

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT Offline BLR April22 |
| Team members | Rahul Patil  Rajath J K  Maheshreddy P  Ganesh M  Devan Thampi S P |
| Domain of Project | Retail |
| Proposed project title | Telecom Churn |
| Group Number | 5 |
| Team Leader | Rahul Patil |
| Mentor Name | Mr. Jayveer Nanda |

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Topic** | **Page No** |
| **1** | **Project Details** | **4** |
| **2** | **Dataset Information** | **6** |
| **3** | **Data Exploration (EDA)** | **10** |
| **4** | **Base Model** | **27** |
| **5** | **Decision Tree** | **30** |
| **6** | **Random Forest** | **33** |
| **7** | **Boosting Methods** | **38** |
| **8** | **Stack Generalization** | **41** |
| **9** | **Reference** | **43** |

### **PROJECT DETAILS**

### **Overview**

The telecommunication sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators have raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.

Customer churn is a considerable concern in service sectors with highly competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase.

**Industry Review**

**Introduction to domain:**

Telecommunications are the means of electronic transmission of information over distances. The information may be in the form of telephone calls, data, text, images, or video. Today, telecommunications are used to organize more or less remote computer systems into telecommunications networks.

Nowadays, telecom industry faces fierce competition in satisfying its customers. The role of churn prediction system is not only restricted to accurately predict churners but also to interpret customer churn behavior.

To stay competitive, TELCOMs must continuously refine everything from customer service to plan pricing and use the power of highly targeted data analytics in helping the company secure or improve their standing in the highly competitive marketplace.

**Impact in Business:**

Telecommunications is an important tool for businesses. It enables companies to communicate effectively with customers and deliver high standards of customer service. Telecommunications is a key element in allowing employees to collaborate easily from wherever they are located, remote or local.

Telecommunications affects how people connect and do business on a global scale. For businesses, in particular, reliable and timely communication is the lifeblood of your company's brand reputation, productivity, and overall success.

**Problem Statement:**

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenue of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn.

**Dataset Information:**

**Target Variable:**

|  |  |  |
| --- | --- | --- |
| **FEATURE** | **DATA TYPE** | **DESCRIPTION** |
| CHURN | Object | Detecting which customers are likely to leave a service or to cancel a subscription to a service |

**Features Understanding:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **DATA TYPE** | **Description** |
| Customer ID | Integer | Primary key of the record. |
| Churn | Object | Detecting which customers are likely to leave a service or to cancel a subscription to a service |
| Monthly Revenue | Float | Revenue of each Customer |
| Monthly Minutes | Float | Number of Minutes call spoken by Customer |
| Total Recurring Charge | Float | The Charges for the Service |
| Director Assisted Calls | Float | When we call an operator to request a telephone number |
| Overage Minutes | Float | Count of Call used over duration to particular post-paid cell phone plan |
| Roaming Calls | Float | The ability to get access to the Internet when away from home at the price of a local call or at a charge considerably less than the regular long-distance charges. |
| Three-way Calls | Float | A way of adding a third party to your conversation without the assistance of a telephone operator. |
| Dropped Calls | Float | Count of Phone calls gets disconnected somehow from the cellular network. |
| Blocked Calls | Float | Count of Telephone call that is unable to connect to an intended recipient. |
| Un-answered Calls | Float | Count of Calling that an individual perceives but is not currently pursuing. |

|  |  |  |
| --- | --- | --- |
| Received Calls | Float | Number of calls received by the customer. |
| Out bound Calls | Float | Call initiated by the call centre agent to customer on behalf of client to know the target customer behaviour and needs. |
| Inbound Calls | Float | In inbound calls, call-centre or customer-care receives call from customer with issues and questions. |
| Peak Calls in Out | Float | Amount of time period with fewer calls than are handled in a busy period. |
| Call Forwarding Calls | Float | Count of Calls Forwarded by user. |
| Dropped Blocked Calls | Float | Number of VM messages customer currently has on the server. |
| Call Waiting Calls | Float | Duration of call-in waiting period |
| Months In Service | Integer | Number of months customer using service. |
| Unique Subs | Integer | subscription of different networks |
| Active Subs | Integer | subscription of the networks that are active or in usage. |
| Service Area | Object | Network service area |
| Handsets | Integer | Count of Handset with user |
| Handset Models | Float | Count of Handsets are used to Contact one to one. |

|  |  |  |
| --- | --- | --- |
| **Feature name** | **Data Type** | **Description** |
| Age HH1 | Float | User aged below 45 |
| Age HH2 | Float | User aged above 45 |
| Children in HH | Integer | Whether there are Children in House hold |
| Handset Refurbished | Object | Are the handsets refurbished or not |
| Handset Web Capable | Object | Are the handsets capable of internet connectivity |
| Truck Owner | Object | Is the user a Truck Owner |
| RV Owner | Object | Is the user an RV owner |
| Home Ownership | Object | Is the house the user is staying, his own |
| Buys Visa Mail Order | Object | Does the user buy Visa Mail order |
| Responds to Mail Offers | Object | Does the user respond to Mail offers |

|  |  |  |
| --- | --- | --- |
| Opt-out Mailings | Object | Did he opt out of the mail offers sent to him |
| Non-US-Travel | Object | Does the user travel to other countries |
| Owns-Computer | Object | Does he have a computer or not |
| Has-Credit Card | Object | Does he have a credit card or not |
| Retention Calls | Integer | No of Retention Calls |
| Retention Offers Accepted | Integer | Customers accepting retaining the retaining offers given by the company. |
| New Cell phone User | Object | Number of customers buying new cell phone. |
| Not New cell phone User | Object | Number of customers uses existing cell phone |
| Referrals Made by Subscriber | Integer | Referrals made by the existing customer to the other customer. |
| Income Group | Integer | The column talks about the customer saying to which category the customer belongs to. |
| Adjustments To Credit Rating | Integer | Rating Scale |
| Handset Price | Object | Its amount paid by the customer for his cell phone. |
| Made call to retention team | Object | User call to Retention in same company |
| Credit Rating | Object | Credit card user rating (out of 7) |
| PrimzCode | object | Grouping of regions according to users |
| Occupation | Object | Occupation of User |
| Marital status | Object | Marital Status Indicated by Yes/No/Unknown |

**Dataset Information:**

Data is taken from Kaggle (Telecom(churn))

No. of features: 56

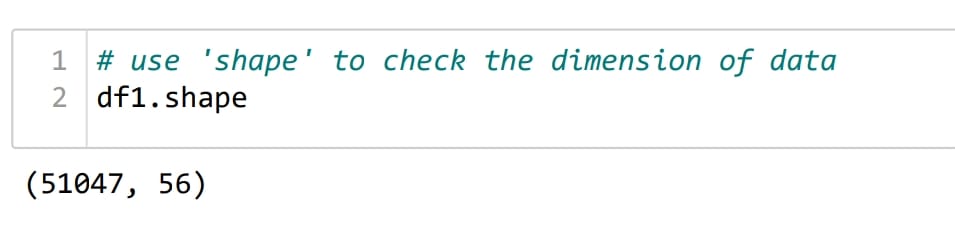
No. of records: 51047

Target Column: churn

Redundant columns: Service area, Customer Id.

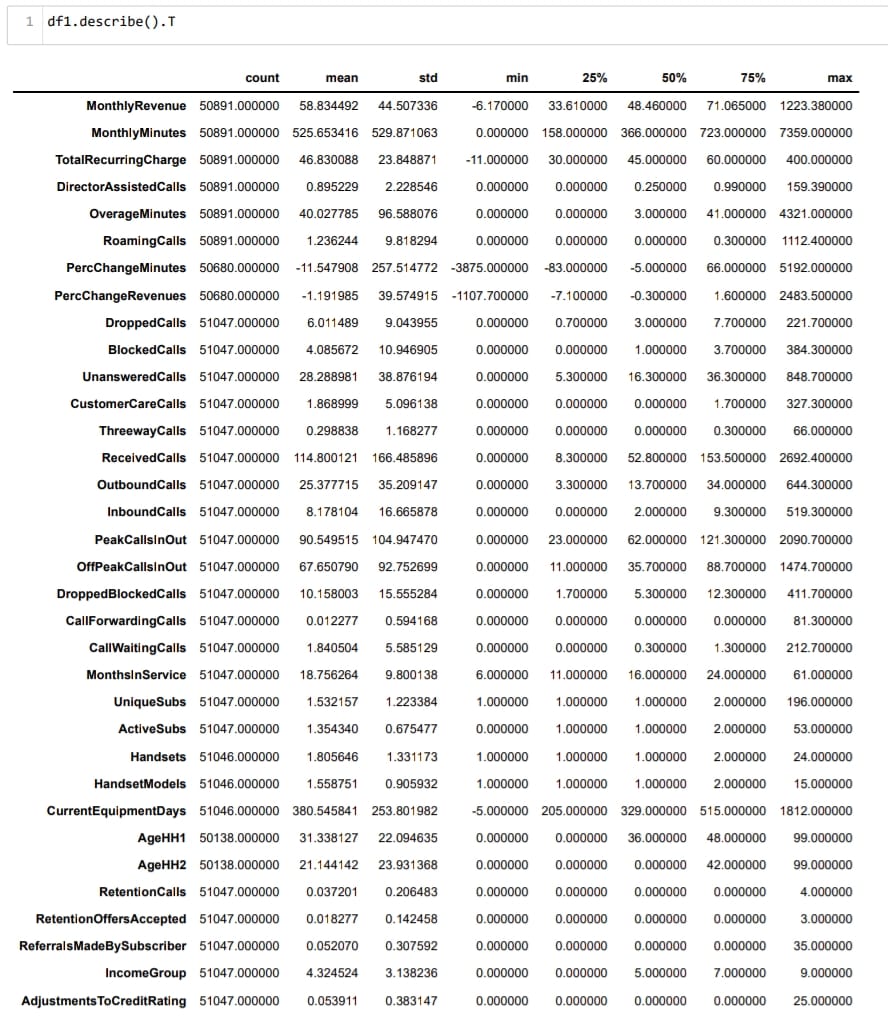
**Understanding the Data:**

Checking Shape of Data:



# **DATA EXPLORATION (EDA)**

**Summary of Dataset:**

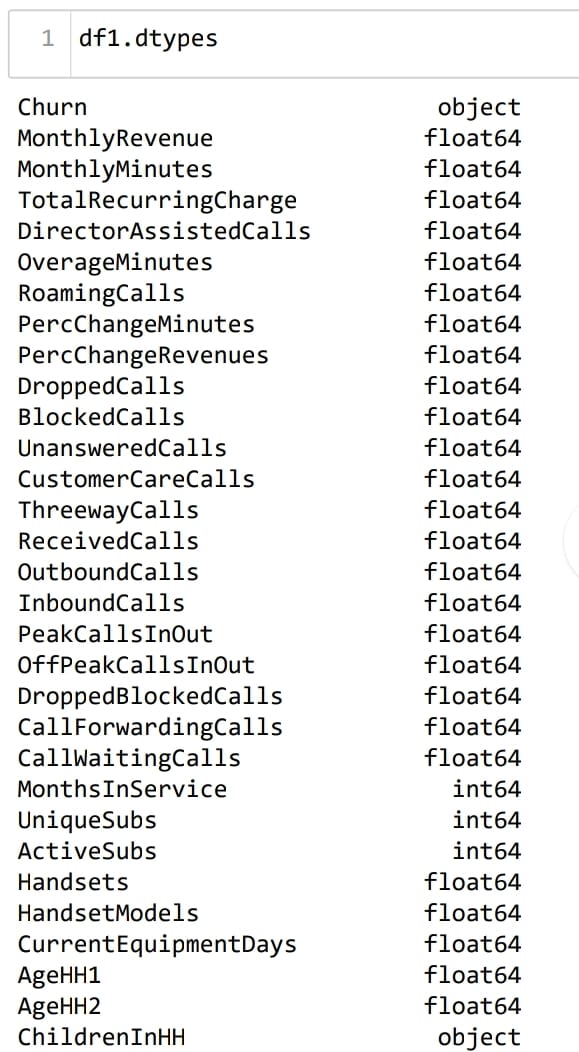


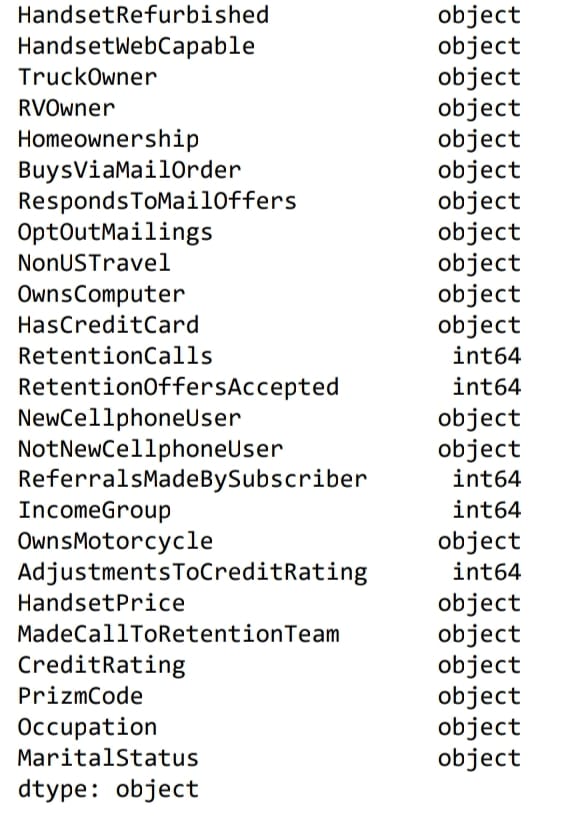
**Interpretation:**

1. Count of all features are not equal so we can say that there are missing values in the Dataset.
2. The difference Between mean and median of each variable is more, so we can say that data is not normally distributed.
3. The difference Between min and max of each variable is more, so we can say that Some of the features also contains potential outliers.

