**Customer Churn Analysis**

**Problem Definition**

Customer Churn is when the subscribers of a company stop using their services. This can be due to multiple reasons. Organizations are exceptionally excited about estimating churn since keeping a current client is definitely more affordable than procuring another client. New business includes working leads through a business pipe, utilizing showcasing and deals financial plans to acquire extra clients. Existing clients will frequently have a higher volume of administration utilization and can create extra client references.

Client maintenance can be accomplished with acceptable client care and items. However, the best path for an organization to forestall whittling down of clients is to really know them. The huge volumes of information gathered about clients can be utilized to assemble churn forecast models. Realizing who is well on the way to abandon implies that an organization can focus on zeroed in advertising endeavors on that subset of their client base.

Forestalling client churn is fundamentally critical to the broadcast communications area, as the hindrances to passage for exchanging administrations are so low.

We have given with a IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

We will be using the scope of machine learning in predicting the churn.

**Data Analysis**

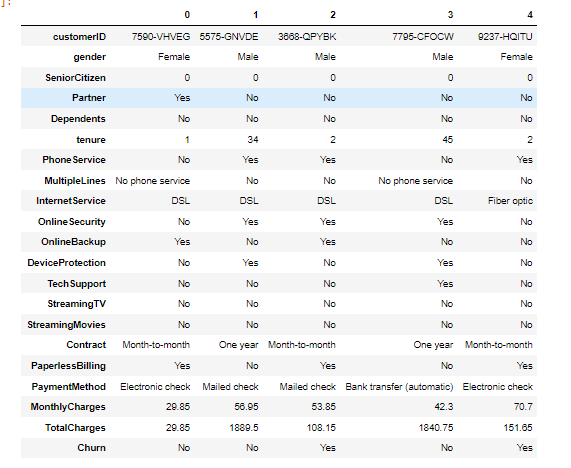
We are importing the data set using the pandas library to a data frame in order to conduct the data analysis. As an initial observation, The data set is found to be having,

* 7043 rows and 21 columns

As we check the initial 5 rows of the data set. We have the following features,

1. customerID
2. gender
3. SeniorCitizen
4. Partner
5. Dependents
6. tenure
7. PhoneService
8. MultipleLines
9. InternetService
10. OnlineSecurity
11. OnlineBackup
12. DeviceProtection
13. TechSupport
14. StreamingTV
15. StreamingMovies
16. Contract
17. PaperlessBilling
18. PaymentMethod
19. MonthlyCharges
20. TotalCharges
21. Churn

The ‘Churn’ Column will be our target column

The initial head of the data set looks,

We will be checking the data types of the given columns, as an initial observation we can see that,

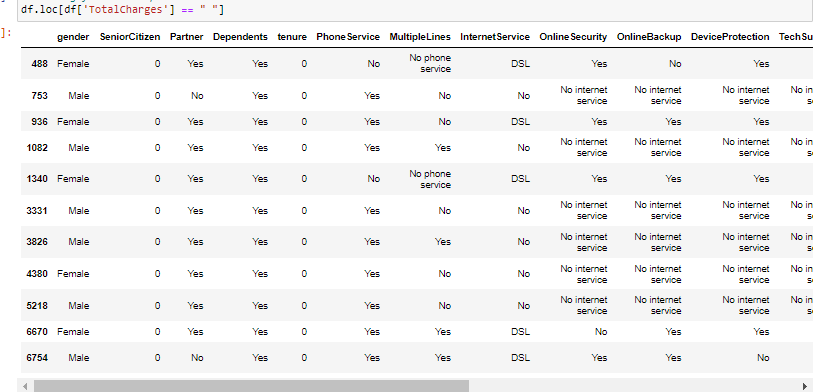
* The columns,Monthly Charges, Senior citizen and Tenure are numerical
* The rest of the columns are shown as ‘object’ data types however, the column total charges is also numerical we will have to change the same later

The customer ID column seems to be having all unique values and would not benefit us in building the model, we will be dropping the column from the date set.

We are checking the data set for any null values however the data set seems not have any.

The column total charges will have to be converted to a numerical datatype for us to perform graphical analysis on the same.

