Analysis Steps

Data Cleaning and EDA

1. We have started with importing Necessary packages and libraries.
2. We have loaded the dataset into a dataframe.
3. We have checked the number of columns, their data types, Null count and unique value_value_count to get some understanding about data and to check if the columns are under correct data-type.
4. Checking for duplicate records (rows) in the data. There were no duplicates.
5. Since 'mobile_number' is the unique identifier available, we have made it our index to retain the identity.
6. Have found some columns that donot follow the naming standard, we have renamed those columns to make sure all the variables follow the same naming convention.
7. Follwing with column renaming, we have dealt with converting the columns into their respective data types. Here, we have evaluated all the columns which are having less than or equal to 29 unique values as catrgorical columns and rest as contineous columns.
8. The date columns were having 'object' as their data type, we have converted to the proper datetime format.
9. Since, our analysis is focused on the HVC(High value customers), we have filtered for high value customers to carryout the further analysis. The metric of this filtering of HVC is such that all the customers whose 'Average_rech_amt' of months 6 and 7 greater than or equal to 70th percentile of the 'Average_rech_amt' are considered as High Value Customers.
10. Checked for missing values.
11. Dropped all the columns with missing values greater than 50%.
12. We have been given 4 months data. Since each months revenue and usage data is not related to other, we did month-wise drill down on missing values.
13. Some columns had similar range of missing values. So, we have looked at their related columns and checked if these might be imputed with zero.
14. We have found that 'last_date_of_the_month' had some misisng values, so this is very meaningful and we have imputed the last date based on the month.
15. We have found some columns with only one unique value, so it is of no use for the analysis, hence we have dropped those columns.
16. Once after checking all the data preparation tasks, tagged the Churn variable(which is our target variable).
17. After imputing, we have dropped churn phase columns (Columns belonging to month - 9).
18. After all the above processing, we have retained 30,011 rows and 126 columns.
19. Exploratory Data Analysis

- The telecom company has many users with negative average revenues in both phases. These users are likely to churn.
- Most customers prefer the plans of '0' category.
- The customers with lesser 'aon' are more likely to Churn when compared to the Customers with higer 'aon'.
- Revenue generated by the Customers who are about to churn is very unstable.
- The Customers whose arpu decreases in 7th month are more likely to churn when compared to ones with increase in arpu.

- The Customers with high total_og_mou in 6th month and lower total_og_mou in 7th month are more likely to churn compared to the rest.
- The Customers with decrease in rate of total_ic_mou in 7th month are more likely to churn, compared to the rest.
- Customers with stable usage of 2g volume throughout 6 and 7 months are less likely to churn.
- Customers with fall in usage of 2g volume in 7th month are more likely to Churn.
- Customers with stable usage of 3g volume throughout 6 and 7 months are less likely to churn.
- Customers with fall in consumption of 3g volume in 7th month are more likely to Churn.
- The customers with lower total_og_mou in 6th and 8th months are more likely to Churn compared to the ones with higher total_og_mou.
- The customers with lesser total_og_mou_8 and aon are more likely to churn compared to the one with higher total_og_mou_8 and aon.
- The customers with less total_ic_mou_8 are more likely to churn irrespective of aon.
- The customers with total_ic_mou_8 > 2000 are very less likely to churn.

1. Correlation analysis has been performed.
2. We have created the derived variables and then removed the variables that were used to derive new ones.
3. Outlier treatment has been performed. We have looked at the quantiles to understand the spread of Data.
4. We have capped the upper outliers to 99th percentile.
5. We have checked categorical variables and contribution of classes in those variables. The classes with less ccontribution are grouped into 'Others'.
6. Dummy Variables were created.

Pre-processing Steps

1. Train-Test Split has been performed.
2. The data has high class-imbalance with the ratio of 0.095 (class 1 : class 0).
3. SMOTE technique has been used to overcome class-imbalance.
4. Predictor columns have been standardized to mean - 0 and standard_deviation- 1.

Modelling

Model 1: Logistic Regression with RFE & Manual Elimination (Interpretable Model)
Most important predictors of Churn , in order of importance and their coefficients are as follows :

- loc_ic_t2f_mou_8 -1.2736
- total_rech_num_8 -1.2033
- total_rech_num_6 0.6053
- monthly_3g_8_0 0.3994
- monthly_2g_8_0 0.3666
- std_ic_t2f_mou_8 -0.3363
- std_og_t2f_mou_8 -0.2474
- const -0.2336
- monthly_3g_7_0 -0.2099
- std_ic_t2f_mou_7 0.1532
- sachet_2g_6_0 -0.1108

- sachet_2g_7_0 -0.0987
- sachet_2g_8_0 0.0488
- sachet_3g_6_0 -0.0399

PCA: PCA : 95% of variance in the train set can be explained by first 16 principal components and 100% of variance is explained by the first 45 principal components.

