Project Report

Project Name: Telecom Customer Churn Analysis & Prediction

Submitted By: Prashant Pathak

Time Duration: 5 June 2023 to 11 June 2023

Introduction

Hello! My dear friends, my name is Prashant Pathak, and I am a student of data science domain.

I like to play with data, And I enjoy getting such important information from the data, apart from this model building is my favorite one.

As we all know that the data science project mainly has to go through these steps:

- Frame the problem
- Data collection
- Data preprocessing
- Exploratory Data Analysis
- Data Visualizations
- Separate data for training & testing
- Model building using different type of algorithms
- Best model selection etc...

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 priorities 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.

So, I have a problem but I don't have any action plan. First, Let's make a perfect plan for my problem.

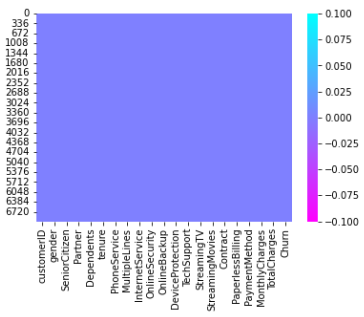
Basic Action Plan: In my first step I want data and after this I will preprocess data according to my problem. After this I will do EDA on data and data visualization for find hidden information and pattern of data. After this I will do train-test split and model building.

First Step: (Data Collection) - As we all know data is basic need for any ML related project, So I searched for perfect data according to my problem and luckily, I got a dataset in csv format from GitHub which is good one for churn analysis. I also put-up dataset in my GitHub repository.

Second Step: (Data Preprocessing) - Basically, I worked on Jupiter Notebook, So I created a new project which name is "Costumer churn analysis" I also mentioned this file in my GitHub repository in IPYNB format. First, I import all basic library and module for my project like: NumPy, pandas, seaborn, matplotlib, SciPy, os etc... After this I made a pandas data frame "df"' using pd.read_csv method. 7043 rows and 21 columns are present in our dataset Some columns are categorical and some are numerical and most important thing is, target column has only 2 categories so it will be termed as "classification problem" where we need to predict the several customer churn using the classification models.

Third Step: (EDA) - In this step, I checked these things; shape, column's name, null values, data type, value count, unique etc.... After this I handle extra space, datatype and Null value in "Total Charges" column. After this I checked heatmap for finding null values but data have not any kind of missing value. I also checked info of data which result are below-

```
[15]:
# let's check the null values again
sns.heatmap(df.isnull(),cmap="cool_r")
plt.show()
```



- There are no missing values present in any column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.1+ MB
```

After this I used describe () method and I got these result -

