**Customer Churn Analysis**

Hello everyone, we are here to learn about Customer Churn Analysis in telecommunication industries, let's start with its introduction.

**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 priorities focused marketing efforts on that subset of their customer base.

**Problem Statement:**

Our problem statement is preventing customer churn which is mostly important to the telecommunications sector, as the barriers to entry for switching services are so low.

Therefore, we have to make a model to predict that which type of customers are likely to churn from telecommunication sector and which are not.

**Data Overview:**

First, we loaded our data on the platform where we have to do our work, here we are going with jupyter notebook and after loading the data we came know that.

We have Dataset which has 7043 rows (customers) and 21 columns (features).

The features are: - customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn.

**Data Analysis:**

Now, we are going to start our data analysis which is the first step to analyze the data.

* First of all, dataset has no null values so it's quite ok for further analysis.
* In this dataset all the columns are object type except 3 columns in which two columns are int type and one column is float type. everything is ok but TotalCharges which is numerical column and it is given as object type, so we have to convert it into float or int type.
* But before converting it we analyze that TotalCharges contain some space due to which it is given as object type so we have to replace it with null values.
* We replaced it with null values and also filled all the null values with its appropriate values. Now, data is ready for further analysis.
* Now, we visualize the data so that we can know about insights of data. And according to that we pre-process the data.

**Data Exploratory Analysis:**

We did some statistical analysis and visualization which are given as.

First, we did univariate analysis : -

After doing **Univariate Analysis** we came to know that some facts i.e.,

* Male customers are little more than the female customer. Without partner customers are more than the with partner customer. Independent customers are more than dependent customer. Customers are more who had phone service. Maximum customers are who do not have multiple lines. Maximum customers took Internet service which is distributed through fiber optic.
* Maximum customers do not take any online security. Maximum customers do not have any online backup. Maximum customer took contract for month to month rather than one year and two years. Maximum customers do paperless billing. Electronic check is the most preferred by the customers. Most of the customers not willing to churn. Lastly our target column (Churn) is imbalanced, we have to balance it for better prediction otherwise it will be biased with that category which has maximum data.

Now, it’s time to do some **Bivariate Analysis** so that we came to know relationship about independent features and with dependent column also.

* We can see that there is no difference between male and female who are willing to churn having high monthly charges. there is not much difference that someone is senior citizen or not but if they are felt monthly charges are high, they are willing to churn.
* Same case for partner and dependent also they are likely to churn if they felt monthly charges is too high.
* It seems whatever will be the tenure churning of customer always happen because of high charges.
* Who have multiple lines are likely to churn less as compare to who do not have, because monthly charges for single line would be costly.
* Owning fiber optic one is less likely to churn as compare to DSL. contract with two years is more prone to churn however their monthly charges are less than other contracts.
* Bank transfer and credit card who use these methods to pay are likely to churn more however their monthly charges are less as compare to electronic check.

We did some statistical analysis and we came to know that,

* Most of the columns have less difference between mean and standard deviation, it means data is not much skewed and likely to be a normally distributed. And it is good for our model because if data will be in standard form, it will be easy to process by the model so prediction will be better as compare to if data is not in standard form.

Now, we did **Multivariate Analysis** and analyzes the correlation between features and dependent column and among features also (i.e., about multicollinearity).

* We use heatmap for analyze the correlation because it is very easy to understand from heatmap. And we came to know that,
* Partner and Dependents columns are lightly correlated to each other.
* Contract and Tenure are quite good correlated to each other.
* Total charges and Tenure are highly correlated to each other.
* Multiplelines, MonthlyCharges and TotalCharges are also quite correlated to each others.
* TotalCharges and MonthlyCharges are good correlated to each other.

After all these analyses of data we are going for data preprocessing which is very important to do, because model accept data in some standard format it can not process the data as it was collected.

**Preprocessing:**

* First, we dropped column customerId because it will not be useful for prediction. Now we have 7043 rows and 20 columns.
* As we already filled all the null values so we are going to do encoding of categorical columns, we did ordinal encoding for those columns which are not in any order and label encoder for those columns which have some order.
* We checked outliers and skewness, and there are only two column SeniorCitizen and PhoneService which have outliers but not much. And we are not going to remove outliers since these columns are categorical type.
* If we talk about skewness, there are few columns which are skewed such as SeniorCitizen, Dependents, PhoneService, Contract, TotalCharges, Churn. But these are either categorical columns or target column except TotalCharges.
* So, we are not going to remove skewness from categorical and target columns. But we tried to remove skewness from TotalCharges.
* After removing the skewness TotalCharges becomes very less with its respective row of MonthlyCharges .
* Logically, it can not possible that Monthlycharges are greater than TotalCharges, so we should not remove the skewness.

