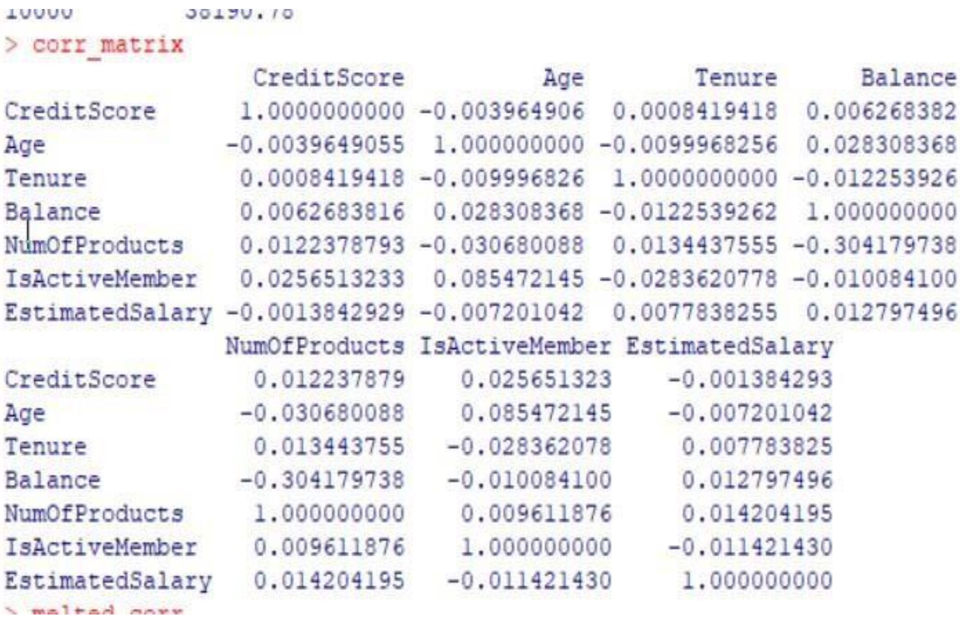
A heat map is basically the representation of two-dimensional information (data) with the help of colors. It gives warm-to-cool spectrum to show which parts of a data has the most attention. We use the heatmap as a correlation matrix. In heatmap correlation matrix, both the axis has same variables and we check the correlation between them by using it. The dark color represents the positive correlation and the medium light color gives no correlation between the variable.

As it gives visual as well as numerical value to check the correlation. The values in the cell indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship. In addition, correlation plots can be used to identify outliers and to detect linear and non-linear relationships. The color-coding of the cells makes it easy to identify relationships between variables at a glance.

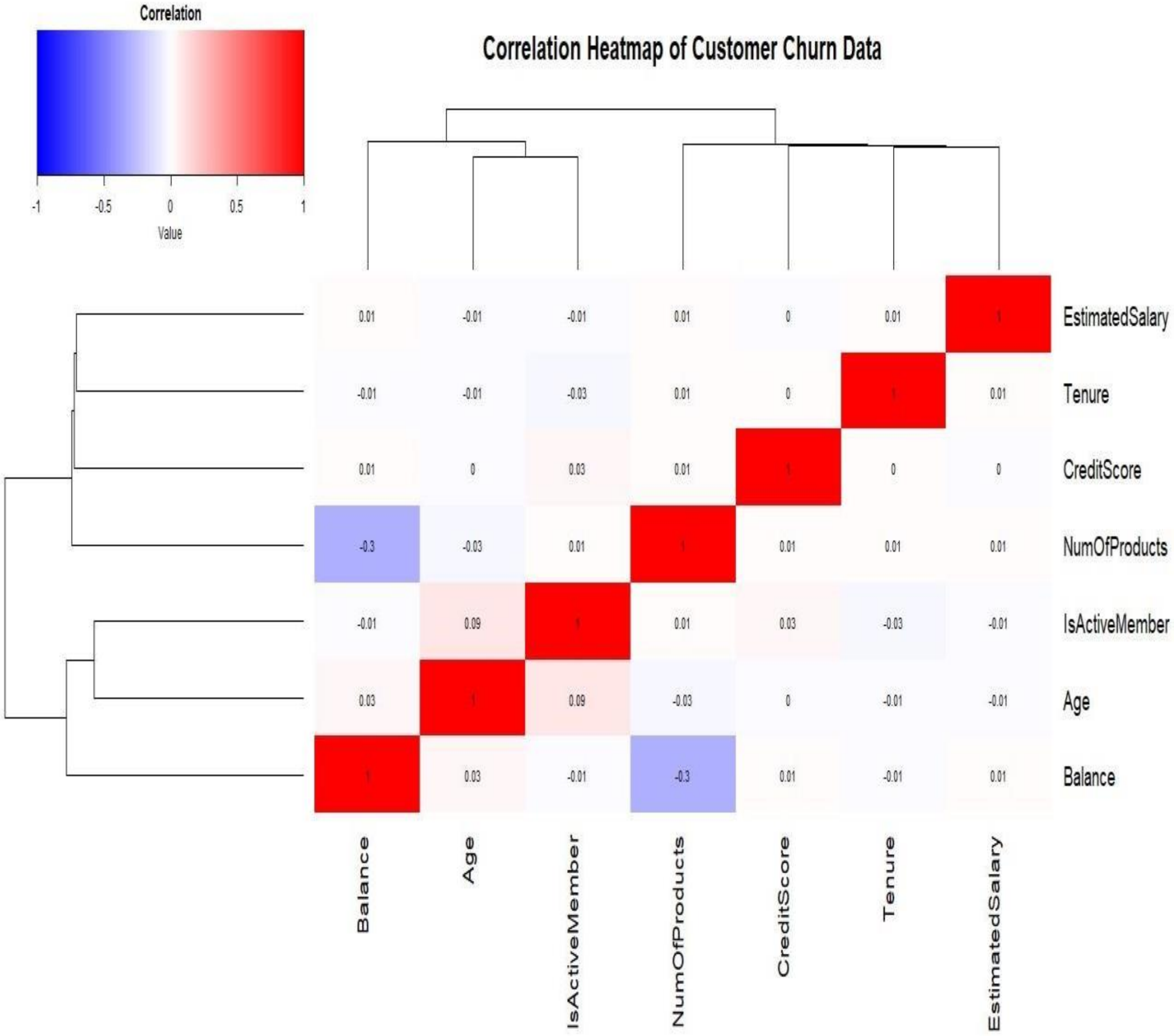
### Check the multicollinearity between the regressors:

By using R PROGRAMMING, we plot the heatmap for our data. The respective commands are as follows:

```
#Correlation using heatmap corr_matrix=cor(num_data)

library(reshape2) melted_corr=melt(corr_matrix)

heatmap.2(corr_matrix,        trace = "none",        col =

colorRampPalette(c("blue",

"white", "red"))(100),        scale = "none",        symm = TRUE,

margins = c(10,10),


     main = "Correlation Heatmap of Customer Churn Data",        key

= TRUE,        keysize = 1.5,        key.title =

"Correlation",        density.info = "none",        cellnote =

round(corr_matrix, 2),        notecol = "black",        notecex =

0.8)
```



```
10000         38190.78
> corr_matrix
                 CreditScore           Age        Tenure      Balance
CreditScore     1.0000000000 -0.003964906  0.0008419418  0.006268382
Age            -0.0039649055  1.000000000 -0.0099968256  0.028308368
Tenure          0.0008419418 -0.009996826  1.0000000000 -0.012253926
Balance         0.0062683816  0.028308368 -0.0122539262  1.000000000
NumOfProducts   0.0122378793 -0.030680088  0.0134437555 -0.304179738
IsActiveMember  0.0256513233  0.085472145 -0.0283620778 -0.010084100
EstimatedSalary -0.0013842929 -0.007201042  0.0077838255  0.012797496
                NumOfProducts IsActiveMember EstimatedSalary
CreditScore       0.012237879    0.025651323    -0.001384293
Age              -0.030680088    0.085472145    -0.007201042
Tenure            0.013443755   -0.028362078     0.007783825
Balance          -0.304179738   -0.010084100     0.012797496
NumOfProducts     1.000000000    0.009611876     0.014204195
IsActiveMember    0.009611876    1.000000000    -0.011421430
EstimatedSalary   0.014204195   -0.011421430     1.000000000
> melted_corr
```



## Conclusion of heat map:

As the correlation coefficient are negligible, we can conclude that the parameters are uncorrelated.

## DECISION TREE:

●What is a decision tree?

The decision tree is a decision support tool that uses a tree-like method of decisions and their possible consequences, including chance event outcomes, and resources. It is a Supervised Machine Learning algorithm that uses a set of rules to make decisions. It is one of the classification algorithms which uses a rule-based approach.

For example: Planning the next vacation depends on various factors such as time, no. of members, budget. It can perform both classification and regression tasks so referred to as the CART algorithm (Classification And Regression Tree).

Intuition: We need to use dataset features to create YES/NO type questions (In our case churn status) until we isolate all data points belonging to each class.

Model characteristics:

1. Fewer the splits more the accuracy.

2. Algorithms assign only one class to each leaf node.

3. It picks the best split to minimize loss function on the basis of purity – "GINI

Impurity"