```
In [22]:
# statistical summary of numerical columns
df.describe()
```

Out[22]:

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2283.300441
std	0.368612	24.559481	30.090047	2265.000258
min	0.000000	0.000000	18.250000	18.800000
25%	0.000000	9.000000	35.500000	402.225000
50%	0.000000	29.000000	70.350000	1400.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

This gives the statistical information of the numerical columns. The summary of this dataset likes perfect since there are no negative/invalid values present

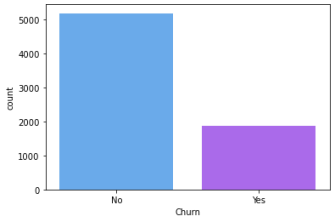
From the above description we can observe the following things

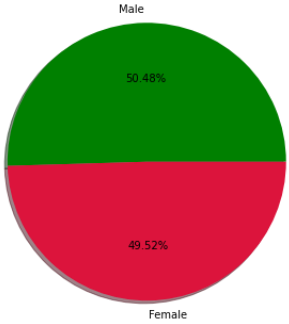
- The counts of all the 3 columns are same which means there are no missing values in the dataset.
- The mean value is greater than the median (50%) in tenure and TotalCharges columns which means the data is skewed to right in these columns.
- The data in the column MonthlyCharges have mean value less than median that means the data is skewed to left.
- By summarizing the data, we can observe there is huge difference between 75% and max hence there are outliers present in the data which we will remove them later using appropriate methods.
- We can also notice the Standard deviation min 25% percentile values from this describe method

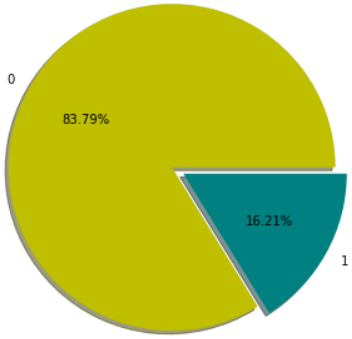
Fourth Step: (Data Visualizations) - There are three types of visualization, on which I worked.

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Univariate Analysis: Plots of Categorical columns-



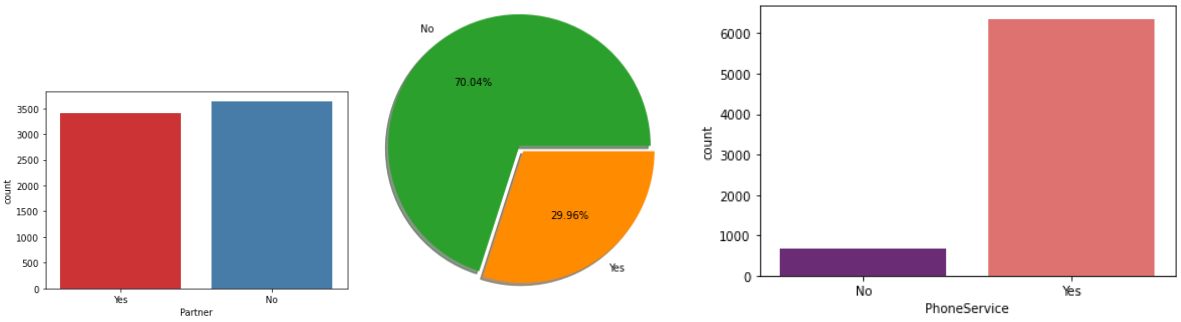




(i) From the count plot we can observe that the count of "No Churn" is high compared to the count of "Yes Churn". That is there are a greater number of customers who have not churned. This leads to class imbalance issue in the data, we will rectify it by using Oversampling method in later part.

(ii) From the plot we can observe the total number of male and female customers are almost same, but still the count of male is 3555 which is high compared to count of female which has 3488 counts.

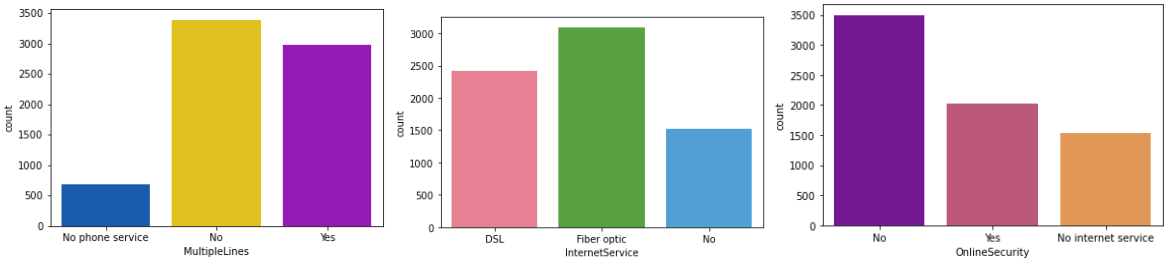
(iii) Here 0 represents the non-senior citizens and 1 represents the senior citizens. The count of 0 is high in data compared to 1 which means the number non senior citizen are quite high compared to senior citizens data in the given dataset Around 83% of the customers are non-senior citizens and only 16% are senior citizens.



(i) The count for partner details is almost similar and the customers who do not have partners are bit higher than who have partners.

(ii) The customers who have dependents are very less in counts that means they do not have anyone dependent on them. Here around 70% of customers have dependents and only 29.96% have no dependents.

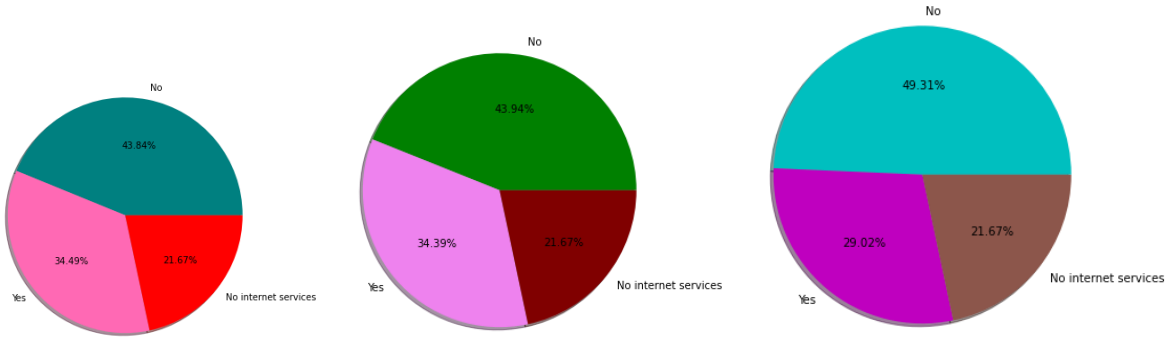
