

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT Offline BLR Oct22 |
| Team members | Venkata Sai Pavan Teja Angina  Akhil Anilkumar  Shivaprasad G  Sahil Kumar Meher  Arun SV |
| Domain of Project | Retail |
| Proposed project title | Telecom Churn |
| Group Number | 9 |
| Team Leader | Sahil Kumar Meher |
| Mentor Name | Mrs. Pranita Mahajan |

**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** | **Conclusions** | **33** |

### **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

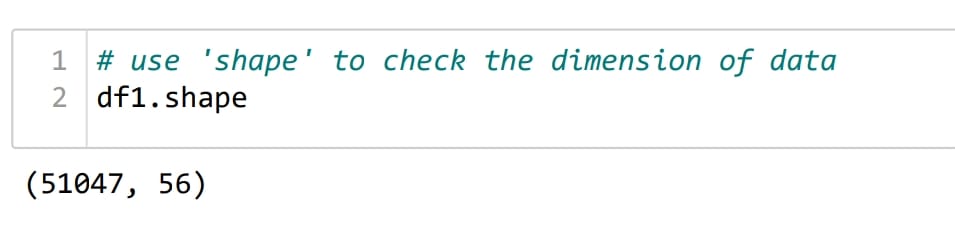
No. of records: 51047

Target Column: churn

Redundant columns: Service area, Customer Id.

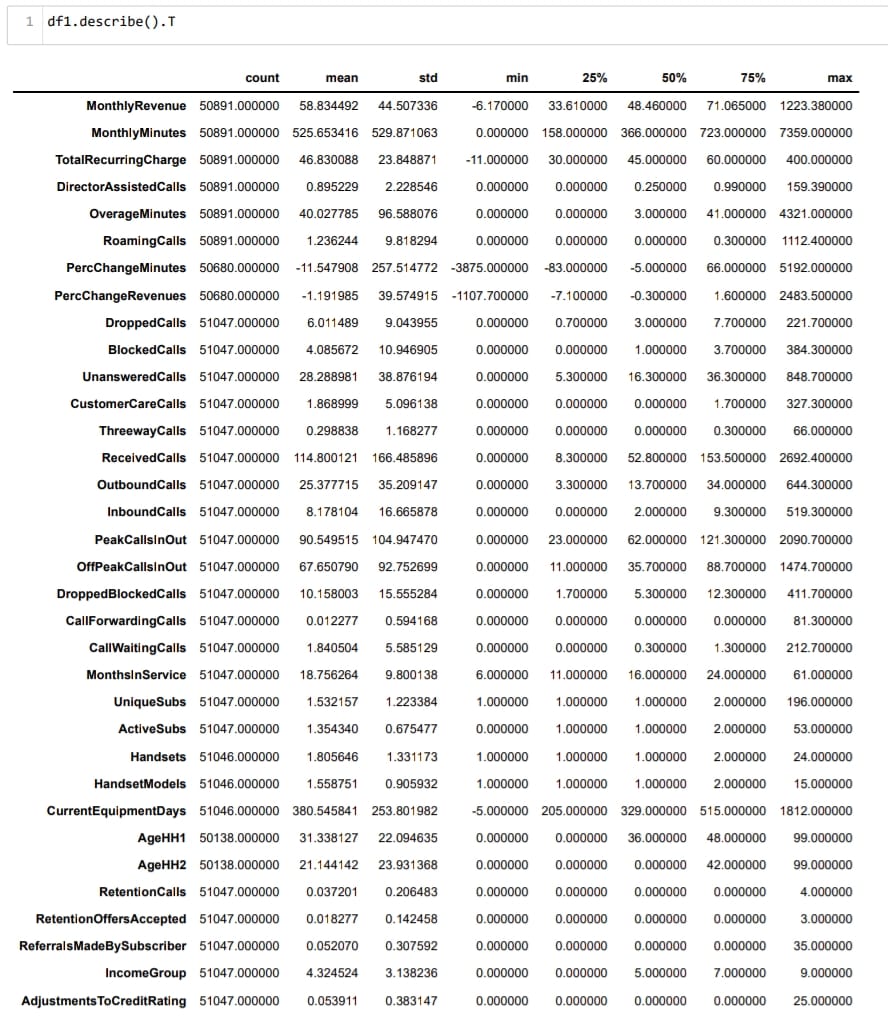
**Understanding the Data:**

Checking Shape of Data:



# **DATA EXPLORATION (EDA)**

**Summary of Dataset:**

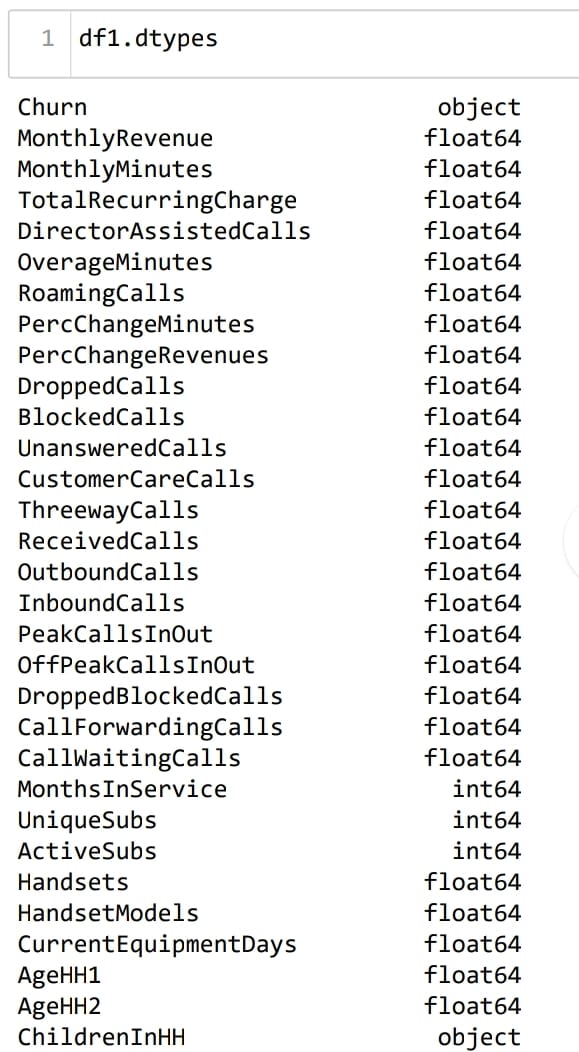


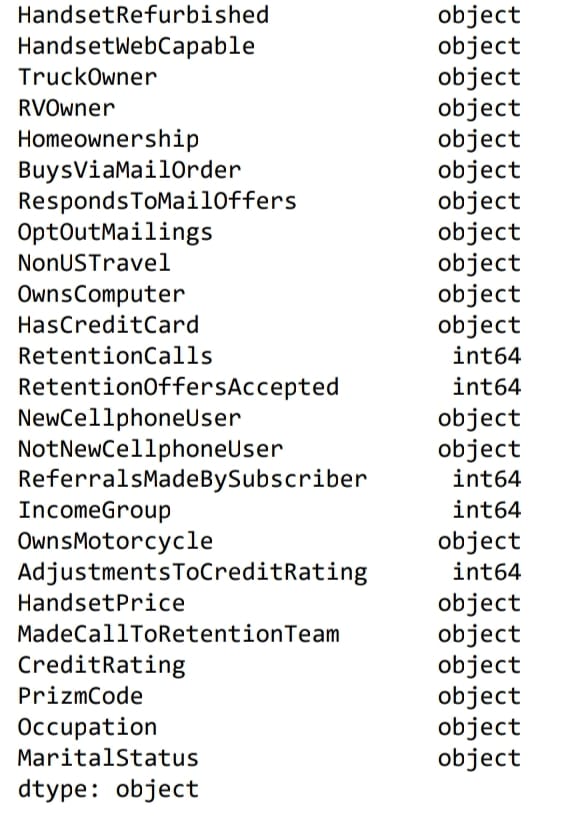
**Interpretation:**

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.

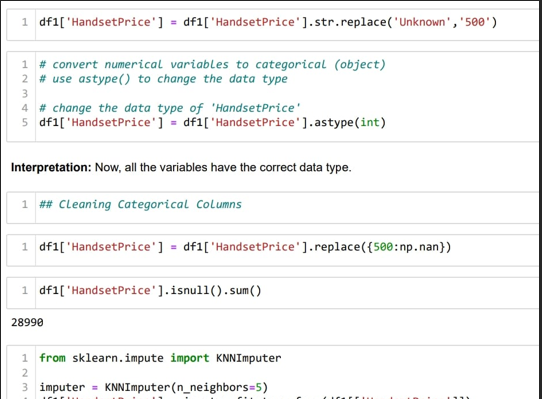
**Check the Data Type:**

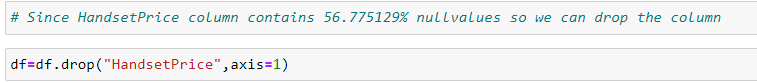
**Check the data type of each variable. If the data type is not as per the data definition, change the data type.**



****

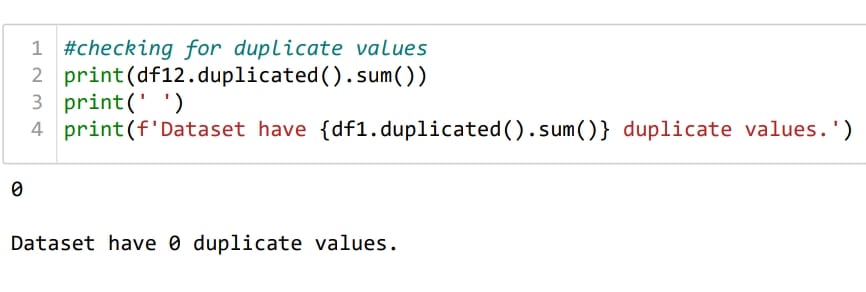
**Recheck the Data type and conversions:**