$$G = 1 - \sum_{k=1}^{c} (P_k^2)$$

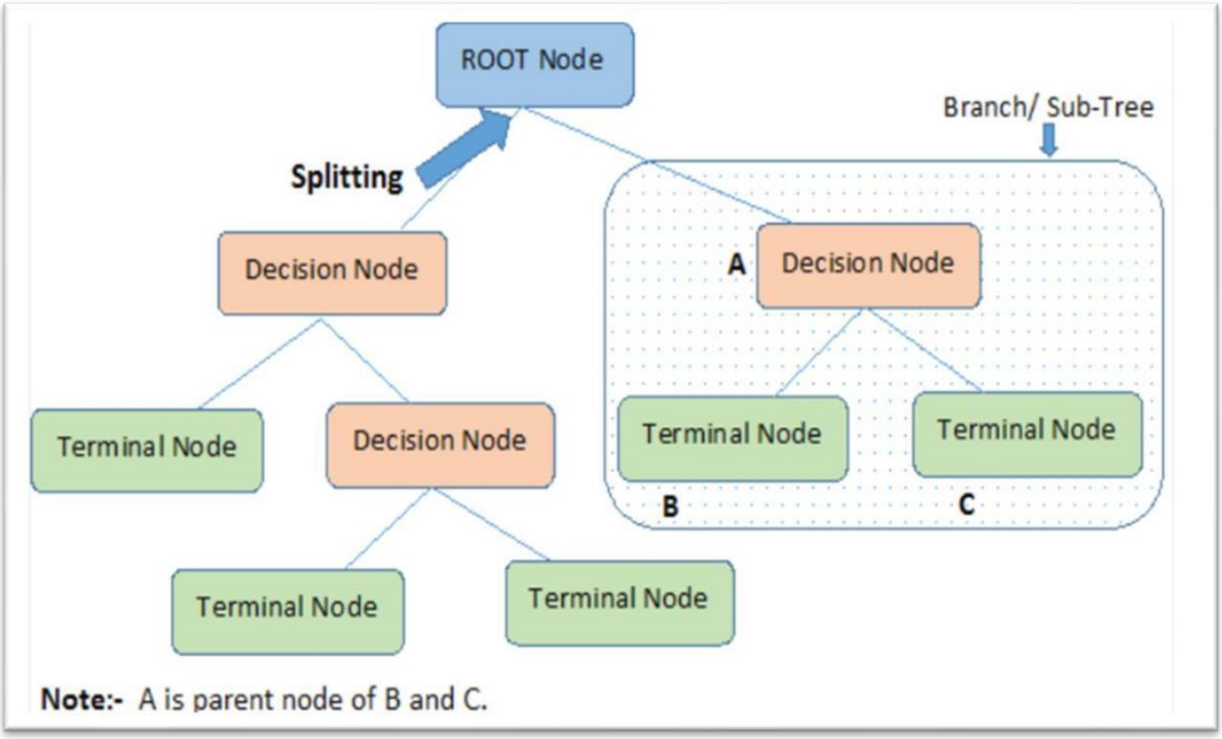
4. Uses greedy approach

5. It can be linearized into decision rules

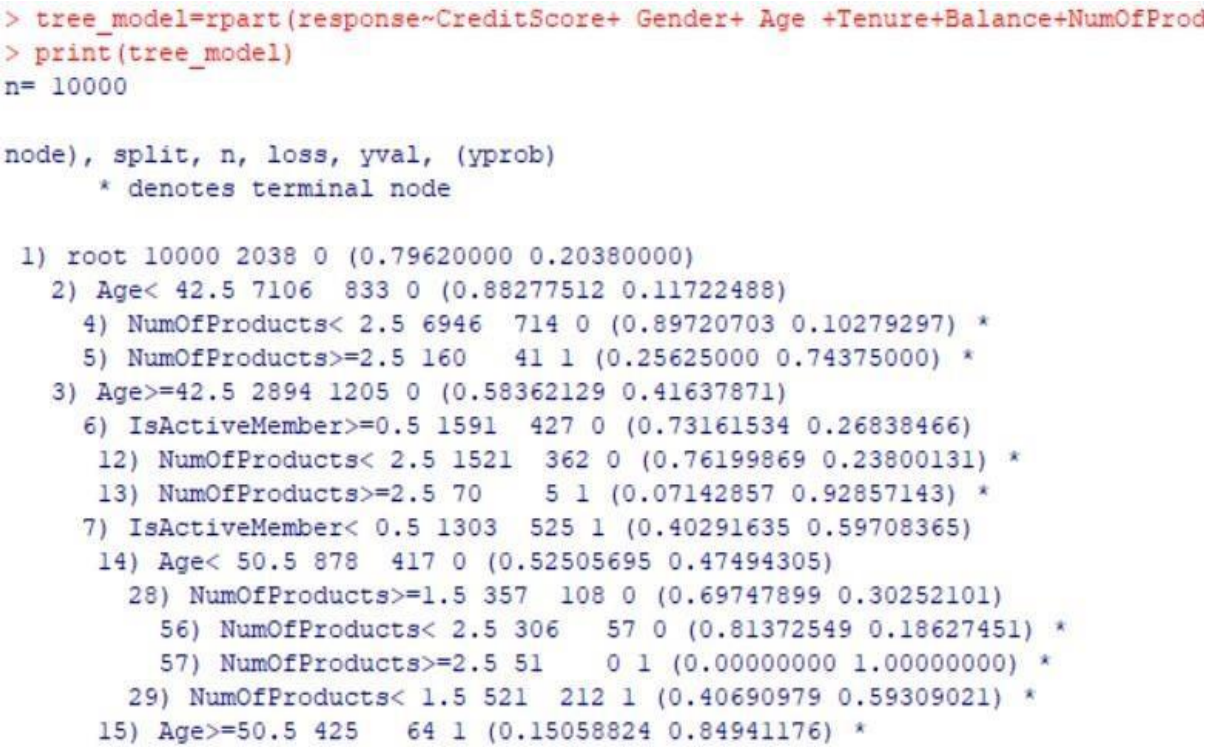
6. It should be parallel by a probability model as a choice model

