TELECOM CHURN CLASSIFICATION MODEL

https://tsrajan29.wixsite.com/data-analytics Website link

PROJECT - CHURN

Abstract: Churn represents the problem of losing a customer to another business competitor which leads to serious profit loss. The increase in the number of churn customers is become the present-day challenge to the telecom industry and such customers create financial burden to the company, identifying such customers is the objective of this Project from the data provided by Acadgild.

Research indicates that the cost of developing a new customer is approximately 5 times higher than retaining the new customer. Many companies looks for Business intelligent solution to predict churn rates for designing effective plans for customer retention.

<u>Scope</u>

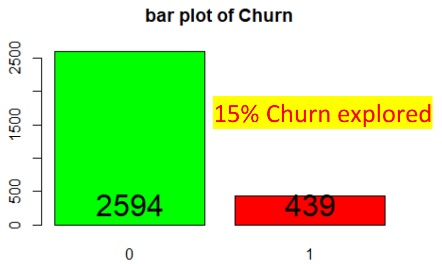
Carryout Data analysis using R – various classification models like Logistic Regression, Decision Trees, Pruning, Bagging, Random Forests, Adaptive Boosting, SVM, ANN, Nearest Neighbor…etc. and provide the Best among the model.

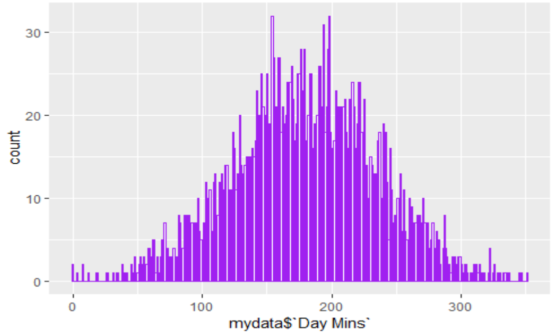
Provide a detailed analysis based on our findings through a power point presentation along with the supporting R Markdown documents.

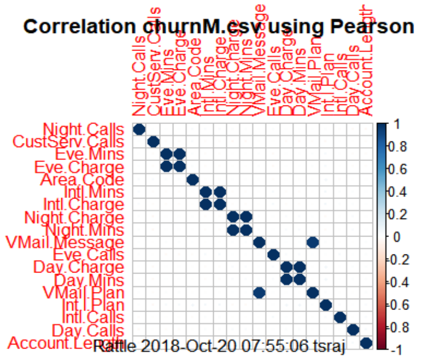
Develop a website of your own as per the guidelines provided and present the data in the website .

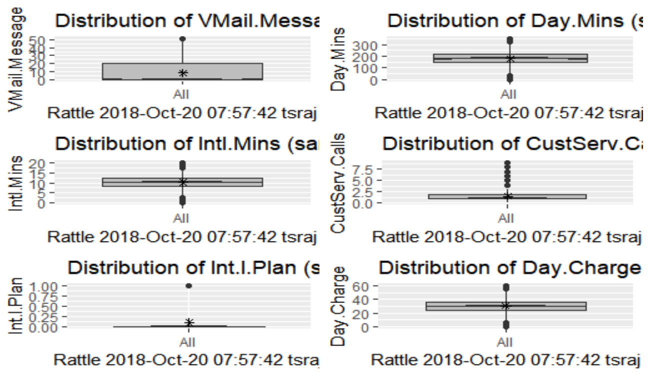
PROJECT – CHURN – CLASSIFICATION MODEL

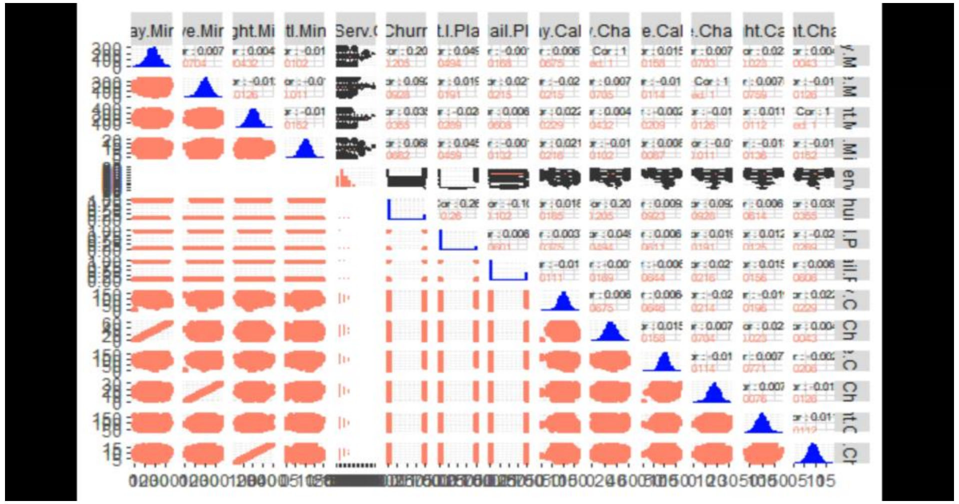
- Model : RV = churn
- Exploratory studies with various visualizations graphs like bar graph, Histogram, Box plot, Ggally gg pairs correlation graphs, heat map and tableau graphs.
- Logistic Regression, Decision Trees, Pruning, Bagging, Random Forests, Adaptive Boosting, SVM, ANN, Nearest Neighbor…etc.
- Various ROC Curves, AUC etc., and the best identified.
- Compared Variable importance plots, information value summary
- Analyzed the findings with that of the Churn excel data and provided a conclusion.
- R Markdown files along with a website link is provided

