SyriaTel Ltd

Customer Retention STRATEGY

Predictive Analysis for Customer Churn Prevention at SyriaTel

Date: 23-Oct-2023

Prepared by: Wayne

Table of CONTENTS

01 Introduction

02 Data Analysis and Preprocessing

03 Model Building and Evaluation

04 Conclusions & Recommendations

05 Q&A and Discussion

INTRODUCTION

Overview

- The aim of this project was to tackle a significant challenge faced by SyriaTel's Customer Retention Department: the reduction of customer churn.
- The project's objective was to develop a predictive model capable of identifying customers at risk of leaving.
- This proactive approach is designed to enhance customer satisfaction and reduce revenue loss due to churn.

Project
OBJECTIVES

The primary objective of this project is to develop a predictive model to help the Customer Retention Department at SyriaTel reduce customer churn. By analyzing historical data and customer behaviors, we aim to build a model that can identify customers at risk of leaving the company.

Churn Prediction

Build a classification model to predict customer churn based on historical data and customer behavior patterns.

Model Performance:

Assess the performance of the predictive model using appropriate classification metrics. Understand how well the model can identify customers likely to churn.

Model Interpretation:

Provide insights into why the model makes certain predictions. Understand the factors contributing to churn risk and communicate these insights to the Customer Retention Department.

Recommendations

Offer actionable recommendations to the Customer Retention Department based on the model's findings. Suggest strategies for retaining at-risk customers and reducing churn.

THE DATA

OVERVIEW

- The dataset, *"Churn in Telecom's dataset*", was sourced from Kaggle.
- It consists of 3,333 rows and 21 columns.
- Target Variable:
  - The target variable for this project is "churn," which is a binary variable indicating whether a customer has churned (1) or not (0).

KEY DATA FIELDS

- **State**: The state in which the customer resides.
- **Account Length:** Duration of the customer's account with SyriaTel.
- **Area Code:** The area code associated with the customer's phone number.
- **International Plan:** Whether the customer has an international calling plan (categorical: "yes" or "no").
- **Voice Mail Plan:** Whether the customer has a voicemail plan (categorical: "yes" or "no").
- **Total Day Minutes:** Total minutes of usage during the day.
- **Total Day Calls:** Total number of calls made during the day.
- **Total Day Charge:** Total charges incurred during the day.
- **Total Eve Minutes:** Total minutes of usage during the evening.
- **Total Eve Calls:** Total number of calls made during the evening.
- **Total Eve Charge:** Total charges incurred during the evening.
- **Total Night Minutes:** Total minutes of usage during the night.
- **Total Night Calls:** Total number of calls made during the night.
- **Total Night Charge:** Total charges incurred during the night.
- **Total Intl Minutes:** Total minutes of international usage.
- **Total Intl Calls:** Total number of international calls.
- T**otal Intl Charge:** Total charges incurred for international usage.
- **Customer Service Calls:** The number of customer service calls made by the customer.
- **Number Vmail Messages:** The number of voicemail messages received by the customer.

DATA CLEANING & EXPLORATION

Data Cleaning:

