Churn reduction

Rohit Raj

2nd November 2018

Contents

**1 Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . .3

**2 Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . ..5

2.1.1 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5.

2.1.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . .9

2.1.3 Feature scaling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . 13

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17

2.2.1 Model evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17

**3 Conclusion**

3.1 Model selection . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19

Chapter 1

Introduction

* 1. 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. 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

Target Variable :

move: if the customer has moved (1=yes; 0 = no)

Chapter 2

Methodology

2.1 Pre Processing

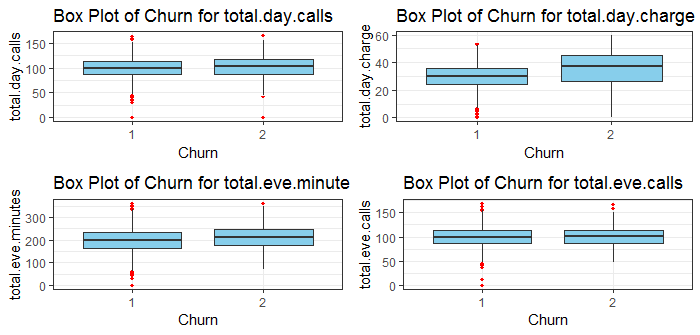
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the

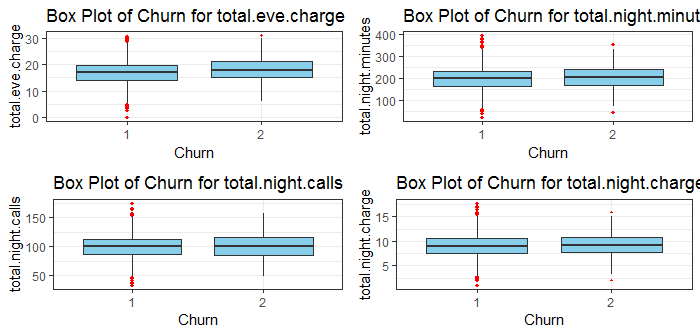
Normal distribution of the variable. We can visualize that by plotting Histogram.

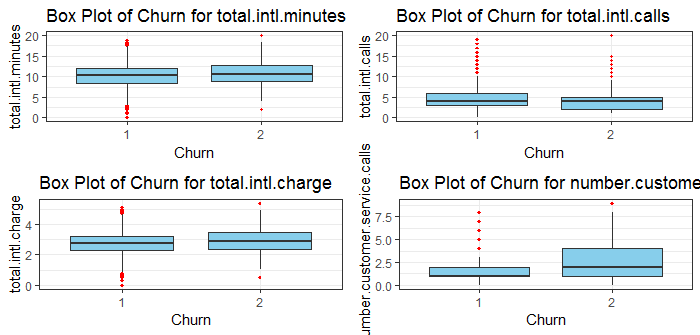
After that we will see that are any missing values there in the data sets or not.

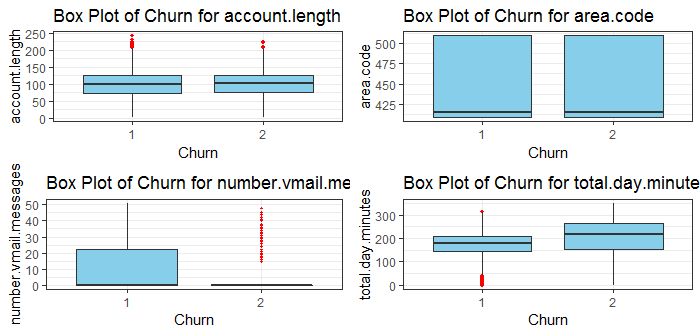
2.1.1 Outlier Analysis

We can clearly observe from these probability distributions that few of the variables are skewed, for example, area code , number.vmail.messages, total intl calls and number.customers.services.calls. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. But we cam see other variables like state, account length, international plan, voice mail plan, total day minutes, total day calls, total eve minute,s, total eve calls, total night minutes, total night calls, total intl minutes, total intl calls, number customer service are normally distributed but also contains few outliers. There fore, we need to remove all those outliers which can badly affect our model’s accuracy. We can see in below fig-

