**Telecom Customer Churn**



#### **1.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 prioritize focused marketing efforts on that subset of their customer base.Although there are many reasons for customer churn, some of the major reasons are service dissatisfaction, costly subscription, and better alternatives. The telecom service providers strive very hard to sustain in this competition. So to sustain this competition they often try to retain their customers than acquiring new ones as it proved to be much costlier. Hence predicting churn in the telecom industry is very important. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

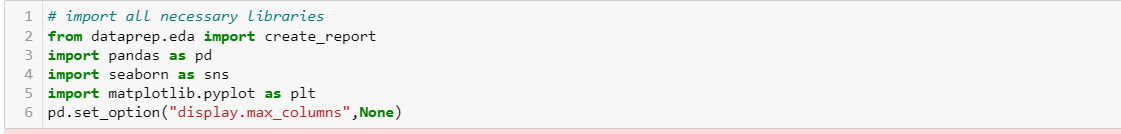
Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

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

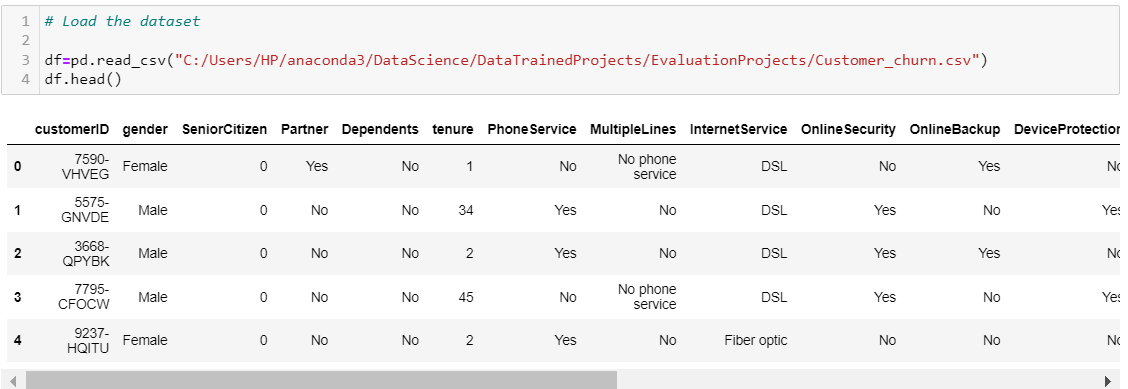
1. **Data Analysis:**

We will get detailed insight of the data set by Analyzing the data carefully. At first we have done a detailed analysis by using the following code.

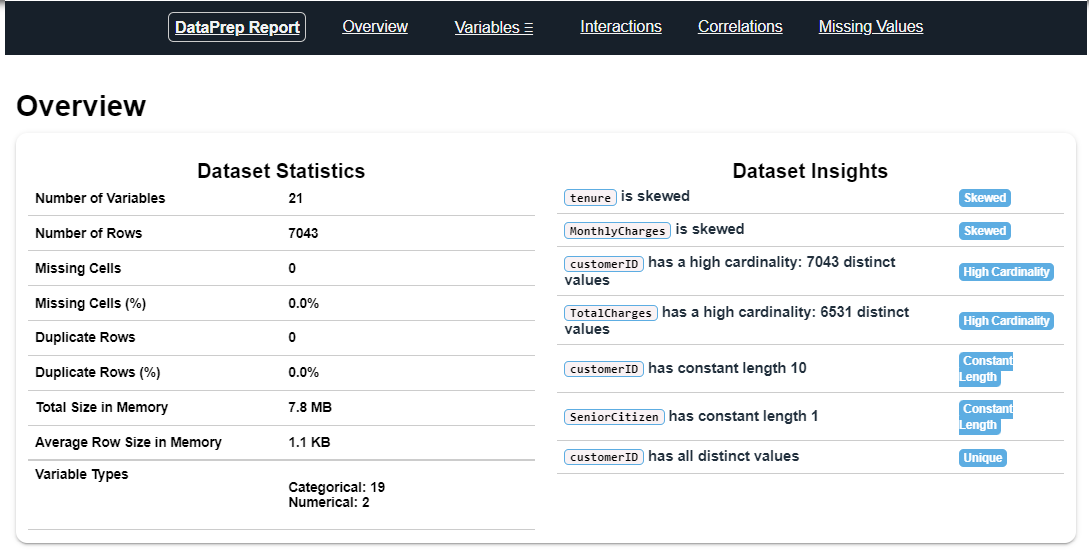
**2.1 Import** all the required libraries.