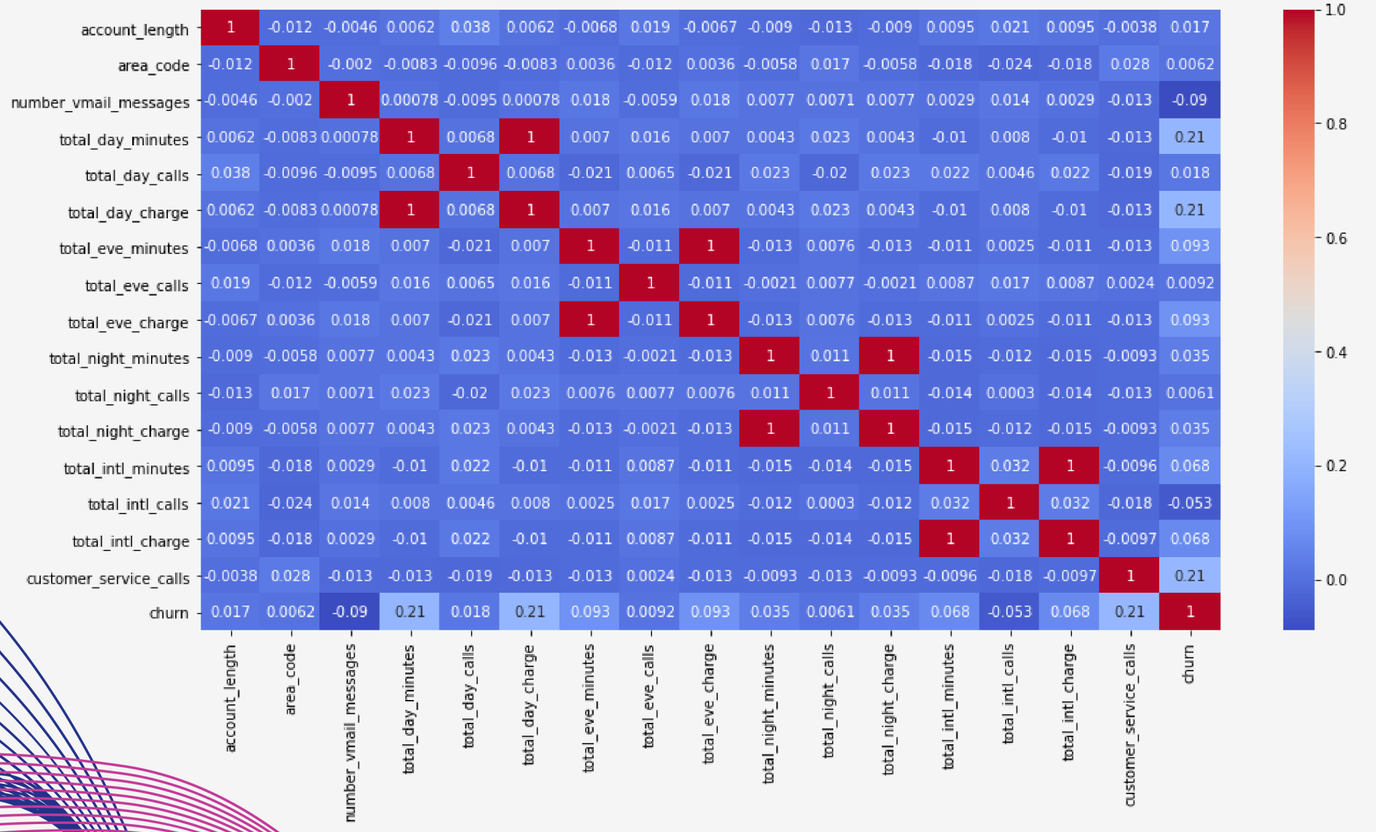
- Initial inspection of the dataset showed that there were no missing values
- The phone number column was dropped as it was deemed to have no predictive importance
- To reduce multicollinearity one from highly correlated pairs was retained

Data Exploration:

- The distribution of key features was visualized to understand the dataset's characteristics.
- The imbalance in the target variable "churn" was discovered, emphasizing the need for specific strategies.

