**Article / Blog on Customer Churn Analysis**

1. ***Problem Definition:*** Customer churn is when a company’s customers stop doing business with that company. Keeping an existing customer is far less expensive than acquiring a new customer. Customer retention can be achieved with good customer service and products. In, Telecommunications sector preventing customer churn is critically important to the barriers to entry for switching services are so low. But the most effective way for a company to prevent attrition of customers is to truly know them.
2. ***Data Analysis:*** We have examined customer data from IBM Sample Data with 7043 nos. of Telecommunication sector. In this Dataset, 7043 and 21 columns, from which 17 columns are categorical and 04 columns are continuous and this dataset is using 1.1 MB memory.
   * 1. This Dataset is separate in Male and Female users who are using Telecom Services like Phone Service, Internet Service, Online Security and Backup, Streaming TV & Movies, Device Protection and Online Technical Support from the company.
     2. Out of Total no. of users approx. 22% is not using any Internet Service so they can not avail any online services which are provided by the company.
     3. Approx. 45 % of customers having one year or two-year contract with company it is showing that these are liable with this Telecom Company and they are regular customer.
     4. Average Monthly Charges of all the customer is approx. 65 (Currency) and we need to predict the Total Charges paid by the customers.
3. ***EDA Concluding Remarks:***
4. As per given data set, 73% customer are happy with the company service and 27% willing to churn. So, we can say churn rate is good.
5. Total Average Charges paid by the customers is 2283. Minimum Charges is 18.80 and maximum paid by the customer 8684.80
6. ***Pre-processing Pipe Line***:
7. As maximum columns are categorical so we have used Encoder for convert the same in continues data type column.
8. After Encoding the data, checked the features which are correlate with our target variable. We have found Monthly Charges and Tenure have positive correlation with Target Variable.
9. Also, other Categorical Columns have positive correlation with Target Variable.
10. We have checked the outliers with Box Plot and found only in Senior Citizen Column which is ordinal data and it will not affect our target variable. However, we deal with outliers by Z-Score Method considering threshold as 3 with data loss approx. 9% which is acceptable.
11. Now Data set shape is 6361 rows and 21 columns.
12. ***Building Machine Learning Models:***
13. Our Target Variable *Total Charges* is continuous data type so we will use Regression Models to predict the label.
14. For predict the label, we have separated the features and label in x, y variable and uses Train Test Split for Training and testing the data. Finally, we will compare the Predicted Data with y\_test data to check the model parameters.
15. We have scaled the feature data (x) and found the random state at which Training and Testing Accuracy is High.
16. We have fit the random state and again split the dataset with Train Test Split.
17. We have fit the training data into our regression model and prediction with x\_test to compare with actual.
18. Calculated the R2 Score with y\_test and perdition score (fitted on Training data) and calculate cross validation score with different type of regression algorithm like lasso regression, Linear regression, Random Forest Regressor.
19. After hyper parameter tuning, we have found that Random Forest Regressor is giving Highest R2 Score is 99.8696569 and CV Score 99.8737723.
20. ***Concluding Remarks:***
21. Difference between R2 Score and CV Score should be minimum for maximum accuracy of model.
22. In our model diff is 0.0041154, which shows we have built a very good model with Random Forest Regressor to predict the Total Charges paid by the customer to Telecom Company.
23. We have plotted a scatter plot between Actual Charges Vs Predicted Charges, found Predicted Charges is very close to Actual Charges which means we have made a very good model with 99.87 accuracy.

***Author Name: Himanshu Sharma***

***Mo. No. 9873553391***