Telecom Churn Case Study

Arun Ramachandran

Problem Statement

Customers in the telecom sector have access to a variety of service providers and can actively switch from one operator to another. The telecoms business has an average annual churn rate of 15 to 25 percent in this fiercely competitive market. Customer retention has now surpassed customer acquisition in importance due to the fact that it is 5–10 times more expensive to gain new customers than to keep existing ones. Retaining highly profitable consumers is the top business objective for many established operator. Analysis needs to be done to predict models to find customers who are likely to leave

Analysis needs to be done for Prepaid and Postpaid model. The Churn is more critical for Prepaid Customers in India and Southeast Asian Market.

There are 2 types of churn Revenue based churn and Usage based churn. We need to do analysis on the usage based definition to define churn. The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months.

There are 3 phases of customer Good, Action and Churn. Churn is defined during the last phase as mentioned and data is discarded corresponding to the tag churned.

AnalysesApproach

1. **Reading and understanding the data**
2. **Data Cleaning**
3. **Filtering the High Value Customer**
4. **Defining Target Variable**
5. **Data Preparation**
6. **Data Modelling**
    I. **Creating dummies**
    II. **Train- Test split**

III. Handling Class Imbalance

7. Logistic Regression
   I. RFEtechnique forvariable selection II. Model Building
   III. Model Evaluation –Accuracy, Specificity, Sensitivity
   IV. Predicting on testdata
   V. Hyperparameter Tuning
8. Model Selection

Data Understanding and Cleaning

- Thedataset telecom_churn_data.csvhas around 99999entrieswith 226 attributes
- The missing valuesare imputed accordinglyfor data recharge ,count recharge columns
- Dropping columns/rowswith high missing values.

- Feature Engineering to create new columns

Filtering the High Value Customers and defining Target Variable

- Imputing the attributes having missing values with advanced imputation technique like `KNNImputer'.
- Two Types of Churn Usage Based churn *"Completely inactive Customers"* and Revenue Based Churn *"Partial Inactive Customers"*
- *9th month is the churn phase , Churn variable is derived using total_ic_mou_9`,`total_og_mou_9`,`vol_2g_mb_9` & `vol_3g_mb_9` attributes*
- Check for correlation of the independent variables and understand their dependencies.

Data Preparation

- Deriving New Variables
- Verifying correlation between target variable (Sale Price) and other variable in the data frame.
- Visualization of data to view the churn rate.