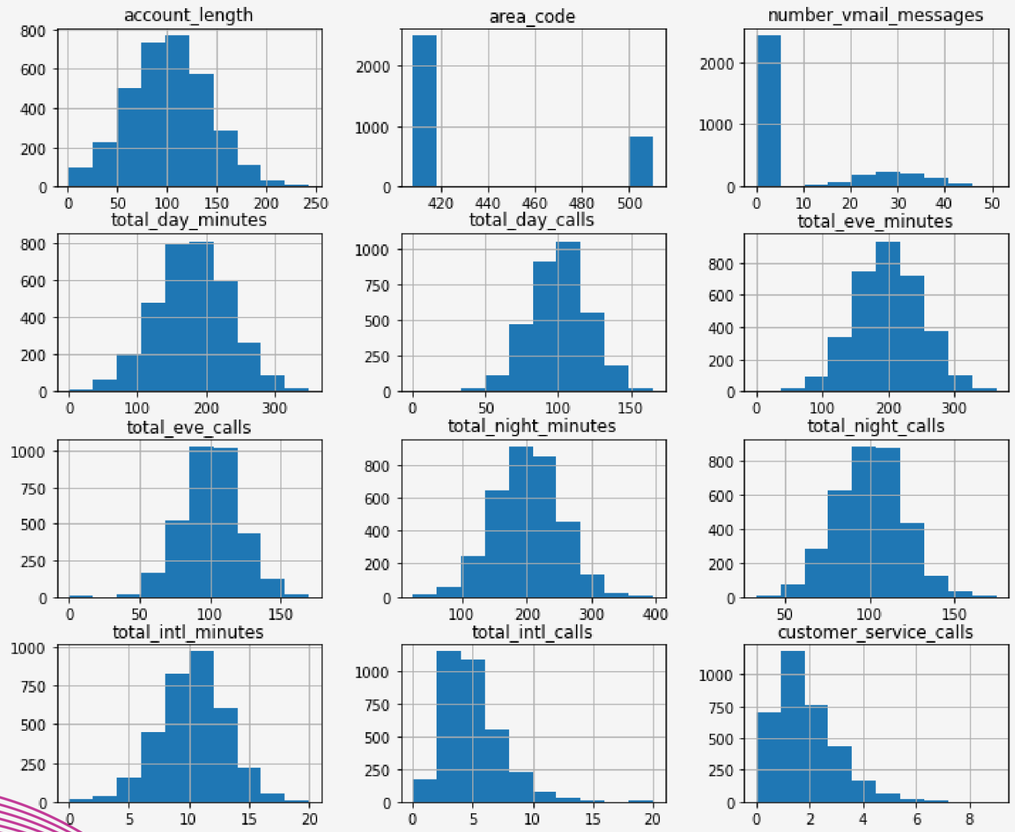
Data Exploration:

- ***Class Imbalance:*** The dataset exhibited a significant class imbalance, with a majority of non-churn (class 0) instances (2850) and a minority of churn (class 1) instances (483).
- Due to the small size of the dataset, the chosen approach to address this imbalance was class weighting



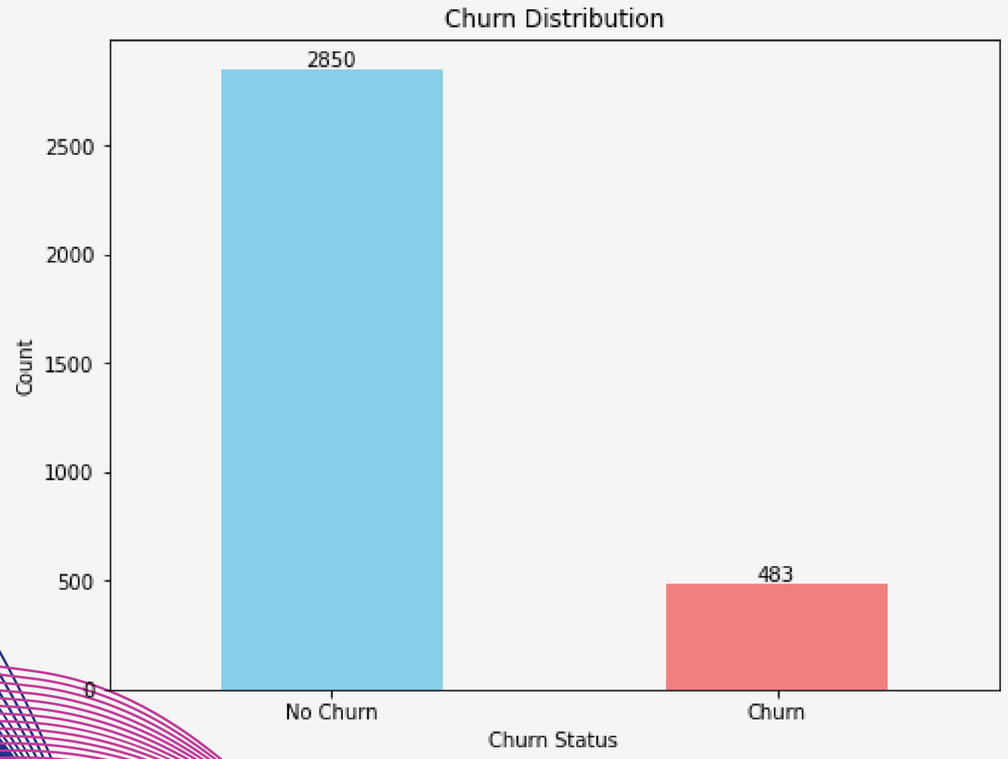
Data EXPLORATION

1. Highly Correlated Features:
   - **total_day_charge** and **total_day_minutes** exhibited a perfect correlation.
   - **total_eve_charge** and **total_eve_minutes** show perfect correlation.
   - **total_night_charge and total_night_minutes** are highly correlated.
   - **total_intl_charge** and **total_intl_minutes** have a perfect correlation.
2. This perfect correlation suggests that one feature in each pair is derived from the other or contains identical information.