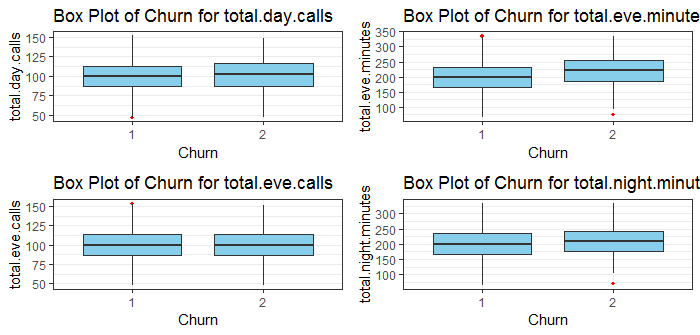


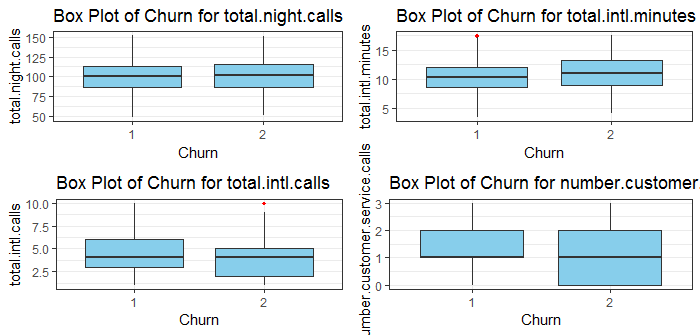


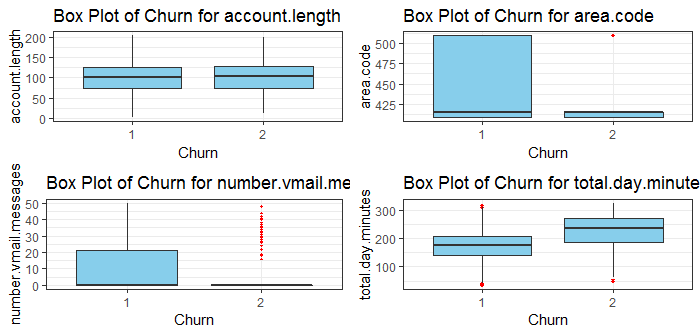




Now we can see the Box plot after removing outliers in below fig-



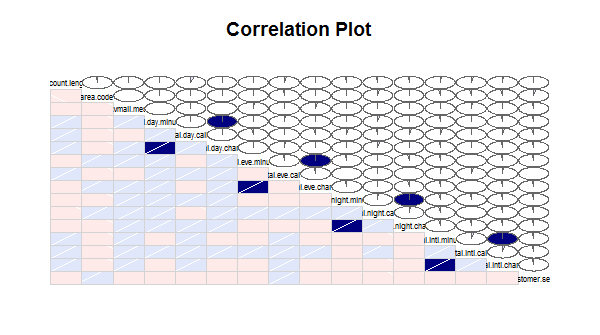




We have removed the outliers value to increase the accuracy of the model like KNN.

2.1.2 Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. There are several methods of doing that. Below we have used correlation plot for numerical data and chi-square test for categorical data.

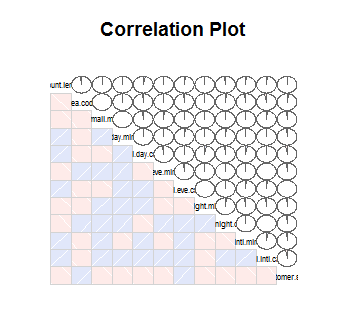


Here, In this correlation plot we can see that in blue box, the variables like” total day charge and total day minutes” are positively correlated with each other that mean both carry the same information.

Similarly , The variables like “total eve charge and total eve minutes” , “total night charge and total night minutes”, and “ total intl charge and total intl minutes”. All these variables corresponding to each other are correlated. This will badly impact on the model accuracy or may be slow down the speed of the model execution.

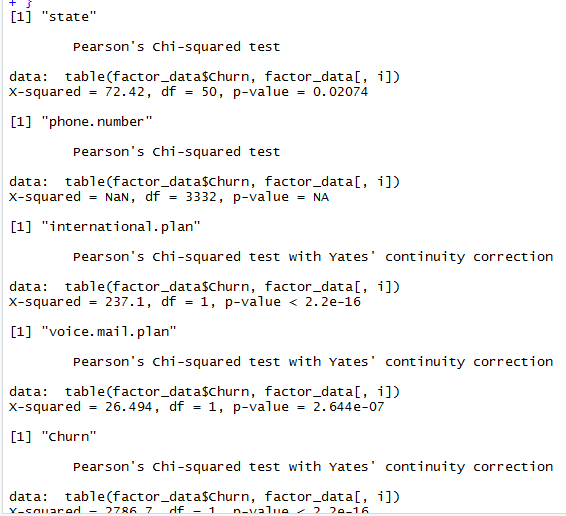
There fore, we need to remove any one side of the corresponding variables. Here we have removed the variables like ( total day charge, total eve charge, total night charge and total intl charge).