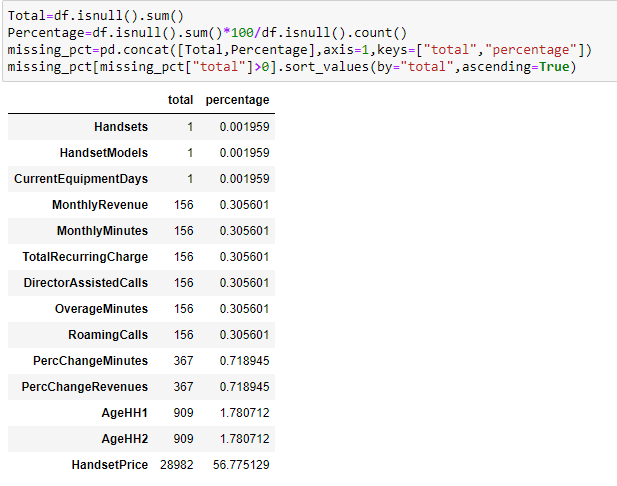


**Data Cleaning**

**Duplicate Values Check:**

****

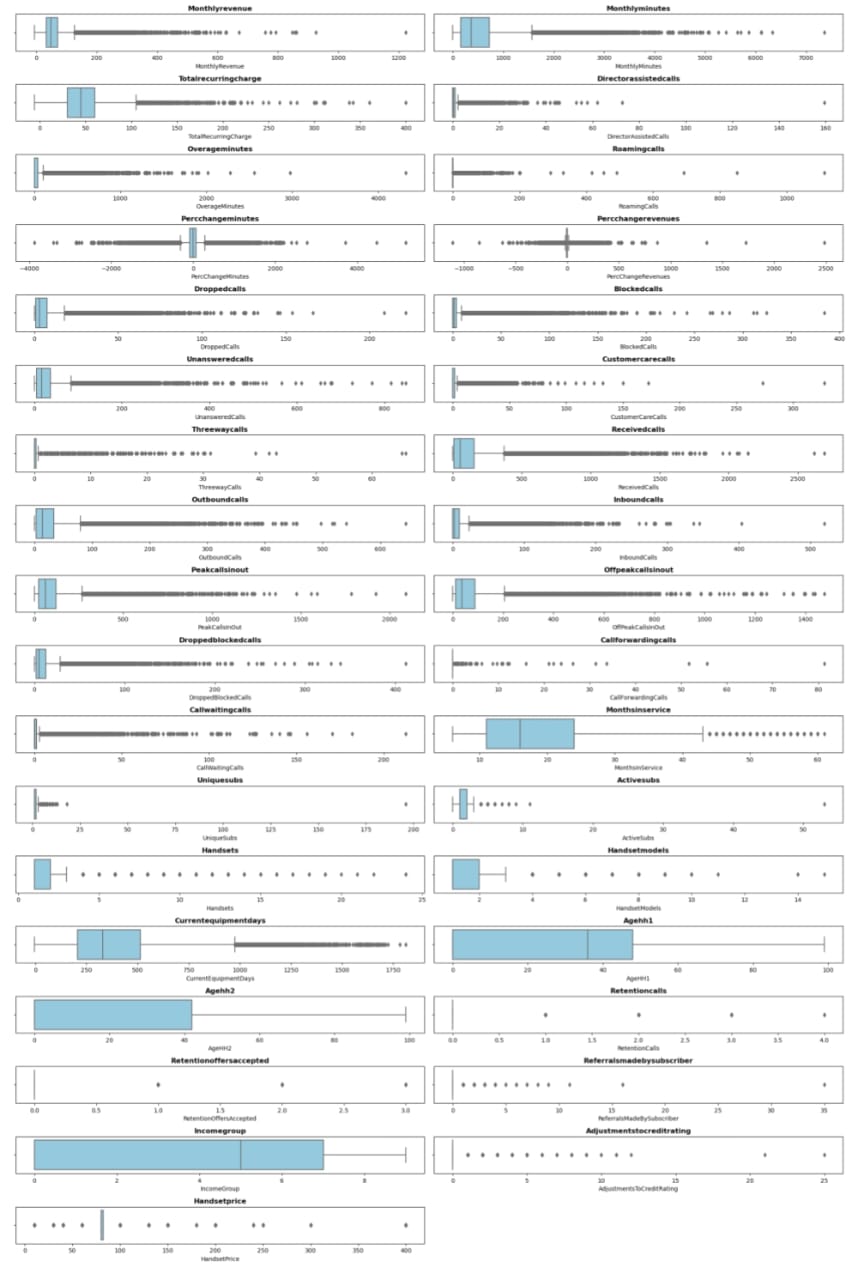
**Missing Values Treatment:**

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

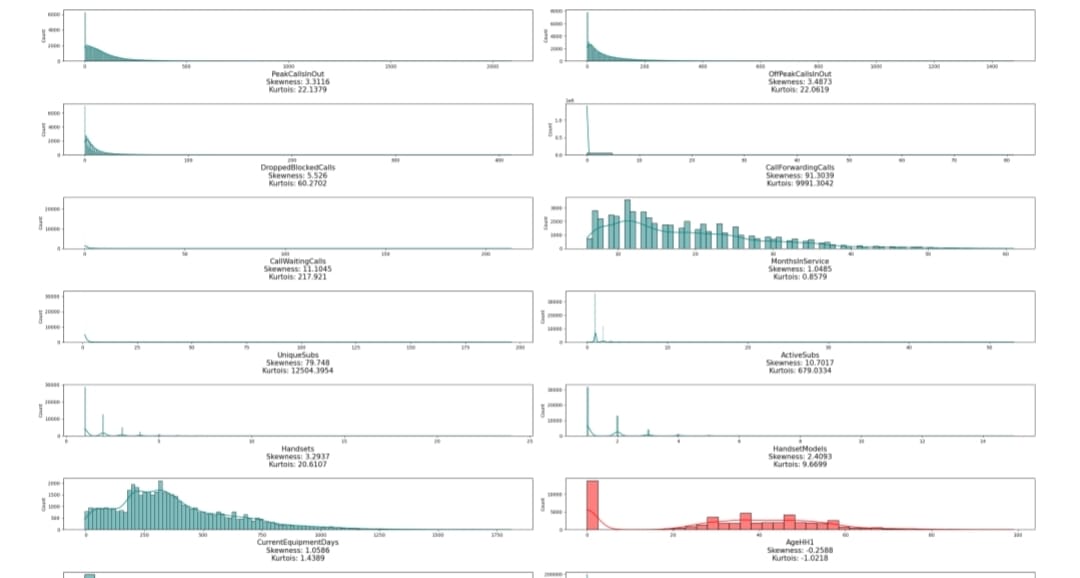
**Reason for outliers exist in the data:**

