Project: Customer Churn

Table of Content

1. Program Objective

The objective of this project is to identify the customer churn for telecom companies. In this project report the exploratory data analysis is done in order to indentify the central tendency of the data and to check the multi co-linearity among different variables. Different models were build viz. logistic regression, Naive Bayes and KNN (K nearest neighbors) to predict whether the customer will continue or discontinue the services provided by the telecom company. From the results obtained after training and testing the build models, the best model is interpreted on the basis of various model performance measurement parameters such as sensitivity, accuracy, specificity, AUC, GINI etc.

2. Exploratory Data Analysis

2.1 Basic Data Summary

The customer data from the telecom company is available in *Cellphone.xls* file. Data given relates the characteristics or behavior of a customer towards the services offered by the telecom company. It contains various attributes like AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee, RoamMins.

Attributes namely ContractRenewal and DataPlan contains binary data (or categorical data with two levels (1’s and 0’s) where ‘1’ represents active and ‘0’ represents Non active). Other attributes has continuous data (Table-1).

Table-1: Basic data summary

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     3333 obs. of  11 variables:
 $ Churn          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AccountWeeks   : num  128 107 137 84 75 118 121 147 117 141 ...
 $ ContractRenewal: num  1 1 1 0 0 0 1 0 1 0 ...
 $ DataPlan       : num  1 1 0 0 0 0 1 0 0 1 ...
 $ DataUsage      : num  2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...
 $ CustServCalls  : num  1 1 0 2 3 0 3 0 1 0 ...
 $ DayMins        : num  265 162 243 299 167 ...