**Check the Data Type:**

**Check the data type of each variable. If the data type is not as per the data definition, change the data type.**



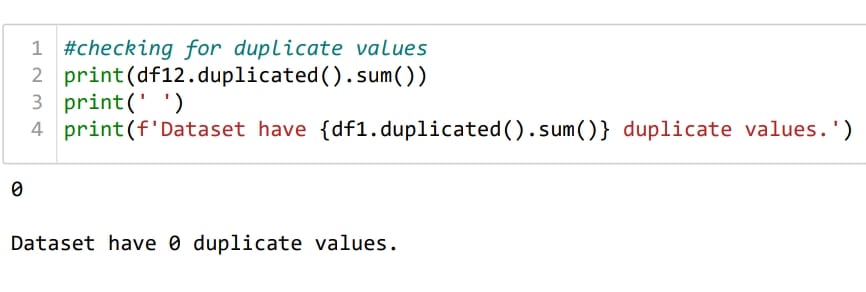
****

**Recheck the Data type and conversions:**

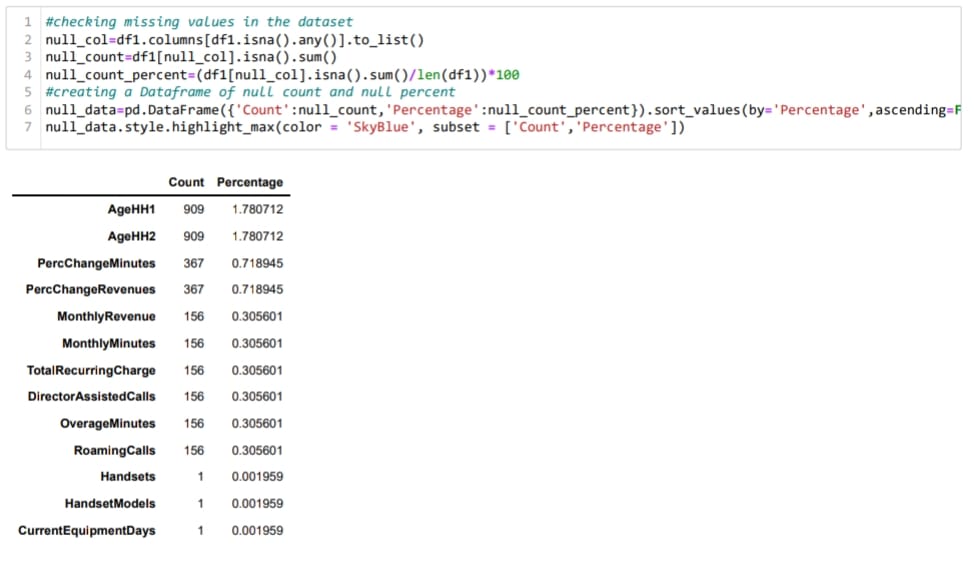
****

**Data Cleaning**

**Duplicate Values Check:**

****

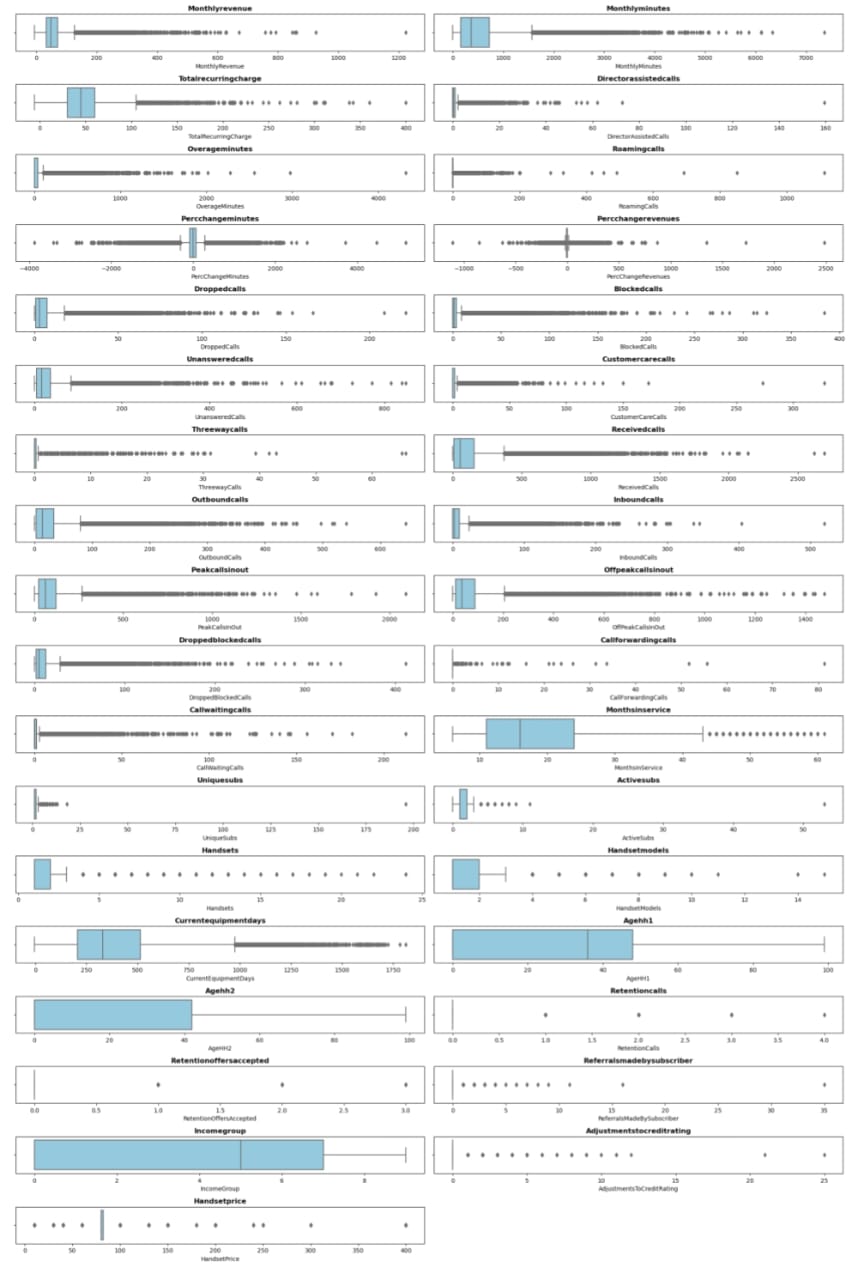
**Missing Values Treatment:**

Missing values plays a prominent role in the dataset. Generally, we can drop the columns or rows depending the percentage of missing values. We can also replace the missing values with optimum values. In order to perform such operations, we will first look into the overall missing values in each column using the below python code. ****

Let us now consider each variable separately for missing value treatment.

### 

**Outlier Analysis:**

****

**Inference:** By Visualizing above boxplot, we can see that all the Features have potential outliers and some features there are extreme values as well.

**Outliers:** Outliers is an observation which deviates so much from the other observations, that it become suspicious that it was generated by different mechanism or simply by error

**Extreme Values:** Extreme Values is an observation with value at the boundaries of the domain

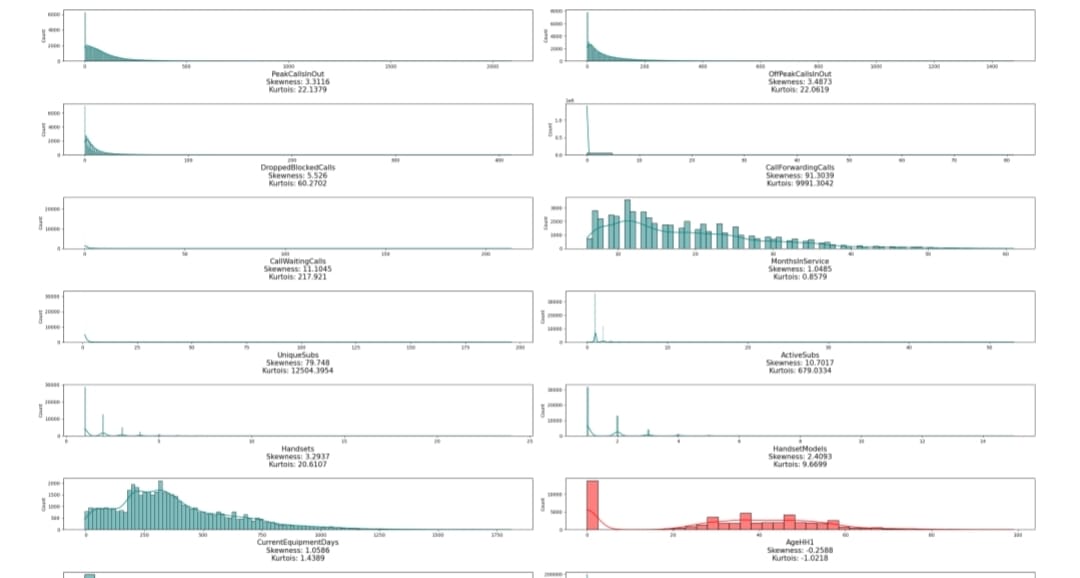
**Reason for outliers exist in the data:**

1. Variability in the Data
2. An experimental measurement errors

**Impact of outliers on Dataset:**

1. It causes various problem during statistical analysis.
2. It effects the mean and standard deviation.

**Skewness Before Transformation:**

****

**Inference:** Here by visualizing dist plot we can see that the Features plotted in Teal colour are positively skewed and Features plotted in red colour are Negatively Skewed.

--: To reduce the impact of skewness we can use various transformation techniques here we are using box cox transformation

