CUSTOMER CHURN PREDICTION FOR SYRIATEL

A Machine Learning Approach to Retain Customer

Presented by: George Mochama Edwin

Project Overview

- **Key Points**:
  - Develop a binary classification model to predict customer churn.
  - Analyze behavior, usage patterns, and demographics.
  - Goal: Enable targeted retention strategies to minimize revenue loss.

Problem Definition

- **Objective**: Predict customer churn (Yes/No) using ML models.
- **Outcome**: Provide actionable insights to:
- Reduce churn rates.
- Enhance customer retention and satisfaction.
- **Metrics for Success**:
- Accuracy, Precision, Recall, F1-Score, AUC-ROC.

Data Collection

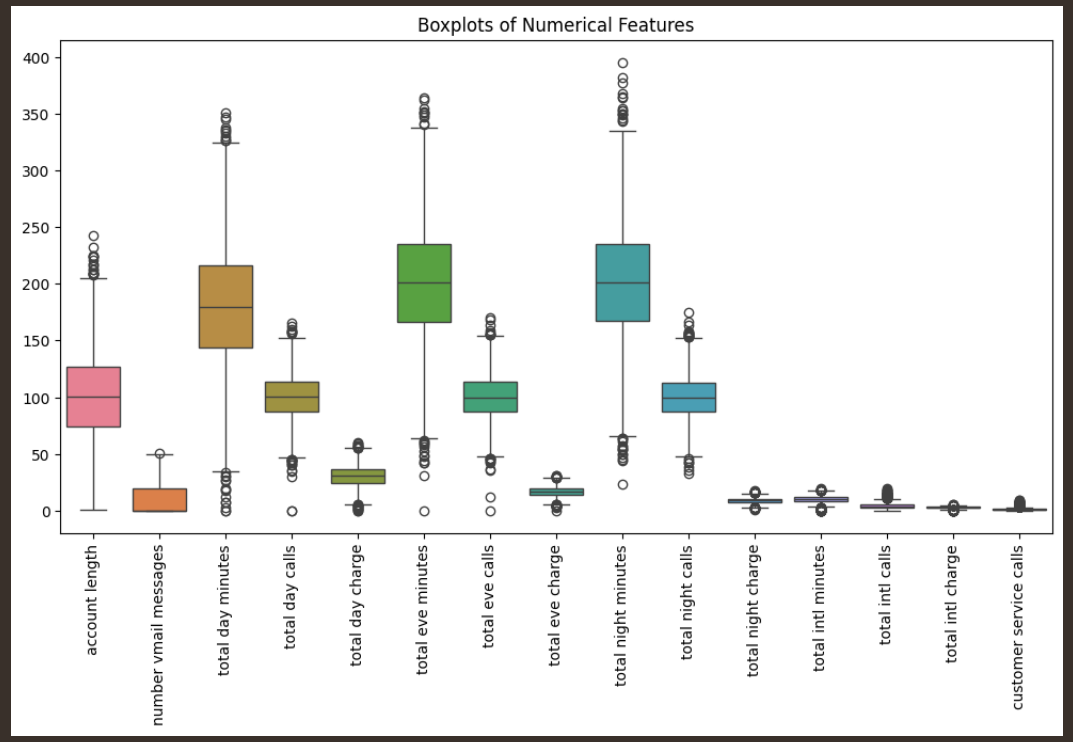
- **Dataset Source**: Kaggle.
- **Dataset Link**: Kaggle Dataset.
- **Data Format**: CSV.
- **Features**: Demographics, account details, usage patterns.