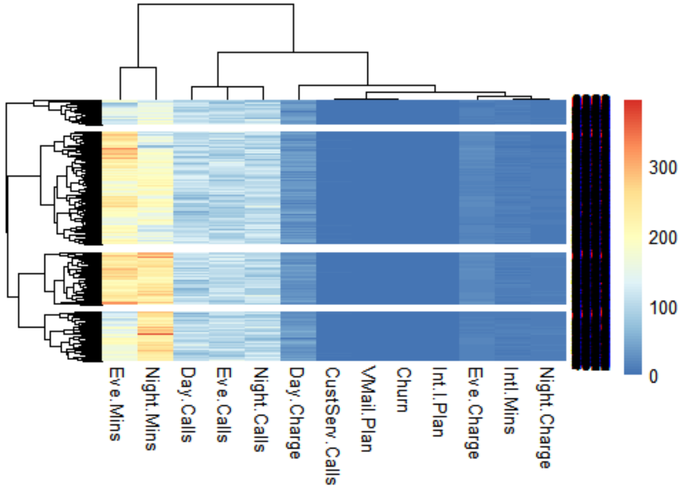


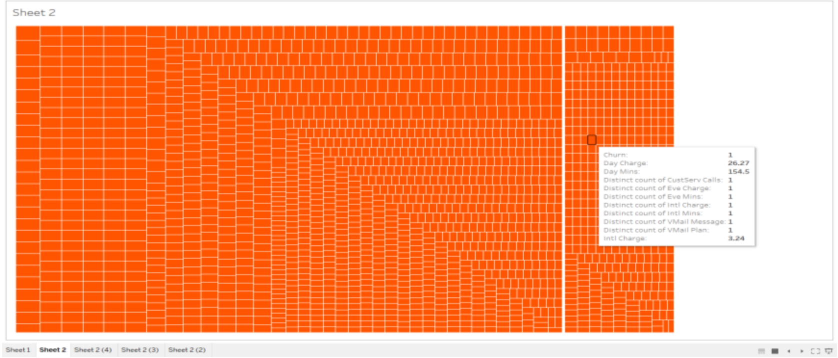


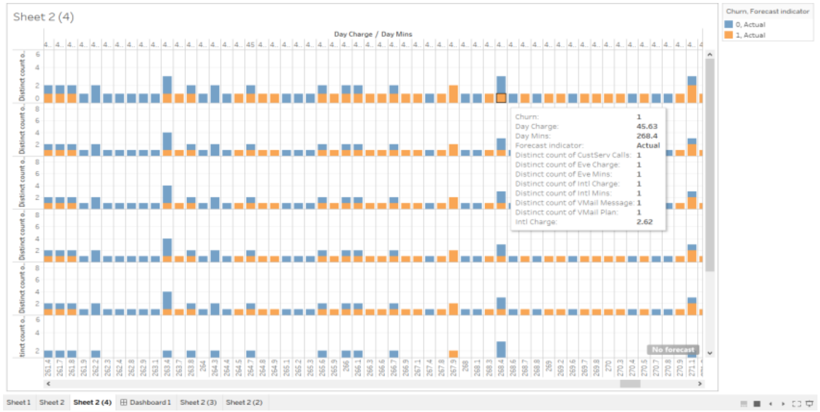








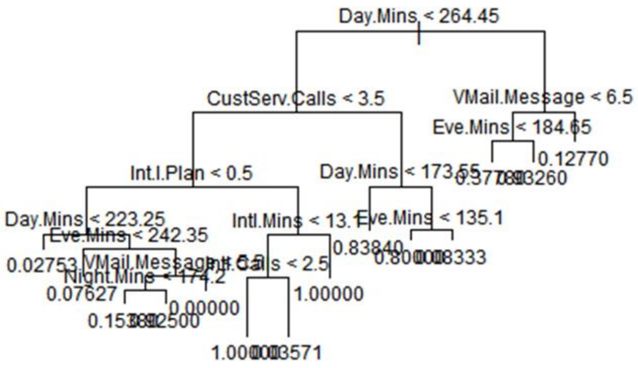


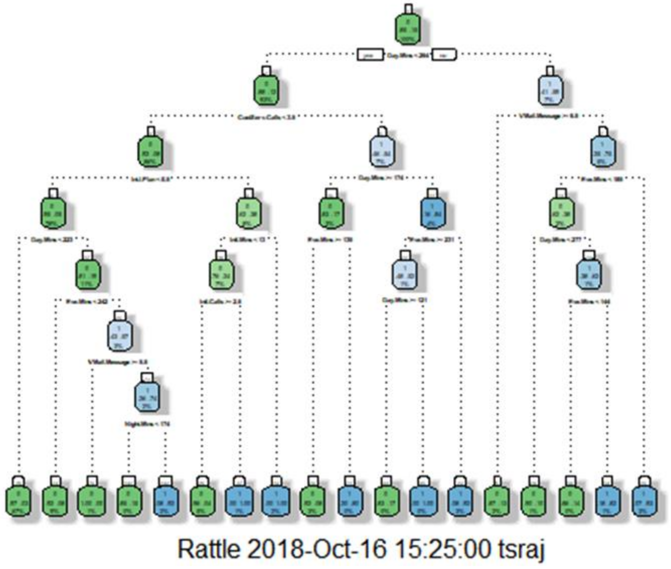


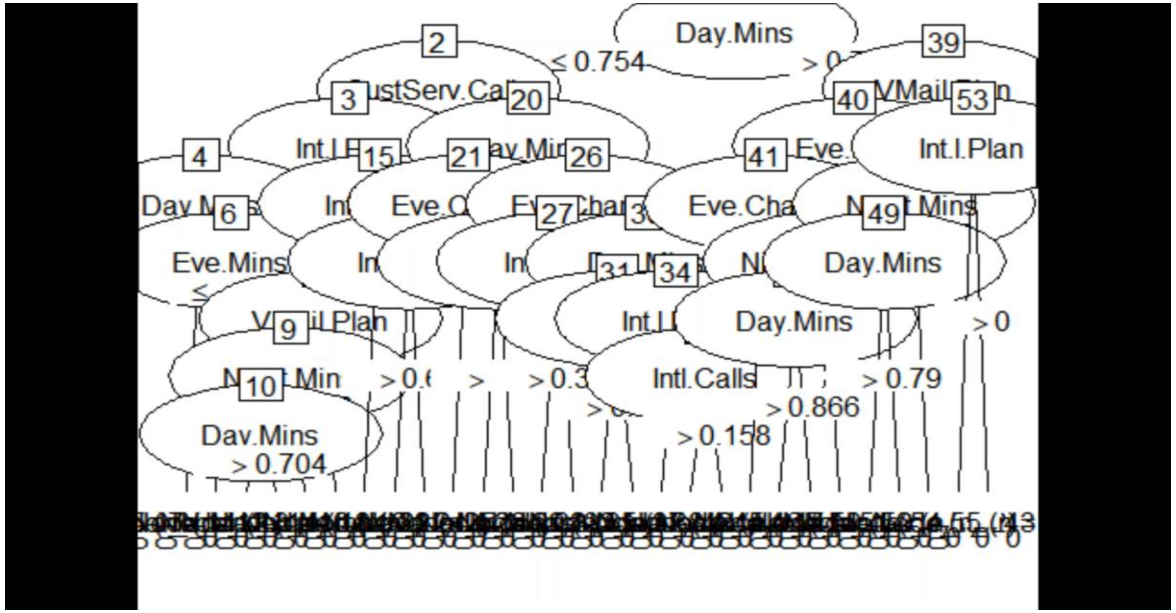
Exploratory Phase – Project CHURN

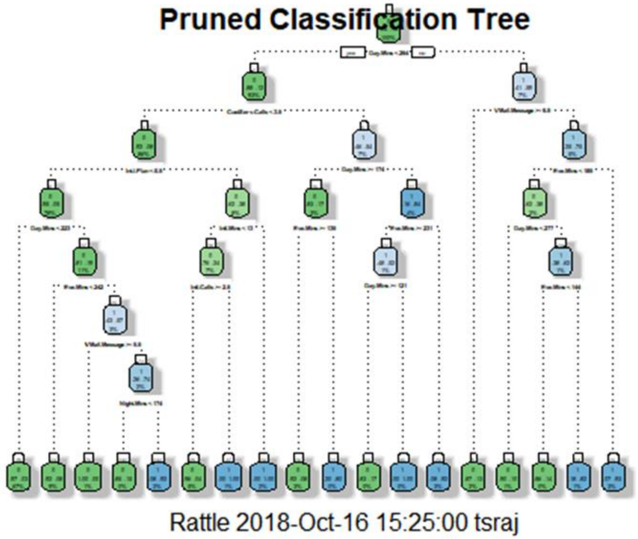
- Assumed ‘1’ as yes for Customer Churn.
- The bar graph indicate around 15% Churn and ‘Churn’ is the response variable.
- Histogram drawn for major activity like Day minutes.
- Box plot shows the Distribution of various continuous variables.
- Ggally ggpairs plots were drawn to show for all the variables and one such video uploaded for complete visualization of the data
- P3 heatmap provides a clear data range
- D3 heatmap segregates churn 1 and 0 with some important information in each section.
- A clear bar graph through Tableau explains about both Churn 1 and 0 details.
- **Decision tree learning** is one of the predictive modelling approaches used in statistics

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics



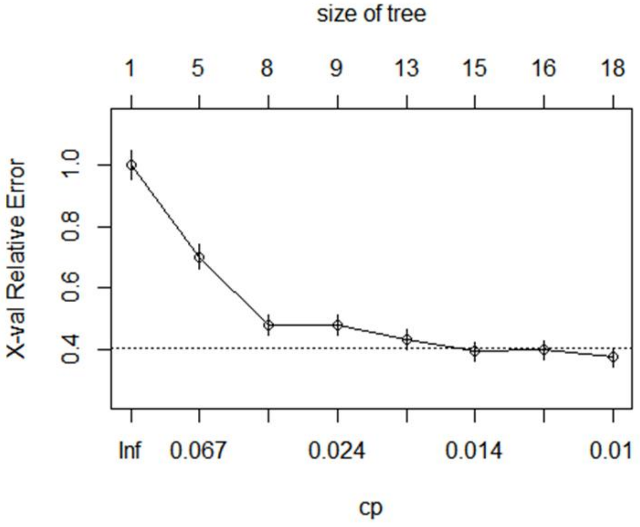




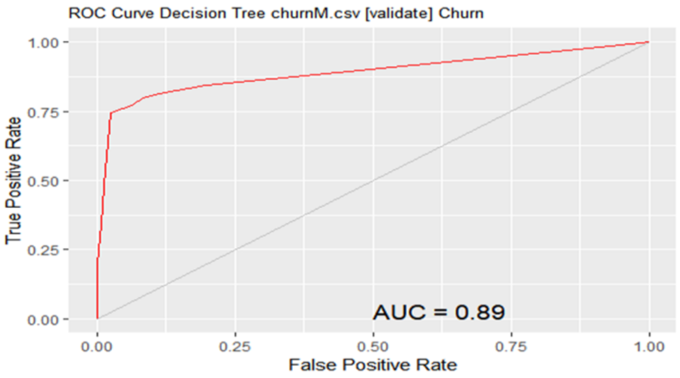


```
printcp(tree)
## Classification tree:
## rpart(formula = Churn ~ Account.Length + VMail.Message + Day.Mins +
##     Eve.Mins + Night.Mins + Intl.Mins + CustServ.Calls + Int.l.Plan +
##     VMail.Plan + Day.Calls + Day.Charge + Eve.Calls + Eve.Charge +
##     Night.Calls + Night.Charge + Intl.Calls + Intl.Charge, data = datatrai
n,
