Churn Modelling

Vinisha   
Post Graduation Program in Data Science  
Praxis Business SchoolBangalore, Karnataka- 560102  
vinisha.maurya@gmail.com

***Problem Statement*** *— The Company wants to understand the factors affecting customer churn*. The analytics done to identify the causes of churn can help company to fix problems and retain customer base and make profit.

Keywords—Exploratory data analysis, Logistic regession.

# Introduction

Churn for any company is the most important factor. Proper control over churn rate can help company to grow exponentially. If customer churn is not maintained properly can lead to downfall of the company. So, every company should focus on all the factors that cause churn, and try to control them to reduce churn and increase customer base.

Data set: Churn Modelling

Steps involve:

1. Data description
2. Exploratory data analysis
3. Logistic model building
4. Validating model
5. Interpretation

# Data DEscription

Dataset consists of:

1. 11 variables and 3333 observations.
2. 11 Numerical variables.
3. Target Variable: ‘Churn’.
4. No missing values.

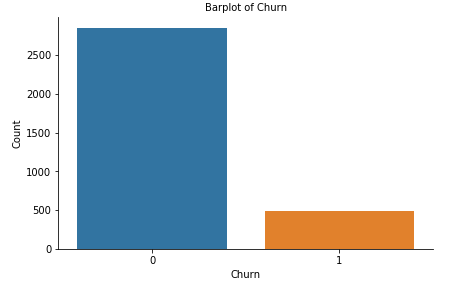
|  |  |
| --- | --- |
| **Variables** | Description of Variables |
| Churn | 1 if customer cancelled service, 0 if not |
| AccountWeeks | number of weeks customer has had active account |
| ContractRenewal | 1 if customer recently renewed contract, 0 if not |
| DataPlan | 1 if customer has data plan, 0 if not |
| DataUsage | gigabytes of monthly data usage |
| CustServCalls | number of calls into customer service |
| DayMins | average daytime minutes per month |
| DayCalls | average number of daytime calls |
| MonthlyCharge | average monthly bill |
| OverageFee | largest overage fee in last 12 months |
| RoamMins | average number of roaming minutes |

# Exploratory data Analysis

## Univariate Analysis

Univariate analysis is done to check the distribution. There are two types of data, categorical and numerical data. For categorical bar charts have been plotted and for numerical data histogram and box plots have been plotted.

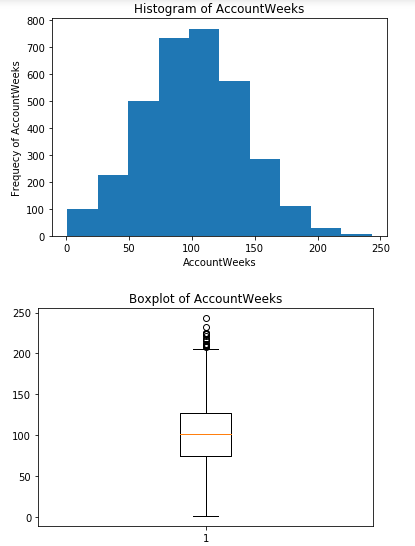
## Target Variable (churn):



* The 1 : 0 ratio is 1:5. So, Data balancing is not required.
* Distribution is Yes – 85.5% and No – 14.5%.

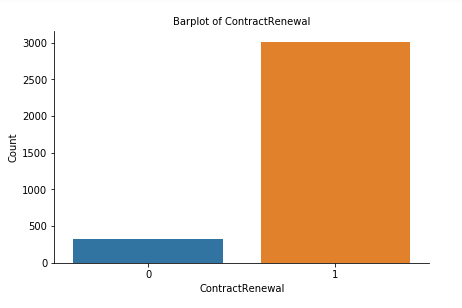
1. *Account Weeks:*

Age is normally distributed with some outliers.



The outliers may is obvious because some customers may be loyal to the company, so they can have more no. of active account weeks

1. *Contract Renewal:*

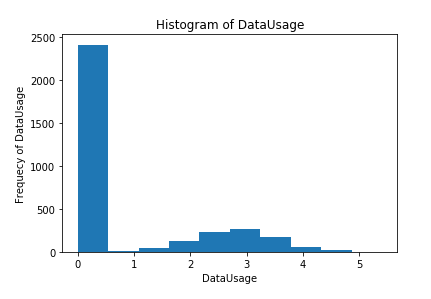


The contract renewal have two values- 1 for recent renewal for contract by the customer and 0 for no contract renewal by the customers.

