Telecommunications Service Provider Customer Churn Data Analysis

Case Study: SyriaTel Telecommunications

Business Understanding

SyriaTel, a mobile telecommunications company based in Damascus, Syria, aims to identify factors leading to customer churn based on historical data and behavioral patterns. The goal is to enhance overall customer satisfaction by mitigating churn factors early, contributing to cost-effectiveness, revenue growth, and long-term business sustainability.

Problem Statement

The telecom industry is experiencing challenges in retaining customers, with an increasing churn rate affecting overall business sustainability. To address this issue, a data analysis project aims to predict customer churn using a dataset from a telecom company. The objective is to create a model that identifies customers likely to churn, providing insights to reduce churn and increase profits. According to

research, a 1% reduction in churn can lead to a 5% profit increase in the telecom industry ([source](https://www.subscriptionflow.com/2022/11/churn-management-in-telecom-industry/)).

Stakeholders

SyriaTel Senior Management: To formulate strategies based on identified churn factors.

Technical Support: To address technical issues highlighted by the model.

Customer Service: To guide customers through hurdles identified for improved satisfaction and retention.

Data Understanding

The dataset provides the following features related to SyriaTel customers in relation to their binary churn values

Unique Identifier (Phone Number): Unique identifier for each customer.

Target (Churn): Binary indicator (yes/no) of customer churn.

State, Area Code: Customer demographics.

Account Length: how long a customer's account is active

International Plan, Voice Mail Plan: Binary indicators of additional services.

Number Vmail Messages, Customer Service Calls: Usage metrics.

Total Day/Evening/Night Minutes, Calls, Charge: Usage and cost metrics.

Featured engineered attribute

Featured engineered attribute was **cumulative daily charges.** The feature sums up each customer's grand total charges per day from the data summed up in the day, evening, night and international daily charges.

Objective

The study aims to identify factors leading to customer churn, provide actionable recommendations, and suggest future improvement areas.

Libraries Used

Python

Scikit-learn

Seaborn

Matplotlib

Pandas

Numpy

Exploratory Data Analysis (EDA)

The EDA was done in two sections, before and after the Modelling section. The first EDA was to lay a foundation of the data understanding by seeing how some features relate to churn. The EDA after modeling was guided by the feature selection method used. Decision Trees feature importance was the preferred method and from its results, three most important features were established and EDA was performed on them for deeper insights.

Insights

Exploratory Data Analysis (EDA) revealed valuable insights into factors influencing customer churn for SyriaTel Telecommunications. Customers churning made fewer than 5 voicemail messages, especially those without a Voicemail Plan. Customer service calls were minimal, with more than half making fewer than 4 calls, suggesting potential dissatisfaction with service assistance or availability. The cumulative daily charge emerged as a significant factor, with most churned customers experiencing high charges, indicating a possible link between pricing and churn. The importance of the engineered cumulative daily charge, along with features like customer service calls and voicemail messages, underscores their role in influencing customer retention.

Models

The dataset was split into an 80/20 ratio for training and testing, respectively.

A baseline model using decision trees was established. Feature selection was performed by extracting feature importance from the decision trees model, selecting features with importance above a 0.1 threshold. The identified important features, in order of significance, are Cumulative daily charges, Customer service calls, Number of voicemail messages, Total international minutes, International plan, and Total international calls.

Tuning efforts were initiated, leading to Model 2, a decision tree model incorporating the top three important features: Cumulative daily charges, Customer service calls, and Number of voicemail messages. Model 3 involved cross-validation plotting against thresholds ranging from 0 to 0.2, with the optimal threshold determined to be 0.011. A decision tree model was then created using this best threshold.

Model 4 was a random forest model featuring only the top three most important features. Further refinement was pursued with Model 5, involving grid search to identify the best parameters for max depth, min samples split, min_samples_leaf, max features, and n estimates. The best parameters were found to be 'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100, and a random forest tree was generated using these optimal parameters.

Evaluation

Our primary focus metric throughout the model evaluation was recall since shows the actual positive instances (churned customers) that the model correctly identifies. In the baseline model, we achieved an accuracy of 0.98 with a recall of 0.87. Model 2, a decision tree incorporating the selected most important features, exhibited a lower accuracy of 0.95 and a reduced recall of 0.68.

Model 3, another decision tree with the best threshold, yielded an accuracy of 0.951 and a recall of 0.871. Model 4, the original random forest model with selected most important features, showed strong performance with an accuracy of 0.9805, recall of 0.8713, and an F1 score of 0.9312. The cross-validation mean recall and F1 score were 0.8559 and 0.9220, respectively.

Model 5, the grid search-tuned random forest model, mirrored the results of Model 4, maintaining an accuracy of 0.9805 and recall of 0.8713. The F1 score was 0.9312, with cross-validation mean recall and F1 score at 0.8571 and 0.9230, respectively. The consistency in performance between the original and tuned models suggests that the hyperparameter tuning had minimal impact. Considering the marginal improvement and resource intensity of the tuned model, the original random forest is chosen as the final model.

Interpretation

The accuracy, confusion matrix, and F1 score remained almost identical before and after grid search. This suggests that the hyperparameter tuning did not result in significant changes in the model's performance. This could mean that the model is at its optimal performance

Key Observations

High Accuracy:

The model exhibits high accuracy, correctly classifying a large portion of instances

Recall:

The recall is consistently high at 0.87, indicating that the model effectively identifies a significant portion of true positive instances (churn). Recall is particularly important in this scenario since identifying all instances of churn is crucial.

Confusion Matrix:

The confusion matrix shows a high number of true positives (88) and true negatives (566) with no false positives (FP = 0). The model has 0 false positive, this is very beneficial, indicating that the model is not incorrectly classifying non-churn instances as churn.

F1 Score:

The F1 score remains high at 0.931, emphasizing a good balance between precision and recall.

Key Considerations

Consistency: Both model 4 and 5 show consistent performance.

Cross-Validation: The grid search tuned random forest model shows slightly better generalization.

Computational Efficiency: The tuned model is resource-intensive with marginal improvement.

Conclusion

Given the marginal improvement and resource intensity of the grid search tuned random forest model, the original Random Forest is chosen as the final model.