##     method = "class")

## Variables actually used in tree construction:
## [1] CustServ.Calls Day.Mins       Eve.Mins       Int.l.Plan
## [5] Intl.Calls     Intl.Mins      Night.Mins     VMail.Message
##
## Root node error: 393/2666 = 0.14741
##
## n= 2666
##         CP nsplit rel error  xerror     xstd
## 1 0.084606      0   1.00000 1.00000 0.046577
## 2 0.053435      4   0.66158 0.70229 0.040025
## 3 0.027990      7   0.46056 0.48092 0.033719
## 4 0.021204      8   0.43257 0.48092 0.033719
## 5 0.015267     12   0.34606 0.43511 0.032189
## 6 0.012723     14   0.31552 0.39440 0.030744
## 7 0.010178     15   0.30280 0.39949 0.030930
## 8 0.010000     17   0.28244 0.37659 0.030084
```

plotcp(tree)

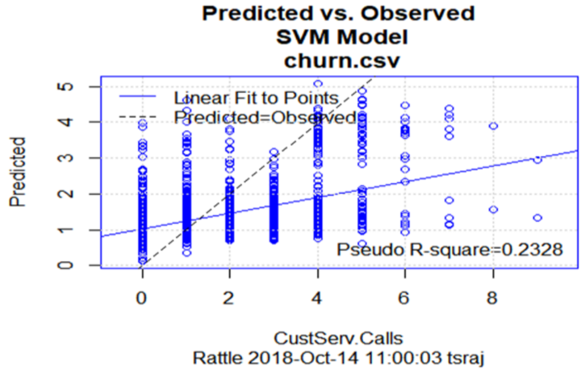



```
improve=1.00000000, (0 missing)
##    Day.Mins       < 264.55 to the left,  improve=0.09769634, (0
missing)
##    Day.Charge     < 44.975 to the left,  improve=0.09769634, (0
missing)
##    CustServ.Calls < 3.5    to the left,  improve=0.09755406, (0
missing)
##    Int.l.Plan     < 0.5    to the left,  improve=0.06796193, (0
missing)
##   Surrogate splits:
##     Day.Mins       < 284.15 to the left,  agree=0.870, adj=0.066,
(0 split)
##     Day.Charge     < 48.305 to the left,  agree=0.870, adj=0.066,
(0 split)
##     CustServ.Calls < 4.5    to the left,  agree=0.869, adj=0.058,
(0 split)
##     Eve.Mins       < 335.85 to the left,  agree=0.862, adj=0.003,
(0 split)
##     Eve.Charge     < 28.545 to the left,  agree=0.862, adj=0.003,
(0 split)
##
## Node number 2: 2153 observations
##   mean=0, MSE=0
##
## Node number 3: 347 observations
##   mean=1, MSE=0
```

