CUSTOMER CHURN

- A Forecast Analysis for Telecom Industry with ACTIONABLE INSIGHTS

Pradeep Panicker 4/5/20 PREDICTIVE MODELLING

TABLE OF CONTENTS

PROJECT OBJECTIVE

DATA INTERPRETATION -

A DATA-SET CONTAINING 'CUSTOMER USAGE & BEHAVIOR PATTERN' – FOR POSTPAID CUSTOMERS HAVING CONTRACTUAL AGREEMENT WITH A TELECOM COMPANY IS COLLECTED FOR EVALUATION.

THE DATA IS COMPILED FOR THREE THOUSAND THREE HUNDRED THIRTY-THREE CUSTOMERS.

PROJECT OBJECTIVE –

- TO PERFORM EXPLORATORY DATA ANALYSIS.
- UNIVARIATE AND MULTI-VARIATE ANALYSIS.
- CHECK MULTICOLLINEARITY & SUMMARISE INSIGHTS OF 'EDA'
- BUILD MODELS ITH – 'LOGISTIC REGRESSION', 'KNN', 'NAÏVE BAYES'
- CREATE MODEL PERFORMANCE PARAMETER CHECK
- GENERATE ACTIONALBE INSIGHTS WITH RECOMMENDATION FOR BEST MODEL – TO PREDICT WHETHER A CUSTOMER WILL CANCEL HIS/HER SERVICE IN FUTURE OR NOT??

WORKING ENVIORNMENT SETUP

I. SOFTWARE TOOL ‘R STUDIO’ IS ESTABLISHED.

II. WORKING DIRECTORY IS SET TO ACCESS WORKING FILE – ‘Cellphone 1’.

  a. setwd (‘C:/Users/prade/OneDrive/Desktop/BABI/GREAT LAKES/PREDICTIVE MODELLING/PROJECT’)

III. Working file imported using [read_xlsx] command – to the Global Environment.

IV. The file is assigned as ‘cellphone’ in R Studio. Dimensions are verified -

  a. dim(bank) – gives a result of 3333 Elements with 11 variables have been successfully imported.

V. The Structure of Data is checked for class of each variables.

  a. ALL 11 VARIABLES are found to be in ‘NUMERIC’ observations
  b. ALL VARIABLES ARE FOUND TO BE SIGNIFICANT.
  c. str(bank) – reveals **NO** PRESENCE of Missing Values.

NOW THE DATA FILE - ‘cellphone’ - IS READY FOR DATA MANUPULATIONS AND STATISTICAL ANALYSIS.

UNIVARIATE ANALYSIS – SUMMARY.

SUMMARY of data –

The SUMMARY of Data is called for and verified for any missing values.

Command - [summary(cellone)]

NOTE – FOR SUMMARY TABLE, PLEASE REFER TO ANNEXURE -I ; TABLE-I

The FIVE Basic Statistical elements of analysis namely – Minimum Value (Min.)of the observed 3333 Values, First Inter Quartile Range (1st Qu.), Median of the 3333 Observations (Median), Average / Mean of the 3333 Observations (Mean), Third Inter Quartile Range (3rd Qu.) and Maximum Value of the observed 3333 Values – for EACH OF THE ELEVEN VARIABLES is listed.

For having better EDA of the VARIABLES Having Discreet data – 'CHURN' ; 'ContractRenewal' & 'DataPlan', the same were converted to FACTOR Variables.

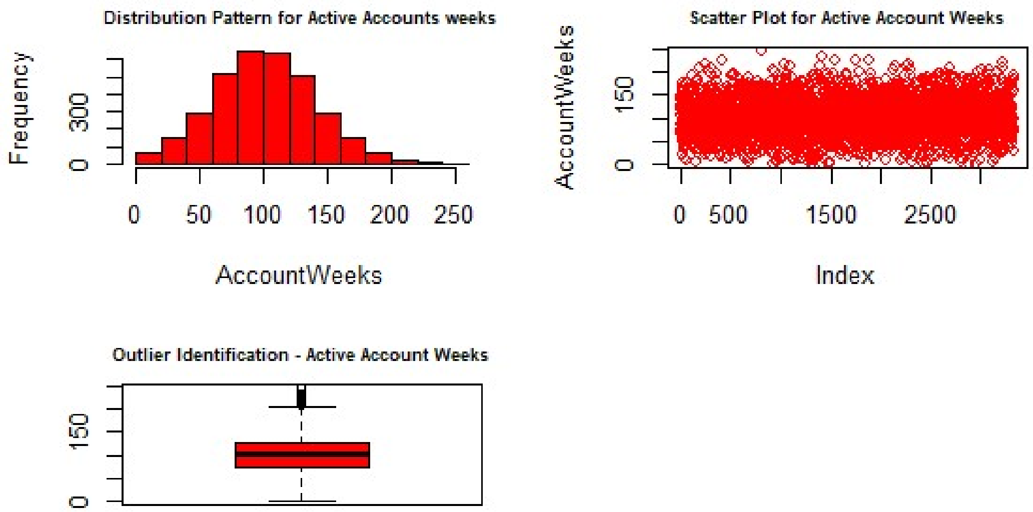
INSIGHTS -

I. **All data points are complete in nature.**
II. **VARIABLES – 'AcountWeeks'; 'DataUsage'; 'MonthlyCharge'; 'OverageFee' & 'RoamMin' seem to have outliers which needs to be treated.**

UNIVARIATE ANALYSIS – 'cellone' Data File.

ACTIVE Account Weeks [AccountWeeks] –

Graph -I



INFERENCES -

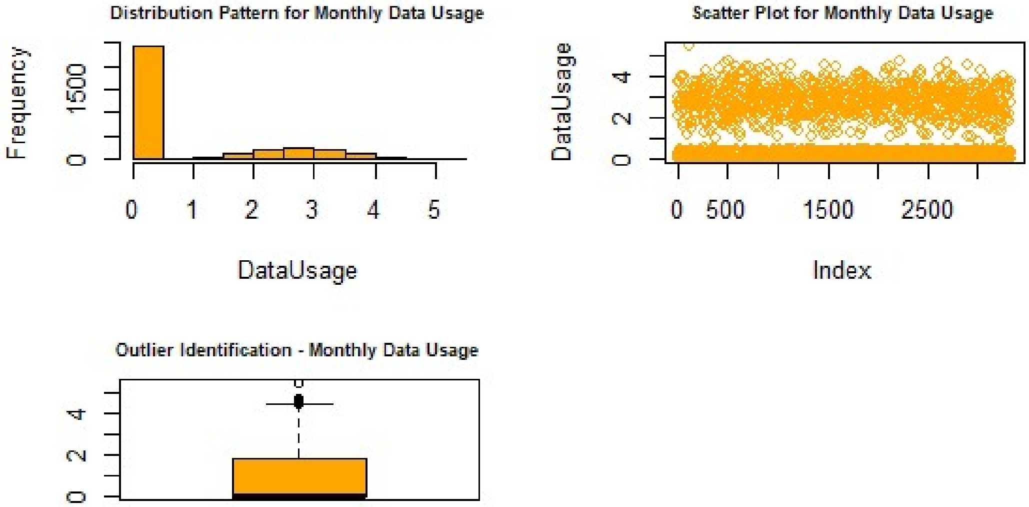
Data Distribution – Histogram [Graph I] shows data is NORMALY distributed. Majority of Customers fall into the category who have chosen to have 75 ~ 130 Weeks [1.5 ~ 2.5 YEARS] of engagement.
Need to see whether any Beneficial Scheme at the Account Opening time – made them stick to the account validity.

Scatter Plot indicates NO LINEAR Relationship.

