Project

Churn Reduction

**Aditya Tiwari**

**Adi138250@gmail.com**

|  |
| --- |
| **Contents** |
| 1. Introduction    1. Problem Statement…………………………………………………………………………………………….1    2. Data…………………………………………………………………………………………………………………1-2 |
| 1. Methodology……………………………………………………………………………………………………….3    1. Pre-Processing…………………………………………………………………………………………………….3   2.1.2 Missing Value Analysis………………………………………………………………………………3-4  2.1.2 Outlier Analysis……………………………………………………………………………………….4-21  2.1.3 Feature Selection…………………………………………………………………………………...21-22 |
| 1. Data Distribution Analysis …………………………………………………………………………………23-25 |
| 1. Modeling………………………………………………………………………………………………………....…….25    1. Model Selection………………………………………………………………………………………………………25       1. Decision Tree Classifier …………………….…………………………………………….…25-26       2. Random Forest Classifier …………………………………………………………………27-28       3. Logistic Regression ………………………………………………….………………….……28-30 2. Conclusion………………………………………………….……………………………………………………………30   Appendix – A – Bar Plots……………...........................………………………………………............31-36  Appendix – B – Python Code……………………………………………………………………………………37-51 |
|  |

1. **Introduction**
2. **Problem Statement**

The objective of this Case is to predict customer behavior. We are providing you a

public dataset that has customer usage pattern and if the customer has moved or not.

It is expected to develop an algorithm to predict the churn score based on usage

pattern.

1. **Data**

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)

Table : Churn Reduction Sample Data (Columns: 1-8)



Table : Churn Reduction Sample Data (Columns: 9-17)



Table : Churn Reduction Sample Data (Columns: 17-21)



Below are the variables present in Churn Reduction dataset

Table: Churn Reduction

|  |  |
| --- | --- |
| **S.no** | **Variables** |
| 1 | state |
| 2 | account length |
| 3 | area code |
| 4 | phone number |
| 5 | international plan |
| 6 | voice mail plan |
| 7 | number vmail messages |
| 8 | total day minutes |
| 9 | total day calls |
| 10 | total day charge |
| 11 | total eve minutes |
| 12 | total eve calls |
| 13 | total eve charge |
| 14 | total night minutes |
| 15 | total night calls |
| 16 | total night charge |
| 17 | total intl minutes |
| 18 | total intl calls |
| 19 | total intl charge |
| 20 | number customer service calls |
| 21 | Churn |

**2. Methodology**

1. **Pre-Processing**

Any predictive modeling requires that we look at the data before we start modeling. We decided to simply remove few variables after loading data set but here we have dropped few variables after correlation test for continuous variables and chi square test for categorical variables. Since our target variable is classified i.e. categorical and few independent variables which are also categorical hence we have applied categorical test.

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 check the presence of missing values in data set.

**2.1.1 Missing Value Analysis**

Missing values Analysis is required to be done so that we can check if there is any missing data. In case data is missing at few places we will impute those missing values by different methods in order to generate appropiate results. In our case we have zero missing values hence no imputation is required. Below table illustrate missing values present in variables of the dataset.

|  |  |  |
| --- | --- | --- |
| **S.no** | **Variables** | **Number of Missing Values** |
| 1 | state | 0 |
| 2 | account length | 0 |
| 3 | area code | 0 |
| 4 | phone number | 0 |
| 5 | international plan | 0 |
| 6 | voice mail plan | 0 |
| 7 | number vmail messages | 0 |
| 8 | total day minutes | 0 |
| 9 | total day calls | 0 |
| 10 | total day charge | 0 |
| 11 | total eve minutes | 0 |
| 12 | total eve calls | 0 |
| 13 | total eve charge | 0 |
| 14 | total night minutes | 0 |
| 15 | total night calls | 0 |
| 16 | total night charge | 0 |
| 17 | total intl minutes | 0 |
| 18 | total intl calls | 0 |
| 19 | total intl charge | 0 |
| 20 | number customer service calls | 0 |
| 21 | Churn | 0 |

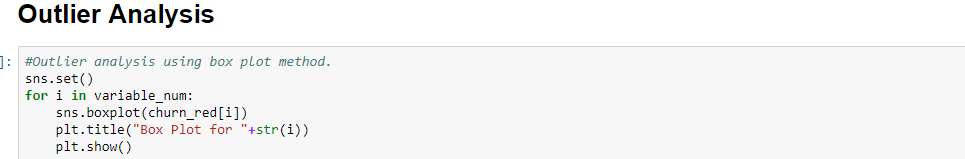
In above table we can clearly see that we don’t have any missing values present in any of the variables of dataset. So, there is no imputation is required and we will move to outlier analysis.

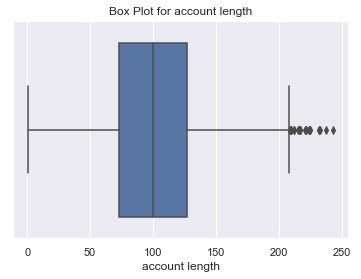
**2.1.2 Outlier Analysis**

The Other steps of Preprocessing Technique is Outliers analysis, an outlier is an observation point that is distant from other observations. Outliers in data can be good and it can be bad as well. Here in our case we there are outliers present. So we will not remove these outliers instead we will be substituting them with other balancing values (such as mean, median or knn method) because we expect them to be relatively random values and replacing them with set values may cause inaccuracy in analysis later.

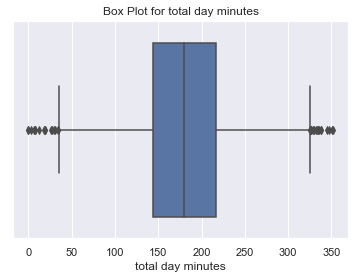
The outlier analysis is done by plotting the box plot. Boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles.

