**Customer Churn Analysis – Machine learning Project**

1. **Problem definition**

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.

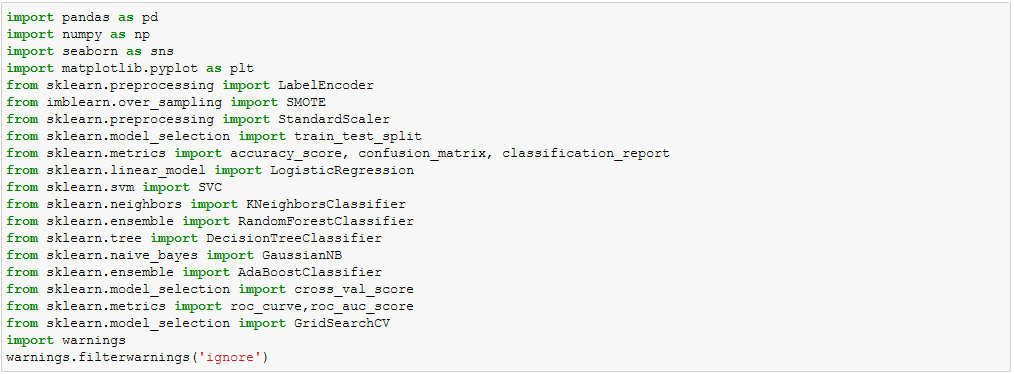
Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to 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.*

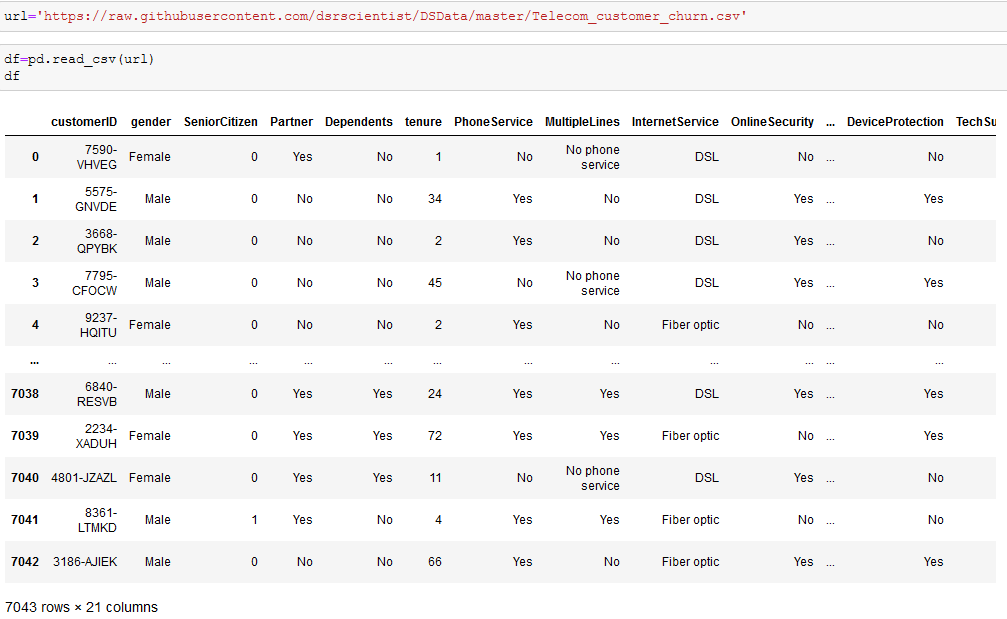
*Here we will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.*

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.**



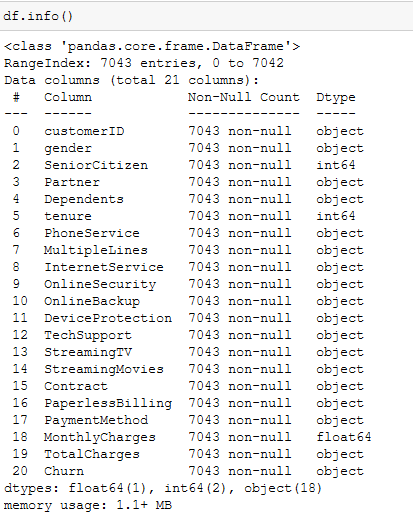
Let’s load the data



The dataset has 7043 rows and 21 columns.