Lets look at top five records within the dataset



Lets look at the Count , missing values, datatypes, skewness of the features in the dataset

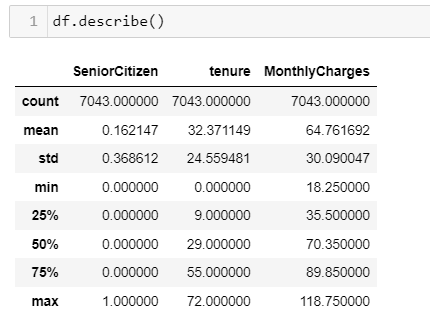


By looking at the above figure, its clear that we have 21 features with 7043 records for individual features

**2.2 Observations:**

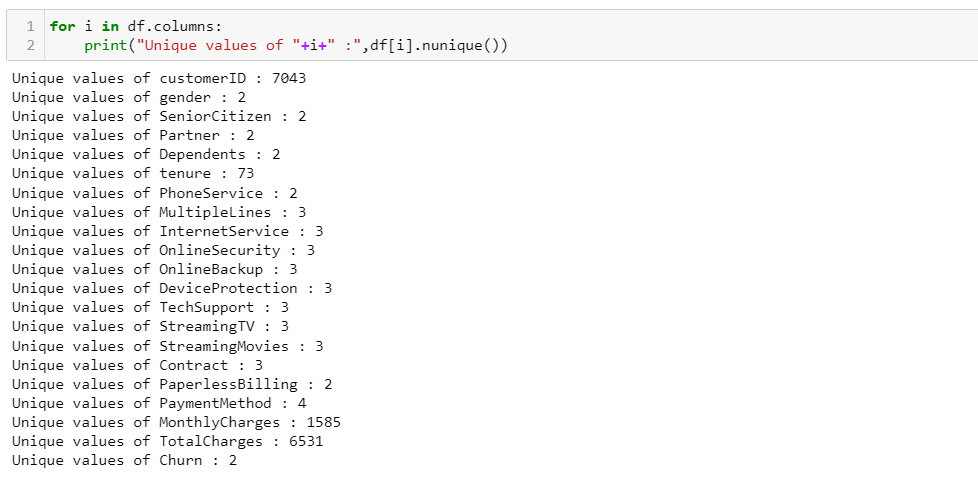
1. Number of features - 21
2. Shape of the dataset - (7043\* 21)
3. No missing values in the dataset
4. No Duplicate values in the dataset
5. Features are of type Categorical (count-19) & Numerical (count-2)
6. Tenure and Monthly charges are having Skewness.

**2.3** We will look at the **statistical parameters.**

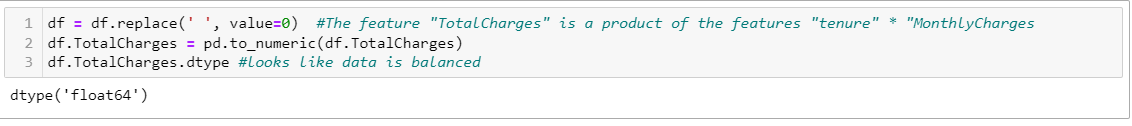


Since we are having difference between mean & median values there is some sort of skewness in both of these variables.

We will look at the unique values of each features



In the dataset, there should be record for "TotalCharges" = 0, but instead there is just a space " " So we will replace it with value '0'



**2.4 Key Insights**

1. We dont have any Null values in the dataset

2. we have replaced " " with value 0

3. We have converted tenure which is of the type Object to float type

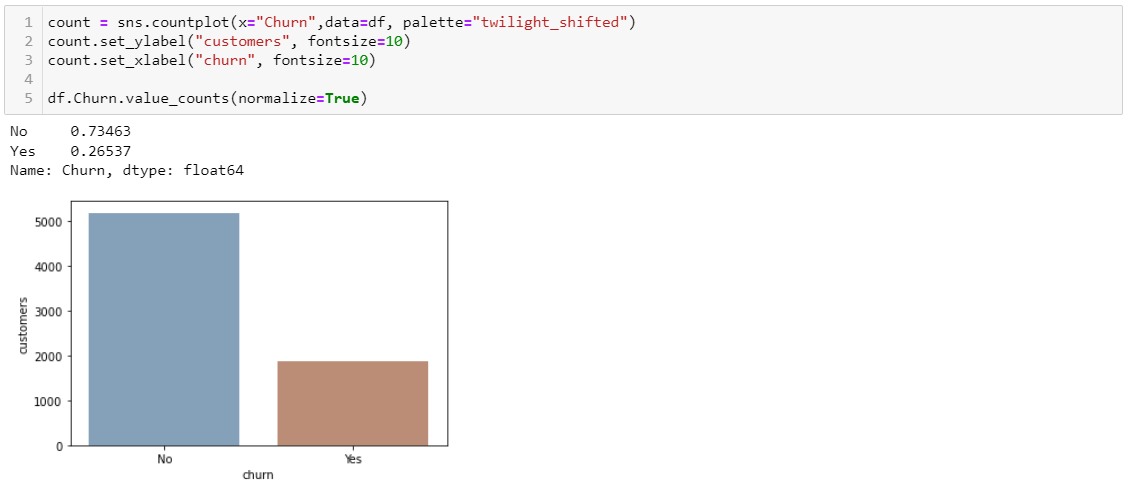
4. There are no duplicated values in any of the feature