(iii) The customers who have phone services are large in numbers and who do not own phone services are very less in number.



(i) The customer who has phone services from single line have high counts compared to the customers having phone services from multiple lines, also the customers who do not have phone services have covered very less data compared to others.

(ii) Most of the customers have chosen to get Fiber optic internet followed by DSL, but there are many customers who do not get an internet service.

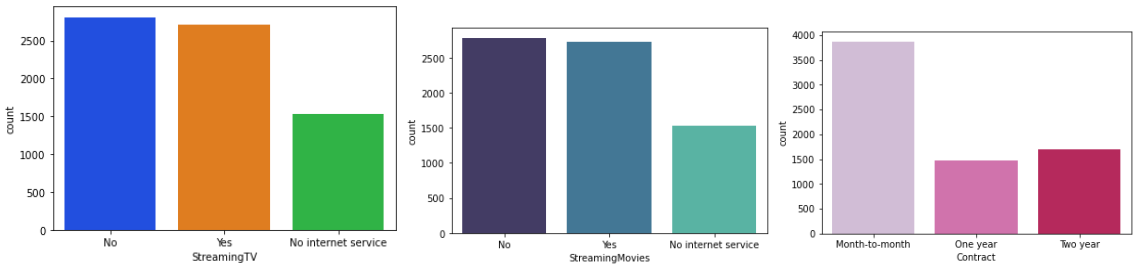
(iii) It is obvious that the customers who have internet services they need online security and who do not own any internet services, they do not need any online security. But from the count plot we can observe many customers who have internet services but they do not use any online security.



(i) It is obvious that the customers who do not own internet services and online security, they do not need online backup usage. From the plot we can see many customers who own internet services they do not have Online backup and the customers who own internet services have very less online backup. Also, the customers who do not have internet services have very less online backup counts compared to others.

(ii) From the count plot we can notice that the customers without any device protection have high counts as compared to the customers who have some kind of device protection. and the customers who do not have internet access they do not need any device protection.

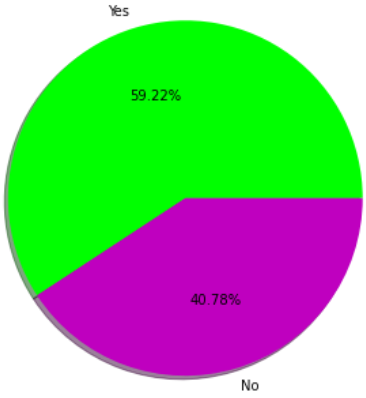
(iii) The customers who do not need any technical support are high in counts compared to the customers who need technical support. Around 49% of the people do not need any technical support and only 29% needs.

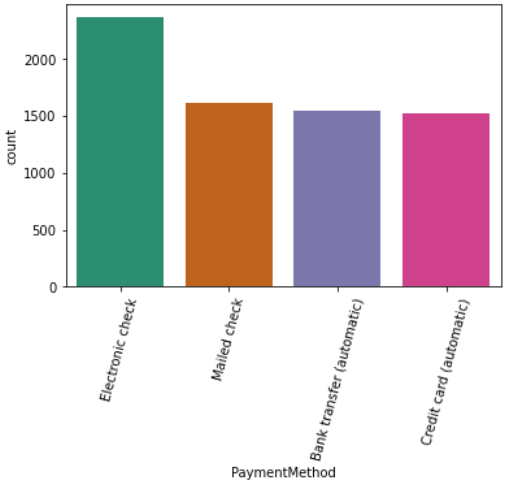


(i) The customers who do not use streaming TV have little bit higher in numbers than the customers who do use Streaming TV and the customers who do not own internet they do not have this service much.

(ii) The customer who doesn't have Streaming movies are high in count followed by the customers who have Streaming Movies services and the customers who do not have internet services they have fewer streaming movies services compared to others.

(iii) Most of the customers prefer Month to Month contract compared to 1 year and 2-year contract.

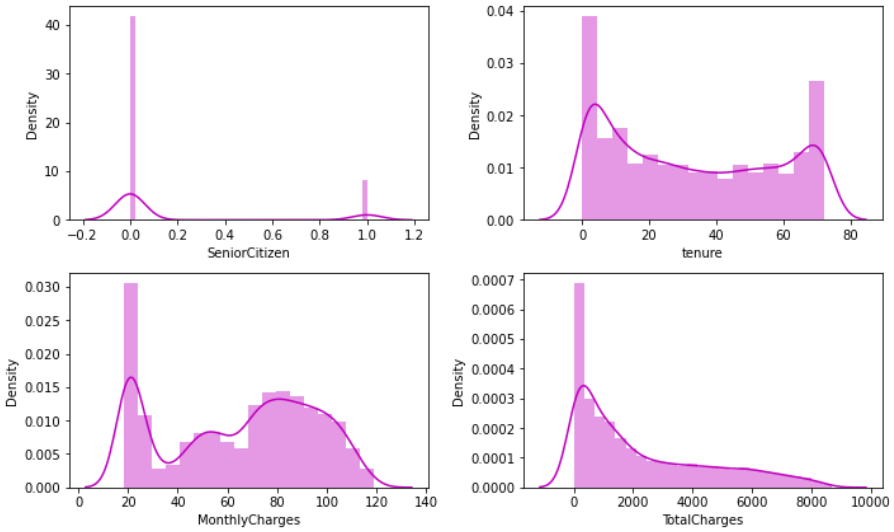




(i) Most of the customers prefer paperless billing and average number of customers who do not prefer paper less billing they may like to receive paper billing.

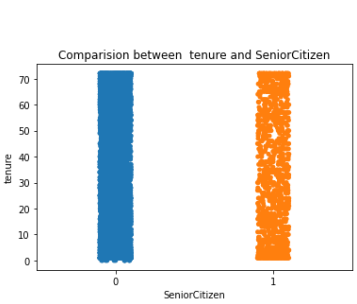
(ii) Most of the customers prefer electronic check payment method and the customers who prefer Mailed check, bank transfer and Credit card have average in count.

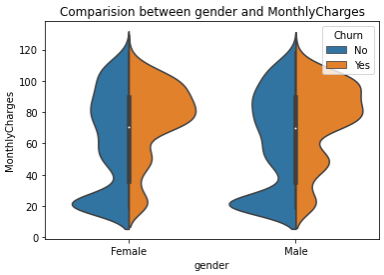
Plots of Numerical Columns:

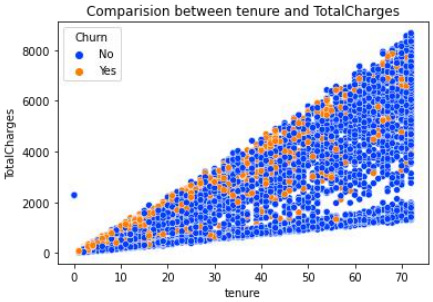