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

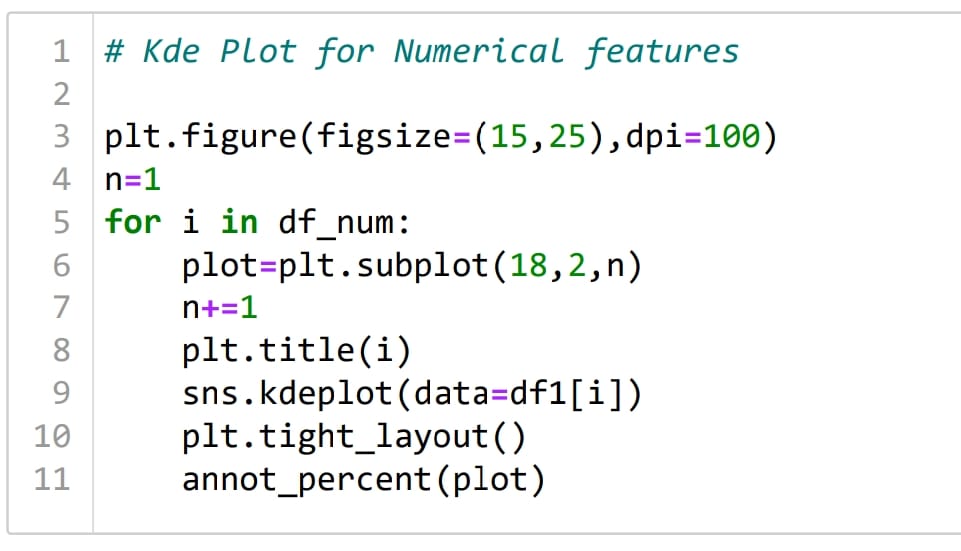
**Skewness After Transformation:**

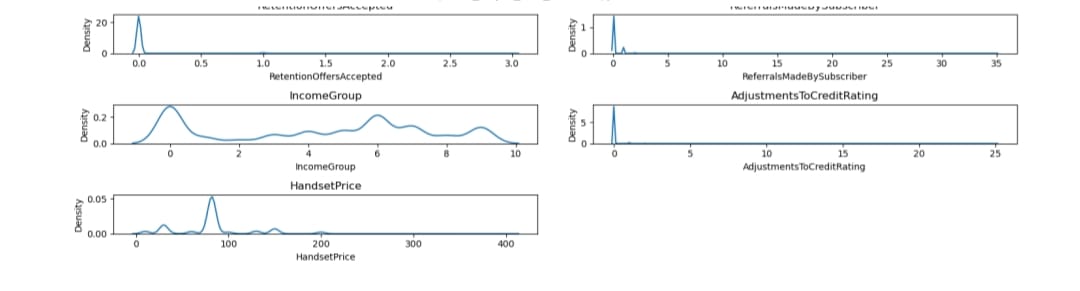


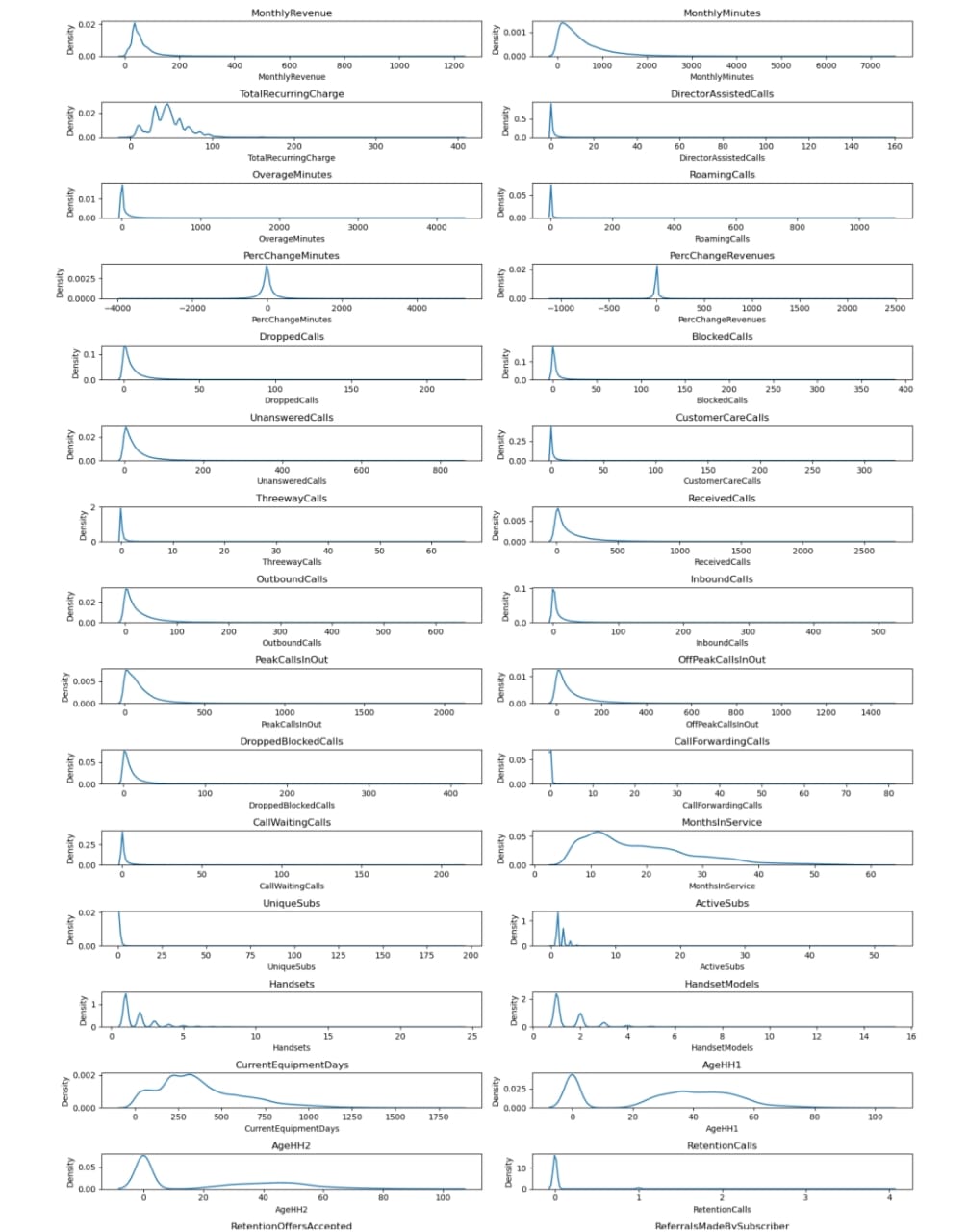
**Inference:** Here by visualizing dist plot we can observe that there is a reduction of skewness after Transformation.

**Descriptive Analysis (EDA)**

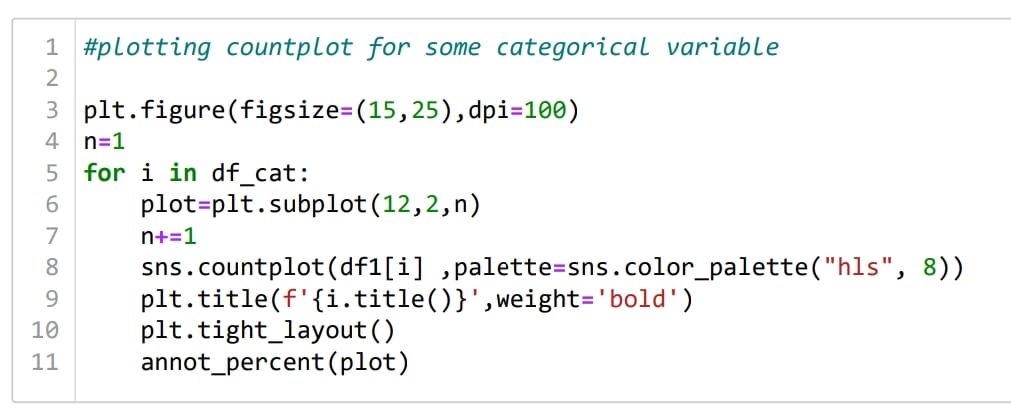
**Univariate Analysis:**

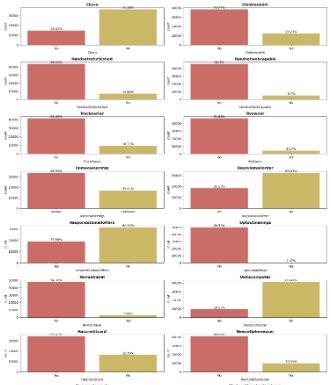
Numerical Columns Visualization:

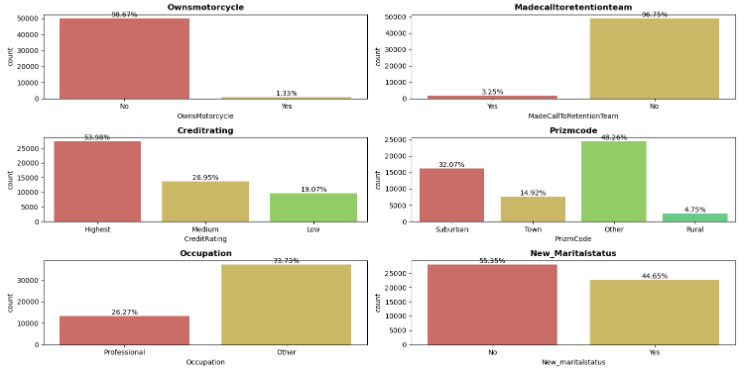




**Categorical Columns Visualization:**

****





#### 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

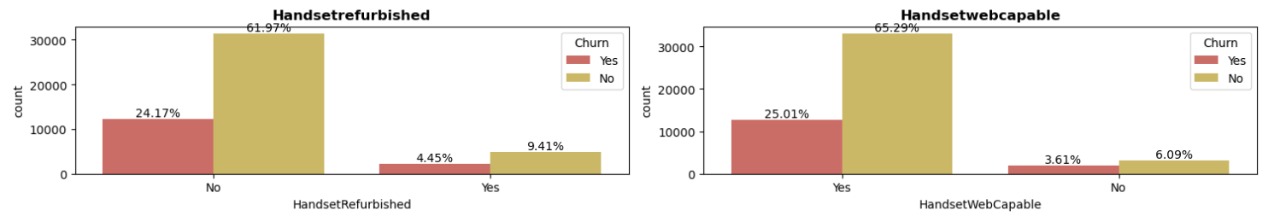
4) Less than 2 percent of them own a motorcycle