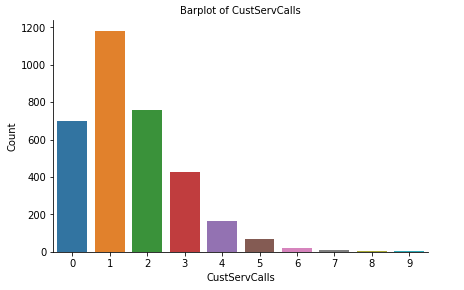
Clearly the distribution is imbalanced. It seems the people are more interested in renewal of contact with the company. This can be plus point for the company.

1. *Data Usage:*

Data usage has a very interesting distribution. It is clearly visible that the data usage is maximum in the 0.5 range. After that it has normal distribution with high standard deviation.

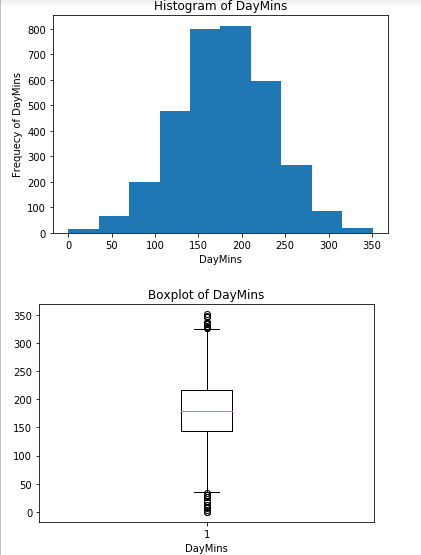


1. *Customer Service Calls :*



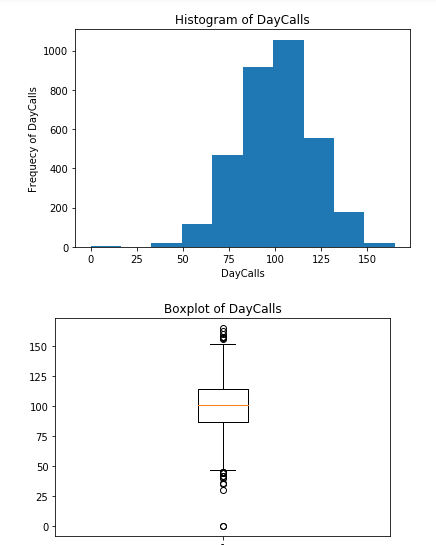
The customer service calls is higher for no. of calls limited to 1, 2, and 4.

1. *Day mins:*



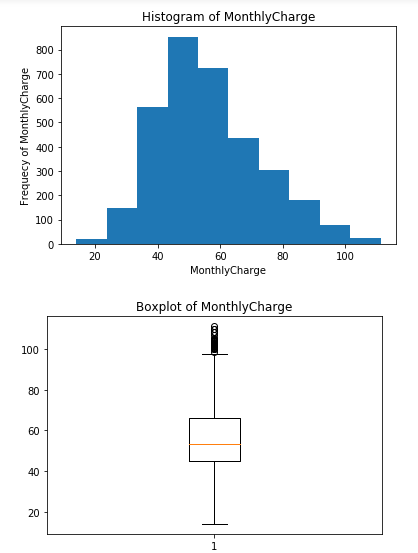
The average day time mins per month is normally distributed with some outlier. The outliers are obvious because average day time min per month can vary from person to person as some may use it for business purpose and at lower end can be students.

