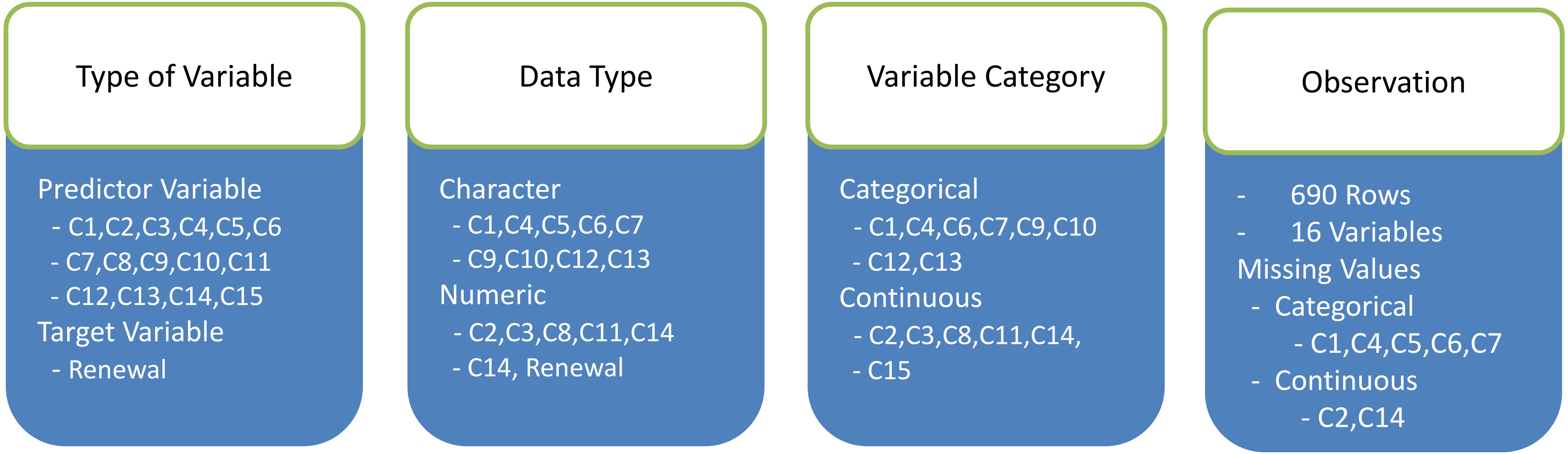
Predicting subscription Renewal

Data Summary



Type of Variable

Predictor Variable
- C1,C2,C3,C4,C5,C6
- C7,C8,C9,C10,C11
- C12,C13,C14,C15

Target Variable
- Renewal

Data Type

Character
- C1,C4,C5,C6,C7
- C9,C10,C12,C13

Numeric
- C2,C3,C8,C11,C14
- C14, Renewal

Variable Category

Categorical
- C1,C4,C6,C7,C9,C10
- C12,C13

Continuous
- C2,C3,C8,C11,C14,
- C15

Observation

- 690 Rows
- 16 Variables

Missing Values
- Categorical
  - C1,C4,C5,C6,C7
- Continuous
  - C2,C14

The following analysis helps to identify factor that affecting subscription renewal and the suggested model will help us to predict subscription renewal status.

Data Pre-Processing

Treating Missing Values

Continuous Variables

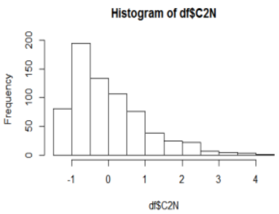
- Missing values in variables C2, C14 are imputed by using Median of all existing values.

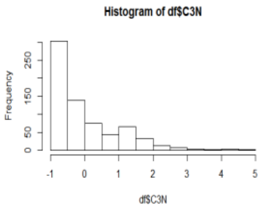
Categorical Variables

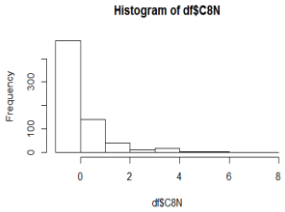
- Missing values in variables C1,C4,C5,C6,C7 are imputed by using KNN imputation method.

