CUSTOMER CHURN PREDICTION FOR SyriaTel TELECOM

Understanding and predicting customer churn is crucial for telecom companies like SyriaTel to maintain profitability and market share.

Project Overview

This project aims to predict customer churn, identify influencing factors, and recommend retention strategies for SyriaTel.

1

Business Understanding

Context, problem, objectives, and success metrics.

2

Data Understanding

Dataset overview and initial observations.

3

Data Preparation

Feature correlation and selection.

4

Exploratory Data Analysis

Target and feature distributions.

5

Modelling & Evaluation

Logistic Regression and Decision Tree models.

6

Conclusion & Recommendations

Key findings and actionable insights.

Business Context & Objectives

Problem Statement

SyriaTel experiences customer churn, impacting profitability. Acquiring new customers is more expensive than retaining existing ones.

The company needs to understand churn patterns and identify at-risk customers for retention.

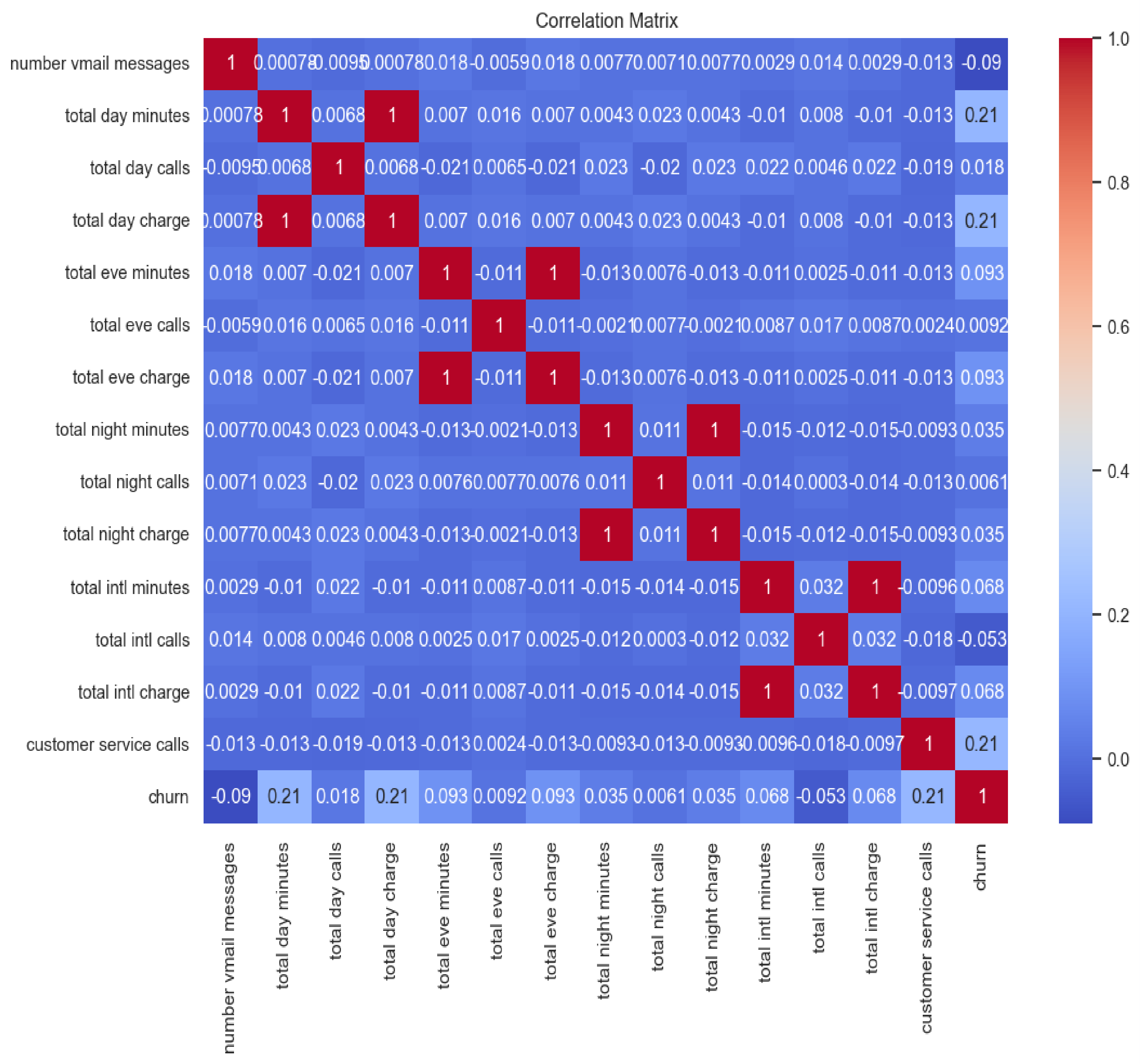
Project Objectives

- Predict customer Churn - (churn = True) or not (churn = False)
- Identify factors influencing churn.
- Advise on key retention strategies.

Success Metrics

- Recall (catch churners).
- Precision (avoid false positives).
- F1 Score (balance between Precision and Recall).
- ROC-AUC (How well model separates the two).

Data Understanding & Preparation



Dataset Overview

- 3,333 customer records from Kaggle.
- 21 features inclusive of demographics, usage, plans, support.
- Target variable: churn (True/False).
- No missing values.

Feature Selection Using Correlation

1. Highly correlated features were dropped to avoid redundancy and multicollinearity.

Dropped Feature	Highly Correlated With
`total day charge`	`total day minutes`
`total eve charge`	`total eve minutes`
`total night charge`	`total night minutes`
`total intl charge`	`total intl minutes`

Data Understanding & Preparation

```
--------------------------------------------
ANOVA for numeric
--------------------------------------------
customer service calls    151.767013
total day minutes         146.350785
total eve minutes          28.932577
number vmail messages      27.035912
total intl minutes         15.583468
total intl calls            9.327945
total night minutes         4.201496
total day calls             1.135412
total eve calls             0.283994
total night calls           0.125631
dtype: float64
--------------------------------------------
Chi2 for categoricals
--------------------------------------------
international plan    203.244178
voice mail plan        25.156959
dtype: float64
--------------------------------------------
```