BOX PLOT Suggests no skewness in the data. OUTLIERS HAVE BEEN IDENTIFIED TOWARDS THE MAXIMUM DATA RANGE.

Monthly Data Usage (in gigabytes)[DataUsage] –

Graph-2



INFERENCES -

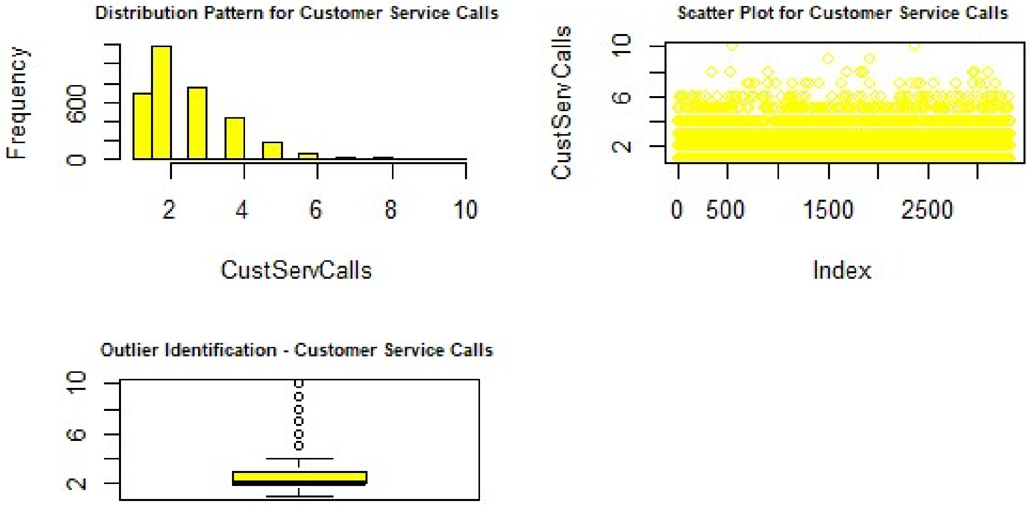
Data Distribution – Histogram [Graph 2] shows data is UNEVENLY distributed. MAJORITY of Customers DO NOT USE DATA PACKAGE AT ALL.
The very less category of customers who use data pack – has an average usage pegged at 3Gigabyte.
This indicates the presence of alternate DATA NETWORK in the area – such as Free Wi-Fi.
The category who still uses Data Pack – should have specific needs – such as high-speed, secure Data Transfer etc…

Scatter Plot - indicates NO LINEAR Relationship.
But clearly indicates TWO BANDWIDTH of usage pattern

BOX PLOT - Suggests high skewness in the data. **OUTLIERS HAVE BEEN IDENTIFIED towards the MAXIMUM Data Range.**

Customer Service Calls Made[CustServCalls] –

Graph 3-



INFERENCES -

Data Distribution –Histogram [Graph 3] shows a RIGHT SKEWEDNESS of the data. Maximum Customers do fall in the range of MAKING 02 CALLS ~ 03 CALLS.

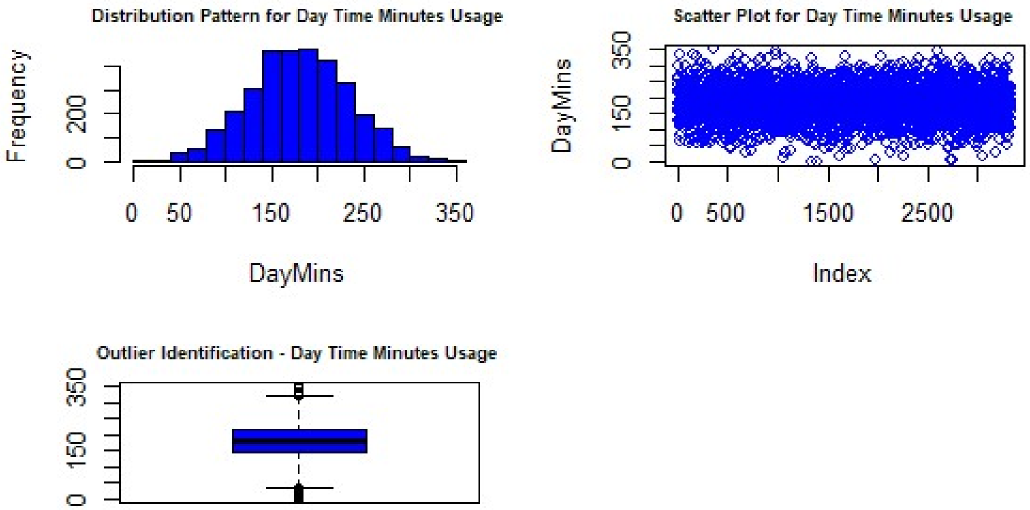
There are customers though, who demand more services / need more clarity to operate their calls – which ranks to 06 CALLS ~ 09 CALLS.

Scatter plot - suggests high density of Customers present below between the 02 CALLS ~ 03 CALLS bandwidth.

BOXPLOT shows RIGHT SKEWNESS of DATA**. Outliers are present, highly-significant at the 'Above the 1.5IQR + Q3 range'.**

DAY TIME MINUTES USAGE[DayMins] –

Graph 4 -



INFERENCES -

Data Distribution –Histogram [Graph 4] shows that data is QUITE NORMALLY DISTRIBUTED. The clustering of data pattern indicates that the maximum bandwidth of usage stays between 150Minutes ~ 250Minutes. Which is otherwise evaluated as 2.5 Hours to 4.0 Hours.

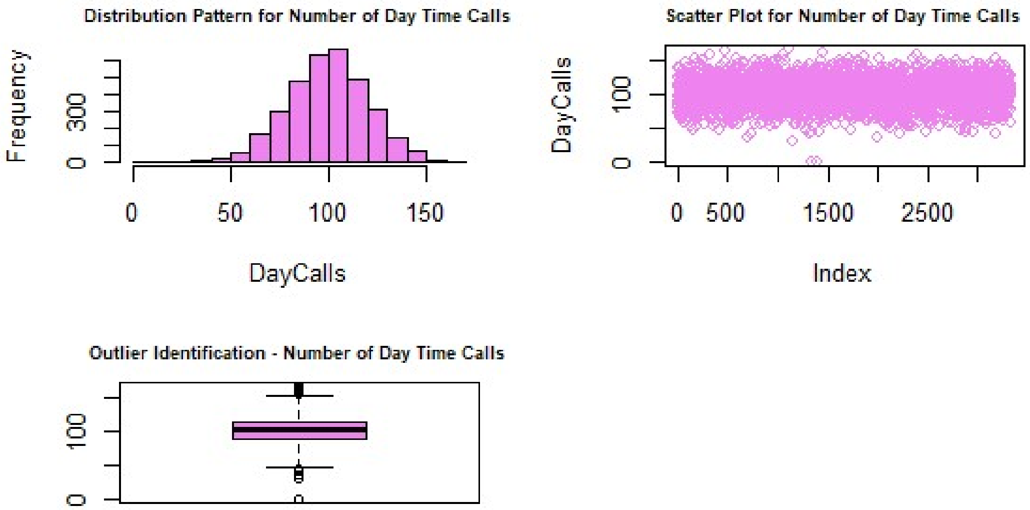
This indicates that a POSTPAID PLAN has been assigned to each customer and it so happens that the PLAN IN USAGE is widely accepted by the majority.

Scatter Plot - does not show any linear distribution pattern.