7. Descriptive means for calculating conditional probabilities

8. Categorical variable decision tree

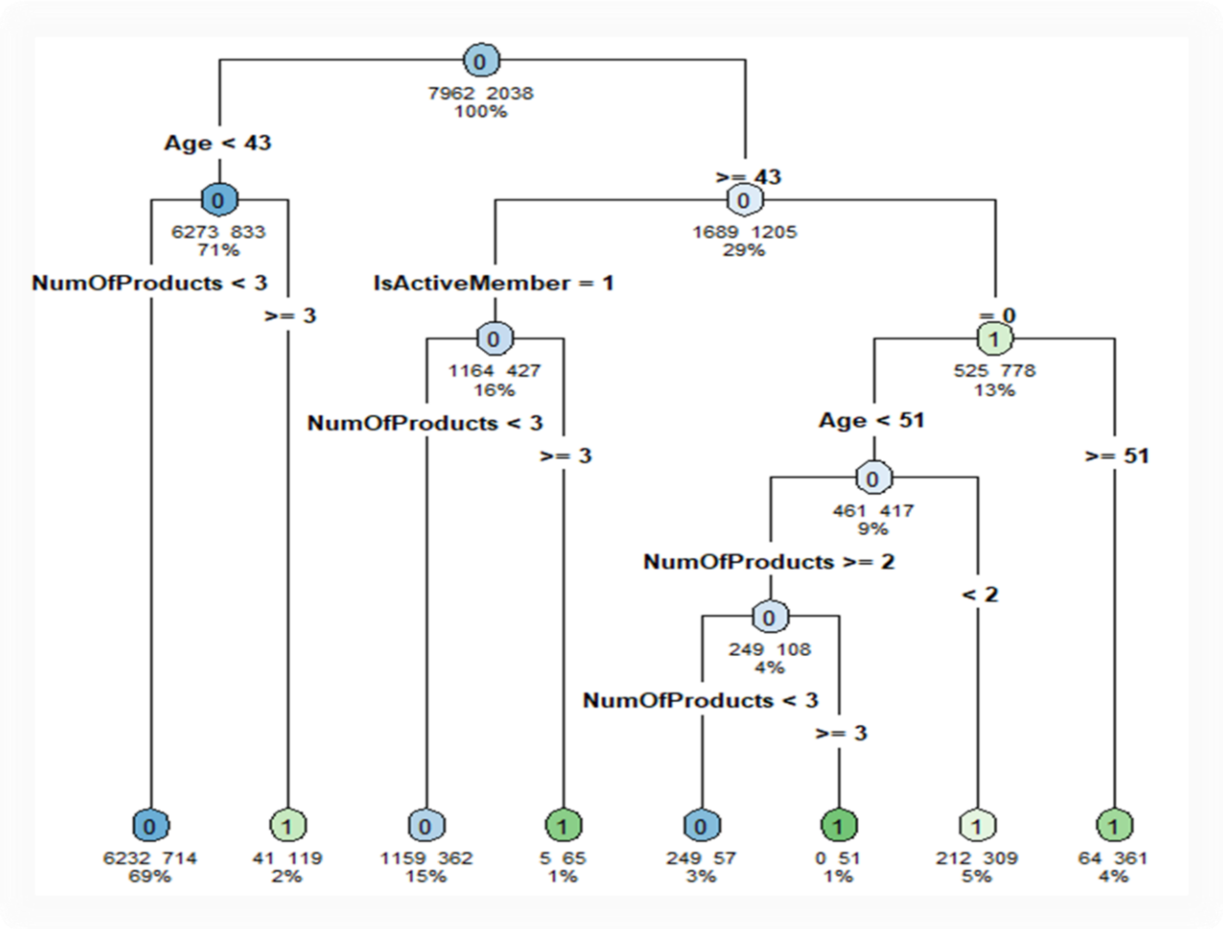


**Note:-** A is parent node of B and C.



```
> tree_model=rpart(response~CreditScore+ Gender+ Age +Tenure+Balance+NumOfProd
> print(tree_model)
n= 10000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 10000 2038 0 (0.79620000 0.20380000)
   2) Age< 42.5 7106  833 0 (0.88277512 0.11722488)
     4) NumOfProducts< 2.5 6946  714 0 (0.89720703 0.10279297) *
     5) NumOfProducts>=2.5 160   41 1 (0.25625000 0.74375000) *
   3) Age>=42.5 2894 1205 0 (0.58362129 0.41637871)
     6) IsActiveMember>=0.5 1591  427 0 (0.73161534 0.26838466)
      12) NumOfProducts< 2.5 1521  362 0 (0.76199869 0.23800131) *
      13) NumOfProducts>=2.5 70    5 1 (0.07142857 0.92857143) *
     7) IsActiveMember< 0.5 1303  525 1 (0.40291635 0.59708365)
      14) Age< 50.5 878  417 0 (0.52505695 0.47494305)
        28) NumOfProducts>=1.5 357  108 0 (0.69747899 0.30252101)
          56) NumOfProducts< 2.5 306   57 0 (0.81372549 0.18627451) *
          57) NumOfProducts>=2.5 51    0 1 (0.00000000 1.00000000) *
        29) NumOfProducts< 1.5 521  212 1 (0.40690979 0.59309021) *
      15) Age>=50.5 425   64 1 (0.15058824 0.84941176) *
```

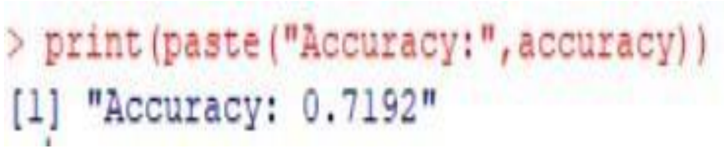
C



Conclusion: Customers with age less than 43 and using products less than 3 are less likely to churn.