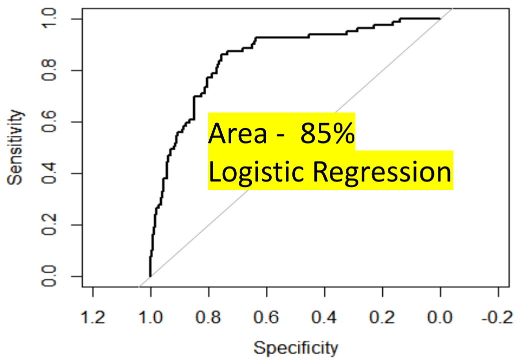


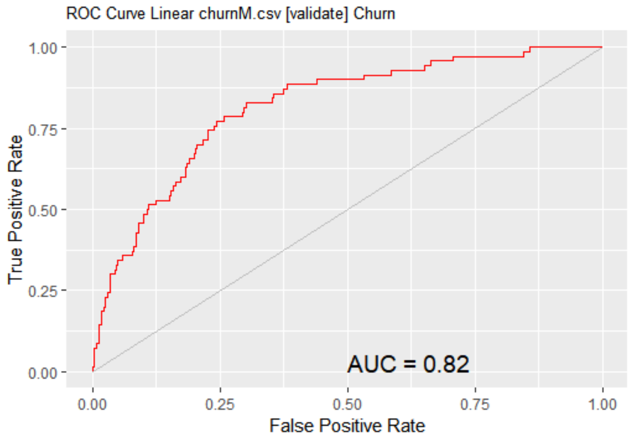
KSVM Classification

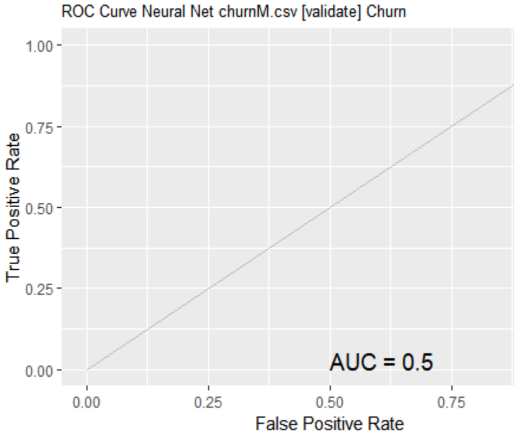
```
set.seed(12345)
churn_classifier_rbf<-ksvm(Churn ~., data = datatrain, kernel='rbfdot')
churn_predictions_rbf<-predict(churn_classifier_rbf,datatest)
agreement_rbf<-churn_predictions_rbf == datatest$Churn
table(agreement_rbf)

## agreement_rbf
## FALSE
##   667

prop.table(table(agreement_rbf))

## agreement_rbf
## FALSE
##     1
```

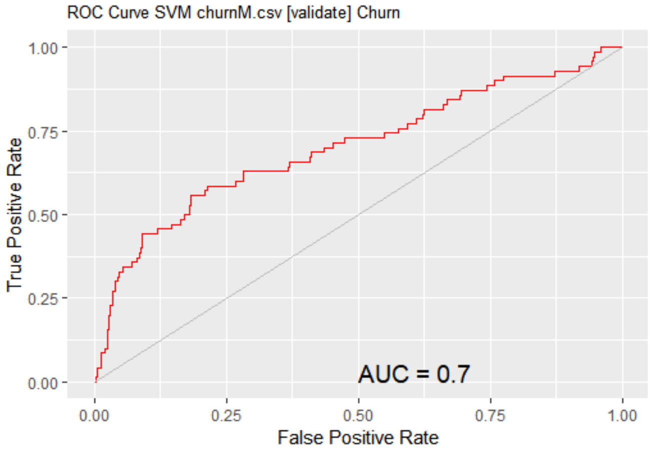


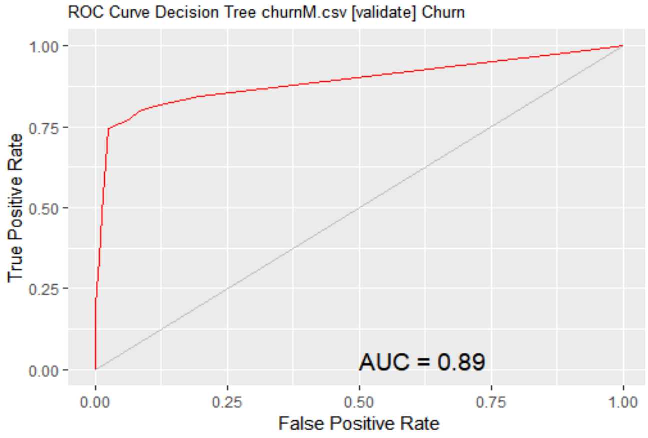
- The complexity parameter (**cp**) is used to control the size of the **decision tree** and to select the optimal **tree** size. If the cost of adding another variable to the **decision tree** from the current node is above the value of **cp**, then **tree** building does not continue.
- Plot cp() provides a graphical representation to the cross validated error summary. The cp values are plotted against the geometric mean to depict the deviation until the minimum value is reached.
- ROC curve provides the true positive rate and False positive rate