(i) From the above distribution plots, we can notice that the data almost looks normal in all the columns except SeniorCitizen and the data in the column TotalCharges is skewed to right Other two columns tenure and MonthlyCharges do not have skewness.

Bivariate Analysis:



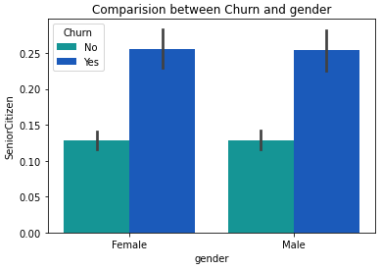


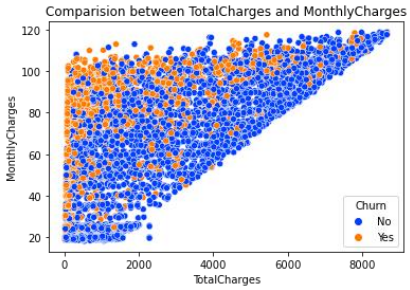


(i) There is no significant difference between the features, here both the feature is in equal length.

(ii) Both male and female customers with monthly charges above 60 have high chances of getting churned.

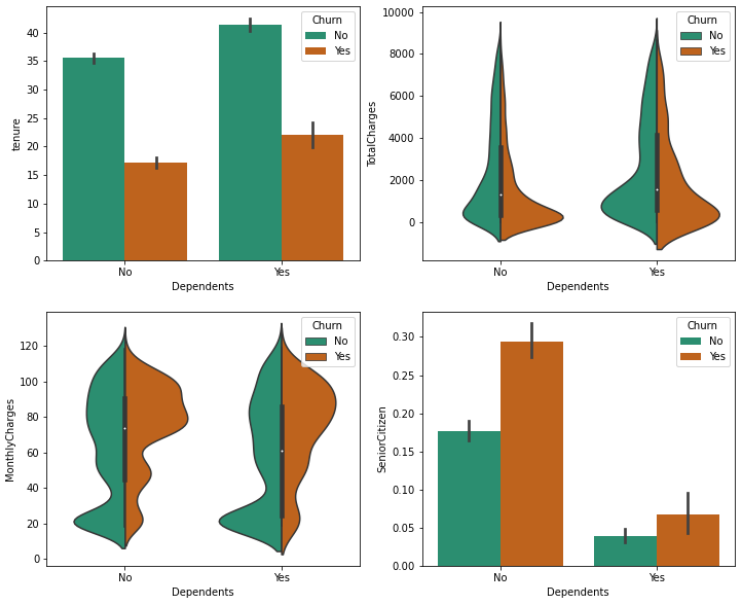
(iii) Here we can notice the strong linear relation between the features. As the tenure increases, TotalCharges also increase rapidly if the customers have low tenure services, then there is high chance of churn.



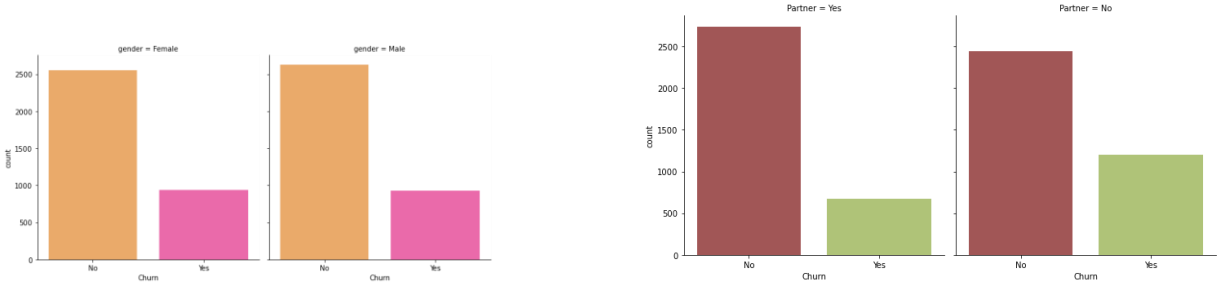


(i) There is no significant difference between the columns. The customer's churns remain unaffected in gender and SeniorCitizen case

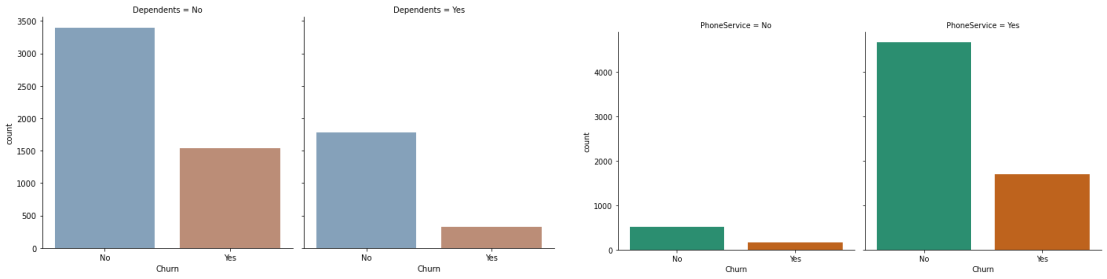
(ii) There is a linear relation between the features. The customers with high monthly charges have high tendency to stop the services since they have high total charges. Also, the if the customers ready to contribute with the monthly charges, then there is an increment in the total charges.



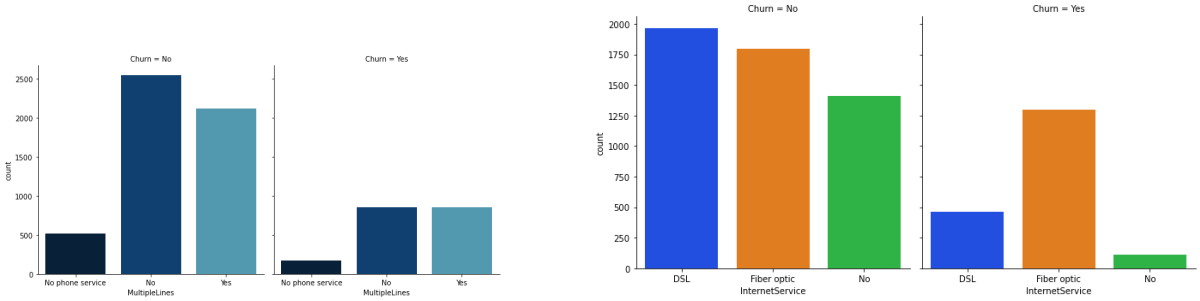
- The customers who have dependents with high tenure, then the churned level is high 80-110.
- The customers who have total charges in the range of 0-2000 with dependents then the chance of getting churned is high
- The customers having Monthly charges between 80-110 with dependents have high churn rate and when the customers have no dependents and having monthly charges around 20 then the ratio of churn is very high.
- If the customer is a senior citizen and has no dependents, then there is a tendency of getting churned.



- In the first plot we can see there is no significance difference in the genders, both the genders have equal churn level
- In the second plot we can see the customers without partners have high churn rate compared to the customers with partners.

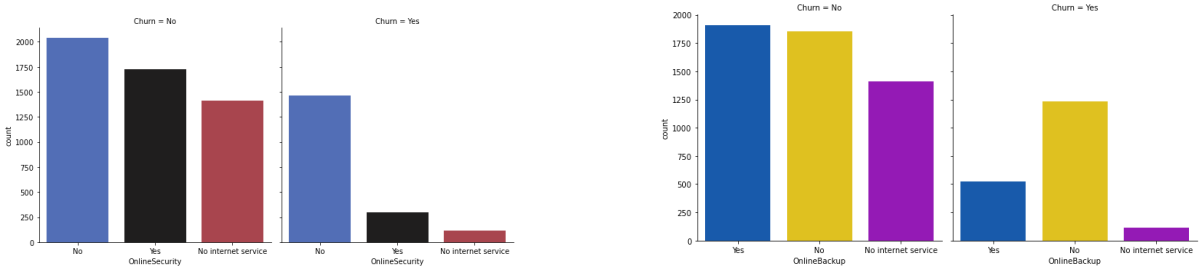


- The customers who do not have any dependency have high churn rate compared to the customers who have dependents.
- In the last plot we can notice the customers who have phone service have high tendency of getting churned



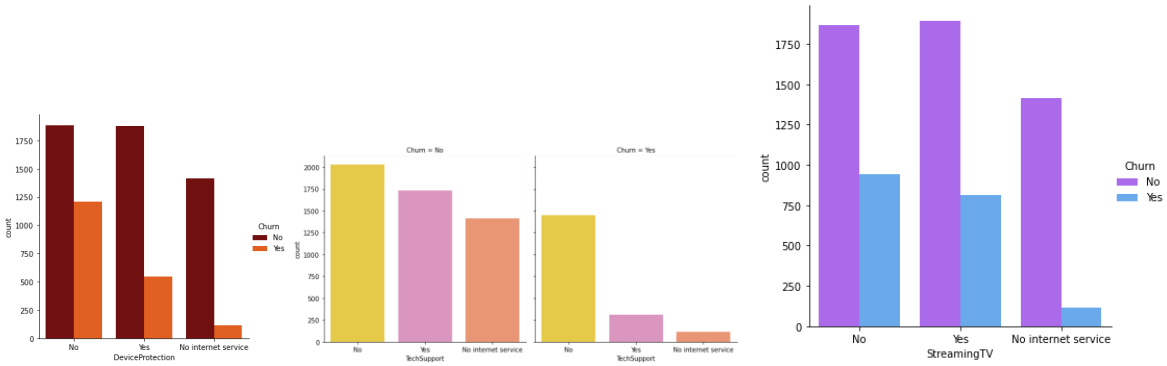
(i) The customers who have phone services from single line have high churn rate compared to the customers having phone services from multiple lines, also there are very a smaller number of customers who do not have phone services.

(ii) The ratio of churn is high when the customers prefer Fiber optic internet services compared to other services, may be this type of service is bad and needs to be focuses on and the customers who own DSL service they have very less churn rate.

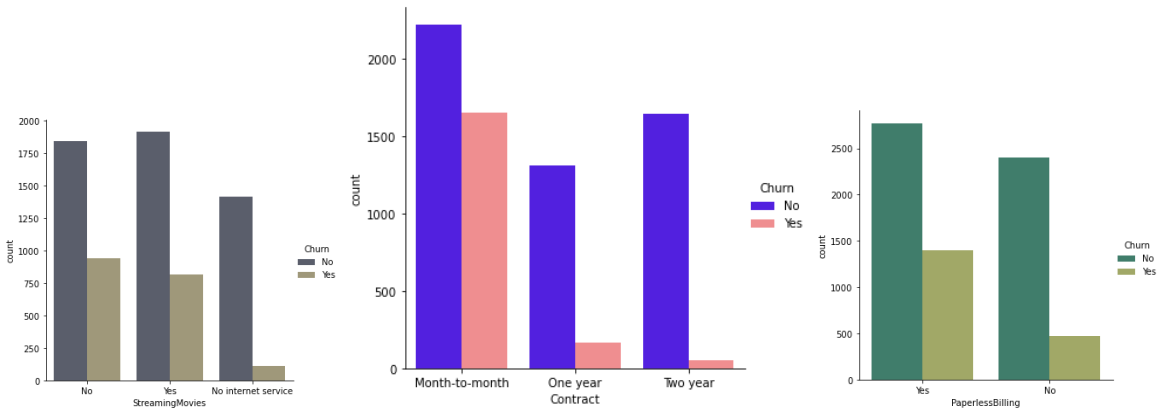


