TELECOM CHURN CASE STUDY

Submitted by-
UJWAL KESHARWANI
NITISH SINGH
SIDDHI KADAM

Problem Statement

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. The telecommunication industry has an average of 15-25% annual churn rate. 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 carried out to find customers who are likely to leave This is done for Prepaid and Postpaid model.

- The two types of churns are Revenue-based churn - Customers who have not utilized any revenue generating facilities and Usage-based churn - Customers who have not done any usage, either incoming or outgoing. We need to do analysis on the usage based definition to define churn.

Business goals

- To develop a model in order to identify the customers at high risk of 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.
- To define high-value customers based on a certain metric and predict churn on high-value customers.

Approach

- Importing The Dataset.
- Data Inspection.
- Data Cleaning.
- Filtering the High Value Customer.
- Defining Target Variable.
- Prepare The Data For Model Building.
- Model Building.
  - Creating dummies.
  - Train- Test split.
  - Handling Class Imbalance.

- Logistic Regression.
  - RFE technique for variable selection.
  - Model Building.
  - Model Evaluation – Accuracy, Specificity, Sensitivity.
  - Predicting on test data.
  - Hyperparameter Tuning.

- Model Selection.

Data Inspection and Cleaning

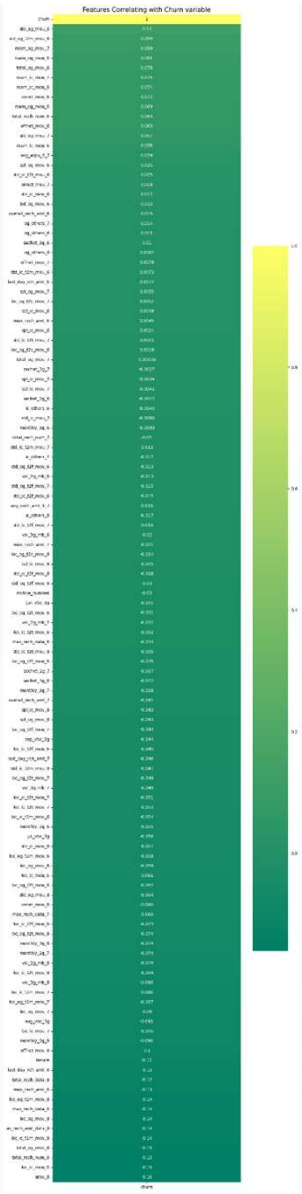
- The dataset used for Telecom churn case study is ‘telecom_churn_data.csv’. It has 99999 entries with 226 attributes.
- The missing value imputation is carried out for “Data Recharge” column and “Count Recharge” column.
- The columns with high missing values are dropped.
- Feature Engineering is carried out to create new columns.

Filtering High Value customers and Defining Target Variables

- The 70th percentile in the overall revenues is defined to determine the high value customer criteria for the company.
- The remaining attributes are imputed with advanced imputation technique called ‘KNNImputer’.
- For defining the Target variables, there are two types of churn: Usage Based churn - Completely inactive Customers and Revenue Based Churn-Partial Inactive Customers.
- Since we are focusing on Usage Based churn, 9th Month is our Churn Phase.
- The churn variables are derived using total_ic_mou_9`,`total_og_mou_9`,`vol_2g_mb_9` & `vol_3g_mb_9` attributes.