**ACCURACY:**

accuracy=mean(predictions==test_data$Exited) print(paste("Accuracy:",accuracy))



```
> print(paste("Accuracy:",accuracy))
[1] "Accuracy: 0.7192"
```

**Conclusion:**

The accuracy rate for fitting the model given by the Decision tree is 71.92% of our data.

**Advantages:**

1. It is easy to implement

2. It is robust to noisy training data

3. It can be more effective if the training data is large

**Disadvantages:**

1. It is computationally intensive

2. Sensitive to outliers

3. Requires feature scaling

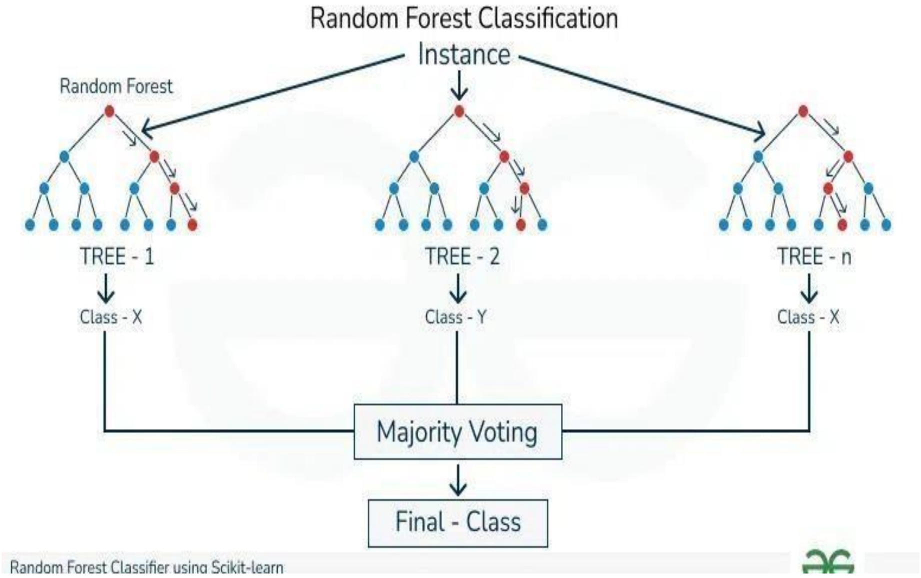
4. Needs a suitable value for k

5. Imbalanced data

6. Curse of dimensionality

**RANDOM FOREST:**

Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.



**ADVANTAGES:**

●Reduced risk of overfitting

●Provides flexibility

●Easy to determine feature importance

DISADVANTAGES:

●Time-consuming process:

●Requires more resources

●More complex: The prediction of a single decision tree is easier to interpret when compared to a forest of them.

**MODEL FITTING:**

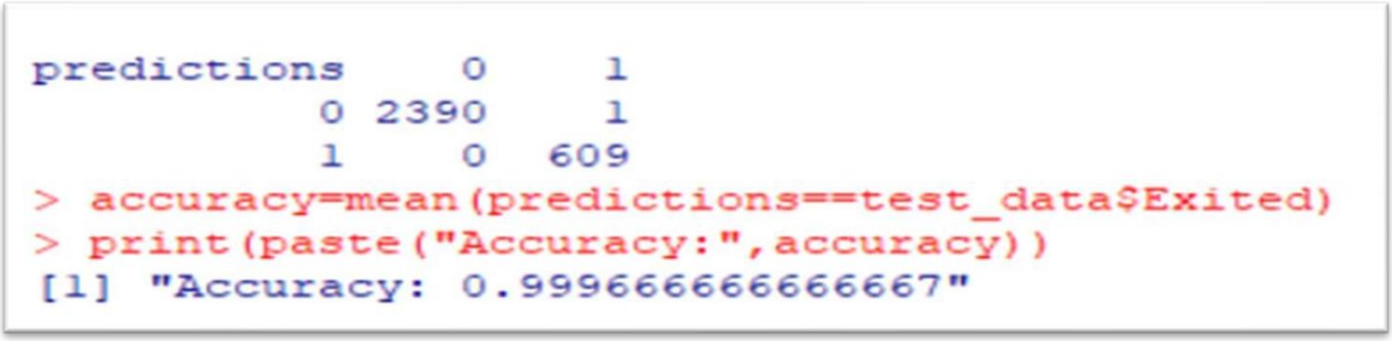
rf_model=randomForest(response~CreditScore+ Gender+ Age +Tenure+Balance+NumOfProducts+ HasCrCard+ IsActiveMember + EstimatedSalary ,data=data,method="class")