Data Preparation

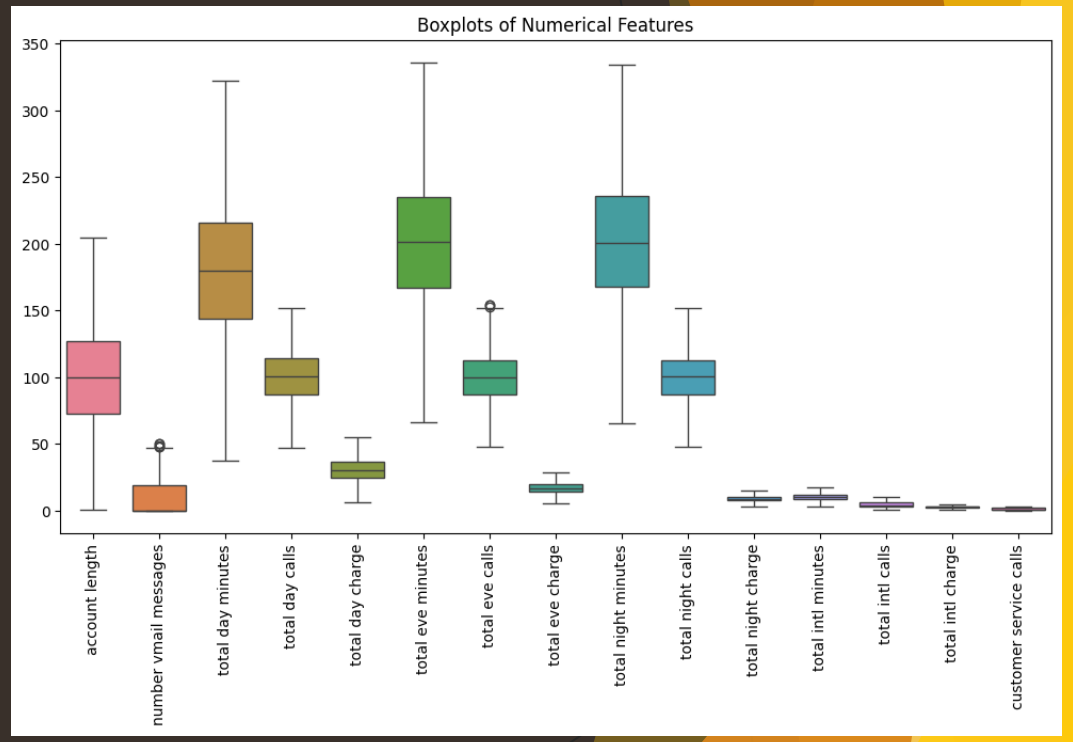
- **Cleaning:**
  - Removed irrelevant columns (phone number, state, area code).
  - Verified no missing or duplicate data.
- **Encoding:**
  - Transformed international plan and voice mail plan into binary (Yes → 1, No → 0).
- **Scaling:**
  - Normalized numerical features using Min-Max Scaling.
- **Outlier Handling:**
  - Detected outliers using boxplots.
  - Removed outliers using the IQR method.

Box plot of detected outliers and removed.