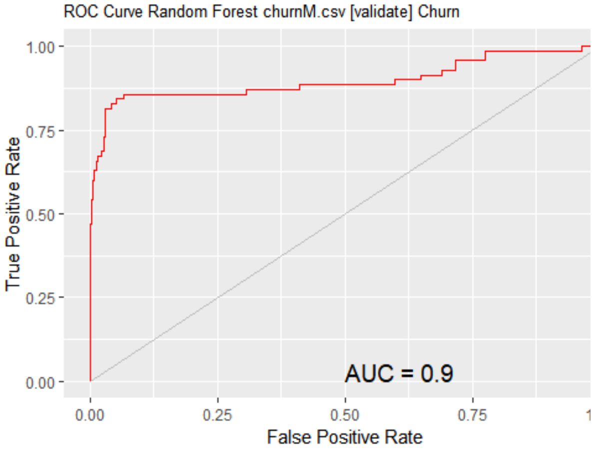


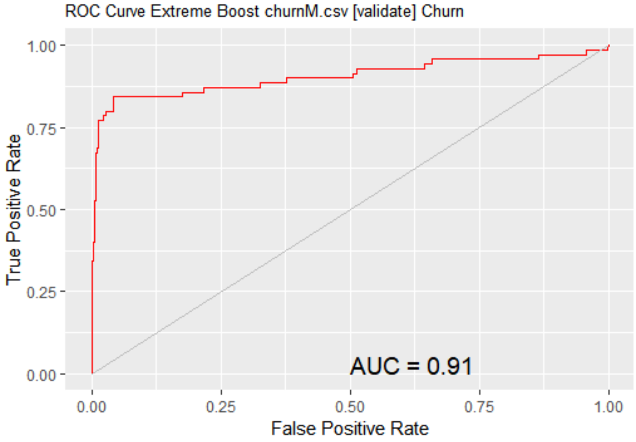




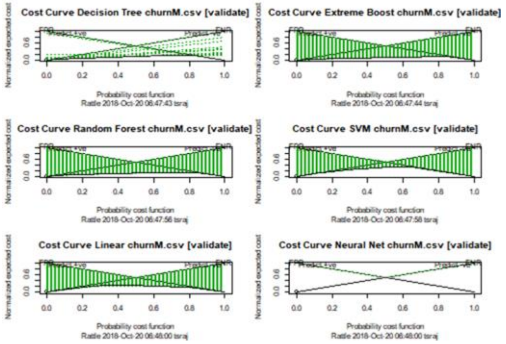


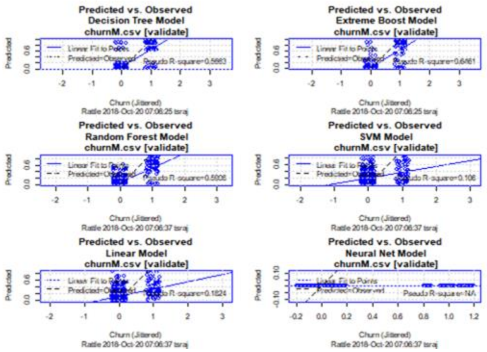


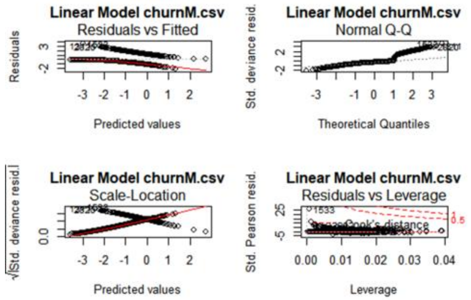


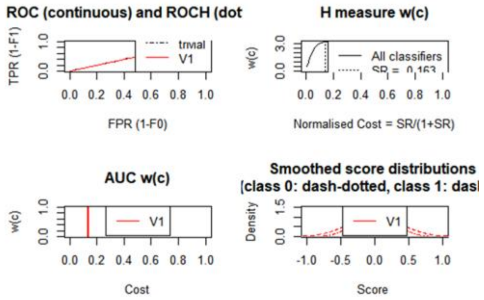


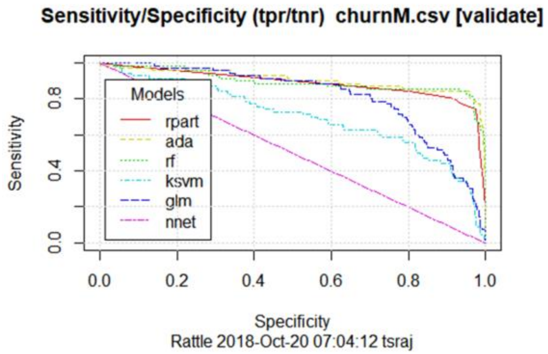
AUC plotted graph for various classification models like Logistic regression, Linear model, ANN, SVM, Decision tree ,Random Forest and Boost models explains true positive rate more than 80% except ANN and SVM model for Data set Churn provided. The Random Forest (RF) algorithm for regression and classification has considerably gained popularity & has grown to a standard classification approach competing with logistic regression in many analysis. Based on the AUC curves, Performance curves, RF performs better in accuracy , we choose to explain the Logistic regression, Tree and Random Forest model. The parameter ntree denotes the number of trees in the forest. The default value is ntree =500 in the package random Forest. The parameter mtry denotes the number of features randomly selected as candidate features at each split

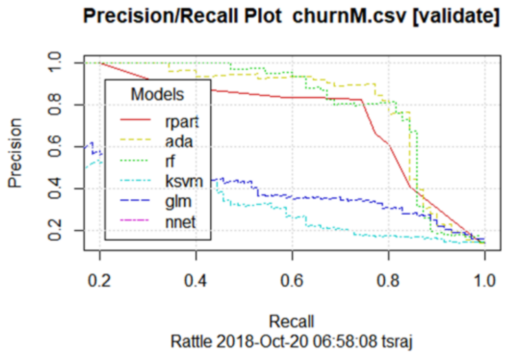


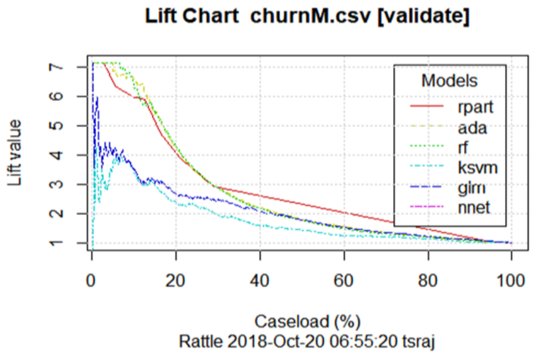


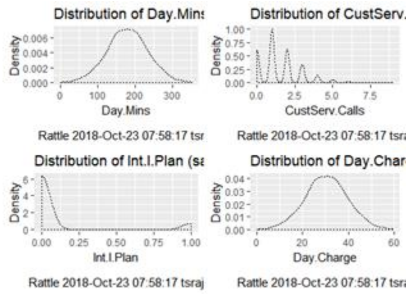












- Sensitivity and specificity are statistical measures of the performance of a binary classification test, sensitivity can also be a true positive rate and specificity as true negative rate.
- The precision-recall curve shows the trade off between precision and recall for different threshold

- Lift charts provides either a total cumulative response or incremental response rate for the purposes of comparing & bench marking the predictive capability of different binary predictive models.