5. Dataset is of the shape (7043, 21)

6. Target variable is of the type categorical so this is a classification problem

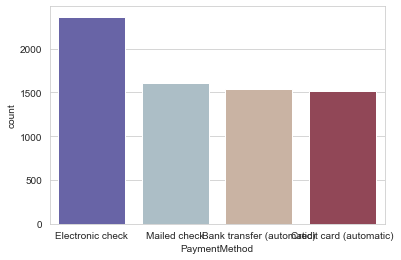
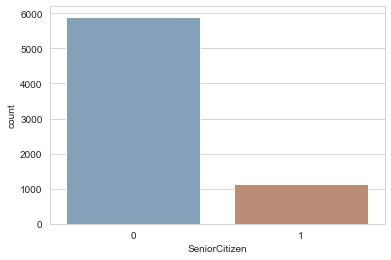
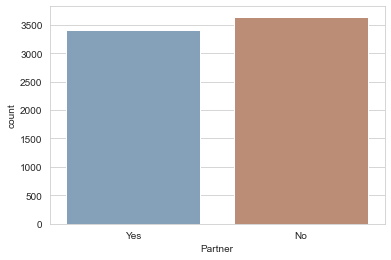
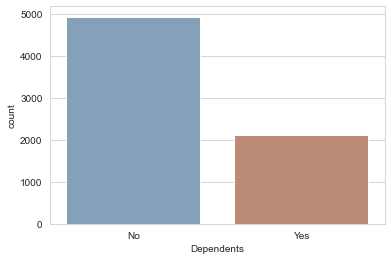
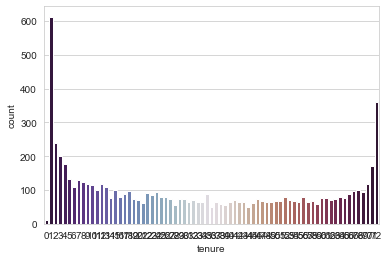
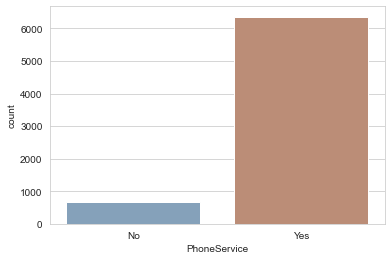
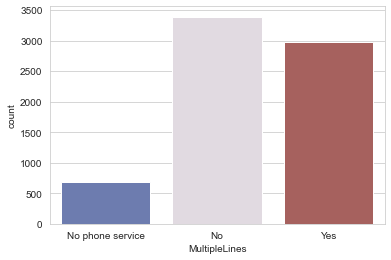
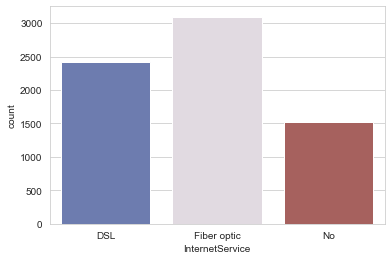
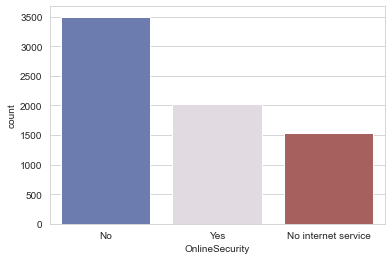
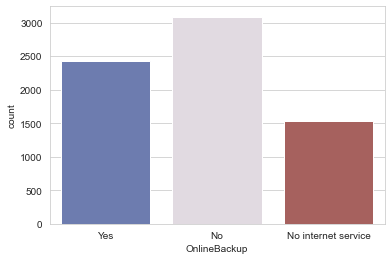
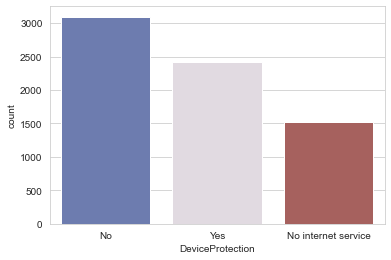
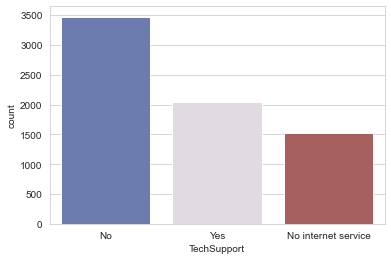
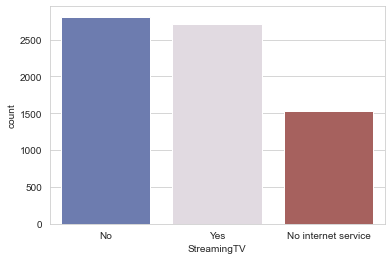
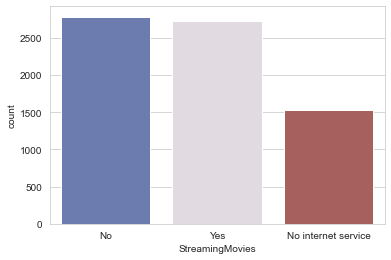
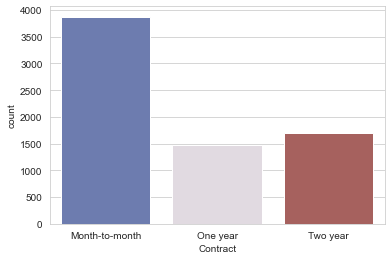
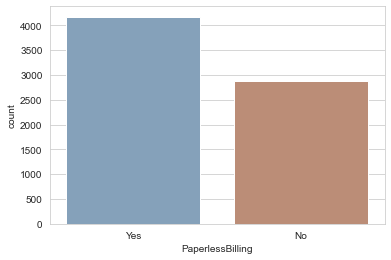
1. **Data Visualization:**

We will visualize the churn & non churn count using countplot so that we will get clear picture to understand the patterns better.



