Telecom Churn Case Study

Study by : Pranav Kohli
Darshan JK
Shivam Baliyan

PROBLEM STATEMENT

In the telecom sector, consumers have the flexibility to choose from various service providers and actively switch between operators. This industry experiences a significant average annual churn rate of 15 to 25 percent, given its highly competitive nature. With the cost of acquiring new customers being 5–10 times higher than retaining existing ones, the emphasis has shifted towards customer retention over acquisition. Retaining high-value customers stands as a paramount objective for many established operators. To achieve this, predictive modeling is imperative to identify customers likely to churn.

This analysis encompasses both Prepaid and Postpaid models, with particular emphasis on the criticality of churn within the Prepaid segment, especially in markets like India and Southeast Asia. Within the realm of churn, two distinct types are recognized: Revenue-based churn and Usage- based churn. Our focus lies on analyzing the latter to delineate churn patterns accurately.

The primary business goal is to forecast churn during the final month (i.e., the ninth month), leveraging data features from the initial three months. This process involves categorizing customers into three phases: Good, Action, and Churn. Churn is specifically defined during the final phase, and corresponding data tagged as churned is disregarded from the analysis.

APPROACH TO PROBLEM

1. Reading and understanding the data

2. Data Cleaning

3. Filtering the High Value Customer 4. DefiningTargetVariable

5. Data Preparation 6. Data Modelling

I. Creating dummies

I. Train- Test split

III. Handling Class Imbalance

7. LogisticRegression

I. RFEtechnique forvariable selection

II. Model Building

III. Model Evaluation –Accuracy, Specificity, Sensitivity

IV. Predicting on test data

V. Hyperparameter Tuning

8. ModelSelection

DATA UNDERSTANDING AND CLEANING

- The dataset named telecom_churn_data.csv comprises approximately 99,999 entries containing 226 attributes.
- Missing values in the data recharge and count recharge columns are filled in appropriately through imputation techniques.
- Columns or rows with significant missing values are removed.
- Feature engineering techniques are applied to generate new columns.

FILTERING THE HIGH VALUE CUSTOMERS + DEFINING TARGET VARIABLE

Using advanced imputation techniques such as 'KNNImputer', missing attribute values are filled in.

- Two categories of churn are identified: Usage-Based Churn, characterized by "Completely Inactive Customers," and Revenue-Based Churn, denoted by "Partial Inactive Customers."

- The ninth month marks the churn phase, determined by attributes including 'total_ic_mou_9', 'total_og_mou_9', 'vol_2g_mb_9', and 'vol_3g_mb_9'.

- Correlation analysis of independent variables is conducted to discern their interrelationships and dependencies.

DATA PREPARATION

- Creating New Variables
- Examining the correlation between the target variable (Sale Price) and other variables in the dataset.
- Visualizing the data to observe the churn rate.

MODEL BUILDING

Generating Dummy Variables for categorical variables.

Addressing class imbalance using the SMOTE method.

Employing Recursive Feature Elimination (RFE) for feature selection.

Constructing the model using Logistic Regression.

Evaluating and refining the model by dropping features based on p-value and VIF values to achieve optimization.

Calculating parameters such as Accuracy, Specificity, and Sensitivity. Applying predictions on the test dataset and conducting evaluations.

MODEL EVALUATION-ACCURACY

Accuracy :83.6

Sensitivity :. 83.6

Specificity :83.6

MODEL EVALUATION ON TEST DATA

Accuracy: 83%

Sensitivity: 80.0%

Specificity: 82.9%

Since the model prioritizes sensitivity, emphasizing the true positive rate in predicting churn by customers, the accuracy stands at 83%.

MODEL EVALUATION –ROC CURVE

ROC Curve

The AUC score for the train dataset is 0.90, and for the test dataset, it is 0.87. This indicates that the model can be deemed as effective.

PCA AND SVM ON LOGISTICS REGRESSION

The logistic regression model's accuracy in training with PCA is 81.8%.

The logistic regression model's accuracy in testing with PCA is 75.4%.

Below is the analysis using Support Vector Machine (SVM):

- Accuracy: 78.1%
- Precision: 23.3%
- Recall: 74.3%

HYPER PARAMETER TUNING

The test score is 86.9%, associated with the following hyperparameters: {'C': 1000, 'gamma': 0.01}.

Below is the analysis using Random Forest: - Accuracy: 93.2%

- Precision: 73.8%

- Sensitivity/Recall: 24.8%

- ROC AUC score: 62.0%

Model Selection

The top-performing model among all is the Logistic Regression model, achieving a recall of 81% and a ROC value of 0.89.

INFERENCES

Accuracy, Sensitivity, and Specificity exhibit similar ranges for both the training and test datasets.

Standard Outgoing Calls and Revenue Per Customer serve as robust indicators of churn.

Local Incoming and Outgoing Calls for the 8th month, along with average revenue in the 8th month, emerge as pivotal columns for predicting churn.

Customers with a tenure of less than 4 years are more prone to churn.

Max Recharge Amount emerges as a significant feature for predicting churn.

Logistic Regression yielded the most accurate predictions after addressing class imbalance.

THANK YOU