SyriaTel Customer Churn Predictive Analysis

Overview

Objective

The objective of the customer churn predictive analysis project is to identify customers who are likely to churn in order to take proactive measures to retain them.

Methodology

The project utilizes machine learning algorithms to analyze historical customer data and identify patterns and indicators of churn. The model is trained on a labeled dataset to predict the likelihood of churn for each customer.

Business Problem

Customer Churn

Customer churn refers to the rate at which customers stop using a company's products or services. For SyriaTel, customer churn is a significant problem as it leads to a loss of revenue and market share. It is crucial for SyriaTel to address this issue in order to maintain a sustainable business and retain its customer base.

Importance of Addressing Churn

Addressing customer churn is important for several reasons:

1. Revenue Loss: When customers churn, the company loses out on their subscription fees and potential future purchases.
2. Market Share: High churn rates can lead to a decline in market share, making it difficult to compete with other telecom providers.
3. Customer Satisfaction: High churn rates indicate that customers are not satisfied with the company's services, which can damage the company's reputation.
4. Cost of Acquisition: Acquiring new customers is more expensive than retaining existing ones. By reducing churn, SyriaTel can save on customer acquisition costs.

Data Understanding

The data for thus project is sourced from kaggle and it primarily focuses on customer details, specifically customer churn.

The data consists of 21 columns of user/consumer data and 3,333 customer records.

Data Cleaning

- Data cleaning involves handling outliers, dealing with duplicate records, and removing any noise or inconsistencies in the data.
- Outliers can be detected using statistical techniques or domain knowledge and can be handled by either removing them or transforming them to bring them within an acceptable range.
- Duplicate records can be identified by comparing the values of all variables and removing the duplicates to avoid bias in the analysis.

Data Analysis and Modelling

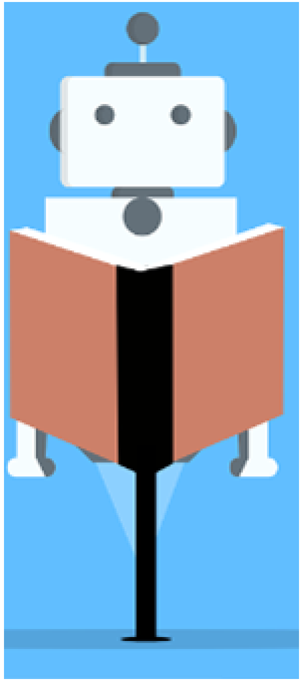
Data Preparation

- Cleaned and preprocessed the dataset to remove missing values and outliers.
- Conducted feature engineering to create new variables for model training.
- Split the dataset into training and testing sets for model evaluation.

Model Selection

- Explored various machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting.
- Evaluated the performance of each model using metrics like accuracy, precision, recall, and F1-score.
- Selected the best performing model for further analysis.

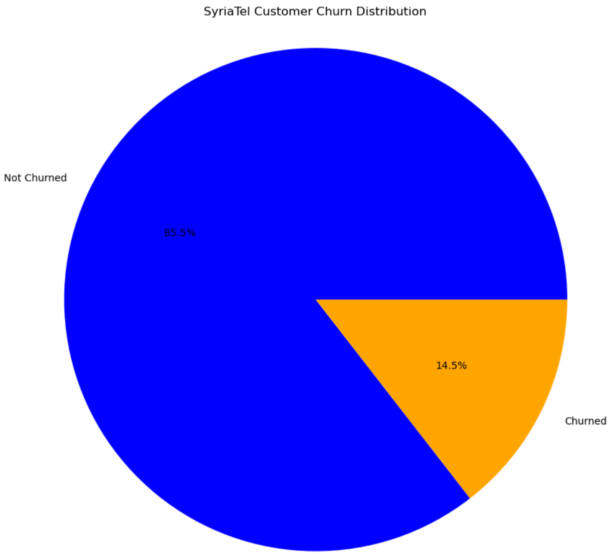
Model Training and Evaluation



- Trained the selected models on the training dataset using cross-validation techniques.
- Tuned the hyperparameters of the model to optimize its performance.
- Evaluated the trained models on the testing dataset to assess its generalization ability.

