Customer Churn Analysis- MACHINE LEARNING PROJECT

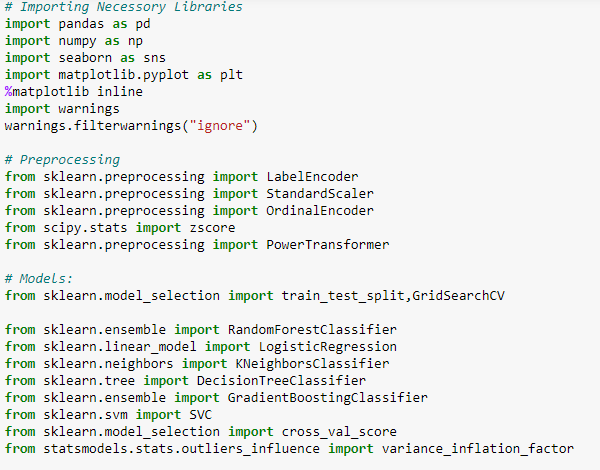
***Problem Statement***:

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals. Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base. Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

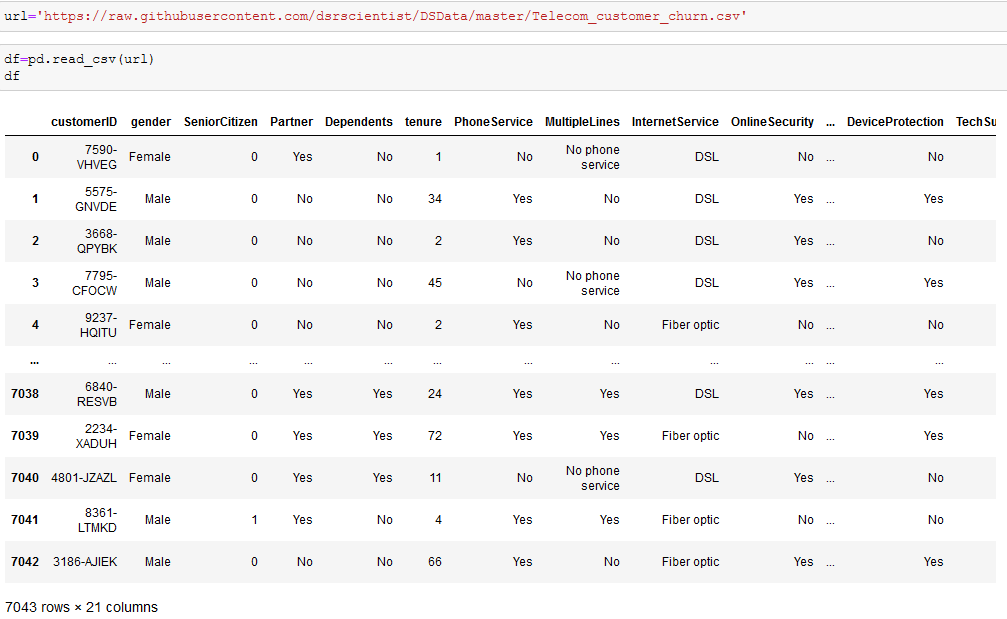
*Here we will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.*

Since we have to predict whether the customer will leave the service or not, we will have to build the classification model to predict it.

***Importing the Libraries, we will use:***

******

***Now we will Load the data***:



With the rapid development of telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, by collecting knowledge from the telecom industries can help in predicting the association of the customers as whether or not they will leave the company. The required action needs to be undertaken by the telecom industries in order to initiate the acquisition of their associated customers for making their market value stagnant. Our project proposes a new framework for the churn prediction model and implements it using by different machine learning models.

we have got one dependendent feature and different independent features for predicting the customer churn in the dataseet.Here the dependent variable, column Churn is our Target which contains only Yes and no,thus this problem is a binary classification problem.

The continuous features are:

1. Senior Citizen-This column suggests that any customer is senior Citizen or not.
2. Tenure – The tenure column suggests the number of months, the number of months the customer is attached with this company.
3. MonthlyCharges.- This column gives us the amount to be paid by the subscribers for monthly subscription.

