Customer Churn Analysis

Problem Statement:

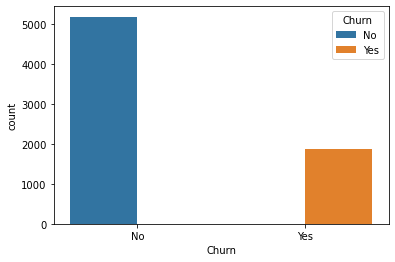
Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales 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.

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

DATA ANALYSIS

In this dataset of over 7000 customers, 26% of them has left in the last month. This is critical to the Telco business because it is often more  expensive to acquire new customers than to keep existing ones

The features in this dataset include the following:  
· **demographic data**: Gender, SeniorCitizen, Partner, Dependents  
· **subscribed services**: PhoneService, MultipleLine, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies  
· **customer account information**: CustomerID, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Tenure

Target is Churn, which has binary classes 1 and 0.

Out[3]:

|  | **0** | **1** | **2** | **3** | **4** |
| --- | --- | --- | --- | --- | --- |
| **customerID** | 7590-VHVEG | 5575-GNVDE | 3668-QPYBK | 7795-CFOCW | 9237-HQITU |
| **gender** | Female | Male | Male | Male | Female |
| **SeniorCitizen** | 0 | 0 | 0 | 0 | 0 |
| **Partner** | Yes | No | No | No | No |
| **Dependents** | No | No | No | No | No |
| **tenure** | 1 | 34 | 2 | 45 | 2 |
| **PhoneService** | No | Yes | Yes | No | Yes |
| **MultipleLines** | No phone service | No | No | No phone service | No |
| **InternetService** | DSL | DSL | DSL | DSL | Fiber optic |
| **OnlineSecurity** | No | Yes | Yes | Yes | No |
| **OnlineBackup** | Yes | No | Yes | No | No |
| **DeviceProtection** | No | Yes | No | Yes | No |
| **TechSupport** | No | No | No | Yes | No |
| **StreamingTV** | No | No | No | No | No |
| **StreamingMovies** | No | No | No | No | No |
| **Contract** | Month-to-month | One year | Month-to-month | One year | Month-to-month |
| **PaperlessBilling** | Yes | No | Yes | No | Yes |
| **PaymentMethod** | Electronic check | Mailed check | Mailed check | Bank transfer (automatic) | Electronic check |
| **MonthlyCharges** | 29.85 | 56.95 | 53.85 | 42.3 | 70.7 |
| **TotalCharges** | 29.85 | 1889.5 | 108.15 | 1840.75 | 151.65 |
| **Churn** | No | No | Yes | No | Yes |

There are some unique values in total charges column we have to remove it by replcing with NaN and then fill Nan value with mean of that column.

We have to drop unnecessary column like customer id which do not contribute.

EDA Concluding Remark.

* Visualisation of data using both matplotlib and seaborn library.
* Finding relationships between features using bar graphs, histograms, box plots, heatmap.
* Analysing both the numerical and the categorical columns separately.

Categorical columns=

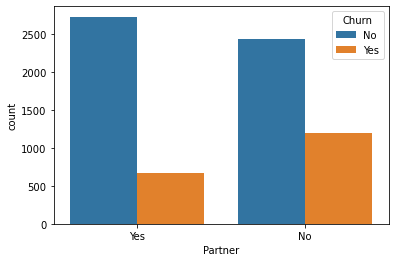
gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines',

'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',

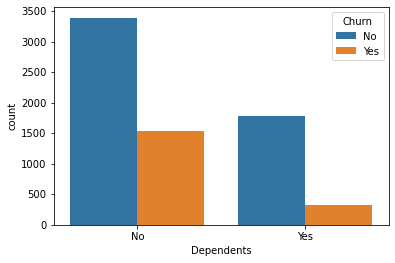
'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',

'PaperlessBilling', 'PaymentMethod', 'Churn']