 $ DayCalls       : num  110 123 114 71 113 98 88 79 97 84 ...
 $ MonthlyCharge  : num  89 82 52 57 41 57 87.3 36 63.9 93.2 ...
 $ OverageFee     : num  9.87 9.78 6.06 3.1 7.42 ...
 $ RoamMins       : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
```

2.2 Bivariate Analysis

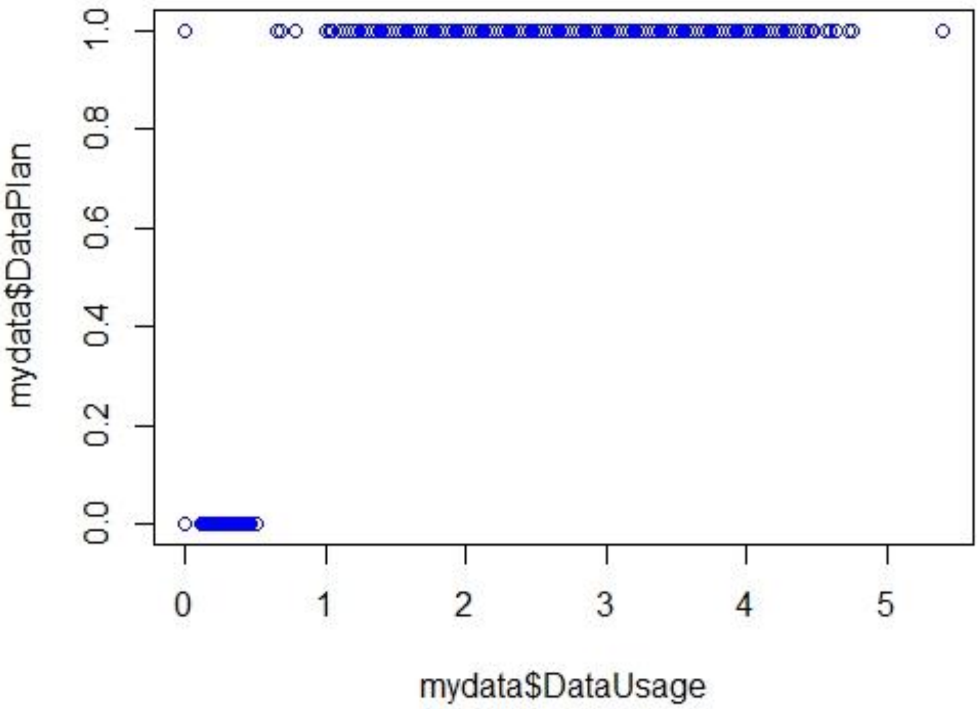
The data represents the Bivariate Analysis as the dependent variable (i.e. Churn) depends upon two or more variables.

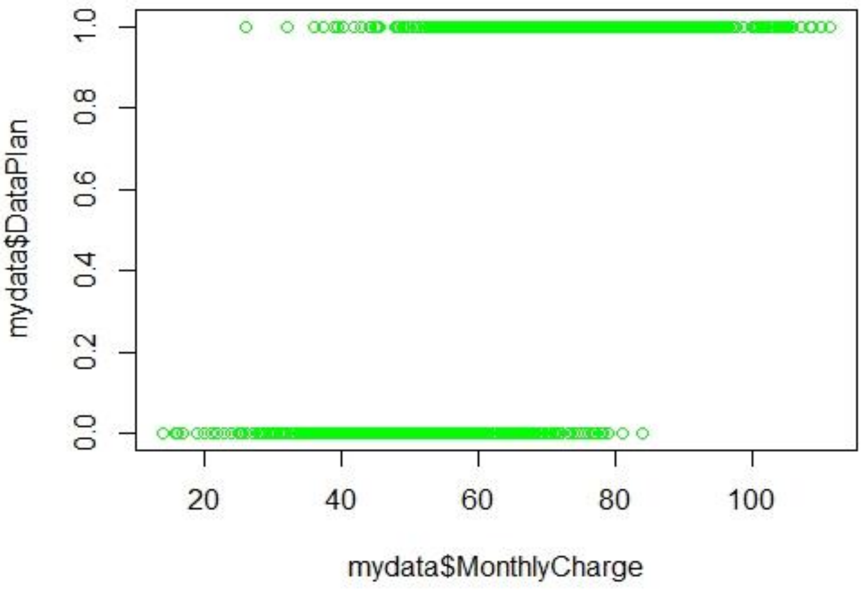
2.3 Summary of the data set

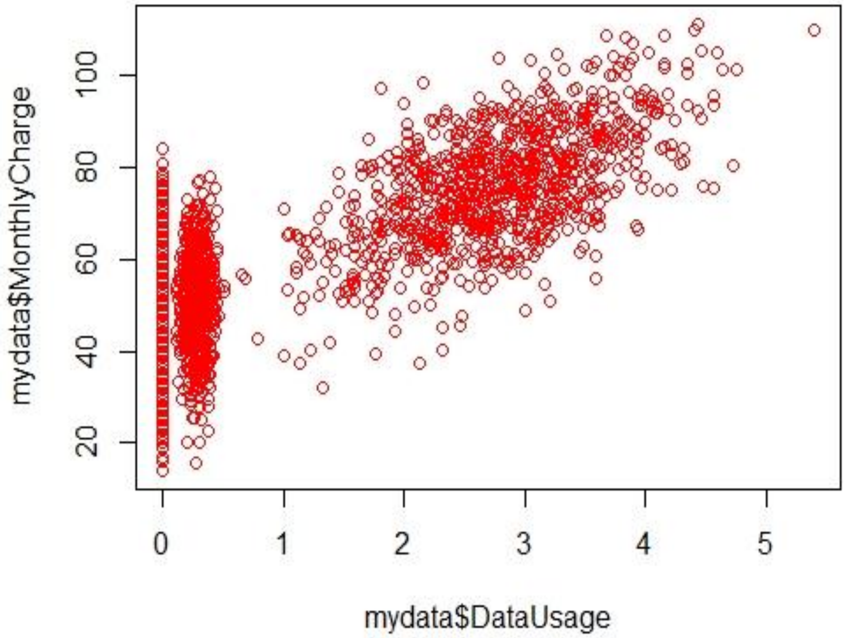
Table-2: Summary

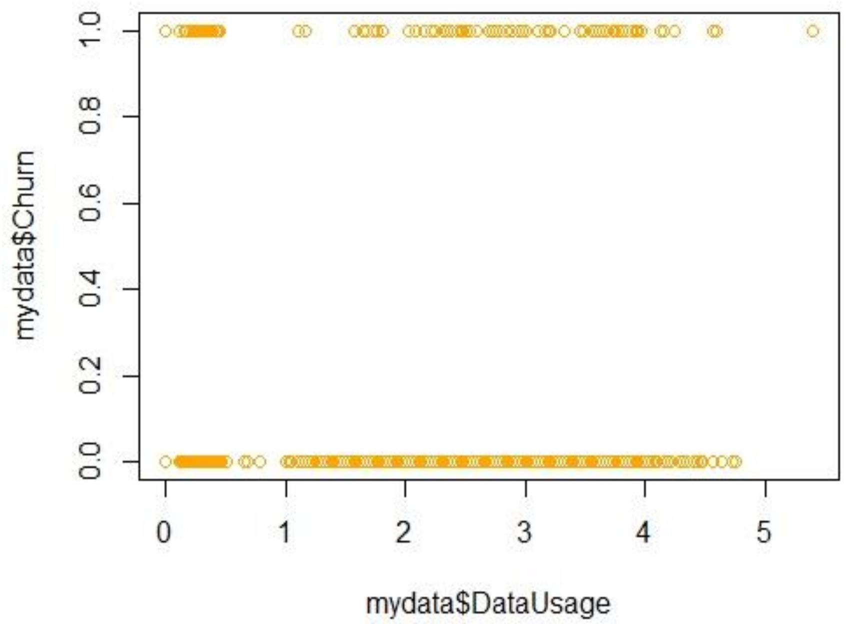
	Churn	Account Weeks	Contract Renewal	Data Plan	Data Usage	Cust Serv Calls	Day Mins	Day Calls	Monthly Charge	Overage Fee	Roam Mins
Min	0.0000	1.0	0.0000	0.0000	0.0000	0.0000	0.0	0.0	14.00	0.00	0.00
1st Qu	0.0000	74.0	1.0000	0.0000	0.0000	1.0000	143.7	87.0	45.00	8.33	8.50
Median	0.0000	101.0	1.0000	0.0000	0.0000	1.0000	179.4	101.0	53.50	10.07	10.30
Mean	0.1449	101.1	0.9031	0.2766	0.8165	1.563	179.8	100.4	56.31	10.05	10.24
3rd Qu	0.0000	127.0	1.0000	1.0000	1.7800	2.0000	216.4	114.0	66.20	11.77	12.10
Max	1.0000	243.0	1.0000	1.0000	5.4000	9.0000	350.8	165.0	111.30	18.19	20.00

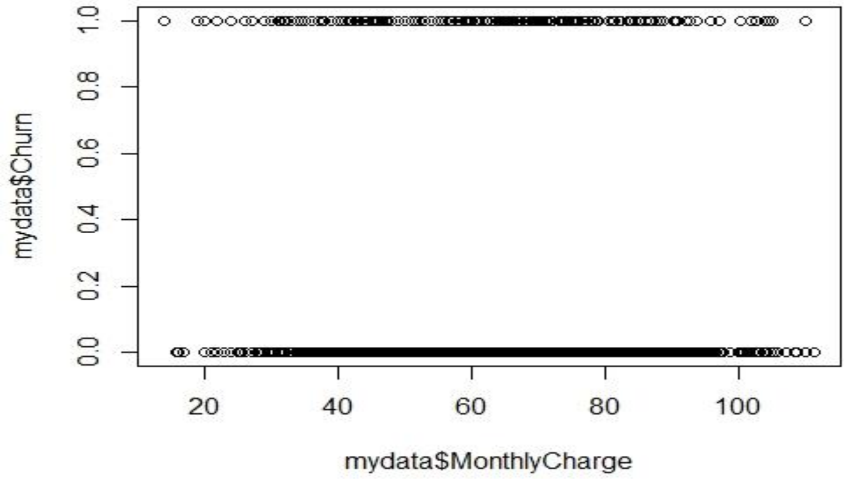
Result obtained from the five point summary of data set (Table-2) it is inferred that the data does not contain any outliers and missing values.




mydata$DataPlan
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
mydata$MonthlyCharge


mydata$MonthlyCharge
20
40
60
80
100
0
1
2
3
4
5
mydata$DataUsage





2.4 Multi co-linearity

There is existence of multi co-linearity in the given data set as shown in Fig.1. Among the 10 attributes (eliminating dependent variable Churn), DataPlan, DataUsage and MonthlyCharge have highly significant correlation.

- Correlation between DataPlan and DataUsage is 95%.