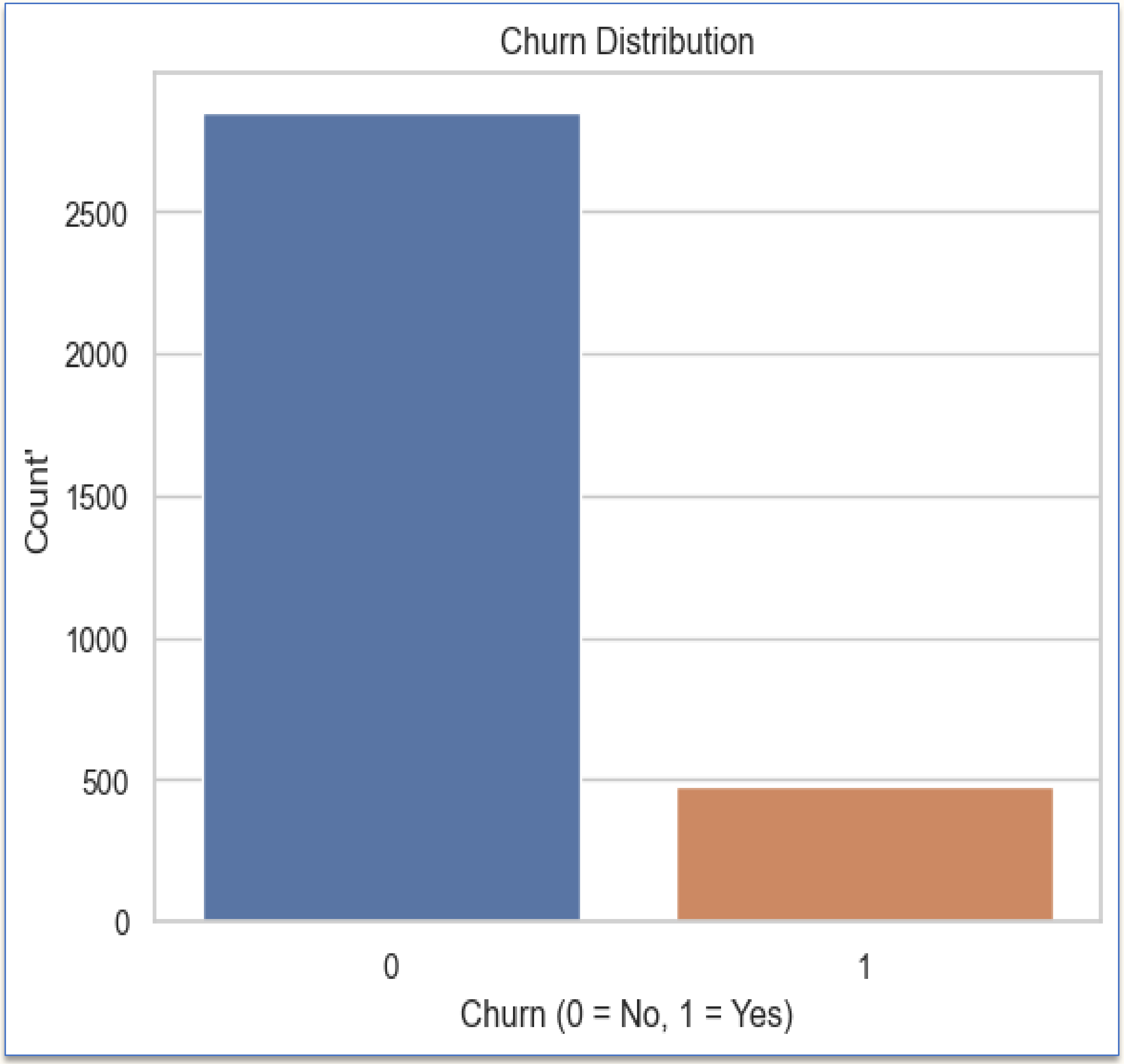
Feature Selection USING Statistical Metrics Evaluation - SelectKBest

- ***Chi-square*** was used for categorical variables to test the independence between features and the target. ***We drop those below 10 score***
- ***ANOVA F-test*** was used for numeric variables to check whether feature means differ significantly across churn categories. ***We drop those less than 2***

Observations

- ***International plan*** showed the strongest association with churn in categorical features, but Voice mail plan was also significant with 25
- ***Customer service calls*** and total day minutes had the highest F-scores, indicating strong influence on churn.
- ***Total day calls, total eve calls, total night calls*** were dropped by ANOVA F-test
- Chi-square did not drop any as it deemed International plana and Voice male plan as significant

Exploratory Data Analysis (EDA)



