“Churn Analysis of Telecom Companies”

Problem Definition:

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.

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

 Data Analysis:

For data analysis first we have to load the data for which we have first import some important libraries.

1. import pandas as pd
2. import numpy as np
3. import seaborn as sns
4. import scipy.stats as s
5. import matplotlib.pyplot as plt
6. #sns.set(style = 'whitegrid')
7. import warnings
8. warnings.filterwarnings('ignore')

Now loading the dataset as csv file.

Data analyzing:

1. Checking the shape of the dataset.
2. Information containing in the dataset. Checking the type of features and target whether it is continuous type or categorical so that we can further change it accordingly. Using .info
3. Checking the number of unique values in the features. Using .nunique
4. Checking the Nulls in the features. Using .isnull
5. Checking the description of the dataset. Checking the interquartile ranges of the features. Using .describe
6. In the Churn analysis I have found that :

* Dataset have 7043 rows and 21 columns.
* There are 18 object type, 2 integer type and 1 float type features and target.
* Features having unique values.
* Senior citizen is categorical column so percentile is not relevant.
* 75% customers having tenure less than 55 months.
* Average monthly charges are 64.76 whereas 25% customer pays more than 89.85.

  EDA Concluding Remarks:

1. Firstly checking the data imbalance of the target variable. Plotting the graph to check this and it is showing highly imbalance of the target variable. The ratio % of churn is 73:27
2. Total charges showing object type which is impossible may be it is containing nulls which are not visible initially so changing it to numeric type using pd.to\_numeric.
3. Now it is showing null values which shows to be very less in percentage so dropping the nulls using .dropna
4. Now EDA the categorical features plotting the graph to analyze the weightage of the unique values.

Conclusion:

* 50.5% are males and 49.5% are females.
* Most of are young generation senior citizens are very less.
* 51.7% having no partner.
* 70.2% having no dependents.
* Large no. of people having phone service.
* 48.1% not having multiple lines, 9.7% not having no phone service.
* Most of the people having fiber optics internet service followed by dsl and 21.6% having no internet service.
* Most people not having any online security.
* Most people not having online backup.
* Most people not having device protection, 34.4% having device protection.
* approx.50% not having techsupport.
* Streaming TV having approx. same yes and no.
* Same as streaming TV streaming movies having same yes and no. those who are streaming TV can stream movies as well.
* Around 55% people contract month to month.
* 59 % people do paperless billing.
* Most people do electronic check payment followed by mailed check and bank transfer.

1. EDA in relation with target variable:

* In gender males are slightly more than females so it is not much important feature to decide churn.
* Senior citizen feature showing the young generation seems to be more likely to churn.
* EDA shows those who are having partner seem to be more likely to churn.
* Those who are having dependents seem to have less churn count.
* Those having phone service seems to have more churn count.
* Multiple lines don’t seem to be good feature to decide churn of the customer.
* Fiber optics customer seems to churn more in respect with DSL.
* The customers those who have not taken internet security seem to churn more in comparison with those who have taken internet security.
* The customers those who have not taken online backup seem to churn more in comparison with those who have taken online backup.
* The customers those who have not taken device protection seem to churn more in comparison with those who have taken device protection.
* The customers those who have not taken tech support seem to churn more in comparison with those who have taken tech support.
* The customers those who have not taken streaming TV seem to churn more in comparison with those who have taken streaming TV.
* The customers those who have not taken streaming movies seem to churn more in comparison with those who have taken streaming movies.
* The customers those who are having month to month contract seem to churn more in comparison with year to year and two year contract.
* The customers having paperless billing seem to churn more in comparison with not having paperless billing.
* The customers those who are having electronic check payment method seem to churn more.

Next is checking the relation between churn with continuous features ie. Total charges and monthly charges.

* Churn is high when monthly charges are high.
* High churn at low total charges. Which is surprising?
* High churn seen in case of month to month contracts. No online security. No tech support. First year of subscription and Fiber Optics internet.
* Low churn is seen in case of long term contracts, Subscriptions without Internet service and the customer engaged for 5+ years.
* Factors like Gender, Availability of phone service and multiple lines have almost No impact on churn.

Now checking the correlation between features:

* **High correlation seen between tenure and Total charges.**

Now next is seeing skewness in the continuous feature.

* Not so much of skewness continuous feature is evenly distributed which shows no outliers in the features.

 Pre-processing Pipeline:

* Firstly, encoding all the categorical features to continuous type. Using get dummies method.
* Encoding the target variable from yes and no to 0 and 1 using replace method.
* Now sampling the target variable using under sampling method to balance the data.
* And now dropping the unnecessary columns which are of less importance in the machine learning models. For eg. Customer id
* And now lastly scaling the data using standard scalar.
* Dividing the feature and target into x and y.

Building Machine Learning Models:

Target variable is classification type so building the classifier models.

To build the models have to first import the necessary libraries.

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import classification\_report

from sklearn.metrics import roc\_auc\_score,mean\_absolute\_error,mean\_squared\_error,accuracy\_score,confusion\_matrix,roc\_curve

from sklearn.tree import DecisionTreeClassifier

1. Splitting the data into train and test data.
2. Defining the function.
3. Training the data and fit into

* Knn model
* Random forest model
* Decision tree model

1. Checking the accuracy score of train and test data.
2. Now checking if the models are under fitting or over fitting by calculating cross validation score.
3. Now lastly using hyper parameter tuning to increase the model performance.
4. Now checking the auc roc curve. And finding the model which has more auc and roc score and accuracy score and less errors and saving the more accuracy score model.

Concluding Remarks:

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

* If you examine closely, it's clear that there is a significant rate of churn among those who have Internet, streaming TV, music, and movie bundles. The business can improve those bundles. With regards to Online Security and Premium Tech Support services, the organization appears to be doing an excellent job.
* Among all the models Random Forest Classifier performing good with accuracy of 80% and area under the curve is also 69% which is pretty good score.