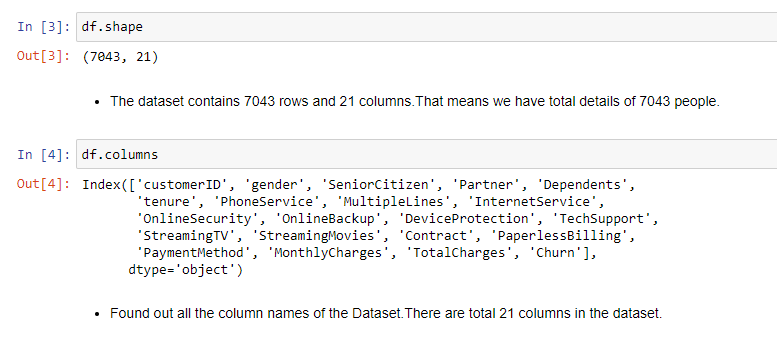
The categorical features are:-

1. Customer id- This is the unique id given to the subscribers for future identification or used while paying the monthly charges for continuing the subscription.
2. Gender- This column suggests that the subscriber is male or Female.
3. Partner- This column gives us the information about the subscriber is having any partner or not. It is Mentioned by Yes or No,
4. Dependents- It means that the subscriber is relying on someone,especially a family member for financial support or not. This column is mentioned by either yes or no.
5. Phone Service- This column gives us the information about the customer is having any phone service or not. This column is also mentioned by yes or no.
6. Multiple Lines- This column gives us the information about the subscriber is having multiple lines or not. This column is also mentioned by yes or no.
7. Internet Service- This column gives us the information about the internet service type used by the customers. The column comprises of two types of connection one is Optical fibre type and the other one is DSL (Digital Subscriber Line),which uses the local phone line to deliver high speed internet. Other type reassembles No, which suggests about the customer which do not have internet services.
8. Online Security- This column suggests that the customer is having online security system or not.
9. Online Backup- This column gives us the information about the subscriber is having Online Back up or not. It is just a strategy for sending a copy of a physical or virtual file or a database to a secondary remote or cloud system.
10. Device Protection- This column gives us the information about the customer is having a Device Protection or not.
11. Tech Support-This column gives us the information about the subscriber is getting any assistance from the tech support department or not, when they face any issues.
12. Streaming Tv- This column gives about the customers are having the Streaming Tv facilities or not.
13. Streaming movies- It suggests that the subscribers are availing the Streaming movies options or not.
14. Contract- This column suggests the type of Contracts availed by the customers. There are three types of Contracts: 1)Month-Month 2)One -year. 3)Two year.
15. Paperless Billing - Paperless billing helps **you eliminate extra paper mail while** also helping the environment. By going paperless, you can receive bills through email or by utilizing your free online account. You can also pay bills online or set up automatic payments. This column suggests that the subscriber is availing Paperless billing or not. (Yes or no).
16. Payment Method- This are the type of methods in which the subscribers pay their monthly subscription fee. The type of methods available are: -

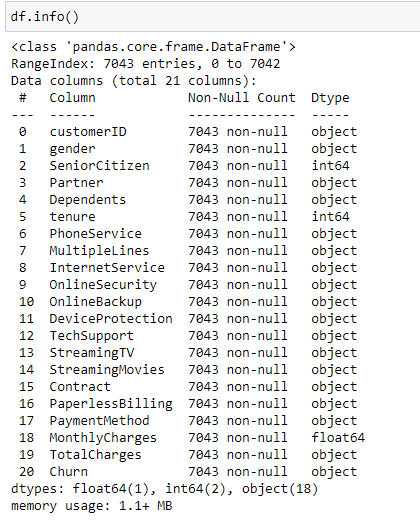
1)Electronic check. 2)Mail check. 3) Bank Transfer (Automatic).4) Credit Card (Automatic).

1. Total charges- This column is all about the Total charges a subscriber needs to pay for availing the services.
2. Churn – (the target column, yes/No)- Churn means to move out or cause to move out vigorously from a company. This is a serious major issue affecting the business of a company and a major challenge to predict or knowing about the customers why they tend to churn. In this dataset we will predict the customer churn rate with the above given features to us.

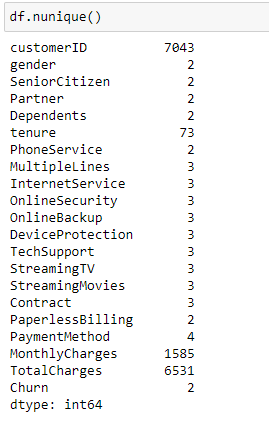
* ***Data Analysis:***



* We can see that the dataset contains 7043 rows and 21 columns. That means that we have the details of 7043 customers.
* In the above image, we used df. columns to find out all the column names present in this dataset.

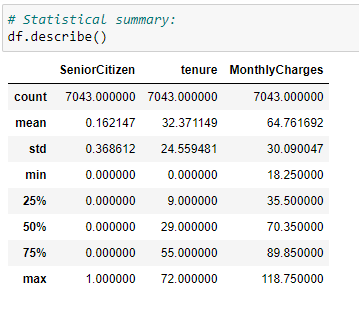


* We can clearly find the columns and their datatype by using df.info ().



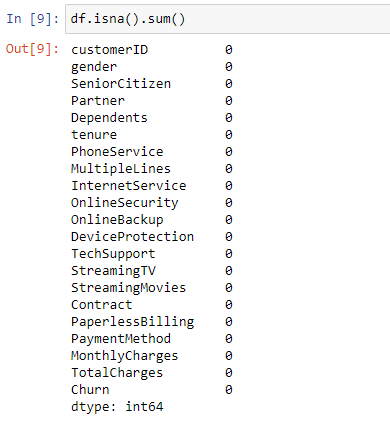
* There is total 7043 customer details in this dataset. ‘customerID’ has 7043 unique values. Our dataset has 7043 rows hence we can conclude that customer Id is unique to each customer and will not have any impact for our machine to learn and understand the feature. Hence going forward we will be dropping this feature.
* we can see the uniqueness of the features or the types of objects present in the features.

Checking the Statistical Summary:



* By df. describe (), we can check the statistical summary of the continuous features.
* We can only see the statistical summary of the continuous features, of the dataset which are SeniorCitizen, tenure and monthly charges.
* There is ony two values in Senior Citizen column, which are 0 and 1.
* The maximum tenure is 72 months and minimum are showing 0, this quite impossible because a telecom company always offers a customer with a tenure.
* The minimum monthly charges are 18.25 and the maximum monthly charges is 118.75.
* The inter Quartile ranges of Tenure and Monthly charges are good in numbers. There is no abnormality.

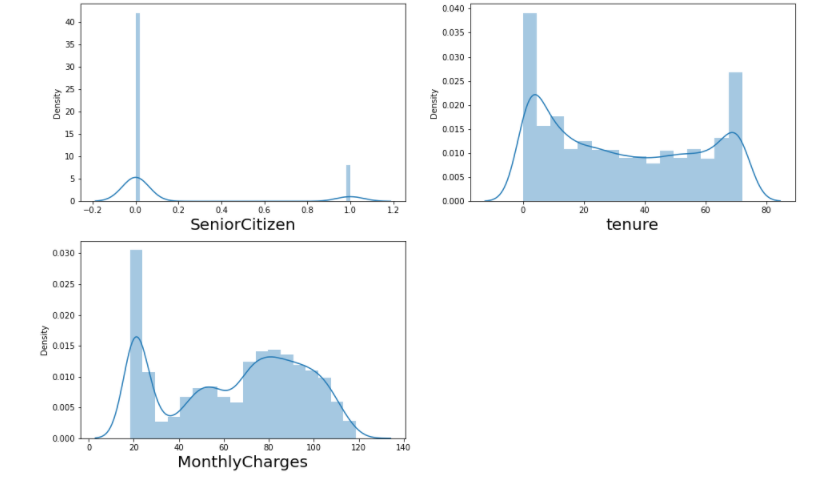
Checking for Null Values:



* There are no missing values present in the dataset.