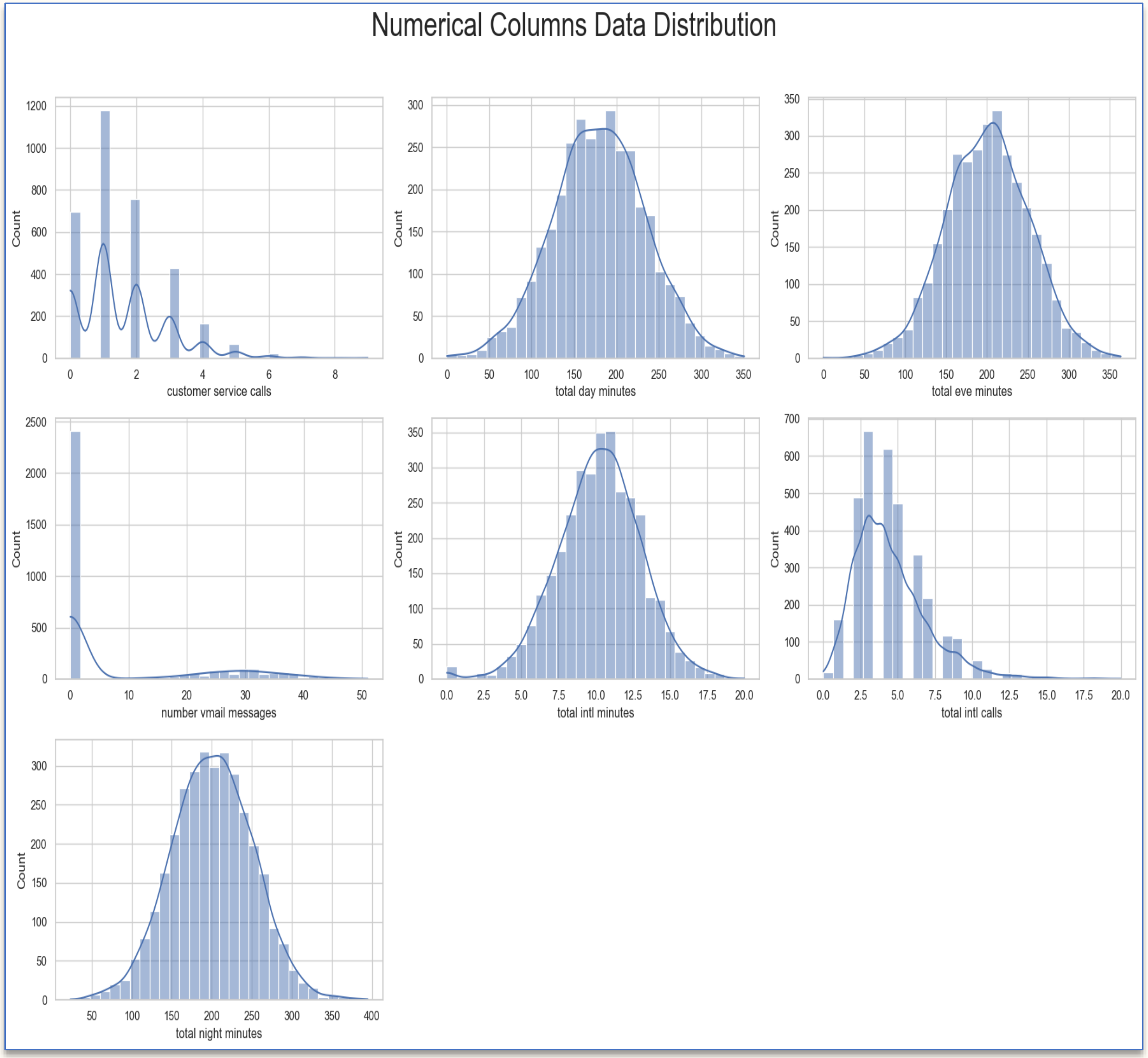
Target distribution to check balance

The target variable is imbalanced, with only 14% churners.

This will be addressed by:

- Balancing **class weights** in the regression model
- Using **SMOTE**(Synthetic Minority Oversampling Technique) to balance the classes to help model understand the minority.