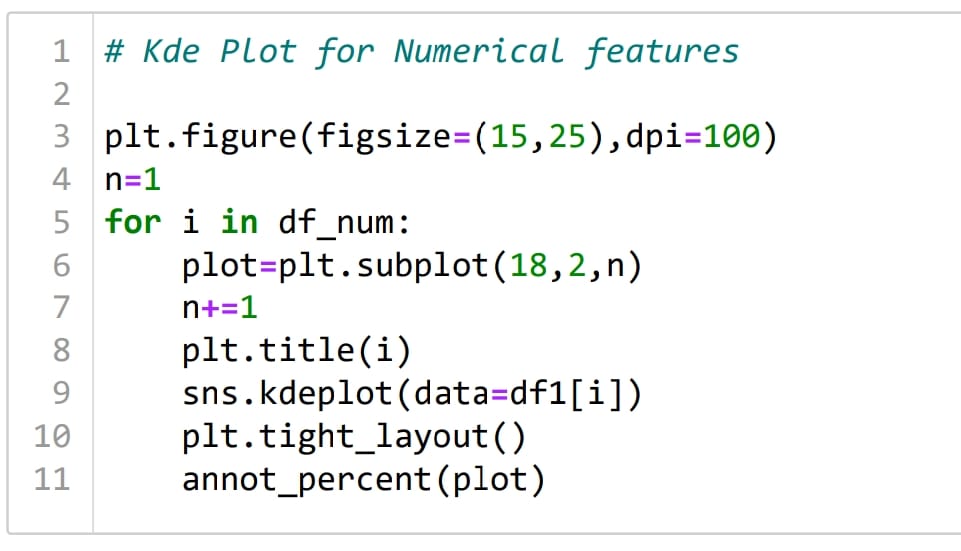
**Skewness After Transformation:**

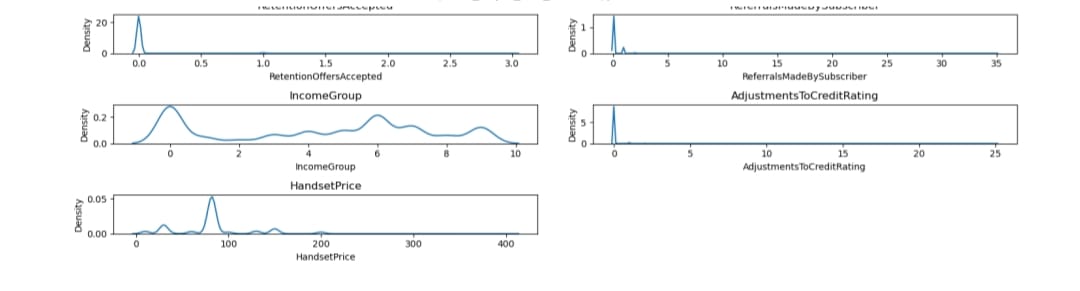


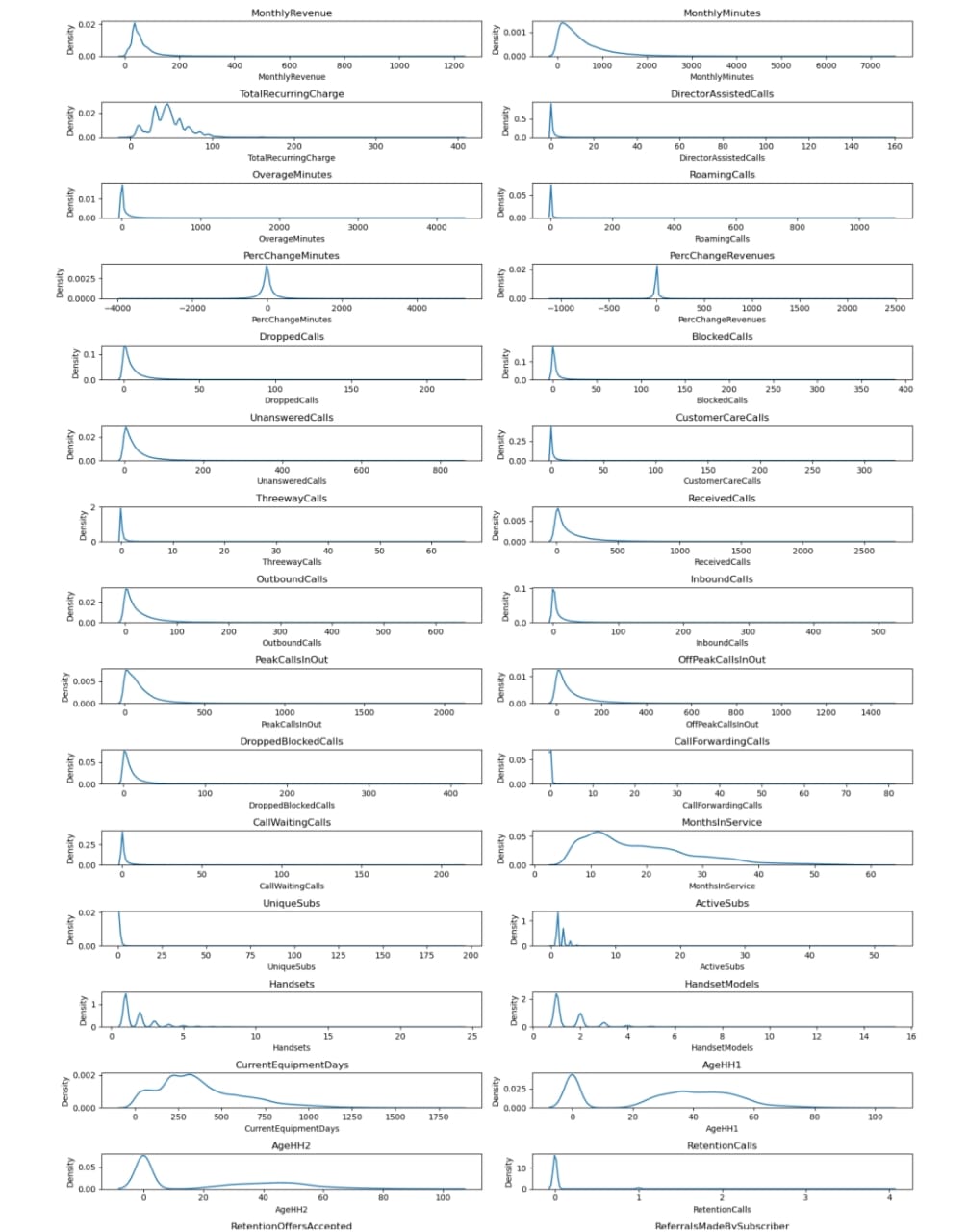
**Inference:** Here by visualizing dist plot we can observe that there is a reduction of skewness after Transformation.

**Descriptive Analysis (EDA)**

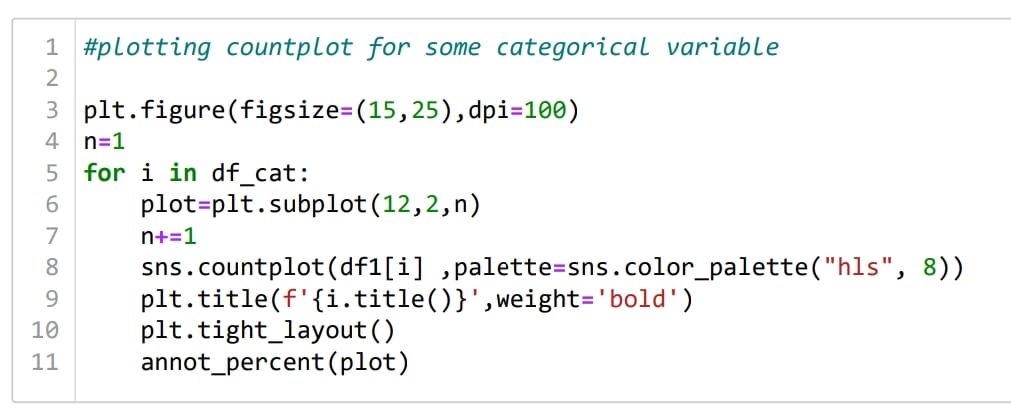
**Univariate Analysis:**

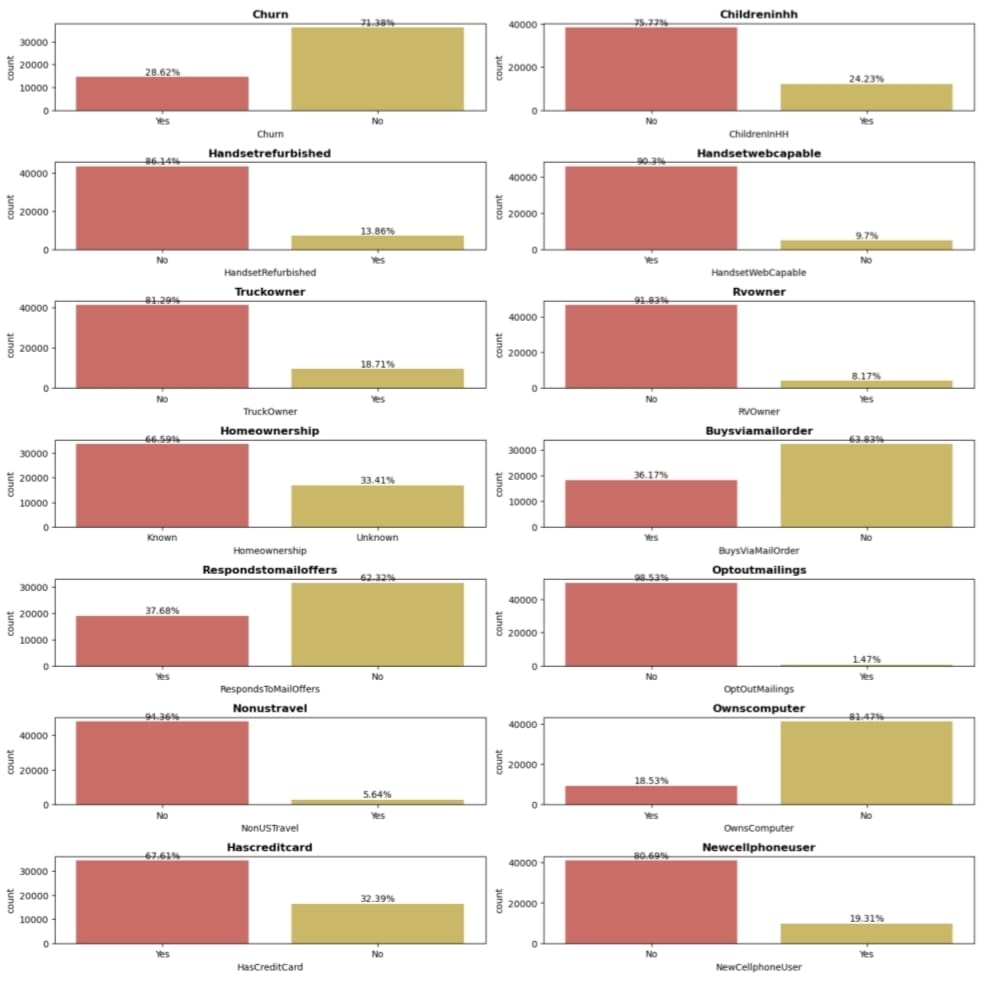
Numerical Columns Visualization:

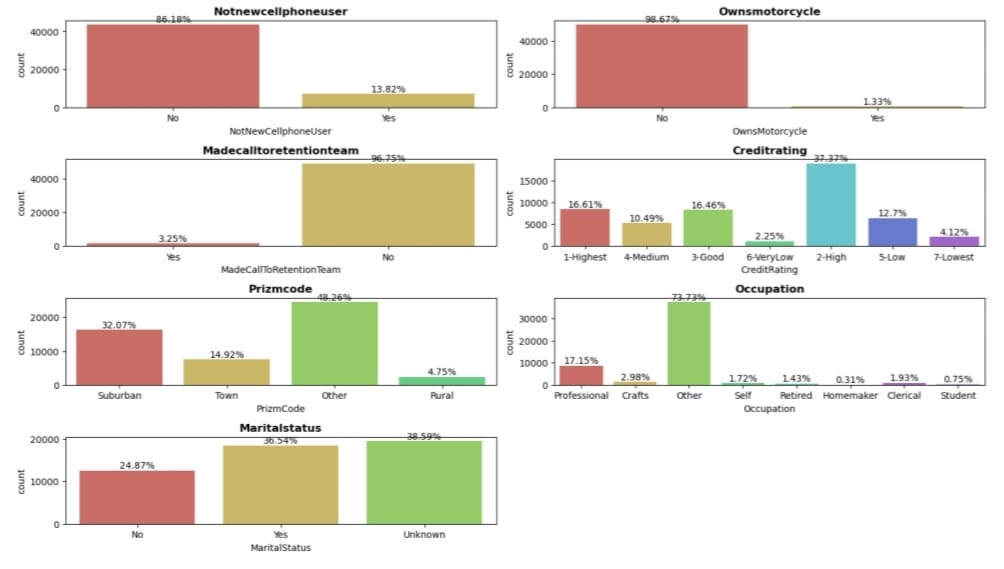




**Categorical Columns Visualization:**

****

****

****

#### Observations:

1) Churn Over 28 percent of people in the data have churned.

2) Handsetwebcapable More than 90 percent of the people in the data have internet support on their phone.