**CONFUSION MATRIX:**

conf_matrix <- table(predictions, test_data$Exited) print(conf_matrix)

**ACCURACY:**

accuracy=mean(predictions==test_data$Exited) print(paste("Accuracy:",accuracy))



```
predictions    0    1
          0 2390    1
          1    0  609
> accuracy=mean(predictions==test_data$Exited)
> print(paste("Accuracy:",accuracy))
[1] "Accuracy: 0.999666666666667"
```

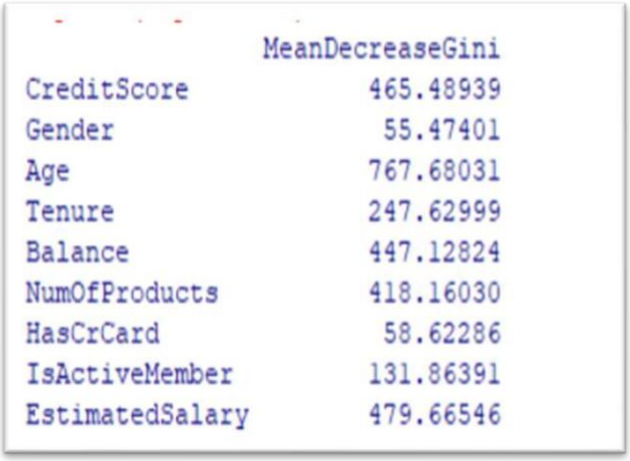
## CONCLUSION:

The accuracy rate for fitting model given by Decision tree is 99.96% of our data.

## FEATURE IMPORTANCE:

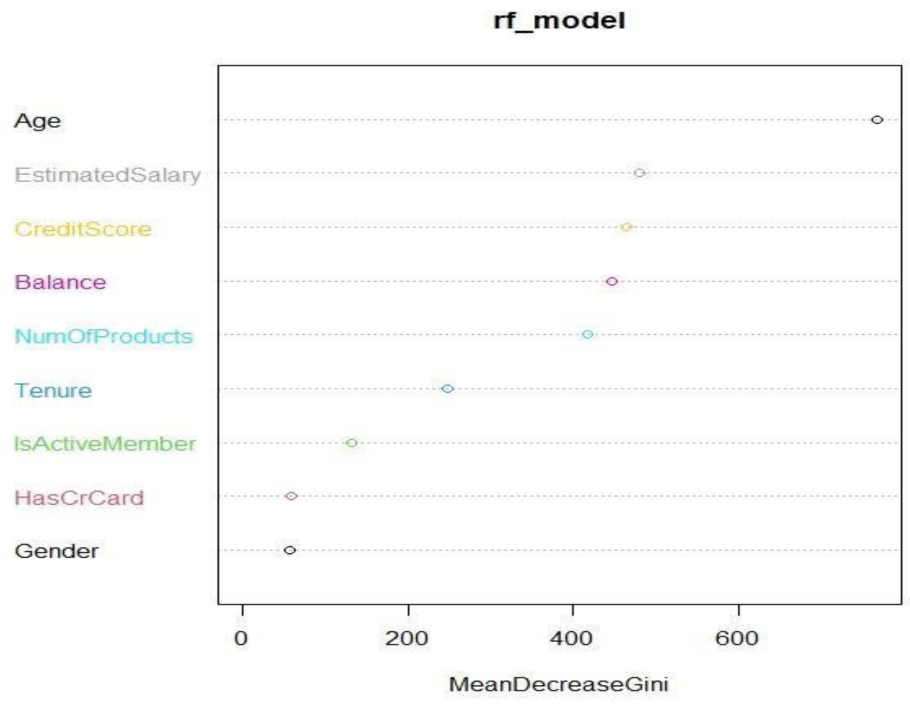
importance <- importance(rf_model) print(importance)

Credit Score, Age, Balance, and Number of Products,Estimated salary are considered the most important features for predicting customer churn, as they have higher values for MeanDecreaseGini. On the other hand, Gender, Has Credit Card and Is Active Member have lower importance values, indicating that they are less influential in predicting churn in this particular model.

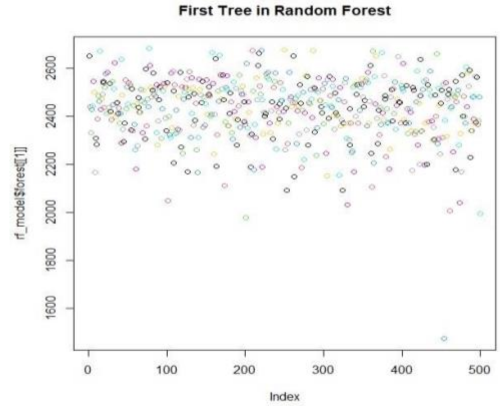


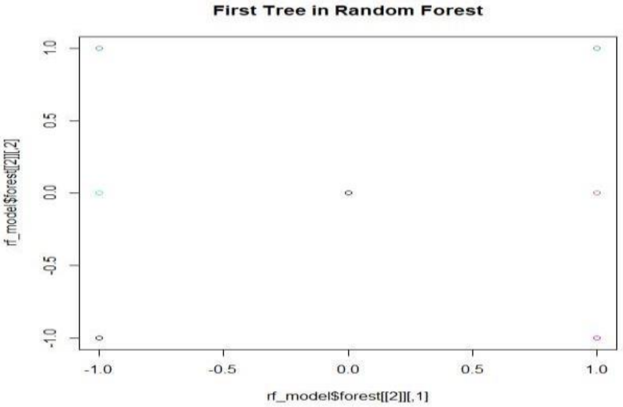
```
                MeanDecreaseGini
CreditScore            465.48939
Gender                  55.47401
Age                    767.68031
Tenure                 247.62999
Balance                447.12824
NumOfProducts          418.16030
HasCrCard               58.62286
IsActiveMember         131.86391
EstimatedSalary        479.66546
```