Model 2 : PCA + Logistic Regression

Train Performance :

Accuracy : 0.627

Sensitivity / True Positive Rate / Recall : 0.918

Specificity / True Negative Rate : 0.599

Precision / Positive Predictive Value : 0.179

F1-score : 0.3

Test Performance :

Accuracy : 0.086

Sensitivity / True Positive Rate / Recall : 1.0

Specificity / True Negative Rate : 0.0

Precision / Positive Predictive Value : 0.086

F1-score : 0.158

Model 3 : PCA + Random Forest Classifier

Train Performance :

Accuracy : 0.882

Sensitivity / True Positive Rate / Recall : 0.816

Specificity / True Negative Rate : 0.888

Precision / Positive Predictive Value : 0.408

F1-score : 0.544

Test Performance :

Accuracy : 0.86

Sensitivity / True Positive Rate / Recall : 0.80

Specificity / True Negative Rate : 0.78

Precision / Positive Predictive Value :0.37

F1-score :0.51

Model 4 : PCA + XGBoost

Train Performance :

Accuracy : 0.873

Sensitivity / True Positive Rate / Recall : 0.887

Specificity / True Negative Rate : 0.872

Precision / Positive Predictive Value : 0.396

F1-score : 0.548

Test Performance :

Accuracy : 0.086

Sensitivity / True Positive Rate / Recall : 1.0

Specificity / True Negative Rate : 0.0

Precision / Positive Predictive Value : 0.086

F1-score : 0.158

Recommendations :

Following are the strongest indicators of churn

Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations , compared to users who don't churn , when all other factors are held constant. This is the strongest indicator of churn. Customers who churn show lower number of recharges done in action period by 1.20 standard deviations, when all other factors are held constant. This is the second strongest indicator of churn. Further customers who churn have done 0.6 standard deviations higher recharge than non-churn customers. This factor when coupled with above factors is a good indicator of churn. Customers who churn are more likely to be users of 'monthly 2g package-0 / monthly 3g package-0' in action period (approximately 0.3 std deviations higher than other packages), when all other factors are held constant.

Based on the above indicators the recommendations to the telecom company are :

Concentrate on users with 1.27 std devations lower than average incoming calls from fixed line. They are most likely to churn. Concentrate on users who recharge less number of times (less than 1.2 std deviations compared to avg) in the 8th month. They are second most likely to churn. Models with high sensitivity are the best for predicting churn. Use the PCA + Logistic Regression model to predict churn. It has an ROC score of 0.87, test sensitivity of 100%.

From the above, the following are the strongest indicators of churn

Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations , compared to users who don't churn , when all other factors are held constant. This is the strongest indicator of churn.

Customers who churn show a lower number of recharges done in action period by 1.20 standard deviations, when all other factors are held constant. This is the second strongest indicator of churn.

Further customers who churn have done 0.6 standard deviations higher recharge than non-churn customers. This factor when coupled with above factors is a good indicator of churn.

Customers who churn are more likely to be users of 'monthly 2g package-0 / monthly 3g package-0' in action period (approximately 0.3 std deviations higher than other packages), when all other factors are held constant.

Based on the above indicators the recommendations to the telecom company are :

Concentrate on users with 1.27 std devations lower than average incoming calls from fixed line. They are most likely to churn.

Concentrate on users who recharge less number of times (less than 1.2 std deviations compared to avg) in the 8th month. They are second most likely to churn.

Models with high sensitivity are the best for predicting churn. Use the PCA + Logistic Regression model to predict churn. It has an ROC score of 0.87, test sensitivity of 100%

Conclusions from Random Forest

Local Incoming for Month 8, Average Revenue Per Customer for Month 8 and Max Recharge Amount for Month 8 are the most important predictor variables to predict churn.

Overall Conclusions

- Std Outgoing Calls and Revenue Per Customer are strong indicators of Churn.
- Local Incoming and Outgoing Calls for 8th Month and avg revenue in 8th Month are the most important columns to predict churn.
- cutomers with tenure less than 4 yrs are more likely to churn.
- Max Recharge Amount is a strong feature to predict churn.
- Random Forest produced the best prediction results followed by SVM.