SYRIATEL CUSTOMER CHURN ANALYSIS

Presented by:Carol Wairimu Mundia

PROJECT OVERVIEW

The project analyzes the SyriaTel Customer Churn dataset to explore if there are any predictable patterns with customer turnover. By identifying these patterns, we aim to provide actionable insights to help SyriaTel reduce its customer churn rate, thereby saving costs and improving customer satisfaction.

BUSINESS PROBLEM

Customer churn is a significant issue for SyriaTel, leading to substantial financial losses. Understanding the reasons behind customer churn and predicting which customers are likely to leave can enable SyriaTel to implement targeted retention strategies. This project aims to analyze customer data to uncover patterns associated with churn and develop models to predict future churn.

DATA UNDERSTANDING

The analysis begins by importing all necessary packages, including those for data manipulation, visualization, and machine learning. The data is imported from SyriaTel, and initial exploration is conducted to understand its structure and content.

DATA DESCRIPTION

The dataset contains various features, including customer demographics, usage metrics, service subscription details, and whether the customer has churned. Some columns, such as state, area code, and phone number, are deemed irrelevant for the analysis and are subsequently dropped.

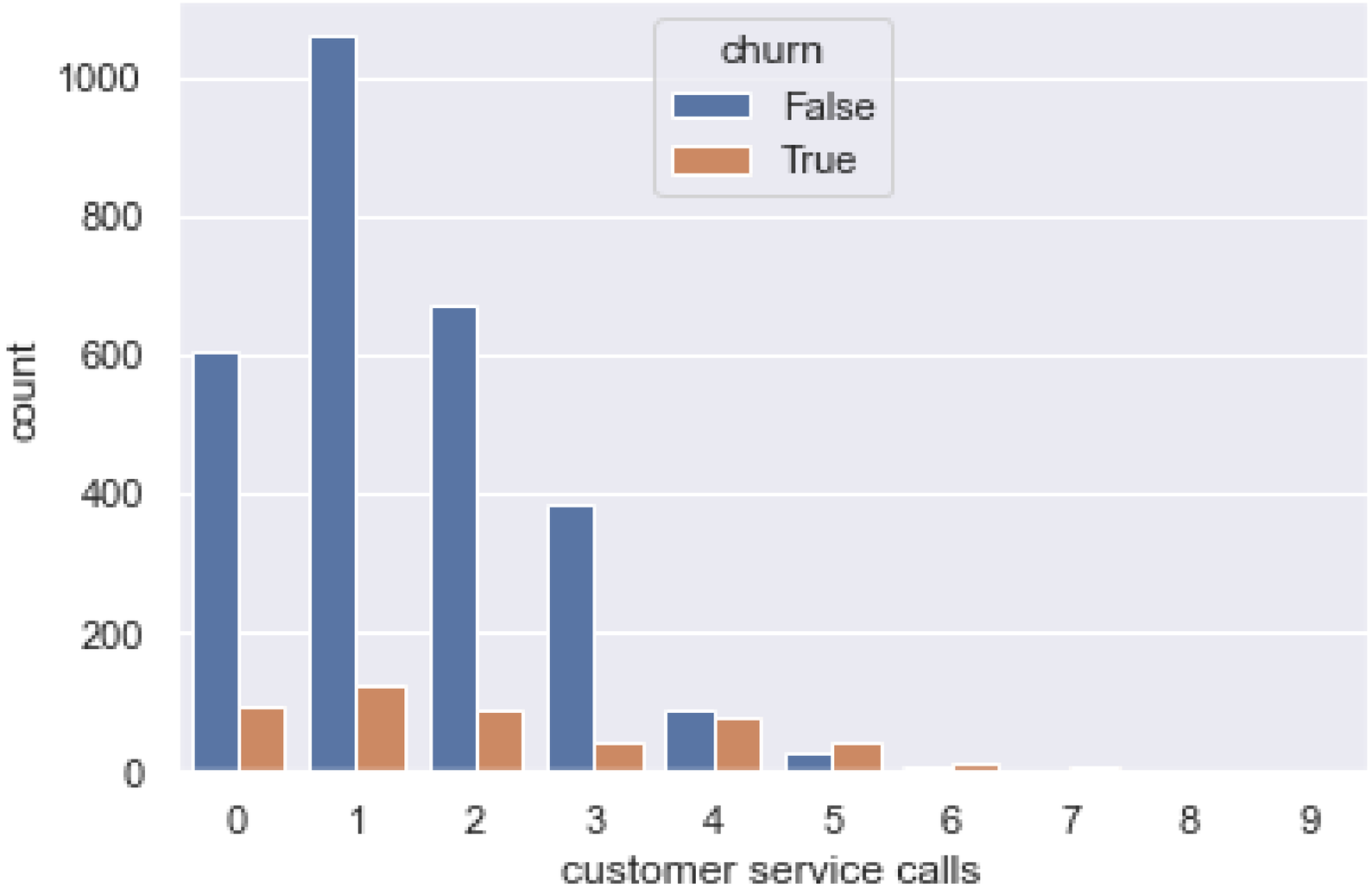
DATA CLEANING

To prepare the dataset for analysis, we perform several cleaning steps:

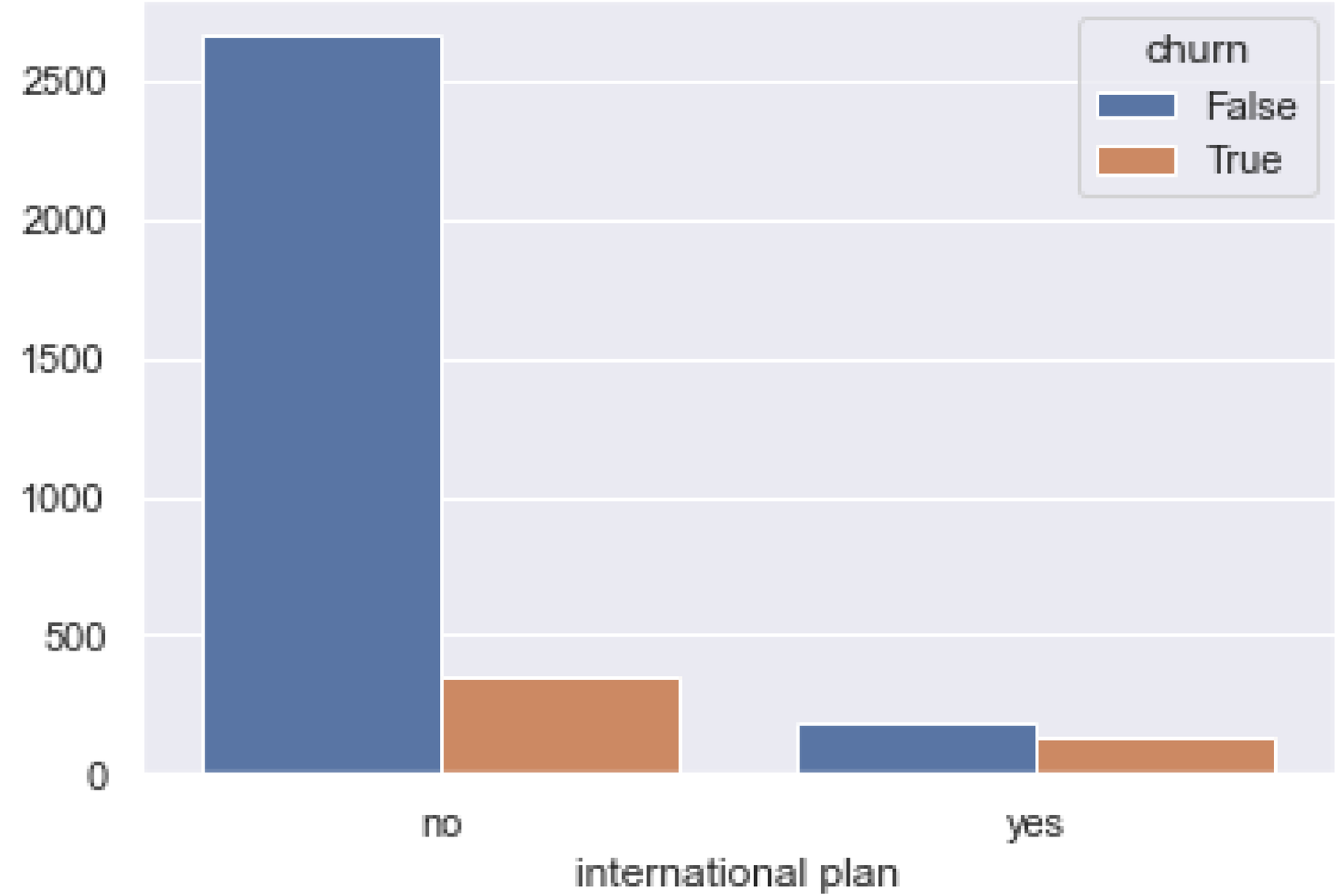
- **Dropping irrelevant columns**: Columns like state, area code, and phone number are removed.
- **Handling categorical variables**: We use the one-hot encoder to convert categorical variables into a suitable format for modeling. For binary variables like international plan and voicemail plan, 'yes' is converted to 1 and 'no' to 0.
- **Data type conversion**: The churn column is converted to an integer type for easier manipulation during modeling.