BOXPLOT indicates – OUTLIERS Present at both ends of the plot, ie. At the **'Less than 1.5IQR – Q1 range' & 'Above the 1.5IQR + Q3 range'.**

AVERAGE Number of DAY TIME Calls[DayCalls] –

Graph 5 -



INFERENCES -

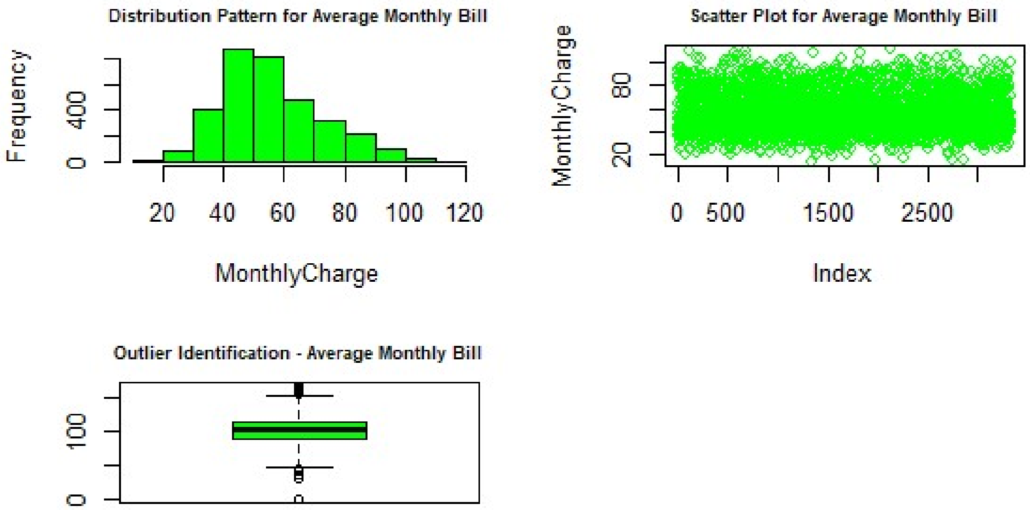
Data Distribution –Histogram[Graph 5] shows that data is NORMALLY Distributed.
An average of 100Calls are made by majority of the customers – per day. Exclusivities are there towards the higher side, but too less.
Daytime calls converging to an average of 100Calls per day – indicates the presence of a call restricting factor, by which the customer is bound to limit his number of calls. This IS due to the POSTPAID Plan offered by the telecom operator – which is a favorite one for the many.

Scatter plot - suggests high density of Customers @ the 100Nos Call per day bandwidth.

BOXPLOT - shows NO SKEWNESS of DATA. Outliers are present, at the 'Less than 1.5IQR – Q1 range' & 'Above the 1.5IQR + Q3 range'.

AVERAGE MONTHLY BILL[MonthlyCharge] –

Graph 6 –



INFERENCES -

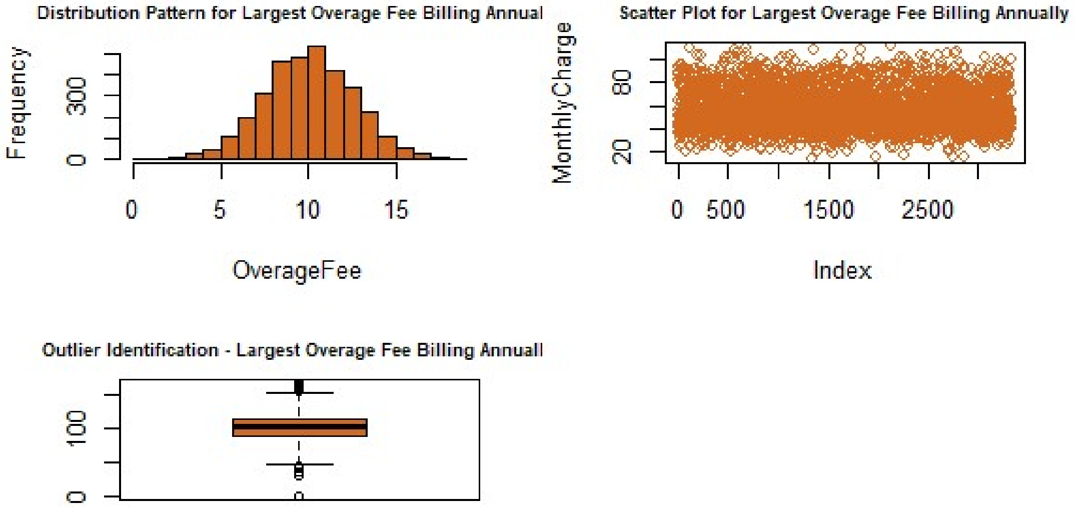
Data Distribution –Histogram[Graph 6] shows that data is NORMALLY Distributed.
Average Billing ranges from 40 Units to 60 Units for majority of the customers – per month. Exclusivities are there towards the higher side, but too less.
Billing trend shows customers are refined to control usage of CALLS, limited to the boundaries of the POSTPAID Plan offered.
Offshoot [> 60 Units of Billing] indicates additional billing than plan, indicating additional usage of CALL TIME / DATA.

Scatter Plot - suggests high density of Customers @ 40 Units to 60 Units Billing Cycle.

BOXPLOT indicates – Outliers are present , at the 'Less than 1.5IQR – Q1 range' & 'Above the 1.5IQR + Q3 range'**, indicating offshoots in billing.**

LARGEST OVERAGE FEE- 12 Months Period[OverageFee] –

Graph 7 –



INFERENCES -

Data Distribution –Histogram[Graph 7] shows that data is NORMALLY Distributed.

Average OVERAGE AGE Fees charged ranges from 07 Units to 12 Units for majority of the customers – annually. Exclusivities are there towards the higher side, but too less.

Billing analysis reveals that ONLY ONE CUSTOMER Exists WITH ZERO ‘OVERAGE FEE’ .

This means , customers use beyond the postpaid plan tariffs and are willing to pay the overage fees.

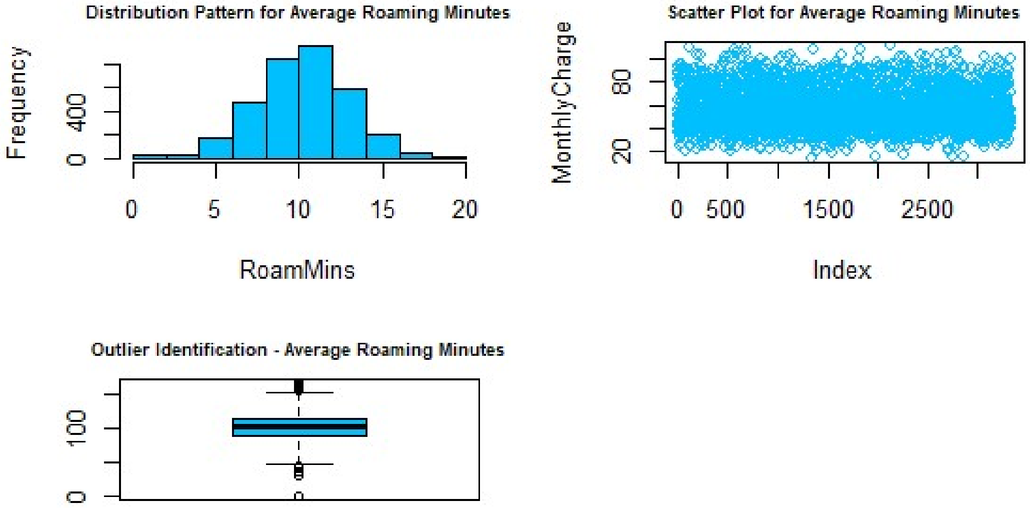
Scatter plot - suggests high density of Customers @ 07 Units to 12 Units Billing Cycle.

.