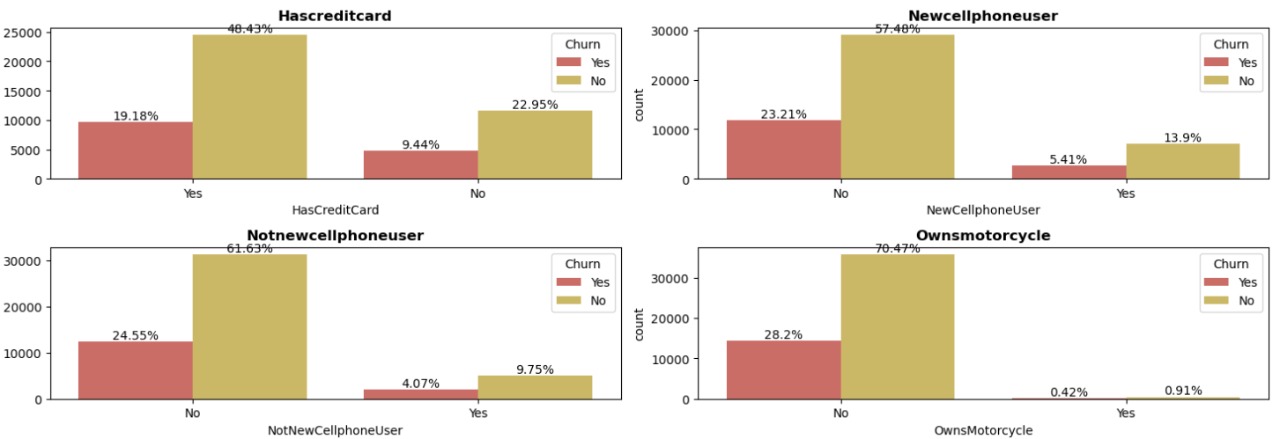
5) More than half of the people's handset price is unknown

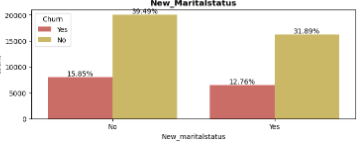
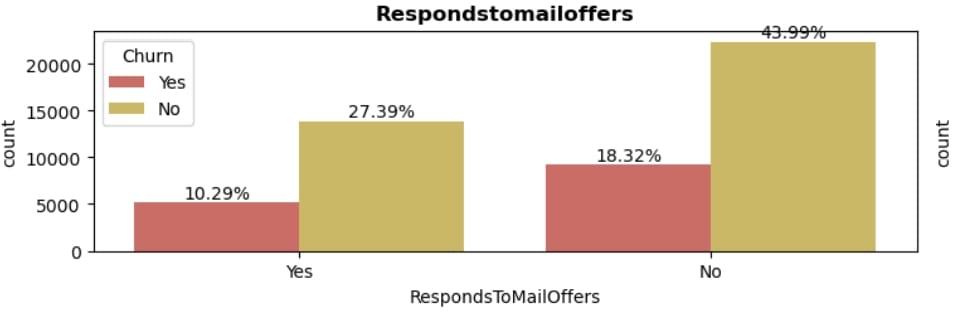
6) Over 70 percent of the data has occupations other than the ones mentioned.

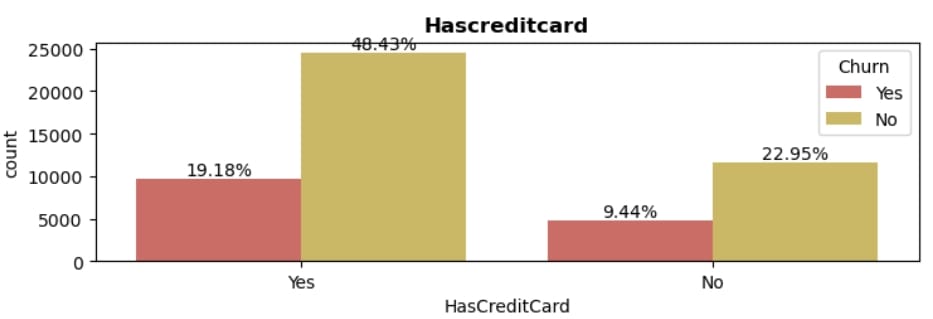
7) New Martial status around 45 percent are married.

