**Customer Churn Analysis – Machine learning Project**

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.*

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

We need to build a classification model to predict the customer’s take on the service.

**1.Importing the basic libraries to begin.**

Text

Description automatically generated

Loading the dataset.

A picture containing table

Description automatically generated

There are 7043 rows and 21 columns in the dataset.

**2. Data Analysis**

We need to see the features available in the dataset to analyse.

Graphical user interface

Description automatically generated with low confidence

The dataset has 3 numeric features and 18 categorical features. Our target variable ‘Churn’ is a categorical feature, so we need to proceed with building the classification model.

Text

Description automatically generated

Checking the columns to have an idea on the dependables on the target variables.

Table

Description automatically generated with medium confidence

We see that the customer ID is a kind of identity of the customers, which doesn’t help much in predicting the target value, so we would drop that column.



Let’s check the description of the dataset.

Table

Description automatically generated

So we observe that the outliers may be less, as the difference between the Max value and the 75th percentile is less.

We will now check the values each features are holding.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

…the output continues with the Churn as the target and featuring the only ‘Yes’ or ‘No’.Also, we could see that only Tenure, Monthlycharges and Totalcharges have continuous values.

Let’s check the Null values.

Text

Description automatically generated

We see that there are no Null values present.Let’s have a pictorial review on the same with the help of the heatmap.

Chart

Description automatically generated

Thus , we could see that there are no Null values in the dataset.

Let's check for each feature present in the dataset and its effect on our target variable. This is very important as we will understand customer behaviour through different patterns and will help the company to fill the gap that has been created resulting in the customer leaving the services.

Now ,we need to understand on how the customers are deciding towards the services.With the help of the dataset we need to analyse which would really help the company to retain the customers.

**3. Data Pre-processing and EDA**

**Churn**



Chart, pie chart

Description automatically generated

This is our target variable which we need to predict. In this dataset we have 26.5% of the customer who have left the service.

**Gender**



Chart, pie chart

Description automatically generated

Text

Description automatically generated

Chart, bar chart

Description automatically generated

The Gender is not playing an important role in deciding whether the customer will continue to use the services or not as they are almost equally distributed.

**Senior Citizen**

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

1 indicating Yes for the senior citizen. We have only 16.2% of the customer as senior citizen and the percentage is higher for them to leave the service.

**Partner**

Chart, pie chart

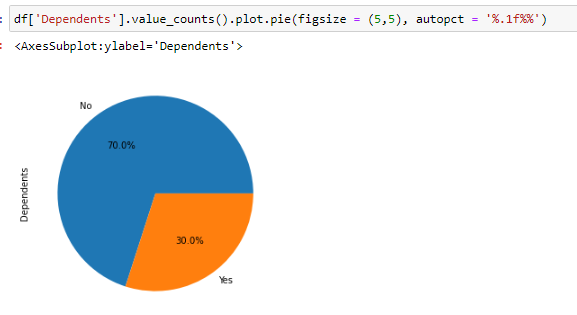
Description automatically generated

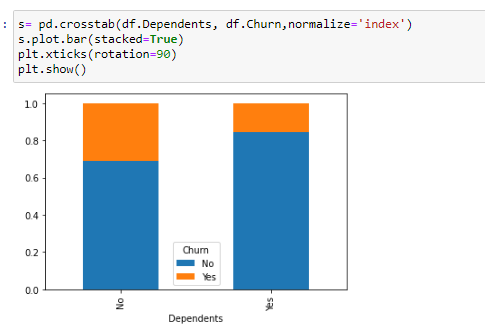
Chart, bar chart

Description automatically generated

48.3% of the customers have partners and they seem to be loyal with the services. Reason could be the family plan which is helping them to reduce the cost of total usage hence not deciding to leave the service.

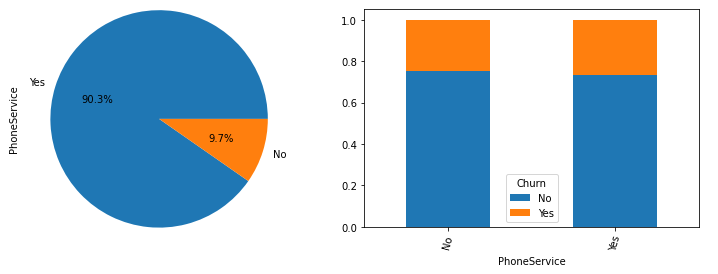
**Dependents**





30% of the users have dependents and they end to stay with the services. This again could be the good family the telecom is providing.

**PhoneServices**



90.30% of the customer uses phone services and it has no impact on the customer’s churn.

**InternetServices**

**Chart, bar chart

Description automatically generated**

21.7% of the customer do not use the internet services and the churn’s is very less among them. Among internet users, churn’s is high for the customer who uses Fibre optics.