BOXPLOT - Outliers are present , at the ‘Less than 1.5IQR – Q1 range’ & ‘Above the 1.5IQR + Q3 range’**, indicating offshoots in billing.**

AVERAGE ROAMING MINUTES[RoamMins]-

Graph 8



INFERENCES -

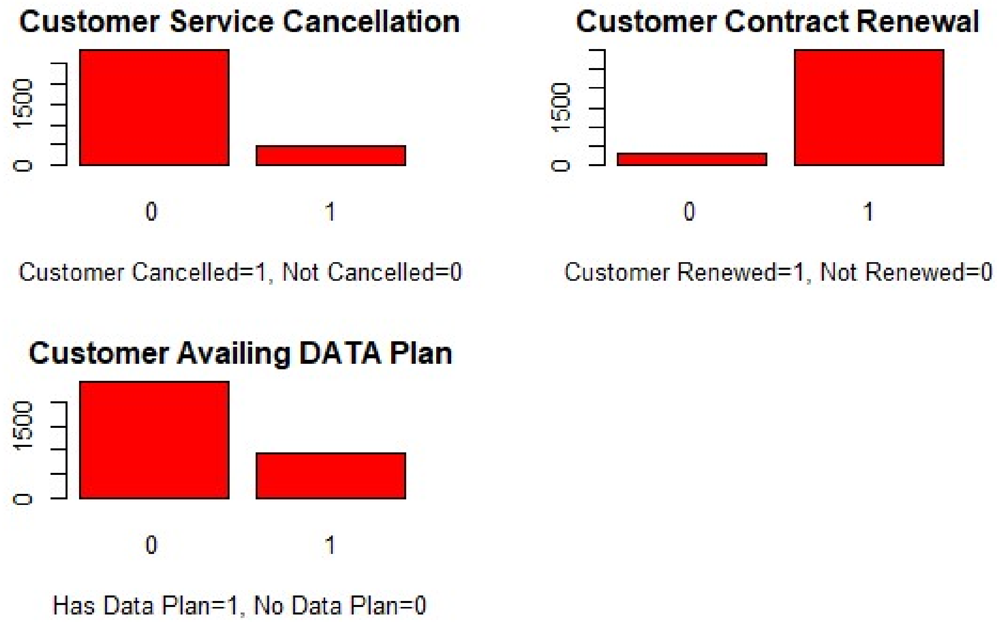
Data Distribution –Histogram[Graph 7] shows that data is NORMALLY Distributed.
Average ROAMING Minutes fall @ 10 Minutes - for majority of the customers. Exclusivities are there towards the higher side.
Billing analysis reveals that EIGHTEEN CUSTOMER Exists WITH ZERO 'ROAMING Minutes' .
It can be inferred from the above that –

a) The postpaid plan used has a limitation to provide 10Minutes of ROAMING Minutes under subsidized scheme.
   OR
b) The CUSTOMERs are largely LOCAL Based.

Scatter plot - suggests high density of Customers @ 10 Units Usage.

BOXPLOT - Outliers are present , at the 'Less than 1.5IQR – Q1 range' & 'Above the 1.5IQR + Q3 range'**, indicating offshoots in ROAMING MINUTES USAGE.**

CUSTOMER CHURN RATIO; CUSTOMER CONTRACT` RENEWAL; CUSTOMER AVAILING DATA PLAN – DISCREET DATA ANALYSIS [Churn; ContractRenewal; DataPlan]-



Inferences -

All the above listed variables having Binary Values, the above plotting depicts clearly the trend of the customers.

Customer CHURN - MAXIMUM Customers have NOT-CANCELLED the network connection with the telecom company.

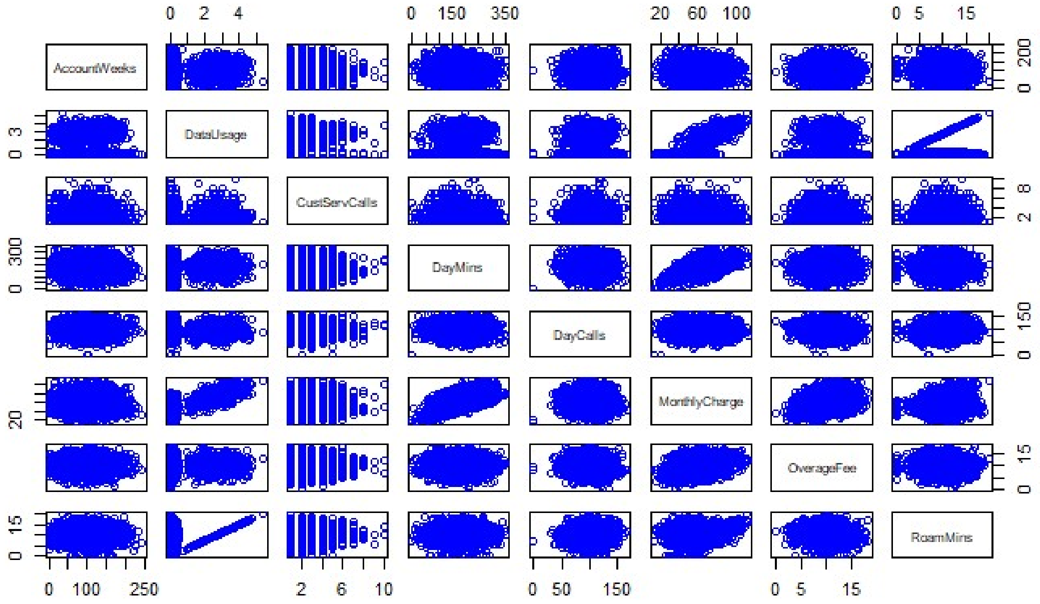
Customer CONTRACT Renewal - MAXIMUM Customers HAVE RENEWED the contract with the same company.

Customer Availing DATA PLAN – Only TWO-THIRD of the Customer base is availing the DATA Plan as provided by the telecom company. This indicates that DATA PLAN Activation is not an integrated part of the Contract.

MULTIVARIATE ANALYSIS – ‘cellphone’ Data File.

OVERALL DATA Scatter Plot Analysis ➔

Graph 9



PLOT ANALYSIS -

CORELATED VARIABLES ➔ Various plots when visually analyzed from GRAPH 9, the below listed MULTIVARIATE Analysis needs to be done for insights, since the same shows collinearity.

1) ‘Data Usage’ Vs ‘ROAMING MINUTES’.
2) ‘DAY Usage’ Vs ‘Monthly Charge’.
3) ‘DAY Minutes’ Vs. ‘Monthly Charge’.
4) ‘Monthly Charge’ Vs. ‘Data Usage’.
5) ‘Monthly Charges’ Vs. ‘Overage Fee’.
6) ‘ROAMING Minutes’ Vs. ‘Monthly Charge’.

NON-CORELATED VARIABLES ➔ The below listed combinations, NO Collinearity is found whatsoever.

1) ‘WEEKS of Active Account (AccountWeeks)’ Vs. ANY Other 7 Variables.
2) ‘Customer Service Calls’ Vs ANY Other 7 Variables.

Inferences -

1) ‘Data Usage’ Vs ‘ROAMING MINUTES’.

Collinearity stands at 16.3% with a positive linear relationship.

As the roaming minutes increases, the Data Usage also is seen increasing. *This also indicates that the Postpaid plan offered has no freebies for ROAMING MINUTES Usage.*

2) ‘DAY Usage’ Vs ‘Monthly Charge’.

Collinearity stands at 78.16% with a positive linear relationship.

The percentage of presence of DAY USAGE w.r.t. BILLING Charges is highly significant. Monthly Billing amount increases as the Day Usage increase. ***This is a logical conclusion of the postpaid contractual terms also – which indicates that the Postpaid plan does NOT HAVE A FIXED USAGE PLAN for any period of day.***

3) ‘Day Minutes’ Vs ‘Monthly Charge’

Collinearity stands at 56.78% with a positive linear relationship.