(i) The customers who have no internet service have very less churn rate and the customers who do not have online security services have high tendency to getting churned

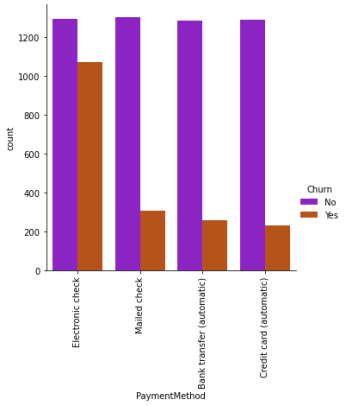
(ii) It is also same as in the case of online security. It is obvious that the customers having who do not have internet services they do not need any online backup. The customers who do not have online backup services they have high churn rate.



(i) The customers who do not own any Device protection have very high churn rate compared to others
(ii) Here we can clearly see that the customers who do not have any Techsupport then they have high churn ration
(iii) The churn rate is nearly same if the customer own StreamingTV or not.

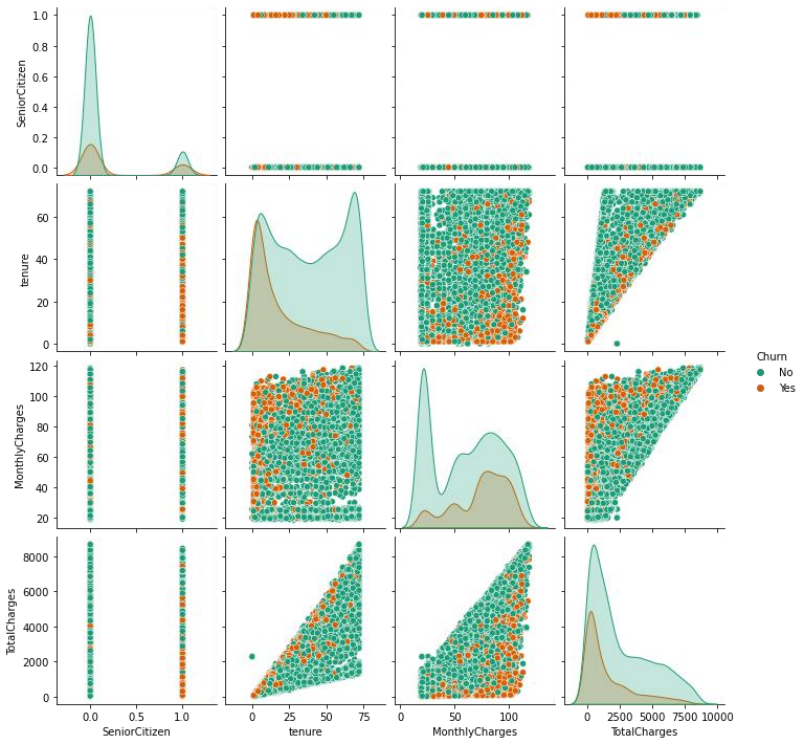


(i) The customers who are existing in the company they do not own StreamingMovies in their devices. And the churn rate is low when the customer does not have internet services.
(ii) The customer who has churned are mostly having month to month contract
(iii) The customers who prefer paperless billing they have high churn rate.



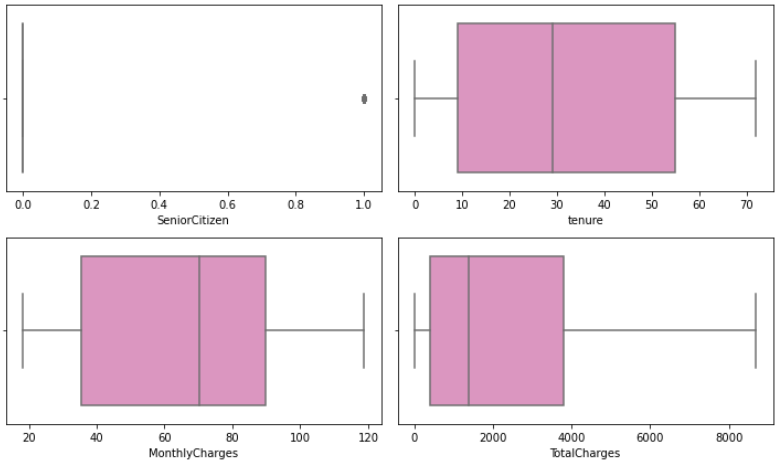
The customers who prefer electronic check have high churn rate also the customers who existing in the company use equal payment method.

Multivariate Analysis:



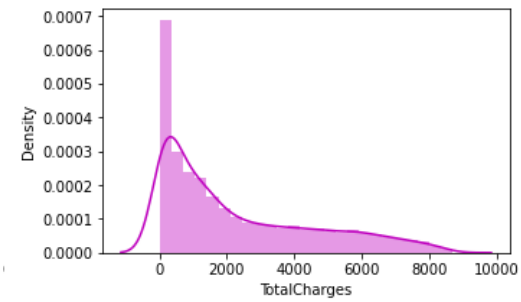
- The pairplot gives the pairwise relation between the features based on the target "Churn" On the diagonal we can notice the distribution plots.
- The features tenure and TotalCharges, MonthlyCharges and TotalCharges have strong linear relation with each other.
- There are no outliers in any of the columns but let's plot box plot to identify the outliers.

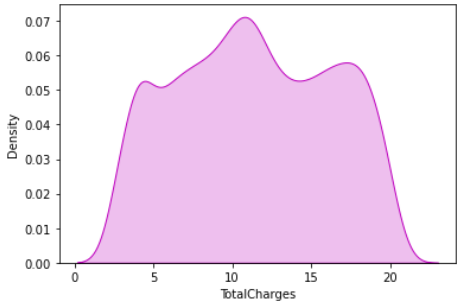
Outlier Detection:



The column Senior Citizen has outliers but it contains categorical data so no need to remove outliers Apart from this none of the columns have outliers.

Removing Skewness: At the time of univariate analysis, we saw Totalcharge column was right skewed.
I used cube root method for removing skewness from total charge column.
We can see skewness before and after-



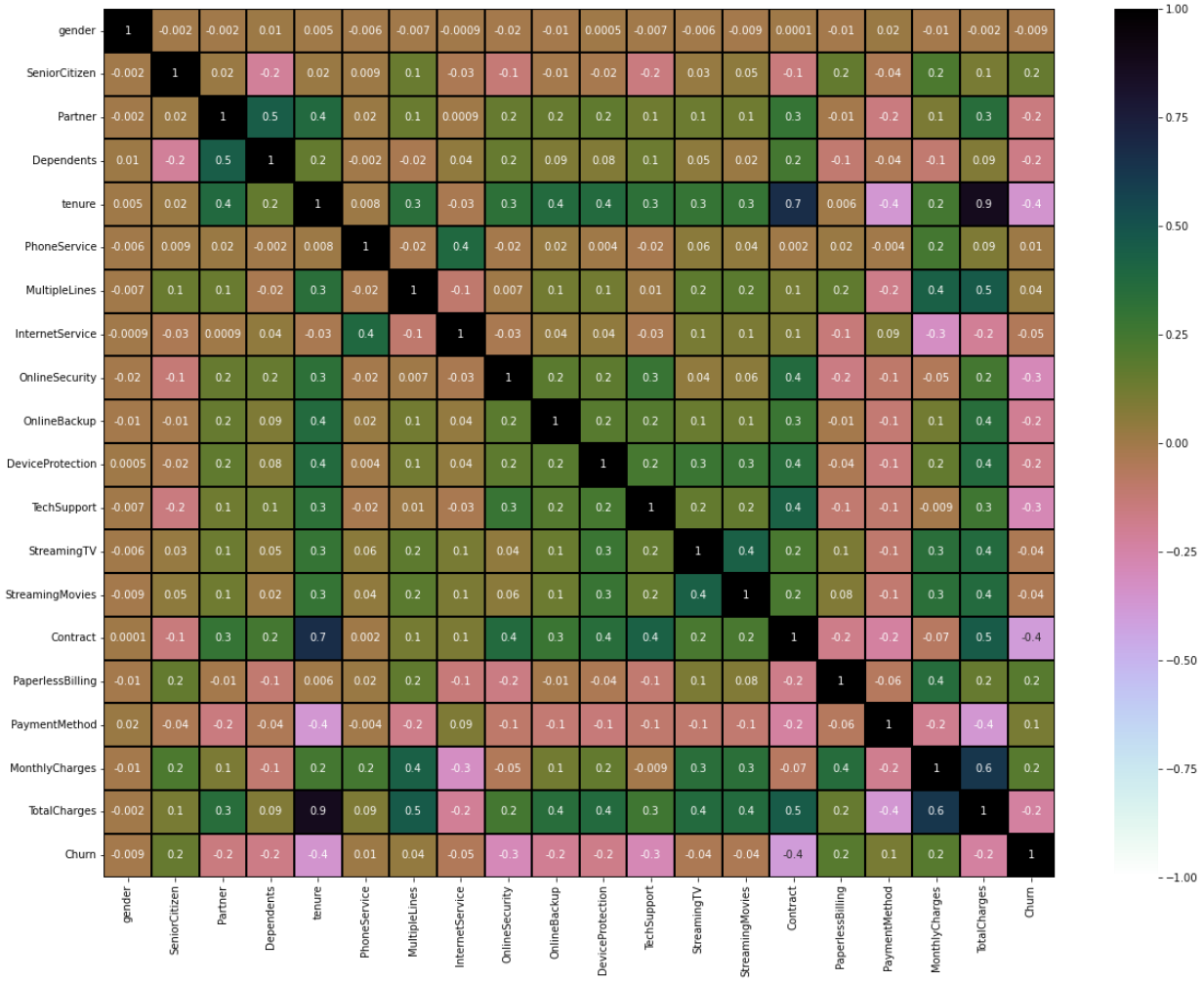


Encoding: Change categorical column into numerical column using ordinal encoder. After encoding the categorical column, we can see all the columns details here. The counts of all the columns are same that means no null values in the dataset. This describes method describes the count, mean, standard deviation, min, IQR and max values of all the columns

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.504756	0.162147	0.483033	0.299588	32.371149	0.903166	0.940508	0.872923	0.790004	0.906432
std	0.500013	0.368612	0.499748	0.458110	24.559481	0.295752	0.948554	0.737796	0.859848	0.880162
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	9.000000	1.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	0.000000	0.000000	29.000000	1.000000	1.000000	1.000000	1.000000	1.000000