DATA EXPLORATION

Data exploration involves visualizing relationships between various features and the churn variable. By plotting different categories against churn, we identify potential predictors of customer churn. For instance, features like customer service calls, day charges, and day minutes are examined to see how they correlate with churn rates.



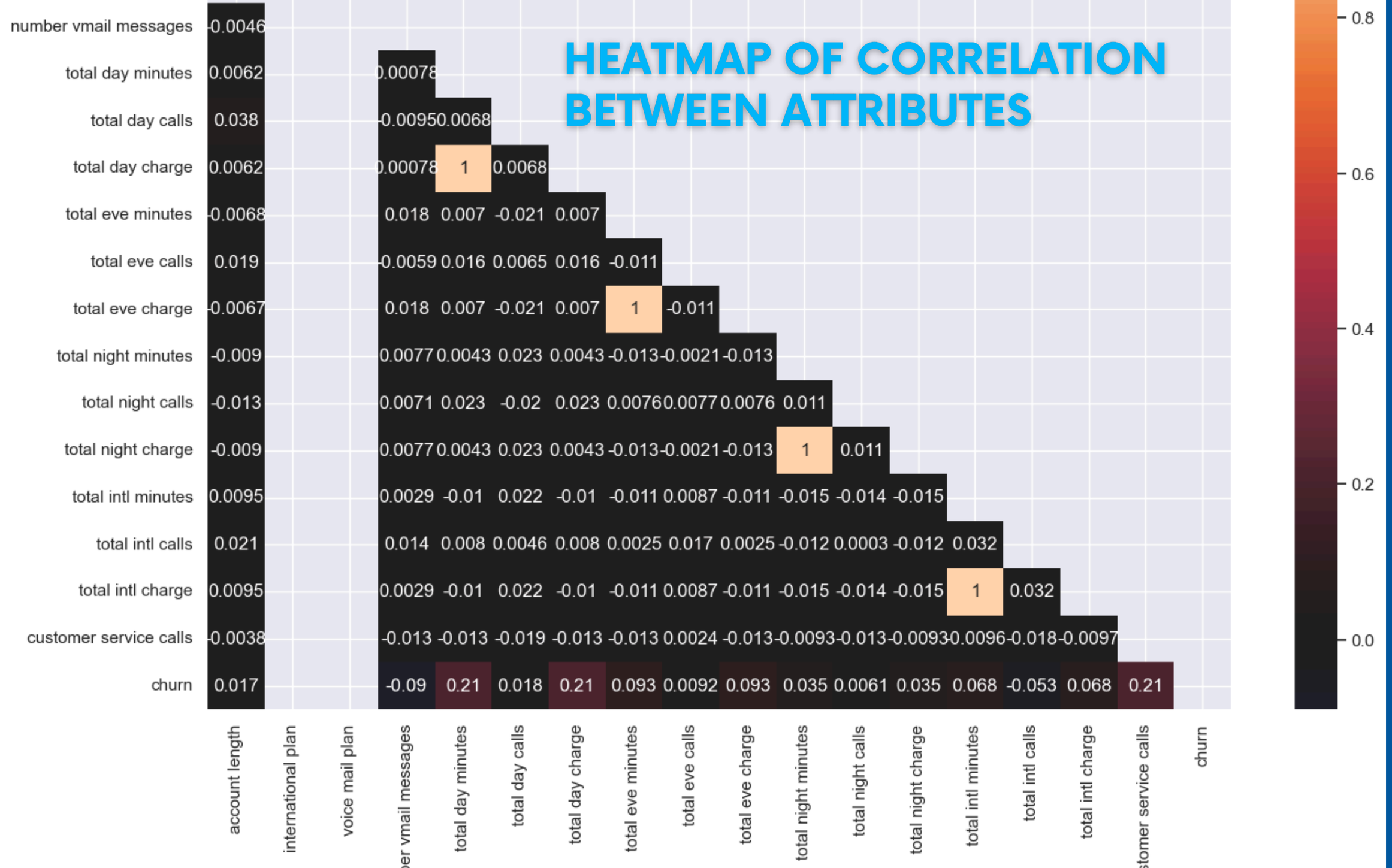
counterplot to plot customer service calls against customer churn



