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 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.

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

Problem Definition-Retention of customer is the biggest problem as customer switch their priority very frequently

And it is very important to know that what are those factors which makes the customer to switch fastly.

Data Analysis- This data set contains 5 rows and 21 columns. Most of the features of the dataset are of Categorical nature. The list of Categorical data- **customerID, gender, SeniorCitizen, Partner, Dependents, PhoneService,**

**MultipleLines, InternetService, OnlineSecurity, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, Churn and rest of the features are in numerical form.**

**In the dataset there is no null value.**

**As per dataset customerID is not of much use so it will be dropped from the dataset.**

**Gender-Female will be replaced with 0 and male will be replaced with 1.**

**Partner- yes will be replaced with 1 and No will be replaced with 0.**

**Dependents- yes will be replaced with 1 and No will be replaced with 0.**

**PhoneService- yes will be replaced with 1 and No will be replaced with 0.**

**MultipleLines- No phone service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**InternetService- No will be replaced with 0, DSL will be replaced with 1 and Fiber optic with 2.**

**OnlineSecurity- No internet service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**DeviceProtection- No internet service will b e replaced with 0, No will be replaced with 1 and yes with 2.**

**TechSupport- No internet service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**StreamingTV- No internet service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**StreamingMovies- No internet service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**Contract-** **Month-to-month will be replaced with 0, One year will be replaced with 1 and Two year with 2**

**PaperlessBilling- yes will be replaced with 1 and No will be replaced with 0.**

**OnlineBackup- No internet service will be replaced with 0, No will be replaced with 1 and yes with 2.**

**PaymentMethod-in dataset it is of no use so it will be dropped.**

**Churn- as it is a output or label or target column in dataset, label encoder will be applied on this.**

**TotalCharges-As it is of object type it should first convert into numerical by converting into float type.**

**It also seen that TotalCharges is rightly skewed and it can be scaled uniformly by using Standard Scaler.**

**Outliers-As there is outlier in SeniorCitizen and PhoneService but it can be ignored and it will not affect the data.**

**Training of data-Now by splitting all features data into training and testing set by train\_test\_split**

**Building Machine Learning Models-As it is a classification problem therefore, we will build try to build all the algorithm related to classification problem.**

**One by one will build all the algorithm and then first will fit the training data and put the testing data for prediction.**

**In this way Various algorithm will be build.**

**Metrics evaluation-By using various evaluation metrics will try to find the best Machine Learning Model like Accuracy Score etc**

**Cross validation technique-allows us to compare different machine learning method and get a sense how well they will work in practice. Now after finding out the cross validation score, we will take out the difference between accuracy score and cross validation score , the machine learning model with greater score will be consider as an best Machine learning algorithm.**

**Now, we can use Hyperparameter Tunning to Test the data by using GridSearchCV.**

**At last we will use ROC AUC CURVE to check the how much model is able to distinguish between the model.**

**Conclusion-In this classification problem Logistic Regression algorithm is the best machine learning model which gives near about 80% accuracy.**