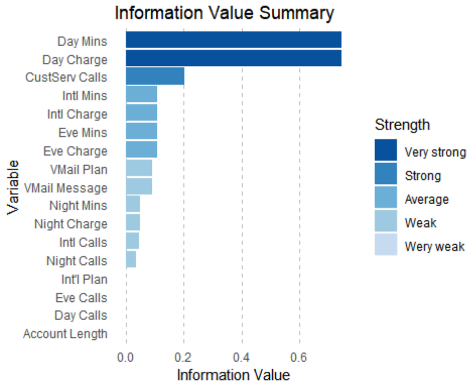
Analysis of Deviance Table

Model: binomial, link: probit

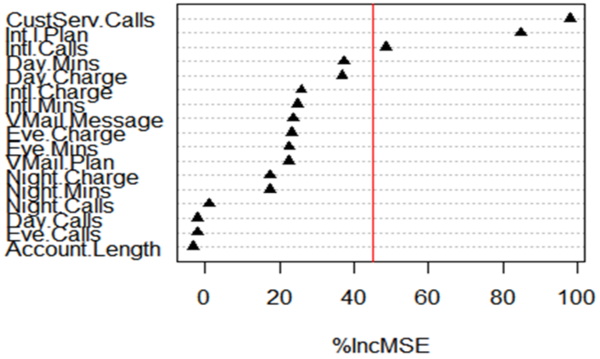
Response: ChurnTerms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2332	1976.0	
Account.Length	1	0.589	2331	1975.4	0.4428532
VMail.Message	1	15.036	2330	1960.4	0.0001055 ***
Day.Mins	1	94.529	2329	1865.8	< 2.2e-16 ***
Eve.Mins	1	21.278	2328	1844.6	0.000003973 ***
Night.Mins	1	2.431	2327	1842.1	0.1189529
Intl.Mins	1	14.557	2326	1827.6	0.0001360 ***
CustServ.Calls	1	135.923	2325	1691.7	< 2.2e-16 ***
Int.l.Plan	1	156.328	2324	1535.3	< 2.2e-16 ***
VMail.Plan	1	9.877	2323	1525.5	0.0016736 **
Day.Calls	1	1.489	2322	1524.0	0.2223286
Day.Charge	1	0.164	2321	1523.8	0.6853416
Eve.Calls	1	0.093	2320	1523.7	0.7599448
Eve.Charge	1	0.020	2319	1523.7	0.8882200
Night.Calls	1	0.043	2318	1523.6	0.8363527
Night.Charge	1	0.061	2317	1523.6	0.8045178
Intl.Calls	1	12.590	2316	1511.0	0.0003877 ***
Intl.Charge	1	0.229	2315	1510.8	0.6319434
Area.Code	1	0.070	2314	1510.7	0.7907806

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1




```
Evaluation on training data (3033 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      28  113( 3.7%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##    2578    16    (a): class 0
##      97   342    (b): class 1
##  Attribute usage:
##
##  100.00% Day.Mins
##   93.67% CustServ.Calls
##   91.03% Int.l.Plan
##   16.85% Eve.Mins
##    9.20% Eve.Charge
##    8.94% Intl.Calls
##    8.80% VMail.Plan
##    6.56% Intl.Mins
##    4.88% Night.Mins
##    1.25% Night.Charge
```



%IncMSE is the most robust and informative measure. Higher number, the more important

A higher **Mean Decrease** in **Gini** indicates higher variable importance.

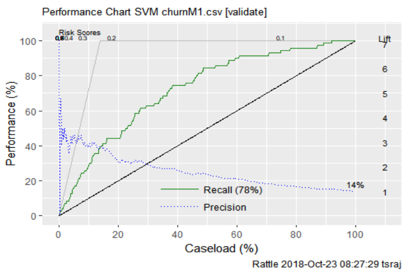
	0	1	MeanDecrease Accuracy	Mean Decrease Gini
CustServ.Calls	57.20	72.51	75.35	46.32
Int.l.Plan	49.69	63.11	62.99	30.26
Day.Charge	30.09	32.57	40.02	57.47
Day.Mins	29.86	32.18	39.73	54.84
Intl.Calls	25.3	31.31	34.28	18.93
Intl.Mins	17.80	15.39	22.04	15.84
VMail.Message	16.64	20.59	21.99	12.06
VMail.Plan	17.32	20.50	21.49	7.92
Intl.Charge	16.89	16.12	21.44	15.82
Eve.Mins	17.20	20.69	21.03	23.82
Eve.Charge	17.21	20.04	20.94	23.14
Night.Mins	12.11	3.90	12.98	12.83
Night.Charge	11.92	2.57	12.47	12.60
Night.Calls	1.16	-0.87	0.69	10.90
Day.Calls	-0.56	1.33	0.10	11.34
Eve.Calls	0.58	-2.18	-0.48	9.54
Account.Length	-0.24	-1.57	-0.88	10.40
Area.Code	-1.85	0.70	-1.42	2.90

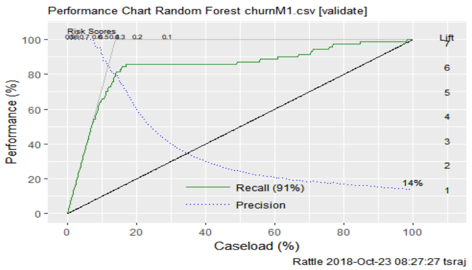
Random Forest using Conditional Inference Trees

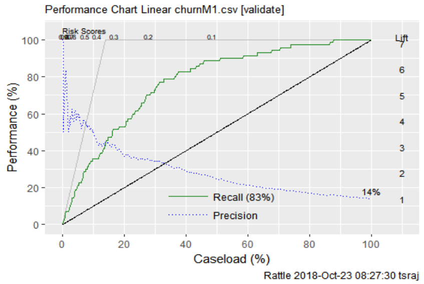
Number of observations: 2333	
Variable Importance	
Importance	
Day.Charge 0.03610955711	
CustServ.Calls	.03498601399
Day.Mins 0.03400233100	
Int.l.Plan 0.02481118881	
Eve.Mins 0.00944988345	
Eve.Charge 0.00888111888	
VMail.Plan 0.00800233100	
Intl.Calls 0.00725407925	
VMail.Message 0.00557342657	
Intl.Charge 0.00525874126	
Intl.Mins 0.00478088578	
Night.Mins 0.00123543124	
Night.Charge 0.00105128205	
Night.Calls 0.00002331002	
Area.Code 0.00003030303	-
Day.Calls	.00023076923
Account.Length 0.00023776224	-
Eve.Calls 0.00024009324	