**EDA:-**

**Checking the normal distribution of the continuous features:**

****

From the above diagrams we can conclude that the above continuous features are not perfectly in normal distribution. There may be skewness present in in it. we will check the skewness latter and if the skewness is present, we will remove it by any skewness removal techniques.

UNI-VARIATE ANALYSIS:

PLOTTING THE GENDER COLUMN:



Male subscriber is high in number, but the difference is very less.

PLOTTING THE SENIOR CITIZEN COLUMN:



As we have seen earlier that the senior citizen column consists of 0 and 1. As in machine learning encoding happens alphabetically, here 0 means No and 1 means yes. That means in this dataset some percentages of subscribers are senior citizens.

PLOTTING THE PARTNER COLUMN:



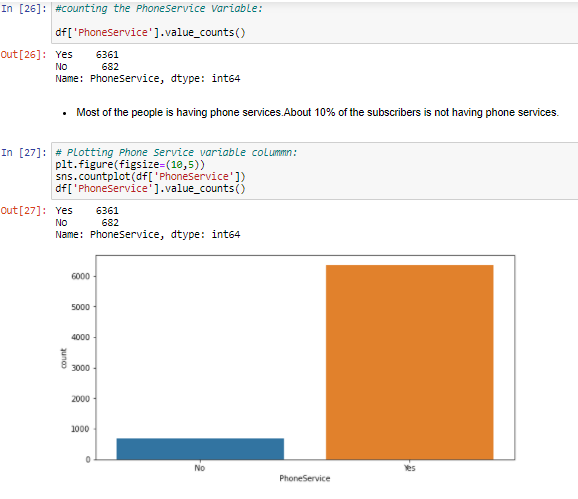
This column tells us about the subscriber is having any partner or not, from above it is clear that no indicates 3641 subscribers and yes indicates 3402 subscribers. The difference is low but most of the subscriber do not have any partner. As most of the subscriber do not have partner, cost may be the burden of oneself and this indicates a reason of churn.

PLOTTING THE DEPENDENTS COLUMN:



We can see that most of the subscribers do not have dependents that means the subscribers are not relying on someone, especially a family member for financial support. This is also a sign of churning process.

PLOTTING THE PHONE SERVICE COLUMN:



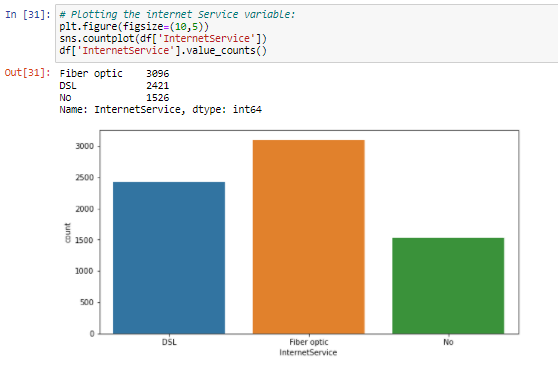
We can see that most of the subscriber’s phone service is in active state.

PLOTTING THE MULTIPLE LINES COLUMN:



Most of the subscribers don’t have multiple lines. The company need to look into this matter, that they can increase in providing the multiple lines to the customer.

PLOTTING THE INTERNET SERVICE COLUMN:



We can see that there are two types of internet connections availing by the subscribers one is Optical fibre type and the other one is DSL (Digital subscriber line). Most of the subscribers use optical fibre connections than the DSL connection, which is good sign because optical fibre connection is the upgraded version of internet connection,

Which can provide high speed internet connection. But the issue is 21% of the subscriber doesn’t have internet connection, which is important feature in our day-to-day life. The company must focus on that 21%, that why those customers doesn’t have internet connection.

PLOTTING THE ONLINE SECURITY COLUMN:



We can see that most of the subscribers don’t avail online Security

System, which is very important in today’s day to day life because most of the people do money online transaction frequently. If the online security is not provided the subscribers will be in great threat like monetary fraud, hence the company must increase the online security system to retain the existing customers.

PLOTTING THE ONLINE BACKUP COLUMN:



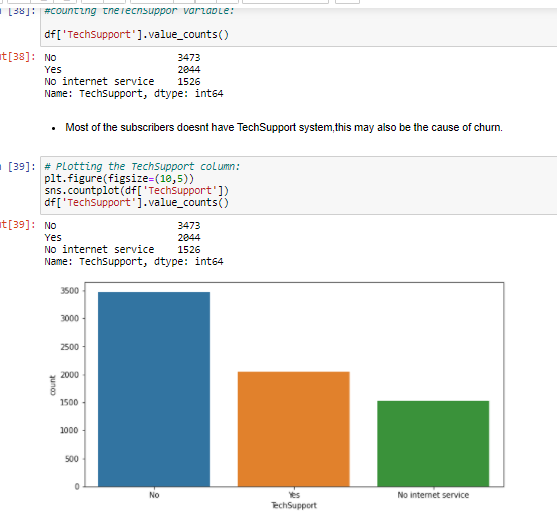
Most of the subscribers don’t have the online backup platform for storing their personal data, which is an important feature for the online community. This is also a major chance of customer churn; hence the company must increase the online backup facility to the subscribers.