There are more number of non-churn cases than churned one's. This is a imbalanced data. so later we will upsample the churned to non-churn count to get a better model.

Since we got clear picture of the ratio between the different features by plotting countplot, we will look in to detail.



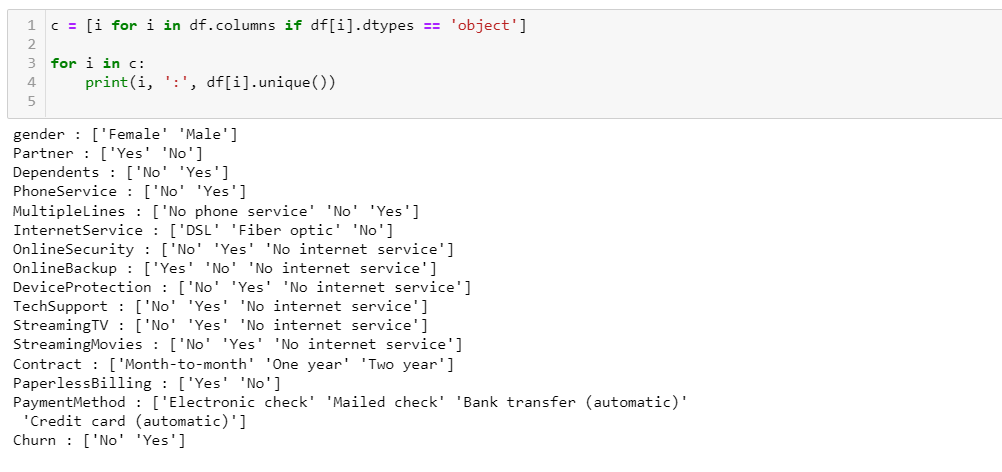
##### **3.1 Inference:**

1. About half of the customers are males and other half of them are females.
2. We have less number of Senior customers than the younger one's
3. There are less Dependants for most of the customers
4. Majority of the customers are there since 1 month. ie, there are more customers who have recently joined our network & comparitively more number of customers who are staying for long tenure
5. Majority of the customers are not availing multiple lines and less are availing no phone service
6. Also, Many of them are availing Multiple Lines
7. Majority of them are using Fibre optic internet services
8. majority of the customers are having No Online security
9. Majority are not having Online Backup & some are having
10. Majority of them are not using customer service suggesting there are no issue in their network
11. There are equal number of people are getting service StreamingTV
12. There are equal number of people are getting service Streamingmovies
13. Majority of them are using month to month contact
14. Majority are paying bills online through Electronic check. and about half of the people are paying by mailing check,bank transfer, creditcard
15. On an average less than half of the customers are about to churn

**4. Pre-processing Data:**

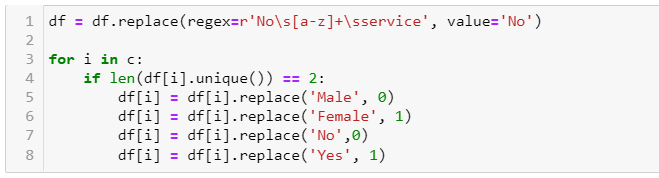
**4.1 Checking of categorical data**

Let's check what kind of catagorical data do we have:

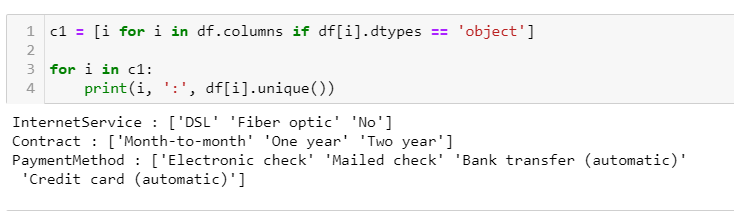


Lets replace No internet service to 'No'.