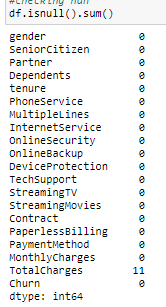
As we have run into error converting the same, the column was subsequently checked for empt spaces and we received the following result,

Since many values have returned the ‘Yes’ values for the check we will have to clear them.

As an initial step we are converting them to ‘NaN’ values to be handled later.

The number of the NaN values in the data set was checked which returned the value 11.

Since the values are continuous, we are filling the ;NaN; values with the total mean of the total charges column.



First, the datatype of the same was changed to float and then we are imputing the ‘NaN’ values with the mean.

**Uni-variate Analysis**

The independend columns are checked using seaborn and matplotlib and are plotted to get an exact idea regarding the data distribution.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Observations** on uni-variate analysis,

* The gender column has male and female sub categories
* Categorical elements categorizing if the user is seniour citizen, majority in the dataset are not senior citizen
* Majority in the datasets doesn't have a partner
* Majority doesn't have any dependence
* All most all the customers have phone service
* There is is significant amount people with multiple phone lines
* Majority of the customers have fiber optic internet connection followed by DSL
* Large amount of customers doesn't have internet security,online backup, Device protection and tech support
* Even though many have streaming tv and streaming movies, the significant number of users doesn't have any
* many users have a month to month plan subscription
* Large amount of users choose paperless billing
* Most elected payment method is electronic check
* The target column ‘Churn’ looks imbalanced
* Tenure ranges from 0--70
* Monthly charge ranges from 20 to around 120, with majority chose 20 plan
* Total charges vary from 0 to 8000+

**Bi-Variate Analysis**

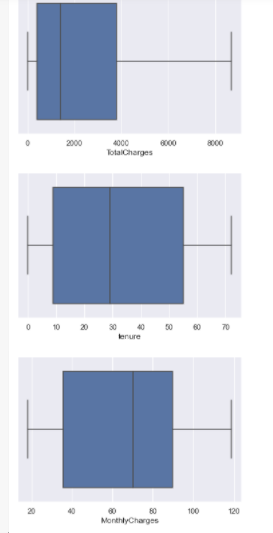
We are trying to analyses various categorical columns in regard with the target columns from which we could try to figure some relationships between various features.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Observations** from Bi-variate Analysis

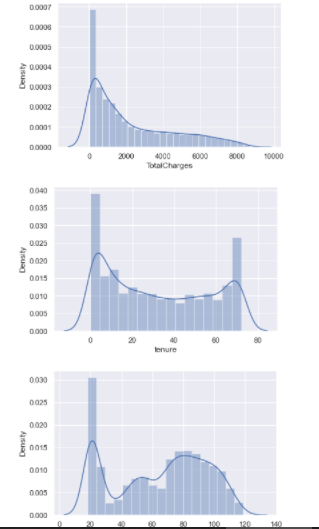
* Churn value in gender seems balanced to male and female
* People who are single seems to leave the service often
* People with no dependents seems to leave the service often,
* Customers with a phone service seems to leave the service often
* customers with fiber optic internet connection seems to leave the service often,
* customers with no online security seems to leave the service often
* subscribers with no online backup seems to leave the service often
* Subscribers with no device protection seems to leave the service often
* Customers with no tech support seems to leave the service often
* Users with no streaming tv or movies seems to leave the service often
* People with multiple lines data seems balanced in terms of churn
* Month-to-month subscribers tends to not reuse the service.
* People who prefer paperless billing seems to leave the service often
* Subscribers who use electronic check tends to leave the service

We will have to check the data for outliers, using boxplots to visualise the presence of outliers in the numerical columns. The plots looks like,

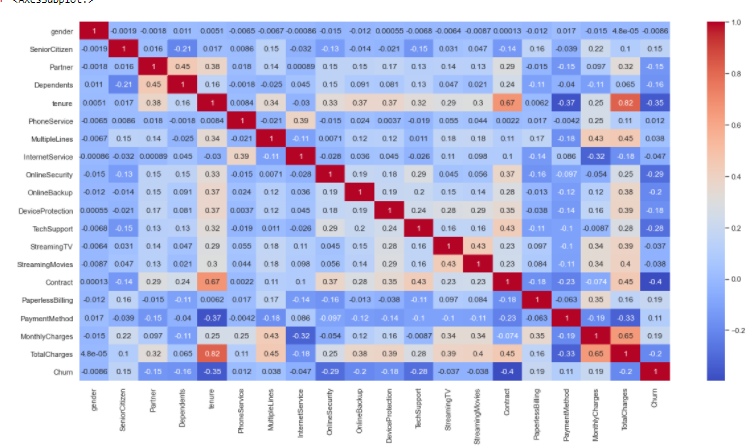
****

Since the plots doesn’t show any outliers in the numerical columns we will have to check the zscore for the same