PLOTTING THE DEVICE PROTECTION COLUMN:



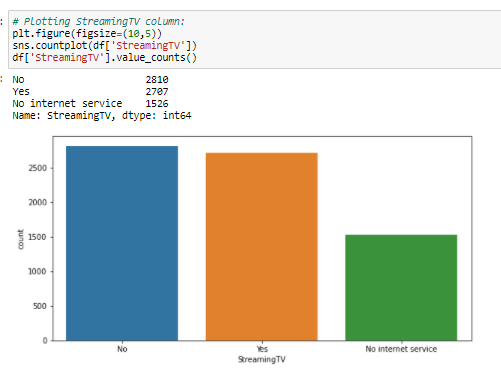
As online security is related to Device Protection, if online security is low, device protection must be strong. As we can see above that most of the subscribers having internet connection don’t have device protection, which is a great threat to the subscribers because their device is not totally protected. The company must keep an eye into it to secure the subscribers devices.

PLOTTING TECH SUPPORT COLUMN:



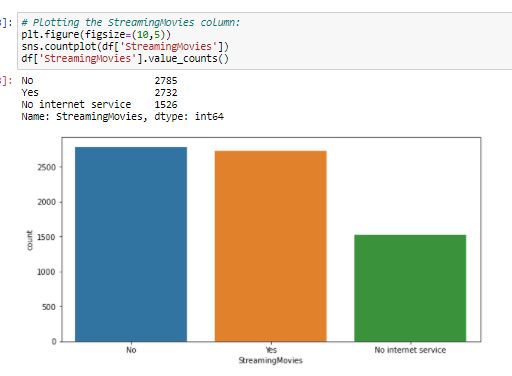
We can see clearly that most of the people don’t get the tech support help when they face any major issues, this may be the biggest cause for customer churn. Till now we found many reasons why customers may leave this company for getting better services available in the market.

PLOTTING STREAMING TV COLUMN:



This is also a major churn reasons because most of the subscribers don’t avail the streaming tv option, TV is a major component of entertainment and customers are always inclined towards entertainment.

PLOTTING STREAMING MOVIES COLUMN:



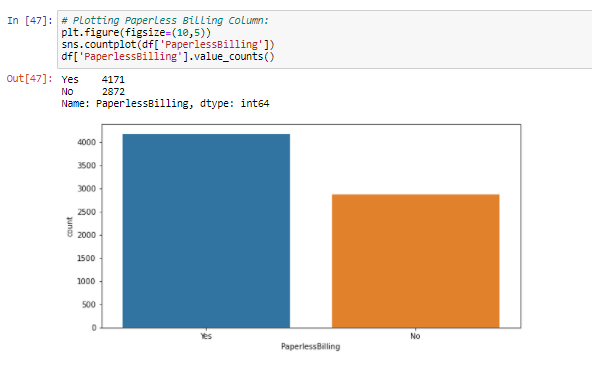
As streaming tv is related to streaming movies, we can see that approx. same amount of people doesn’t avail the streaming movies option. This is a negative sign because subscribers always love entertainment.

PLOTTING THE CONTRACT COLUMN:



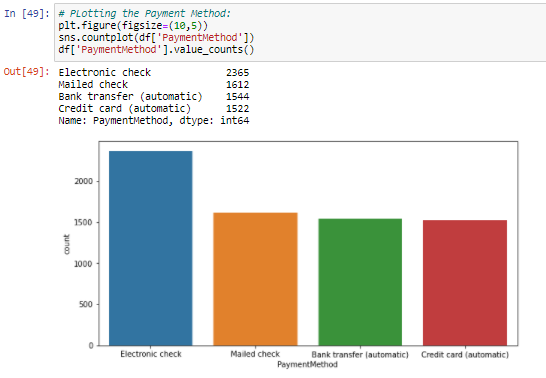
We all know that contract is an important attribute between a company and the customers. Here most of the subscribers falls in monthly contract, hence customer can leave the company if he or she gets better opportunity in the market. So, company must launch exciting offers for one year contract and two-year contract so the customers can stay with this company for a long run.

PLOTTING THE PAPERLESS BILLING COLUMN:



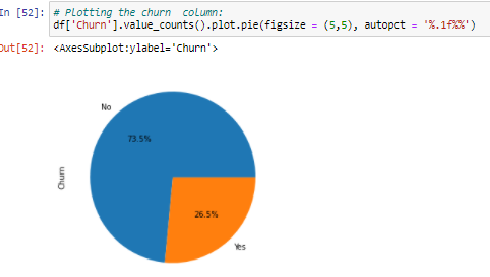
Paperless billing helps you eliminate extra paper mail while also helping the environment. By going paperless, you can receive bills through email or by utilizing your free online account. You can also pay bills online or set up automatic payments. This is a good sign that most of the subscribers use Paperless Billing Services.

PLOTTING THE PAYMENT METHOD COLUMN:



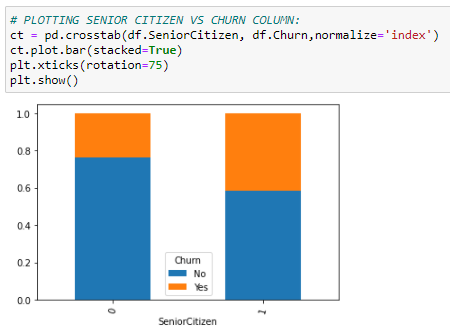
We can see most of the people use electronic cheque for payment. The company can focus on increasing the payments through Bank Transfer and Credit Card, which is in automatic mode and will be carefree for the customers while paying. They can do it buy giving offers in online transactions.

PLOTTING THE CHURN COLUMN:

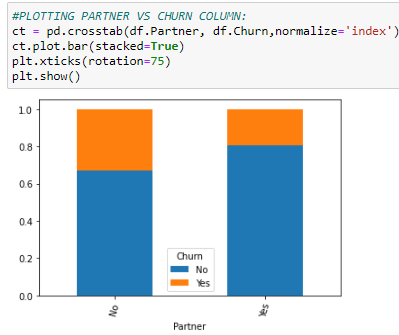


We can see that the customer churn rate of this company is 26.5%, which is very high and it directly effects the growth of the company.