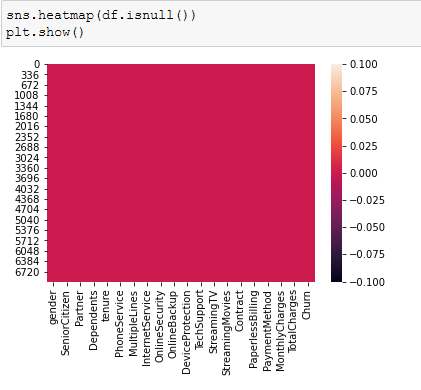
**2. Data Analysis**

Let’s see all the features we have in the dataset.



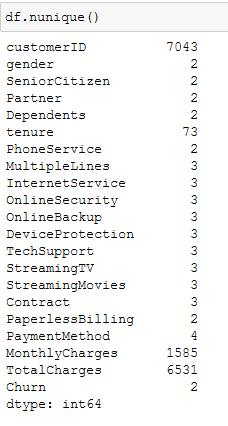
Here is our data. We can observe it has 3 numeric features and 18 categorical features. Our target variable ‘Churn’ is a categorical feature, indicating the use of building the classification model to predict it.

Let’s check for any missing values.



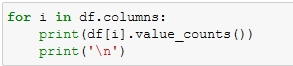
From the above heat map we can conclude that our dataset does not have any missing values.

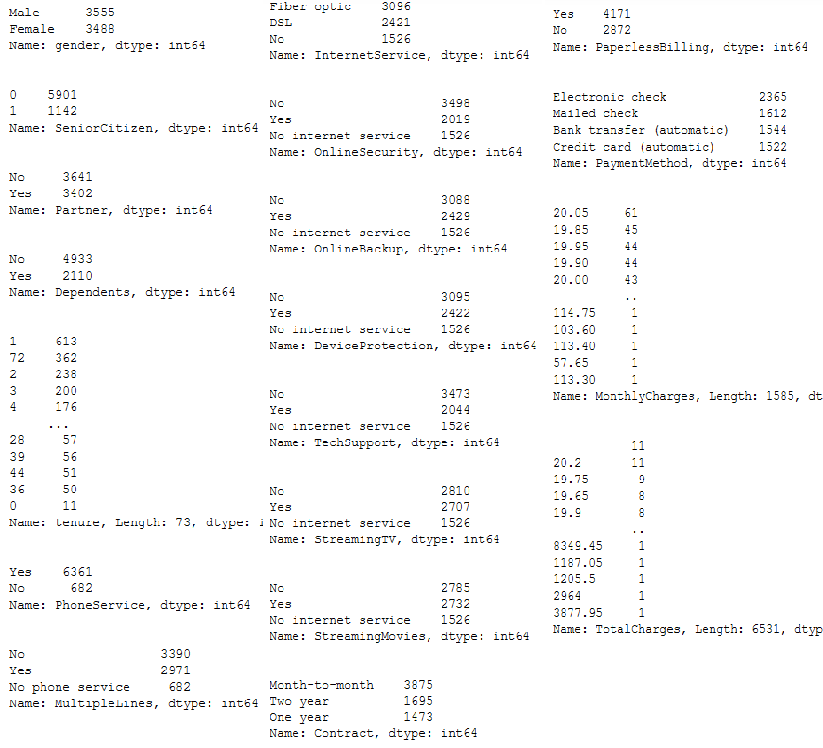
Moving forward we will check the number of unique value each feature has.



‘customerID’ has 7043 unique vales. 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 will now check the values each features are holding.