After removing these variables we can see again in the correlation plot below-



Now here we will find that there is no any variable which are either positively or negatively correlated. We have now reduced the burden of model having unnecessary variables.

Now look at the chi-square test for the categorical data below to find out the irrelevant variable.



Here, we can see that the p value of phone number is NA that means the variable “phone number” does not describe the target variable so we need to remove that also.

2.1.3 Feature Scaling

Feature scaling is the method used to standardize the range of independent variables or features of data. In data preprocessing it is also called data normalization and is generally performed in the data preprocessing step.

Why do we need feature scaling?

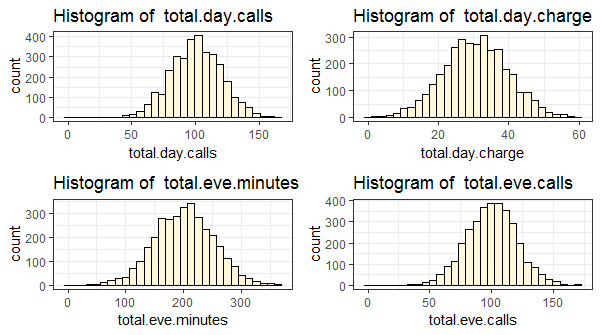
Since the range of values of raw data varies widely , in some machine learning algorithm like in KNN will not work properly without normalizing.

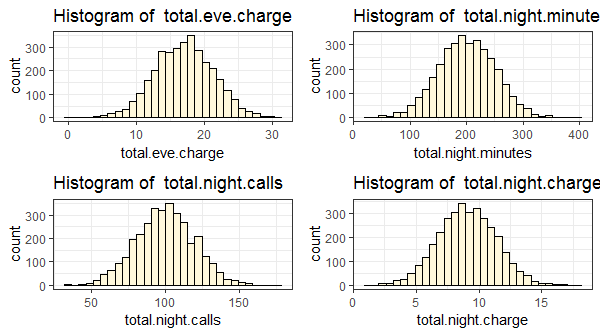
Here we have performed both Normalization and standardization. Normalization for the data which are not normally distributed and the standardization for normally distributed data sets. In below figure we can see most of the variables are normally distributed such as “total day calls”, “total day charge” , “total day minutes”, “total eve calls”, “ total eve charge”, “ total night calls” and so on.

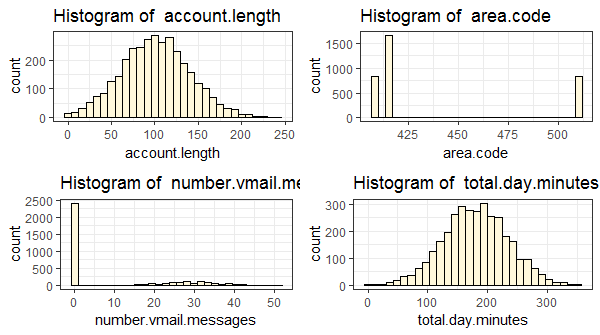
For such variables we have standardized to come at the same scale.

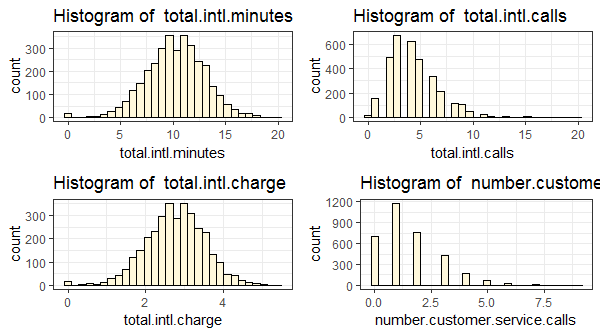
But few of the variables such as “area code”,” number vmail messages”, “total intl calls”,” number customer services calls” are not normally distributed. For these variables Normalization is the best option to make these at the same scale.