**Bivariate Analysis:**

****



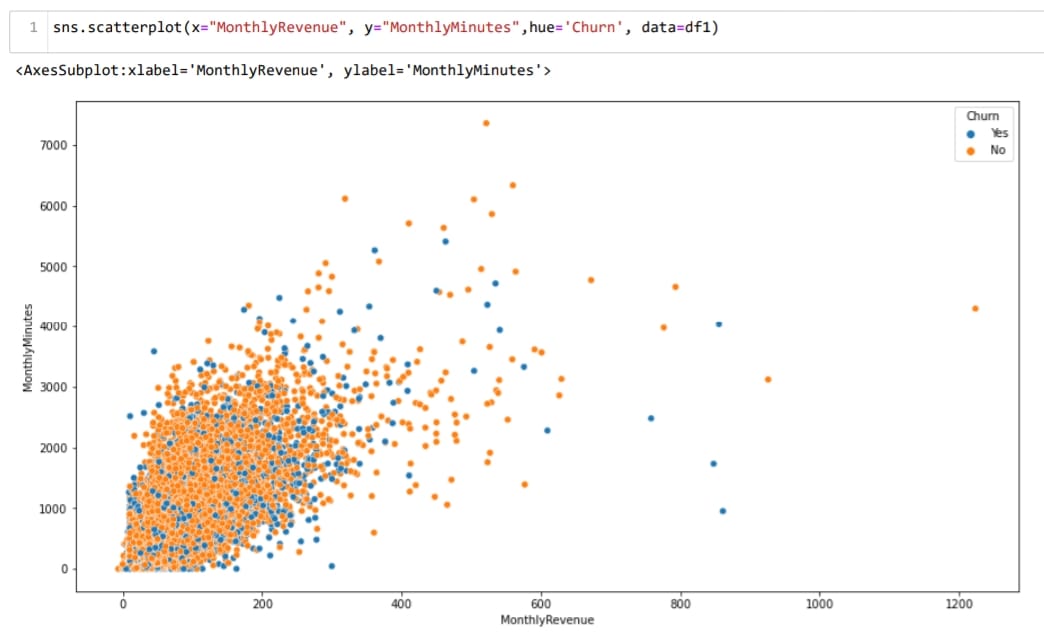




### **Observation:**

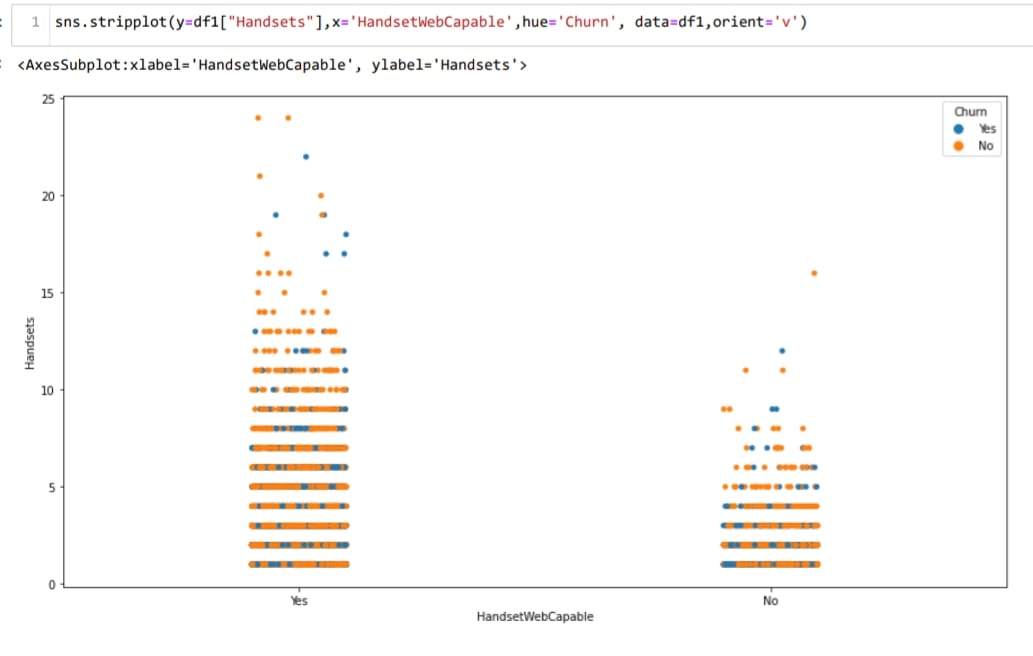
1. In Handset web capability over 25% of people who have churned has more than 90% of Internet capability on their phone.
2. Less than 6% of people who own new phone have churned.
3. Data shows that people who have Credit Cards are more likely to Churn
4. Marital Status of people churning is independent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

**Multivariate Analysis:**

****

### **Observation:**

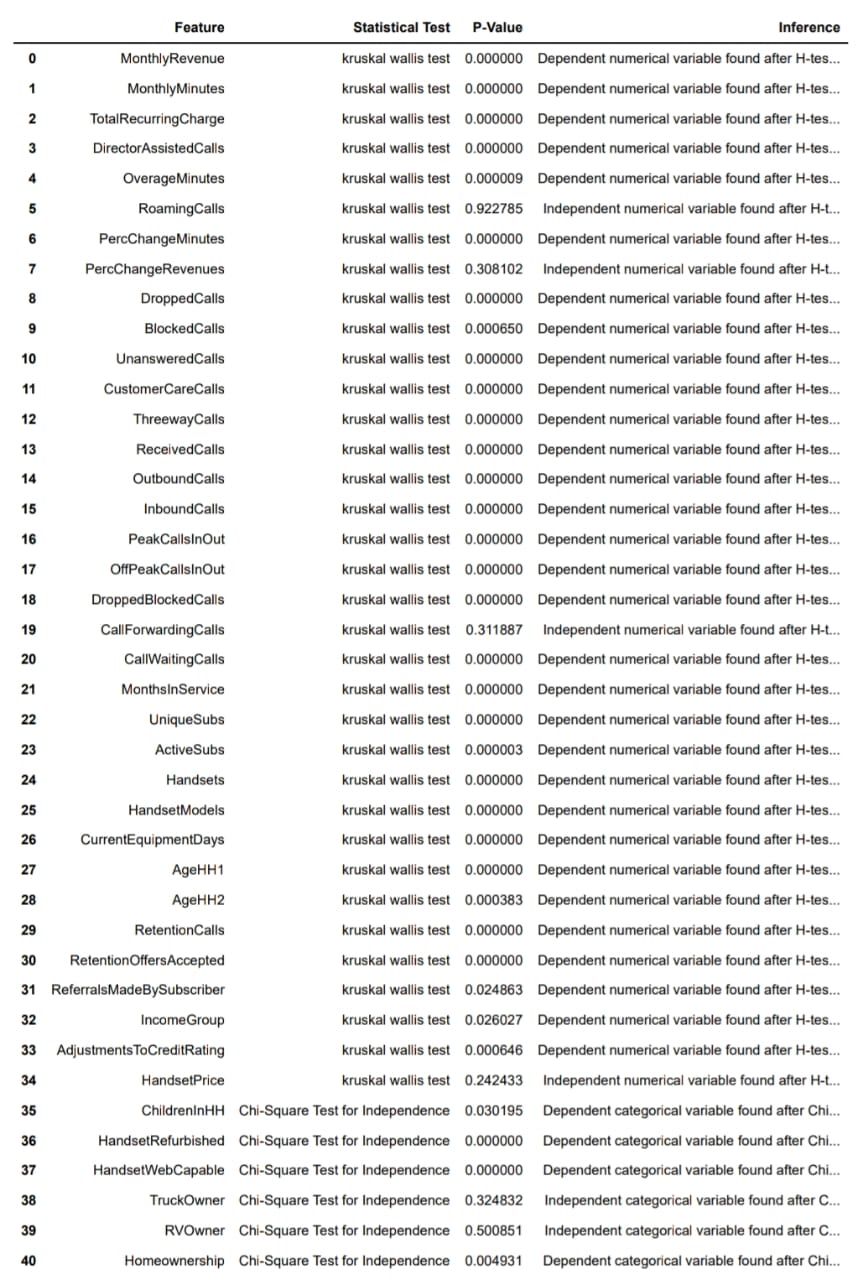
According to plot, as Monthly Revenue Increases, then number of Monthly Minutes increases, but we could not draw any conclusion on churn.