After all these pre-processing we are ready for splitting the data into training and testing part.

After splitting the data, we have to take care of its imbalance nature, for that we used SMOTE (Synthetic Minority Oversampling Technique) which is used to oversample the minority class at the level of majority class.

Now, dependent column balanced and after this we scale the data so that all the features came onto the same scale otherwise our model will be biased towards high range of feature column.

We used MinMaxScaler to scale our feature column.

**Building Machine Learning Model:**

We used several models for training our data such as Logistic Regression, Support Vector Machine, Decision Tree Classifier, KNeighbors Classifier, Random Forest Classifier, AdaBoost Classifier.

**Logistic Regression:**

Logistic regression is the statistical method to analyze a dataset in there are one or more independent variables that determines an outcome. This is the measure of only two possible outcomes.

**Support Vector Machine:**

Support vector machine is a supervised machine learning algorithm which can be used for both classification and regression problems. For classification we use support vector classifier and for regression we use support vector regressor.

It separates the classes with the help of maximum margin hyperplane.

**Decision Tree Classifier:**

Decision Tree Classifier is a supervised machine learning algorithm which can be used for both classification and regression problems. It splits the population to make a tree and it uses the criterion Gini or Entropy to know the impurity level on the basis of impurity it decided the root node and its child node.

**Random Forest Classifier:**

Random forest is the ensemble technique for both classification and regression.

It makes multiple decision trees at the training time and provides input as divided samples from dataset which is known as bagging technique. Output is the mode of classes from all the decision trees or mean of all output on the basis of problem whether it is classification or regression.

**KNeighbors Classifier:**

The K in the name of this classifier represents the k nearest neighbors, where k is an integer value. This classifier implements learning based on the voting by nearest k neighbors, the choice of k depends on dataset.

**AdaBoost Classifier:**

AdaBoost algorithm, short for Adaptive Boosting. It is a boosting technique that is used as ensemble method. It works on the principle where learners are grown sequentially, except for first each learner grown from previous grown learners. All means weak learners are converted into strong ones.

* All these models are used for the training purpose and we also introduced the cross-validation score so that we can understand the overfitting or underfitting nature of our model.
* We use metrices like accuracy score, confusion matrix and classification report for the analysis of accuracy of our model prediction.
* On the basis of scores (accuracy score and cross validation score), we choose few models for hyperparameter. we drop decision tree for highest difference between accuracy score and cross validation core as compare to other models. Rest of the model we choosed.
* We also try to train our model without handling class imbalance but accuracy score was quite low that is why we choose to go with after handling class imbalance.

It's time to do hyper parameter tuning of rest of the models, it need to be tuned so that the model can optimally solve the machine learning problems.

* There are different types of technique available to so hyper parameter tuning and here we did hyper parameter tuning using Grid Search CV.
* After doing hyper parameter tuning, we drop logistic regression because of its low accuracy and rest of the model have quite good accuracy.
* So, for choosing best model among all we use one more metric which is ROC AUC curve and its score.
* Auc-roc curve is a performance measurement for the classification problem at various threshold settings.
* Roc is the probability curve and auc represents the degree or measure of separability. Higher the auc, better the model prediction. Roc curve plotted between true positive rate and false positive rate.
* Lastly, we have support vector classifier, kneighbors classifier, random forest classifier and adaboost classifier for roc auc curve and score.

So, on the basis of roc auc score adaboost classifier worked well having better accuracy than other models. Therefore, it will be our final model.

Now, we have saved our model for further predictions, we can use pickle and joblib to save our model, here we are going to use joblib to save the final model.

**Conclusion:**

Churn rate is a health indicator for subscription-based companies. The ability to identified customers that are not satisfy with provided service recognizes businesses to concern about product, pricing plans and its weak nature, operation issues and expectation to proactively reduce reason for churn.

It's important to define data sources and observation period to have a full picture of history of customer interaction. Selection of the most significant features for the model would influence its predictive performance.

Company with a large customer base and numerous offerings would benefit from customer segmentation. The number and choice of machine learning models may also depend on segmentation results. Data scientist also need to monitor deployed models, revise and adapt features to maintain desired level of prediction accuracy.