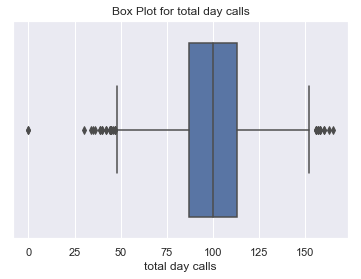
Fig. 1.0

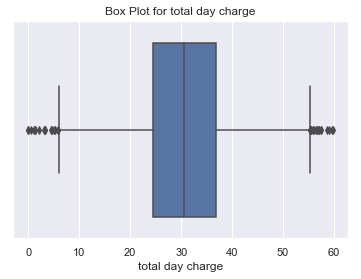


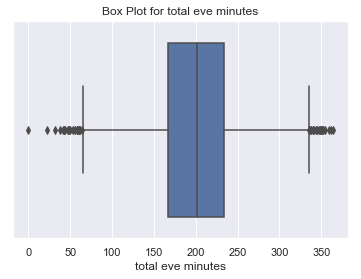


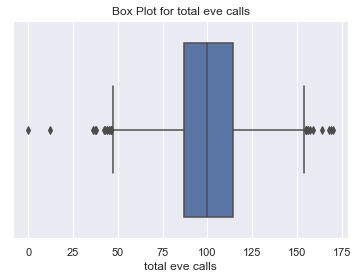


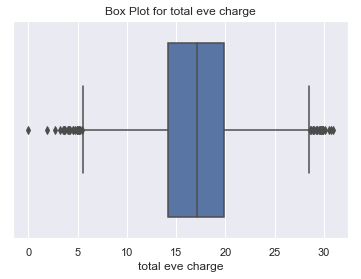


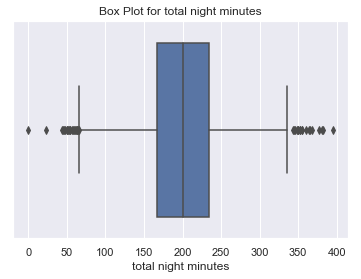


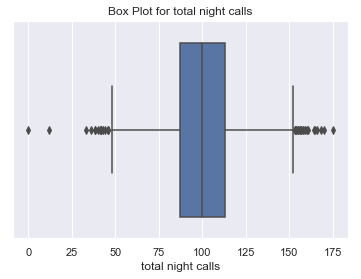


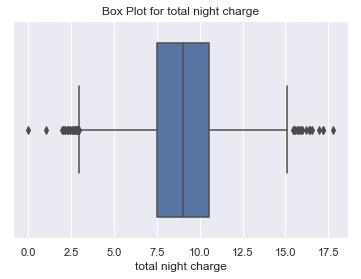


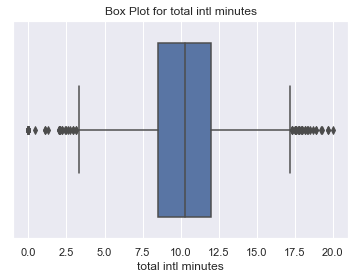


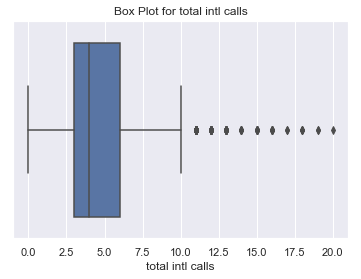


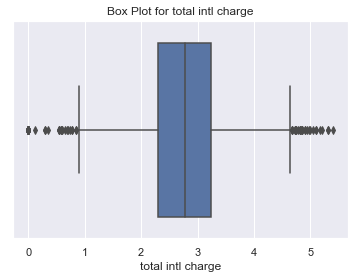