The percentage of presence of DAY Time MINUTES w.r.t. Billing Charges is quite significant. Monthly Billing amount increases as the Day Time Minutes increase. **With respect to the *postpaid contractual terms – this indicates that no special fixed usage billing SLAB provided to customers.***

4) ‘Monthly Charge’ Vs. ‘Data Usage’

Collinearity stands at 78.16% with a positive linear relationship

The percentage of presence of DATA USAGE w.r.t. Billing Charges is highly significant. Monthly Billing amount increases as the DATA USAGE increase. **With respect to the *postpaid contractual terms – this indicates that no fixed DATA PACK is made available for the customers, AT A FIXED PRICE BANDWIDTH.***

5) ‘Monthly Charges’ Vs. ‘Overage Fee’.

Collinearity stands at 28.18% with a positive linear relationship.

The percentage of presence of OVERAGE FEE w.r.t. MONTHLY Billing Charge is found.

Deep Analysis of ‘OVERAGE FEES’ –

Interesting to also note that ‘OVERAGE FEE’ has no correlation to any of the other variable such as –

MULTI VARIATE Analysis	COR. Factor
Overage Fees vs DATA USAGE	1.96%
Overage Fees vs DAY MINUTES	0.70%
Overage Fees vs DAY CALLS	-2.10%
Overage Fees vs ROAMING MINUTES	-1.10%
Overage Fees vs ACCOUNT WEEKS	-0.67%

From the above analysis, it is quite evident that OVERAGE FEES is not caused by the variations in DATA USAGE; DAY MINUTES; DAY CALLS; ROAMING MINUTES & ACCOUNT WEEKS.

This indicates that the variable by which the OVERAGE FEES is caused, IN NOT PRESENT IN THE DATA GIVEN FOR ANALYSIS.

This is to be treated as a point of concern since OVERAGE FEES can be interpreted as against the Postpaid contractual terms.

6) ‘ACCOUNT WEEKS’ vs All other Variables –

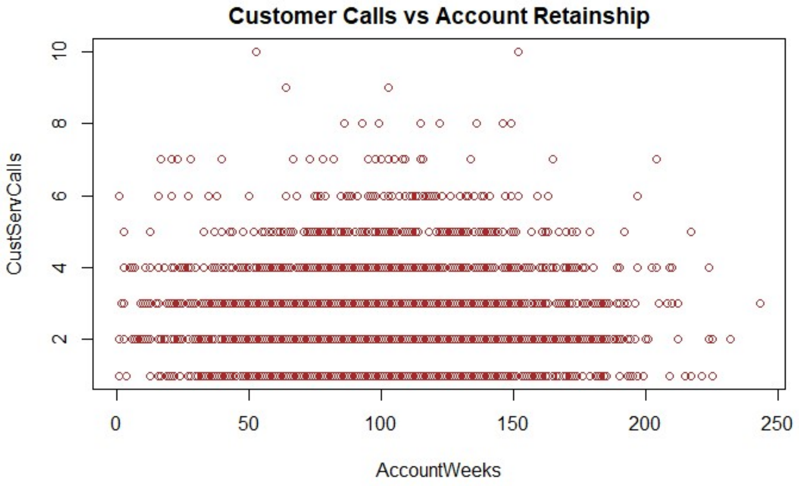
MULTI VARIATE Analysis	**COR. Factor**
ACCOUNTS WEEKS vs DATA USAGE	1.43%
ACCOUNTS WEEKS vs DAY MINUTES	0.62%
ACCOUNTS WEEKS vs DAY CALLS	3.84%
ACCOUNTS WEEKS vs ROAMING MINUTES	0.95%
ACCOUNTS WEEKS vs OVERAGE FEES	-0.67%

The above table shows a VERY WEAK Correlation of the Number of weeks a Customer held account – DUE TO THE VARIABLES ‘Data Usage’; ‘Day Minutes’; ‘Day Calls’; ‘Roaming Minutes’; ‘Overage Fees’.

Analysis between ‘CUSTOMER CALLS’ to the telecom company and the ‘ACCOUNT WEEKS’, shows evidently that Customers who have made Service Calls more than SIX have stayed less WEEKS with the company. Refer GRAPH 10 below.

This can also be treated as an off-shoot case scenario wherein the Customers who called in several times – 07 Times and above – had some special need that the telecom company could not satisfy. Need deep analysis with this chain of customers!

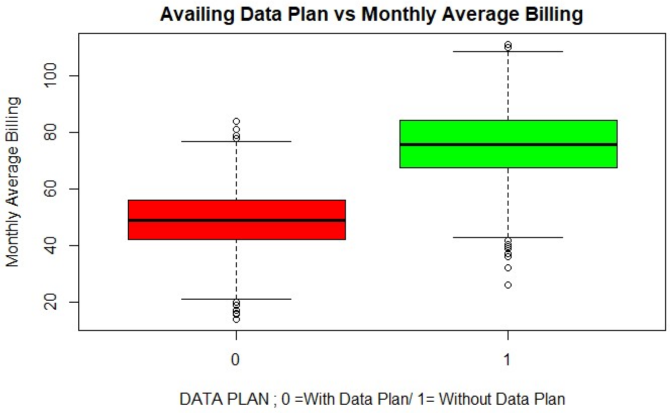
GRAPH 10 -



Also, the below tabular analysis gives insight to the fact that the CUSTOMER CALLS made by the customers had NO RELATION to the factors such as 'Data Usage'; 'Day Minutes'; 'Day Calls'; 'Roaming Minutes'; 'Overage Fees'.

MULTI VARIATE Analysis	**COR. Factor**
CUSTOMER CALLS vs DATA USAGE	-2.17%
CUSTOMER CALLS vs DAY MINUTES	-1.34%
CUSTOMER CALLS vs DAY CALLS	-1.89%
CUSTOMER CALLS vs ROAMING MINUTES	-0.96%
CUSTOMER CALLS vs OVERAGE FEES	-1.29%

7) DATA PLAN (Availed / Not Availed) vs MONTHLY BILLING –

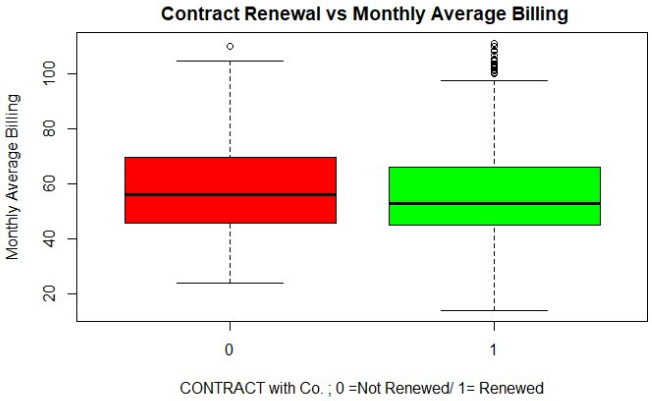


Inferences –

The above analysis indicates that customers WITHOUT DATA PLAN provides a higher MONTHLY BILLING Volume than the customers with DATA PLAN.

If not a LOCALITY Issue, the telecom company needs to investigate reviving the DATA PLAN for customer friendly packages.

8) CONTRACT RENEWAL vs MONTHLY BILLING –



Inferences –

Interesting to find that the customers who DID NOT RENEW THE CONTRACT with the telecom company WERE ACTUALLY HAVING HIGH VOLUME OF BILLING AMOUNT per month than those customers who renewed the contract.