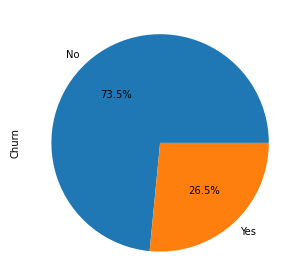


So only Tenure, Monthlycharges and Totalcharges have continuous values. Rest all the features have either 2 or 3 value present.

Let's check for each feature present in the dataset and its effect on our target variable. This is very important as we will understand customer behaviour through different patterns and will help the company to fill the gap that has been created resulting in the customer leaving the services.

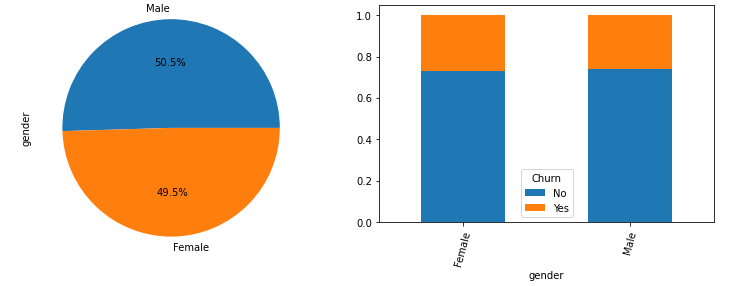
**3. Data Pre-processing and EDA**

**Churn**



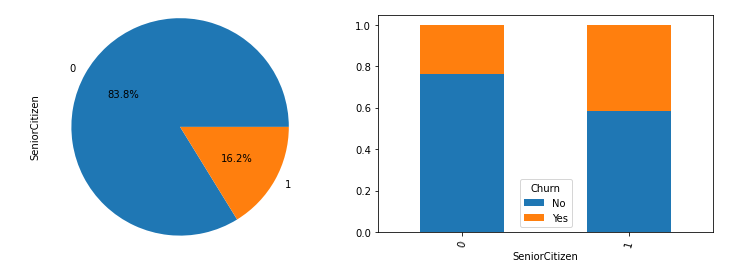
This is our target variable which we need to predict. In this dataset we have 26.5% of the customer who has left the service.

**Gender**



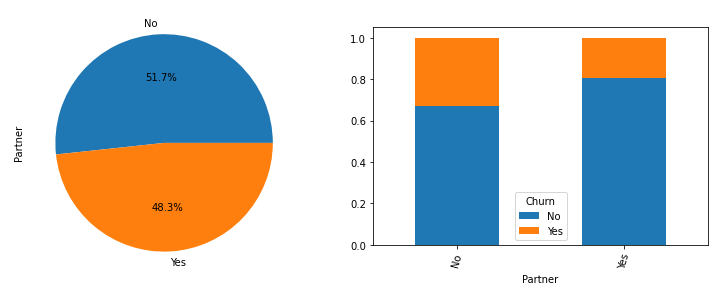
Male and female are equally distributed and if we compare it with Churn, gender does not play an important role to decide whether the customer will continue to use the services.

**Senior Citizen**



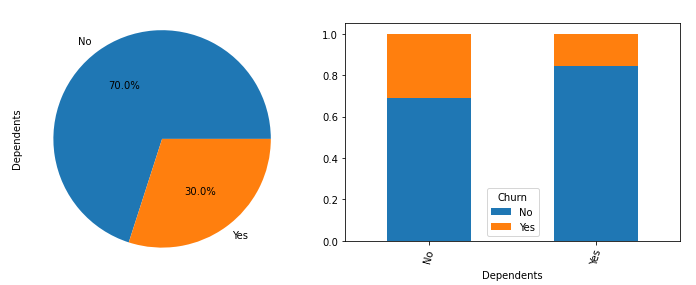
1 indicating Yes for the senior citizen. We have only 16.2% of the customer as senior citizen and the percentage is higher for them to leave the service.