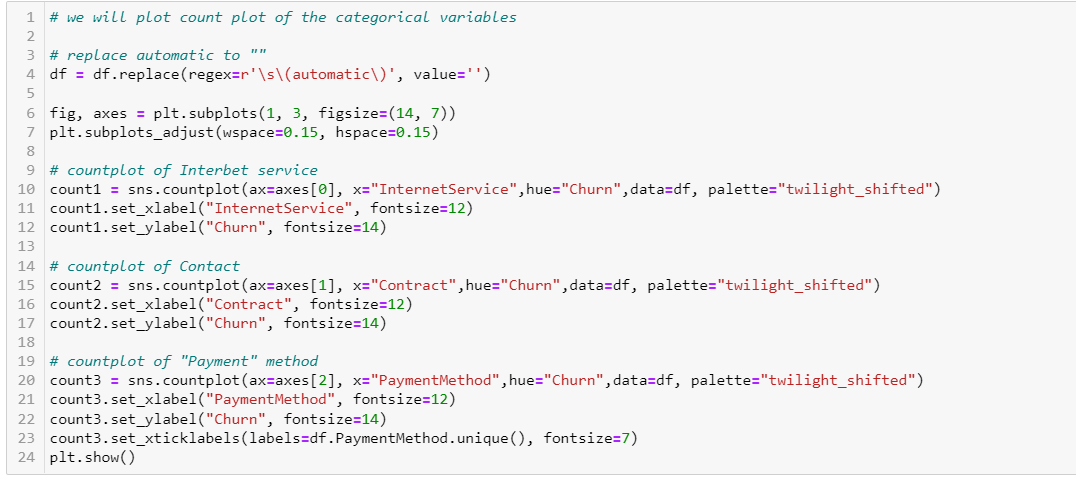
we will replace Categorical variable to Binary

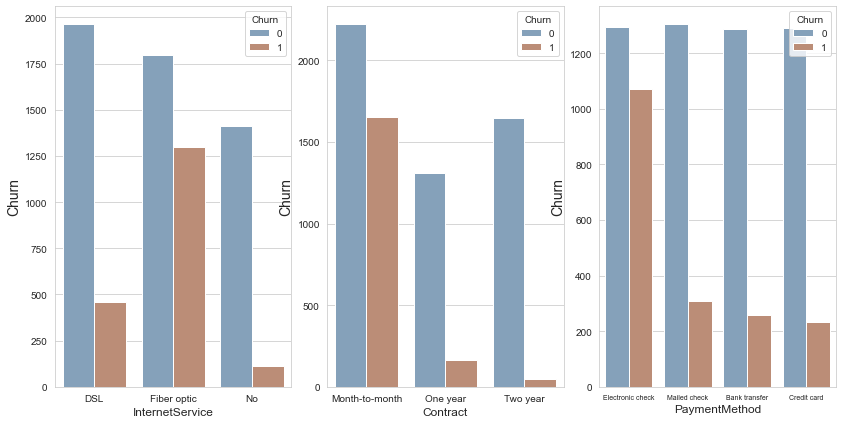


Now lets look at other Categorical features



We have found the unique values of categorical variables. Now lets plot the count plot for these categorical features based on churn. Below is the code

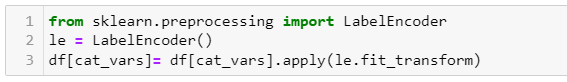




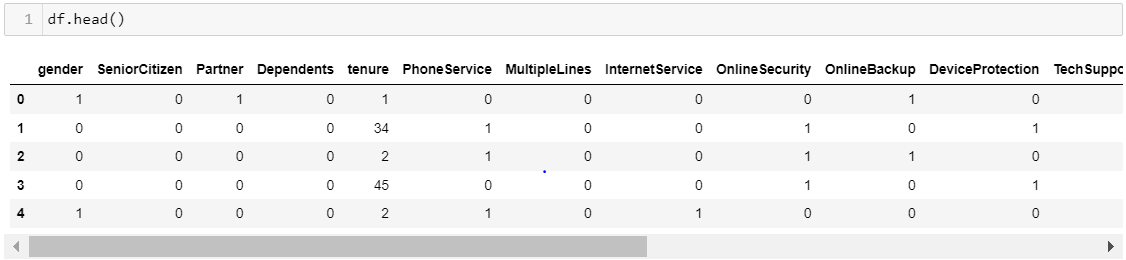
We can clearly see that the features "InternetService" and "Contract" can be considerd ordinals and they have good correlation with the churn labels. So I will change them accordingly.

**4.2 Label Encoding:**

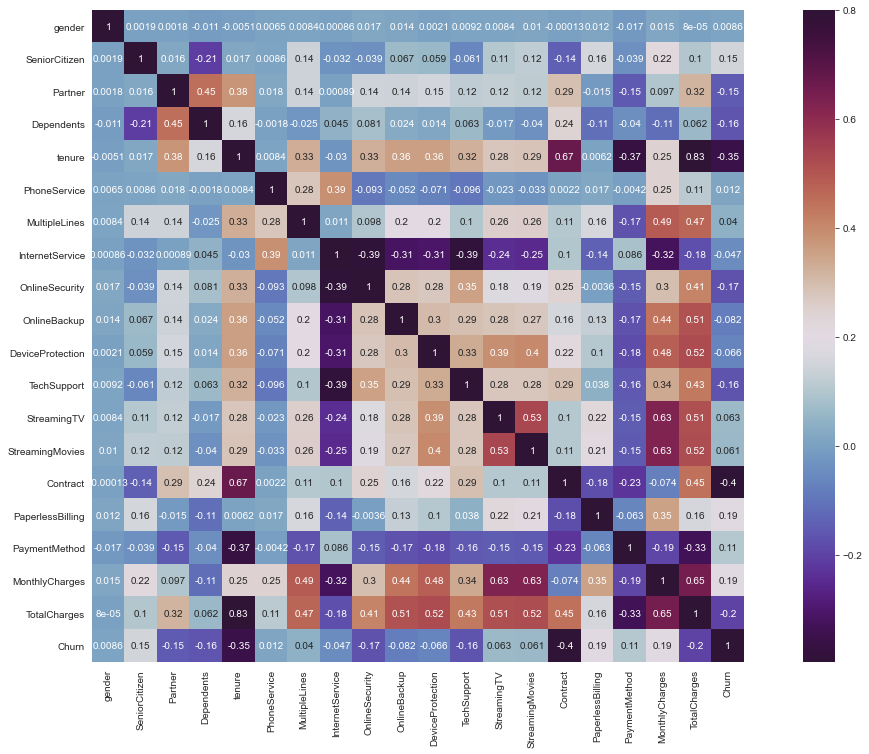
We will encode the categorical labels to Numerical values