This can be an indication that -FOR CUSTOMERS WITH HIGH USAGE OF DAY CALLS/ DAY MINUTES, THE EXISTING POSTPAID PLAN IS NOT EFFICIENT TO REATIN THE CUSTOMERS. The company really needs to implement strategies to retain this category of customers.

OUTLIERS AND MISSING VALUES – ANALYSIS

	1st Quart[A]	3rd Quart[B]	IQR	MIN Value	[A] - 1.5IQR	MAX. Value	[B] + 1.5IQR	Inference	Low Fence range	High Fence Range	DATA Count (Low Fence Range)	DATA Count (High Fence Range)
Account Week	74	127	53	1	-5.5	243	206.5	Outlier Present	NIL	206.6 to 243	0	18
Data Usage	0	1.78	1.78	0	-2.67	5.4	4.45	Outlier Present	NIL	4.45 to 5.4	0	11
Cust. Serv. Calls	1	2	1	0	-0.5	9	3.5	Outlier Present	NIL	3.6 to 9.0	0	696
DAY Call Minutes	143.7	216.4	72.7	0	34.65	350.8	325.45	Outlier Present	0 to 34.64	325.46 to 350.8	14	11
DAY Calls	87	114	27	0	46.5	165	154.5	Outlier Present	0 to 46.4	154.6 to 165	15	8
AVG. Monthly Charges	45	66.2	21.2	14	13.2	111.3	98	Outlier Present	NIL	99 to 111.3	0	34
Overage Fee	8.33	11.77	3.44	0	3.17	18.19	16.93	Outlier Present	0 to 3.16	16.94 to 18.19	14	10
Roaming Minutes	8.5	12.1	3.6	0	3.1	20	17.5	Outlier Present	0 to 3.0	17.6 to 20.0	31	14

The above tabular data was derived from the basic rules of OUTLIER Detection –

a) Low Fence region of OUTLIER = 1st Quartile – 1.5 x IQR.
b) High Fence region of OUTLIER = 3rd Quartile + 1.5 x IQR.

Inferences –

Maximum Outliers are present with the variable 'Customer Service Calls' – this being a discreet variable and not having any influence over the rest of the SIX variables, OUTLIER TREATMENT IS NOT REQUIRED. Also, OUTLIER Treatment needs to be recommended only of 'Customer Service Calls' is to be used for model building**.**

All the rest of the variables show a presence of less that 10% of Customer Base as OUTLIERS. Also, the MEAN of these variables is not pulled away from the MEDIAN by the presence of these outliers – exception can be found in 'Average Charges', which can be treated exclusively by removing those lists of customers.

MISSING VALUES → NO MISSING VALUES WERE FOUND TO BE PRESENT IN THE COMPLETE DATA.

MULTICOLLENEARITY

For the analysis of MULTICOLLENEARITY, a Linear Regression model needs to be created with the continuous variables in the data.

'Data Usage'; 'Day Minutes'; 'Day Calls'; 'Roaming Minutes'; 'Overage Fees' were used as Independent variables for analyzing its impact of the dependent variable – 'MonthlyCharge'

Analysis results as below –

The below commands were utilized to check the multi-collinearity from the Variation Inflation Factors derived –

1. vif(lm(DataUsage~ ., data = cellphone))
2. vif(lm(DayCalls~ ., data = cellphone))
3. vif(lm(DayMins~ ., data = cellphone))
4. vif(lm(OverageFee~ .,data = cellphone))
5. vif(lm(RoamMins~ ., data = cellphone))
6. vif(lm(MonthlyCharge~ ., data = cellphone))
7. vif(lm(AccountWeeks~ .,data = cellphone))

```
REFER ANNEXURE – II for the VIF Values.
```

HUGE Multicollinearity was observed and thus the variables need to be chosen wisely, for model building.

vif (lm(AccountWeeks~ Churn+ContractRenewal+CustServCalls+DayMins+DayCalls+RoamMins, data = cellphone))

The above MODEL gave VIF values which were quite satisfactory-
New dataset was created by omitting the Variables which gave high VIF Values.
cellphone2= cellphone[,-c(4,5,9,10)]

```
vif(lm(AccountWeeks~ Churn+CustServCalls+DayMins+DayCalls+RoamMins+ContractRenewal,
 data = cellphone2))
          Churn   CustServCalls         DayMins        DayCalls        RoamMins 
ContractRenewal 
       1.184796        1.057748        1.048353        1.001295        1.007109        1
.080818
```

SUMMARISATION OF INSIGHTS FROM EDA

Inferences from Multivariate Analysis –

1. The postpaid plan offered to the customer **has NO Freebies** w.r.t. DATA Usage, DAY TIME Calls duration, Data Pack & Roaming Minutes.
2. The **MONTHLY CHARGES** billed has a direct positive relation with DATA Usage, DAY TIME Calls duration, Data Pack & Roaming Minutes. 99.9% of variations in the MONTHLY CHARGES is explained by these four factors. Significant among those are DAY MINUTES, DATA USAGE & OVERAGE FEES.
3. Interesting to note that **OVERAGE FEES**, even though it has a positive relation over BILLING, OVERAGE FEES shows no correlation with any other variables such as - DATA Usage, DAY TIME Calls duration, Data Pack & Roaming Minutes. OVERAGE Fees is usually charged against a billing factor which is excessively used over the contractual terms. **If no correlation with the existing factors such as -DATA Usage, DAY TIME Calls duration, Data Pack & Roaming Minutes – means the FACTOR LIABLE FOR OVERAGE FEES IS NOT LISTED IN DATA.**
4. **SATISFACTION level** of customers is shown with the analysis that LESS CUSTOMER SERVICE CALLS for customers who have maintained MORE PERIOD of ACCOUNT WEEKS with the telecom company**.**
5. It is also inferred that the existing postpaid plan DOES NOT HAVE SUFFECIENT SATISFACTORY PARAMETERS TO HOLD A RANGE OF CUSTOMERS WHO USE HIGH VOLUME OF DAY CALLS / DAY MINUTES. The company needs to re-look into the postpaid plan and renew the same to retain this category of customers.

Inferences from OUTLIERS analysis –

1. OUTLIERS are present in all the variables, but the percentage of presence w.r.t. to its correlation is insignificant. Hence NO OUTLIER Treatment is required for modelling purpose**.**

Inferences from MULTI-COLLENEARITY –

1. Upon building various Linear regression models with inter-changing dependent variable, the multicollinearity factor was found very significant. A new dataset was created with selective variables as independent variables – by which VIF Values were well within the allowable range to NULLIFY Multi-collinearity.

LOGISTICS REGRESSION

First, the basic rules validation of the dataset for Logistics Regression needs to be done –

1) The target variable should be in Binary form –
    a. The target variable of ‘cellphone’ is “churn” and is found to be classified in ‘1’ & ‘0’. Also ‘1’ represents the condition of predictable output by the model to be built.
2) Multicollinearity between the predictor variables –

    While analysis with all possible combinations created through multiple Linear Regression model, HIGH Multicollinearity was found within the Independent Variables.

    REFER ANNEXURE – II for the VIF Factors derived.

    Cautious interpretations could sort and select possible list of independent variables w.r.t. the Target Variable – ‘Churn’.

```
Churn  ContractRenewal     CustServCalls              DayMins             DayCalls
RoamMins
        1.184796        1.080818        1.057748        1.048353        1.001295
1.007109
```

3) The independent variables and linearly related to the log-odds –
    a. Not checked since the Logistics Regression will take of this abnormality (if any) by itself.
4) Logistics Regression requires a large number of observations.
    a. 3333 Rows ÷ 11 Column = 303 Data per column is good enough of volume of data for Logistic Regression [Data below the 150 Data per column will reduce accuracy of the LOGIT MODEL].

Alternate Data set was created after removing the variables with high multicollinearity – ‘cellphone2’.