**Partner**



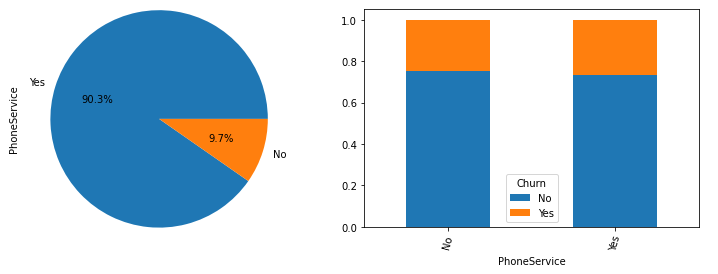
48.3% of the customers have partners and they seem to be loyal with the services. Reason could be the family plan which is helping them to reduce the cost of total usage hence not deciding to leave the service.

**Dependents**



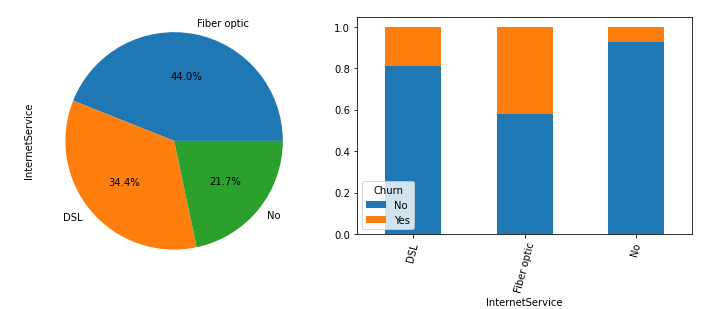
30% of the users have dependents and they end to stay with the services. This again could be the good family the telecom is providing.

**PhoneServices**



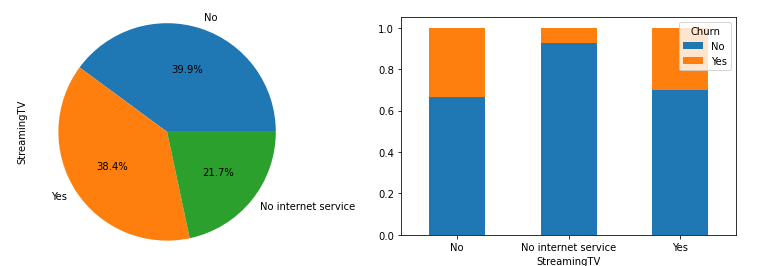
90.30% of the customer uses phone services and it has no impact on the customer’s churn.

**InternetServices**

****

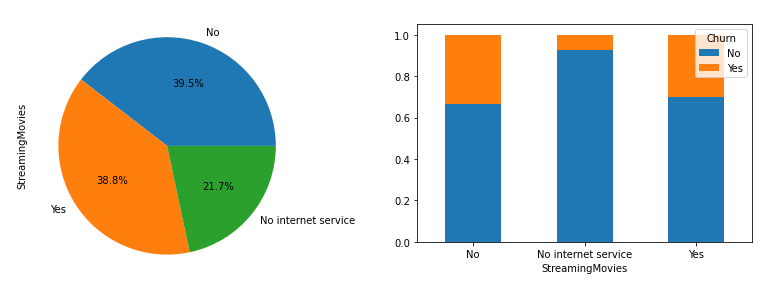
21.7% of the customer do not use the internet services and the churn’s is very less among them. Among internet users, churn’s is high for the customer who uses Fibre optics.

**StreamingTV**



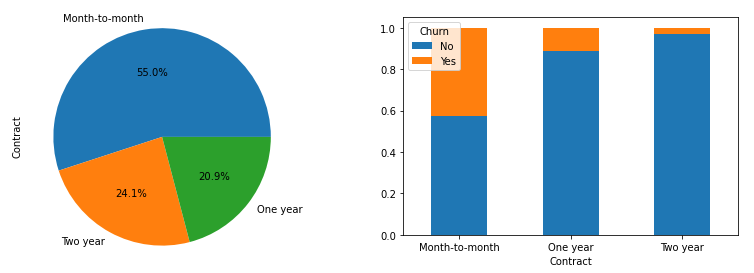
38.4% customer uses the stream TV services and they tend to stay with the telecom services when we compare it with customers who do not use this service.