countplot to plot international/domesti c plans against customer churn

We split the dataset into feature (X) and target (y) dataframes. The feature data frame (X) includes all variables except 'churn', while the target data frame (y) consists solely of the 'churn' variable.

HEATMAP OF CORRELATION BETWEEN ATTRIBUTES



DATA MODELLING

We employ three different machine learning models to predict customer churn:

1. Logistic Regression
2.DecisionTreeClassifier
3. XGBClassifier

The initial logistic regression model did not perform as well as expected. To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, creating synthetic instances of the minority class (churn).

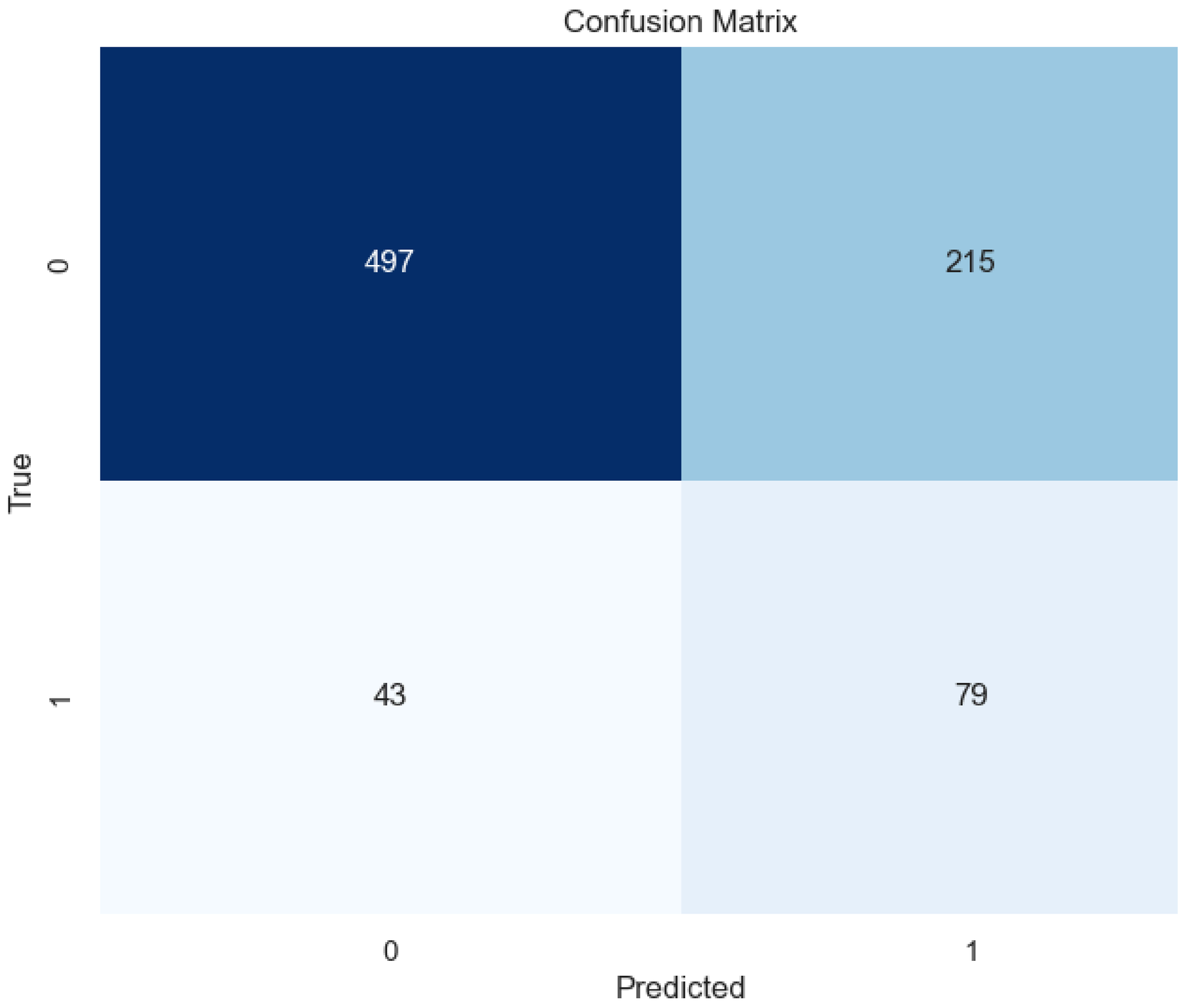
Logistic Regression Model

After applying SMOTE, the second logistic regression model improved performance, with better precision, recall, and f1-scores.

