Project Report

Churn Reduction

*Madhur Dhawan*

*14 oct 2018*

* **Problem Statement:** The objective of this Case is to predict customer behaviour. We are providing you a public dataset that has customer usage pattern and if the customer has moved or not. We expect you to develop an algorithm to predict the churn score based on usage pattern.
* **Data:** There are 21 variables in our data in which 20 are independent variables and 1 (Churn) is dependent variable. Since our target variable is categorical in nature, this is clear it’s a classification problem.

**The predictors provided are as follows:**

● account length

● international plan

● voicemail plan

● number of voicemail messages

● total day minutes used

● day calls made

● total day charge

● total evening minutes

● total evening calls

● total evening charge

● total night minutes

● total night calls

● total night charge

● total international minutes used

● total international calls made

● total international charge

● number of customer service calls made

And out target varable is “Churn” which is categorical.

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analysing data sets to summarize their main characteristics. In the given data set there are 21 variables and data types of all variables are either float64 or int64. There are 3333 observations and 21 columns in train set and 1667 observations in test set in our data set. Missing value is also present in our data.

In the full dataset the variables are divided into two type **factor** and **numeric,** so first convert all variables to proper type.

**Categorical variables are:-** state, international plan, voice mail plan, churn

**Continues variables are:-** account length, area code, phone number, number vmail messages, total day minutes, total day calls, total day charges, total eve minutes, total eve calls, total eve charges, total night minutes, total night calls, total night charges, total inter minutes, total inter calls, total inter charges, number customer service calls

**Missing Value Analysis**

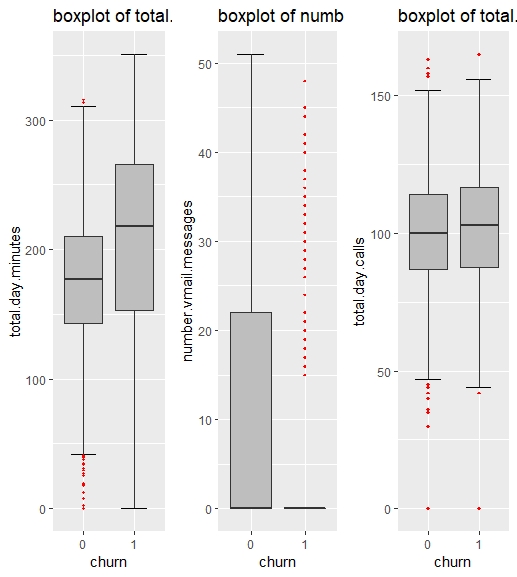
In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observations and else we need to compute those missing value.

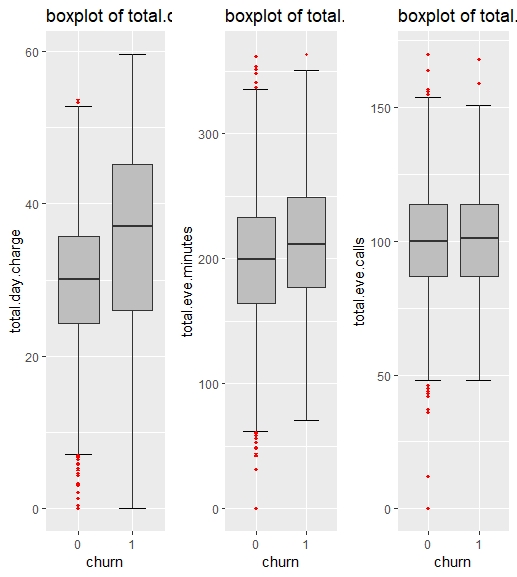
There is no missing value present in the dataset.