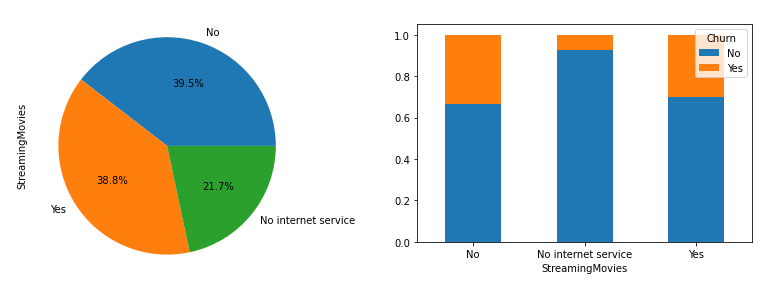
**StreamingTV**

Chart, bar chart

Description automatically generated

38.4% customer uses the stream TV services and they tend to stay with the telecom services when we compare it with customers who do not use this service.

**StreamingMovies**



Like StreamingTV even for customer using StreamingMovies there is less chance of them leaving the telecom services.

**Contract**

**Chart

Description automatically generated**

55% customer have month to month customer and they have greater chances of switching the telecom services. Customers with longer contract are not too keen to leave the services.

**4. EDA concluding remarks:**

1.The company needs to introduce more plans suiting the other sets of people other than senior citizen.More flexible plans are available for senior customers alone.

2. We have high number for churn from the customer who uses Fibre optics as internet services.

3. Company could focus on introducing more discounts on the long term plans as churn is less for the customer who are on 2 year contract.

1. **Building Machine Learning Models.**

Let’s check for the outliers.

Graphical user interface, text, application

Description automatically generated

We could see that the skewness is not present in the dataset.

We need to change the categorical data to the numerical data ,so we will proceed with the Label encoding method.

Text

Description automatically generated

Having an imbalanced data set for a model will not help in generate higher accuracy models, higher balanced accuracy and balanced detection rate.Using SMOTE we will balance the dataset.

Before that we will split the data, as X and Y.

Text

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with low confidence

Thus, we have 7043 rows and 19 columns in the X and 7043 rows and 1 column i.e Target column as Y.

Importing the SMOTE library for over sampling the data.Text

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

We can see that the class is now balanced.

Now we will scale the values using the Standard Scalar method.Importing the library and implementing the Standard Scalar method.

Text

Description automatically generated with medium confidence

Now the data is ready to pass it for the Machine Learning.

We will import all the necessary libraries for Machine Learning.

Text

Description automatically generated

We will check the best possible random state to train our model. Using LogisticRegression to check and then later will run with other models

Graphical user interface, text, application

Description automatically generated

We can see that the best random state is maxRS. We will split the data with this random state and train it.

Graphical user interface, text, application

Description automatically generated

From the output we can see that the RFC has performed well with test accuracy of 86% .

**Cross Validation**

We will check the Underfitting or Overfitting of the model using Cross Validation.

Text

Description automatically generated

To finalise the model we will use the ROC\_AUC score.

Text, letter

Description automatically generated

We get the o/p as below.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

On summarizing the performance ,we get the following result.

Graphical user interface, application

Description automatically generated

From the above metrics we can observe that:

1. DTC and KNN has performed well with least difference on CV score.
2. RFC has given us the best accuracy of 86% Cv score of 83% and ROC AUC score of 86%
3. Adaboost has also given best accuracy of 84% Cv score of 8% and ROC AUC score of 84%
4. Extratree has also given best accuracy of 84% Cv score of 83% and ROC AUC score of 84%
5. DTC has test accuracy of 79% CV score of 78% and ROC AUC score of 79% So we will do Hyper tuning for DTC,RFC,Adaboost and Extratree

**HyperTuning of the model**

Importing GridSearch to hypertune the model



Text

Description automatically generated with medium confidence

Text

Description automatically generated

Chart, line chart

Description automatically generated

The model has not improved.

HyperTuning of DTC

Text

Description automatically generated

Text

Description automatically generated

Chart, line chart

Description automatically generated

DTC has test accuracy incresed by .7 % of 79%, CV score of 78% and ROC AUC score of 80%

HyperTuning of ExtraTree

Text

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Chart, line chart

Description automatically generated

Extratree has test accuracy incresed by .2 % of 84.5% CV score of 82.5% and ROC AUC score of 84.5%. Since, DTC has test accuracy incresed by .7%, we will save DTC as the best model.

Graphical user interface, text, application

Description automatically generated

**Concluding Remarks**

We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we converted features into numeric ones. Afterwards we started training 7 different machine learning models, tuned it’s performance through optimizing it’s hyperparameter values and finally selected the Decision Tree Classifier.

But the concluding part is that the company can make changes in its plans accordingly to retain its customers by filing the gaps that were clearly visible while analysing the data.