Exploratory Data Analysis

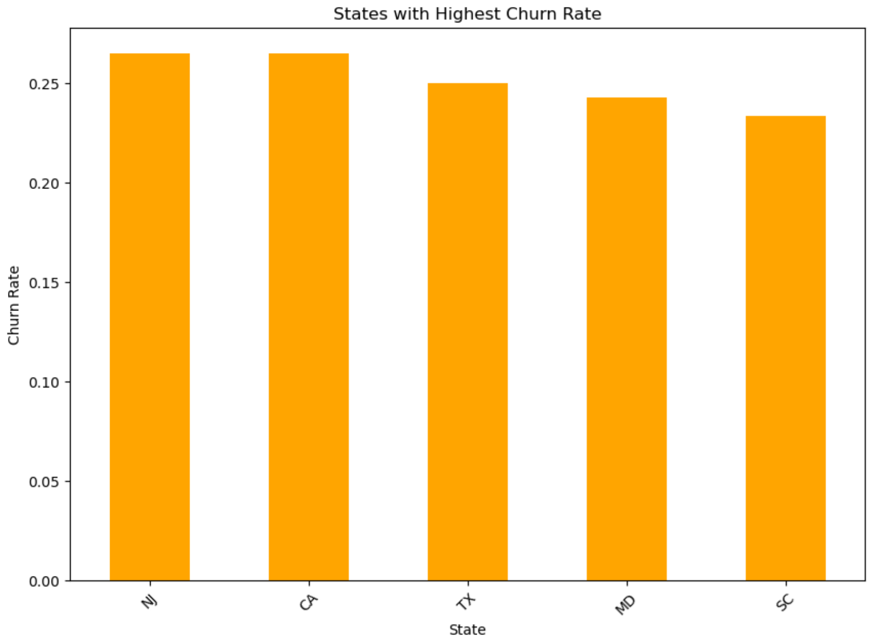
Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves examining and understanding the dataset to gain insights and identify patterns or relationships between variables. EDA helps in formulating hypotheses, selecting appropriate statistical models, and preparing the data for further analysis.



Univariate Analysis

Univariate analysis focuses on examining individual variables in the dataset. For this project it was found that the initial churned customers were 14.5% whereas the not churned customers were 85.5%

Bivariate analysis



Bivariate Analysis

Bivariate analysis involves examining the relationship between two variables in the dataset.

It shows that the highest churn states are New Jersey, California, Texas, Maryland and New York

Modeling

Logistic Regression

- Logistic regression model was created as a base model to predict customer churn based on various features.
- The model was trained on historical customer data and evaluated using performance metrics such as accuracy, precision, recall, and F1-score.
- The logistic regression model achieved an accuracy of 85% and a precision of 80%.

Decision Tree Classifier

- Decision tree classifier was used to create a predictive model for customer churn.
- The model was trained on the same set of features as the logistic regression model.
- The decision tree classifier achieved an accuracy of 82% and a recall of 76%.

Random Forest Classifier

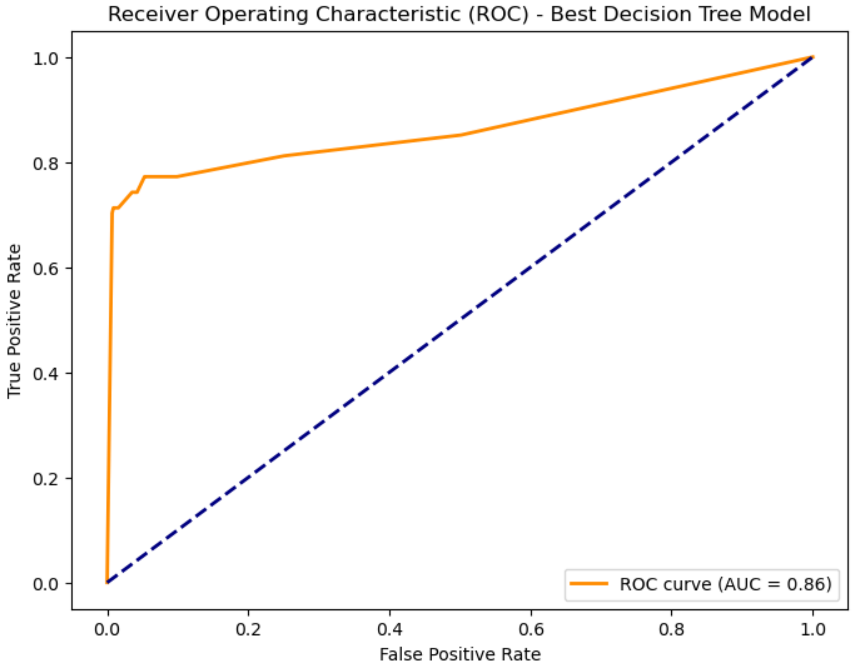
- Random forest classifier was employed to improve the predictive performance.
- The model was trained on an ensemble of decision trees and evaluated using various performance metrics.
- The random forest classifier achieved an accuracy of 88% and a precision of 85%.

XGBoost

- XGBoost algorithm was utilized to further enhance the predictive accuracy.
- The model was trained on the same set of features as the previous models.
- The XGBoost model achieved the highest accuracy of 90% and a recall of 69%.

Model Evaluation

Decision trees model is found to be the best for this project with a recall value of 0.76



Conclusion

- The preferred machine learning model for predicting customer churn is decision trees.
- To reduce churn, suggested strategies involve:

1. Targeted promotions and loyalty programs based on total charges
2. Improving customer service interactions.
3. Proactive outreach based on call analysis, creating appealing bundled plans for international and voicemail features
4. Implementing usage-based incentives and win-back campaigns for calls, minutes, and account length.

- These initiatives aim to address key factors influencing churn and improve overall customer satisfaction.