Telecom Churn

Capstone Project

Objectives

- to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months

Overview of the dataset

- ‘telecom_churn_data.csv' : Contains customers data like Outgoing and Incoming calls minutes of usage, mobile internet usage, Recharge amounts, Recharge numbers, etc for four months (i.e. June, July, August, September).

Data Understanding

- Explored data using **.head()** function.
- The shape of the data was found out to (99999, 226)

Data Preparation

Filtering High-value customers

- Customers who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months are considered to be High-value customers
- This value X was found out to 368.5.
- Dropped the rows where the customers have recharged less than 368.5.
- After dropping the rows, the shape of the dataset was found to be (30011, 226).

Tag churners and Removing Attributes of the churn phase

- Based on the fourth month, we tagged the churned customers as follows: (churn=1, else 0) . Those who have not made any calls (either incoming or outgoing) and have not used mobile internet even once in the churn phase.
- After tagging churners, we removed all the attributes corresponding to the 9^{th} month (i.e. all attributes having ‘ _9’ in their names).

Handling Missing Values

- Identified null values for each column through **isnull()** function.
- Columns having more than 50% missing values were dropped.
- Rows were dropped for columns having less than 5% missing values
- The date type of columns were also dropped
- Finally the shape of the dataset that we got was (28487, 137)

Visualization

- Visualized distribution plot and Scatter plot for some of the variables
- The Distribution plots for variables with different months almost have same kind of distribution
- A linear relationship was seen between Total Recharge Number and Average Revenue per user (ARPU)
- The data usage for 3gb data is more than that of 2gb data but in any way its not affecting the ARPU

Model Building

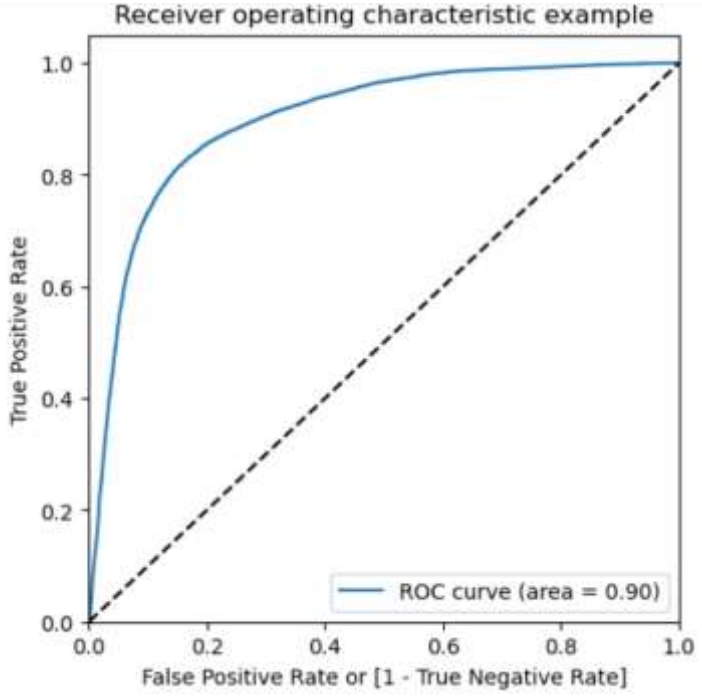
- The dataset was splitted into train and test sets.
- The churn rate was found out to be around 6% indicating class imbalance.
- To handle this class imbalance, resampling technique specifically Oversampling was done on train set using **SMOTE() function**.
- Feature scaling was done on this resampled set using **StandardScalar()** function.
- A logistic regression model was created.

Feature Selection

- Feature selection was done with the help of RFE and VIF.
- Using RFE, we selected 30 features out of 137 whereas VIF helped us to eliminate 5 more columns.
- Finally, we got 26 features as strong predictors of churn.