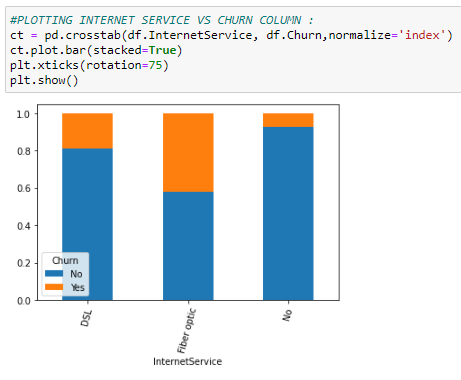
BI-VARIATE ANALYSIS:



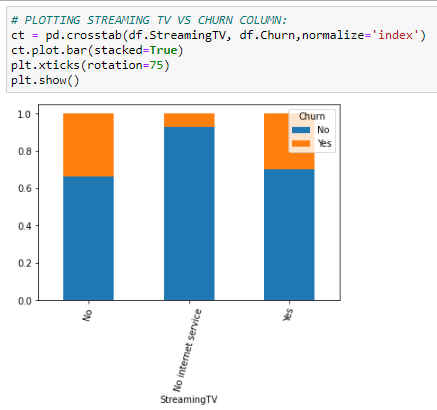
Senior citizen is more likely to leave the services. The company may give some senior citizen subsidies to retain the senior citizens.

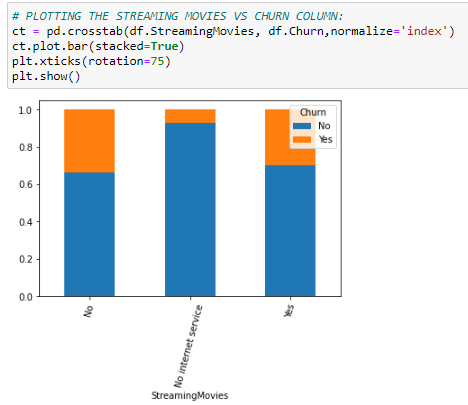


Having partner make them less likely to leave the service because dependency increases.

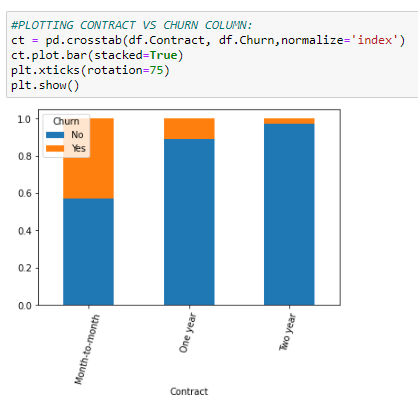


Having No internet service can make the user stay with the service.





We can see in both the cases with streaming tv and streaming movies,’ no internet service’ have less chance in leaving the services. I guess is there any thing wrong with the internet services? We can also see that churn is more among the users whose internet services are fibre optics or DSL. the telecom company needs to check if the quality of service needs to be improved or not.



As mentioned above it is clear yet again that having a long contract can make the users stay with the services so the company must make the long-term plan more attractive so that the customer can grab it in a higher manner.

EDA concluding remarks:

1. Have more flexible plan for senior citizen. Maybe they are not the one who uses internet services so much hence telecom should introduce more plans suiting their needs.

2. Customer who are single and do not have dependent, company should introduce more flexibility for such users.

3. We have high number for churn from the customer who uses Fibre optics as internet services. Company should identify the reason for them leaving the services.

a) Is it because the price is not competent with other service providers?

b) Are we able to meet the customer expectation in terms of quality that they expect?

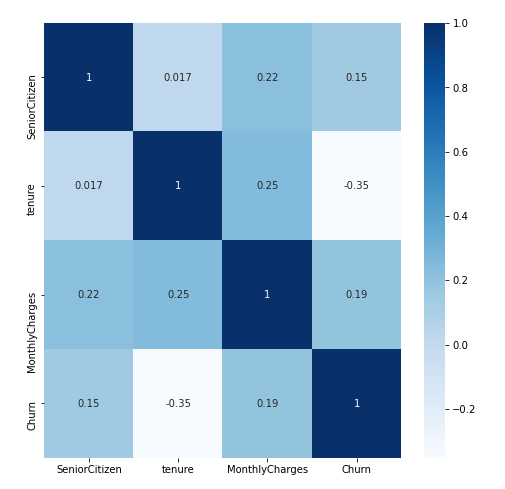
c) Any other feedback through survey we can identify.

1. Introduce more lucrative long terms plans as churn is less for the customer who are on 2-year contract. We need to highlight the services, benefits and discount they enjoy if they are on a long-term plan.

PRE-PROCESSING PIPELINE:

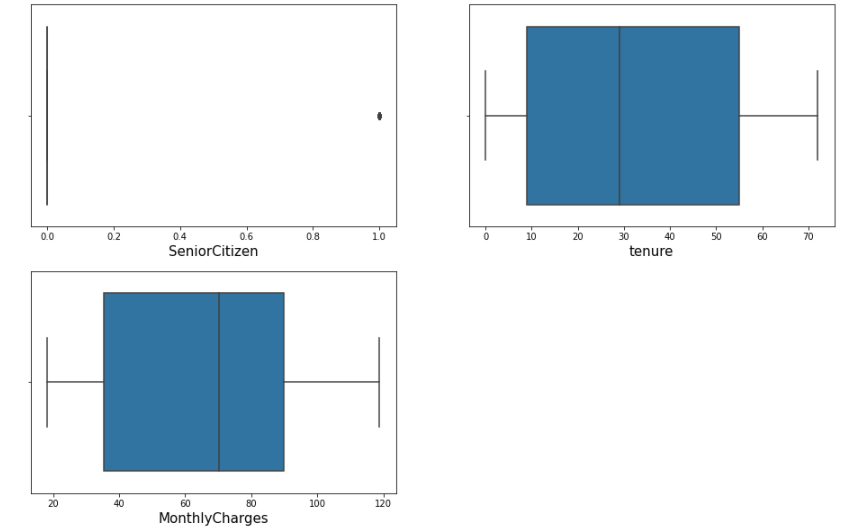
To build the model we need to check for few things:

**CHECKING MULTI-COLINIEARITY**-(Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.)



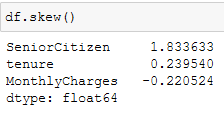
We can see that most of the features are categorical and we have few data to check the multicollinearity. from the above heat map, it is clear that none of the numerical features are co related to each other.

**CHECKING OUTLIERS**: (Outlier is a datapoint that differs significantly from other observations).



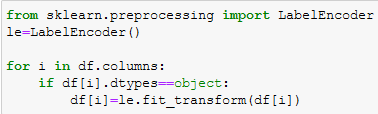
There is no outliers present. Since senior citizen has only 2 variables,1 is showing as outliers which is not correct.

**SKEWNESS:** (Skewness is the measurement of the symmetry of the distribution).

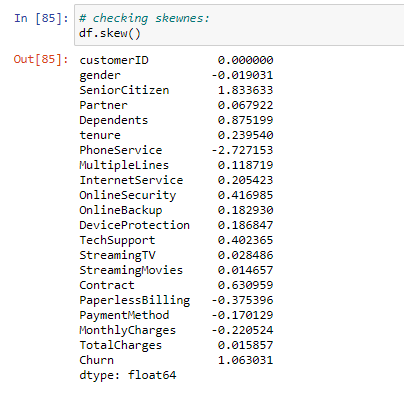
There is no skewness present in the dataset. Senior citizen is more of the categorical data rather than numerical because it has only two values 0 and 1.

**LABEL ENCODING:**

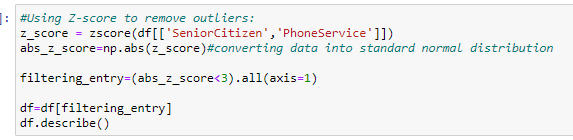
Since we have to pass the numerical data to machine to learn, we will now change the categorical features to numerical using Label Encoder.



All the columns are converted into numerical data. Now it is ready to implement this data to put into the model for further prediction. We will again check the skewness because the data is now converted into numerical.

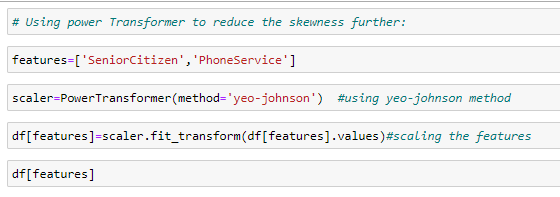


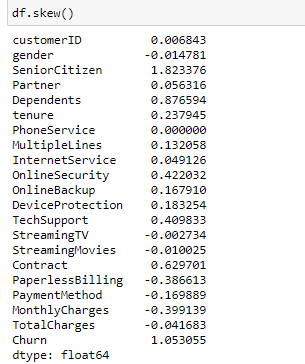
we can see that phone service column and Senior Citizen columns are highly skewed. We will use z-score to remove the skewness.



We can see that there is still skewness in phone service column and senior citizen column. we will use power transformer to remove the skewness.

Now we will use the Power transformer to reduce the skewness:





Now the skewness is removed and thus the dataset is ready for building models for prediction.

**Imbalance Data**

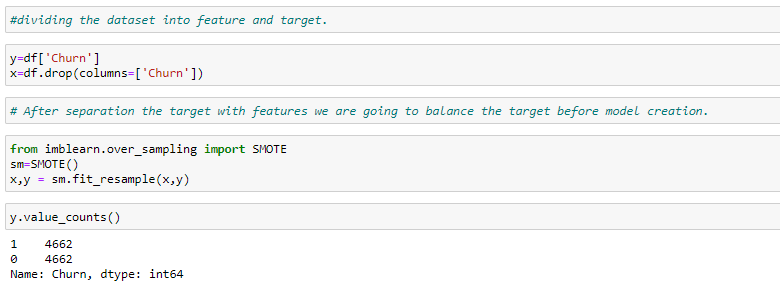
Our target variable ‘Churn’ is not distributed evenly. Having a balanced data set for a model would generate higher accuracy models, higher balanced accuracy and balanced detection rate. Hence, it’s important to have a balanced data set for a classification model.

There are two techniques we can use, either we use under sampling or oversampling of data.

The simplest implementation of over-sampling is to duplicate random records from the minority class, which can cause overfitting.

In under-sampling, the simplest technique involves removing random records from the majority class, which can cause loss of information.

For this project I will be using oversampling using SMOTE. But before that we will have to separate the input and target variable.



**Standardizing**

Data standardization is about making sure that data is internally consistent and it comes to the common scale.

We will use StandardScaler to standardize the data.