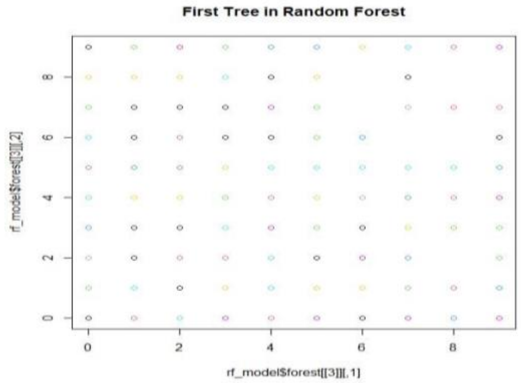
## Representing feature importance visually:

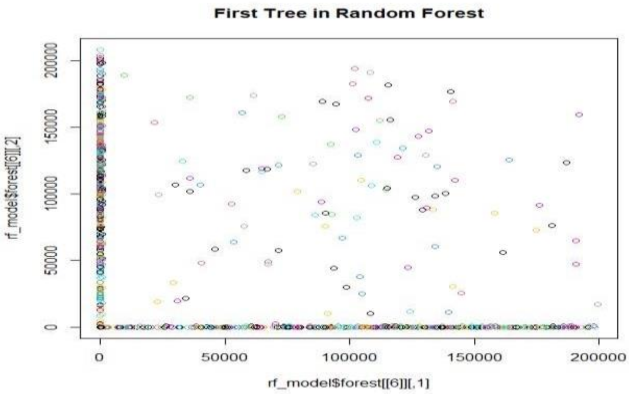


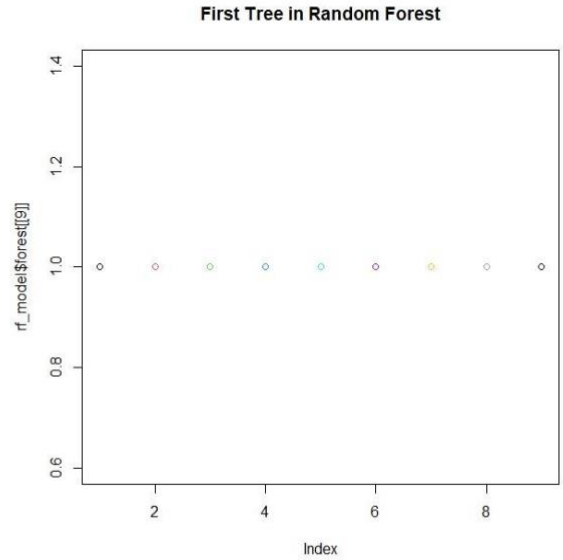
## Plotting individual random forest tree:











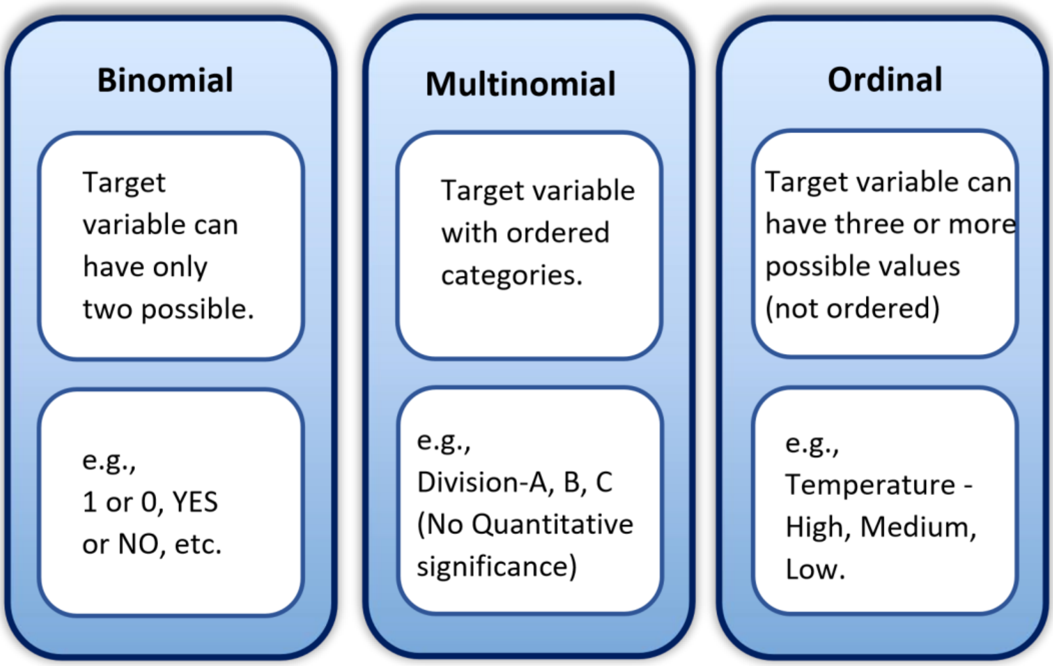
## LOGISTIC REGRESSION MODEL:

What is the logistic model?