**StreamingMovies**



Like StreamingTV even for customer using StreamingMovies there is less chance of them leaving the telecom services.

**Contract**

****

55% customer have month to month customer and they have greater chances of switching the telecom services. Customers with longer contract are not too keen to leave the services.

**4. 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?

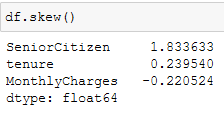
c) Any other feedback through survey we can identify.

4. 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.

1. **Building Machine Learning Models.**

To build the model we need to check for few things

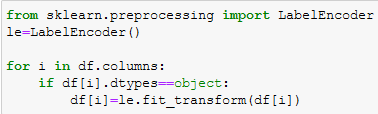
**Skewness**



There is no skewness present in the dataset. Senior citizen is more of the categorical data rather than numerical.

**LabelEncoder**

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



**Imbalance Data**

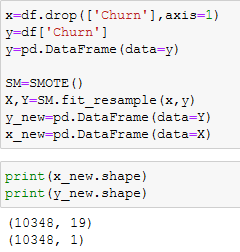
Our target variable ‘Churn’ is not distributed evenly. Having a balanced data set for a model would generate higher accuracy models, higher balanced accuracy and balanced detection rate. Hence, it’s important to have a balanced data set for a classification mode.

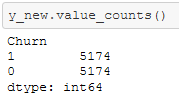
There are two techniques we can use, either we use undersampling or oversampling of data.

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

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.



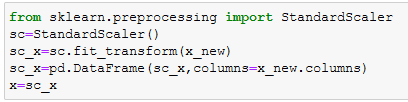


If you notice that now we have 10348 rows instead of 7043 rows and value in our target variable is equally distributed.

**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.



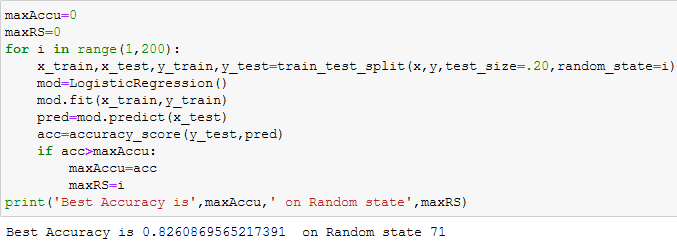
Now our data is ready for machine learning.

**Model Fitting**

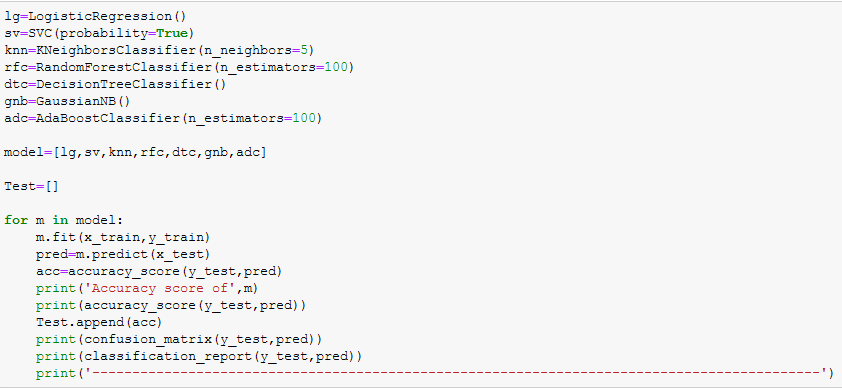
We will train our model using below Algorithtm

1. LogisticRegression
2. SVC
3. KNeighborsClassifier
4. RandomForestClassifier
5. DecisionTreeClassifier
6. GaussianNB
7. AdaBoostClassifier

We will 1st select the best random state to run our model.

