Customer Churn Analysis

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 prioritize 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.

Data Analysis :

From the above Problem statement we can come to know that Problem is explaining about the categorical data outcome. Categorical data can be classified into 2 parts either Yes or No, 1 or 0, True or False or anything which have 2 equal and opposite sides. Initially we need to understand the Problem Statement and then find the Label i.e nothing but the Output from the desired Problem statement.

As we know that Customer churn is the issues may Telecom companies are facing now a days. Our role comes into picture to find and reduce the risk of Churn. Due to which Customer can be retained such that business can be improved to greater extent. We Data Scientist/ Analyst have the major roles into finding the proper set of data and providing the Customer Best Feature from the Analysis

Now initially we have to scrap the data from the dataset available or will be shared by the customer. If data is not available with the customer then we have the choice with our skills and experience we can check the history of the Customer and guide for the feature for next 6 months to 1 year and gather the data.

Most of the cases the data is available in the form of DataFrame and we need to Load the data and analyse the data with the help of Algorithms.

After fetching the data the next step is to check and verify the data health. Health refers data availability. The more data availability chances of accuracy is more for the prediction

If the data insufficiency is there in the dataset then we need to clean the data with the help of some techniques. The more we clean the data the more we can make the analysis perfect and make the accurate.Cleaning method include having the mean, median or mode of the data or we have the correct data physically we can feed into the Data-frame.

After cleaning the data we need to move on to the next step i.e checking if the data is in the integer format because our algorithm do not understand the language of object or text. We need to convert the object into the integers and then perform the Algorithm on the dataset. Always make sure that using Encoder is the best way for data conversion else we can convert each unique object into integers.

As we have completed the cleaning Part further we will move into the next step i.e understand the Problem again and choose the appropriate algorithm for the set of data to be perform.

EDA Concluding Remark :

Now we need to find if the Problem is Categorical or Continuous.As we know that Output we want is Yes or No then this problem is of categorical type and we need to use the categorical Algorithms.

We need to conclude with the suitable algorithm dataset.

Preprocessing the data

We have multiple Algorithm for the Categorical dataset but I have used KNN Algorithm which I prefer in Categorical dataset because it finds the best parameters required for the desired output. The desired output is the Customer Churn. Now we have to apply import the necessary libraries for the Algorithm and then initially find the best parameters for the Algorithm using SelectKBest Classifier. The dataset is divided into the two variables one in which entire columns are available except Output (Label) and other variable we need to feed output Column (Label).

Now we have to find the best parameters from the divided dataset and after that from the available best parameters we need to train the model

Building Machine Learning Models :

Now we need to train the data from the above best parameters, we can take 75% of the data for testing. Above best parameters will train the data with all the possible ways

Now we need to test the data with the available 25% of the data. Testing data helps to find the accuracy in our model.Before the accuracy we need to find the Confusion matrix for the data such that accuracy, precision, etc can be found out easily

After the development of Algorithm we have to find the accuracy of the model.

Concluding Remarks :

We have to Cross validate the data with K fold validation method and find Multi-collinearity. Multi-collinearity is the term to find if model is repeating the same answer from the train data.

For the best Algorithm output Multi-collinearity should not exist in the Model

To find the multi-collinearity we have various techniques VIF, Heatmap, etc.

Now with the help of Hyper parameter tuning we have to find our Best parameter which suits the Algorithm and increases the Accuracy.

Finally with the help of classification report we can find the accuracy, precision, F1 Score

After we have satisfied with the Algorithm Accuracy we need to save the model in the pickle format and we can use the Algorithm for the parameters and then we can hand over to the Customer for the Production Trail purpose and in the Production Trial phase we need to find that the shortcomings in the model and then we need to overcome the shortcomings.

By - KAUSTUBH KUNKAVLEKAR