Data EXPLORATION

- **Feature Distributions:**
  - The majority of numeric features in the dataset exhibit normal distributions, which is a positive aspect of modeling.
- **Skewed Features:**
  - Two features, **'total_intl_calls'** and **'customer_service_calls,'** are left-skewed. This skew indicates that the majority of customers make fewer international and customer service calls. When modeling these features, it was important to consider transformations to address the skewness.



Data EXPLORATION

- **Churn Distributions:**
  - Class Imbalance: The dataset exhibited a significant class imbalance, with a majority of non-churn (class 0) instances (2850) and a minority of churn (class 1) instances (483).
- **Consideration for Handling Churn Imbalance:**
  - To address the class imbalance issue in our dataset, we adopted the strategy of class weighting. Due to the relatively small dataset, class weighting provides an effective method to balance the impact of each class during model training.

PREPOCESSING

Data Splitting:

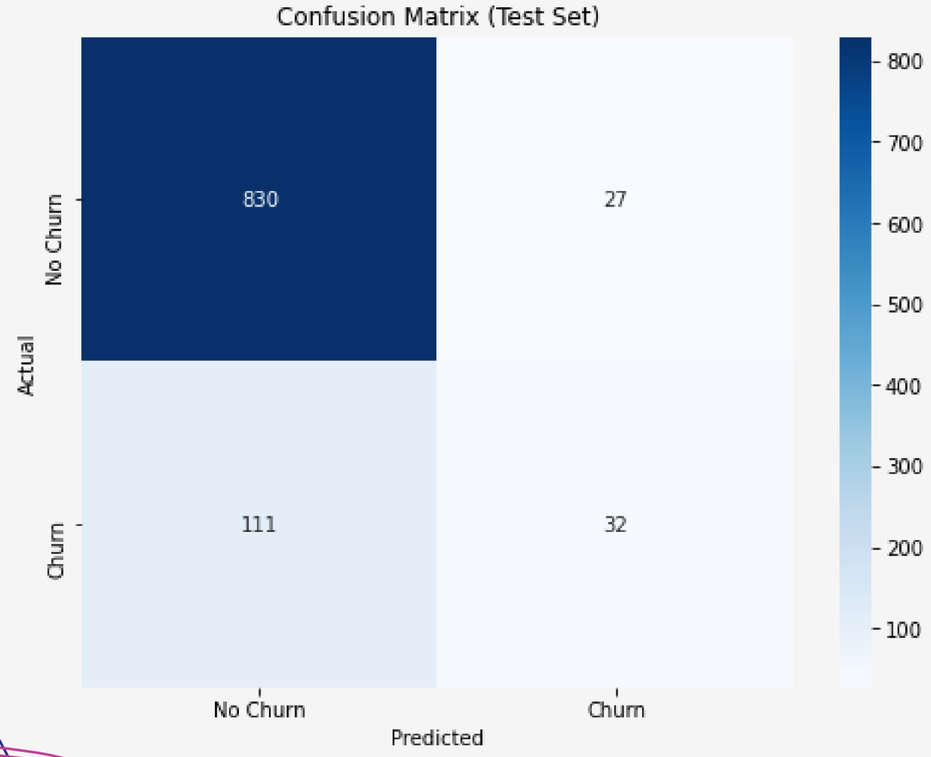
- We divided the dataset into training and testing sets with a 70/30 split. This allowed us to train our models on one portion and evaluate their performance on another.

Numerical Data Transformation:

- We applied Standard Scaler to our numerical features, ensuring that their values had a mean of 0 and a standard deviation of 1. This step is crucial for models like SVM that are sensitive to feature scales.