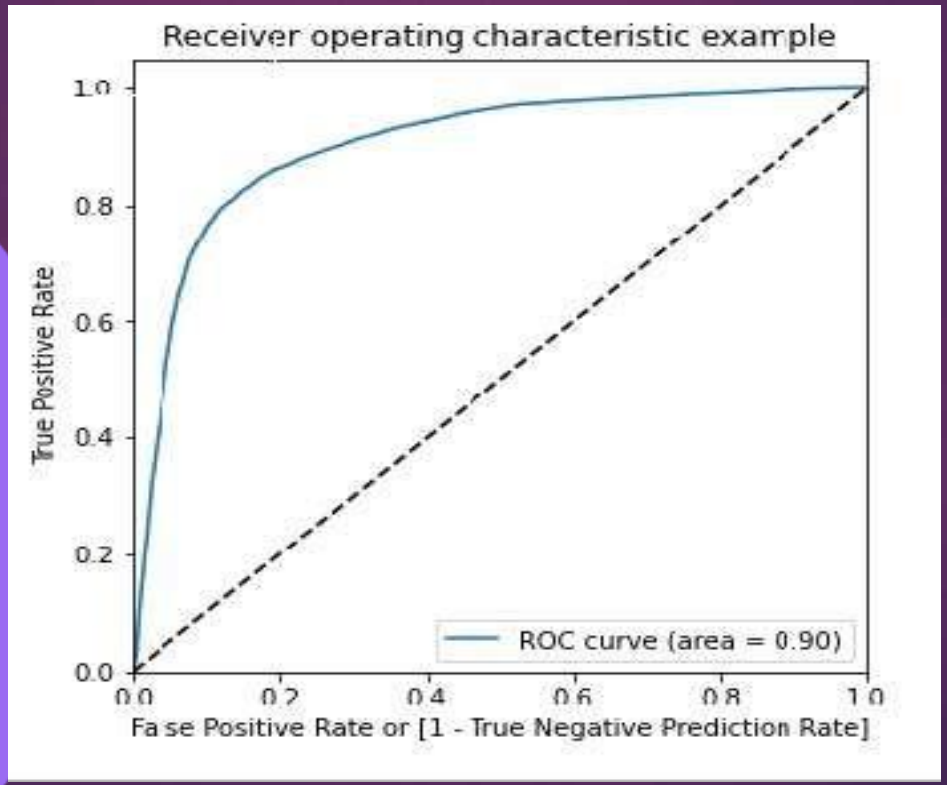
Model Building

- Creating Dummy Variables for categorical variables

Handling Class imbalance, using SMOTE method.

- Featuresareselected using RFE
- Modelisformed using Logistic Regression
- Bychecking p value and vifvalues, features are dropped and optmal model isobtained
- Parameters likeAccuracy, Specificity, Sensitivityare calculated
- Predictionsareapplied on Testdata and evaluated