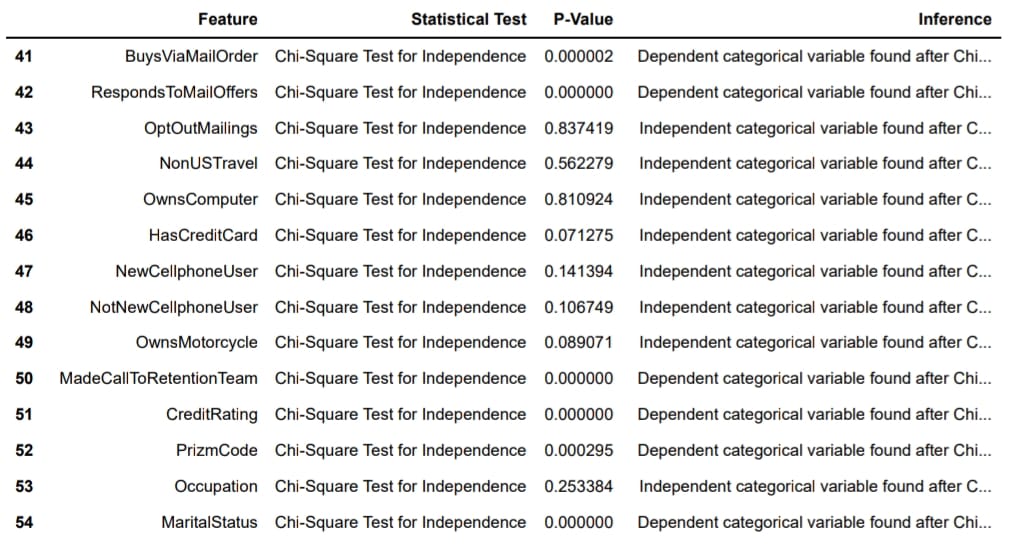


### **Observation:**

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).

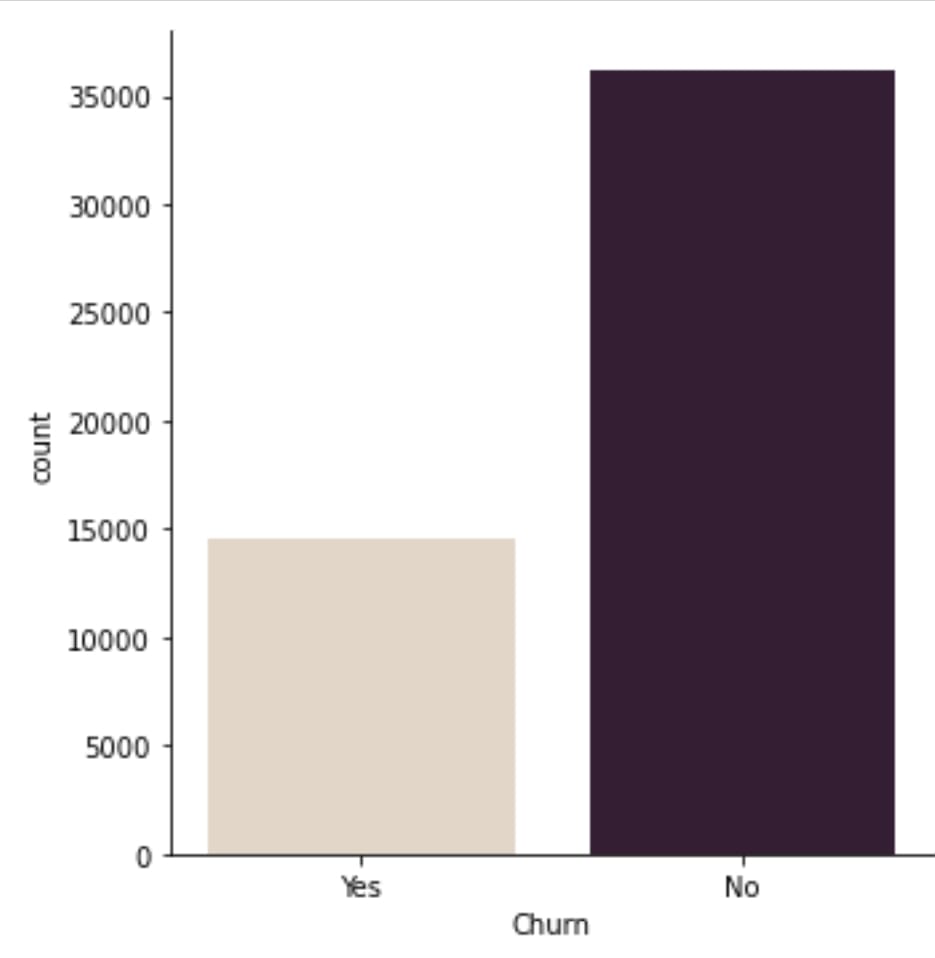
We have used Jarque-bera test to check the normality of data

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

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

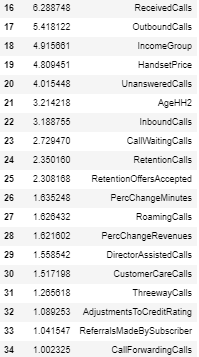
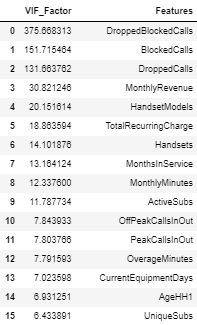
We found that data is not normal therefore we use Kruskal Wallis test to check its dependency on the target variable

**Class Imbalance and its Treatment:**

****

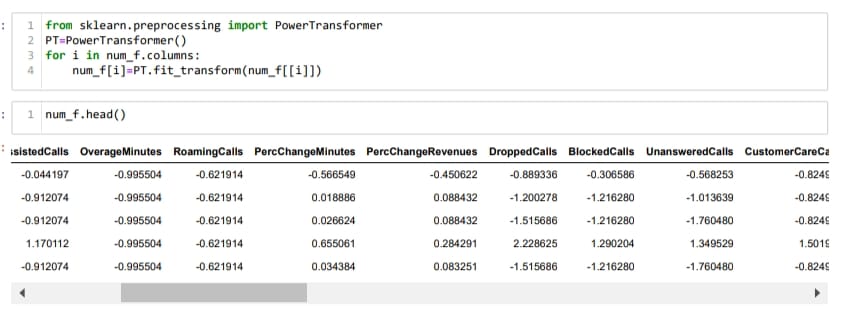
Here we can see that our target variable is too imbalanced, and to treat that we are going to use oversampling techniques like smote.