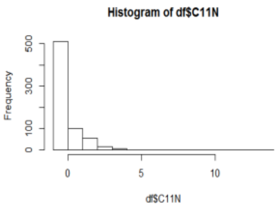
Normalization & Outlier Detection

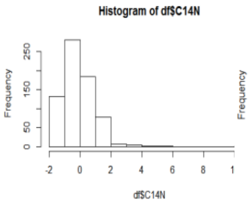
- Transforming all continuous variables to Z-score scale, i.e. $z = \dfrac{x - \mu}{\sigma}$

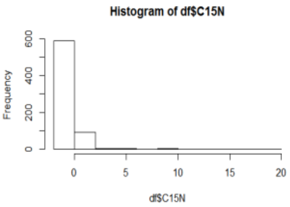






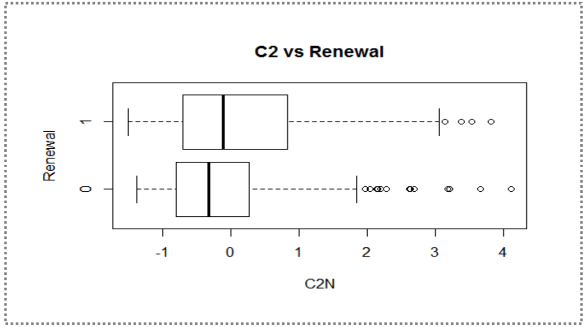


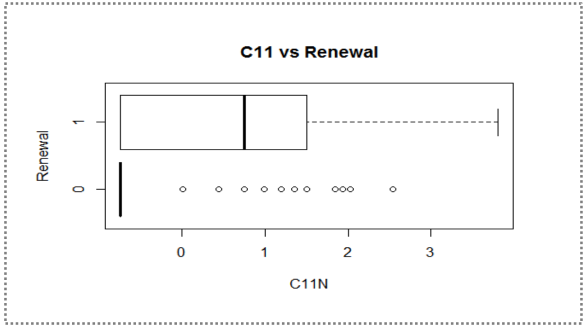


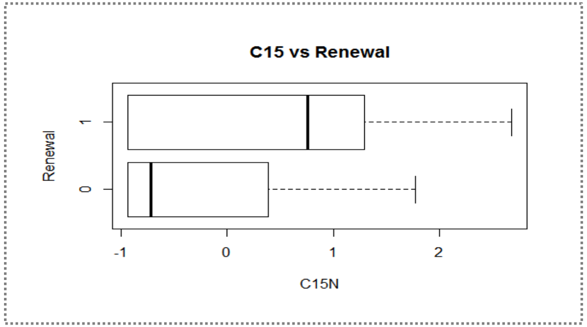


- Histogram clearly states that, data is skewed to right indicating mean is greater than median.

Variable Importance Analysis

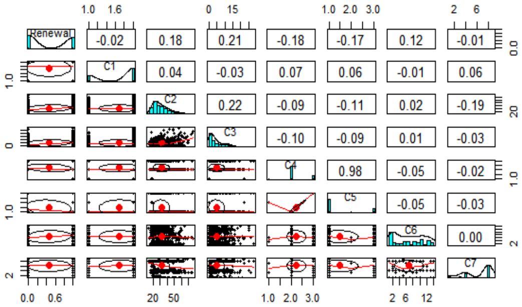






Subscription are renewable when C2 has low values but still some outlier exist.

Subscription renewal is distributed from mean of those not renewal, those who are all not renewal are skewed to right than those renewal.



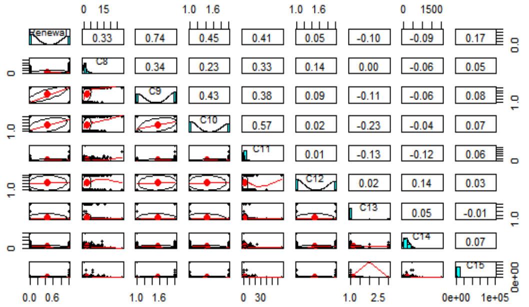


Chart in the left depict the variable correlation between predictor vs. target variable.

- ○ C8, C9, C10, C11 have high correlation with Renewal subscription.
- ○ C5, C7 have low correlation with Renewal subscription when compared to other variables.

