# Problem Definition:

The dataset contains 21 features having different impact on the customer\_churn. Column 1 contains customerID which uniquely identifies the customer. Gender of the customer male or female. Partner column indicates that the whether customer have a partner or not. Dependent column indicates that whether customer have a dependent or not, dependents could be children, parents, grandparents etc. tenure column indicates about the months that a customer has been with the company by the end of the quarter specified. PhoneService column indicates whether customer has subscribed the home phone service with the company or not. MultipleLines indicates whether the customer have multiple telephone lines subscribed with the company or not along with no phone service also. InternetService column indicates whether the customer has subscribed company's internet service or not or type of internet service like DSL, FiberOptic, Cable. OnlineSecurity column indicates that whether the customer has subscribed additional online security service of the company or not. OnlineBackup column indicates that whether the customer have subscribed online backup service provided by the company or not. DeviceProtection column indicates that if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company. TechSupport indicates whether the customer have an additional device protection plan for their internet equipment provided by the company or not. StreamingTV column indicates about the internet service to stream television programming from the third-party provider. StreamingMovies indicates whether the customer uses their internet service to stream the movies from the third party provider. Contract column indicates about the current contract type month-to-month,one year,two year. PaperlessBilling column indicates whether the customer has chosen the paperless billing or not. PaymentMethod indicates about the mode of payment done by the customer to the company like Bank Withdrawal, Credit Card, Mailed Check. MonthlyCharges column indicates about all the monthly charges for all the services to the company. TotalCharges column indicates about the total charges calculated at the end of the quarter specified above. Churn Label indicates whether the customer left the company this quarter or not, Yes or No.

We have to predict about the Customer Churn is Yes or No.

# Data Analysis:

In the data analysis I have checked for the data type of the features by checking head and tail elements. Then I have checked for the datatype of the features and try to find any anomaly in that. I have found that the dataset is having object, int64 and float64 type in the column. After matching the datatype with the features data, I have checked for null values. Then I have checked for the unique values in the target feature and found that it is having unique values with object datatype. Then I have checked for the empty values in the features and target variable and found no empty values in the target column but found 11 empty values in the ‘TotalCharges’ feature. After that I have tried to find out the general information about the dataset and found that ‘TotalCharges’ column was having object datatype which needs to be converted into float64. After that I have replaced the empty values with the np.nan value and then converted the object type to float64 type using astype function. Then I have used the SimpleImputer for imputing the missing values, there I have used the strategy median since the data was skewed in that feature.

Now after imputing I have started the Univariate Analysis of the data by making a DataFrame for nominal data and named it ‘Customer\_Churn\_nominal’ and put some of the features having the nominal data. Then I have done the visualization of the nominal data using countplot, in the gender column I found that customer churn for both male and female are approximately equal. From the Partner column we can say that customers having partner having churn value ‘No’ is greater than the customers don’t have any partner and vice-versa for ‘Yes’ churn value. From the dependent column we can say that customers not having dependents are having greater value of churn as ‘No’ as compared to the customers having dependents in the same category. Customers subscribing the PhoneService have greater value of churn ‘No’ as compared to customers not subscribed the PhoneService. From the MultipleLines columns we can say that the customers not subscribing multiple lines have highest number of customer churn having value ‘No’ followed by the customers subscribed single line and customers having no phone service have least number of customers having churn value ‘No’. InternetService column tells us that the customer’s having DSL is having maximum number of churn value as ‘No’ as compared to FiberOptic and customers not having any InternetService. In the online security service column we have customers with security connection having more number of customer churn value as ‘No’ then other columns. From the online backup service column we can say that customers having subscribed the online backup of the company have more value of customer churn as ‘No’ as compared to other columns and No internet service. From the device protection column we can say that customers that don’t have DeviceProtection subscribed have highest value of customer churn value as ‘No’, followed by other columns, and same order for customer churn value ‘Yes’. Customers having TechSupport not subscribed have maximum value of customer churn as ‘No’ followed by customers subscribed it and customer having no internet service have least value of ‘No’ for customer churn, and customer churn value ‘Yes’ is following the same trend as followed by the customer churn ‘No’. Customers subscribed to StreamingTV service have maximum number of customer churn as ‘Yes’ followed by customers having not subscribed and least with ‘No internet service’. Same trend is followed by customers having customer churn value as ‘Yes’. From the streaming movies column, we can say that we can say that customers subscribed to streaming movies service have highest value of customer churn as ‘No’ followed by customer churn customers who are not subscribed to the service and followed by the customer with ‘no internet service’. For the customer churn value ‘Yes’ we have maximum value for customer not subscribed to StreamingMovies service followed by customers subscribing the service and least value for customers with No internet service. From the contract column we can say that customers having month-to-month contract have maximum value of customer churn as ‘No’ followed by One-year-contract and followed by two-year-contract. Same pattern is followed by the customer churn value ‘Yes’ for different customers having contracts. From the paperless billing we can say that customers having customer churn value as ‘No’ is greatest for customers with paperless billing and same trend is followed by the customers with ‘Yes’ as customer churn value. In the payment method column, we can say that customers having different payment method have different customer churn value customers having churn value as ‘No’ for various methods of customer churn.