3) More than 65 percent of them don't have a credit card

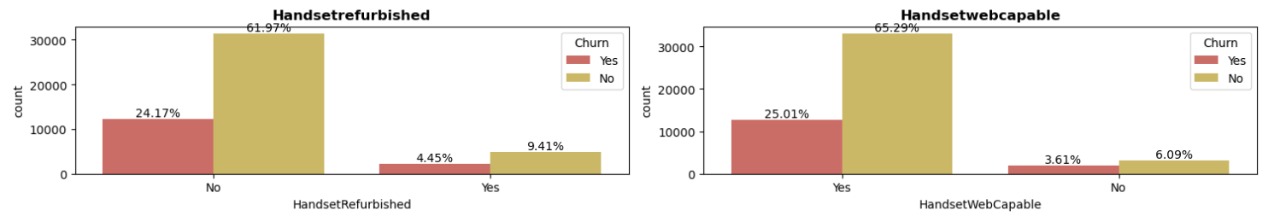
4) Less than 2 percent of them own a motorcycle

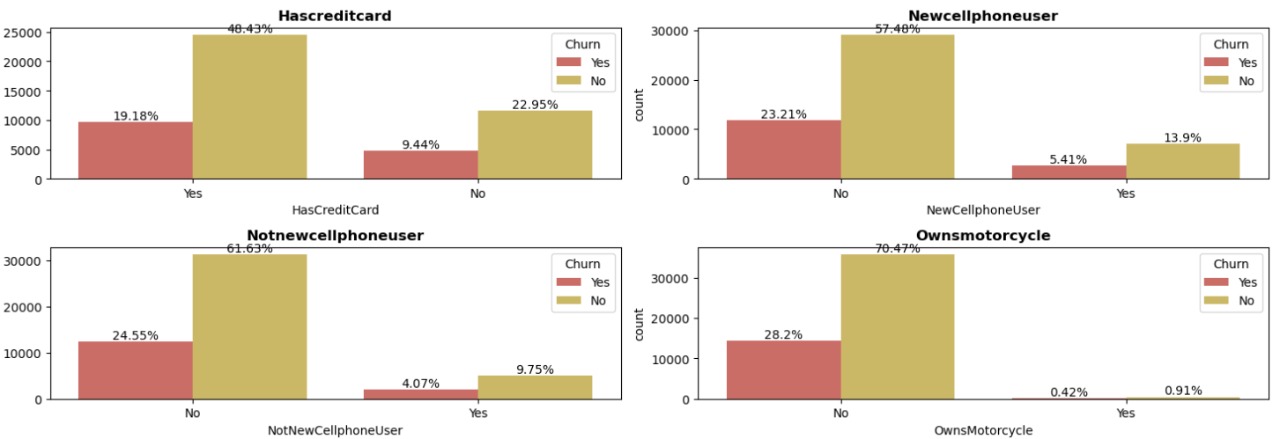
5) More than half of the people's handset price is unknown

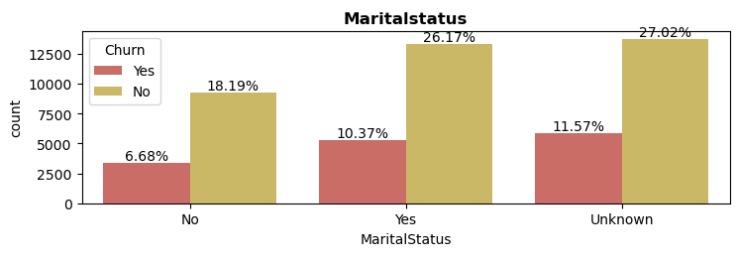
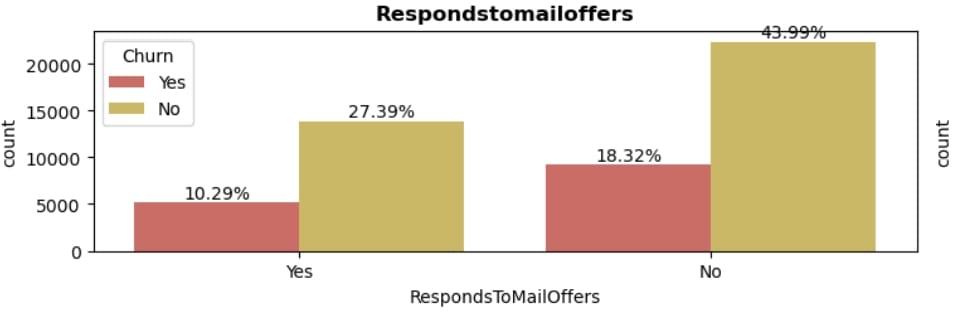
6) Over 70 percent of the data has occupations other than the ones mentioned.

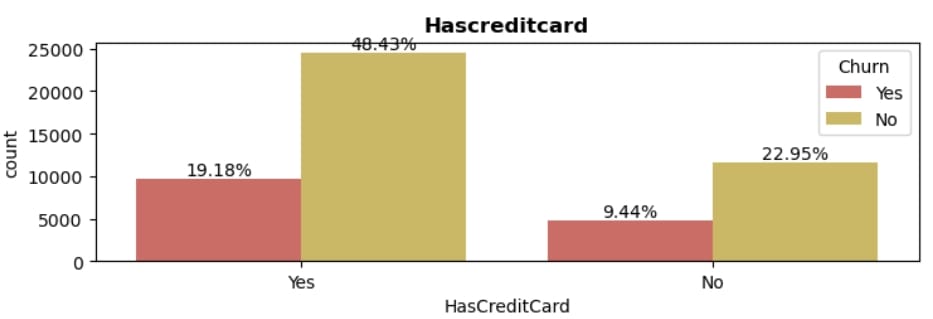
7) Martial status of 60 percent of the data is known out of which, 25 percent are not married. The rest are unknown.

**Bivariate Analysis:**

****



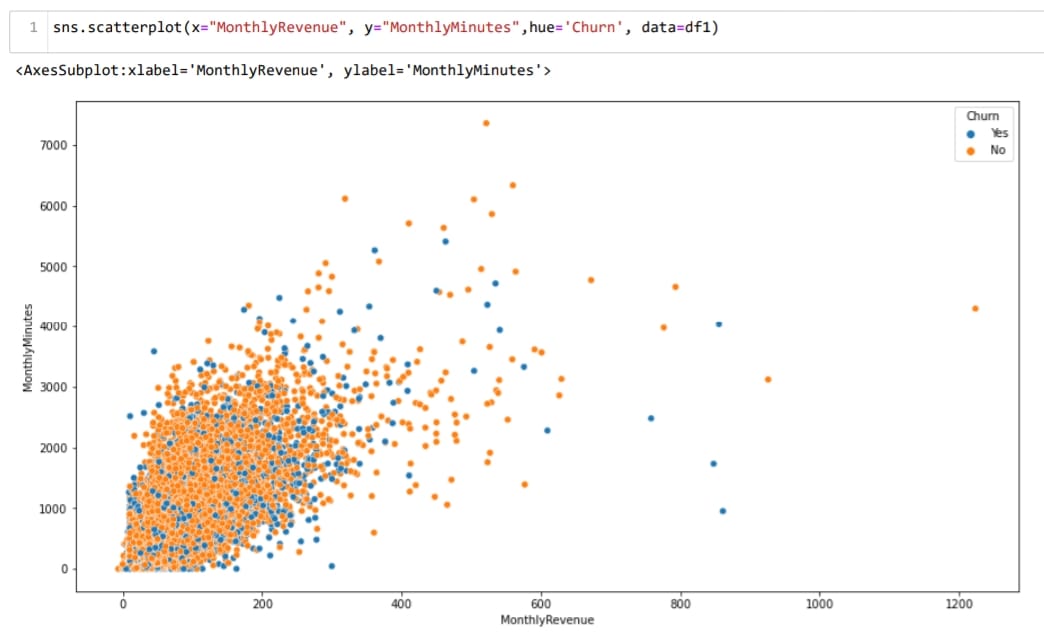




### **observation:**

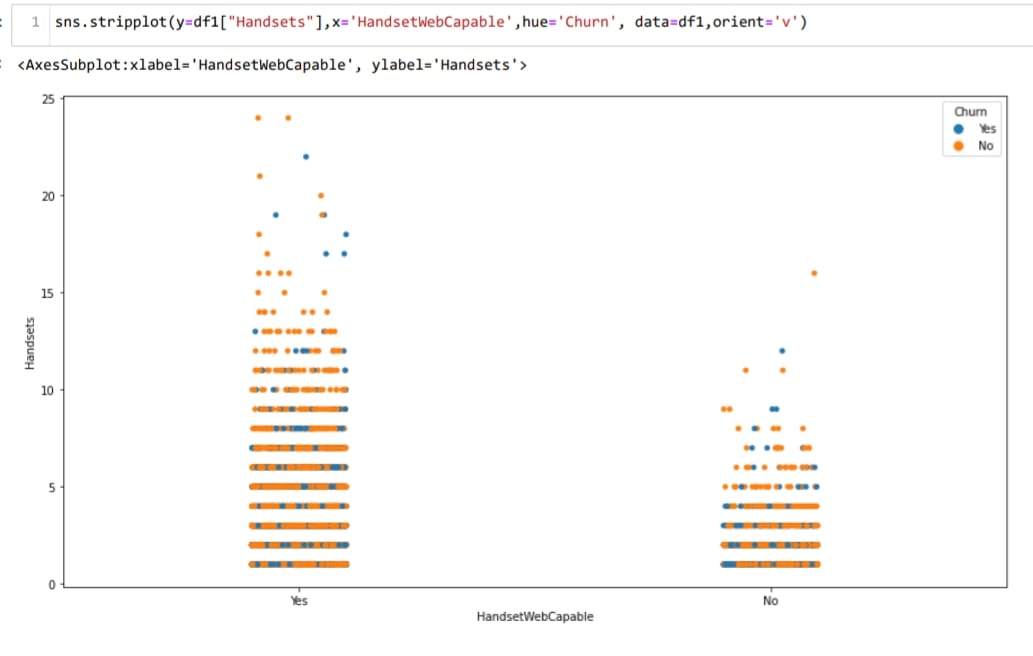
1. In Handset web capability over 25% of people who have churned has more than 90% of Internet capability on their phone.
2. Less than 6% of people who own new phone have churned.
3. Data shows that people who have Credit Cards are more likely to Churn
4. Marital Status of people churning is independent
5. People who have responded mail offer are less likely to churn

**Multivariate Analysis:**

****

### **Observation:**

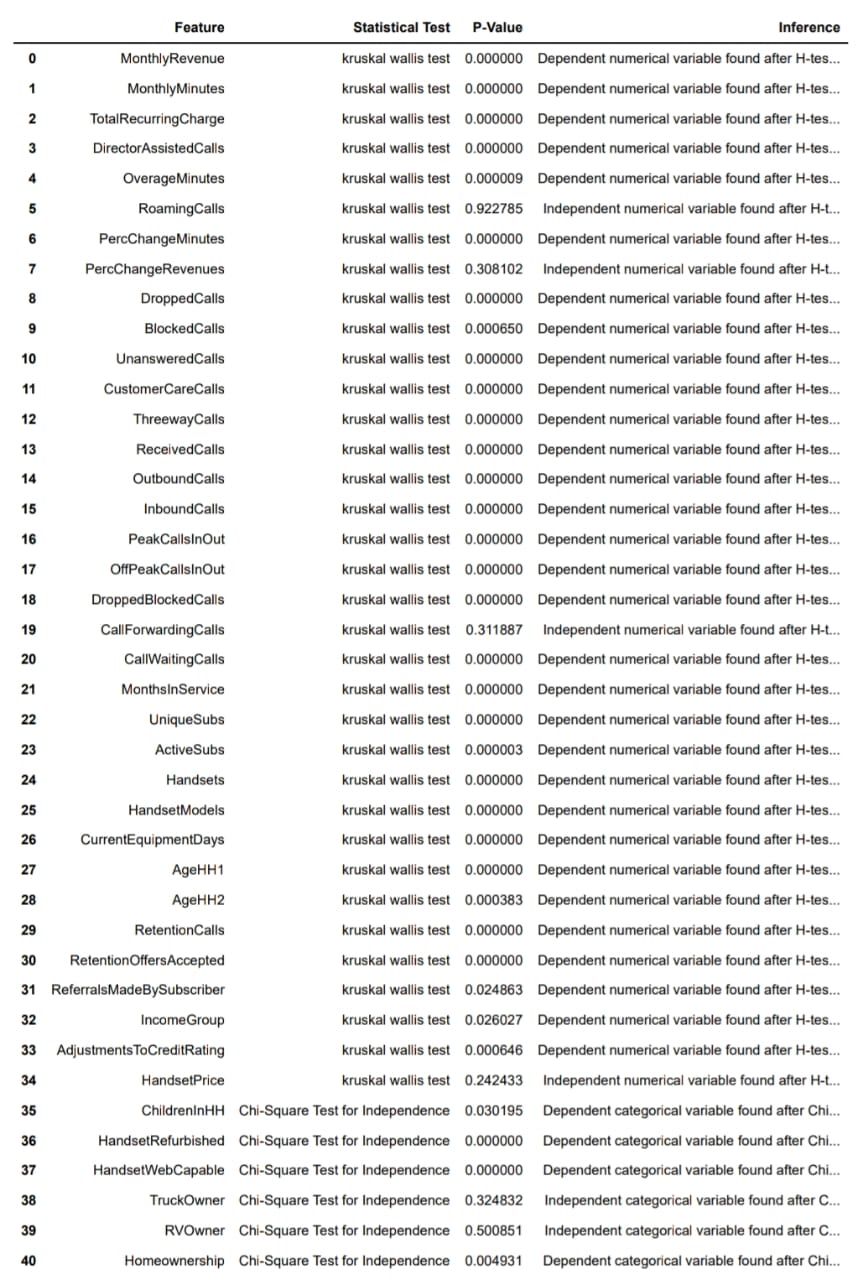
According to plot, as Monthly Revenue Increases, then number of Monthly Minutes increases, but we could not draw any conclusion on churn.

