Telecom Churn Case Study

Problem Statement

Business Problem Overview

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

For many incumbent operators, retaining high profitable customers is the number one business goal.

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

In this project, we will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Understanding the Business Objective and the Data

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

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 do this task well, understanding the typical customer behaviour during churn will be helpful.

Understanding Customer Behaviour During Churn

Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of customer lifecycle :

1. The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.
2. The ‘action’ phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
3. The ‘churn’ phase: In this phase, the customer is said to have churned. We define churn based on this phase. Also, it is important to note that at the time of prediction

(i.e. the action months), this data is not available to us for prediction. Thus, after tagging churn as 1/0 based on this phase, we discard all data corresponding to this phase.

In this case, since we are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month is the ‘churn’ phase.

Deriving New Features

Filtering only High Value Customers

In this segment, we'll identify high-value customers aligning with our business objectives, focusing solely on prepaid customers experiencing usage-based churn. We'll evaluate the recharge amounts for June and July, selecting only the top 70% of customers as high-value.

We can create a new feature named Total Data Recharge Amount by multiplying the values of total_rech_data and av_rech_amt_data, which represents the amount recharged by the customer for data usage.

- Total Data Recharge Amount = Total Data Recharge * Average Data Recharge Amount

Created another column for total recharge done Total Amount for the months 6 and 7.

- Total Amount = Total Data Recharge Amount + Total Recharge Amount

Additionally, we can compute the Total Average Recharge Amount during the 'Good Phase', which includes months 6 and 7.

- Total Average Amount = (Total Data Recharge Amount + Total Recharge Amount) / 2

Impute missing values using KNN Imputer - Training Set

We cannot discard these missing values because doing so would result in the loss of valuable information. Instead, we will employ imputation techniques such as KNNImputer.

KNNImputer is a technique used to impute missing values in a dataset based on the values of its nearest neighbors. It works by identifying the k nearest neighbors of each data point with missing values, then averaging or taking a weighted average of the available values from those neighbors to fill in the missing value. This approach leverages the similarity between data points to estimate the missing values more accurately.

Observations:

- A substantial correlation of 74% exists between the total recharge amount in month 7 and month 8.
- Similarly, a notable correlation of 68% is observed between the maximum recharge amount in month 8 and the last day's recharge amount in the same month.

- This suggests that customers who are unlikely to churn tend to recharge higher amounts in month 8.

Key Observations:

Model 1 (Logistic Regression with Recursive Feature Elimination - RFE):

Training: Good accuracy (85.69%), high sensitivity (83.51%), and specificity (87.87%).

Testing: Slightly higher accuracy on the test set (86.67%) but lower sensitivity (77.45%) and precision (35.03%).

Issue: The low precision on the test set could indicate that the model has a problem with false positives, or it is misclassifying a significant number of negative cases as positive.

Model 2 (Logistic Regression with PCA and Hyperparameter Tuning):

Training: Lower accuracy (80.96%) and precision (27.73%) compared to Model 1. Sensitivity and specificity are somewhat balanced.

Testing: Accuracy (80.98%) is close to the training accuracy. Precision drops significantly (26.48%) on the test set.

Issue: Like Model 1, the precision is quite low on the test set, indicating that the model may be predicting more false positives.

Model 3 (Decision Tree with PCA and GridSearchCV):

Training: Excellent performance on training with very high accuracy (95.23%), sensitivity (94.85%), and specificity (95.26%).

Testing: Accuracy drops to 82.86%, sensitivity drops to 54.30%, and precision is very low (24.43%).

Issue: The drastic drop in performance on the test set shows that the model might be overfitting, as it is not generalizing well. The significant decrease in sensitivity and precision indicates that the model struggles with correctly classifying positive cases during testing.

Model 4 (Random Forest Classifier with PCA and Hyperparameter Tuning):

Training: High accuracy (90.91%) and sensitivity (94.21%) on the training set.

Testing: Performance drops to 82.86% accuracy with very low sensitivity (54.30%) and precision (24.43%).

Issue: Similar to Model 3, the performance drop indicates overfitting, with the model failing to generalize well to the test set. Low precision could indicate a high number of false positives.

Model 5 (AdaBoost Classifier):

Training: Lower accuracy (82.62%) compared to the other models, but sensitivity (79.18%) and specificity (82.94%) are still reasonable.

Testing: The performance on the test set (Accuracy 82.86%, Sensitivity 54.30%, Precision 24.43%) is similar to that of Models 3 and 4.

Issue: AdaBoost struggles with generalization as well, similar to the other models.

Business Recommendations

Based on the analysis of our logistic regression model with RFE, here are some business ideas to improve churn rate:

1. Roaming Offers: Provide personalized roaming packages to frequent roamers.
2. Local Call Promotions: Offer competitive rates and bonuses for local calls.
3. Data Recharge Strategies: Promote data packs with targeted marketing campaigns.
4. High-Value Recharge Incentives: Offer discounts for high-value recharges to retain customers.
5. Service Engagement Initiatives: Enhance engagement through loyalty programs and personalized offers.
6. Retention Campaigns: Target customers with low recharge activity with special offers.
7. Non-Data User Promotions: Encourage non-data users to try data services with bundle offers.
8. Night Pack Revival: Revive night pack usage through attractive offers and incentives.

Implementing these strategies can effectively reduce churn and improve customer retention in your telecom business.

Summary

After experimenting with various models, including Logistic Regression with Recursive Feature Elimination (RFE), Logistic Regression with hyperparameter tuning, and PCA, as well as Decision Tree, Random Forest, Adaboost, and XGBoost classifiers with hyperparameter tuning and PCA, it's evident that only Logistic Regression with PCA consistently demonstrates the highest sensitivity in both the train and validation sets. Consequently, this model should be considered as the final choice. Other models, although showing promising accuracy in the training phase, perform poorly on the test set, suggesting overfitting.

In the context of telecom churn, where minimizing churn rate is crucial, sensitivity emerges as the most pertinent metric. Hence, based on this criterion, the Logistic Regression model with PCA stands out as the most suitable choice among all alternatives.