Customer Churn Analysis

Introduction:

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.

Problem statement:

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

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

Here step by step outline for the project

1. Data Analysis.
2. EDA Concluding Remark.
3. Pre-Processing Pipeline.
4. Building Machine Learning Models.
5. Concluding Remarks.

Data analysis:

First we download all the necessary libraries .Then we import the dataset .

Dataset includes various features with the help of them we will do analysis.

Firstly we will chek various attributes like shape,which columns are present in dataset,dtypes of columns present in dataset, info of columns ,unique values present in colunmns.The target variable 'Churn' is not numerical so it's a classification type problem .

Then we will go for the analysis of target columns . we chek its dtypes,unique values present in column. Then we chek Description of Data set.This gives statistical information of the dataset.There is no any negative or invalid values in dataset. so from we can conclude There is no any null value since the count of all columns is same. So then we further move to assure that there is no any kind of null values present in the dataset by isnull method .We found 11 null values in total charges column.It is to less data as compared to whole dataset so we removed it.Then we drop the columns which are not required for the anayalysis. Then we converted the target variable in to the numeric form.So that we can easily do further analysis.

Then we will chek the relationship between monthly charges and total charges with the help of lmplot and we can conclude that there is high churn is seen where the monthly charges are high and total charges are low .

Then we checked the correlation between the variables and visualize it with the help of bar plot and we can conclude that ,

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, Subscription without internet service and the customer engaged for five plus years.

Factors like Gender, Availability of phone service Have almost no impact on Churn.

Then we further move to bivariate analysis . With the help of uniplot function we can chek the distribution for various attributes .From that we can conclude that

For the important attribute like payment method we can see the electronic chek are the highest churners . Most female users using credit card are more churners.

Then for contracs monthly contracs are more churnres.

EDA Concluding Remark :

From the EDA part we can conclude that,

Electronic chek medium are highest churners.

Monthly customers are more likely to churn because of no contract terms, as they are free to cutomers.

No online security,No tech support category are high churners.

Non senior citizens are high churners.

Building Machine Learning Models:

We import the necessary libraries first from sklearn.

Then we separate the dependent ad independent variable as x and y .Then we convert the data in to training and testing data.

Then we use Decision tree classifier .we will fit the data in it . then we create a prediction variable y_pred , we compare actual y value and predicted y value **.** As our dataset is imbalnced so first we will do SMOTEENN analysis on data, then chek the classification report and confusion matrix for resampled data.

And then we got good result.Simmilerly we chek the different classifiers.Like RandomForestClassifier,KNClassifier,ADAboostclassifier etc.

Then we save the model,to use it again.We build a model to see that the at which condtion the customer churned.