WITH OUTLIERS



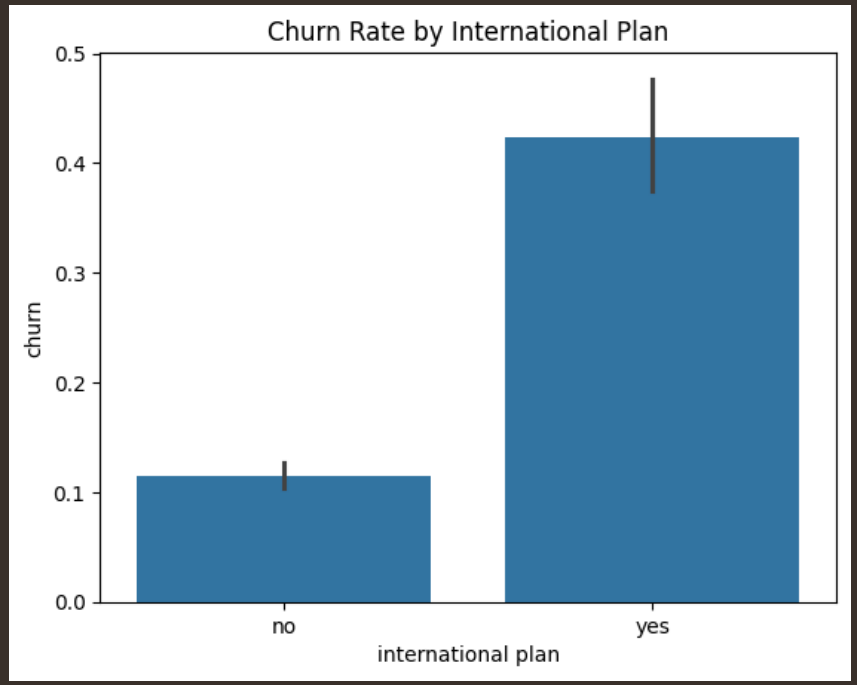
WITHOUT OUTLIERS



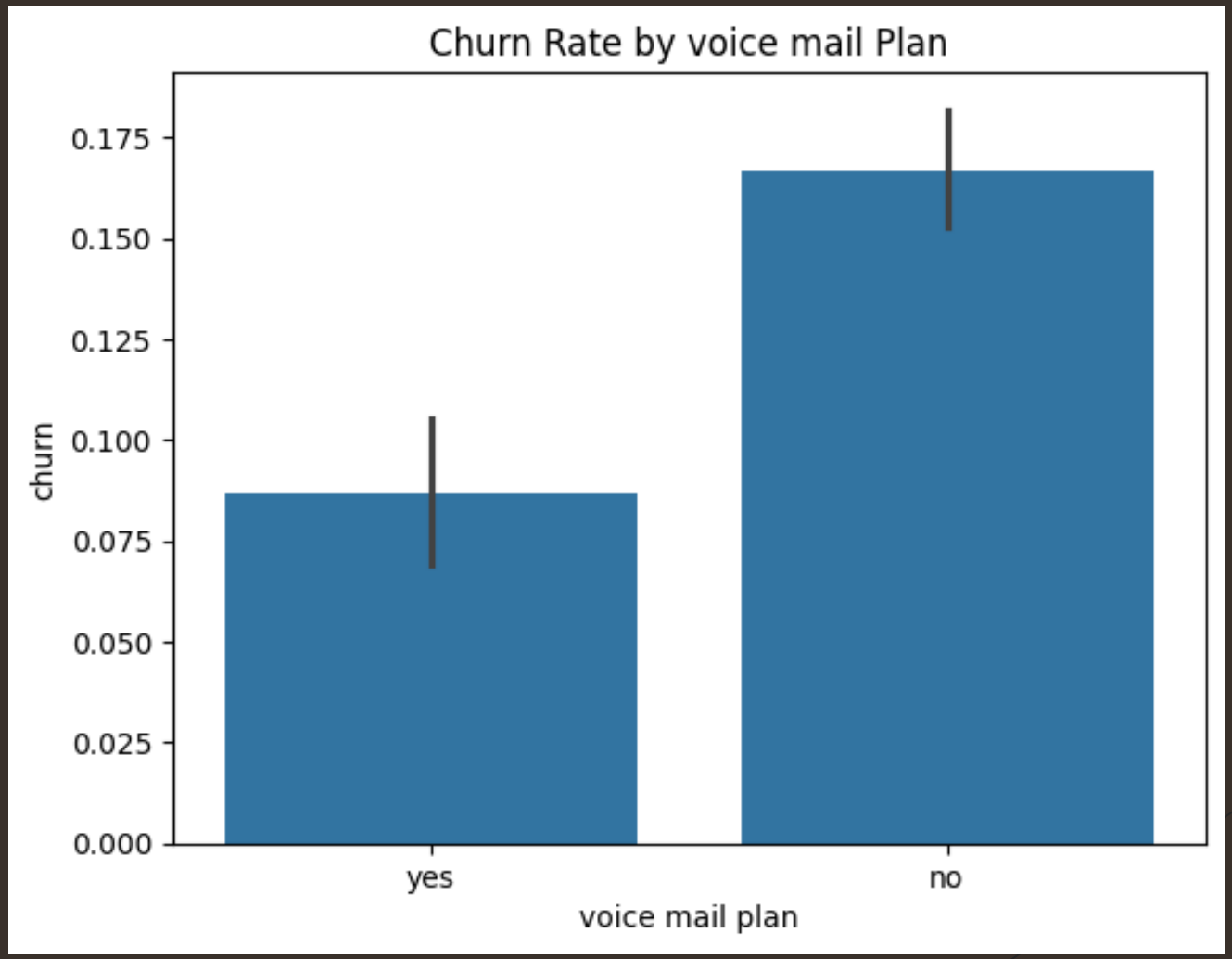
Exploratory Data Analysis (EDA)

Key Insights:

- Customers with an international plan are more likely to churn (42% churn rate vs. 11% without the plan).



Customers with a voice mail plan churn less (8.6% vs. 16.7% without the plan).



Train-Test Split

- **Details**:
  - Split dataset into training and test sets (70/30 split).
  - Observed class imbalance:
    - Majority: No churn (0) - 2493 instances.
    - Minority: Churn (1) - 304 instances.

Modeling Approach

- **Models Evaluated:**
  1. Logistic Regression: Baseline model for simplicity and interpretability.
  2. Random Forest Classifier: Combines multiple decision trees for robustness.
  3. Decision Tree Classifier: Simple and interpretable but prone to overfitting.
- **Evaluation Metrics:**
  - Precision, Recall, F1-Score, AUC-ROC.

