Churn Prediction Assignment

Business problem overview:

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.

For many incumbent operators, *retaining high profitable customers is the number one business goal*.

To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

Description of the dataset:

The dataset is quite huge dataset with 99999 rows and 226 columns among which ID Columns are 2 , Date Columns are 12, Numeric columns are 204 and categorical columns are 8. It also has recharge details of last 4 months. If we assume that this month is 10^{th} month then the recharge details for the month 6,7,8 and 9 are given also.

Feature Engineering:

For Null Imputation, we dropped all those columns which have more than 40% missing values.

The columns which have only single unique values have been dropped. All date columns were dropped.

Now, Total Recharge Number columns, Average Recharge Number, Average Recharge Amount, Maximum Recharge Amount were used to derive meaningful insights, e.g: Total Data Recharge Amount was calculated by multiplying total recharge data and total recharge amount. After that, we took the 70^{th} percentile of average amount of data recharge for 6^{th} and 7^{th} months. The whole data has been filtered on this number which helped to reduce the data volume and this new volume of data has 30001 rows and 185 columns.

Then we created a Churn variable where the customers who neither have used calls nor used internet in September are marked as 1 i.e, churned customers and rest are marked as 0. This will be our target variable having 91.86% True (1) and 8.13% False (0) which shows a high imbalance. Then we removed all the columns which are corresponding to churn and the columns which are having below 5% null values are dropped again. Final Data volume became rows with 28504 and columns with 140.

Exploratory Data Analysis:

We performed some univariate and bivariate analysis where we found that some variables have right skewness and there is presence of outliers. We used K-Sigma technique for imputation of outliers.

From the bi-variate analysis we understood that there is a drop in the total recharge amount for churned customers in the 8th Month. We also observed that there is a huge drop in maximum recharge amount from data in the 8th Month for churned customers. We also found that there is a huge drop in the total recharge number in the 8th Month for churned customers.

Model Building:

As the number of variables is too high, so there is a possibility of curse of dimensionality. So we performed PCA (Principal Component Analysis) to understand how many variables are actually explaining a threshold amount of variability – here we chose the threshold as 90% and got to know that 60 variables are enough to explain 90% variance of the data.

After applying PCA followed by Logistic Regression, we got Sensitivity, Specificity and AUC-ROC score as 81%, 81% and 87% respectively.

Post this we applied Random Forest Classification with 5 fold cross validation and performed grid search where AUC score we got as 91% but Sensitivity was too low (28%) and specificity as 99%. This means out of all true or actual positives (i.e, actual churned customers) it can classify only 28% actual classes(TP/TP+FN), whereas 99% specificity indicated=s that, out of all positive(True Positive and False Positive) it's able to identify 99% positive cases(TP/TP+FP)

Then we tried to understand the variable importance and after understanding that, we took total 30 top features and based on those top features we reran Logistic Regression, but this time sensitivity fell down a bit but rest metrices showed similar score as before- sensitivity came as 69%, specificity came as 90% and ROC came as 88%. This model we can take as an optimum model.

Future Step:

We can do more hyperparameter tuning and also apply boosting algorithms to get better result.

Business Strategies:

- Customers with less than 4 years of tenure are more likely to churn and company should concentrate more on that segment by rolling out new schemes to that group.
- Telecom company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone.
- Incoming and Outgoing Calls on roaming for 8th month are strong indicators of churn behaviour
- The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of STD and ISD packages.
- To look into the issues stated above, it is desired that the telecom company collects customer query and complaint data and work on their services according to the needs of customers.