Lets see how the features look like after doing Label Encoding



we will plot the heatmap



**4.3 Observations from the Heatmap:**

1. Also, The service provided like Contract,StreamingTV,TechSupport,DeviceProtection,OnlineBackup,OnlineSecurity,Multiple lines is a having very strong corelation with Total charges. So it suggests all these services comes with very High Total charge and to avail all these service you have to pay more.

2. There is a strong correlation between 'tenure' (or 'MonthlyCharges') and 'TotalCharges' because TotalCharges = tenure\* MonthlyCharges.

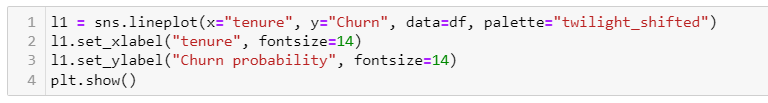
3. In addition, we can see that there is a strong correlation between 'OnlineSecurity' and 'MonthlyCharges'. This is probably because the "fiber optic" is the most expensive service and this is the reason for high monthly charges. It's also can explain why the churn probability is higher for the customers who own this service - the price is too high for them.

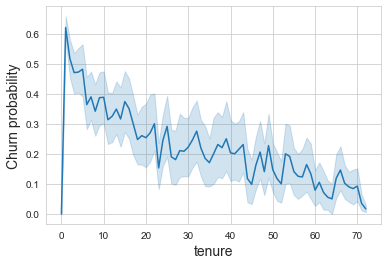
4. Another strong negative correlation is between "tenure" and "Contract".It make sense that the higher the tenure - the contract will be yearly and not monthly.

5. So if we will look at the features affecting or having stronget correlation with 'Churn'.

**4.4 Pattern based on tenure:**

Next we will see the trends of churn based on tenure of the customer associated with our network





As the figure explains churn probability will reduce as the tenure of the customers gradually increases. So our aim is to retain the customers within the same network by providing value added services.

**5.Data modelling:**

Let’s start explaining our ****choice** to build model based on the different demographics.We will first split the dataset into train & test dataset to first train the model and then we will test the testset using the learnings of the trained dataset.**

**Before splitting the dataset we will resample the imbalanced dataset. Remember, we had less churn rate than non-churn date. So doing resampling is best way to make it balanced so that we can get more accuracy,f1 scores & precision.**

**You can use any of the following to resample the dataset**

**5.1** To deal with the imbalanced data, we can use any of thewor following two techniques:

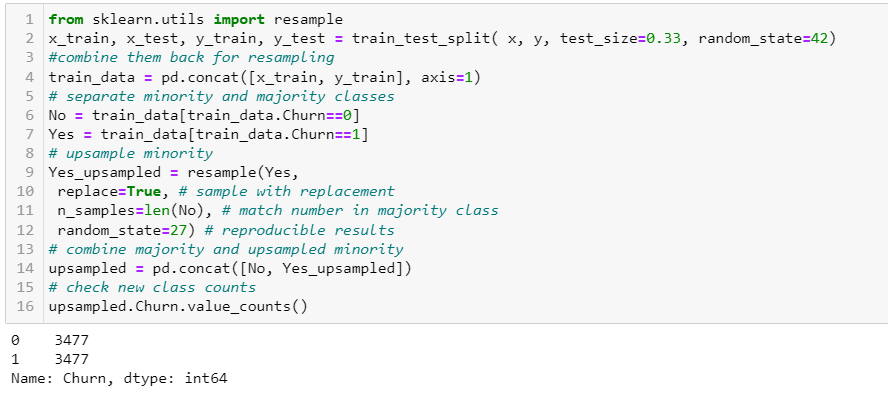
****1.**** ****Over-sampling - SMOTE:**** Synthetic Minority Over-sampling Technique SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b. (Page 47, Imbalanced Learning: Foundations, Algorithms, and Applications, 2013.)