Data Preparation

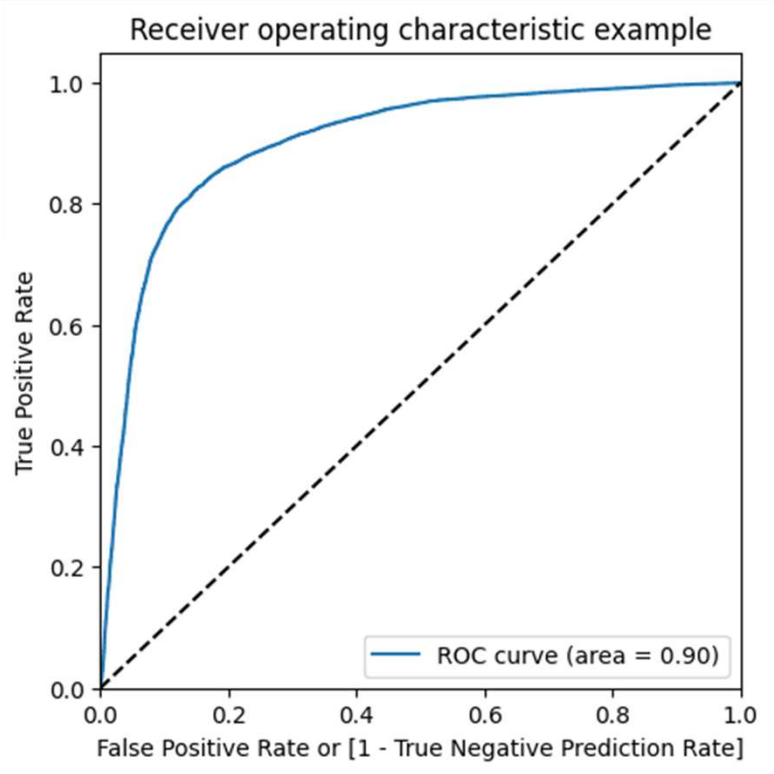
- New variables are derived.
- Correlation is verified between target variable (Sale Price) and other variable in the data frame as shown.
- Data visualization is carried out.



Model Building

- Dummy variables for Categorical variables are created.
- Class Imbalance is handled using SMOTE method (balance the data w.r.t. churn variable).
- Feature selection using RFE.
- Model building using Logistic Regression.
- The p-values and VIF values are evaluated to drop certain features and an optimal model is obtained.
- The values of Sensitivity, Specificity, False Positive Rate, Precision and True Positive Rate are calculated.