After doing the countplot I have converted the string data to numeric type since the machine learning models don’t understand the string data using Label Encoder. Then I have used the SMOTE analysis for the imbalanced class since the machine learning model may become biased towards majority class of the target variable. Now the dataset value has increased approximately 1.5 times. After that I have joined both target variable and features of the dataset to make it again a single dataset. Then I have used distribution plot for continuous values. Some of the columns have skewed value and some columns have uniform distribution in the target dataset. After that I have checked for the Outliers for various columns of the features. And from the outlier checking I found that SeniorCitizen, PhoneService, TotalCharges columns have come outliers in them. After that I have done Bivariate Analysis on the data for various columns using scatterplots. From the bivariate analysis we can say that some correlation can be found in TotalCharges and MonthlyCharges.

# EDA Concluding Remarks:

After that I have used describe method of pandas for checking various Statistical parameters like count, mean, standard deviation, minimum value, various quartiles like 20,25,40,50,60,75,80 and maximum value. From the analysis I found that Mean is higher than median in SeniorCitizen, Partner, Dependents,tenure,PhoneService,OnlineSecurity,DeviceProtection,TechSupport,Contract,TotalCharges so the data is positively skewed in these columns.

Small gap can be found between 75 percentile and max in customerID, tenure, InternetService, OnlineSecurity, TechSupport, Contract,PaymentMethod,MonthlyCharges,TotalCharges,so some outliers are present in these columns.

Then I have used the Multivariate Analysis for various columns using the pairplot and also used corr() function to find out the correlation coefficient for various features. Then I have used the correlation heatmap for visualizing the correlation value for various columns and found out that Columns having positive correlation are: tenure and TotalCharges with value=0.86, tenure and Contract with value=0.69. Then I have used the zscore method to find out the outliers and found out that 762 rows are having outliers which are 7.3 percent of the total rows, meaning that a total of 7.3 percent of data will be lost if we remove the outliers, so I have used the original DataFrame.

# Pre-processing Pipeline:

In the Pre-processing Pipeline, there are two steps that we follow. First is Gathering data which can be real time data or the data collected from various sources such as a file, database, survey and other sources. Second is data pre-processing which includes checking for missing data and imputing it, skewness detection and removal. After that I used the skew() method for skewness checking for various columns of the dataset and found out that Columns having positive skewness are: SeniorCitizen with value=2.190890, Dependents with value=1.091246, tenure with value=0.514831, OnlineSecurity with value=0.718351,TechSupport with value=0.692289,Contract with value=1.097111,TotalCharges with value=1.130017.Columns having negative skewness are: PhoneService with value= -2.783234,PaperlessBilling with value=-0.602445.

After that I have separated the target column and features columns, then used the power\_transorm from the sklearn library of machine learning with method ‘yeo-johnson’ for complete removal for positive and negative skewness both. Then I have used the Standard Scaler technique for scaling the normalised data after power\_transform method. Since there was not much of the multicollinearity problem in the dataset as was seen from the heatmap we will not use any collinearity removing technique like VIF or PCA for removing multicollinearity. Thus, putting the data directly to train\_test\_split.