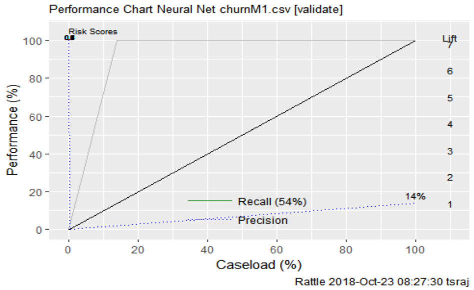
```
summary(fit1)

## Call:
## rpart(formula = Churn ~ ., data = churnTrain[, -1])
##   n= 2500
##
##     CP nsplit rel error   xerror       xstd
## 1 1.00      0         1 1.000371 0.04180395
## 2 0.01      1         0 1.187941 0.05840154
##
## Variable importance
##          Phone      Day.Charge        Day.Mins CustServ.Calls
##             84               6               6              5
##
## Node number 1: 2500 observations,    complexity param=1
##   mean=0.1388, MSE=0.1195346
##   left son=2 (2153 obs) right son=3 (347 obs)
```

```
importance(rf)

##                   %IncMSE IncNodePurity
## Account.Length -3.3667171      8.958597
## VMail.Message  23.5093488     10.984626
## Day.Mins       37.3007328     46.899902
## Eve.Mins       22.6072179     22.906599
## Night.Mins     17.2330540     12.429405
## Intl.Mins      24.8707290     13.953814
## CustServ.Calls 97.9485792     33.708149
## Int.l.Plan     84.7632991     27.756730
## VMail.Plan     22.4330568      7.813428
## Day.Calls      -2.0564407      9.410889
## Day.Charge     36.6905800     44.824857
## Eve.Calls      -2.2526506      7.513130
## Eve.Charge     23.1060040     23.041325
## Night.Calls     0.9573658      8.763461
## Night.Charge   17.2960451     11.731147
## Intl.Calls     48.4776494     18.678216
## Intl.Charge    25.7659394     14.681465
```

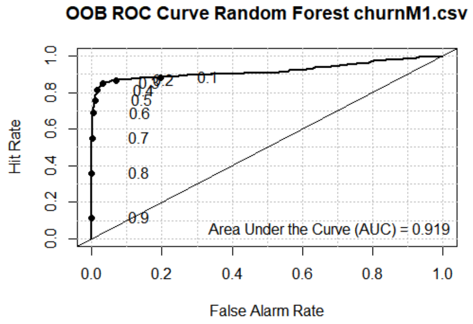


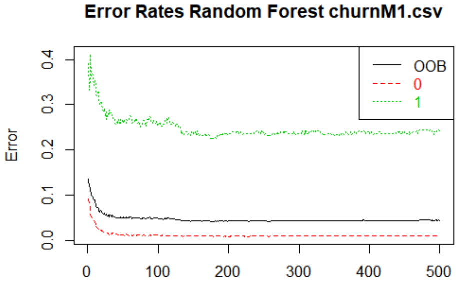


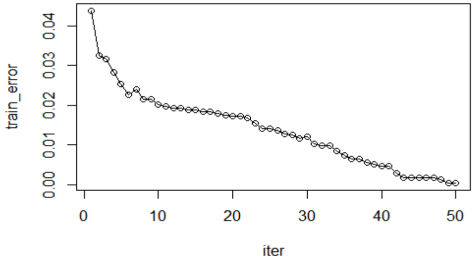


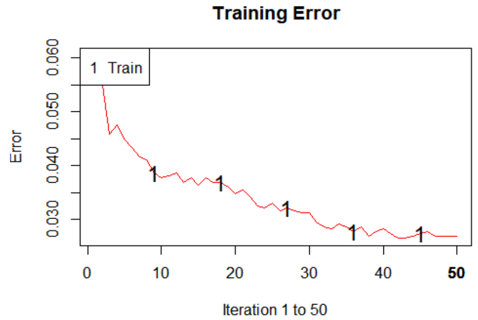


A risk chart plots performance against caseload. Suppose we had a population of just 100 entities (audit cases). The case load is the percentage of these cases that we will actually ask our auditors to process. Like ROC curves, we need to be careful in the use of risk charts for selecting cut points. We note that the area under the ROC curve is a good measure of predictive discrimination, but in comparing models it may not be sensitive enough

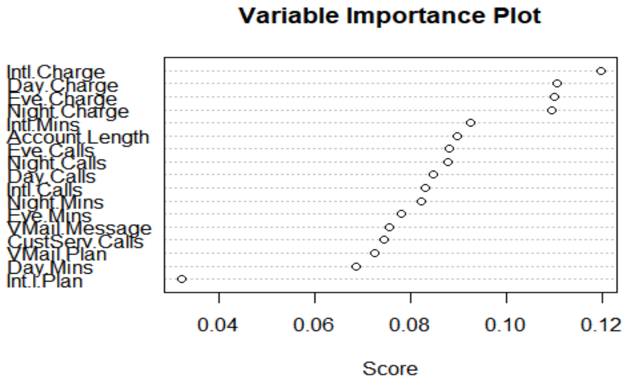


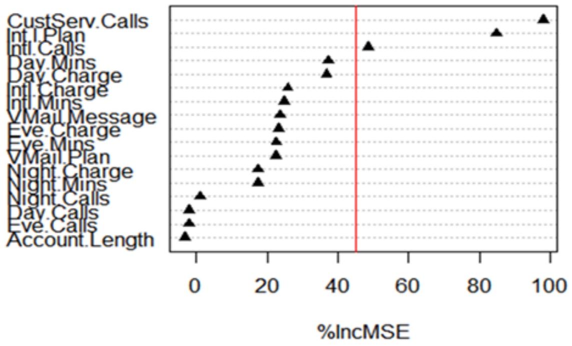


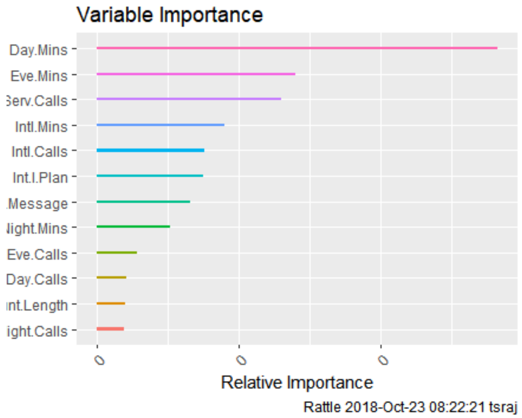


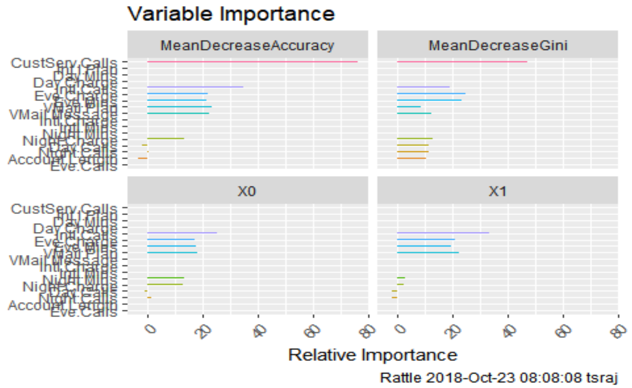


Out-of-bag (**OOB**) error, also called out-of-bag estimate, is a method of measuring the prediction error of **random forests**, boosted **decision** trees etc utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training..