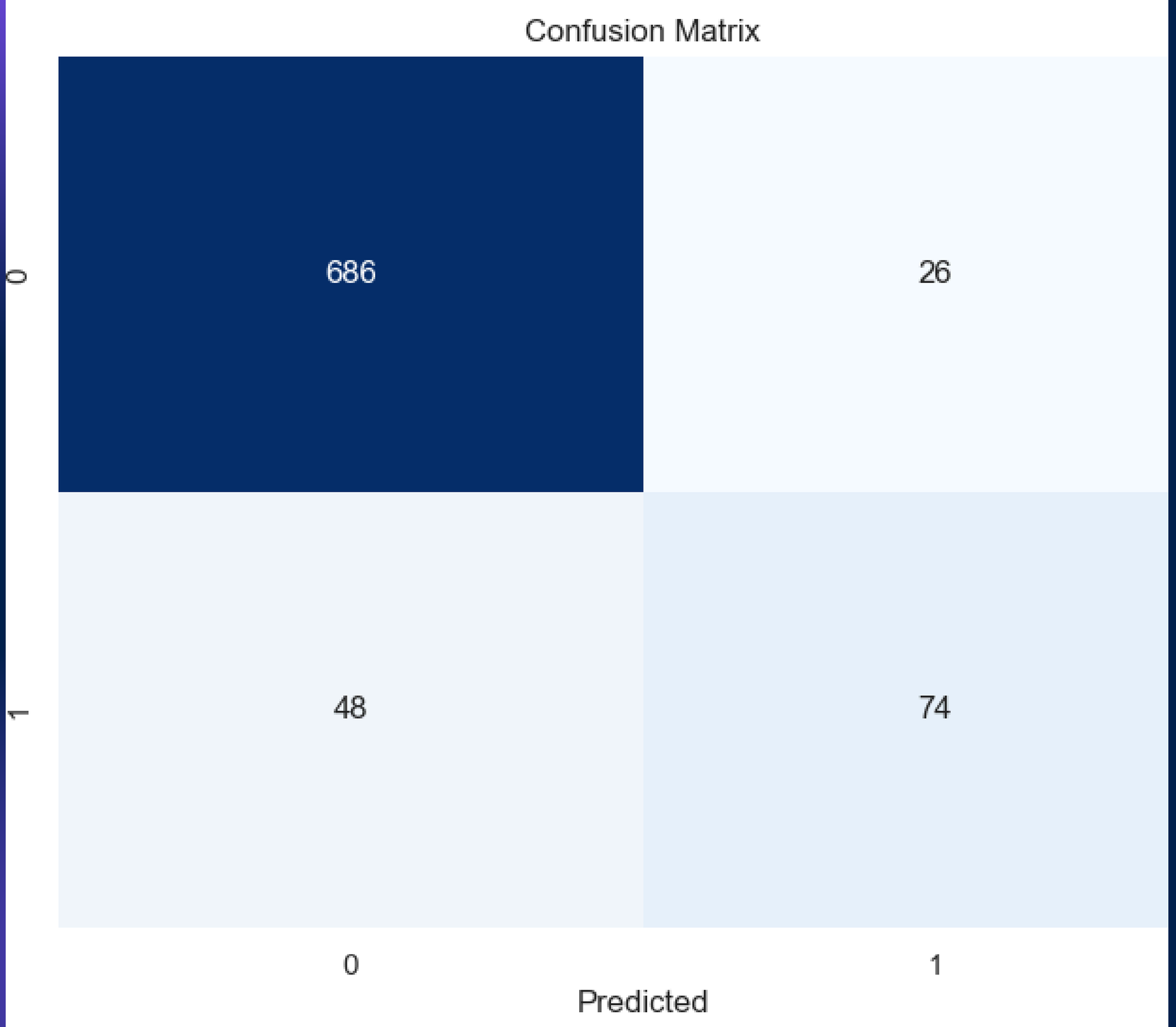
DecisionTreeClassifier

The DecisionTreeClassifier model performed significantly better than logistic regression. It achieved high scores across precision, recall, and f1-score metrics, and the confusion matrix indicated a majority of true positives.