- Correlation between DataPlan and MonthlyCharge is 74%.

- Correlation between DataUsage and MonthlyCharge is 78%.

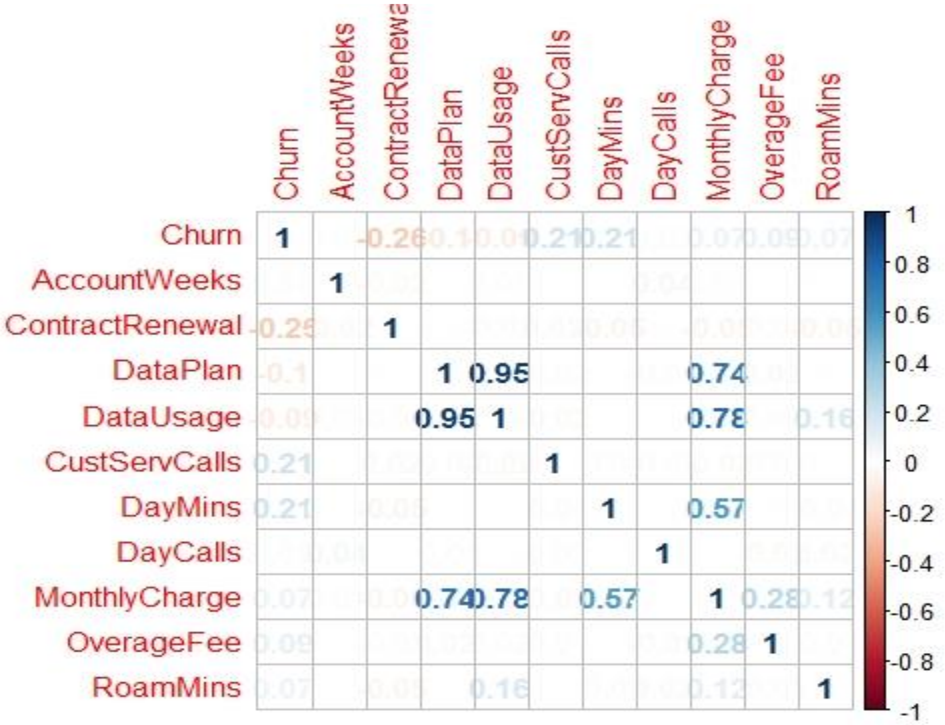


Fig.1: Correlation plot

Treatment of Multi co-linearity:

From Fig.1 it is evident that there exist a strong correlation between the variables DataPlan, DataUsage and MonthlyCharge. Hence to remove or treat the problem of multi co-linearity, the concept of dimensionality reduction is utilized. Variables DataPlan, DataUsage and MonthlyCharge are grouped together into a single variable named as “ServiceUsage”.

Factor analysis method is utilized to group the variables and the results obtained are tabulated in Table-3.

Table-3: Correlation

	DataPlan	DataUsage	MonthlyCharge
DataPlan	1.0000	0.9459	0.7374
DataUsage	0.9459	1.0000	0.7816
MonthlyCharge	0.7374	0.7816	1.0000

Table-4: Eigen value

2.6468	0.3020	0.0511

Table-5: Factor analysis solution without rotation

	PA1	h2	u2	com

DataPlan	0.95	0.90	0.10	1
DataUsage	1	0.99	0.0085	1
MonthlyCharge	0.78	0.61	0.3889	1

Multiple R square of scores with factors 0.99

Table-6: Factor analysis solution with rotation

	PA1	**h2**	**u2**	**com**
DataPlan	0.95	0.90	0.10	1
DataUsage	1	0.99	0.0085	1
MonthlyCharge	0.78	0.61	0.3889	1

Multiple R square of scores with factors 0.99

2.5 Insights from Exploratory Data Analysis

- Data set contains the categorical and continuous data.
- No oultiers exists in the data set.
- The data set represents the bivariate analysis as the dependent variable (Churn) depends on two or more variables.
- There is existence of multi co-linearity between the variables DataPlan, DataUsage and MonthlyCharge.
- Correlation between DataPlan and DataUsage is 95%.
- Correlation between DataPlan and MonthlyCharge is 74%.
- Correlation between DataUsage and MonthlyCharge is 78%.