- Through Logistic regression model we can generate Analysis of Deviance table, Alternate hypothesis and important value to identify & consider the important variable for the model, Random forest model provides variable importance plot with %incMSE, variable importance through Mean decrease accuracy and Mean decrease Gini etc.,
- From these charts we can identify the important variables which are closely impacted the churn of the customers

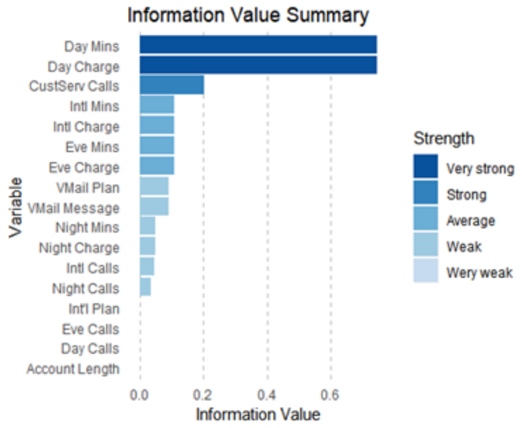
R findings compared to the Excel data Churn

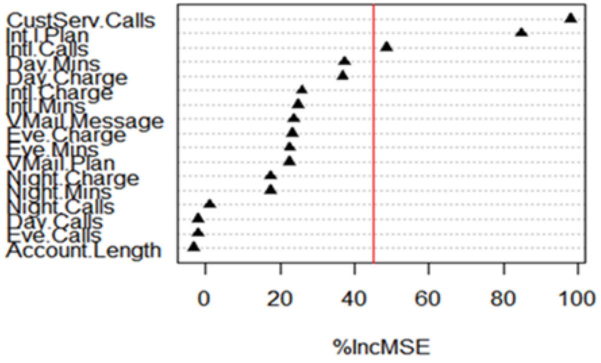
	All customers		Churn1		% churn of count	% Churn of sum	per min charges	% loss of total charges
	count	sum	count	sum				
Day Mins	3333	599190	483	99940	14	16.7		
Eve Mins	3333	669867.5	483	102594.1	14	15.3		
Night Mins	3333	669506.5	483	99126.9	14	14.8		
Intl Mins	3333	34120.9	483	5168.1	14	15.1		
CustServ Calls (non zeros)	2636	5209	389	1077	15	20.7		
Int'l Plan (1)	323	323	137	137	42	42.4		
VMail Plan (1)	922	922	80	80	9	8.7		
Day Calls	3333	334752	483	48945	14	14.6		
Day Charge	3333	101864.17	483	16989.97	14	16.7	0.170	8.6
Eve Calls	3333	333681	483	48571	14	14.6		
Eve Charge	3333	56939.44	483	8720.55	14	15.3	0.085	4.4
Night Calls	3333	333659	483	48493	14	14.5		
Night Charge	3333	30128.07	483	4460.76	14	14.8	0.045	2.3
Intl Calls	3333	14930	483	2011	14	13.5		
Intl Charge	3333	9214.35	483	1395.65	14	15.1	0.270	0.70
All Charges collected		198146		39385.63				19.9

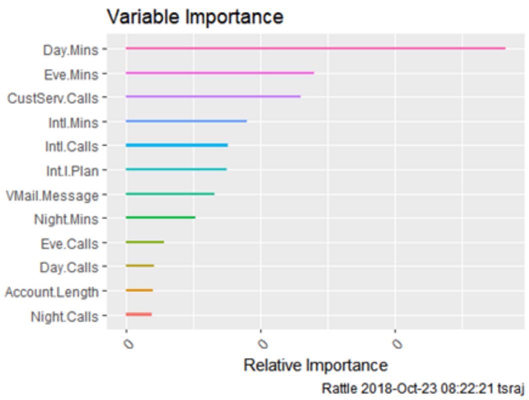
As per most of the classification models the variable importance are namely Day minutes, International minutes, Customer service calls, evening min etc

As per the excel data of churn , Days minutes reduces(8.6% of total charges) the major revenue due to churn, followed by Evening, night and international etc
The international plan had a churn rate of 43% and these customers also contribute income to Day and other minutes and hence the issue need to be looked into seriously.

There is a poor resolution of customer issues and the churn rate due to the customer service call is 20.7% which is a very high rate of churn for the business.
Hence the findings of the classification model clearly indicates the important variables as given above.







Summary

FINDING:

From exploratory findings there is overall 14% churn on telecom company.

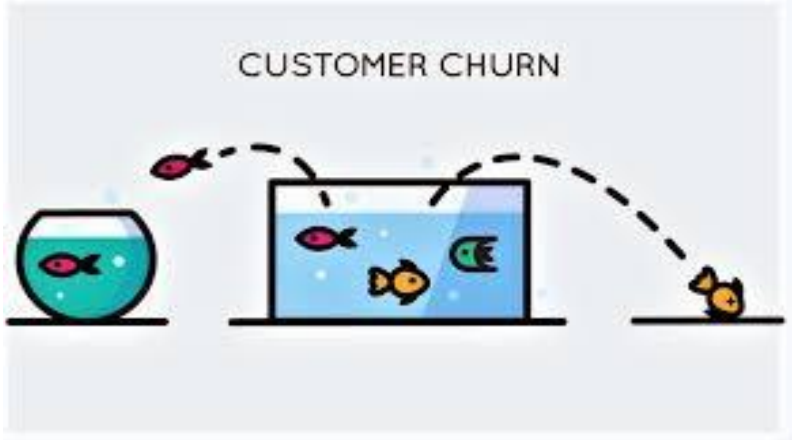
Various classification models were performed and based on the detailed analysis Random forest model appears to be a better one.

The variable important chart from RF and information value summary from Logistic model are comparable.

The important variables are compared with the various calculations made in Excel Churn data provided in the previous slide and found to be in alignment.

This data need to be further studied with other market data for the major reason for the churn rate and understand the market issue like higher charges, service issues & competition edge to set right the same.

The cost of getting a new customer is 5 times than loosing an existing customer. Hence churn Management is an Business intelligence requirement for better customer retention.



Telecom CHURN Classification Model

Please visit Project website as per below link.
https://tsrajan29.wixsite.com/data-analytics