Exploratory Data Analysis (EDA)



Numerical Columns Data Distribution

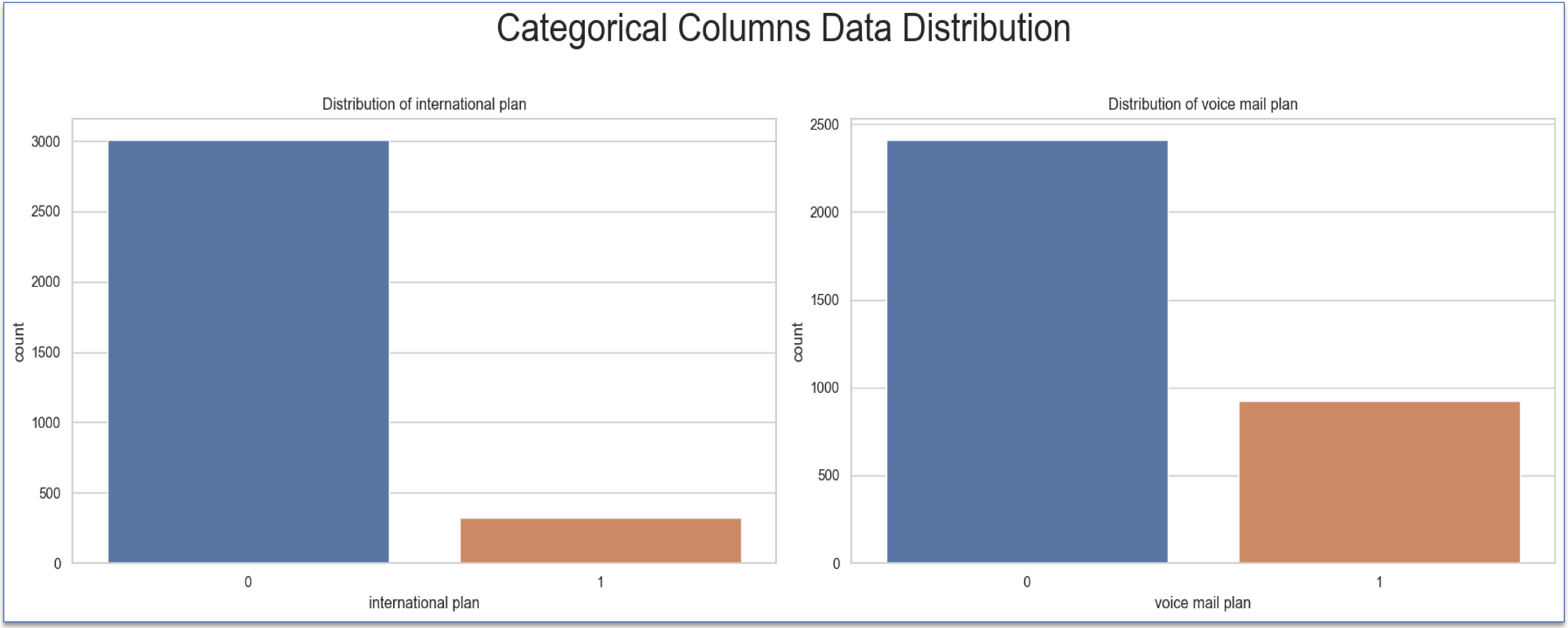
Observations from Numerical Feature Distributions

- **Customer Service Calls**: Most customers made 1–3 service calls, Frequent customer service calls may signal dissatisfaction.
- **Total Day Minutes**: Follows a normal distribution, around 180–200 minutes.
- **Total Evening Minutes**: Slight Right-skewed, peaks around 200 minutes.
- **Number Vmail Messages**: Highly right-skewed with most customers having 0 messages, potential low predictive measure.
- **Total International Minutes**: Normally distributed around 10 minutes.
- **Total International Calls**: Most customers made 3–5 international calls; skewed right. May indicate Niche usage patterns
- **Total Night Minutes**: Near-normal distribution centered around 200 minutes.