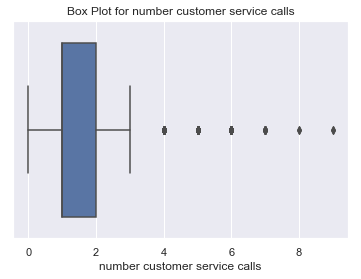






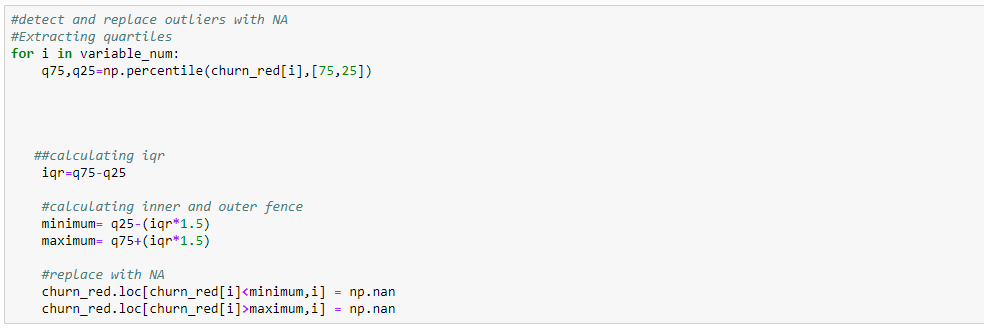


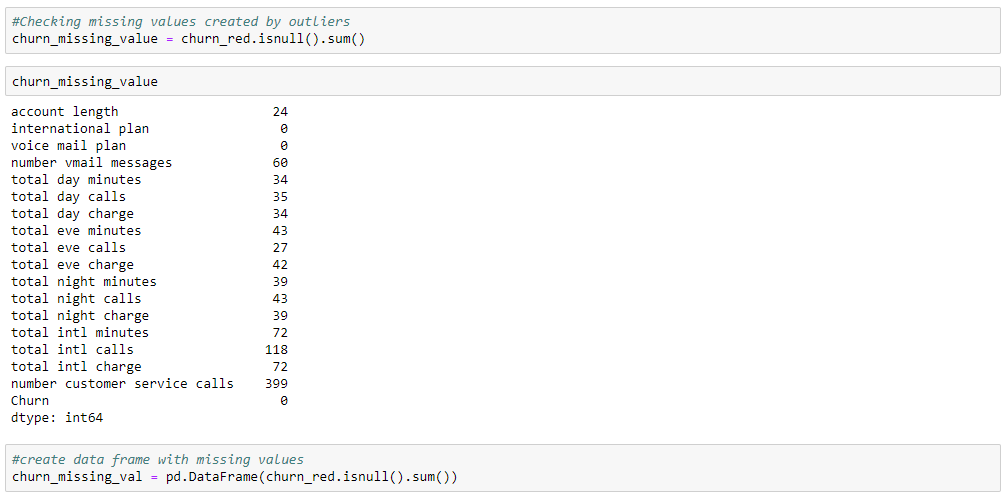


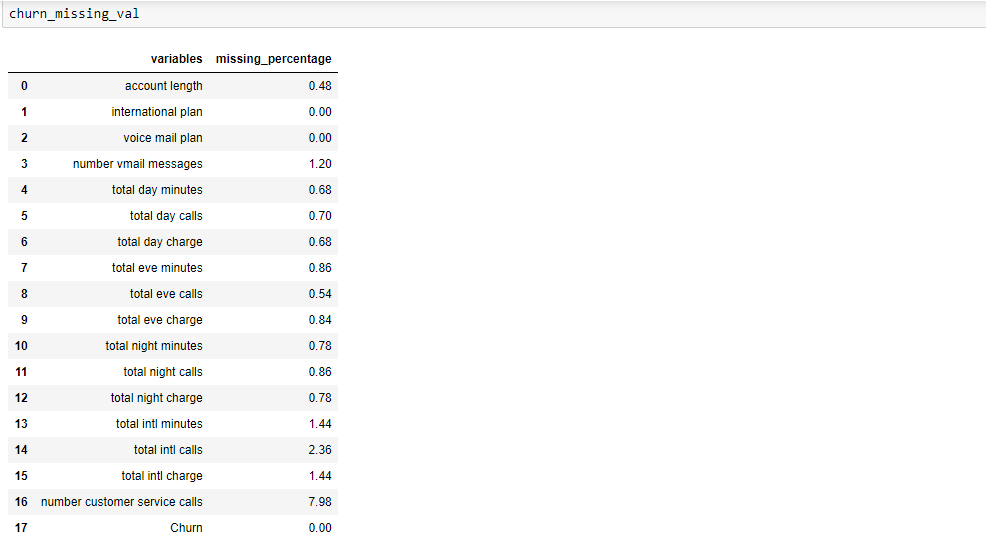
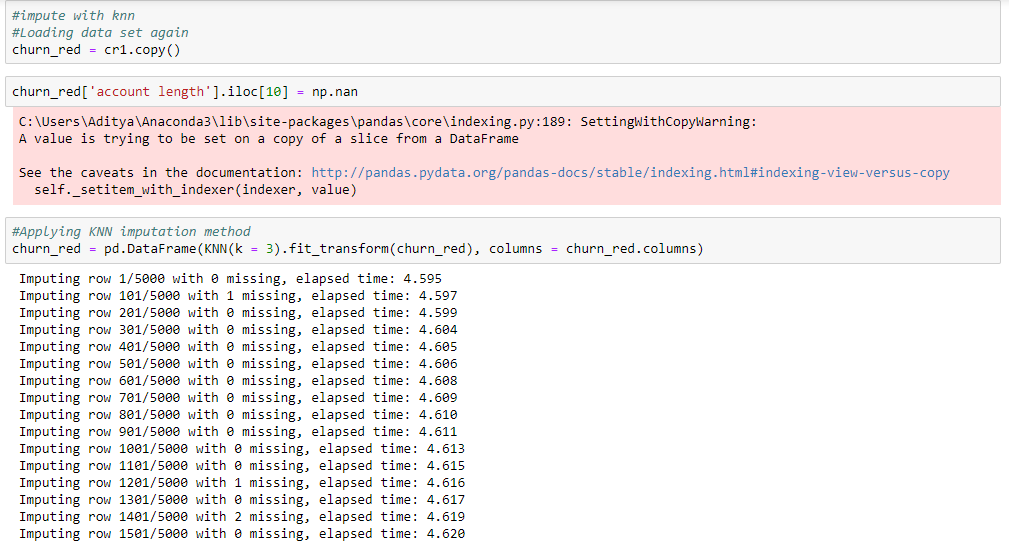


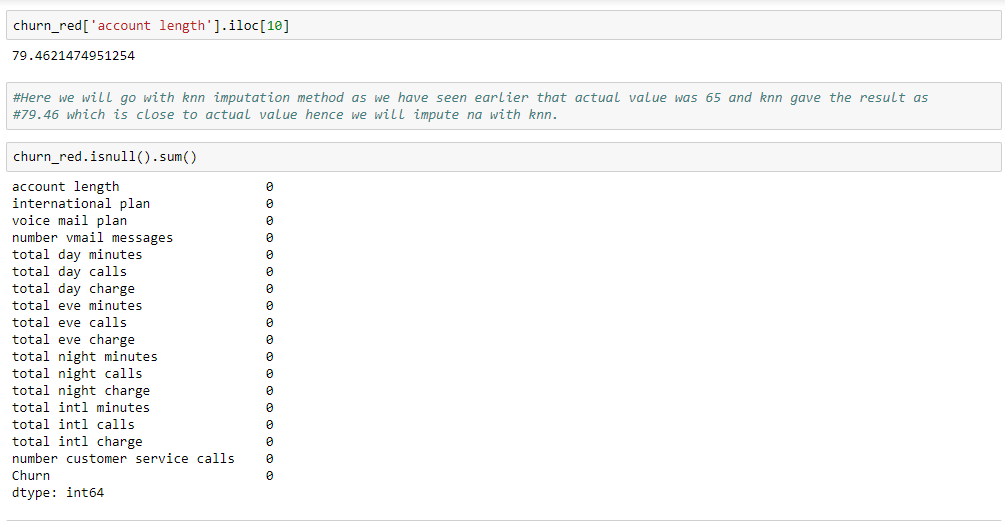
The above boxplots clearly shows the outliers present in the variables. So to treat the outlier we calculated upper and lower quartile then the interquartile range and then we created nan value in place of outliers which are later replaced by knn imputation method.