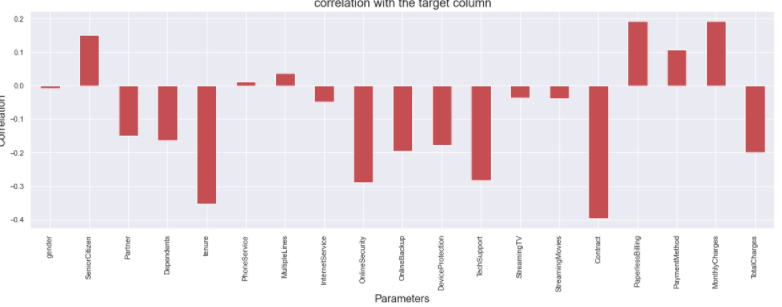
We will be checking the skew of the data using distribution graphs. Plotting the same produced the below results,

****

As per the plots we can see that the data might have a slight skew, However we are leaving the same untouched as doing the same produced better results in the long run.

****

The correlation plot, since the number of features where lot higher, we have plotted the correlation graph to the target column.

The correlation to the target column plot is showing that,

* The paperless billing seems to be having high correlation to the churn value
* Monthly charges and paperless billing has a positive correlation
* Tenure and contract have the highest negative correlation

As we have previously seen the Churn column, The number of ‘Yes’ and “No’ values are highly imbalanced, this might cause a bias in the overall performance of the models.

In order to handle the imbalance issue we will be using the most commonly used technique which is known as SMOTE: Synthetic Minority Over-sampling Technique. The method involves supplementing the training data with multiple copies of some of the minority classes.

The sample code is given as,

x\_over,y\_over=sm.fit\_resample(x,y)

The x\_over and y\_over is our new x and y. The x part will be treated with standard scaler After splitting the X and Y variables, we need to scale them to the common level. The values are brought to common level and then we can apply further machine learning algorithm to the input data. The values in the continuous columns are quite high when compared to the smaller encoded values in the categorical columns. Hence, this is an important step to get a good model.

StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature. Since there are no outliers, the scaling should work perfectly.

**EDA Concluding Remarks**

* Customer value is dependent upon many features
* Monthly charges and paperless billing has a positive correlation to the churn value
* Tenure and contract have the highest negative correlation to the churn column
* People who has dependents and are in a relationships tends to stay with the subscription often
* People who tends to leave more often chose the least lengthy subscription plans

## **Pre-processing Pipelines**

## We have removed the null values

## Customer ID column was removed from the datasets

## The data was checked for skewness and outliers

## The datatype of the total charges column was changed

## The data was split into x and y

## Over sampling was used on the data to handle bias

## The standard scaler was used on the data

**Building Machine Learning Models**

* We are splitting the features as x and the target variable as y as the initial step.
* Since the target column has categorical values, we are importing all the classification algorithms.
* We are implementing a small function to identify the best random state while splitting the data.

 The best accuracy is 0.785829307568438 on random state 140

* Splitting the data for training and testing with the previously identified random state
* Defining the functions for model building and cross validation.
* Calling different models and checking the model accuracy score and cross validation score

**Selecting the best model:**

After Using **Logical regression** algorithm we were able to achieve the following results

Accuracy score: 0.785829307568438

Mean cross validation Score: 0.767589280919392

difference between crossvalidation score and actual score: 0.018240026649046004

The accuracy score seems decent and the difference between the actual score and the cross validation score is minimal

Using **GaussianNB**

Accuracy score: 0.7761674718196457

Mean cross validation Score: 0.7659474226154201

difference between cross validation score and actual score: 0.01022004920422559

We are getting a stable result using Gaussian NB

The **DecisionTreeClassifier** Model also produced a similar result of,

Accuracy score: 0.814170692431562

Mean cross validation Score: 0.7979455640312597

difference between cross validation score and actual score: 0.016225128400302324

We are getting the following results using the **Random forest Classifier,**

Accuracy score: 0.8618357487922705

Mean cross validation Score: 0.8397938746109466

difference between crossvalidation score and actual score: 0.022041874181323973

We are also checking the accuracy score of **AdaBoostClassifier**

Accuracy score: 0.8454106280193237

Mean cross validation Score: 0.8318642579789532

difference between crossvalidation score and actual score: 0.013546370040370403

The **decision Tree Classifier** gave as the best score of them all,

After running almost 6 different regression algorithms.