**Check of Multicollinearity:**

****

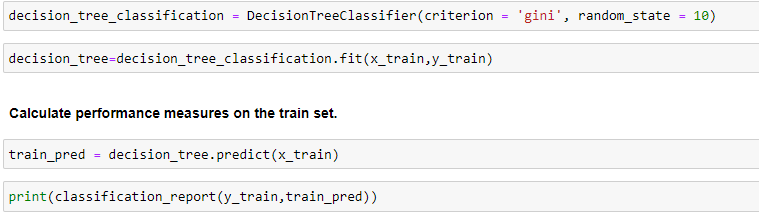
**Transformation:**

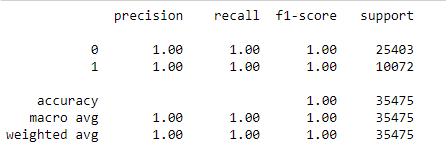
Transformation is a process that can be used to change the scale of the original data to get more accurate results. We used Power transformation, as we can see that there is large number of outliers present so we use Yeo-Johnson transformation technique to reduce the outliers and make the data more normally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



**Decision Tree (Base Model)**

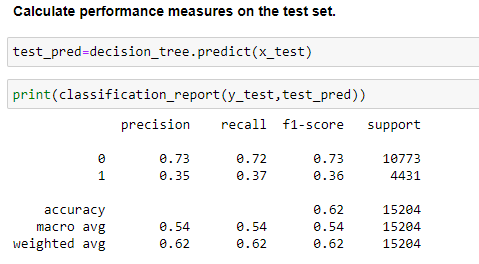
#### Build a full decision tree on a training dataset.





**Interpretation:** The performance measure on the train set model gives 100% accuracy.

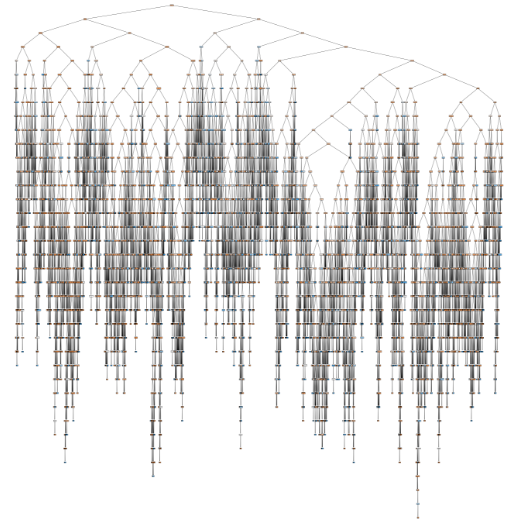
We will also calculate performance measures in the test set.



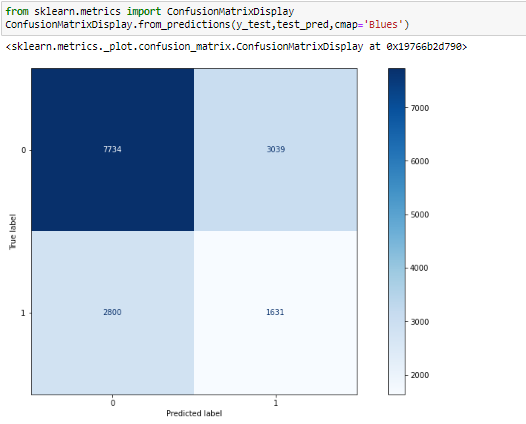
**Interpretation:** The performance measure on the test set model gives 62% accuracy.

We can clearly observe that the model is overfit. We will perform hyperparameter tuning and check the accuracy.

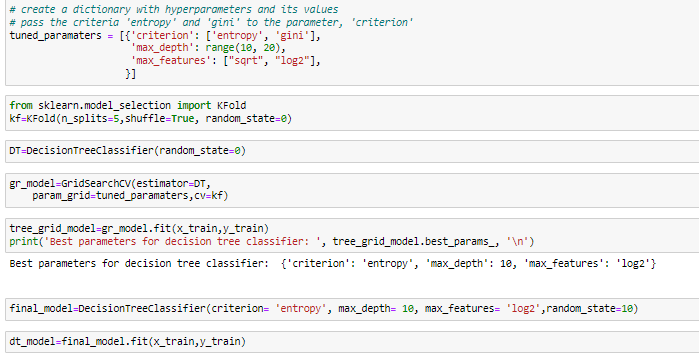
**Decision Tree:**



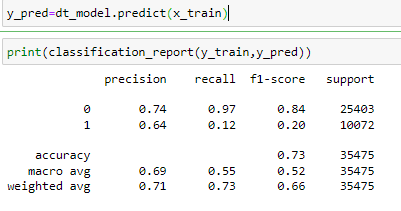
**Confusion Matrix:**



**Tune the hyperparameters using Grid search CV( Decision tree):**



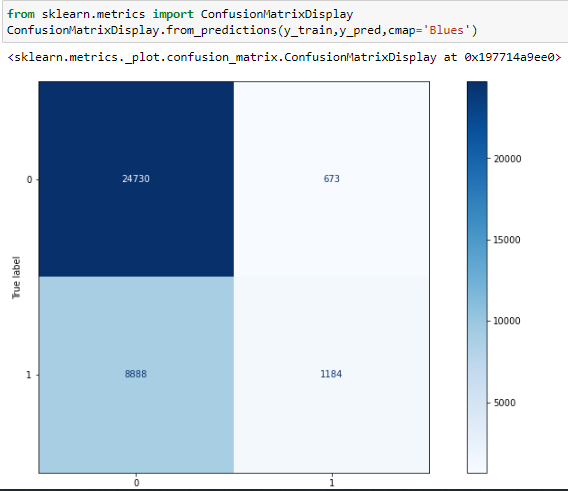
**Performance of train set after hyperparameter tuning:**



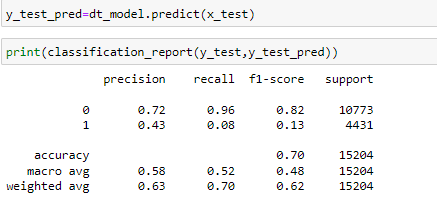
**Interpretation:**

From the above output, we can infer that the accuracy of the train set after hyperparameter tuning is 73%.

**Confusion Matrix:**



**Performance of train set after hyperparameter tuning:**



**Interpretation:** From the above output, we can infer that the accuracy of the test set after hyperparameter tuning is 70%.

# **Conclusions:**

* Base model performance is good, but it can be improved by different algorithms.
  + We have to implement various classiﬁcation machine learning algorithms and take feedback for them. This will take us to the best suited model