Outlier treatment

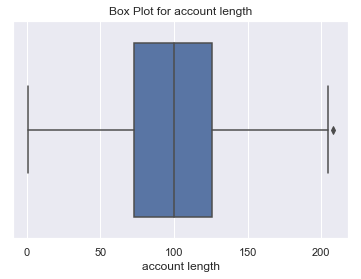


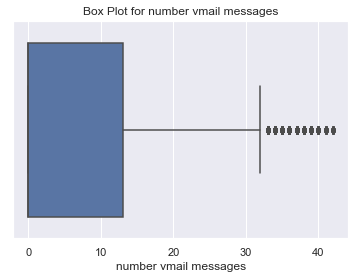


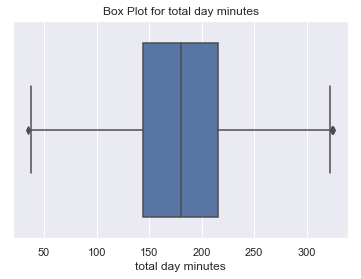
 

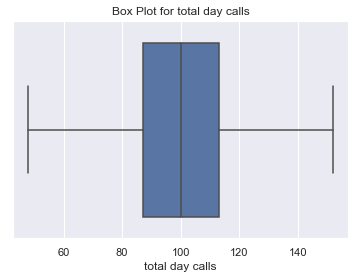


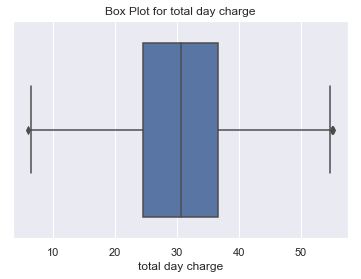
Boxplots after outlier treatment

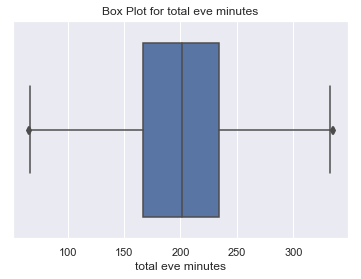


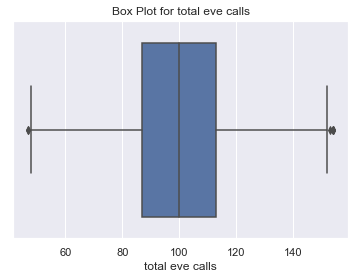


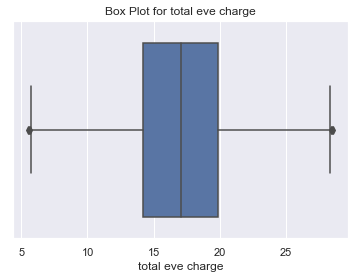


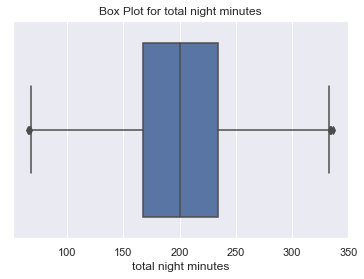


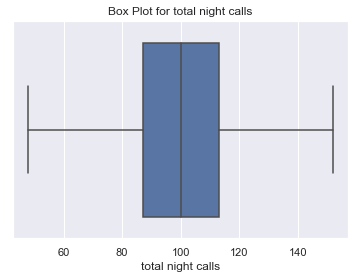


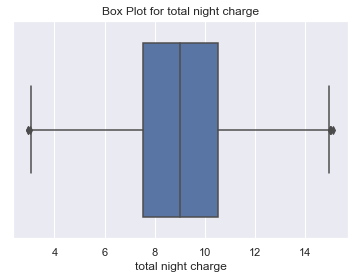


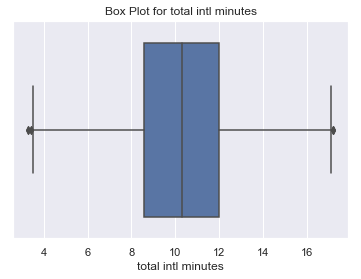


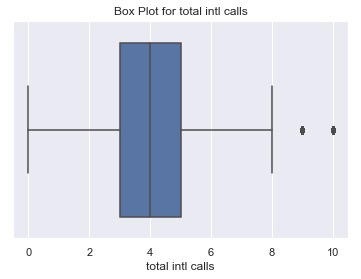


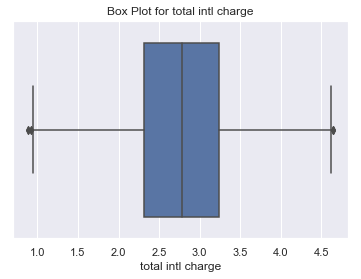


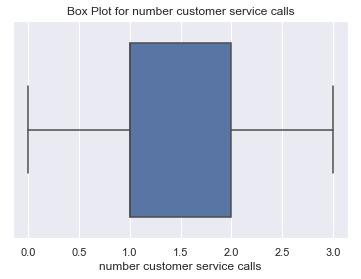












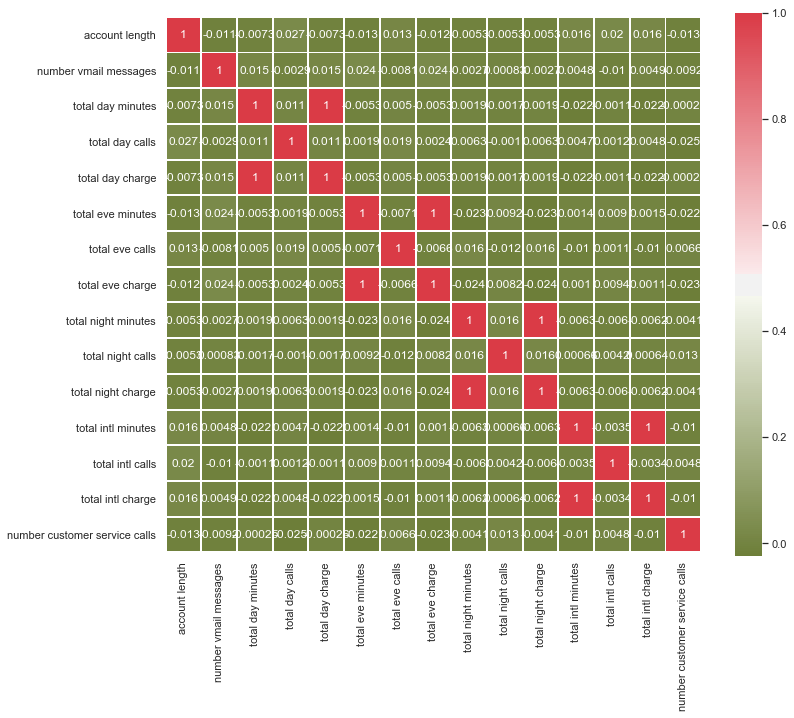
**2.1.3 Feature Selection**

Machine learning works on a simple rule of GIGO i.e. Garbage In Garbage Out. Here garbage refers to the noise or redundant values.

This becomes even more important when the number of features are very large. We need not use every feature at our disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are important. Feature subsets gives better results than complete set of features for the same algorithm or “Sometimes, less is better!”.

We should consider the selection of feature for model keeping in mind that there should be low correlation between two independent variables otherwise there will be problem of multicollinearity.