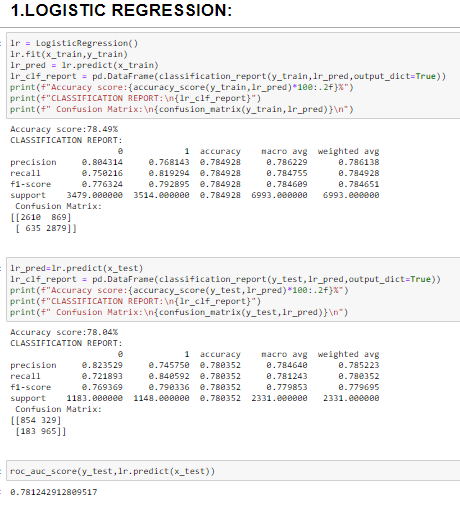


**HOLD- OUT -METHOD-** The holdoutmethod is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only.

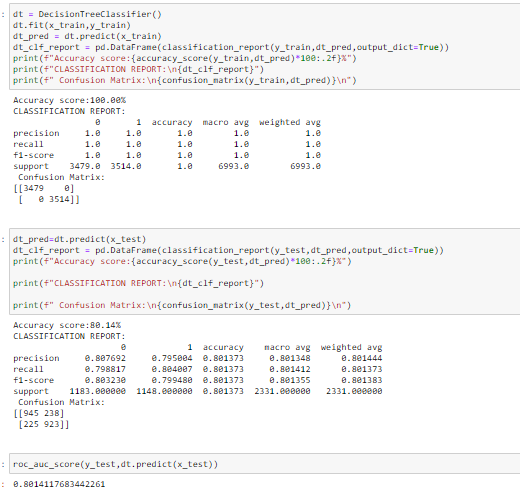
****

**BUILDING MODELS:**

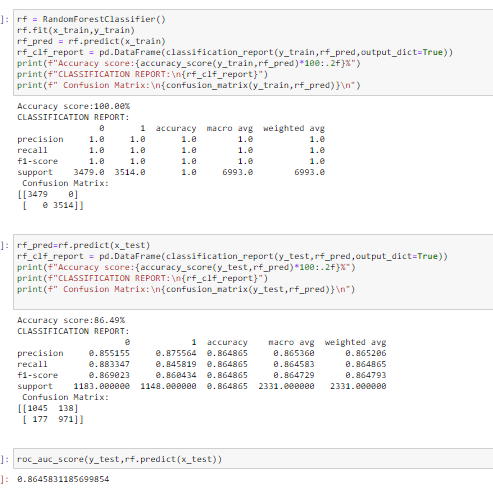
**1.LOGISTIC REGRESSION:**

****

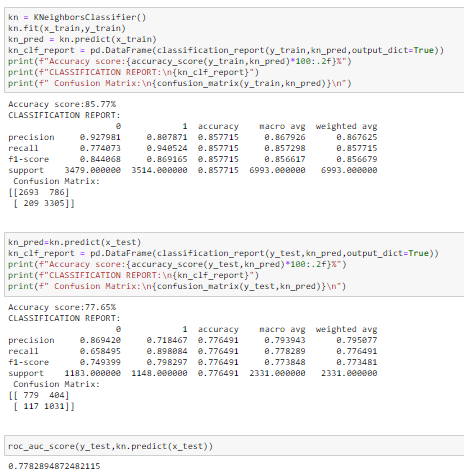
**2. DECISION TREE CLASSIFIER:**

****

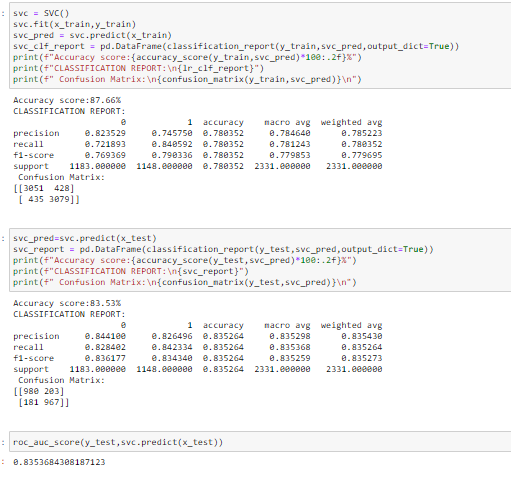
**3.RANDOM FORREST CLASSIFIER:**

****

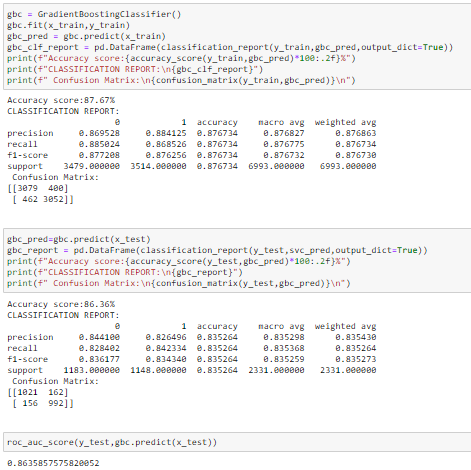
**4.K-NEIGHBORS CLASSIFIER:**

****

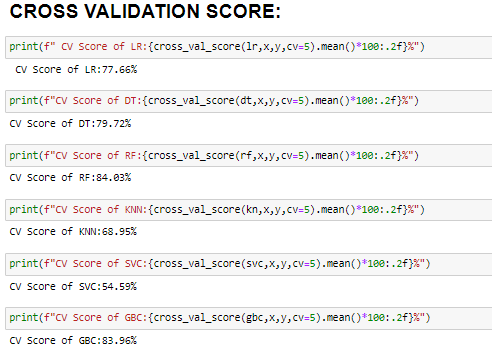
1. **SUPPORT VECTOR CLASSIFIER:**

****

**6.GRADIENT BOOSTING CLASSIFIER:**

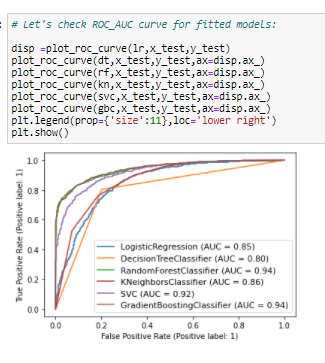
****

**CROSS VALIDATION SCORE-** We will use cross validation method to check the performance of the model for underfitting and overfitting. It is important particularly in a case where the amount of data may be limited. IT is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modelling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

****

**ROC-AUC CURVE:**

AUC-ROC curve helps us visualize how well our machine learning classifier is performing. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

****

**CONFUSION MATTRIX:**



A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

The first row is classified as Not leaving the company – 1045 customers were correctly classified, called as True Positive and 138 were wrongly classified as leaving the company (False negative).