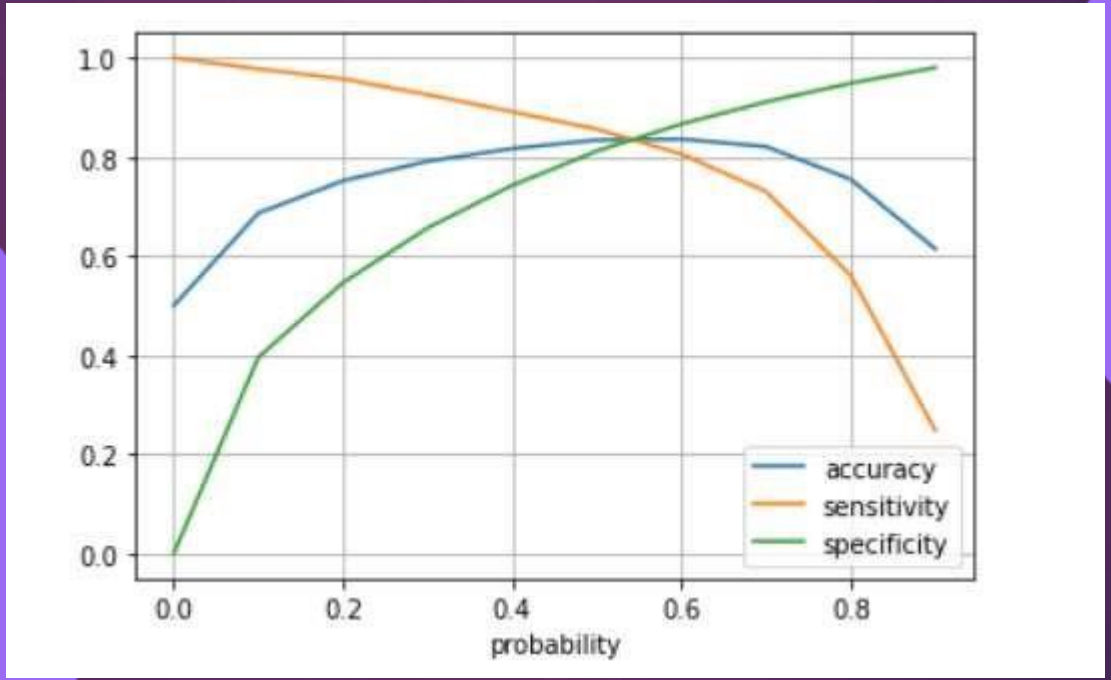
Model evaluation –ROC Curve



- ROC Curve

Model Evaluation-Accuracy

- Accuracy :83.6
- Sensitivity:. 83.6
- Specificity:83.6



From the above graph, we can see the optimum cutoff is around 0.54**

Model Evaluation on Testdata

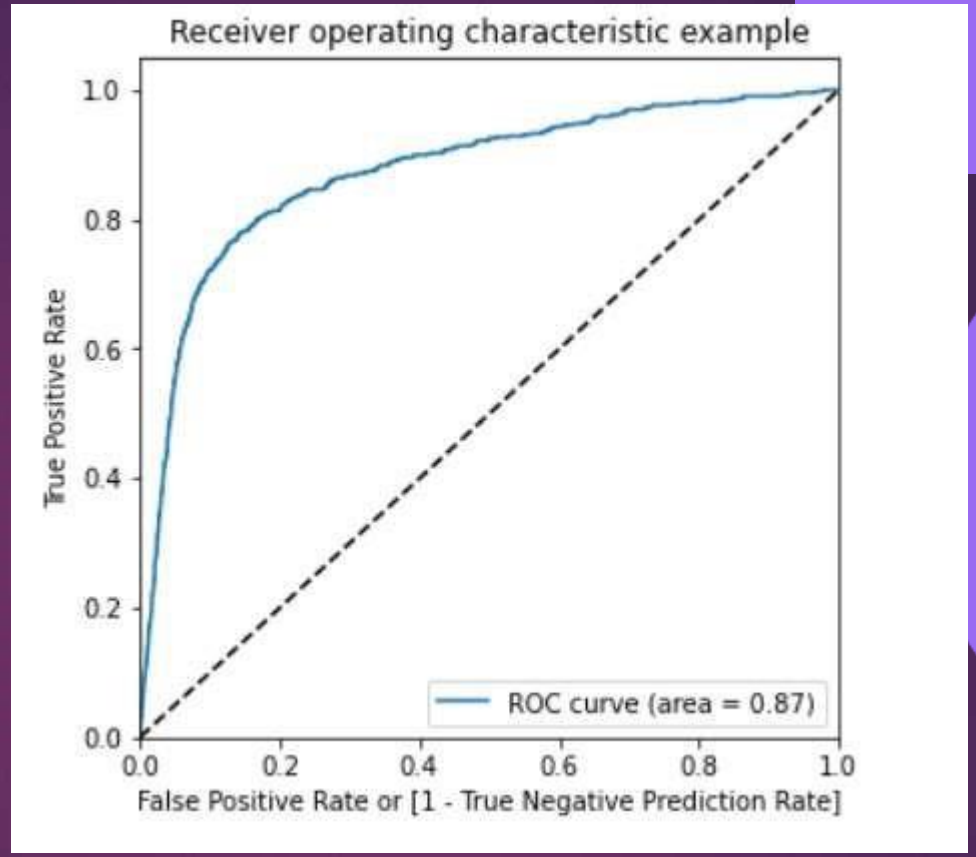
- Accuracy :83
- Sensitivity:80.0
- Specificity:82.9

As the model created is based on a sensitivity model, i.e. the True positive rate is given more importance as the actual and prediction of churn by a customer

Model evaluation –ROC Curve

- ROC Curve

The AUC score for train Data Frame is 0.90 and the test Data Frame is 0.87.This model can be considered as a good model.



PCA and on Logistics Regression

SVM

- The Accuracy of the logistic regression model in train with PCA: 81.8
- Accuracy of the logistic regression model in test with PCA: 75.4

- SVM Analysis Below.

Accuracy 78.1
Precision 23.3
Recall 74.3

Hyper parameter Tuning

- The Test score is 86.9 corresponding to hyper parameters {'C': 1000, 'gamma': 0.01}
- Random Forest Analysis Below.

Accuracy 93.2

Precision 73.8

Sensitivity/Recall 24.8

Roc_auc_score 62.0

Model Selection

- The best model out of all is Logistic regression model which gives recall of 81% and ROC value of 0.89

Inferences

- Accuracy, Sensitivityand Specificityare insimilarrange forTrainand Test data
- Std Outgoing Calls and Revenue Per Customer are strong indicators of Churn.
- Local Incoming and Outgoing Calls for 8th Month and avgrevenue in 8th Month are the most important columns to predict churn.
- Customers with tenure less than 4 yrs are more likely to churn.
- Max Recharge Amount is a strong feature to predict churn.
- Logistic Regression produced the best prediction results after tackling Class Imbalance