Fig. 3.0



From above correlation plot we can clearly see that variables like total day minutes and total day charge, total eve minutes and total eve charge, total night minutes and total night charge, total intl minutes and total intl charge have high correlation. It means that we must drop one variable out of two having high correlation. So, in our study here we will drop variables 'total day minutes', 'total night minutes', 'total eve minutes', 'total intl minutes'.

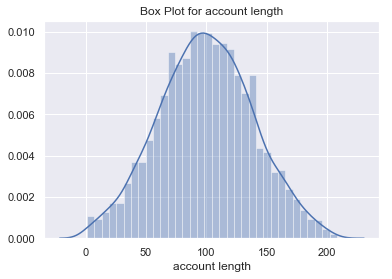
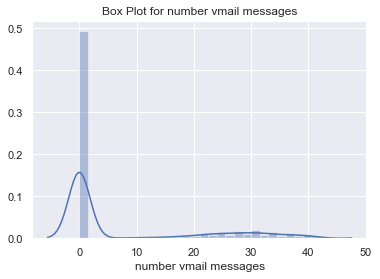
Color dark Red indicates there is strong positive correlation and dark green indicates negative correlation.

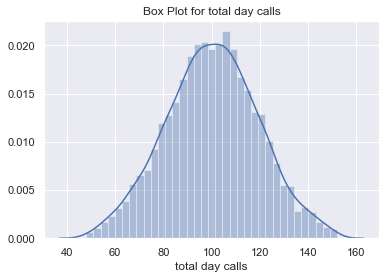
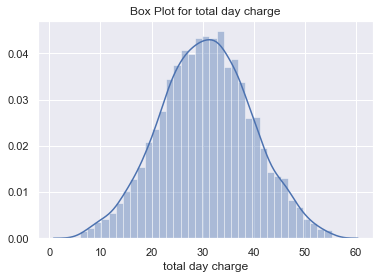
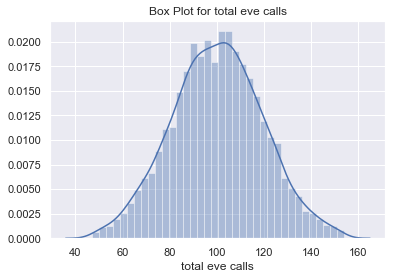
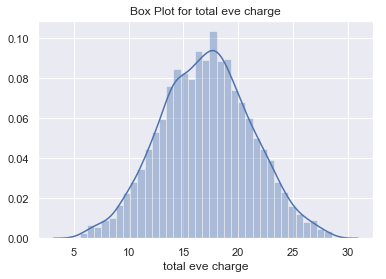
1. **Data Distribution**

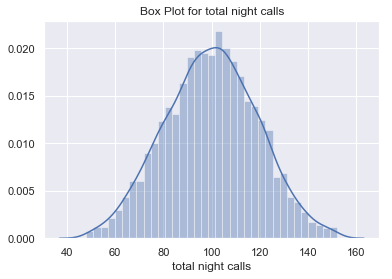
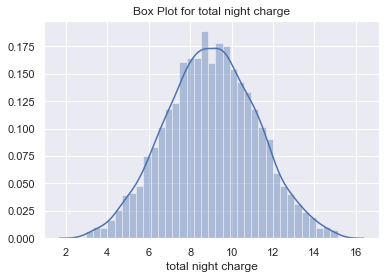
**Checking distribution of variables with help of distribution plot**

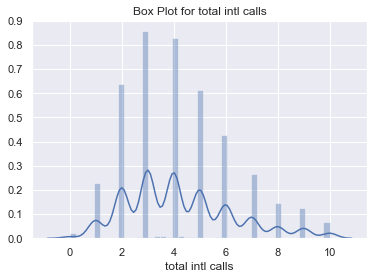
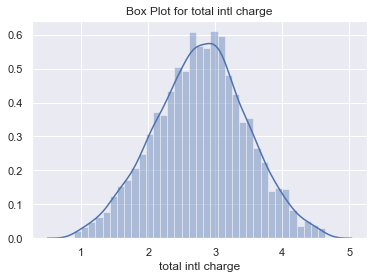
Distribution of continuous predictor variables

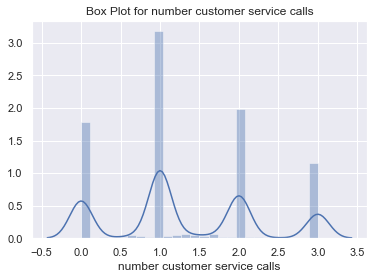
Fig. 4.0



From above distribution plots we can clearly see that data is normaly distributed. Since data is normally distributed hence we will standardise data i.e. making data revolve near about it’s mean point giving more appropiate results.



1. **Modelling**

**4.1 Model Selection**

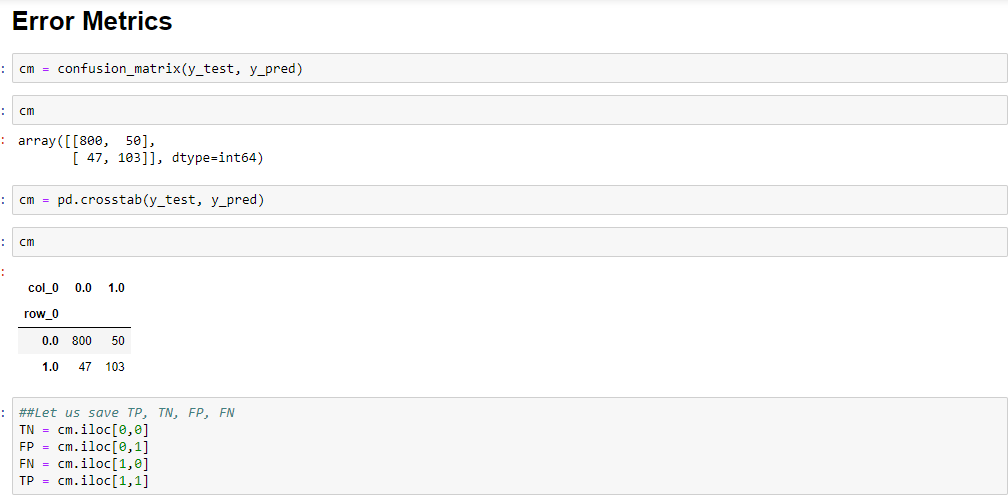
Model Selection is a process of selecting the model which have better accuracy and can work on train and test data. We must select a model where algorithm works well and shows low error rate. We have also used error metrics here; error metrics can be defined as an Error Metric a type of Metric used to measure the error of a forecasting model. They can provide a way to forecast and quantitatively compare the performance of competing models. We made Decision Tree Classifier, Random Forest Classifier and Logistic regression model. When we executed all three models the results were decision tree classifier showed much better results compared to other two.