Exploratory Data Analysis (EDA)

Categorical Columns Data Distribution Observations

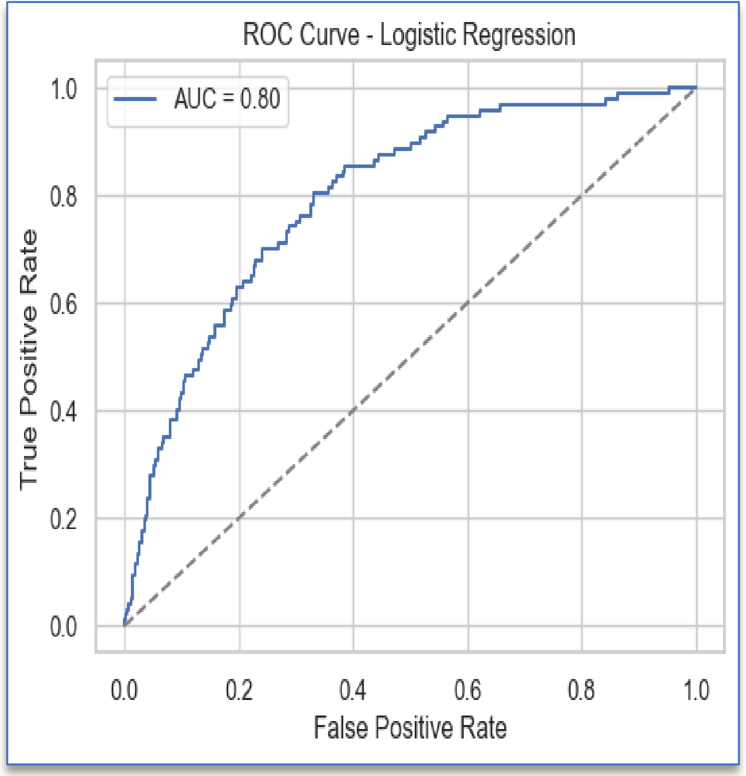
- Most customers have no international plans
- Churn customers are less

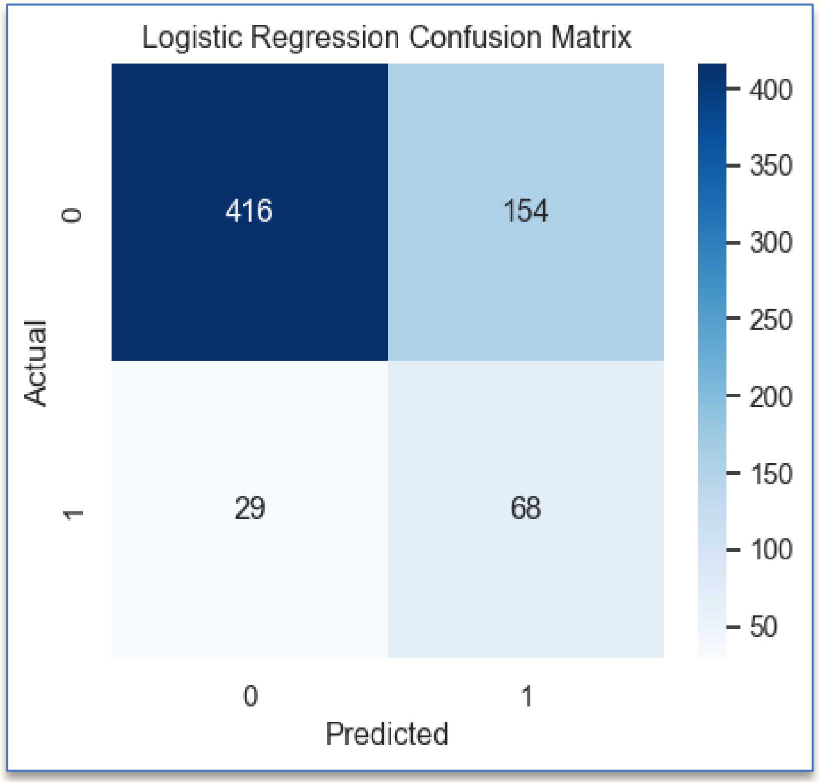


Modelling: Logistic Regression

A logistic regression model was built with preprocessing, SMOTE for class balancing, RFECV for feature selection, and GridSearch for tuning.

```
Best Parameters: {'classifier__C': 0.01, 'classifier__max_iter': 100}
F1 Score on Test Set: 0.42633228840125387

 Classification Report:
               precision    recall  f1-score   support

           0       0.93      0.73      0.82       570
           1       0.31      0.70      0.43        97

    accuracy                           0.73       667
   macro avg       0.62      0.72      0.62       667
weighted avg       0.84      0.73      0.76       667

 ROC-AUC Score: 0.797069994574064
```