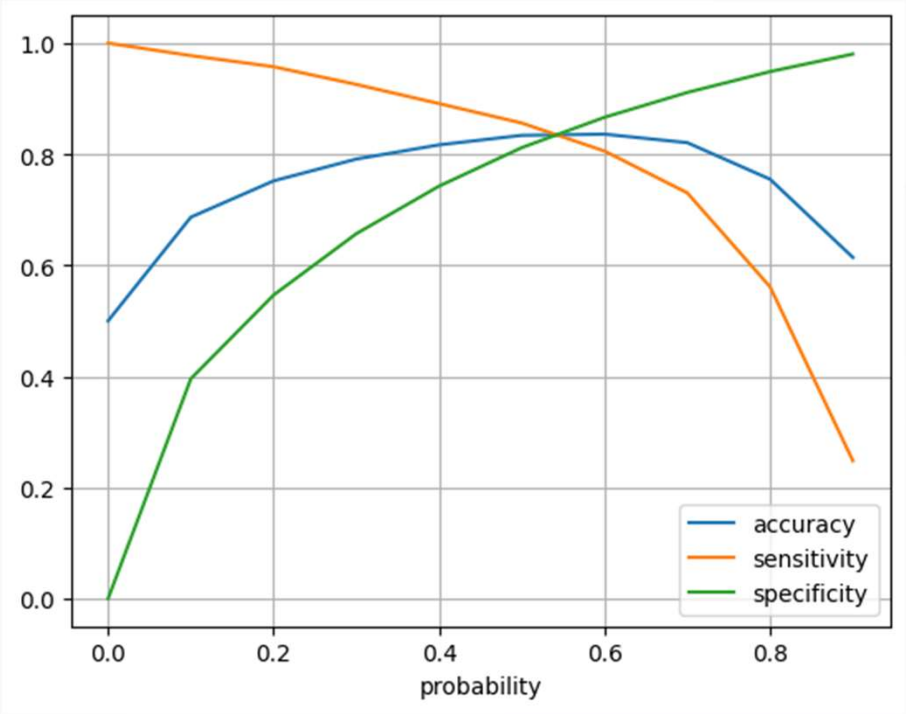
Model evaluation- ROC curve



Model evaluation

- Accuracy- 0.83
- Sensitivity- 0.83
- Specificity- 0.83

- The optimal cut-off from the graph obtained is 0.54

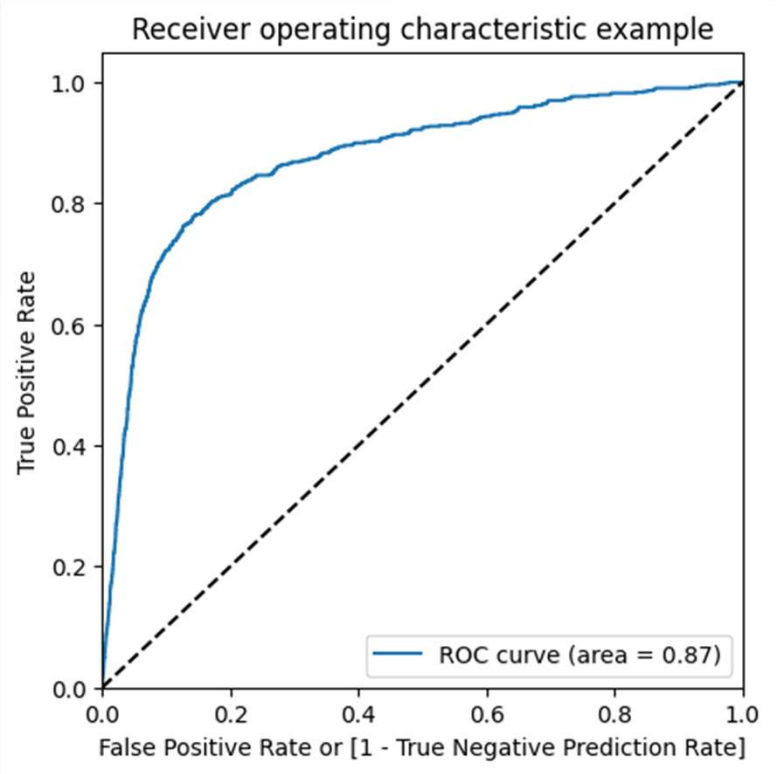


Model evaluation- Test data

Evaluating the model on test data, the values obtained are,

- Sensitivity- 0.80
- Specificity- 0.83
- False Positive Rate-0.17
- Precision-0.29
- True Negative Rate-0.97

Model evaluation- ROC curve



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

Logistic Regression using PCA and SVM model

- The Accuracy of the Logistic regression model on train dataset with PCA: 81.8
- The Accuracy of the Logistic regression model on test dataset with PCA: 75.4
- Using SVM Logistic regression model- Accuracy 78.1, Precision 23.3 and Recall 74.3

Hyperparameter Tuning

- GridSearchCV() is used to tune the hyperparameters.
- The Test score is 86.9 corresponding to hyper parameters {'C': 1000, 'gamma': 0.01}.
- Using the Random Forest model, the Accuracy is 93.2, Precision is 73, Sensitivity is 24.8 and ROC score is 62.0

Model Selection

- The best model out of all the models evaluated is the Logistic regression model which gives Recall of 81% and ROC value of 0.89

Results

- Accuracy, Sensitivity and Specificity are in similar range for the train data and test data.
- Std Outgoing Calls and Revenue Per Customer are strong indicators of Churn.
- Local Incoming and Outgoing Calls for 8th Month and average revenue in 8th Month are the most important columns to predict churn.
- Customers with tenure less than 4 years are more likely to churn.
- Max Recharge Amount is a strong feature to predict churn.
- Logistic Regression produced the best prediction results which gives Recall of 81% and ROC value of 0.89