75%	1.000000	0.000000	1.000000	1.000000	55.000000	1.000000	2.000000	1.000000	2.000000	2.000000
max	1.000000	1.000000	1.000000	1.000000	72.000000	1.000000	2.000000	2.000000	2.000000	2.000000

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
0.904444	0.797104	0.985376	0.992475	0.690473	0.592219	1.574329	64.761692	11.358079	0.265370
0.879949	0.861551	0.885002	0.885091	0.833755	0.491457	1.068104	30.090047	4.896177	0.441561
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	18.250000	2.659006	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	35.500000	7.381699	0.000000
1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	2.000000	70.350000	11.188354	0.000000
2.000000	2.000000	2.000000	2.000000	1.000000	1.000000	2.000000	89.850000	15.586542	1.000000
2.000000	2.000000	2.000000	2.000000	2.000000	1.000000	3.000000	118.750000	20.555116	1.000000

Correlation between the target variable and independent variables using HEAT map

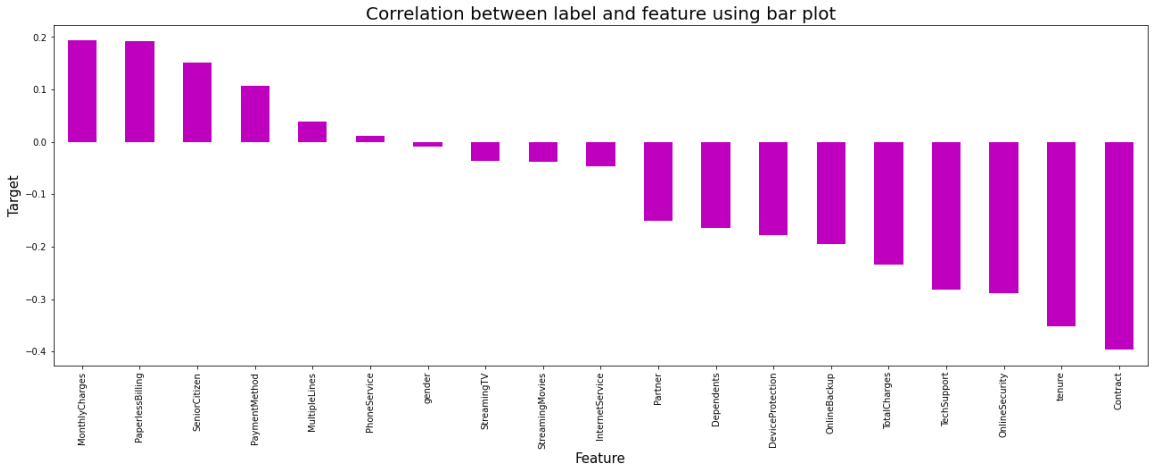


This heatmap shows the correlation matrix by visualizing the data we can observe the relation between feature to feature and feature to label. This heat map contains both positive and negative correlation.

- There is no much positive correlation between the target and features.
- The columns MonthlyCharges , PaperlessBIlling, SeniorCitizen and PaymentMethod have positive correlation with the label Churn.
- The label is negatively correlated with Contract, tenure, OnlineSecurity, TechSupport, TotalCharges, DeviceProtection, OnlineBackup, Partner and Dependents.
- Also, the column gender has very little correlation with the label we can drop it if necessary.
- The columns TotalCharges and tenure, Contract and tenure, TotalCharges and MonthlyCharges and many other columns have high correlation with each other.

This leads to multicollinearity issue, to overcome with this problem we will check VIF values and then we will drop the columns having VIF above 10.

correlation between label and features using bar plot



From the above bar plot, we can notice the positive and negative correlation between the features and the target. Here the features gender and PhoneService have very less correlation with the column.

Fifth Step: (Data separating, feature scaling) - In this step, I separated data into feature & label. After this I used feature scaling with help of "StandardScaler". We have scaled the data using standard scalarization method to overcome with the issue of data biasness.
In the heat map we have found some features having high correlation between each other which means multicollinearity problem so let's check the VIF value to solve multicollinearity problem.

Variance Inflation Factor (VIF)

	VIF values	Features
0	1.001696	gender
1	1.149704	SeniorCitizen
2	1.462974	Partner
3	1.383950	Dependents
4	12.357252	tenure