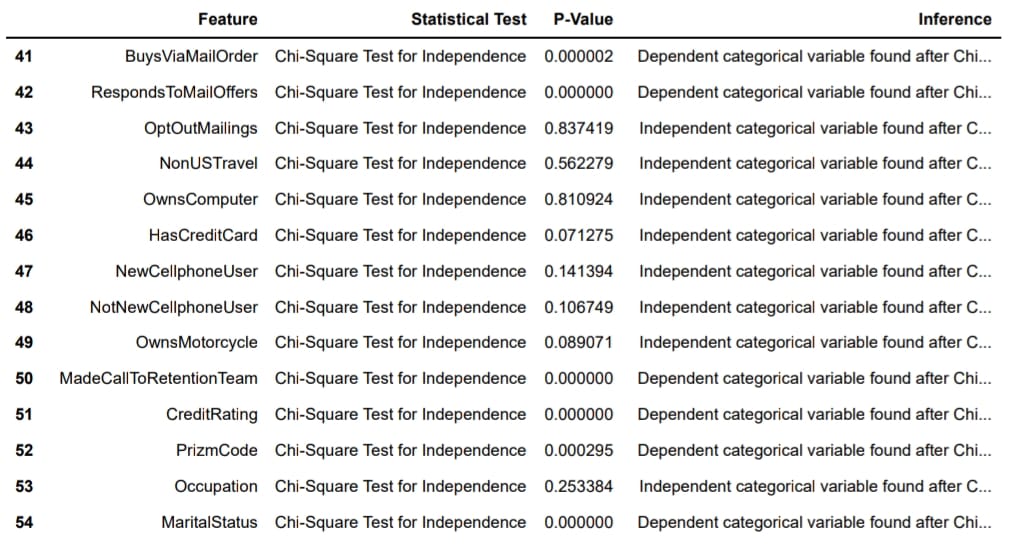


### **Observation:**

As the number of handset Increases, with this certain percentage peoples are more likely to churn.

**Statistics (Stats)**

****

****

We have used Chi-Square Test for Independence to test whether the categorical variables are independent or not.

**𝐻0** : The variables are independent.

**𝐻1**: The variables are not independent (i.e., variables are dependent).

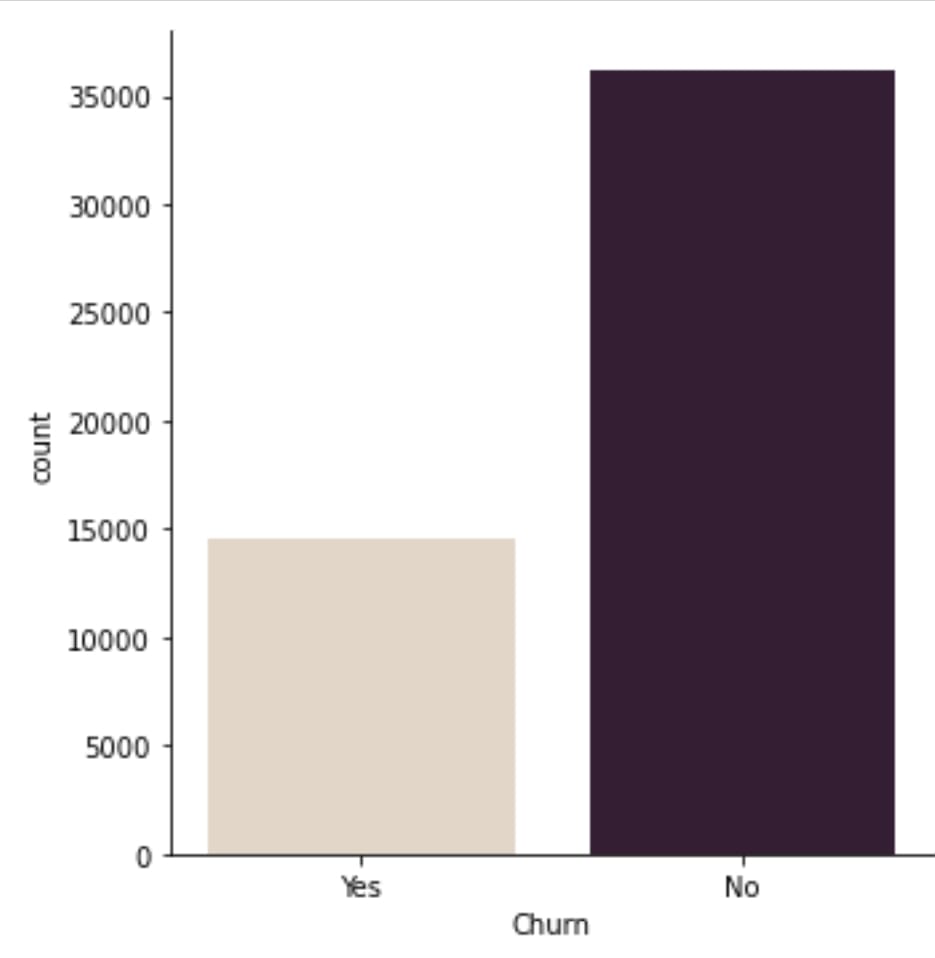
We have used Jarque-bera test to check the normality of data

**𝐻0** : The data is normally distributed.

**𝐻1**: The data is not normally distributed.

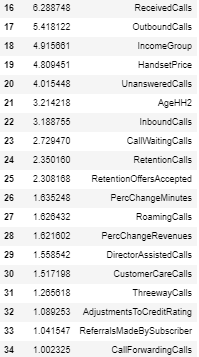
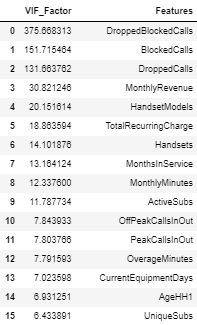
We found that data is not normal therefore we use Kruskal Wallis test to check its dependency on the target variable

**Class Imbalance and its Treatment:**

****

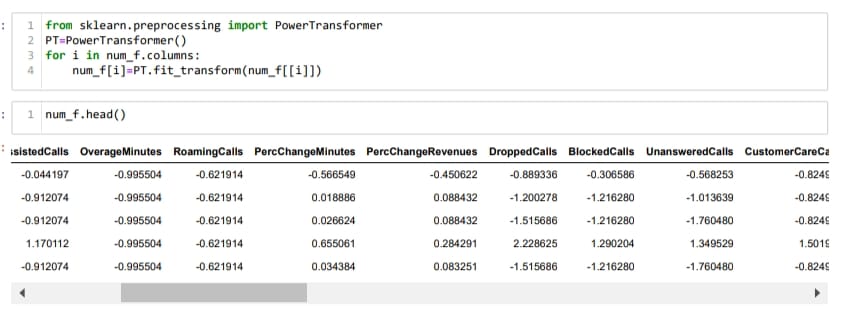
Here we can see that our target variable is too imbalanced, and to treat that we are going to use oversampling techniques like smote.

**Check of Multicollinearity:**

****

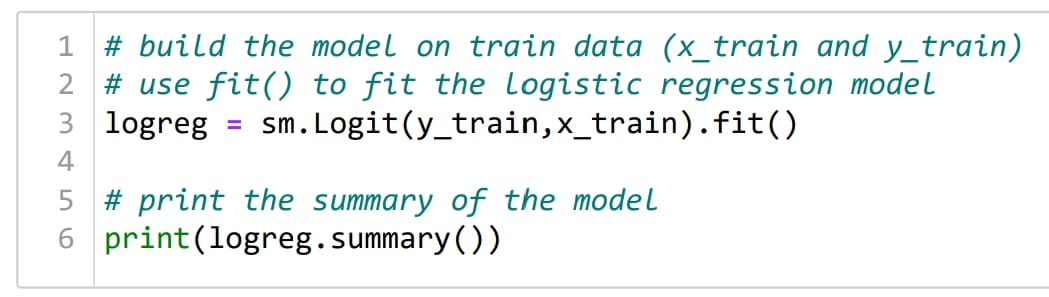
**Transformation:**

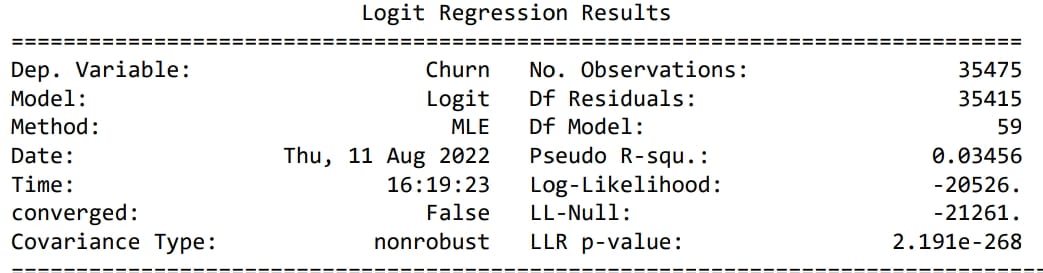
Transformation is a process that can be used to change the scale of the original data to get more accurate results. We used Power transformation, as we can see that there is large number of outliers present so we use Yeo-Johnson transformation technique to reduce the outliers and make the data more normally distributed.



**Logistic Regression (Base Model)**

#### Build a full logistic model on a training dataset.





**Interpretation:** The Pseudo R-squ. obtained from the above model summary is the value of McFadden's R-squared. This value can be obtained from the formula:

**McFadden's R-squared =**1−(𝐿𝑜𝑔−𝐿𝑖𝑘𝑒𝑙𝑖ℎ𝑜𝑜𝑑/LL−Null)

Where,  
Log-Likelihood: It is the maximum value of the log-likelihood function  
LL-Null: It is the maximum value of the log-likelihood function for the model containing only the intercept

1. The LLR p-value is less than 0.05, implies that the model is significant.

**Cox & Snell R-squared:** The convergence of the logistic model can be determined by the R-squared value. It is one of the types of Pseudo R-square.

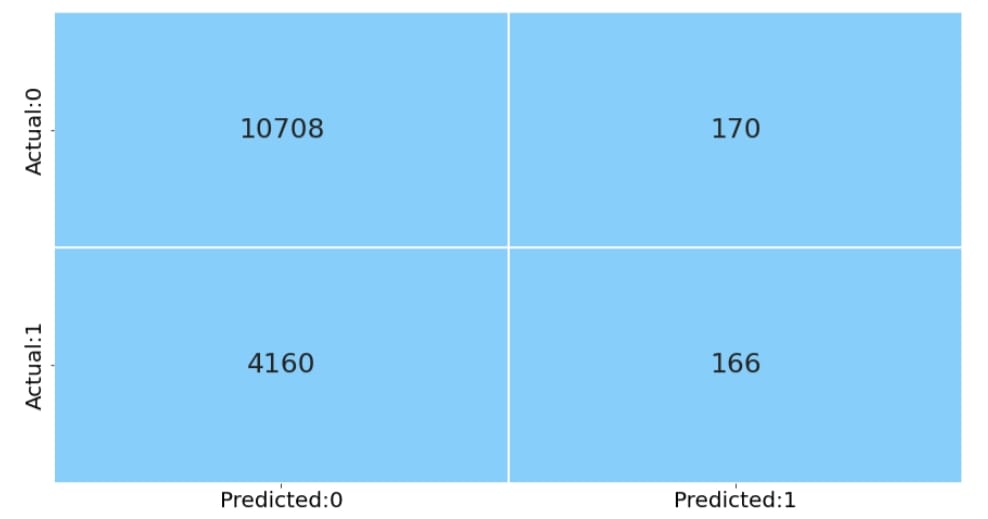
1. The maximum of Cox & Snell R-squared is always less than 1. By above model Cox & Snell R-squared is less than 1 i.e. (0.03456).

#### The AIC (Akaike Information Criterion) value:

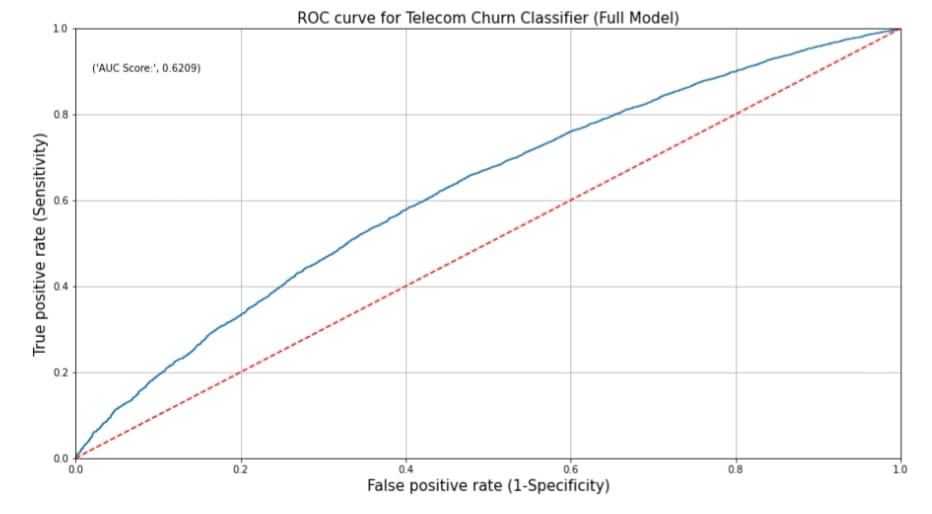
It is a relative measure of model evaluation. It gives a trade-off between model accuracy and model complexity.