Observations

- Precision (non-churners): 93% (excellent).
- Precision (churners): 31% (low, high false positives).
- Recall (non-churners): 73%.
- Recall (churners): 70% (relatively good).
- F1 Score (churners): 0.43 (struggles to identify).
- Accuracy: 73%.
- ROC-AUC: 0.797 (good separation ability).

➢ *The model performs well in identifying non-churners with high Accuracy (73%) and good ROC-AUC 0.80,indicating strong separation capability between churners and non-churners.*

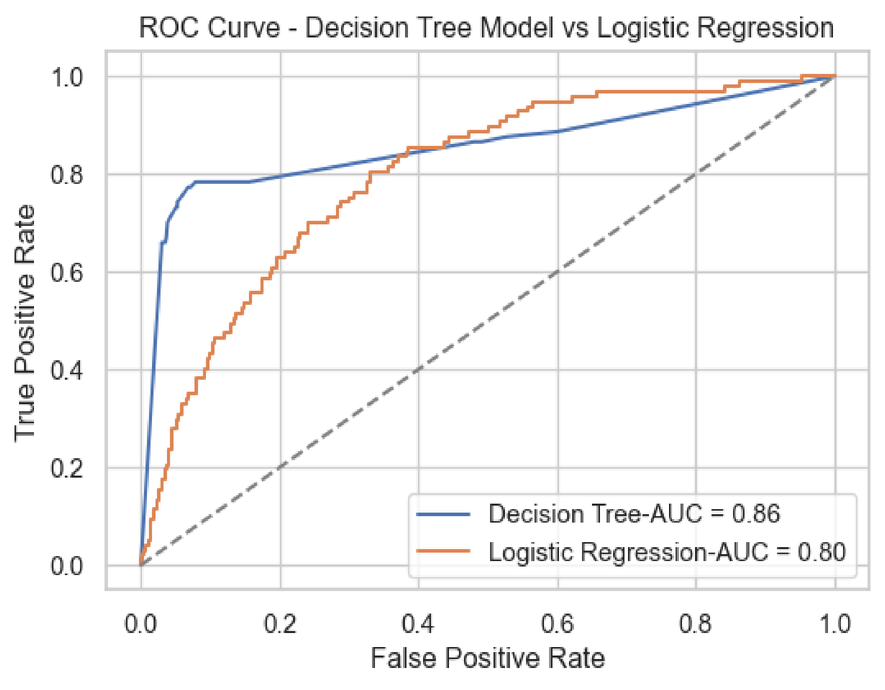
➢ *However, it struggles with precision for churners (31%), meaning it often wrongly predicts churn. Further tuning or trying other models*

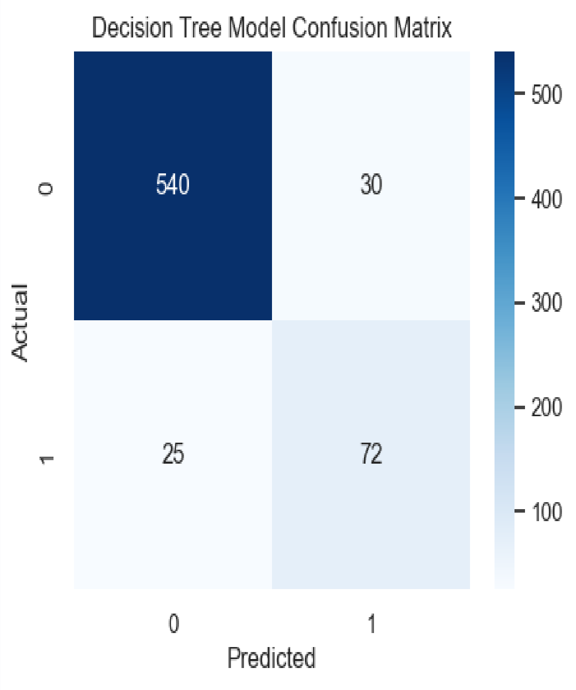
I will therefore try another tree-based model “Decision tree” and compare the findings