Variable Inflation Factor

- Detecting Multi-collinearity
- VIF of jth independent variable
- Removing the variable with highest VIF (>5)

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Round #1

- Variable C5 is removed

Round #2

- Variable C7 is removed

Considering above analysis, following variables are considered for modelling.

Categorical
- C1,C4,C6,C7,C9,C10
- C12,C13

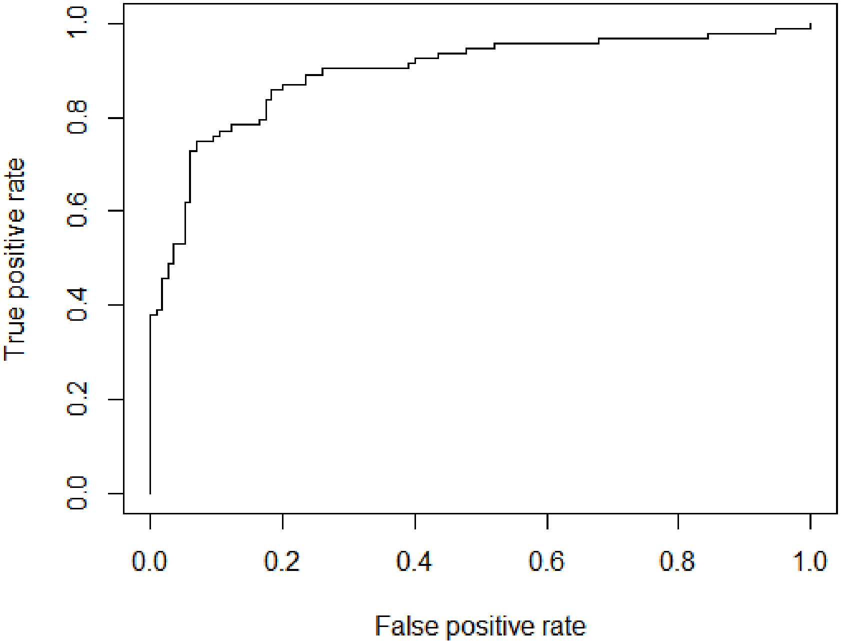
Continuous
- C8,C11,C15

Pre-Modelling Summary

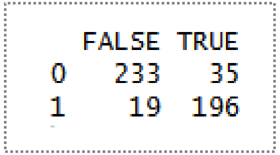
- 690 observation were considered.
- No missing value and outliers.
- Considered Variable ,
  - Categorical
    - C1,C4,C6,C7,C9,C10,C12,C13
  - Continuous
    - C8,C11,C15
- The 70% (483) of data was taken as training set, remaining 30% (207) as test period.
- For deciding accuracy, Confusion Matrix : It creates confusion matrix and error images for each class. In addition it calculates the classification accuracy assessment indices (accuracy measure - adding true positive and true negative, divide by total).

Logistic Regression

ROC plot – Testing Set



- Training



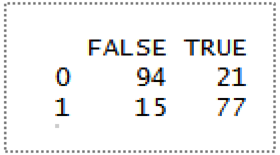
```
     FALSE  TRUE
  0    233    35
  1     19   196
```