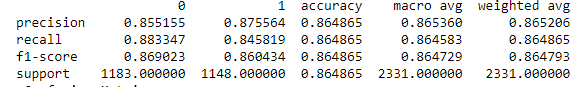
The second row is about the leaving the company – 177 customers were wrongly classified as leaving the company (false negative) and 971 were correctly classified as leaving the company.

A confusion matrix gives you a lot of information about how well your model does, but there is a way to get even more, like computing the classifiers precision.

**Precision and Recall**

Precision – 0.86

Recall – 0.86



Precision (also called positive predictive value) is **the fraction of relevant instances among the retrieved instances**, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.

Our model predicts 86% of the time customer retention correctly (precision). The recall tells us that it predicted the retention of 86% of the customer who actually did not churn.

**F1 Score – 0.86**

You can combine precision and recall into one score, which is called the F-score. The F-score is computed with the harmonic mean of precision and recall. Note that it assigns much more weight to low values. As a result of that, the classifier will only get a high F-score, if both recall and precision are high.

From the above metrics we can conclude that:

1. DTC have performed well with least difference on CV score.

2. RFC has given us the best accuracy of 86.4% CV score of 84% and ROC AUC score of 94%. we will do hyper parameter tuning of this model.

3. KNN has test accuracy of 77% CV score of 68% and ROC AUC score of 86%¶

4. DTC has test accuracy of 80% CV score of 79% and ROC AUC score of 80%.

5.SVC has test accuracy of 83.53% CV score of 58.9% and ROC AUC score of 92%.

6.. Logistic Regression has test accuracy of 78% CV score of 77.9% and ROC AUC score of 85%.

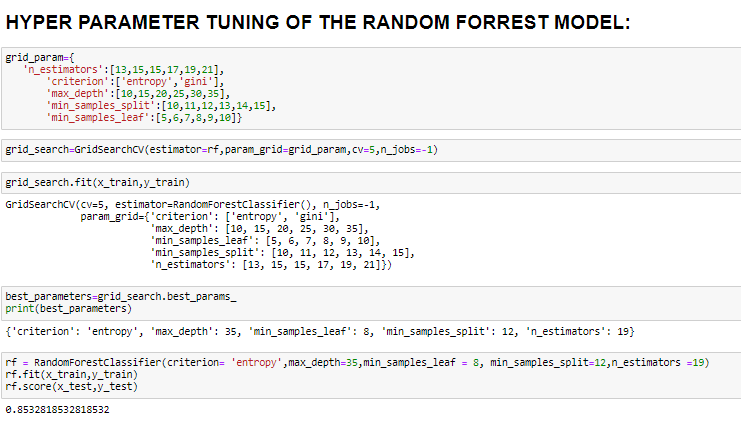
7.Gradient Boosting classifier model has test accuracy of 78% CV score of 83.96% and ROC AUC score of 94%.

8. None of the model is overfitting.

We will do Hyperparameter tuning of the Random Forest Model because we got the less difference between the accuracy score and cross validation score, and the accuracy score, cross validation score and ROC AUC score is higher among the other models.

**HYPER PARAMETER TUNING: -**

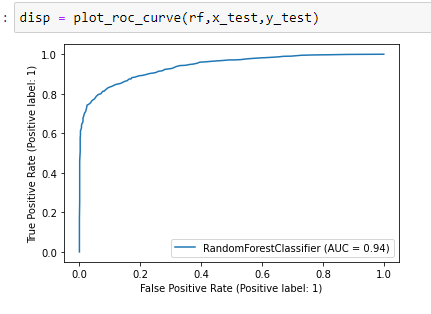
Hyperparameters are crucial as they control the overall behaviour of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

****

To select the final model, we don’t only have to look only at the test accuracy score, we also need to check the Cross validation (CV) score. The least difference between the test and CV score indicates that the model is performing well without being underfit or overfit.

Although the KNearestNeighbors model has performed well with the least difference between test accuracy and CV score, we select RandonForestClassifier as our final model. It has the maximum test accuracy of 86.4% with CV score of 84% and also the AUC ROC score of 94%. After hyper parameter tuning, we got the score of 85.32%.

PLOTTING THE ROC AUC CURVE:



**SAVING THE MODEL:**

****

**Summary**

We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data pre-processing part, we converted features into numeric ones. Afterwards we started training 6 different machine learning models, picked one of them (Random Forest) and tuned its performance through optimizing it’s hyperparameter values and finally selected the RandomForestClassifier.

**Concluding Remarks**

There is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features. The performance of the model can still be improved.

Telecom Company should also look at the analysis about the customer leaving the services. It can conduct various surveys to understand the customer behaviour and implement the necessary changes to its services.

The predictions from the ML model can help in understanding the customers who might leave and their service. With this information the company can do the following:

Relationship Managers (RM) get a daily feed on who has the propensity to churn and what are the influencing factors:

This can help trigger a conversation with the customer and understand their pain points and possibly fix the situation even before the churn occurs.

This can create more customer entanglement as the RM would reach out to them anticipating issues will eventually increase their lifetime value.

The model can be deployed to run on a regular basis to understand the changes in the behaviour of the customers and the relationship managers can act accordingly.

Collect more data from customers going forward such as reviews, ratings etc. to understand more about why a certain group of customers might be leaving.

Capture the geographic information of the customer and correlate with some of the demographic attributes to find out patterns around customer churn and geography.

By deploying data systems which can capture all these data sets we will be able to understand a complete profile of each customer and their activity and not only reduce customer churn but can also drive incremental revenue.