Churn Prediction in Telecom using ML

By

P Raja Vamsi Krishna

Objective of the Project : Develop a churn Prediction model in telecom to predict customers who are most likely subject to churn

Importance of the study :- Cost of acquiring a new customer is more than the retaining existing customer. Hence predicting churn rate becomes very important and can take appropriate business actions for reducing the churn.

Data Source for the project: Customer's dataset in CSV file containing 7043 rows and 21 columns.

The Process

Basic Processes

1. Importing required libraries like pandas, numpy, matplotlib and seaborn.
2. The CSV file was imported and read into dataframe named as df.

Analyzing and Visualizing the data set

1. Checked the top 5 rows of the dataset.
2. Checked the number of rows and columns of the dataset.
3. Checked the type of data under each column of the dataset.
4. Converted the datatype of the column "Total Charges" from object to float64.
5. Checked for null values, there were 11 null values under the column "Total Charges" and hence removed the rows this null values.
6. Checked the descriptive statistics like count, percentiles, max, min etc for the dataframe.
7. Dropped the column 'customerID' from the dataframe as it is not useful for further analysis.

Outlier Identification using Box Plot

Checked for outliers for the columns 'SeniorCitizen','tenure','Monthlycharges','TotalCharges' by using boxplot. No outliers found.

Data Processing

1. Seperating the dataframe df into independent variables(x) and dependent or target variable(y)
2. .Label encoding was done to convert the output variable into 0 and 1.
3. One Hot Encoding was done to convert all categorical columns(columns with object datatype) into numbers(0's and 1's)

Data Split

1. Data Split was done to split the data first into Training and Testing Data in a 75/25 ratio.

Feature Scaling

1. Standard Scaler has been used for feature scaling or normalization of independent variables(x : x_train & x_test).

Model Building using different ML algorithms

1. Logistic Regression, Decision Tree, Random Forest and SVC model were built on train data and prediction was done on test data.

Model Evaluation

1. Evaluation of all the models were done by using Confusion Matrix and accuracy_score, recall_score, precision_score, f1_score, rc_auc_score.
2. Logistic Regression, Decision Tree, Random Forest and SVM Classifiers returned accuracy scores of 79%, 79%, 78% and 80% respectively.
3. Confusion matrix was drawn using Seaborn library for each model.
4. For Decision Tree, Criterion taken was 'Entropyi' and Maximum Depth taken was 3.
5. For Random Forest , n_estimators was 15 with Criterion as 'Entropy'.

SVC – Grid Search – Cross Validation

1. For SVC, model parameters and grid parameters were obtained.
2. Cross Validation was done by importing GridSearchCV.
3. The grid was fitted into the training dataset.
4. Inspected the best parameters using GridSearchCV in the best_params attribute and best estimator in the best_estimator_attribute
5. The predictions were re-run and confusion matrix was re-obtained using Y_test and grid_predictions.
6. The re-estimation returned an accuracy score of 81%.

Conclusion

All the ML models gave good prediction of about 80% accuracy on churn.