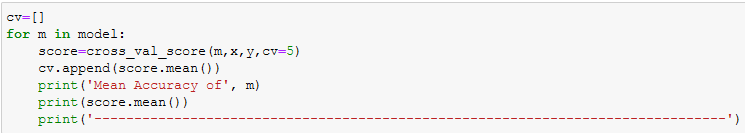


Now lets the train the data with remaining model.



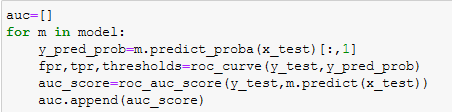
**Cross Validation**

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.

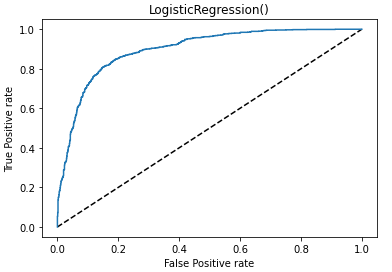
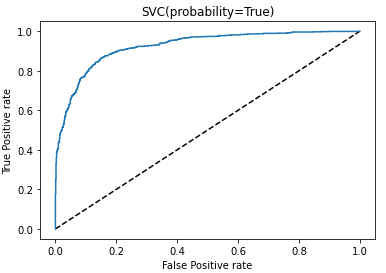


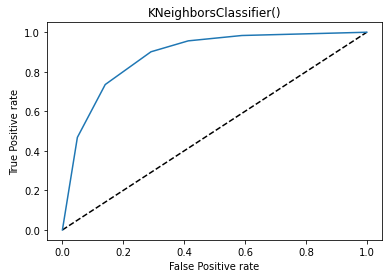
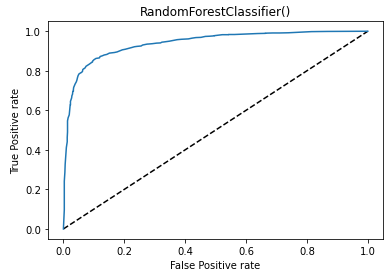
**AUC-ROC**

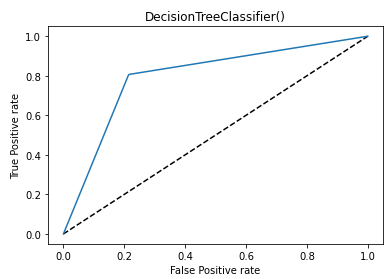
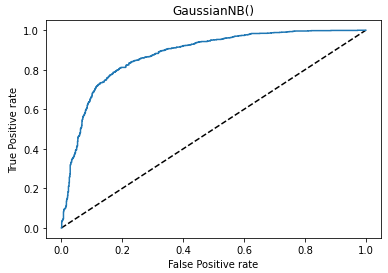
AUC-ROC curve helps us visualize how well our machine learning classifier is performing. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates 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.

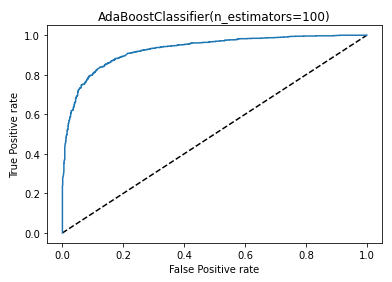


Let’s check the AUC-ROC curve for all the models.

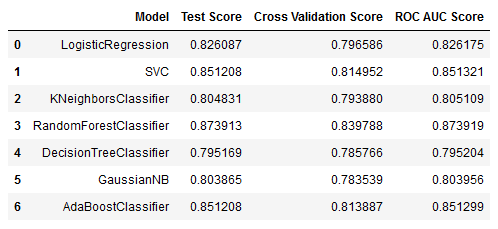
 



Now we will check the performance of our models against all the metrics.