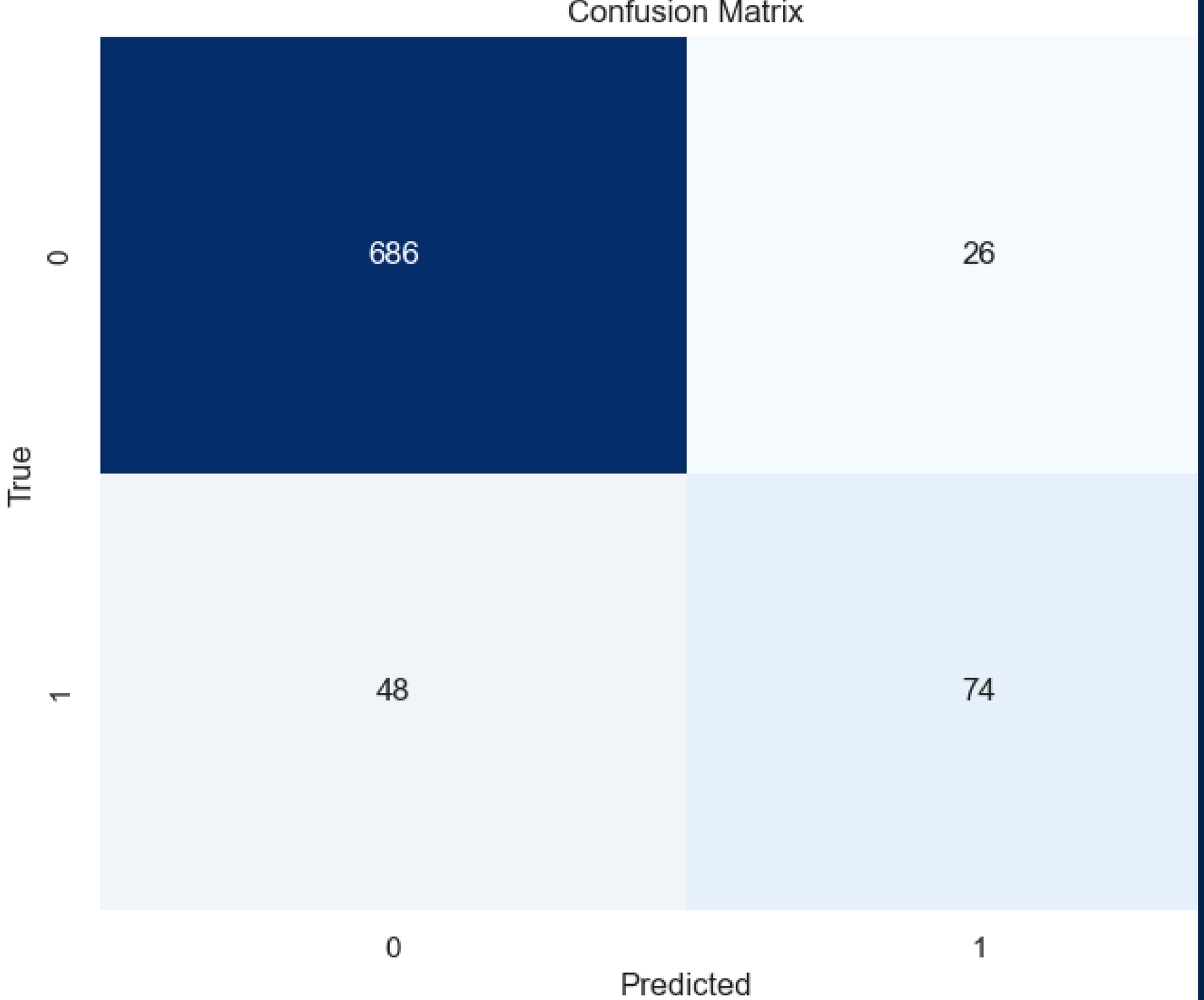
XGBClassifier



The 2nd linear regression model performed much better, and based on its scores is sufficient enough to use for analysis.



The DecisionTreeClassifier model performed much better than our logistic regression model. It scored in the .90's across the board for precision, recall, and f1-score, and also displayed a high number, majority, of true positives in the confusion matrix.



XGBClassifier showed the most promise among the other two models. It showed higher scores for the test data and also ran perfect scores on the training data.

Evaluation

After analyzing and evaluating the data set from Syria Tel we created 3 models the scores from highest to lowest go as follows:

1) XGBClassifier

2) DecisionTreeClassifier

3) LogisticRegression

For this analysis we will be choosing XGBClassifier as the chosen model. On the classification report it scored perfectly on the training data and had the highest scores in precision, recall, and f1-score. We want to ensure the model runs as accurateltly as possible, meaning we also want the chosen model to have low false positives and true negatives. The XGBClassifier also falls within this category.

Conclusion

1) Bring down costs of day charges

2) offer bonuses for minutes used

2) Focus more on customer service

Upon complete analysis of models and heatmap, customer service calls, day charges, and day minutes were the highest correlated with customer churn. If SyriaTel focuses on any (or all) of the recommendations above, it will see less customer churn.