We can see in the histogram plot of all these variables below-

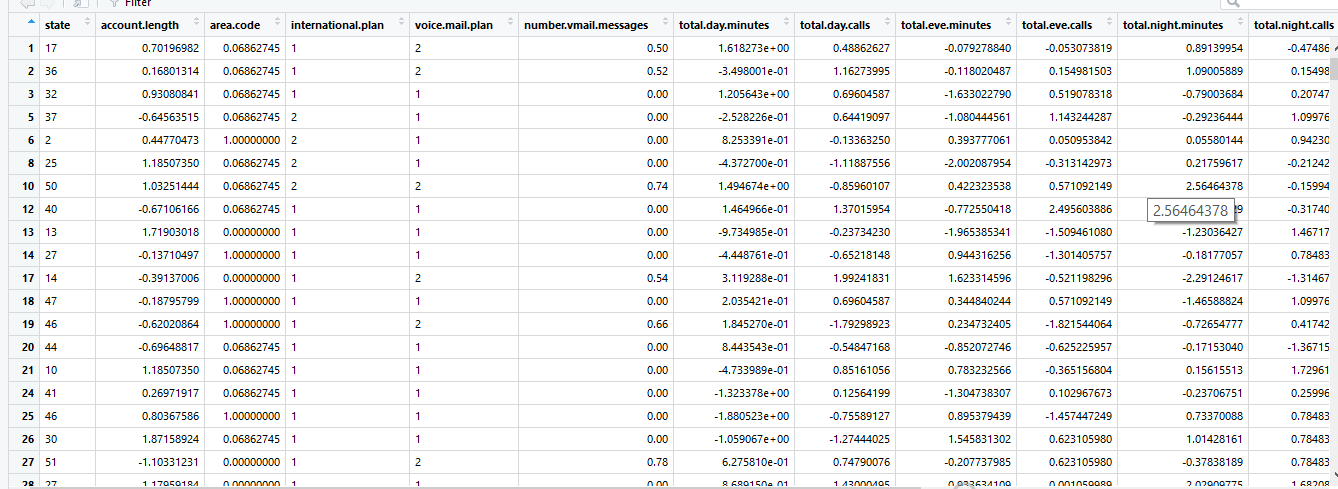


t





After Normalization and standardization we can see our data sets completely changed into the same scale below-



2.2 Modeling

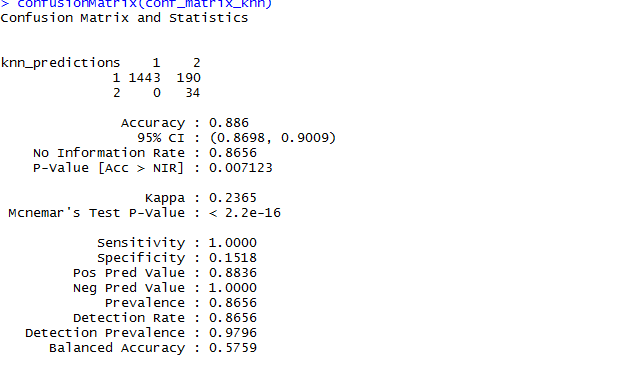
2.2.1 Model evaluation

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. We can compare the models using the best Error matrix also called confusion metrix.

Out of all the classification problem KNN works well in churn reduction data sets.

As we can see below both KNN and Decision Tree Confusion metrix-

For KNN-

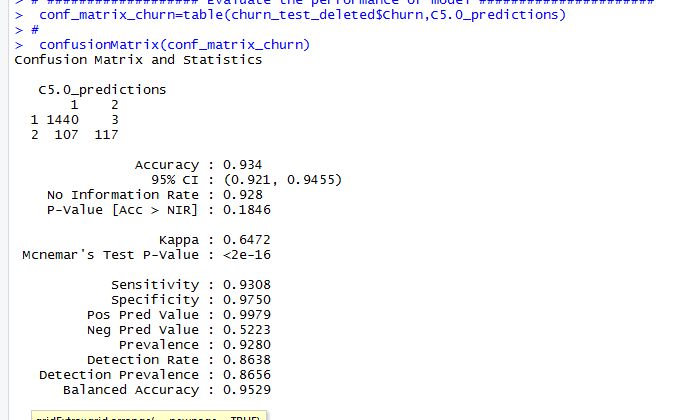


Accuracy =88.6 %

FNR=0 % which is good for the model

FPR= 11.69%

For Decision Tree-



Here, accuracy of Decision tree is higher than the KNN model but look at the FNR and FPR. After analyzing all these evaluation we have con

Chapter 3

Conclusion

3.1 Model selection-

.

Now we will see Here, accuracy of Decision tree is higher than the KNN model but look at the FNR and FPR. After analyzing all these evaluation we have concluded that KNN works well in the churn reduction data sets