5	1.622391	PhoneService
6	1.398354	MultipleLines
7	1.870013	InternetService
8	1.256219	OnlineSecurity
9	1.192694	OnlineBackup
10	1.288549	DeviceProtection
11	1.312649	TechSupport
12	1.445353	StreamingTV
13	1.443816	StreamingMovies
14	2.519511	Contract
15	1.203384	PaperlessBilling
16	1.180762	PaymentMethod
17	5.384672	MonthlyCharges
18	16.702806	TotalCharges

By checking VIF value we can find the features which causing multicollinearity problem.
Here we can find the feature TotalCharges and tenure have VIF value greater than 10 which means they have high correlation with the other features.
We will drop one of the columns first, if the same issue exists then we will try to remove the column having high VIF (above 10).

	VIF values	features
0	1.001684	gender
1	1.149639	SeniorCitizen
2	1.460856	Partner
3	1.382106	Dependents
4	2.754468	tenure
5	1.622282	PhoneService
6	1.391652	MultipleLines
7	1.825876	InternetService
8	1.247696	OnlineSecurity
9	1.185932	OnlineBackup
10	1.280152	DeviceProtection
11	1.303573	TechSupport
12	1.443671	StreamingTV
13	1.442276	StreamingMovies
14	2.459201	Contract
15	1.202918	PaperlessBilling
16	1.180664	PaymentMethod
17	2.733024	MonthlyCharges

All the columns have VIF less than 10 which means the data is free from multicollinearity problem. So here we can move further to build our machine leaning models.

Oversampling

In [79]:
```
# oversampling the data
from imblearn.over_sampling import SMOTE
SM = SMOTE()
x, y = SM.fit_resample(x,y)
```

In [80]:
```
# checking value vount of target column
y.value_counts()
```

Out[80]:
```
0.0    5174
1.0    5174
Name: Churn, dtype: int64
```

We have used oversampling method to balance the data and checked the value count . Since the highest count of Churn column is 5174 so the data is balanced by oversampling all the categories to the count 5174.

Finally the data is also balanced then we can build our machine learning classification models.

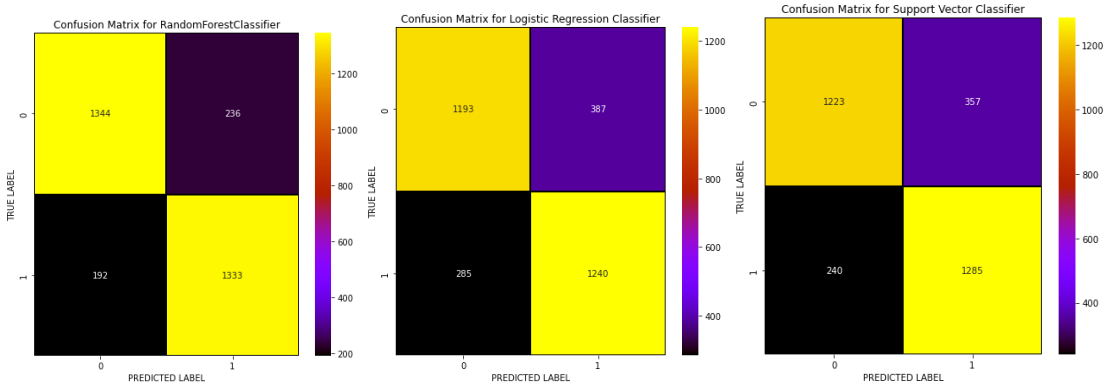
Sixth Step: (Modeling) - This is the most important step, in which we will make classification model using many techniques. Some steps are below-

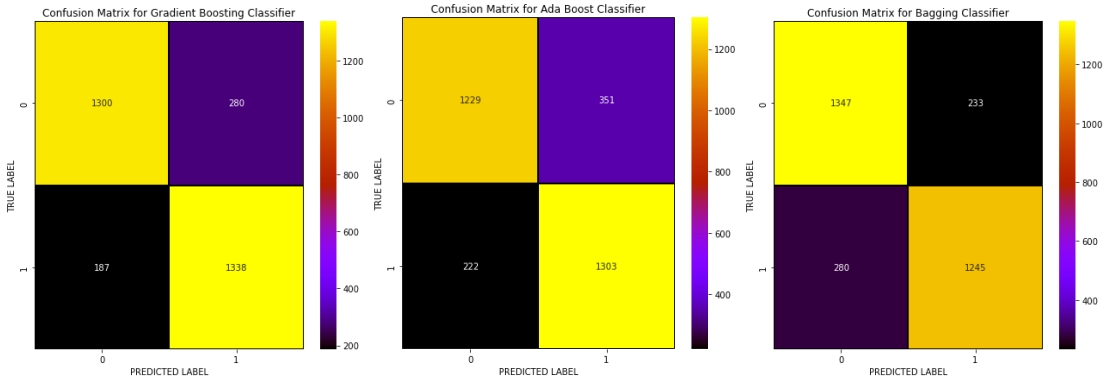
1. Find best Random State with highest Accuracy Score.
2. Separate data into training part and testing part.
3. Import classification algorithms & useful modules.

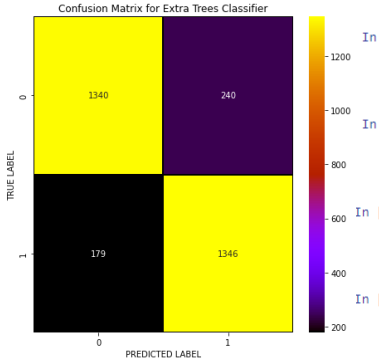
In [83]:
```
from sklearn.ensemble import RandomForestClassifier,ExtraTreesClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import GradientBoostingClassifier, AdaBoostClassifier, BaggingClassifier
from sklearn.metrics import classification_report, confusion_matrix, roc_curve, accuracy_score
from sklearn.model_selection import cross_val_score
```

4. Train each algorithm and check accuracy, confusion matrix and classification report.

Confusion matrix and Cross validation score of each model are below







Cross Validation Score