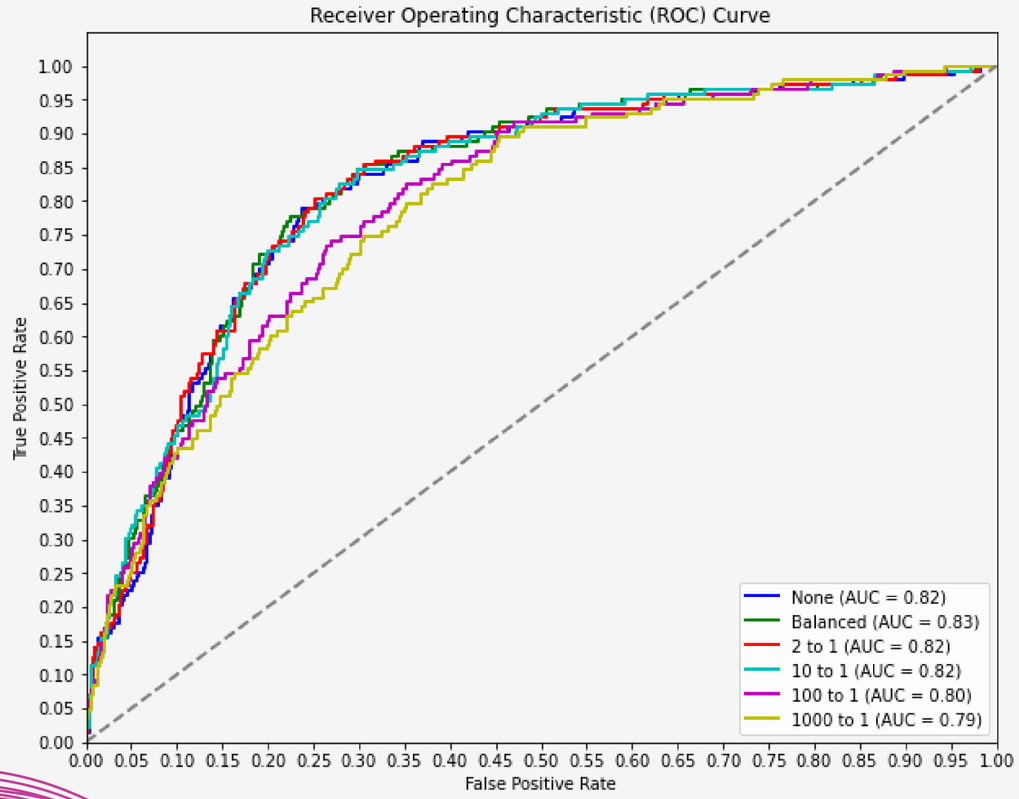
Categorical Data Encoding:

- We employed one-hot encoding for our categorical variables, enabling the models to work with these features effectively.



Base
LOGREG MODEL

- The first model built was a base Logistic Regression model, employing default scikit learn settings.
- Although the base model had a reasonable testing accuracy of 0.862, the recall for churned customers was 0.22, indicating a challenge in identifying customers likely to churn.
- A low churn recall is a concern as it means we are missing many actual churn cases, affecting our ability to proactively address customer retention.
- The suspected source of the low recall in the base Logistic Regression model is the class imbalance issue.