****2.**** ****Under-sampling - NearMiss3:**** When instances of two different classes are very close to each other, it removes the instances of the majority class to increase the spaces between the two classes. It works in 2 steps: Firstly, for each minority class instance, their M nearest-neighbors will be stored. Then finally, the majority class instances are selected for which the average distance to the N nearest-neighbors is the largest.

****3. Resampling****

**5.2 Resampling:**

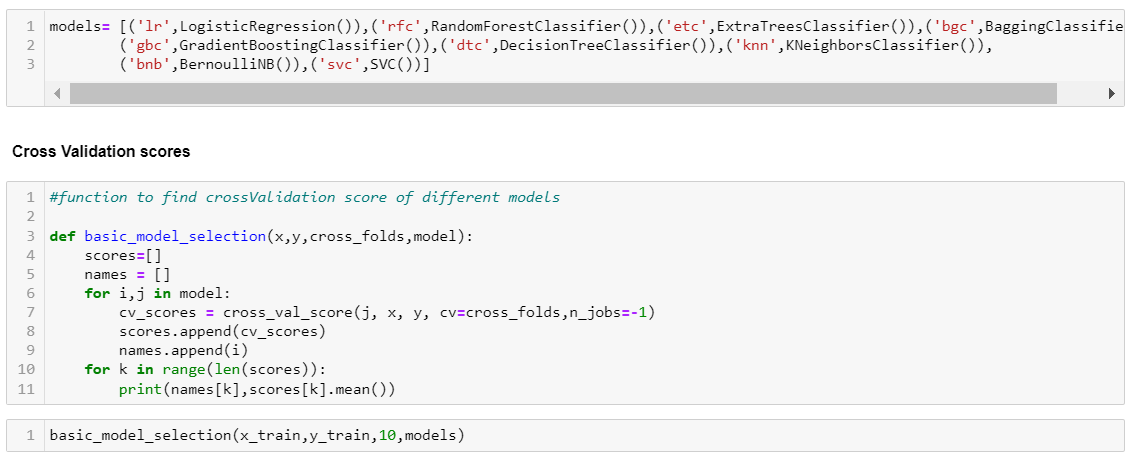
**Below is the code for Resampling:**

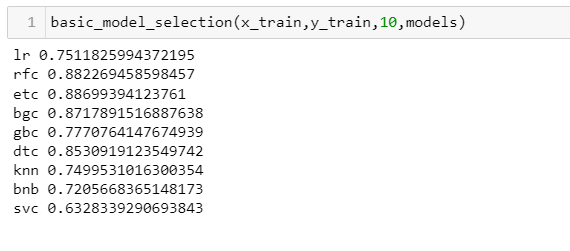


As we know that we have got the balanced churn & non-churn rates at the end of resampling.

**5.3 Splitting the data:**

Now, we will split the data into train & test sets. Most of the fellas will take models one by one. But I m using all the models passed into a generic function which gives the cross validation scores for each of the models.





We are getting comparitively high cross validation scores for RandomForestClassifier & ExtraTreesClassifier. So we will consider both for Hyper parameter tuning.

**5.4 Hyperparameter Tuning:**

At first we will define the parameters specific to each model that we have selected for tuning. In our case it is RandomForestClassifier and ExtraTreesClassifier.



After using GridSearchCV to fit a particular model we will get the best parameter as output. So we have got best score and parameter keeping that in mind we will decide whether which model is giving high score when used with the best possible parameters.

**5.5 Model evaluation:**

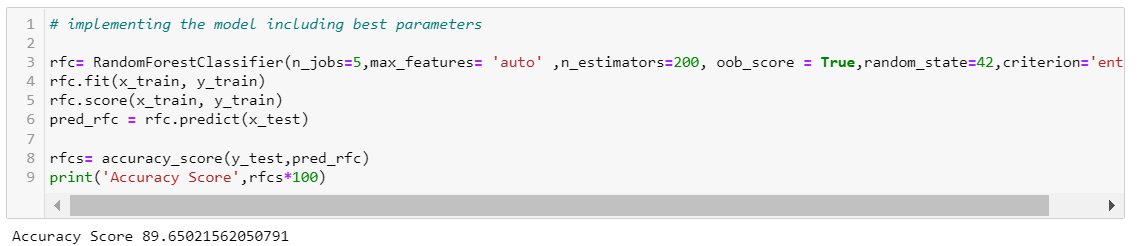
We first check the model performance with Random Forest Classifier by predicting the cross validation score on the test dataset. This is giving ~87.8 validation score



The second chosen model Extra Trees Classifier is giving ~ 88.4 validation score which is quite high than RFC

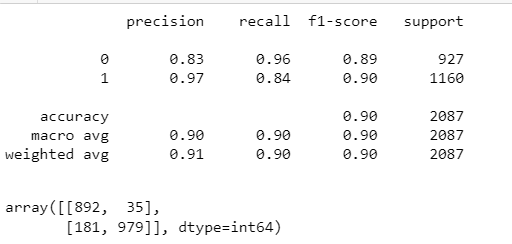


Since we cant just choose the model for final evaluation by simply considering the cross validation score. We will implement the model using best parameters



After implementing the RFC model we got Accuracy score of 89.6 which is quite high than predicted ETC score.