# Building Machine Learning Models:

In building the machine learning model, I have used the train\_test\_split method from sklearn library of machine learning to find out the best random state from 200 random states. I have used the DecisionTreeClassifier for my best random state selection and also calculated the accuracy score for that random state and finally found out that maximum accuracy was found out at random state 70 with value 82.7. Now I have used five classifier algorithms e.g., SupportVectorClassifier, RandomForestClassifier, AdaboostClassifier, GradientBoostingClassifier and KNeighborsClassifier for calculating evaluation metrics like accuracy\_score, confusion\_matrix and classification\_report and found out that various parameters like precision, recall, f1-score, macro-avg and weighted avg. From the accuracy data for various algorithms, I found that GradientBoostingClassifier and RandomForestClassifier both have same accuracy with maximum value of 87. After that I have used the Cross-Validation-score and in that I have used k-fold cross-validation taking the value of cross-validation 5. After using the cross-validation score I found score, score mean value and standard deviation of the score. And finally, I found that cross-val-score of the RandomForestClassifier is the highest with value 84. Now I have compared the value of cross-val score and model accuracy and found out that RandomForestClassifier have minimum difference between the model’s accuracy and cross-val-score so we can say that the problem of overfitting and underfitting is least in this algorithm, so I have chosen this algorithm for my model. If a model has overfitting problem, then it will try to capture all the points in a dataset in this process it will also capture the noises present in the dataset making it less accurate for prediction. So best model will be that having least underfitting and also least overfitting.

After that I have done hyperparameter tuning for, hyperparameters in a model express important properties of the model such as its complexity or how fast it should learn. Hyperparameters are those parameters having major impact on accuracy and efficiency of the result. There are basically two types of search methods first is GridSearch and second is RandomSearch. In the GridSearch different hyperparameters are defined in the dataset and they are combined using Cartesian product to form a multidimensional grid. After that we will try all the parameters in the grid and then select the hyperparameter setting with best result. In the RandomSearch Random all the points in the Grid are not searched rather only Random points are searched making the process less exhaustive.

Since the number if rows in the dataset are large so I have chosen the Randomized Search to make the hyperparameter less computationally exhaustive. I have imported the RandomizedSearchCV from sklearn.model\_selection. After that I have selected four parameters from the list of parameters for RandomForestClassifier. After that I have fit the training x and y data in the RCV and found the best parameters. Finally, I have created my Final Model using those best parameters of RnadomForestClassifier and found out the accuracy for that model which comes out to be 80. After that I have calculated the true positive rate and false positive rate to find out the roc curve. Roc curve is basically used to evaluate our machine learning model. It is graph that shows the performance of a classification model at all possible thresholds. It is plotted between two parameters tpr and fpr. It is plotted at y-axis having ‘sensitivity’ and at x-axis ‘1-specificity’. An AUC measures how well a model is able to distinguish between two classes. After that I have plotted the roc curve and found out the ROC Curve area that is 0.80.

Finally, I have saved my model using the joblib library in python to save my model.

In the conclusion part I have made a made two arrays first one is y\_test for the testing data of target column and other one is predicted which is predicted y value in the dataset using x test value.

# Concluding Remarks:

From the above project of building a Machine Learning model for classification problem, I have found that data cleaning is an important part for building any machine learning model because cleaned data will make our machine learn data more accurately, so I have my focus on data cleaning more using Simple Imputer for imputing missing values, removing the outliers using z-score method and removing the skewness also for normalizing the data. Data visualization is also a very important for building Machine Learning model because without visualization we will not be able to know the insights of data and its hidden patterns. In the visualization part for univariate analysis countplot and distplot are used. In Bivariate analysis we use scatterplots since it tells us about the the correlation between columns. And the third one is Multivariate Analysis which tells us about the correlation between different columns. Imbalanced dataset is also a problem in the dataset since if target column is imbalanced then our algorithm will be biased while learning datapoints. So SMOTE analysis is best practice to balance the target column if it is having less than 0.1 or 1 million rows, so I have used SMOTE analysis. Choosing the correct model as our final model is also a challenge as we have to train our model in 4-5 algorithms and then finally selecting the model with least overfitting and underfitting. Choosing the correct model will give accurate results. Overall result of prediction of Customer churn was obtained and was used along with y\_test data to check the model accuracy.