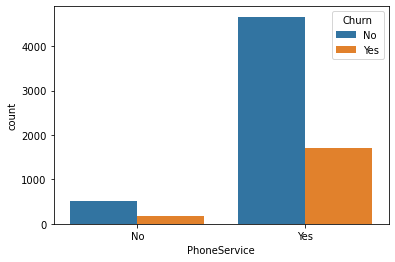
Partner



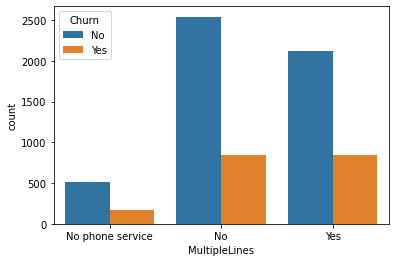
Dependents



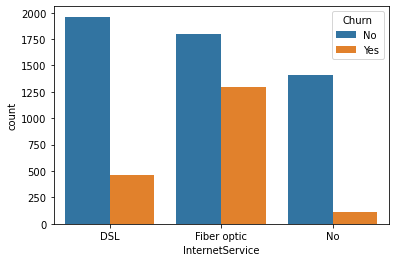
PhoneService



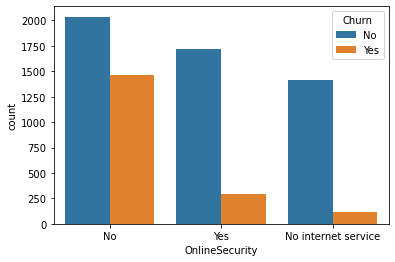
MultipleLines



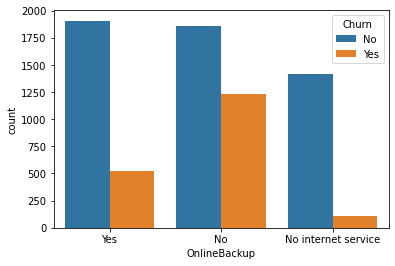
InternetService



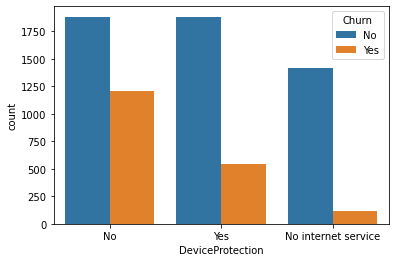
OnlineSecurity



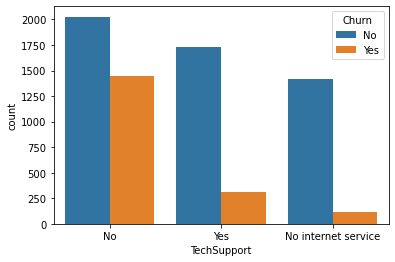
OnlineBackup



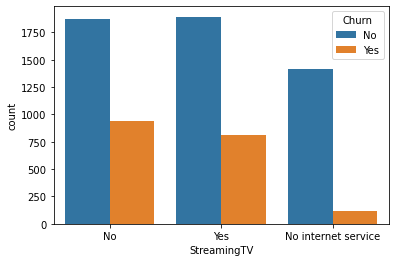
DeviceProtection



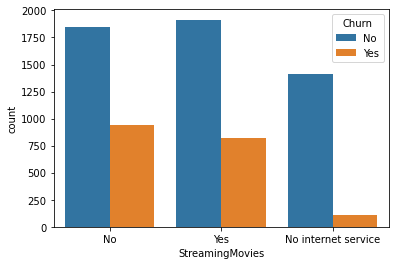
TechSupport



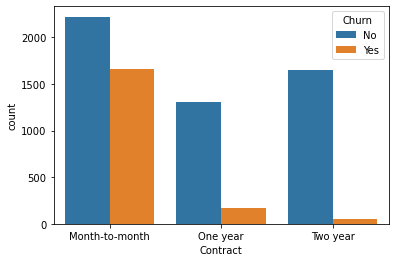
StreamingTV



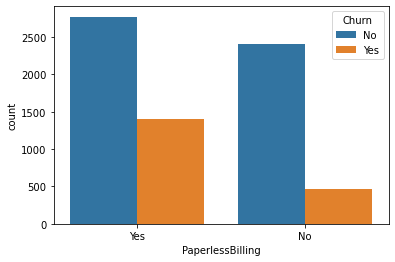
StreamingMovies



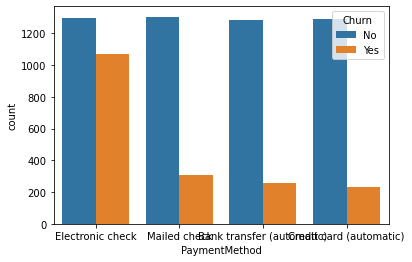
Contract



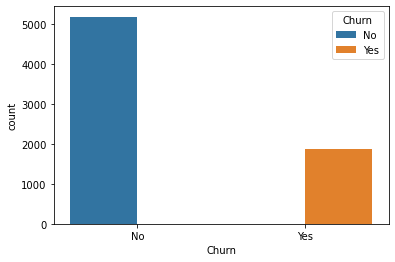
PaperlessBilling



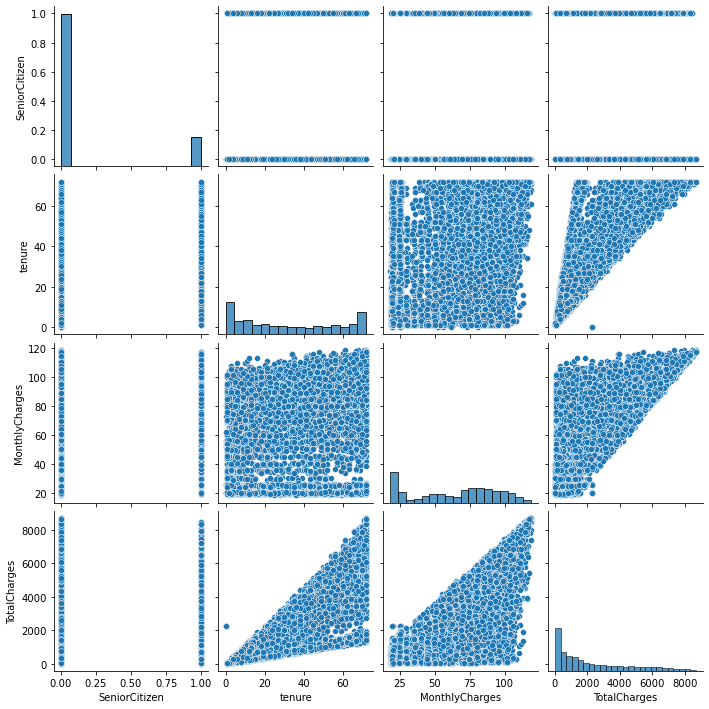
PaymentMethod



Churn



These 3 features Tenure, Monthly Charges and Total Charges are continuous data to be split into categories. When I looked at the pair plot



Correlation



Although the correlation matrix does not indicate any high degree of correlation with the dependent variable, it does serve us with a clear view of all the factors.

Here we can see some features are highly correlated and some are less correlated.some feature are also negatively correalated

From above categorical and numerical data visualisation againt churn and using heatmap we acan come to some conclusion that some are highly and some are weekly coreelated but don’t have to drop any column

# Pre-Processing Pipeline:

This is essential to perform since the machine learning algorithms only work on the numerical data. So, the necessity to convert the categorical column into a numerical one is must.because machine leraning cannot read categorical data