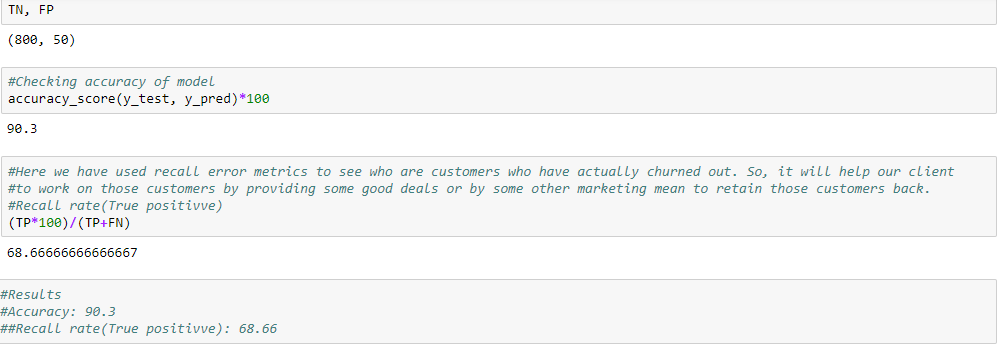
**4.1.1 Decision Tree Classifier**

A tree has many analogies in real life and turns out that it has influenced a wide area of machine learning. In decision analysis, a decision tree classifier can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem.

Decision Tree Algorithm







**4.1.2 Random Forest Classifier**

Random forests classifier is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification) that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

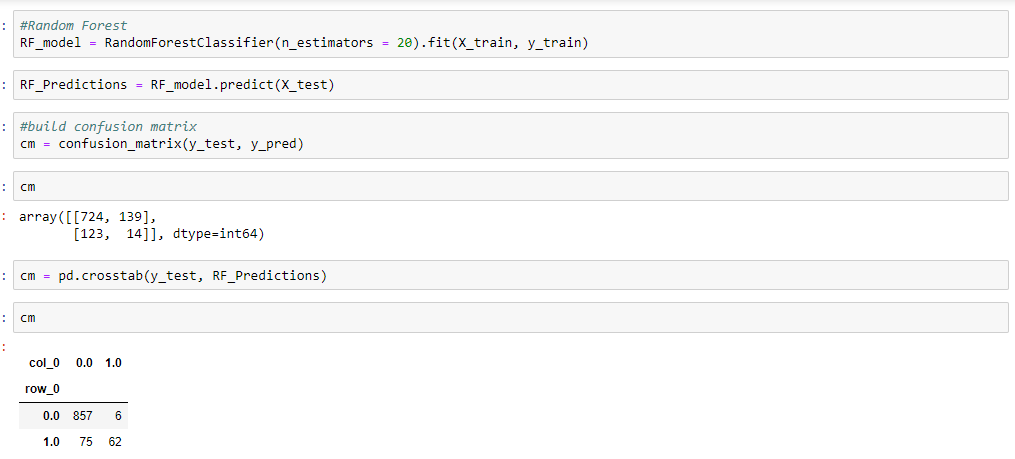
Random forest functions in following way

* Draws a bootstrap sample from training data.
* For each sample grow a decision tree and at each node of the tree

1. Ramdomly draws a subset of variable and p total of features that are available
2. Picks the best variable and best split from the subset of mtry variable
3. Continues until the tree is fully grown.