- Factor analysis method is employed for the treatment of multi co-linearity.

3. Logistic Regression Model Building

3.1 Interpretation from Logistic Regression Model

- Logistic regression model was build with considering churn as the dependent variable and all other variables AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee, RoamMins as the independent variables.
- Multi co-linearity is tested and the variables DataPlan, DataUsage and MonthlyCharge are found to be highly correlated.
- Factor analysis method is employed for the treatment of multi co-linearity.

- All the three variables DataPlan, DataUsage and MonthlyCharge are merged into group called ServiceUsage.
- For intercept only model, maximum value of log likely-hood function is found to be -897.8.
- For full model, maximum value of log likely-hood function is found to be -711.48.
- Mcfaden test is done and the Mcfaden value is found to be 0.2075 which gives the information that 20.75% of uncertainties in the intercept only model was explained by full model. This suggests that the build logistic regression model is a good.
- Individual coefficient test was performed and it is found that the variables AccountWeeks and DataCalls are not significant with p values as 0.72 and 0.75 respectively which is greater than 0.05 (95% level of significance).
- Logistic regression model performance is calculated on the basis confusion matrix and tabulated (table-7).
- ROC plot is drawn and area under the curve is found to be 59.14%.
- Accuracy of the model is found to 86.90%.

Table-7: Logistic regression model performance

Accuracy	Sensitivity	Specificity
0.8690369	0.206451613	0.976464435

3.2 KNN Model

3.2.1 Interpretation from KNN model

- Since the data set contains very large value therefore choosing K value to be equal to 19 (i.e. K=19).