It is a statistical method that is used to predict a ‘binary output such as Yes or No (in our case 1 or 0). The logistic regression model predicts the dependent variable of data using independent regressors.

It is a supervised classification algorithm used in classification problems. As in linear regression, it is assumed that the data follows linear function similarly logistic model builds a regression model to predict the probability that a given data entry belongs to a Category numbered “1” OR “0”.

Types of Logistic Regression:



Binomial

Target variable can have only two possible.

e.g., 1 or 0, YES or NO, etc.

Multinomial

Target variable with ordered categories.

e.g., Division-A, B, C (No Quantitative significance)

Ordinal

Target variable can have three or more possible values (not ordered)

e.g., Temperature - High, Medium, Low.

**Why this model?**

As in our data, response variable is in the form of binary type and also there is no collinearity between the regressors (Using heatmap we can observe), hence we have used this model for testing quality of water i.e., whether it is potable or not.

**Model of Logistic Regression:**

1. Y = E(Y|x) + ε

2. Y = Π(x) + ε

Where, ε is Bernoulli random variable with

a. E(ε) = 0

b. Var(ε) = π(x)(1-π(x))

$\pi(x)$=1+*eβe*0*β*+0β1+β1 X1 X1+β2+β2 X2 X2+β3+β3 X3 X3+β4+β4 X4 X4+β5+β5 X5 X5+β6+β6 X6 X6+β7+β7 X7 X7+β8+β8 X8 X8+β9+β9 X9 X9

Where β1, β2, β3, β4, β5, β6, β7, β8, β9 are regression coefficients and variables are

βO Exited β1 CreditScore β2

Gender β3 Age

β4 Tenure β5

Balance

β6 Number Of Products β7

Has Credit Card β8

Is Active Member

Logistic model considers probability using which we are going to allocate new observation to specify class. For this purpose, the threshold probability is decided and by default it is consider as **P=0.5**

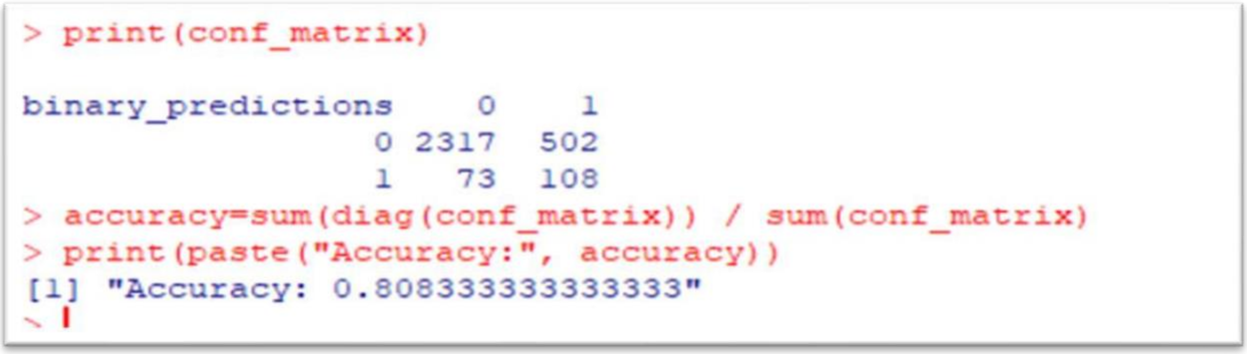
## Split the dataset into train and test data:

train_indices=sample(1:nrow(data), 0.7 * nrow(data)) train_data=data[train_indices,

] test_data=data[-

train_indices, ]

## Confusion matrix and accuracy:



```
> print(conf_matrix)

binary_predictions    0    1
                 0 2317  502
                 1   73  108
> accuracy=sum(diag(conf_matrix)) / sum(conf_matrix)
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.808333333333333"
>
```

**Conclusion:**

The accuracy rate for fitting the model by logistic regression model is **80.83%.**

## OVERALL CONCLUSION:

* ACCURACY COMPARISON*

| | |
|---|---|
| Decision Tree | 71.92% |
| Random Forest | 99.96% |
| Logistic Regression model | 80.83% |

Of the three models that were fitted to this data, the Random forest model proved to be the best fit with accuracy 99.96%. With this accuracy, it concludes that our data is correct fitted.

## REFERENCES:

### Books

1) T.Y.B.Sc statistical computing using R software by Vishwas R. Pawgi.

### Links

1) https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction
2) https://www.slideshare.net/SOUMITKAR/customer-churn-analysis-andprediction245216850
3) https://www.ibm.com/topics/machine-learning