1. *Day Calls:*



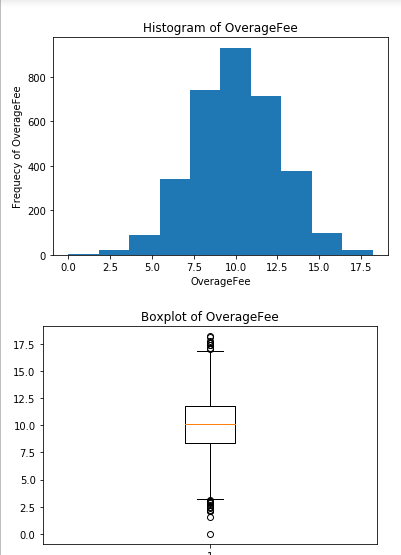
Average number of day time call is slightly left skewed with some outliers.

1. *Monthly Charges:*



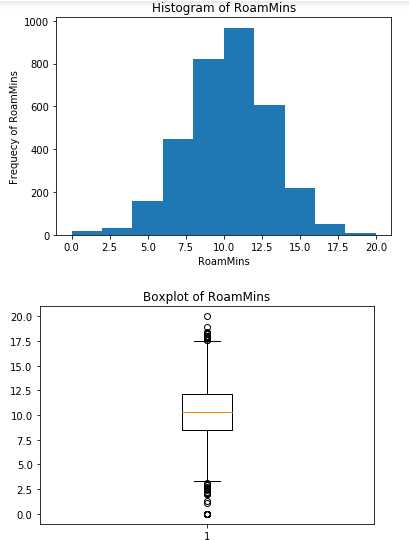
Average monthly bill is slightly right skewed and have some outliers towards right end that can be of business officials.

1. *Overage Fee:*



Overage fee in last 12 months is normally distributed with some outliers.

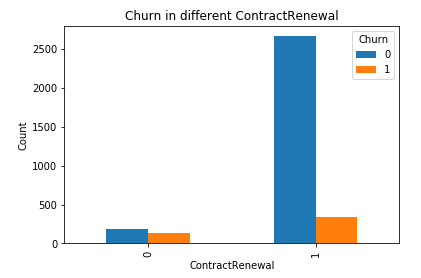
1. *RoamMins:*



Average no. of roaming no. is normally distributed with some outliers.

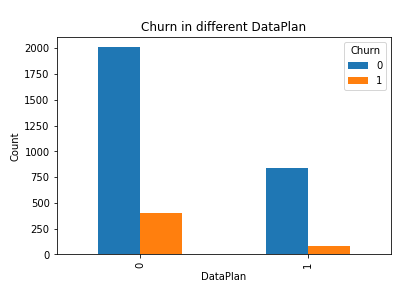
## Bivariate Analysis:

1. *Churn in different Contract Renewal:*



This is very clear that the customers who have not renewed their contract are more likely to churn.

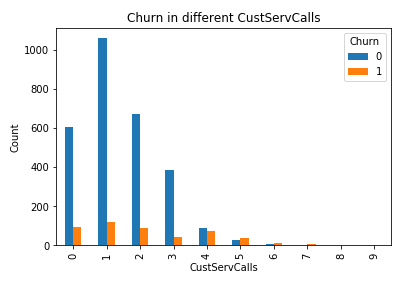
1. Churn in different Data plan:



Customers are more likely to churn if they have data plan.

This is very unlikely. It indicates that there is some problem with Data Plan that needs to be checked.

1. Churn in different customer service calls:

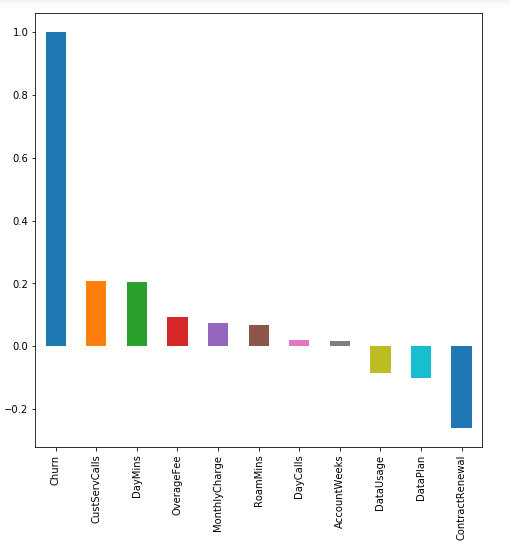


As the customer service calls to customers increases, the churn rate increases.