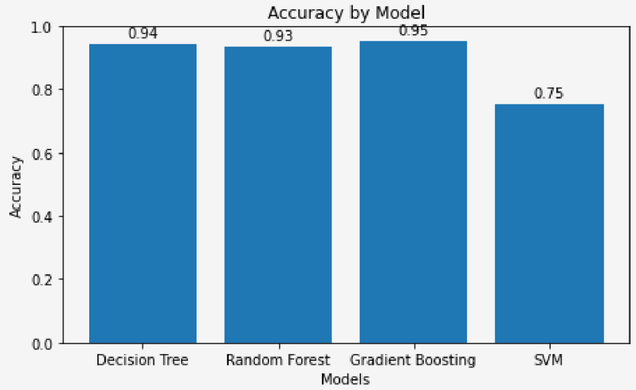


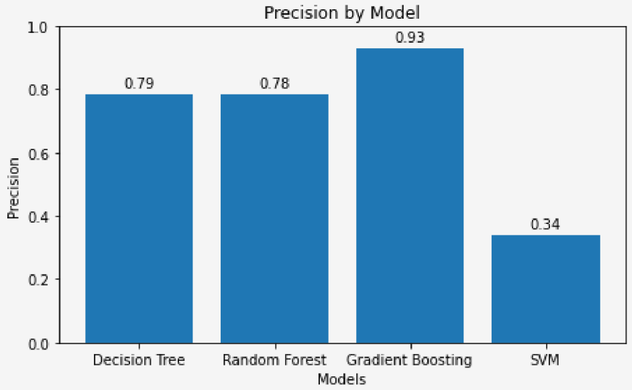
Modified Base

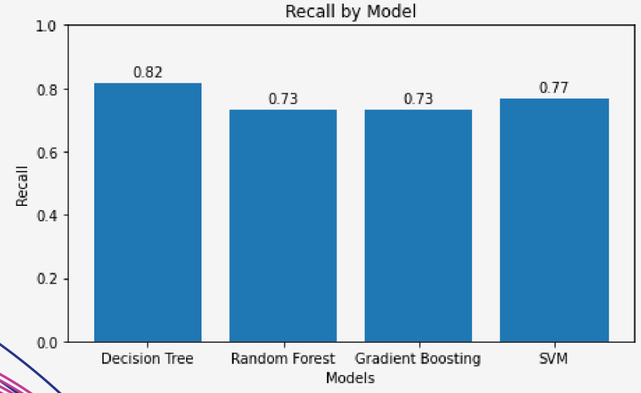
LOGREG MODEL

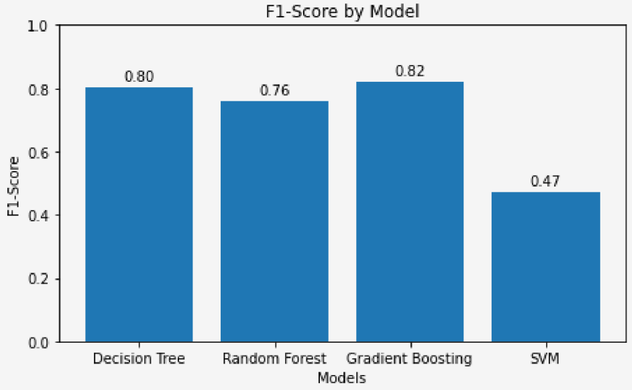
Handling Class Imbalance:

- We employed class weighting to address the imbalance issue
- The 'balanced' class weight setting was employed to achieve a notable improvement in churn recall (0.82).
- However, this approach resulted in a trade-off: While recall improved, it led to lower accuracy (0.65) and precision (0.28). This resulted in an F1-Score of 0.42, indicating the challenge of striking a balance between different performance metrics.









Advanced MODELLING

Decision Trees & Ensemble Methods

- To enhance churn prediction, we considered various models, aiming for high recall and improved precision.

Model Insights:

- Gradient Boosting excelled with the highest accuracy (95%), strong precision (93%), and balanced recall (73%).
- Decision Tree showed good performance with high recall (82%) and precision (79%).
- Random Forest achieved a reasonable balance between precision and recall.
- Support Vector Machine (SVM) had a high recall but lower precision, resulting in more false positives.

