Telecom Churn Prediction:

Main Objective: To predict the customer churn in upcoming month or not.

Data Introduction: So basically the data is of telecom sector which consists fields like mobile number, circle id, monthly call usages, incoming and outgoing calls, recharge counts and amounts, mobile internet recharge details etc.

Data Understanding: So the dataset consists of user call usages for 4 months i.e. July-Sept.  It  has features like month wise incoming calls, outgoing calls, STD calls, roaming recharges, Last recharge date, recharge amount and many more details.

There were around 1L rows and 226 columns.

Now we need to identify the target variable.

For the same we took the last month (i.e. September) data as our target variable.  So for the same we considered summation of 4 cols which is Total\_ic\_mou\_9, Total\_og\_mou\_9, vol\_2g\_mb\_9, vol\_3g\_mb\_9. Wherever the sum = 0 that are considered as churned as customers and rest are considered as non-churned customers.

So we observed that churn rate was around 10.19% and rest were non churned customers which clearly shows Class imbalance problem.

Data segregation:

To simplify data and better understanding we segregated the data monthly and then sub categorized them into call usage, recharge columns, incoming and outgoing related cols.

We did it using a 2 functions where in first function we searched for the character  “6” in column name and append that column to June  list and same for other months too.. Then using second function we further bifurcated into sub categories.

Handling Missing Values:

While data bifurcation we observed that there were 214 numeric columns and 12 non numeric columns. From which we excluded sept month data as it was target variable. And removed more columns like mobile number and circle id etc. on the basis of unique count. [ If they ask how then : mobile number was having unique count of 99999 and rest cols like circle\_id, std\_og\_t2o\_mou, etc. had unique count as 1 which when observed was having value has 0 only]

So for the numeric columns where data was not available we imputed that columns with 0 [ why : because there was no dependency with other columns as we don’t have geographical features]

And for cols like last recharge date we didn’t impute missing values because it shows the last recharge information of consumer.

EDA:

1. Age on Network - We observed that people who have subscribe from 1-3 years are more likely to churn then rest.
2. We observed that there were outliers in many of major columns  like Incoming call usage, outgoing call usage, Recharge amount, 2G and 3G data pack and also the outliers where in the top percentile

Outlier Treatment:

Now as we observed that most of the outliers are covered on the top most percentile we have capped them at 99th percentile and 1st percentile to retain most of our data.

Feature Engineering:

We have created new features like:

1. Considering the last recharge date from 3 months
2. Taking mean of 3 months for all numeric columns

Preparing data for Model:

1. Classified categorical and numerical data
2. For categorical : data[night\_pck\_user\_6','monthly\_2g\_6','sachet\_2g\_6','monthly\_3g\_6','sachet\_3g\_6','fb\_user\_6',] we created dummy encoding and dropped original features
3. For Numerical :
   1. We did log transformation and then standardized[re-scaled] the numerical features for getting our mean as 0 and STD as 1

1. After which we took care of class imbalance problem by applying SMOTE/ADAYSN
   1. After which the count of churn increased

So after data prep we applied split our train and test data into 70-30 ratio.

We applied PCA for feature selection and dimensionality reduction.[why pca ]

After that we used X\_PCA\_TRAIN as our X variable and Churn as our Y variable and applied 2-3 models like KNN, logistic regression, Random forest and XGboost.

From above algorithms KNN showed best performance.

CONCLUSION-

The importance of this type of research in the telecom market is to help companies

make more profit. It has become known that predicting churn is one of the most

important sources of income to telecom companies. Hence, this research aimed to

build a system that predicts the churn of customers in telecom company. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 70% for training and 30% for testing. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. In addition, we encountered another problem: the data was not balanced. This problem was solved by Oversampling or using trees algorithms not affected by this problem. These algorithms are Logistic Regression , Random Forest, KNN classifier , and XGBOOST algorithm.