This is a very useful observation that more customer service calls irritates customers and leads to churn.

This area is something the company should focus to stop churn rate.

1. Correlation Plot:

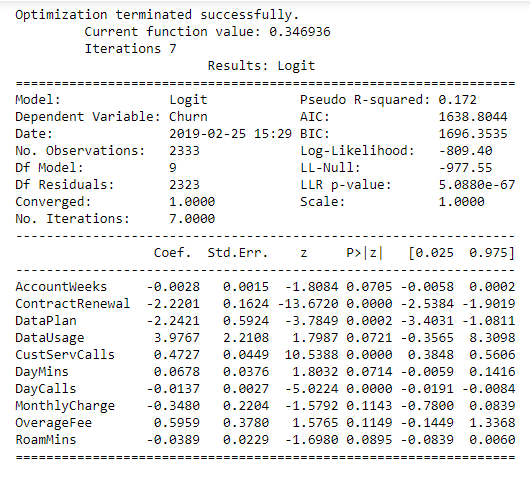


1. Here it is clearly visible that Contract Renewal, Data plan, Data Usage are inversely related to Customer churn.
2. Account weeks, Day Calls have almost no correlation with churn.
3. Customer Service Calls, Day Mins, Overage fee and Roam mins have high correlation with customer churn.

# Logistic model building

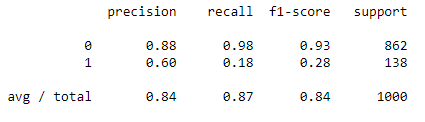
Steps:

1. Splitting X and Y dataframe with response variable and target variables respectively.
2. Train and Test split in 70:30 ratio.
3. Used statsmodel.api to build Logit function: This is done to check the variable importance based on p-value.



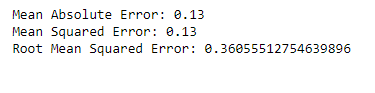
1. Validating model.
   1. Logistic Regression Model before balancing the data:

Accuracy of model: 0.87



Since Logistic model is predicting 0’s properly, so accuracy is not a very good measure to validate the model.

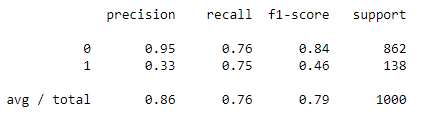
Since, sensitivity is very low for this model. The model is not predicting the target variable accurately.



RMSE is very high, accepting this RMSE would be a matter of concern for the company.

* 1. Logistic regression model after SMOTE.

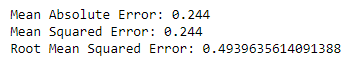
Accuracy of model : 0.76.



Accuracy for this Logistic regression model after balancing the data has reduced, which is a good sign for the model, that means it is predicting 1's properly.

Sensitivity is very high for this model 0.75.

Interpreting.



RMSE is very high 0.49, accepting this RMSE would be a matter of concern for the company.

The model is predicting the target variable better than the previous model but not in an efficient way.

Some other ensemble model like XGboost, Random Forest would predict better than Logistic model.

# Insights and Suggestions to company

1. Attributes Contributing to the Customer Churn in the company:

a. MonthlyCharge p-value 0.1143, OverageFee p-value 0.1149, DataUsage p-value 0.0721, DayMins p-value 0.0714 and AccountWeeks p-value 0.0705, from statsmodel.api logit(), shows that these variables do not have significatant affect on customer churn. So, these attributes can be ignored in building further models. That may give better accuracy.

b. Customer Service calls have highest correlation with customer churn.From bivariate analysis also, as the customer service calls to customers increases, the churn rate increases.This is a very useful observation that more customer service calls irritates customers and leads to churn.This area is something the company should focus to stop churn rate.

c. Contract Renewal also have high negative correlation. It is contributing to customer churn. This is also a very meaningful insight to the company to focus more on contract renewal. Customers with no contract renewal have higher churn rate. So, company should focus on providing offers to increase customers attraction towards the renewing contracts of the plan.

d. Daily Mins i.e average daytime minutes per month is also a key feature which is contributing the customer churn. Company should focus on providing discounts and variety of plans and offers so that customers stick to the companies plan and churn rate can be reduced.