We can use here ordinal encoding

#converting categorical data into numeric ones

from sklearn.preprocessing import OrdinalEncoder

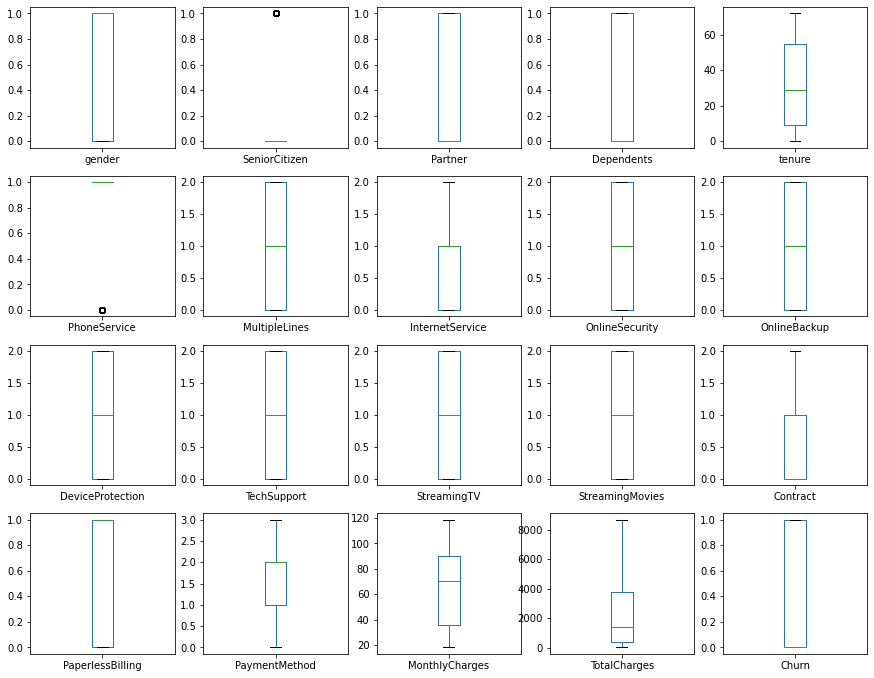
enc=OrdinalEncoder()

for i in df.columns:

if df[i].dtypes=="object":

df[i]=enc.fit\_transform(df[i].values.reshape(-1,1))

Using the above ordinal encoding method, categorical data can be replaced with number.

We have to check outlier 

So there are no outlier so we can preoceed further otherwise we have to remove it

# 5. Building Machine Learning Models.

We have to split in x and y .x should contain all features and y should have label

Now we will apply split into train test and apply some model to get prediction

This is classification problem so we have to use binary classification model here

1)logistic regression

2) knn

3)decision tree

4) GradientBoostingClassifier

5) RandomForestClassifier

Cross validation score of Logistic Regression model : 0.8023577125943608

Cross validation score of Random Forest model : 0.789011387831473

Cross validation score of knn model : 0.7537976240402606

Cross validation score of gb model : 0.804772082069811

Cross validation score of dt model : 0.7298022251435577

When I apply all model I got GradientBoostingClassifier as best model as it was giving high accuracy of 80%

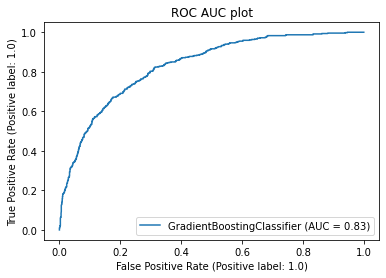
Confusion matrix=

array([[1172, 123]

[ 227, 239]]

So to increase accuracy we have to do hypertuning with the help of

grid search cv method .we will get best parameter then we will plot

roc curve  


conclusion:

I would propose to implement the gradient boostingclassifier model. And business can also focus on this list of features to understand whether a customer will likely to churn or not.

1. After auc curve our accuracy increased to 83%
2. The visualisation we were aiming to achieve at the beginning of this project has now become attainable
3. All of the performed analysis helped me a lot to know the **features on which price is highly positively and negatively coorelated with.**