From the above metrics we can conclude that:

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

2. RFC has given us the best accuracy of 87% CV score of 83% and ROC AUC score of 87%

3. KNN has test accuracy of 80% CV score of 79% and ROC AUC score of 80%¶

4. DTC has test accuracy of 79% CV score of 78% and ROC AUC score of 79%

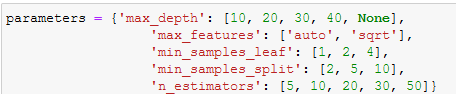
5. None of the model is overfitting.

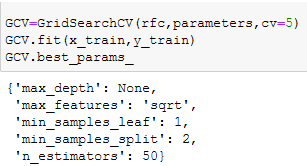
We will hypertune both the model and see if there is an improvement in the model performance.

**HyperParameter Tuning**

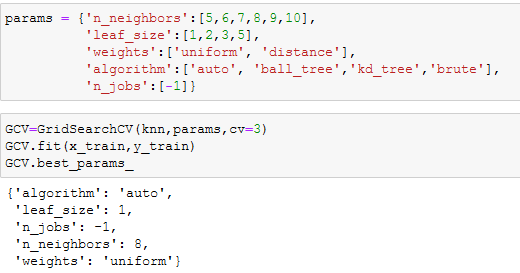
Hyperparameters are crucial as they control the overall behaviour of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

**HyperTuning of RFC**

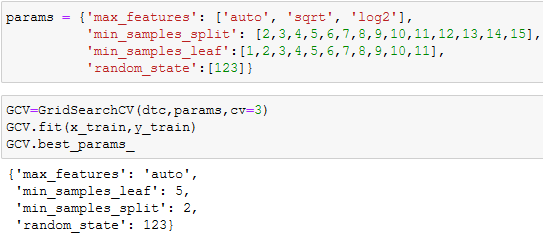




**HyperTuning of KNN**

****

**HyperTuning of DTC**

****

Below is the performance of the models after Hypertuning.

**RFC**

Accuracy Score - 86.42%

Cross Validation Score - 83.57%

AUC ROC Score - 86.42%

**KNN**

Accuracy Score - 80.77

Cross Validation Score - 79.88

AUC ROC Score - 80.79

**DTC**

Accuracy Score - 82.07

Cross Validation Score- 78.95

AUC ROC Score- 82.07

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 undefit 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% with CV score of 83% and also the AUC ROC score of 86.42%

Further Evaluation of our final model

**Confusion Matrix**

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 – 903 customer were correctly classified , called as True Positive and 135 were wrongly classified as leaving the company (False negative)

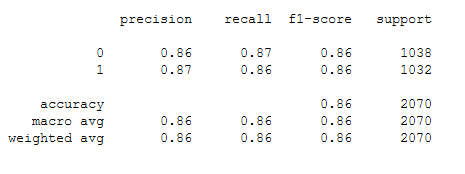
The second row is about the leaving the company – 147 customers were wrongly classified as leaving the company (false negative) and 885 were correctly classified as leaving the company.

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

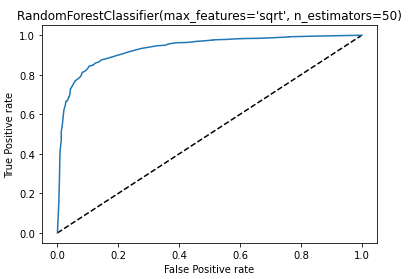


Our model predicts 86% of the time customer retention correctly (precision). The recall tells us that it predicted the retention of 86% of the customer who actually did not churn.

**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.

We now see the AUC ROC curve of our final model.



86% is a good accuracy and we can say that we will be able to predict 86% correctly if the customer will continue to use the telecom services or going to leave it.

**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 preprocessing part, we converted features into numeric ones. Afterwards we started training 7 different machine learning models, picked three of them (RandomForest, KNearestNeighbors and DecisionTreeClassifier) and tuned it’s performance through optimizing it’s hyperparameter values and finally selected the RandomForestClassifier.

1. **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.