Modelling: Decision Tree Modeling

Model was built with preprocessing, SMOTE for class balancing and GridSearch for tuning.

```
Best Parameters: {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 10}
F1 Score on Test Set: 0.7236180904522612

 Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.95      0.95       570
           1       0.71      0.74      0.72        97

    accuracy                           0.92       667
   macro avg       0.83      0.84      0.84       667
weighted avg       0.92      0.92      0.92       667

 ROC-AUC Score: 0.8558690540784952
```





Observations and Comparison with Logistic regression

Precision (non-churners):
Decision Tree: 96% | Logistic Regression: 93% → (Slight improvement)

Precision (churners): Decision Tree: 71% | Logistic Regression: 31% → (Significantly fewer false positives)

Recall (non-churners): Decision Tree: 95% | Logistic Regression: 73% → (Much better detection)

Recall (churners): Decision Tree: 74% | Logistic Regression: 70% → (Slight gain)

F1 Score (churners): Decision Tree: 0.72 | Logistic Regression: 0.43 → (Better balance of precision & recall)

Accuracy: Decision Tree: 92% | Logistic Regression: 73% → (Significant improvement)

ROC-AUC: Decision Tree: 0.856 | Logistic Regression: 0.797 → (Stronger class separation)

The Decision Tree ROC curve is closer to the top-left, showing a better trade-off between true positives and false positives than Logistic Regression. While Logistic Regression performs reasonably well, the Decision Tree offers superior churn classification.

The results demonstrate the Decision Tree capability to correctly identify churn customers while maintaining high reliability for loyal customers.

Model Comparison

The Decision Tree model significantly outperformed Logistic Regression across all key metrics.

96%

DT Precision (Non-Churn)

vs. 93% for Logistic Regression

71%

DT Precision (Churn)

vs. 31% for Logistic Regression

95%

DT Recall (Non-Churn)

vs. 73% for Logistic Regression

74%

DT Recall (Churn)

vs. 70% for Logistic Regression

92%

DT Accuracy

vs. 73% for Logistic Regression

0.856

DT ROC-AUC

vs. 0.797 for Logistic Regression

The Decision Tree provides more reliable identification of churners, crucial for targeted business action.

Conclusion & Recommendations

The Decision Tree model's capability to correctly identify churn customers makes it the superior choice.

Deploy Decision Tree Model

Use as the core engine to score and flag customers likely to churn.

Prioritize High-Risk Customers for campaigns

Focus retention resources on identified high-risk churners.

Integrate Churn Scores into CRM

Embed predictions into dashboards for real-time monitoring and faster decisions.

Monitor Heavy Daytime Users & Intl. Plan Holders

Offer custom bundles or cheaper rates to retain these high-usage or globally connected users.

Prioritize High Call Volume Customers

Flag customers with many service calls for priority support or follow-up.

By implementing these recommendations, SyriaTel can proactively reduce customer churn and enhance customer loyalty.

Thank You!

We value your input! Please take a moment to provide feedback

on the analysis using the following questions:

1. Did the analysis of the data resonate with the company’s goal?

2. Any other insight on the data that you find useful?

"Thank you for your valuable feedback! Your input will help us

refine the analysis.

Contacts

Name: Erastus Kaiba Njuguna

Email: Erastusnjuguna24@gmail.com