The date was split into ‘traindataset2’ & ‘testdataset2’.

A model was generated, on the ‘traindataset2’ with ‘Generalized Linear Model’ command and parameter as ‘logit’ –

model2 = glm(Churn~.,data= trainingdata2, family=binomial(link = 'logit'))

AIC Value was 1602

The McFadden value came to be 17.3%

‘AccountWeeks’ & ‘DayCalls’ were identified as INSIGNIFICANT Variables to the model.

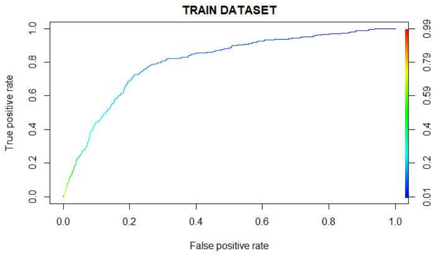
Hence model was regenerated, removing ‘AccountWeeks’ & ‘DayCalls’.

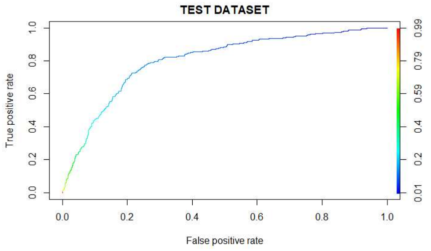
model3= glm(Churn~ ContractRenewal+CustServCalls+DayMins+RoamMins,data= trainingdata2, family=binomial(link = 'logit')

AIC VALUE REDUCED TO 1599.7

Performance analysis of this MODEL on the TRAIN Data Set & TEST Data Set provided results as below –

	train - Logit	test - Logit
sensitivity	0.1432	0.1612
specificity	0.9719	0.9835
accuracy	0.8529	0.862
AUC	0.802	0.8007
Optimum Cuttoff	0.5	0.5





The sensitivity is evaluated as LOWER THAN specificity since – WE ARE MORE CONCERENED OF THE 'ONES[1]' THAN THE 'ZEROS[0]' OF THE TARGET VARIABLE 'Churn' – since 'ONES' representing the Customers who have LEFT / CHURNED away from the telecom company

The model is robust since it has performed well with the test data.

K NEAREST NEIGHBOURS [Knn]

For the ‘Knn’ method of model building for Classification purpose, we need to clean the data set for achieving optimum accuracy and increasing the model performance parameters.

Approach methods –

1) Outlier Treatment
2) NORMALISATION of DATA
3) SPLITING of data set TRAIN + TEST
4) MODEL Building.

Outlier Treatment –

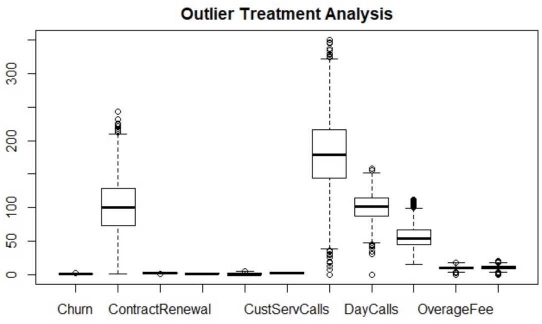
Refer page 19 of this report.

It is found that ‘CustServCalls’ & ‘MonthlyCharges’ have significant outliers.

We will now nullify these outliers by removing the respective rows containing the uppermost outlier values.

```
cellphoneknn= cellphone[cellphone$CustServCalls<3.6,]
```

```
boxplot(cellphoneknn, main= 'Outlier Treatment Analysis')
```



Further treatment will be decided upon the model performance.

Normalization of Dataset –

Prior to normalization, the target variable is removed and kept safe – creating a new data subset for normalization.

```
cellphoneknnscale= cellphoneknn[,-c(1,3,4)]
cellphoneknnscale= scale(cellphoneknnscale, center = TRUE, scale = TRUE)
```

After scaling down the complete dataset, the target variable, in its original form, is joined back to the scaled dataset.

```
cellphoneknnscaled= cbind(cellphoneknnscale,cellphoneknn$Churn)
names(cellphoneknnscaled)[9]= 'Churn'
```

SPLITTING Data -

Data is split into train dataset & test dataset @ the ratio of 70:30.
Proportion check is done w.r.t. to the Target Variable.

```
prop.table(table(cellphoneknnscaled$Churn))
        0         1 
0.8858551 0.1141449 

> prop.table(table(trainknn$Churn))
        0         1 
0.8856988 0.1143012 

> prop.table(table(testknn$Churn))
      0       1 
0.88622 0.11378 
```

The dataset 'cellphoneknnscaled' is now ready for model building.

MODEL Building & Performance measures -

The kNN model is built using commands from the library(class).

The command is such that the predicted values of the Target variable, in the test dataset, is got directly.

predtest= knn(trainknn[-1],testknn[-1],trainknn[,9],k=19)

We now need to optimize the Accuracy of the model, by running the model for different K values. When done so, the following results were observed –

K value	Specificity	Sensitivity	Accuracy %	Remarks
19	0.9986	0.2556	0.9140	MAX. Preferred K Value
10	0.9943	0.3111	0.9166	
15	0.9957	0.3000	0.9166	
5	0.9886	0.3222	0.9126	
3	0.9772	0.3667	0.9077	
2	0.9330	0.3778	0.8698	

The OPTIMUM ACCURACY – WITH THE BEST SENSITIVITY FOR THE 'ONES[1]' THAT CAN BE ACHIEVED – IS A MODEL BUILT WITH 'K' VALUE AT '15'.

NAÏVE BAYES – MODEL BUILDING

Approach Method –

The primary and most import condition to build a NAÏVE BAYES Model is **–**

All the Independent variables should be in categorical format.

The secondary condition is that the NAÏVE BAYES Model to succeed in performance, a LARGE VOLUME of DATA is required for analysis – such as dataset with 10,000 Rows and above.

Our dataset – cellphone, contains THREE Variables (One TARGET + TWO Independent Variables) in Factor Format with binary values and rest of the EIGHT Variables in Numeric formats. **NAÏVE BAYES cannot be applied to datasets with numeric variables.**

Thereby, we need to subset a dataset with Only THREE Variables – 'Churn', Customer Renewal' & 'Data Plan'.

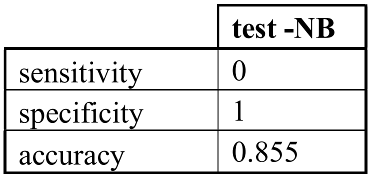
'cellonenb' dataset is created with THREE Columns – 01 Target variable & Two Independent Variables.

The ZEROs & ONEs are suitable converted to 'Nos' & YES' for Naïve Bayes analysis.

The target variable is retained in Factor Format.

The other two variables are now in Categorical Format.

Model is built and the following model performances are noticed **–**



	test -NB
sensitivity	0
specificity	1
accuracy	0.855

MODEL VALIDATION EXERCISE

LOGISTICS REGRESSION -

	TRAIN Data	TEST Data
sensitivity	0.1432	0.1612
specificity	0.9719	0.9835
accuracy	0.8529	0.862
AUC	0.802	0.8007

kNN CLASSIFICATION -

K value	Specificity	Sensitivity	Accuracy %	Remarks
19	0.9986	0.2556	0.9140	MAX. Preferred K Value
10	0.9943	0.3111	0.9166	
15	0.9957	0.3000	0.9166	
5	0.9886	0.3222	0.9126	
3	0.9772	0.3667	0.9077	
2	0.9330	0.3778	0.8698	

NAIVE BAYES CLASSIFICATION

	test -NB
sensitivity	0
specificity	1
accuracy	0.855

The problem statement we have is to classify the Customers from the NEW DATA as to whether they will CHURN OUT from the company.