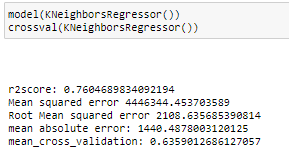
The Decision Tree Classifier gives the best stable model based on the lowest variance with the cross validation score, even though algorithms like random forest regressor gave as the better r2 scores.

The scores are,

Accuracy score: 0.814170692431562

Mean cross validation Score: 0.7979455640312597

difference between cross validation score and actual score: 0.016225128400302324

****

We are selecting the Decision Tree Classifier for hyper parameter tuning.

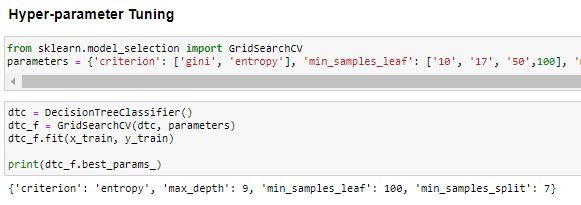
We are running GridSearchCV for The Decision Tree Classifier.

GridSearchCV tries all the combinations of the values passed in the dictionary (which I’ve set manually) and evaluates the model for each combination using the Cross-Validation method. Hence after using this function, we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

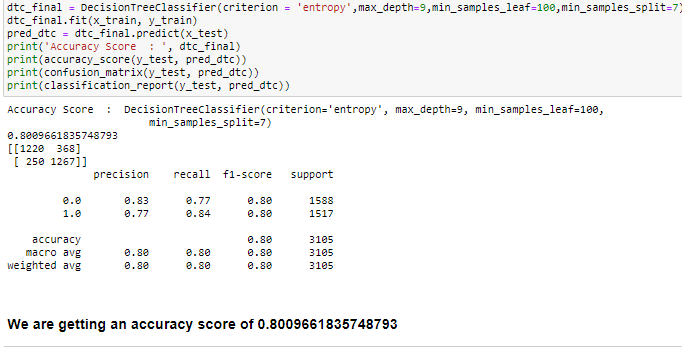
It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters

The parameters are declared as a dictionary and checking the best values for 4 different features. After running the same the best values for the parameters we passed to the algorithm are,

{'criterion': 'entropy', 'max\_depth': 9, 'min\_samples\_leaf': 100, 'min\_samples\_split': 7}

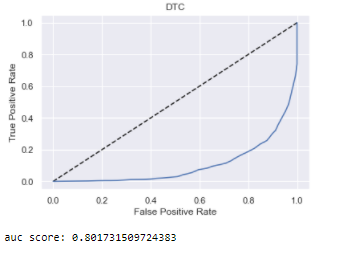
s

We will be running the best model using the new parameters,

****

The hyper parameters has produced a model with an accuracy score of 80% which is a pretty stable model. We can plot the AUC -ROC curve,

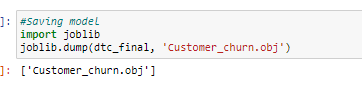
AUC-ROC Curve: This is frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition, the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

****

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represent the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

The AUC Score is 0.80 which also seems decent. We can decide on this final model.

**Saving the model**: I’ve used the Joblib module to build the ML model. It provides utilities for saving and loading Python objects that make use of NumPy data structures, efficiently.



## **Concluding Remarks**

## The subscribers stays with the service with many reasons and right marketing and targeted advertising techniques could enhance the both the longevity and the customer base

* Churn rate is a health indicator for subscription-based companies. The ability to identify customers that aren’t happy with provided solutions allows businesses to learn about product or pricing plan weak points, operation issues, as well as customer preferences and expectations to proactively reduce reasons for churn.

The Model created gives as an accuracy score of 80% in predicting the customer churn status