Accuracy Measure

= (233+196)/483

= 88%

- Testing



```
      FALSE  TRUE
  0      94    21
  1      15    77
```

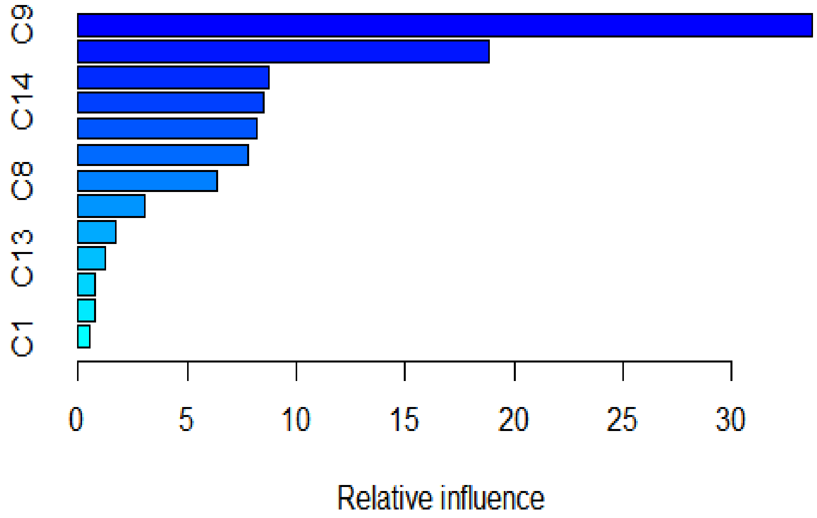
Accuracy Measure

= (94+77)/207

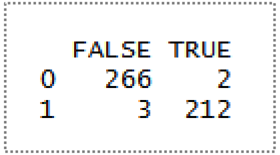
= 83%

Random Forest – ntree = 150

Variable Importance Plot



○ Training



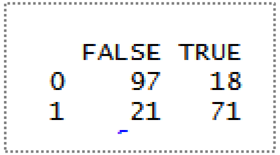
```
   FALSE  TRUE
0    266     2
1      3   212
```

Accuracy Measure

= (266+212)/483

= 98%

○ Testing



```
   FALSE  TRUE
0     97    18
1     21    71
```

Accuracy Measure

= (97+71)/207

= 92%

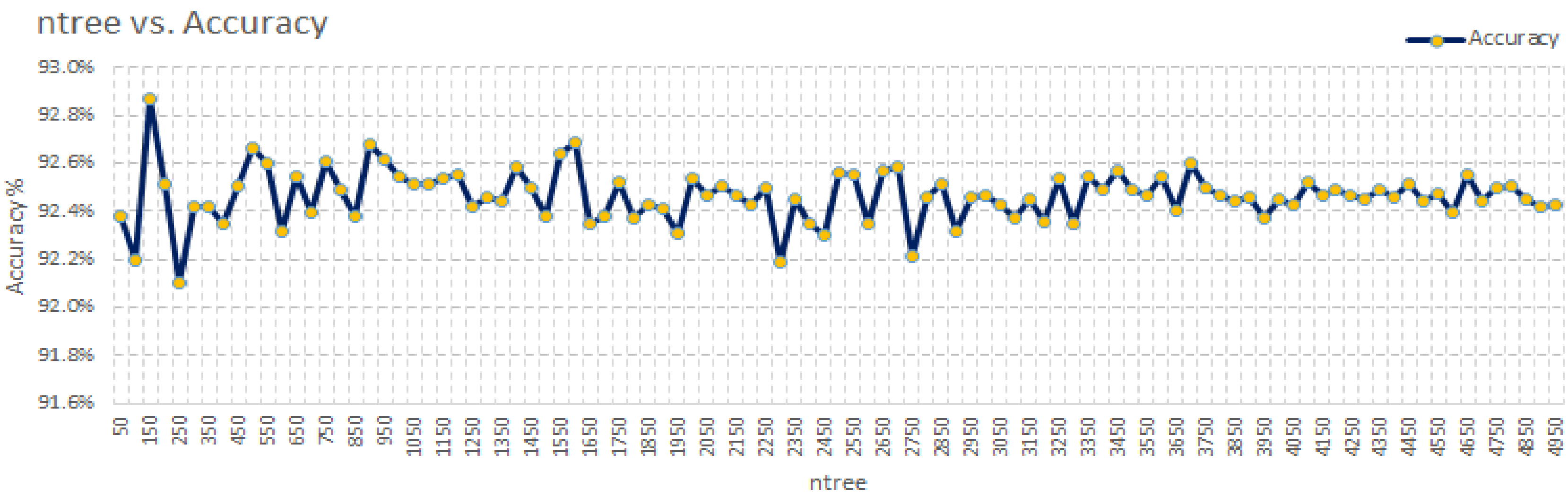
NOTE: ntree 150 explain given in appendix

Insights

- Variable C9, C14, C8, C11 has significant impact in subscription renewal.
- Other variable did not have much influence with the subscription renewal.

APPENDIX

ntree – Random Forest



- ○ Above graph depict, ntree=150 is more optimal when compared to others.