Also we will look at f1 score, precision & confusion matrix:



****5.5 Performance Metric****

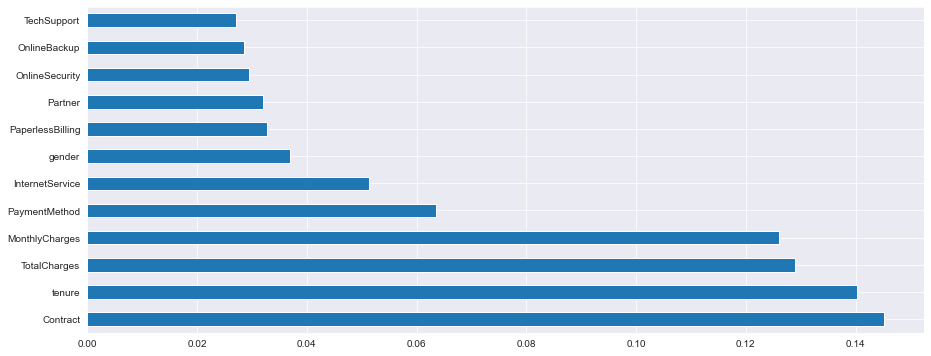
Metrics:

1. Auc score
2. Accuracy score
3. F1 score
4. Precision
5. Recall
6. Binary confusion matrix

So we got high accuracy score as well as quite good score for f1score, precision & recall. no model can perform better than the Random Forest Classifier in our case. So we will choose all the 3 models to work on auc\_score.

**5.6 Importance:**

Before proceeding to selecting our final model for production we will consider ETC & draw importance so that we can get info on key drivers for churn.

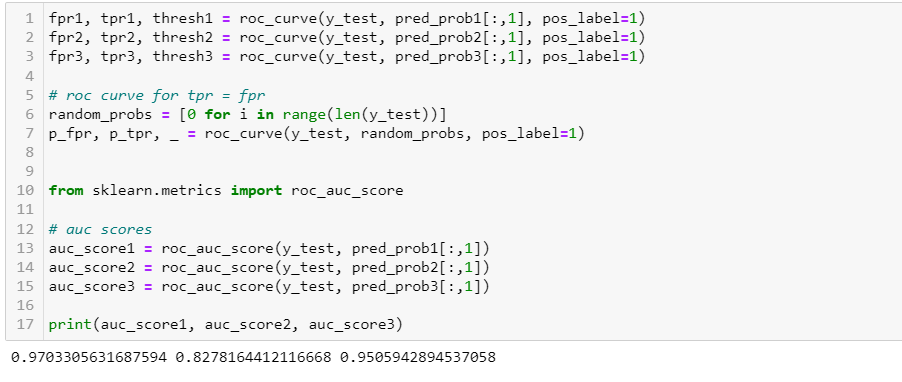


By the above graph, we are sure that tenure,contract,Monthly charges,Total charges has high correlation with the Target variable. so any variance in this will impact on the prediction of customer churn[¶](http://localhost:8890/notebooks/anaconda3/DataScience/DataTrainedProjects/EvaluationProjects/CustomerChurn_predict.ipynb" \l "By-the-above-graph,-we-are-sure-that-tenure,contract,Monthly-charges,Total-charges-has-high-corelation-with-the-Target-variable.-so-any-variance-in-this-will-impact-on-the-prediction-of-customer-churn)

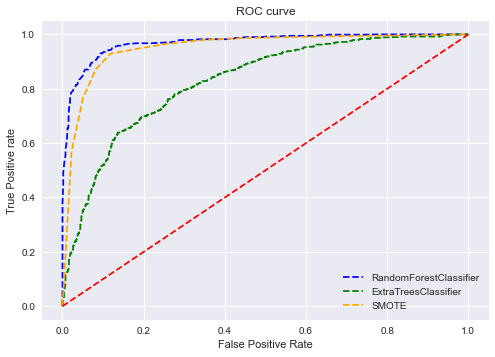
So all 4 features are considered as the key drivers for churn.

**5.7 ROC AUC Curve:**

Next step is to predict the probability of test dataset and get the auc\_roc\_score for each of the model.



We got comparitively high auc\_score for Random Forest Classifier, we will further plot the ROC curve to see which graph gives better performance. As the area under the curve is more for Random Forest Classifier, we will consider the same as our final model for production.



**6.Conclusion:**

Although the ExtraTreeClassifier model provided the best interpreters on the key drivers for churn by knowing the importances, the accuracy score & recall value is low making more mistakes in predicting the churned customers as non churned customers. So we are using random forests as our final model for this problem. These ensemble models can provide high accuracy values in other classification problems.

* 1. **Key Drivers for churn:**

Contract, Tenure, Monthly charges, Total charges are the predictors for the most of the cases whether a customer is going to churn or not.

Higher the Tenure and Contract lesser is the churn rate.

Lesser the Monthly charges & Total charges lesser is the churn rate.

1. **References:**

You can find the source code for this case study [here](https://github.com/Sinchana2/TelecomCustomerChurn)

If you like my work please don’t forget to up vote!