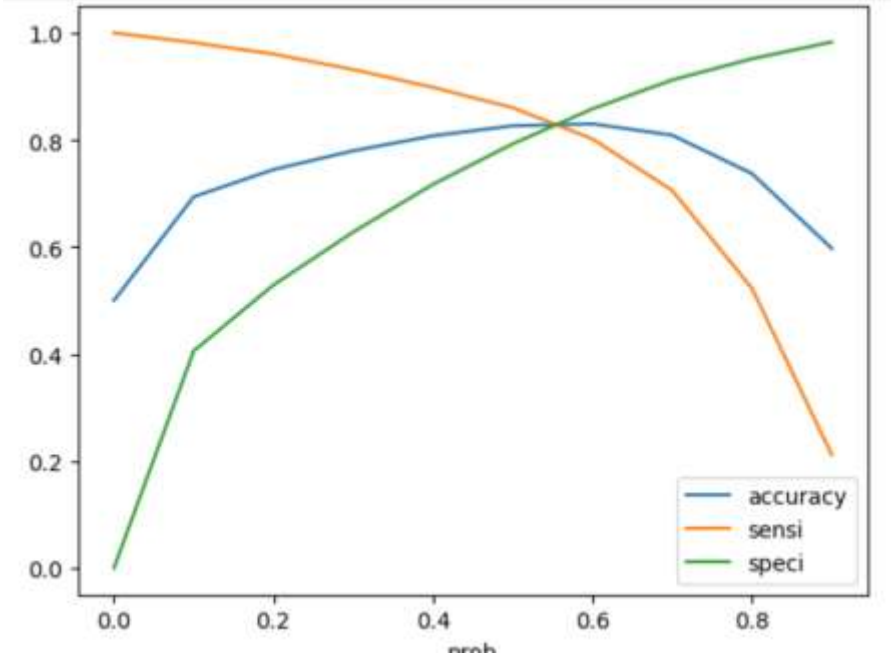
Plotted ROC curve

- From the graph we can see that our model got an AUC of 0.9 indicating overall model performance is perfect
- Finally, we got 26 features as strong predictors of churn.



Finding Optimal cutoff point

- From the graph we can see that 0.6 is the optimum point to take it as cutoff probability.

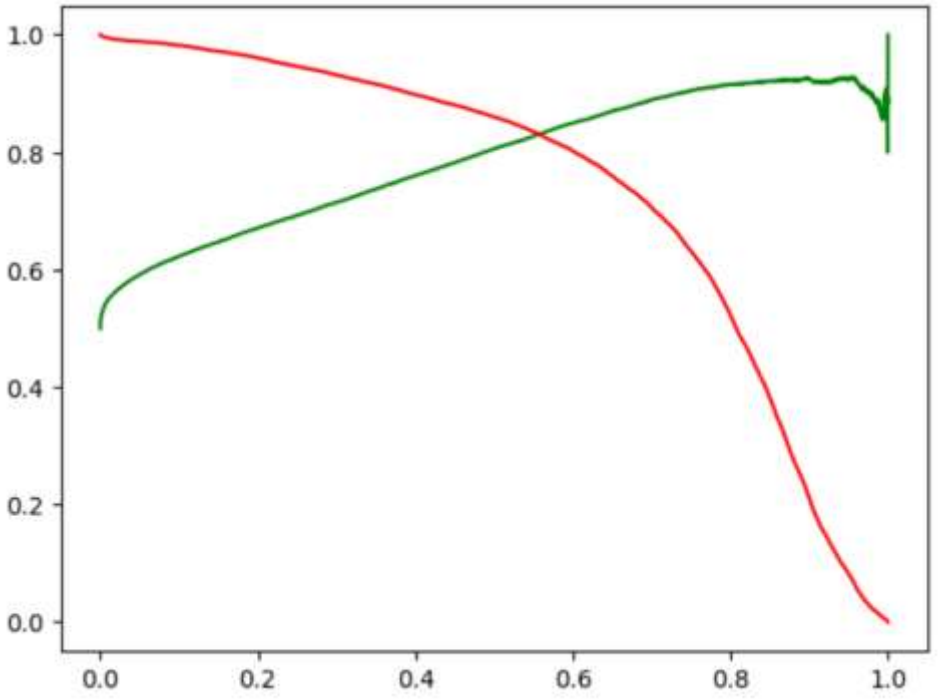


Accuracy, Sensitivity and Specificity

- As we confirmed our model performance to be perfect and Optimal cutoff point to be 0.6, we proceeded to make changes in our predicted churn column accordingly.
- After that we calculated Accuracy, Sensitivity and Specificity for train set.
- The Accuracy, Sensitivity and Specificity for train set was found out to be 82.9%, 80.1% and 85.8% respectively.

Precision and Recall

- From the precision recall curve we can see that we again got the cutoff point to be 0.6.