Model SELECTION

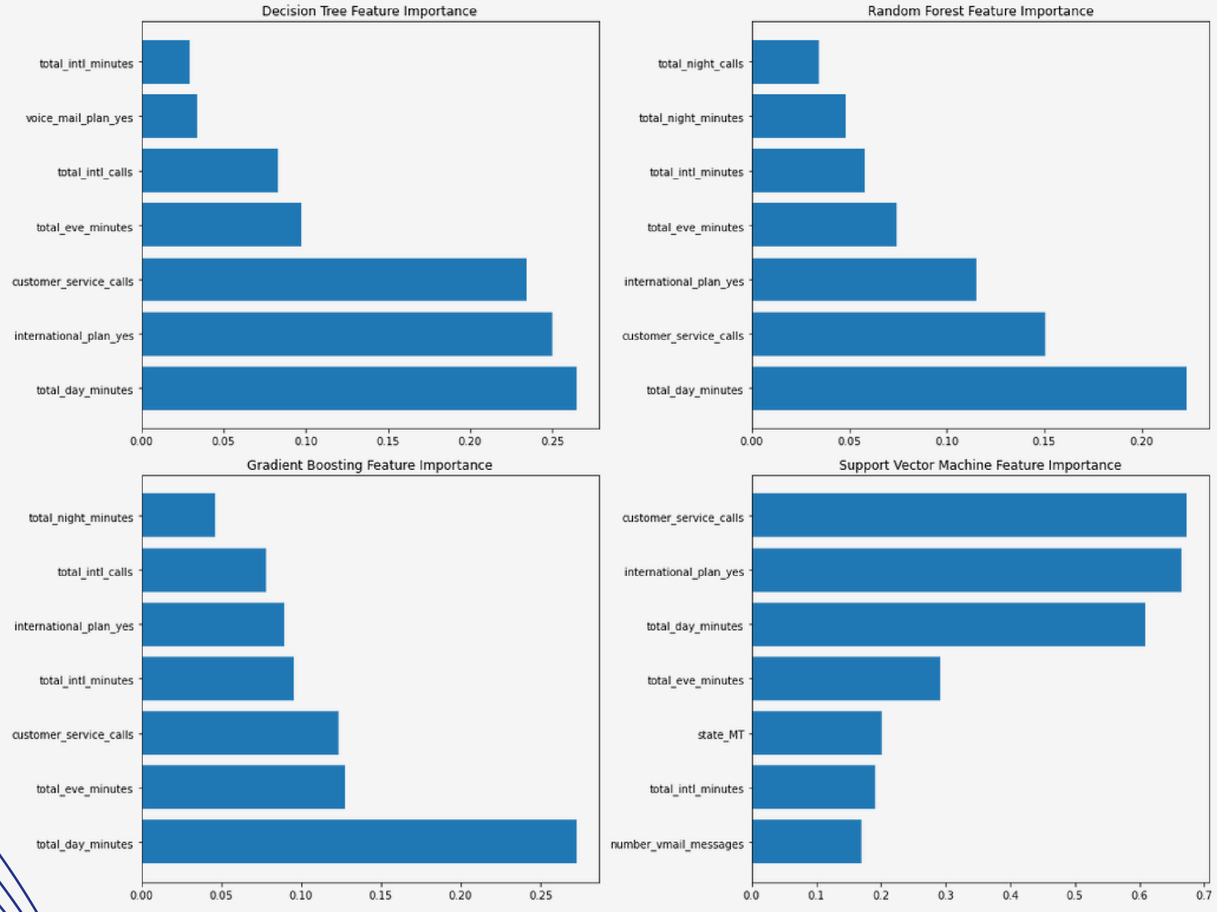
Decision Tree vs. Gradient Boosting: A Trade-off Perspective for Syriatel

- Decision Tree offers high recall but results in more false positives, suited for prioritizing genuine churn cases.

- Gradient Boosting balances precision and recall, minimizing false alarms while accurately identifying at-risk customers.

Model Selection CONSIDERATIONS

1. **Alignment with Priorities:** Choose a model that aligns with the department's priorities and resources.
2. **Cost-Effective False Alarms:** Consider the cost implications of false alarms. Decision Tree yields more false alarms, while Gradient Boosting strikes a balance.
3. **Accurate Churn Identification:** Focus on accurate identification of churn risk. Gradient Boosting excels in this aspect.
4. **Cost-Benefit Analysis:** Conduct a cost-benefit analysis to determine the model that best suits department objectives and capacity.