- Partitioning the data set into train and test data using the 70:30 rule. Which means 70% of the data is used to build the model and the remaining 30% of the data is to test the performance of the build model.
- KNN model was build for all the odd values of K starting from K =19 to K =1.
- Model performance was observed from the confusion matrix corresponding to each value of K and tabulated in the table (Table-7).
- Optimum value of K is found to be K=7 corresponding to the model which has the highest accuracy.

Table-8: KNN model performance

	KNN		
Test data	**Accuracy**	**Sensitivity**	**Specificity**
k=19	0.8756	0.2316	0.9978
k=17	0.8783	0.2598	0.9957
k=15	0.881	0.2824	0.9946
k=13	0.8801	0.2881	0.9924
k=11	0.8828	0.3163	0.9903
k=9	0.8855	0.3333	0.9903
k=7	0.8936	0.3954	0.9882
k=5	0.8909	0.3954	0.9849
k=3	0.8819	0.3841	0.9764
k=1	0.8639	0.4802	0.9367

3.3 Naïve Bayes Model

3.3.1 Interpretation from Naïve Bayes Model

- It takes less time to build a model.
- The accuracy with the naïve bayes model for training data set is found to be 90.57%.
- Hence among the three models i.e. logistic regression model, KNN and Naïve Bayes was found to be the best as it has the highest accuracy.

Table-9: Naïve Bayes model performance

	NB		
	Accuracy	**Sensitivity**	**Specificity**
Train data	0.9057	0.3958	0.9945
Test data	0.8845	0.3863	0.9823

3.4 Actionable Insights and Recommendation

- Results obtained from the exploratory data analysis explains that the DataPlan, DataUsage and MonthlyCharge are highly correlated which means that if a company wants to reduce the customer churn, it should improve its data quality w.r.t to monthly charges.
- Increase in the data plan with reduced monthly charge is recommended.
- Although in order to reduces the loss beared by the company after providing data services at low cost, the company can increase their roaming charges, overage fee and variables.