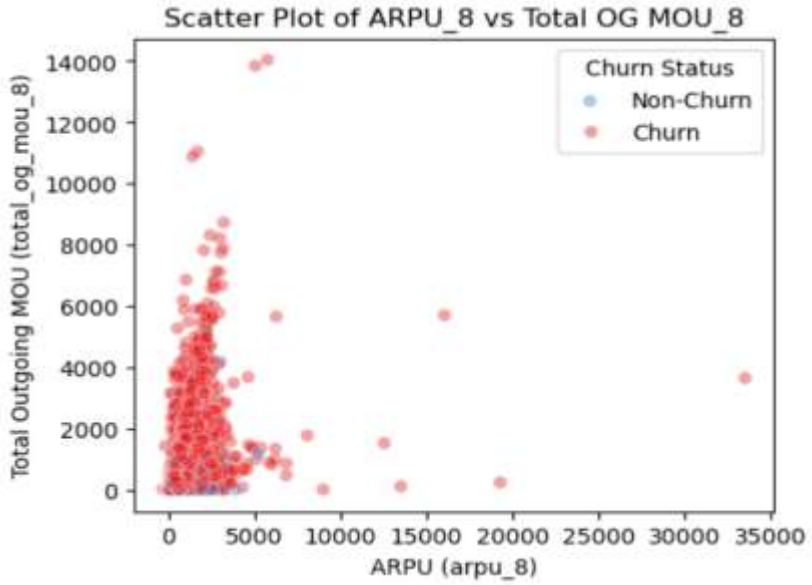
Making predictions on test set

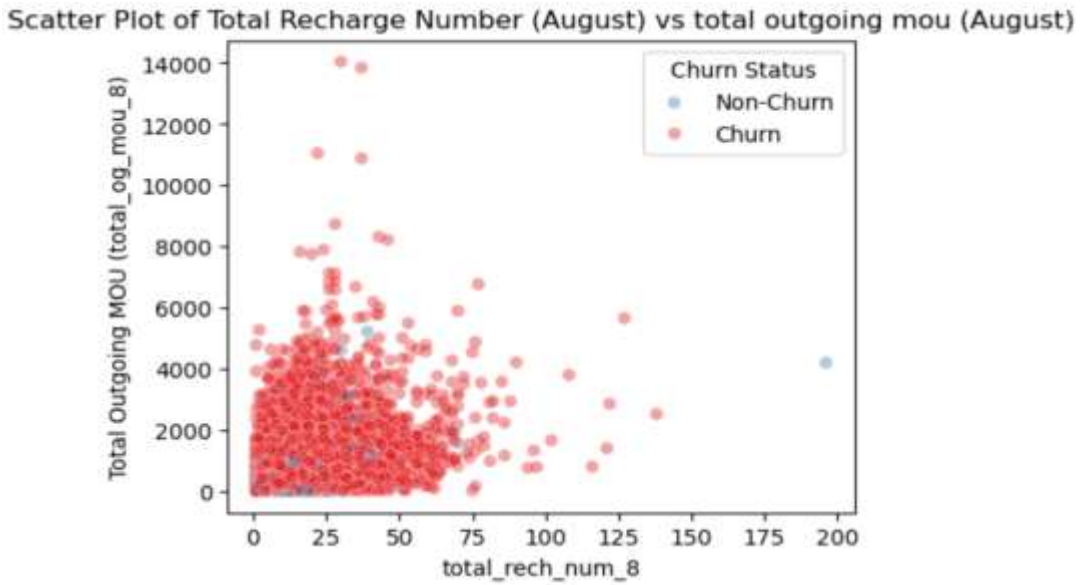
- The predicted values of the test set was calculated using built model and also Accuracy, Sensitivity and Specificity was calculated.
- The Accuracy, Sensitivity and Specificity for test set was found out to be 84.8, 70.1 and 85.7 respectively.
- As the difference between Accuracy for train and test set is not much, we can say that the model built was perfect with no Overfitting or Underfitting.