```
In [98]:
# cv score for Random Forest Classifier
print(cross_val_score(RFC,x,y,cv=5).mean())
```

```
0.8501260148079657
```

```
In [99]:
# cv score for Logistic Regression Classifier
print(cross_val_score(LR ,x,y,cv=5).mean())
```

```
0.7773502567227745
```

```
In [100]:
# cv score for Support Vector Classifier
print(cross_val_score(svc ,x,y,cv=5).mean())
```

```
0.8072136881454552
```

```
In [101]:
# cv score for Gradient Boosting Classifier
print(cross_val_score(GB ,x,y,cv=5).mean())
```

```
0.8353401839437942
```

```
In [102]:
# cv score for Ada Boost Classifier
print(cross_val_score(ABC ,x,y,cv=5).mean())
```

```
0.813497150248784
```

```
In [103]:
# cv score for Bagging Classifier
print(cross_val_score(BC ,x,y,cv=5).mean())
```

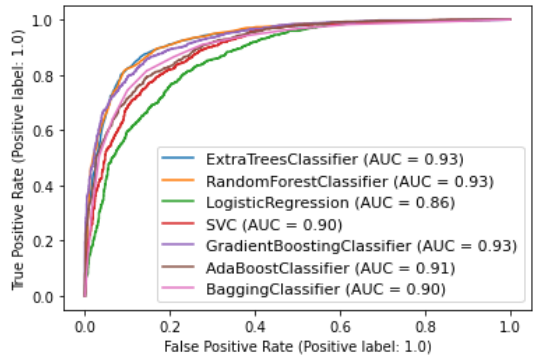
```
0.8228759955450018
```

```
In [104]:
# cv score for Extra Trees Classifier
print(cross_val_score(XT ,x,y,cv=5).mean())
```

```
0.86094755103518
```

Above are the cross validation score for the models

Plotting ROC and compare AUC for all the models used



Note: - From the difference between the accuracy score and the cross validation score we can conclude that ExtraTrees Classifier is our best fitting model which is giving very less difference compare to other models. So, I selected ExtraTrees as a final model.

Seventh Step: (Hyperparameter tuning, ROC and AUC curve, Model saving) -

Result of hyperparameter tuning;

```
In [108]: 1 GCV.best_params_

Out[108]: {'criterion': 'entropy',
           'max_depth': 20,
           'n_estimators': 300,
           'n_jobs': -2,
           'random_state': 50}
```

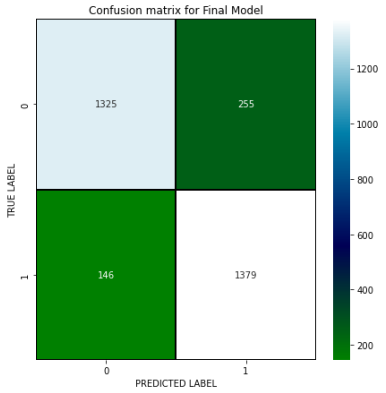
These are the best parameters values that we have got for Extra Trees Classifier

Final model selection;

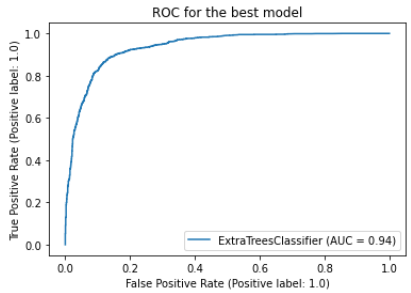
```
In [109]: 1 FinalModel = ExtraTreesClassifier(criterion ='entropy', max_depth=20, n_estimators=300, n_jobs=-2, random_state=50)
          2 FinalModel.fit(x_train, y_train)
          3 pred = FinalModel.predict(x_test)
          4 acc=accuracy_score(y_test,pred)
          5 print(acc*100)

87.085346215781
```

The accuracy of best model increased after tuning and giving 86.92% which is very good

Confusion Matrix for final model;



ROC-AUC Curve for final model;



Saving the Model

In [113]:
```
# saving the model using .pkl
import joblib
joblib.dump(FinalModel,"Telecom_Customer_Churn_Analysis.pkl")
```

Out[113]: ['Telecom_Customer_Churn_Analysis.pkl']

We have saved our model using joblib library

Predicting the saved model

In [114]:
```
# let's load the saved model and get the prediction

# loading the saved model
model=joblib.load("Telecom_Customer_Churn_Analysis.pkl")

# Prediction
prediction = model.predict(x_test)
prediction
```

Out[114]: array([0., 0., 0., ..., 0., 0., 0.])

These are the predicted churned values of the customers.

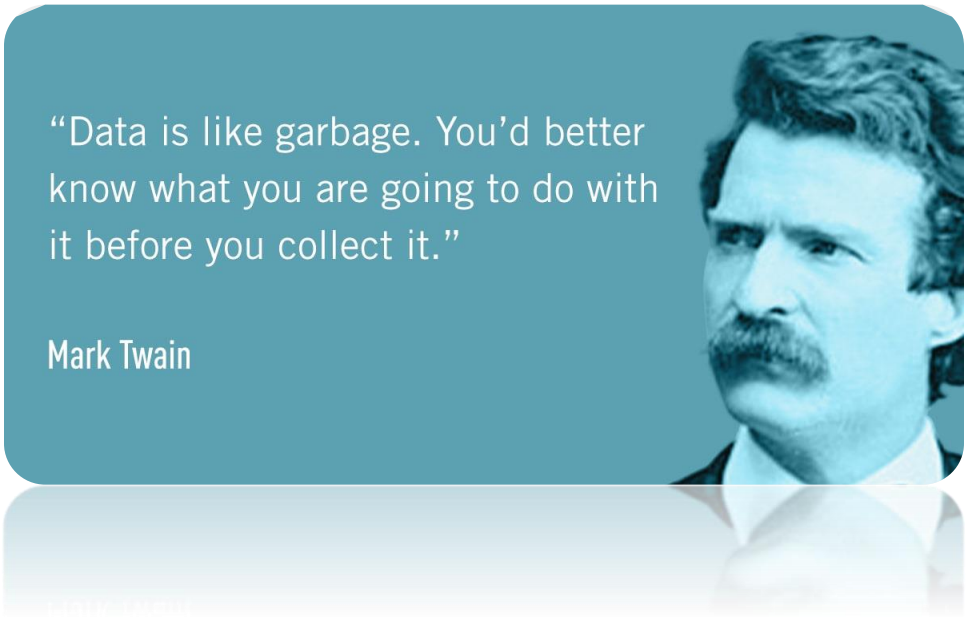
In [120]:
```
df1=pd.DataFrame([model.predict(x_test)[:],y_test[:]],index=["Predicted","Original"]).T
df1.sample(10)
```

Out[120]:

	Predicted	Original
2490	0.0	0.0
1536	0.0	0.0
2996	0.0	0.0
2411	0.0	0.0
3065	1.0	0.0
132	1.0	1.0
147	1.0	1.0
2472	0.0	0.0
2780	0.0	1.0
1029	1.0	1.0

Both actual and predicted values are almost same.

Now Our Churn Prediction Model is ready to use....



Thankyou....