'kNN' method of classification stands out as the best model that can be used for this given scenario – this conclusion arrived from reviewing the above Model performance measures.

ACTIONABLE INSIGHTS

We will now attempt to prepare actionable insights from the Logistics Regression Coefficients.

Below tabular information is an output derived from the Logit model built –

	Intercept [a]	Exponential Coeff.[b]	P / 1-P[c] = 1(+/-) b	Probability of Y(Churn), if X=1
Contract Renewal YES (X)	0.007273	0.16808	0.83192	0.5459
Customer Serv. Calls (X)	0.007273	1.6489	0.6489	0.6065
DAY MINUTES (X)	0.007273	1.0124	0.0124	0.9878
ROAM MINUTES (X)	0.007273	1.0873	0.0873	0.9197

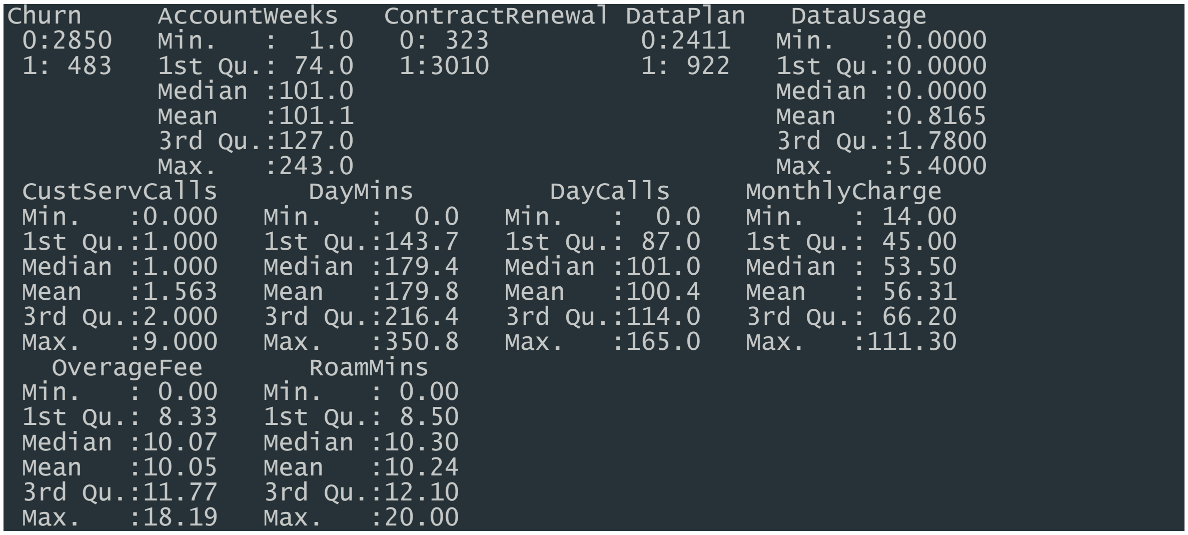
The above variables were identified as the most significant ones that contribute to the CHURN-OUT of the customers. Thereby, interpretations as follows –

1) For a customer who has RENEWED THE ACCOUNT with the telecom company, 54.59% Chances are there that the customer will leave the company, under the given circumstances.
2) For an existing customer, CUSTOMER SERVICE CALLS RESPONSE contributes to 60.65% chances that the customer will leave the company, under the given circumstances.
3) For an existing customer, 98.78% chances exist that HE/SHE will leave the company due to the DAY MINUTES contribution in the postpaid plan given by the company.
4) For an existing customer. 91.97% chances exist that HE/SHE will leave the company due to the ROAMING MINUTES contribution in the postpaid plan given by the company.

The above calls for a review of the existing postpaid plan that the company has rolled out to the customers – ON THE ABOVE LISTED FOUR PARAMETERS.

ANNEXURE – I

SUMMARY -



```
Churn     AccountWeeks   ContractRenewal DataPlan   DataUsage
 0:2850   Min.   :  1.0   0: 323          0:2411   Min.   :0.0000
 1: 483   1st Qu.: 74.0   1:3010          1: 922   1st Qu.:0.0000
          Median :101.0                            Median :0.0000
          Mean   :101.1                            Mean   :0.8165
          3rd Qu.:127.0                            3rd Qu.:1.7800
          Max.   :243.0                            Max.   :5.4000
 CustServCalls      DayMins         DayCalls     MonthlyCharge
 Min.   :0.000   Min.   :  0.0   Min.   :  0.0   Min.   : 14.00
 1st Qu.:1.000   1st Qu.:143.7   1st Qu.: 87.0   1st Qu.: 45.00
 Median :1.000   Median :179.4   Median :101.0   Median : 53.50
 Mean   :1.563   Mean   :179.8   Mean   :100.4   Mean   : 56.31
 3rd Qu.:2.000   3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.: 66.20
 Max.   :9.000   Max.   :350.8   Max.   :165.0   Max.   :111.30
   OverageFee       RoamMins
 Min.   : 0.00   Min.   : 0.00
 1st Qu.: 8.33   1st Qu.: 8.50
 Median :10.07   Median :10.30
 Mean   :10.05   Mean   :10.24
 3rd Qu.:11.77   3rd Qu.:12.10
 Max.   :18.19   Max.   :20.00
```

ANNEXURE - II

VIF VALUES FOR Multicolleniarity –

```
vif(lm(DataUsage~ ., data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan   CustServCal
ls         DayMins        DayCalls 
       1.211655        1.003513        1.083083       12.364215        1.0592
57        7.901265        1.003385 
  MonthlyCharge      OverageFee        RoamMins 
      21.156834        2.482759        1.349063 
> 
> vif(lm(DayCalls~ ., data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan       DataUsa
ge   CustServCalls         DayMins 
       1.211110        1.002465        1.083885       12.476811     1964.7981
90        1.058835     1031.515505 
  MonthlyCharge      OverageFee        RoamMins 
    3243.288106      224.643231        1.351938 
> 
> vif(lm(DayMins~ ., data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan       DataUsa
ge   CustServCalls        DayCalls 
       1.211627        1.003617        1.083202       12.473226       15.0501
14        1.059289        1.003387 
  MonthlyCharge      OverageFee        RoamMins 
       3.300405        1.224746        1.352348 
> 
> vif(lm(OverageFee~ .,data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan       DataUsa
ge   CustServCalls         DayMins 
       1.211626        1.003637        1.083219       12.472866       21.7148
05        1.059301        5.623725 
       DayCalls   MonthlyCharge        RoamMins 
       1.003377       14.612091        1.352350 
> 
> vif(lm(RoamMins~  ., data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan       DataUsa
ge   CustServCalls         DayMins 
       1.206419        1.003761        1.083766        9.571448     1959.9134
89        1.059105     1031.453876 
       DayCalls   MonthlyCharge      OverageFee 
       1.003024     3243.032319      224.632439 
> 
> vif(lm(MonthlyCharge~ ., data = cellphone))
          Churn    AccountWeeks ContractRenewal        DataPlan       DataUsa
ge   CustServCalls         DayMins 
       1.211661        1.003619        1.083184       12.472980       12.8169
28        1.059241        1.049681 
       DayCalls      OverageFee        RoamMins 
       1.003387        1.012102        1.352322 
> 
> vif(lm(AccountWeeks~ .,data = cellphone))
          Churn ContractRenewal        DataPlan       DataUsage   CustServCal
ls         DayMins        DayCalls 
       1.211515        1.083508       12.463066     1964.030679        1.0593
93     1031.216812        1.001945 
  MonthlyCharge      OverageFee        RoamMins 
    3242.354928      224.584936        1.352231 
```

Thank You!