AIC: 41172.911

**Confusion Matrix:**

****

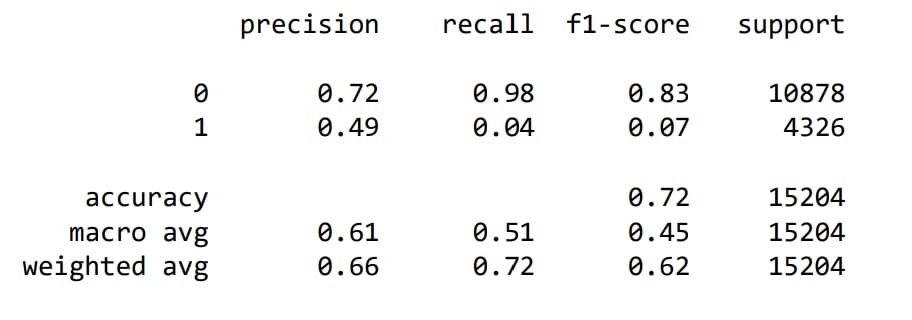
**ROC Curve:**

****

**Inference:**

* The red dotted line represents the ROC curve of a pure random classifier; a good classifier stays as far away from that line as possible (towards top-left corner).
* From the above plot, we can see that our classifier (logistic regression) is away from the dotted line; with the AUC score 0.6209.

**Classification Report:**

****

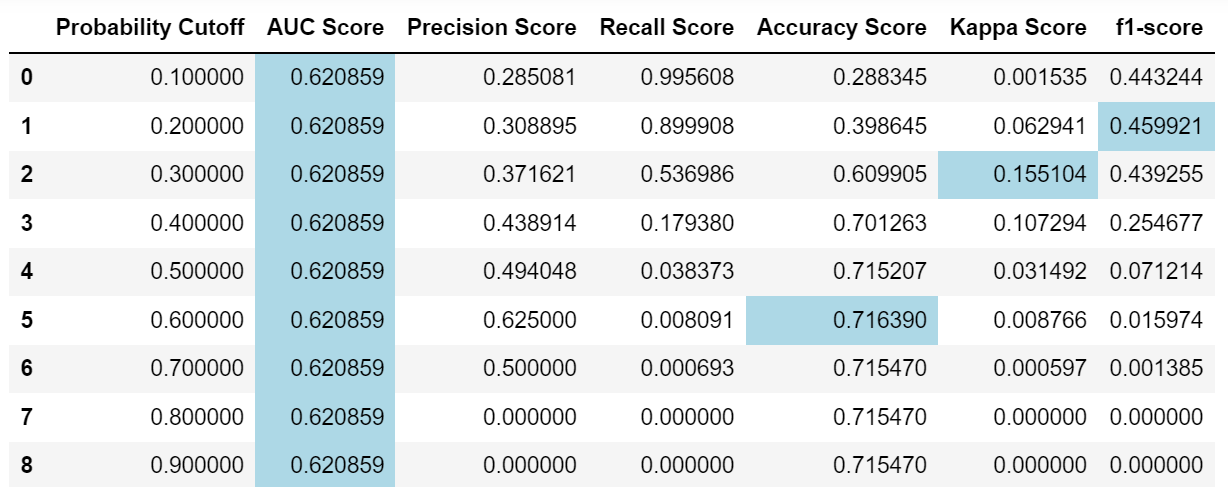
**Interpretation:**

From the above output, we can infer that the recall of the positive class is known as sensitivity and the recall of the negative class is specificity.

support is the number of observations in the corresponding class.

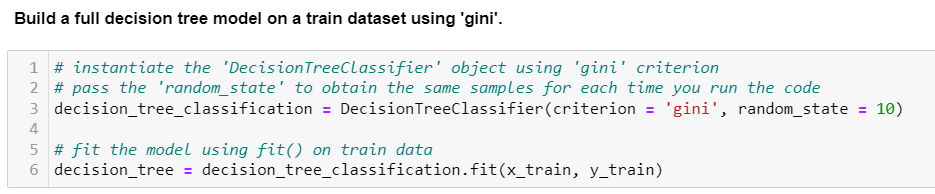
The macro average in the output is obtained by averaging the unweighted mean per label and the weighted average is given by averaging the support-weighted mean per label.

**Score Card for Logistic Regression:**

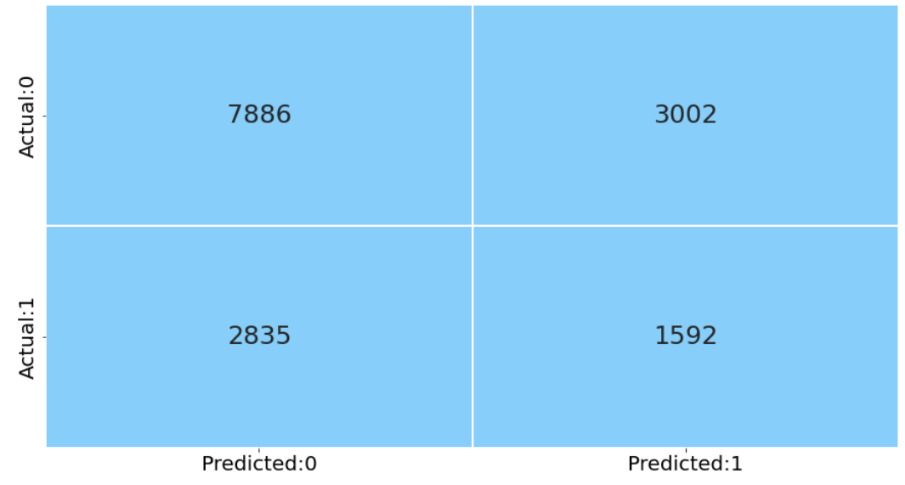
****

**Interpretation:** The above data frame shows that, the model cutoff probability 0.6, returns the highest AUC score, f1-score, kappa score and accuracy.

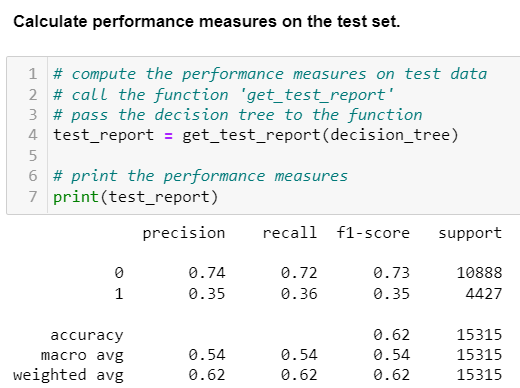
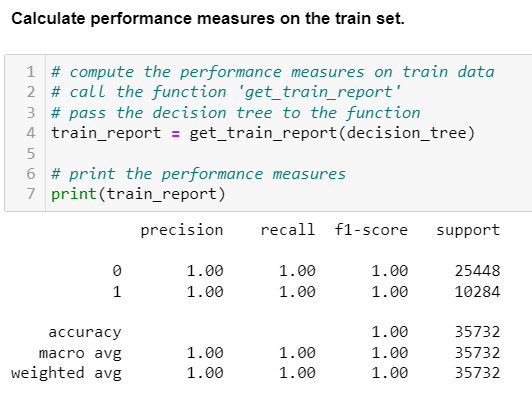
**Decision Tree**

****

**Model Performance: -**

**1.Confusion Matrix**

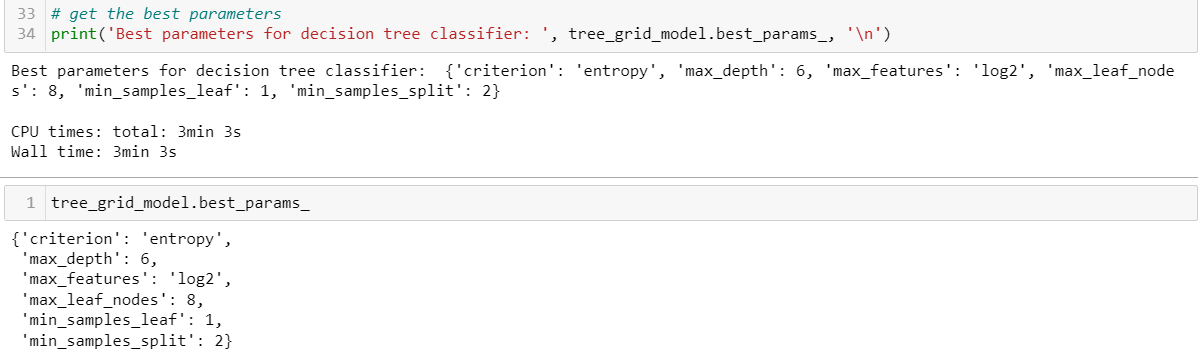
**2.Report: -**

****

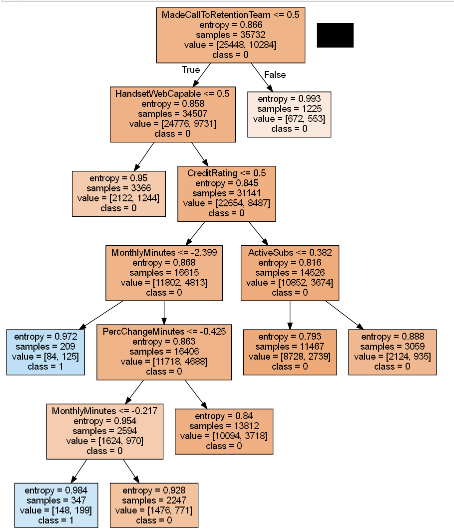
**Inference: -**

* From The above model, our train accuracy is 1 and test accuracy is 0.62, result is Overfitting
* As the model is over fitted, our false Negative and false Positive is inaccurate
* In the next step we tuned the Hyperparameters and rebuild the model.

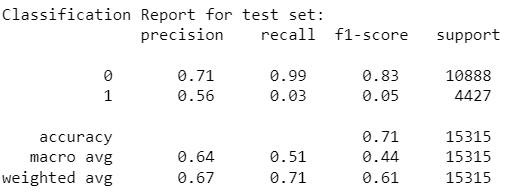
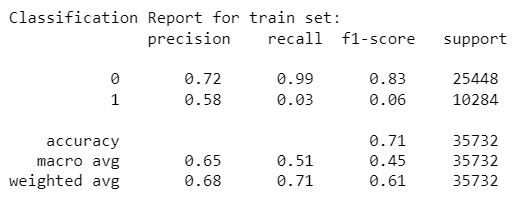
## Tune the Hyperparameters using GridSearchCV (Decision Tree)



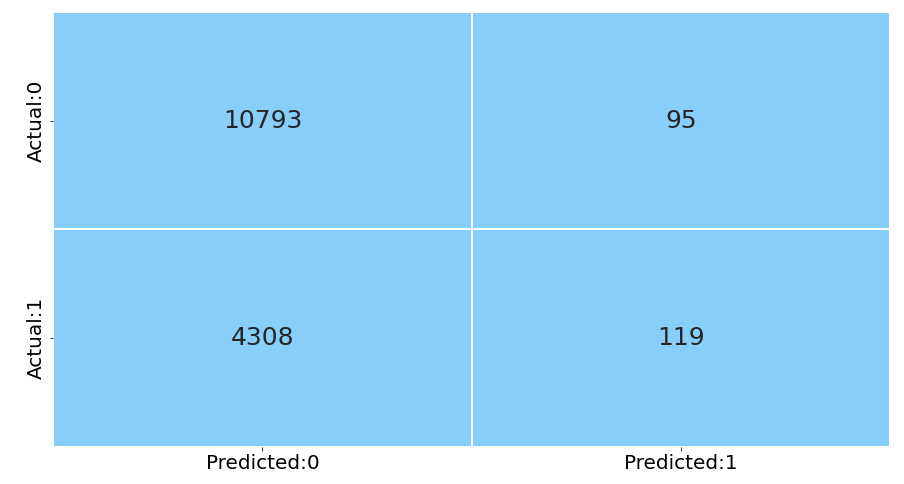
Hyper Tuned Decision Tree:



**Model Performance after Tunning:**

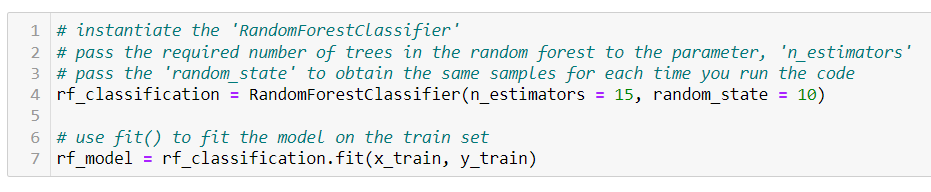
****

**Confusion Matrix:**

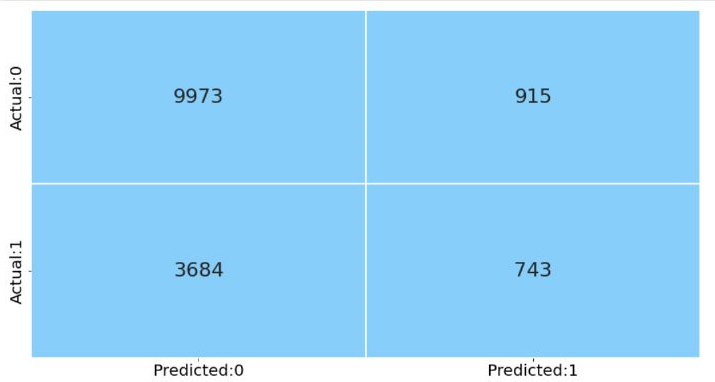
****

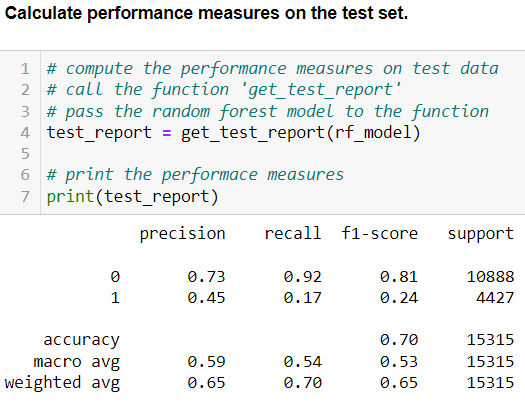
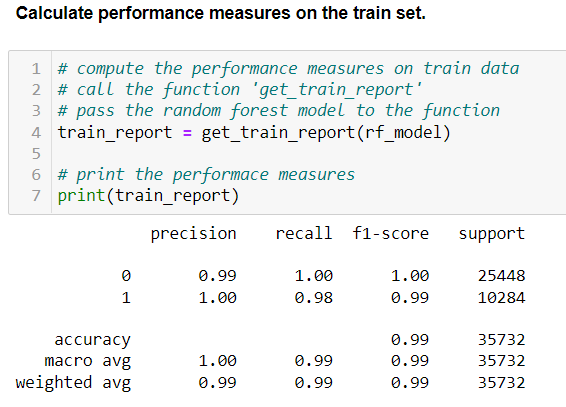
**Inference: -**

* The train and test Accuracy are comparable, which shows the reduction in overfitting.
* In this case the false negative and false positive values can be trusted and the FN value are quite High, but as our focus is on reduction of False negative values
* Typically, Random Forest classifier is more accurate than a single decision tree, we rebuild the model using the same to reduce the FN and increase the accuracy.

**Random forest for classification**

**Confusion matrix:**

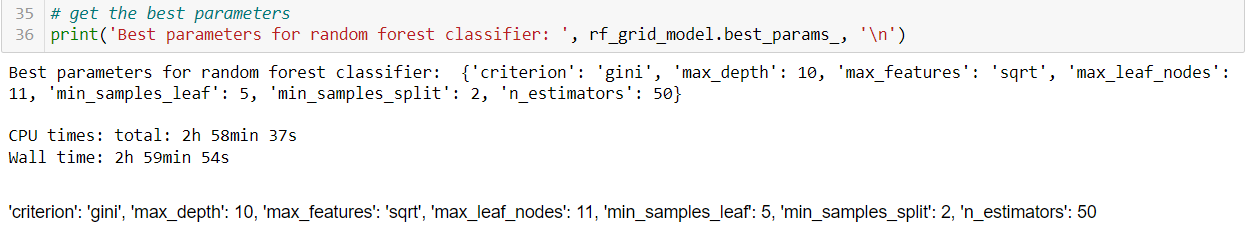
****

**Report: **

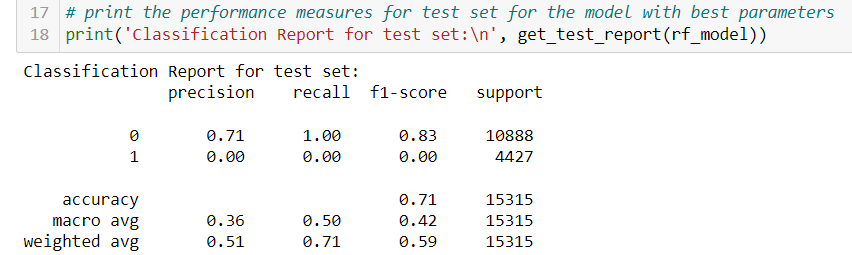
**Inferences:**

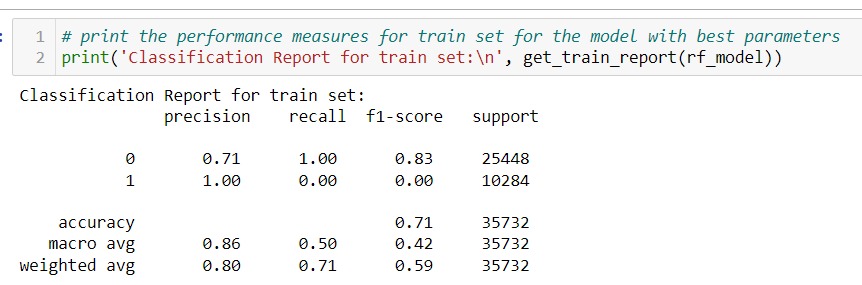
* From The above model, our train accuracy is 0.99 and test accuracy is 0.70, result is Overfitting
* As the model is over fitted, our false Negative and false Positive is inaccurate
* In the next step we tuned the Hyperparameters and rebuild the model.

## **Tuned the Hyperparameters using GridSearchCV (Random Forest)**

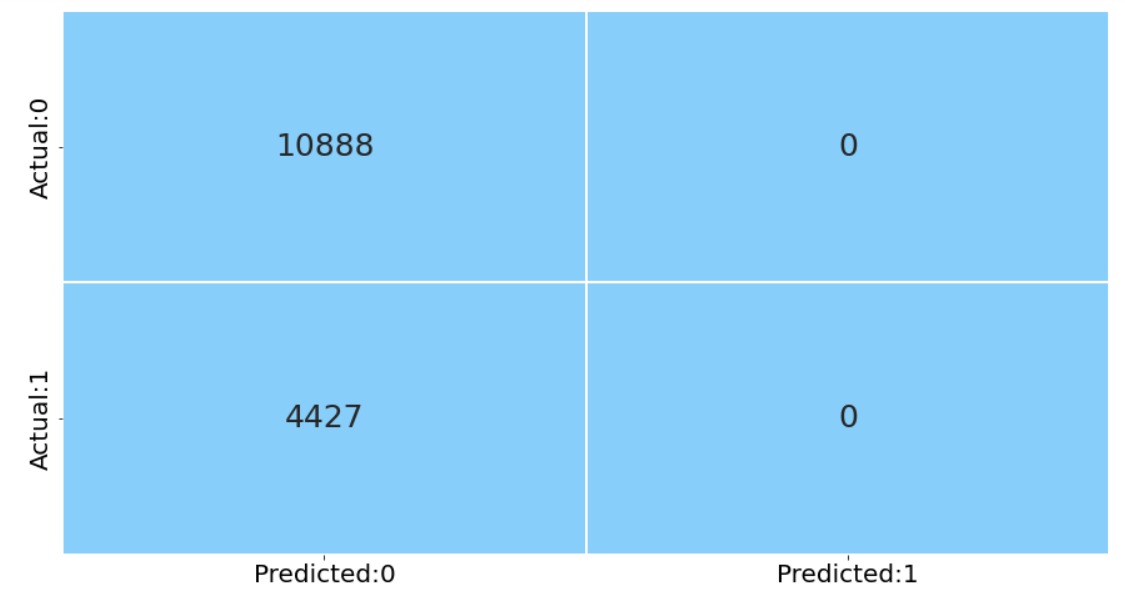


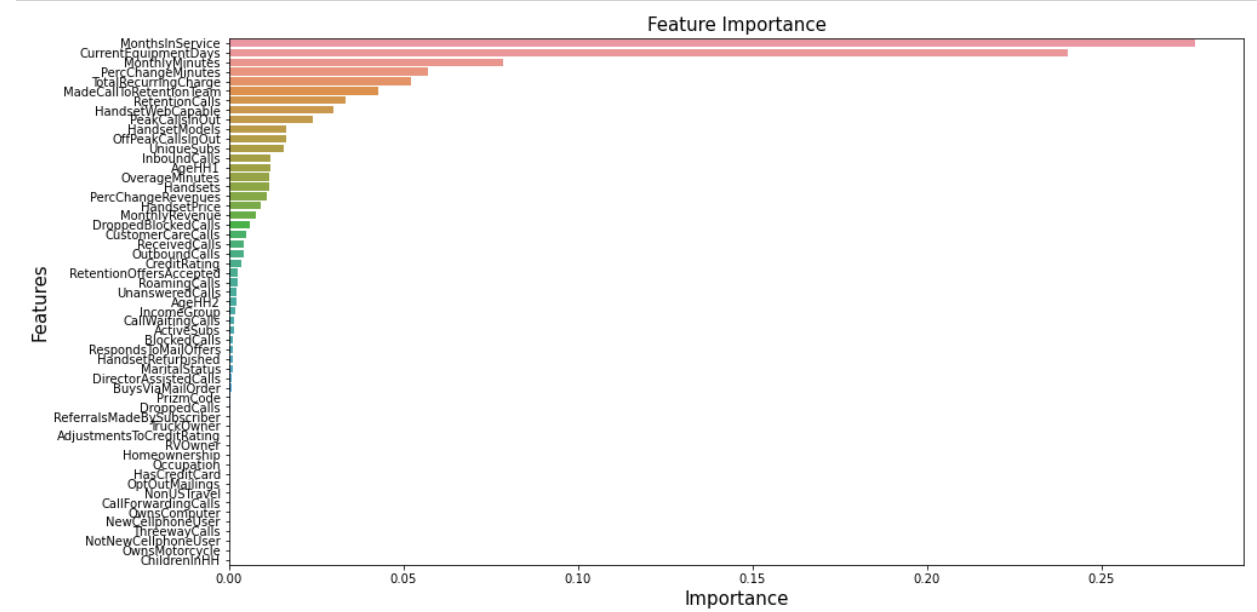
**Model performance after tuning:**

****

****

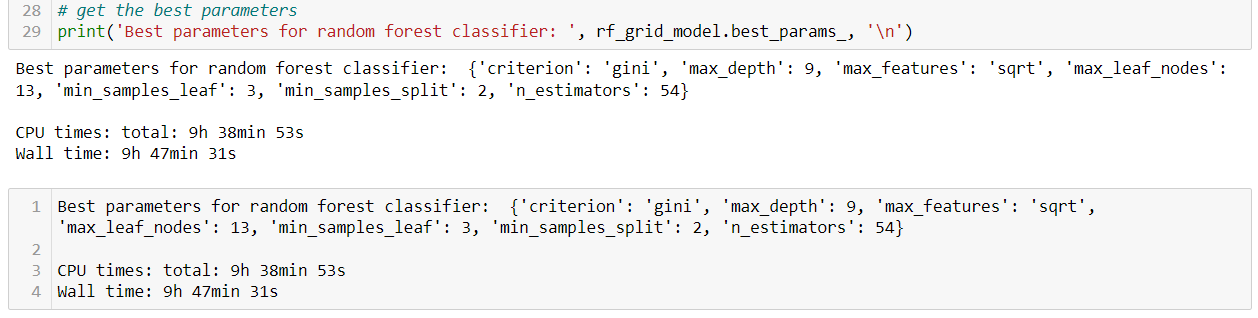
**Confusion matrix:**

****

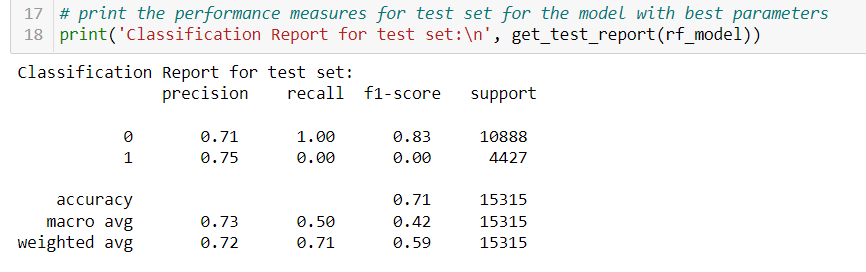
**Feature importance:**

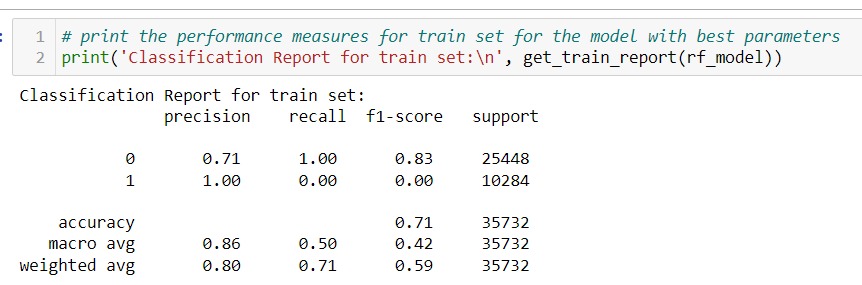
The method feature-importance returns the value corresponding to each feature which is defined as the ratio of total decrease in Gini impurity across every tree in the forest where the feature is used to the total count of trees in the forest. This is also called as, Gini importance.