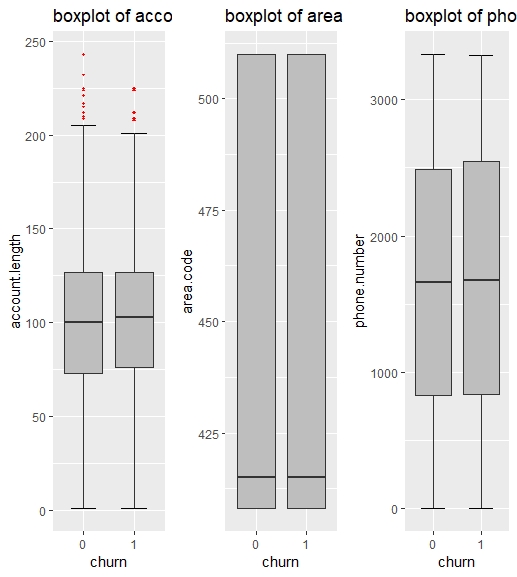
**Outlier Analysis**

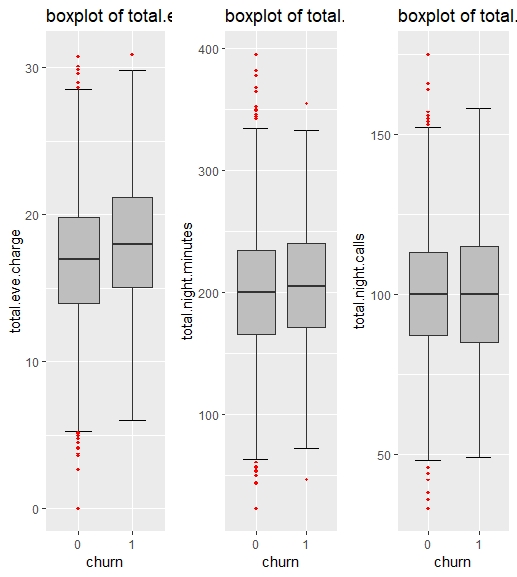
We can clearly observe from these probability distributions that most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

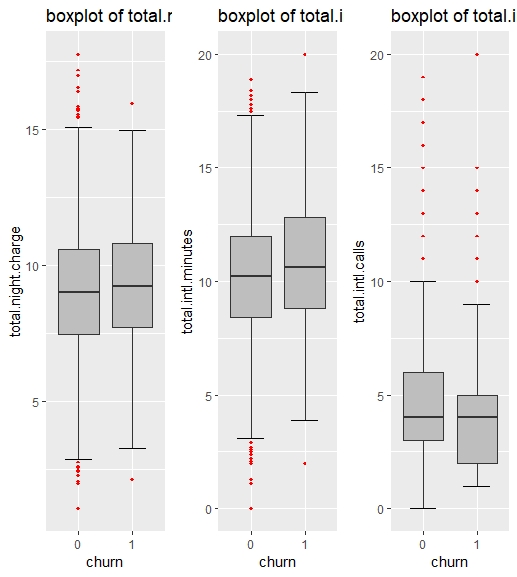
In figure we have plotted the boxplots of the 17 predictor variables with respect to **Churn**. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.

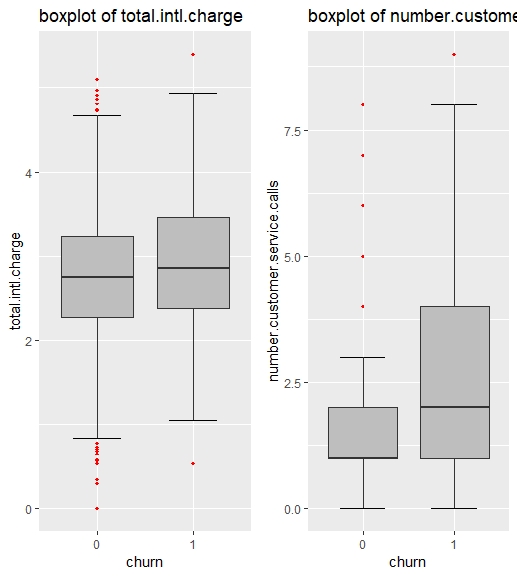










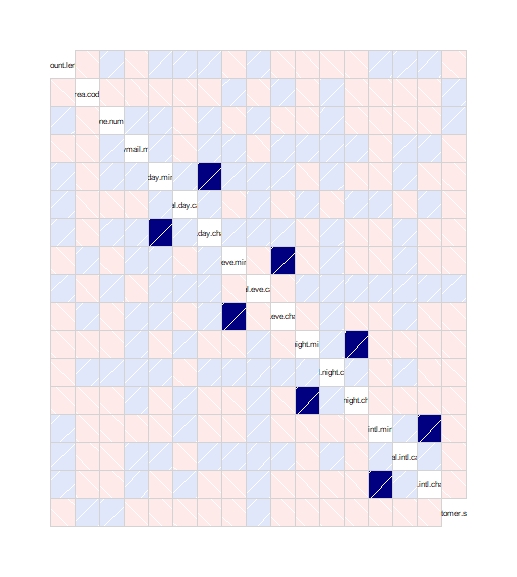


* As it is clear from the boxplot that most of the varable contains some outliers and to remove them replace it with the **maximum value and minimum value** of that particular varable.

**Features Selection**

In this part we check which varable is contributing to tell about the target varable. This can be done in two parts :-

1. **For Continues varable :-** In this we will check only continues varables i.e we will select those varable who have highly correlate with target varable and less correlate with each other.

**Correlation plot is very helpful to check the correlation between to numeric varables** 

It is clear from plot that varable **total day minutes & total day charges,** **total eve minutes & total eve charges,** **total night minutes & total night charges,** **total intl minutes & total intl charges**  are highly correlated with each other. So its better to remove one varable from it.