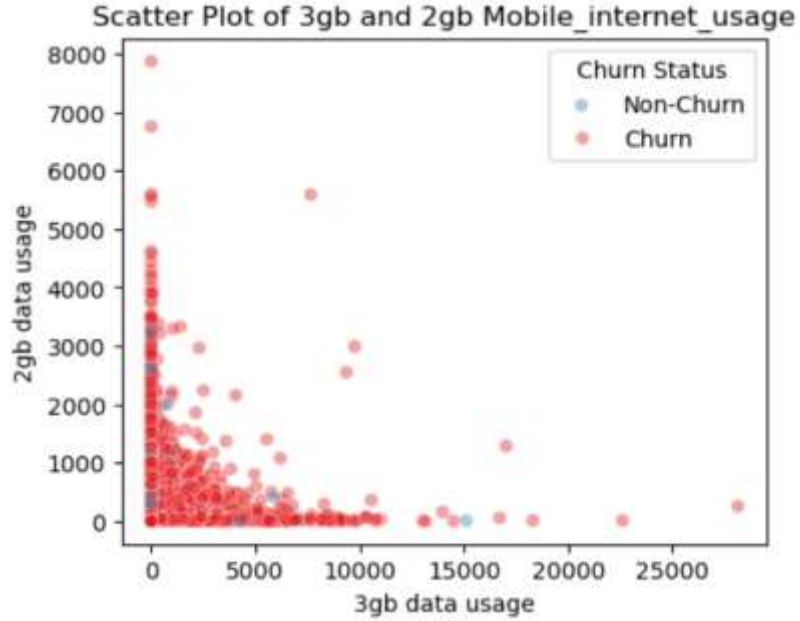
The following are the strong predictors of churn:

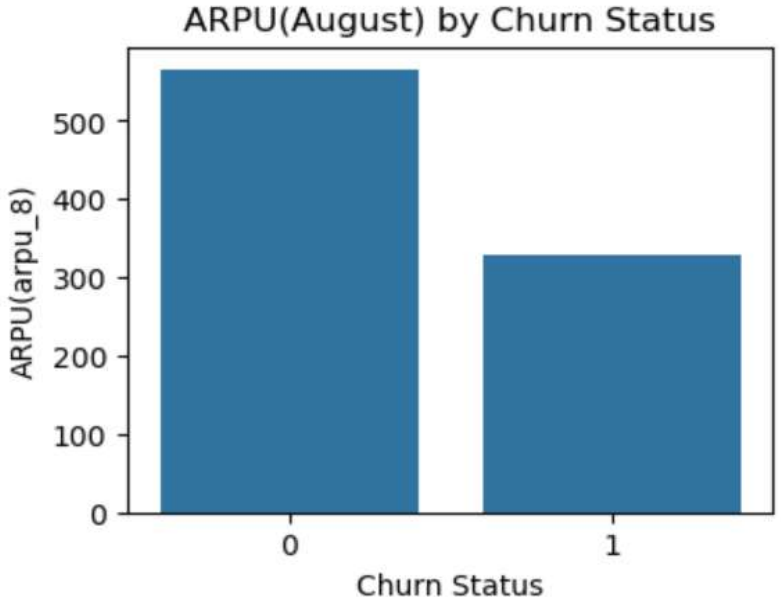
- 'loc_ic_t2m_mou_7', 'arpu_8', 'total_og_mou_8', 'loc_og_mou_8', 'loc_ic_t2m_mou_8', 'loc_og_t2m_mou_8', 'total_ic_mou_7', 'total_rech_amt_7', 'std_og_t2m_mou_7', 'std_og_t2t_mou_7', 'monthly_3g_6', 'monthly_2g_7', 'monthly_3g_8', 'vol_3g_mb_7', 'total_rech_num_8', 'monthly_2g_6', 'monthly_2g_8', 'jun_vbc_3g', 'last_day_rch_amt_8', 'sachet_2g_8', 'vol_2g_mb_7', 'sachet_3g_8', 'aon', 'std_ic_t2t_mou_8', 'roam_ic_mou_7', 'sep_vbc_3g'

Some Visuals after identifying the strong predictors









Conclusion

- We successfully load the dataset, handled the missing values, visualized some of the variables using plots, handled class imbalance for our target variable, done feature scaling, built a model using Logistic Regression, carried out feature selection using RFE and VIF, displayed the values of metrics like Accuracy, Sensitivity and Specificity, Plotted ROC curve, found Optimal cutoff point, done Precision and Recall tradeoff and finally made predictions for the test set.
- The churn rate was found out to 6% indicating a class Imbalance. So, Oversampling using SMOTE function was done to resample the train sets.
- From feature selection, we found 26 variables out of 134 variables to be strong predictors of churn.
- From both Optimal cutoff point and Precision-Recall curve, we found that Churn-probability of 0.6 is the best cutoff point to consider. So, the value of Churn-probility greater than 0.6 was considered as Churn and below 0.6 was considered as Non-churn.
- The Accuracy, Sensitivity and Specificity for train set was found out to be 82.9, 80.1 and 85.8 respectively.
- The Accuracy, Sensitivity and Specificity for test set was found out to be 84.8, 70.1 and 85.7 respectively.
- As the difference between Accuracy for train and test set is not much, we can say that the model built was perfect with no Overfitting or Underfitting.

Thank You