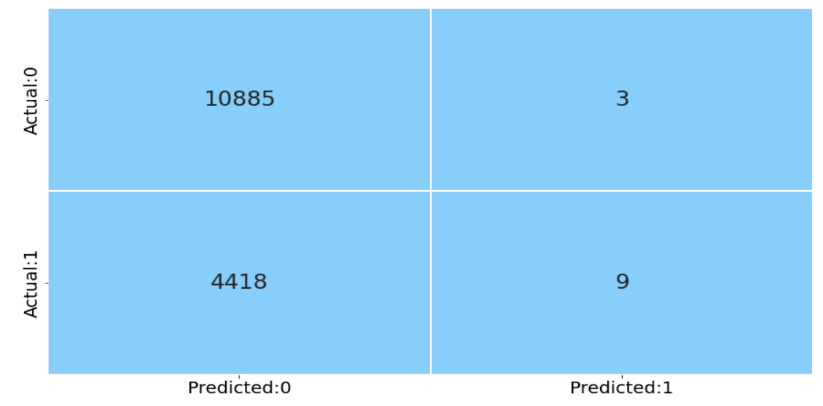
## **Tune the Hyperparameters using GridSearchCV (Random Forest) -2**



**Model performance after tuning:**





**Confusion matrix:**

**Inference:**

The train and test Accuracy are comparable, which shows the reduction in overfitting.

* In this case the false negative and false positive values can be trusted and the FN value are quite High, but as our focus is on reduction of False negative values
* Typically, Random Forest classifier is more accurate than a single decision tree, we rebuild the model using the same to reduce the FN and increase the accuracy.

**Boosting**

# **Boosting Methods:**

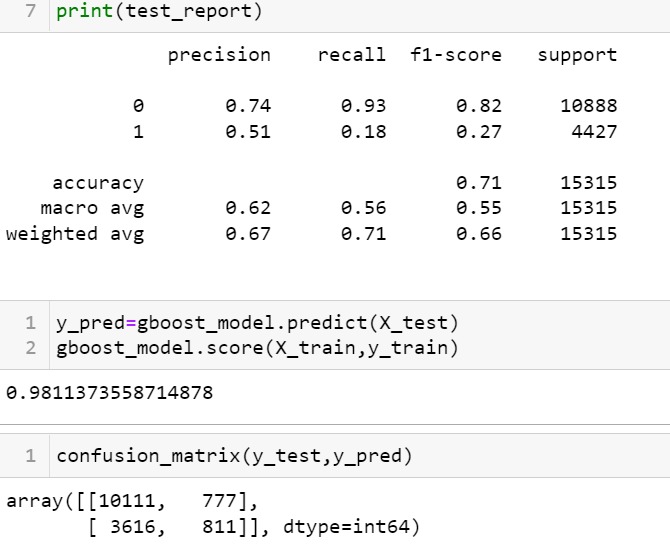
The Ensemble technique considers multiple models for predicting the results. Bagging and Boosting are two of the types of ensembles. The bagging methods construct the multiple models in parallel; whereas, the boosting methods construct the models sequentially.

Earlier, we have studied one of the bagging (bootstrap aggregating) technique i.e., Random Forest.

The boosting method fits multiple weak classifiers to create a strong classifier. In this method, the model tries to correct the errors in the previous model. In this section, we learn some of the boosting methods such as AdaBoost, Gradient Boosting and Boost.

## **1 Gradient Boosting:**

This method optimizes the differentiable loss function by building the number of weak learners (decision trees) sequentially. It considers the residuals from the previous model and fits the next model to the residuals. The algorithm uses a gradient descent method to minimize the error.

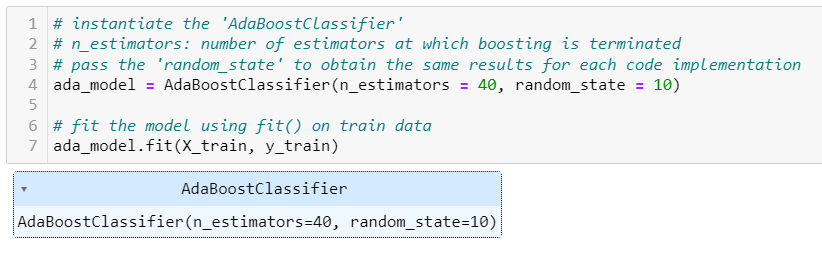


**Inference:**

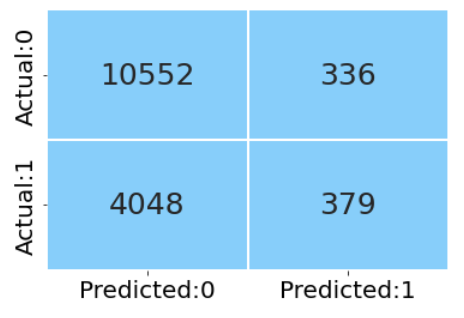
* According to the train-test report, train value shows 98% and test value shows 71% from this we can say the model is overfitted.

## **2 AdaBoost:**

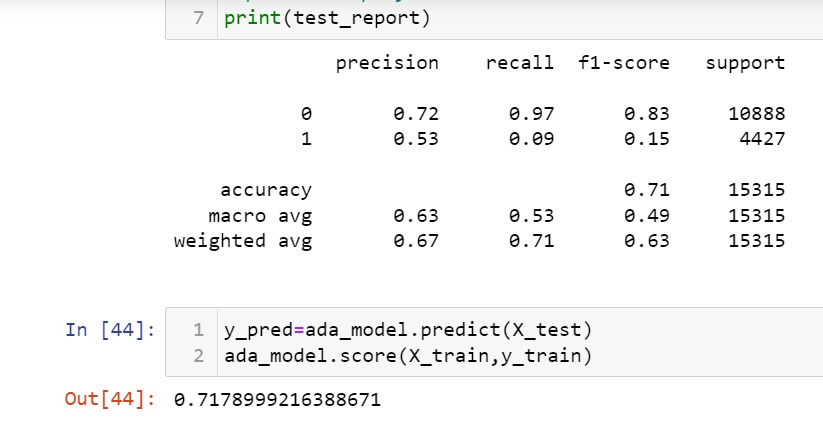
Let us build the AdaBoost classifier with decision trees. The model creates several stumps (decision tree with only a single decision node and two leaf nodes) on the train set and predicts the class based on these weak learners (stumps). For the first model, it assigns equal weights to each sample. It assigns the higher weight for the wrongly predicted samples and lower weight for the correctly predicted samples. This method continues till all the observations are correctly classified or the predefined number of stumps is created.



**Confusion matrix:**



**Model Performance:**



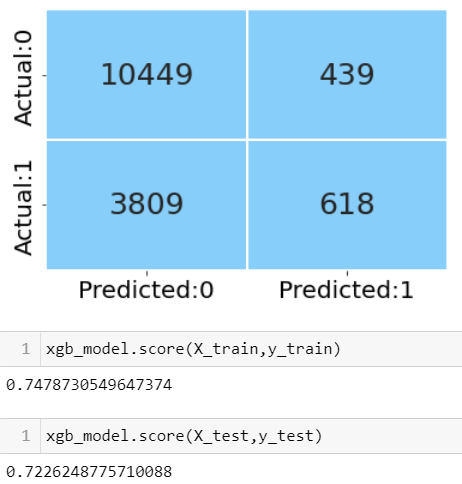
# **Inference:**

* According to Adaboost report, the train-test report shows comparable values with each other.
* The model fit is good, but still need to improve the model by reducing the FN.

## **3. XGBoost:**

XGBoost (extreme gradient boost) is an alternative form of gradient boosting method. This method generally considers the initial prediction as 0.5 and build the decision tree to predict the residuals. It considers the regularization parameter to avoid overfitting.

**Model Performance:**



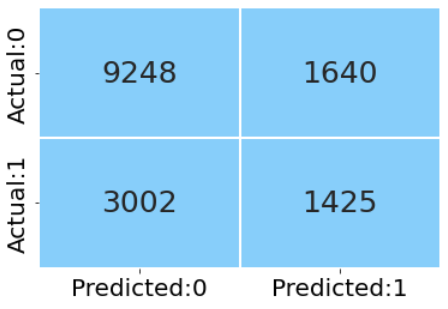
**Inference:**

* According to XGBoost model report, we can observe that train-test values are good.
* Compared to previous models, the errors have been reduced.
* Compare to all the boosting methods this is best.

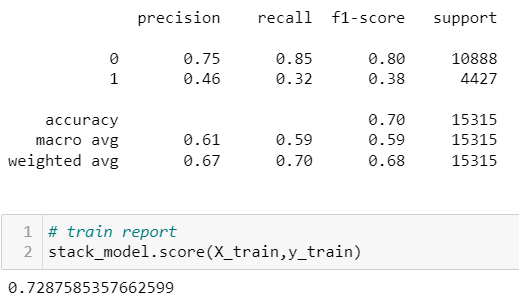
# **Stack Generalization:**



**Confusion Matrix:**



**Model Performance:**

****

**Inference:**

* According to stack Generalization report, we can observe that train test values are good.
* The model is good fit, type II errors are reduced by 807
* **Compared to all the models built this model is the best model.**

**Limitations:**

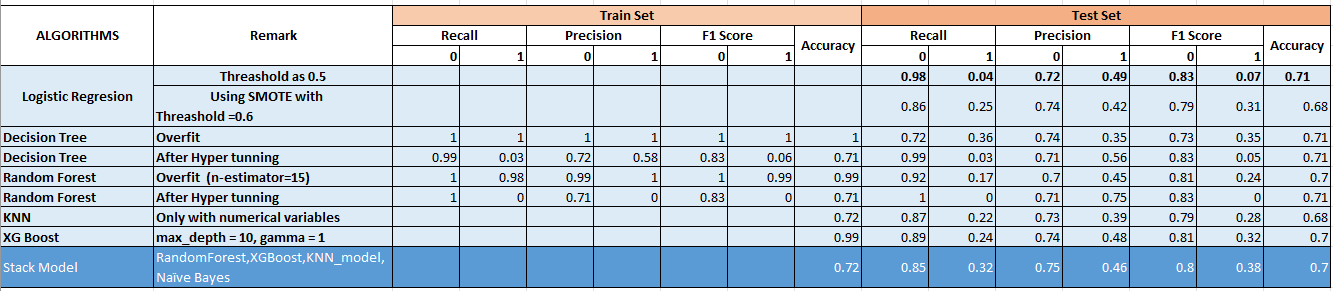
* The data which we have is highly imbalanced this might lead to inaccurate predictions.
* To enhance the data quality and to reduce errors we have transformed the data using power transformer, getting Business insights out of this would be difficult.
* To proceed with Feature Engineering, we need to have domain knowledge

**Conclusion:**

* At first, we dealt with the null value imputation and then we proceeded with Exploratory data analysis to analyse the univariant and bivariant features to understand why the customers are churning.
* As the data was not normal, we use non parametrical statistical test **Kruskal Wallis test**
* This test is used to check features are dependent or independent to Target variables.
* We have built various classification algorithms and final outcomes are as follows
* Compare to base logistic model, the overfitting is reduced and FN errors are reduced by

nearly 32%

* Comparatively the recall value has been boosted from 4% to 32%
* Compare to base Decision model, the overfitting is reduced and FN errors are reduced by nearly 30%

****

**Reference:**

1.Customer Churn Analysis

Brief Overview of Customer Churn Analysis and Prediction with Decision Tree Classifier.

Retrieved from https://towardsdatascience.com/customer-churn-analysis-4f77cc70b3bd

2. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry (2006).

Retrieved from <http://people.stern.nyu.edu/shan2/customerchurn.pdf>

3. Kiran Dahiya, Surbhi Bhatia. Customer Churn Analysis in Telecom Industry.

Retrieved from <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7359318>

4.Teemu Mutanen. Customer churn analysis – a case study.

Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.7169&rep=rep1&type=pdf>

5.Churn Analysis: 3-Step Guide to Analysing Customer Churn Dominique Jackson (March 31, 2020).

Retrieved from <https://baremetrics.com/blog/churn-analysis>

6.Customer Churn Analysis: A Comprehensive Guide Amit Phaujdar on Churn Analysis, Marketing Analytics (March 15th, 2021).

Retrieved from <https://hevodata.com/learn/understanding-customer-churn-analysis/>

7.Understanding Random Forest How the Algorithm Works and Why it Is So Effective. Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

8.Logistic Regression. Retrieved from https://www.sciencedirect.com/topics/computer-science/logistic-regression

9.Decision Tree Algorithm, explained. Retrieved from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

10.Prashant Gupta : Decision Trees in Machine Learning(May 18, 2017).

Retrieved from <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

11.Logistic regression.

Retrieved from <https://www.ibm.com/topics/logistic-regression>

12.Gradient Boosting from scratch.

Retrieved from https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*