Cor(total.day.minutes,total.day.charge) 🡪 0.999

cor(total.eve.minutes,total.eve.charge) 🡪 0.998

cor(total.night.minutes,total.night.charge) 🡪 0.998

cor(total.intl.minutes,total.intl.charge) 🡪 0.984

**Feature Scaling**

**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

**Model Evaluation:**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models.

There are various algorithm available to predict the churn score based of above pattern. So on the behalf of error matrix we chose the best algorithm

* **Decision Tree**

With using Algorithm **C5.0**

Rule 1: (2221/60, lift 1.1)

international.plan = 0

total.day.minutes <= 223.2

number.customer.service.calls <= 3

-> class 0 [0.973]

Rule 2: (45/1, lift 1.1)

international.plan = 0

voice.mail.plan = 1

total.day.minutes > 264.4

-> class 0 [0.957]

**Confusion Matrix**

pred\_value

0 1

0 1427 16

1 74 150

Accuracy : 0.946

95% CI : (0.9341, 0.9564)

No Information Rate : 0.9004

P-Value [Acc > NIR] : 1.108e-11

Kappa : 0.7394

Mcnemar's Test P-Value : 1.874e-09

Sensitivity : 0.9507

Specificity : 0.9036

Pos Pred Value : 0.9889

Neg Pred Value : 0.6696

Prevalence : 0.9004

Detection Rate : 0.8560

Detection Prevalence : 0.8656

Balanced Accuracy : 0.9272

* **Ensembling Model**

With using Algorithm **Randam forest**(ntrees=2000)

len freq err

[1,] "6" "0.011" "0.105"

[2,] "6" "0" "0"

**Condition**

[1,] "state %in% c('NJ') & area.code<=462.5 & total.day.minutes<=265.85 & total.day.calls<=147.5 & total.intl.minutes<=13.35 & number.customer.service.calls<=3.25"

[2,] "state %in% c('NJ') & area.code<=462.5 & total.day.minutes<=265.85 & total.day.calls<=147.5 & total.intl.minutes<=13.35 & number.customer.service.calls>3.25"

pred

[1,] "0"

[2,] "1"

Confusion Matrix

0 1

0 1421 22

1 68 156

**Accuracy : 0.946**

95% CI : (0.9341, 0.9564)

No Information Rate : 0.8932

P-Value [Acc > NIR] : 1.751e-14

Kappa : 0.7459

Mcnemar's Test P-Value : 2.101e-06

Sensitivity : 0.9543

Specificity : 0.8764

Pos Pred Value : 0.9848

Neg Pred Value : 0.6964

Prevalence : 0.8932

Detection Rate : 0.8524

Detection Prevalence : 0.8656

Balanced Accuracy : 0.9154

**🡪 Statistical Model**

**Logistic Regression**

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8770 -0.5086 -0.3247 -0.1797 3.0730

Coefficients:

Estimate Std. Error z value Pr(>|z|)

account.length 1.002e-03 1.526e-03 0.656 0.51154

area.code -6.652e-04 1.340e-03 -0.497 0.619515

phone.number -2.973e-06 5.868e-05 -0.051 0.959595

international.plan1 2.109e+00 1.509e-01 13.973 < 2e-16 \*\*\*

voice.mail.plan1 -2.080e+00 5.817e-01 -3.577 0.000348 \*\*\*

number.vmail.messages 3.834e-02 1.824e-02 2.102 0.035583 \*

total.day.minutes 1.368e-02 1.129e-03 12.116 < 2e-16 \*\*\*

total.day.calls 5.497e-03 3.086e-03 1.782 0.074813 .

total.eve.minutes 7.935e-03 1.227e-03 6.465 1.01e-10 \*\*\*

total.eve.calls 1.386e-03 3.125e-03 0.443 0.657457

total.night.minutes 3.681e-03 1.213e-03 3.034 0.002417 \*\*

total.night.calls 2.171e-04 3.137e-03 0.069 0.944843

total.intl.minutes 8.646e-02 2.285e-02 3.783 0.000155 \*\*\*

total.intl.calls -1.128e-01 2.729e-02 -4.132 3.60e-05 \*\*\*

number.customer.service.calls 6.054e-01 5.602e-02 10.807 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

pred\_logit

**Confusion Matrix**

0 1

0 1393 50

1 168 56

**Accuracy : 0.8692**

95% CI : (0.8521, 0.8851)

No Information Rate : 0.9364

P-Value [Acc > NIR] : 1

Kappa : 0.277

Mcnemar's Test P-Value : 2.295e-15

Sensitivity : 0.8924

Specificity : 0.5283

Pos Pred Value : 0.9653

Neg Pred Value : 0.2500

Prevalence : 0.9364

Detection Rate : 0.8356

Detection Prevalence : 0.8656

Balanced Accuracy : 0.7103

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Decision tree(c5.0)** | **Randam forest** | **Logistic regression** |
| **Accuracy** | **94%** | **96%** | **86%** |
| **FNR** | **33%** | **25%** | **79%** |

As it is clear from the **Accuracy** and **False negative rate** **Randam forest** algorithm is giving the best result so we freeze this model.