Random Forest Implementation

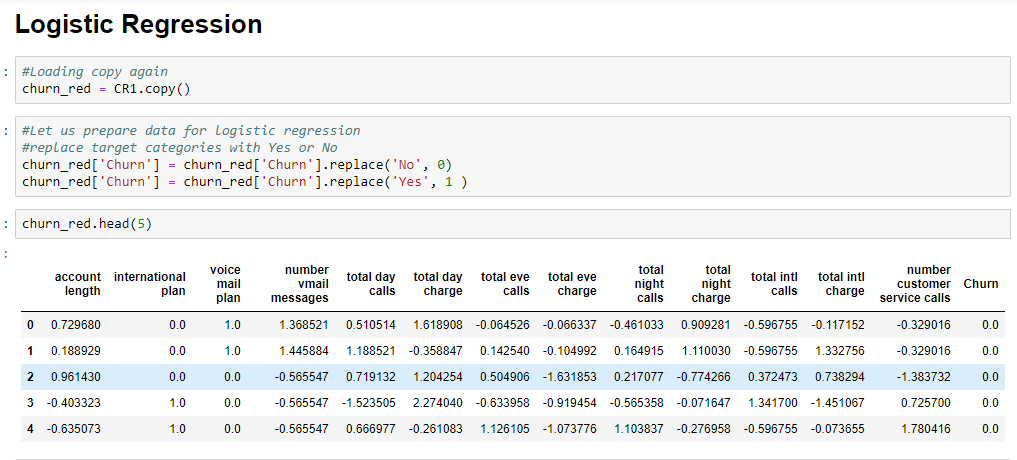


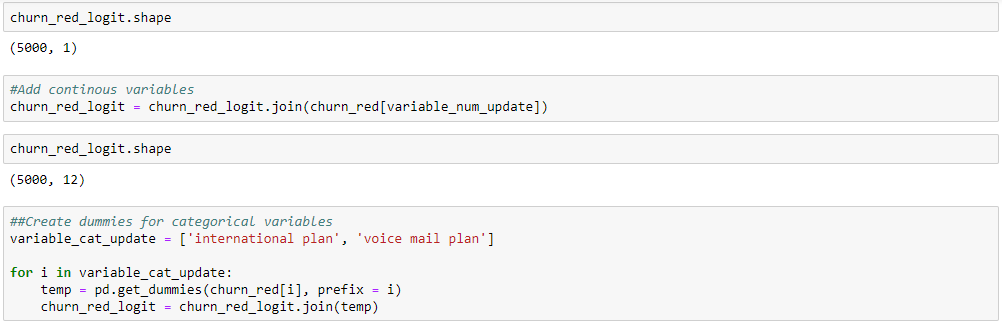


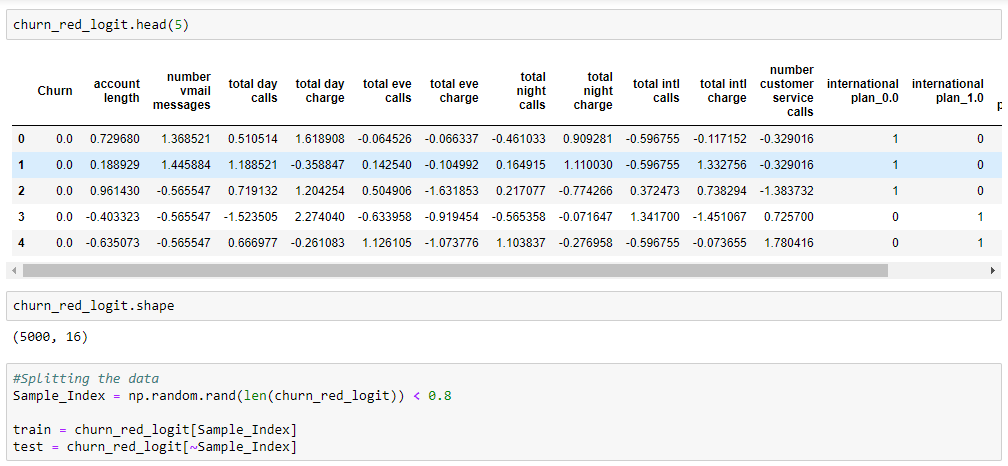


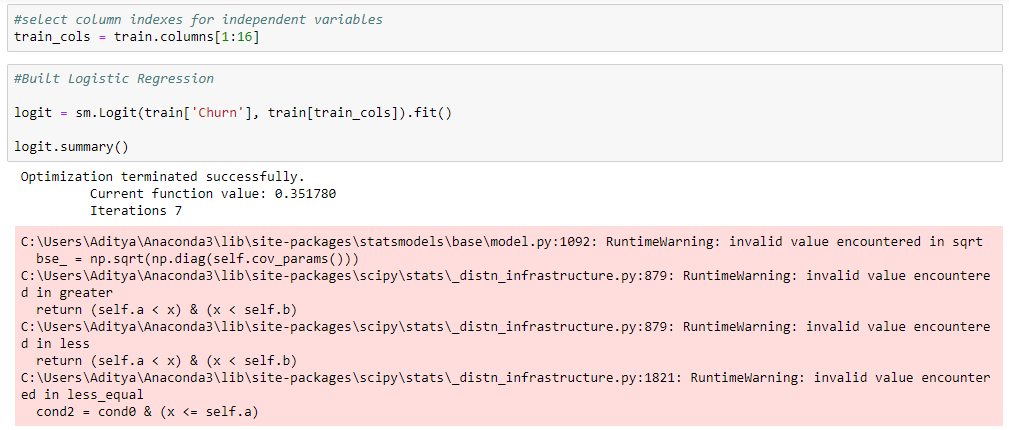
**4.1.3 Logistic Regression**

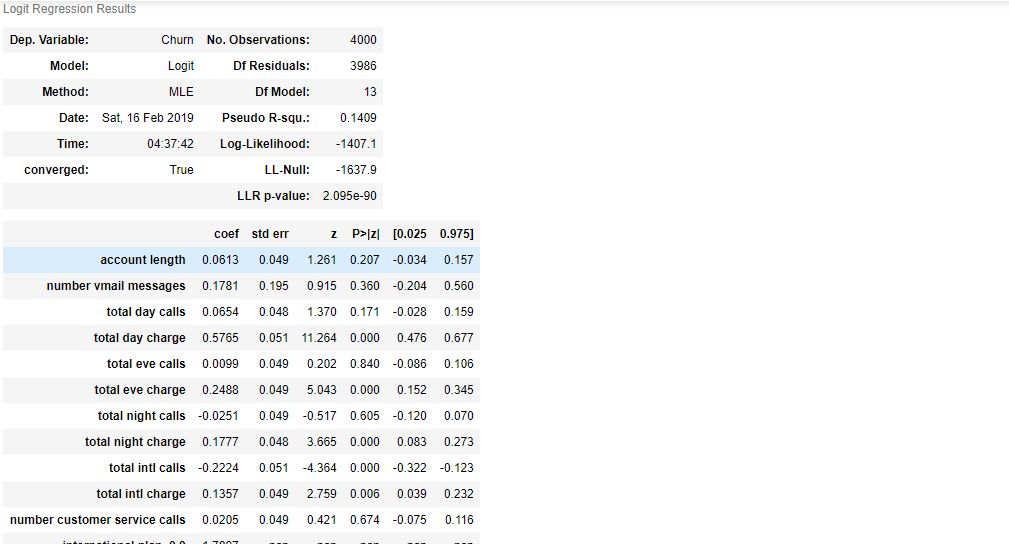
The logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