Feature IMPORTANCE

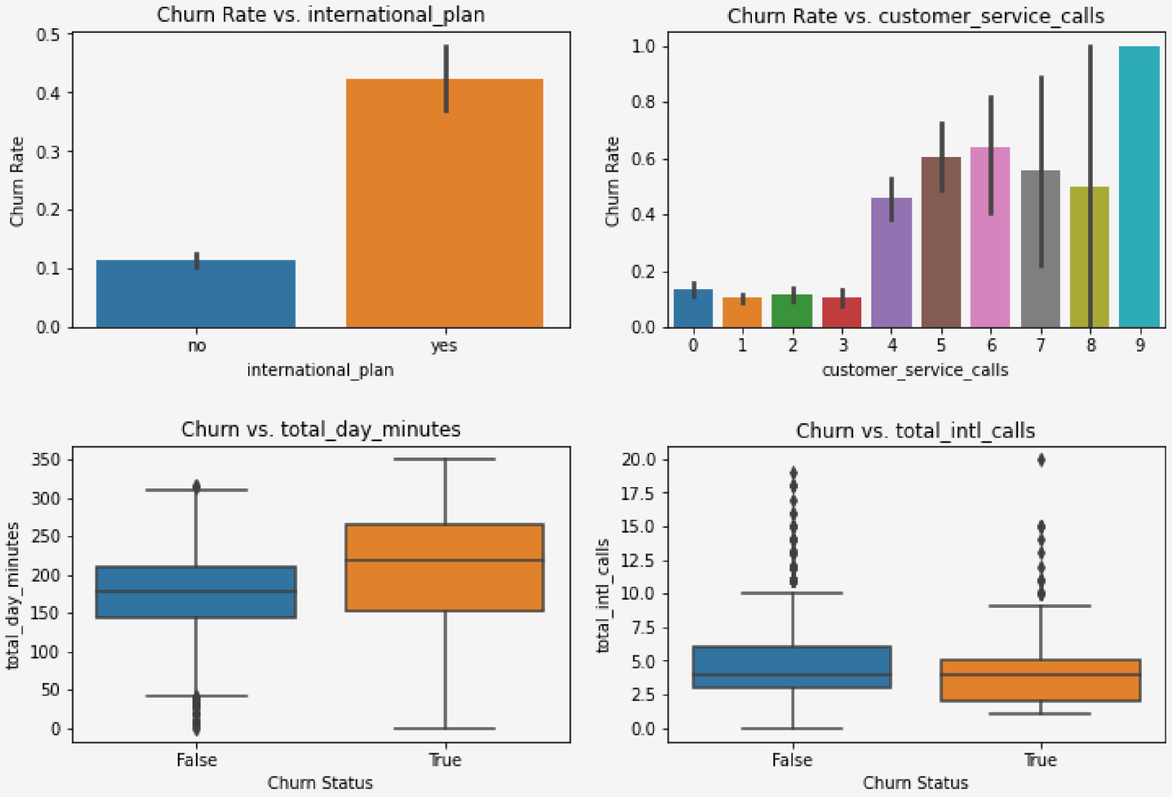
Feature importance provides valuable insights into which factors influence customer churn. Here are the top three features from our two best models:

Decision Tree:

- Total_day_minutes
- International_plan_yes
- Customer_service_calls

Gradient Boosting:

- Total_day_minutes
- Total_eve_minutes
- International_plan_yes
-



Feature - Target

RELATIONSHIP ANALYSIS

Taking a closer look at the top three features by importance from the two best model, we found the following relationships with the target variable:

- The churn rate for clients with international plans is higher, suggesting further investigation into plan pricing and call quality.
- The churn rate is high for customers with more customer service calls, indicating potential issues that need resolution.
- The churn rate is slightly higher for customers with more total day minutes, suggesting the need to review rate structures or provide incentives.

Conclusions & RECOMMENDATIONS

In summary, our analysis has revealed valuable insights into predicting customer churn for Syriatel. Our models, particularly Gradient Boosting, offer strong performance with high accuracy, precision, and recall. Feature importance analysis highlights key drivers of churn, such as 'Total_day_minutes,' 'International_plan,' and 'Customer Service calls.' These insights guide us toward actionable recommendations for enhanced retention efforts and customer satisfaction.

Recommendation 1

International Plan Optimization:
Given that the churn rate is higher for clients with international plans, the department should investigate the reasons behind this. It could be related to pricing, call quality, or other factors. By addressing **these issues, Syriatel can reduce churn among international plan users.**

Recommendation 2

Customer Service Escalation Threshold:
High churn rates associated with a high number of customer service calls suggest that some issues may not be resolved effectively. To address this, the department can implement an escalation threshold. When a customer calls more than the threshold, the case is escalated for review and further assessment, potentially leading to improved issue resolution.

Recommendation 3

Rate Structure Review: :
Since churn rates are slightly higher for customers with more total day minutes, the department can review rate structures. Introducing lower rates when customers cross a specific threshold of day minutes or offering other incentives can help retain customers with higher usage patterns.

SyriaTel Ltd

THANK YOU

Q & A

https://github.com/waynekipngeno

kipngenowayne@gmail.com

+254792003993