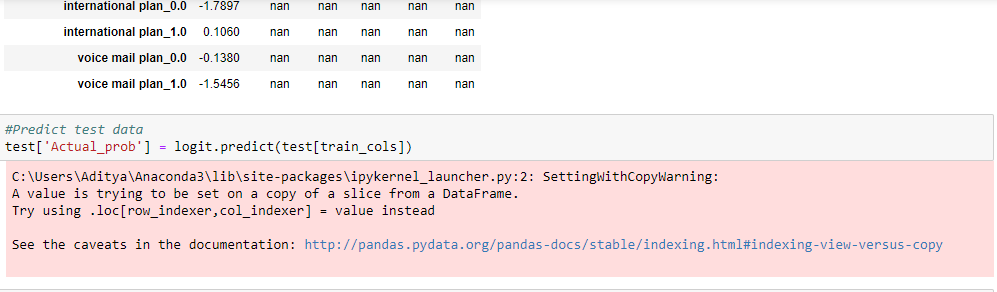


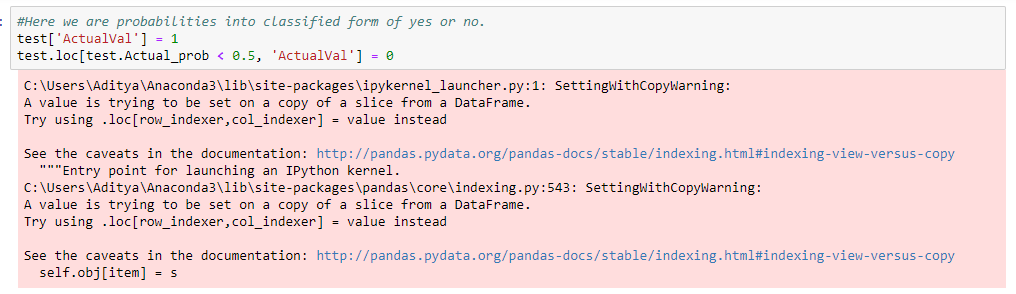




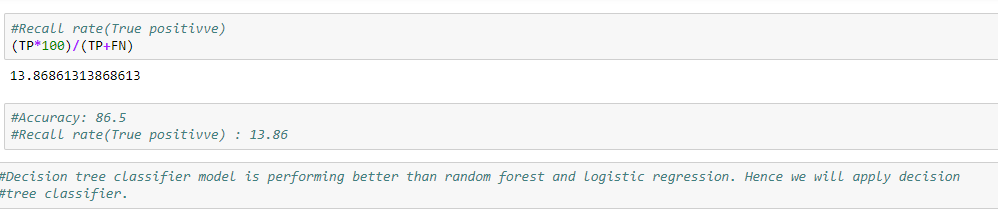










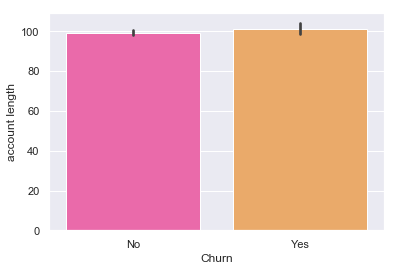


**6. Conclusion**: - For the Churn reduction decision tree classifier Model is appropriate to predict the churn score based on usage pattern. Company should also focus on giving customers some extra minutes or launching some new tariffs to retain their customers.

**Appendix - A**

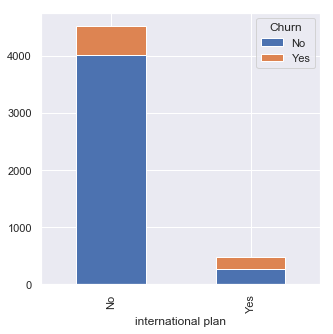
**Bar Plots**

**Bar plot between account length and Churn**



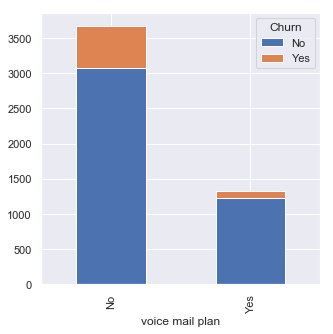
From above bar plot we can see that those customers having more account length have churned out.

**Bar plot between international plan and churn**



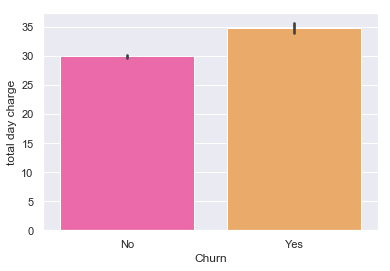
As we can see from above bar plot that whether customers are having international plan or not, still few have churned out. So, we can infer here that there might be some other reasons as well because of which few customers have churned out even they were not having the international plan.

**Bar plot between voice mail plan and churn**

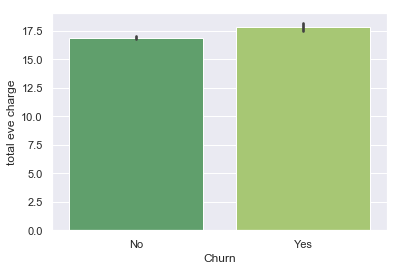


From above plot we can clearly see that customers not having voicemail plan have even churn out. So it’s like same as international plan and company should look after it.

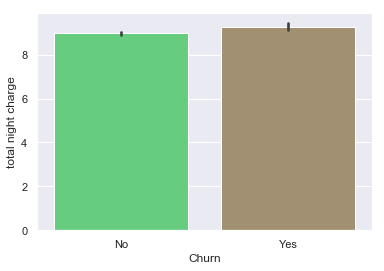
**Bar plot between total day charge and churn**



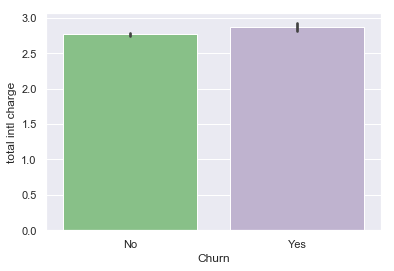
**Bar plot between total eve charge and churn**



**Bar plot between total night charge and churn**

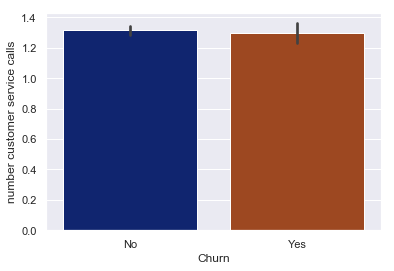


**Bar plot between total intl charge and churn**



From above plots we can clearly see that there is a churn where charges are more. So, company should do something about it either by introducing some new tariffs or through providing some extra minutes to retain their customers.

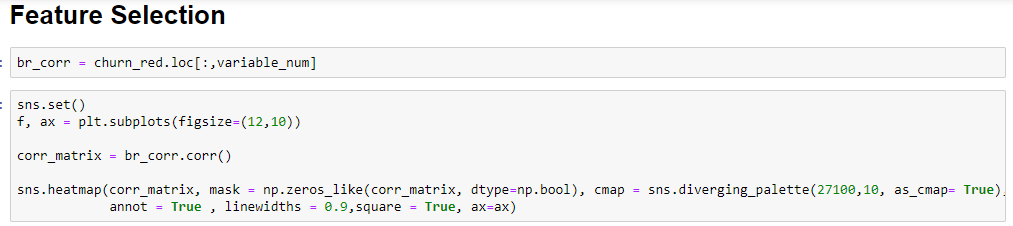
**Bar plot between number customer service calls and churn**



In above bar plot we can clearly see and infer that the number of customer service call has resulted in a no to the customer churn.

**Appendix- B - Python Code**

**Fig 3.0 Python Code**



**Complete Python File**