Logistic Regression Performance

- **Results:**
  - Training Accuracy: 89.32%.
  - Test Accuracy: 88.57%.
  - AUC: 0.861.
- Challenges:
  - High performance for "No Churn" (97% recall).
  - Low performance for "Churn" (22% recall).

Random Forest Performance

- **Results:**
  - Training Accuracy: 100% (Potential overfitting).
  - Test Accuracy: 94%.
  - AUC: 0.871.
- Highlights:
  - Best performance among models.
  - Balanced Precision (68%) and Recall (74%) for "Churn".

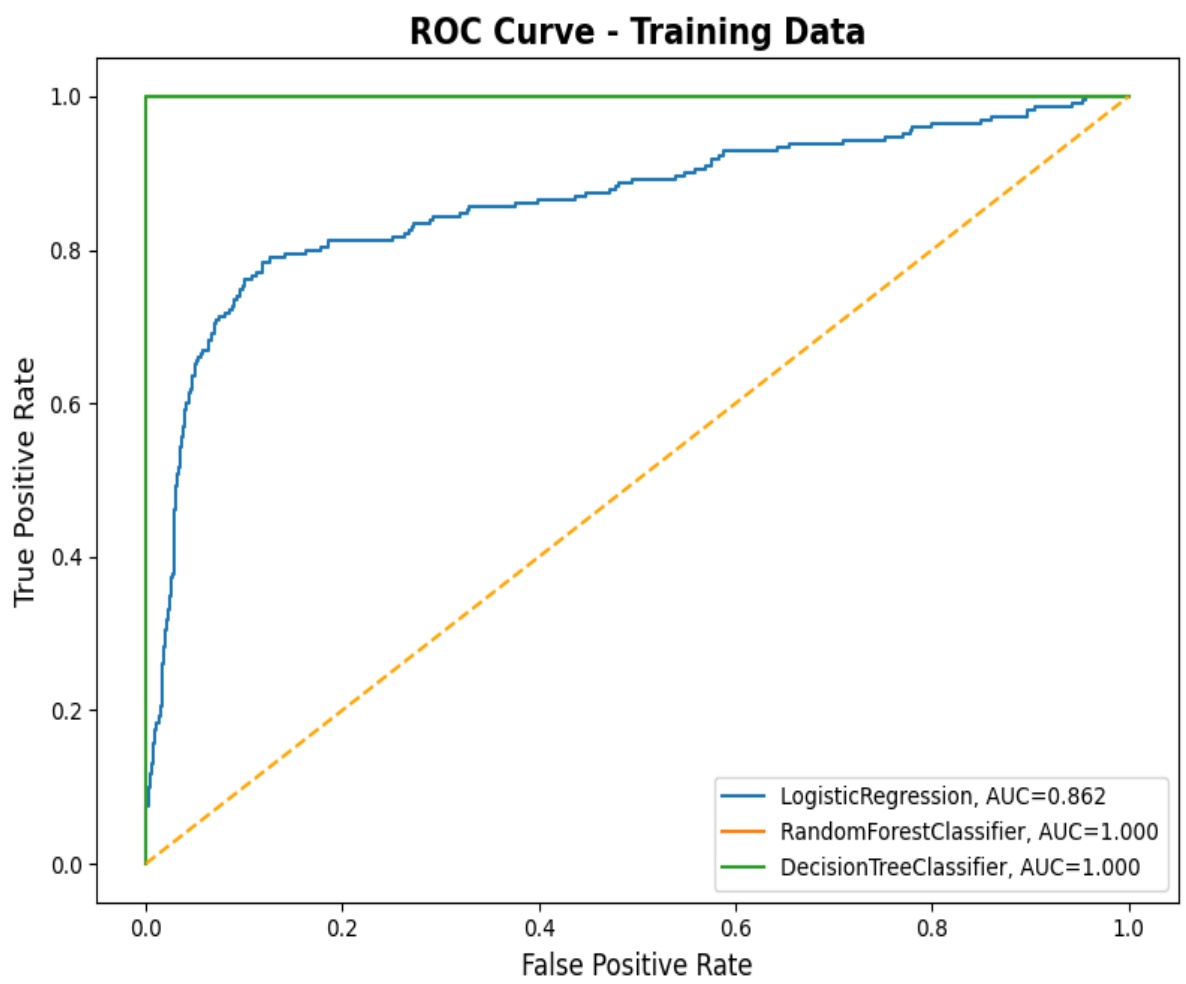
Decision Tree Performance

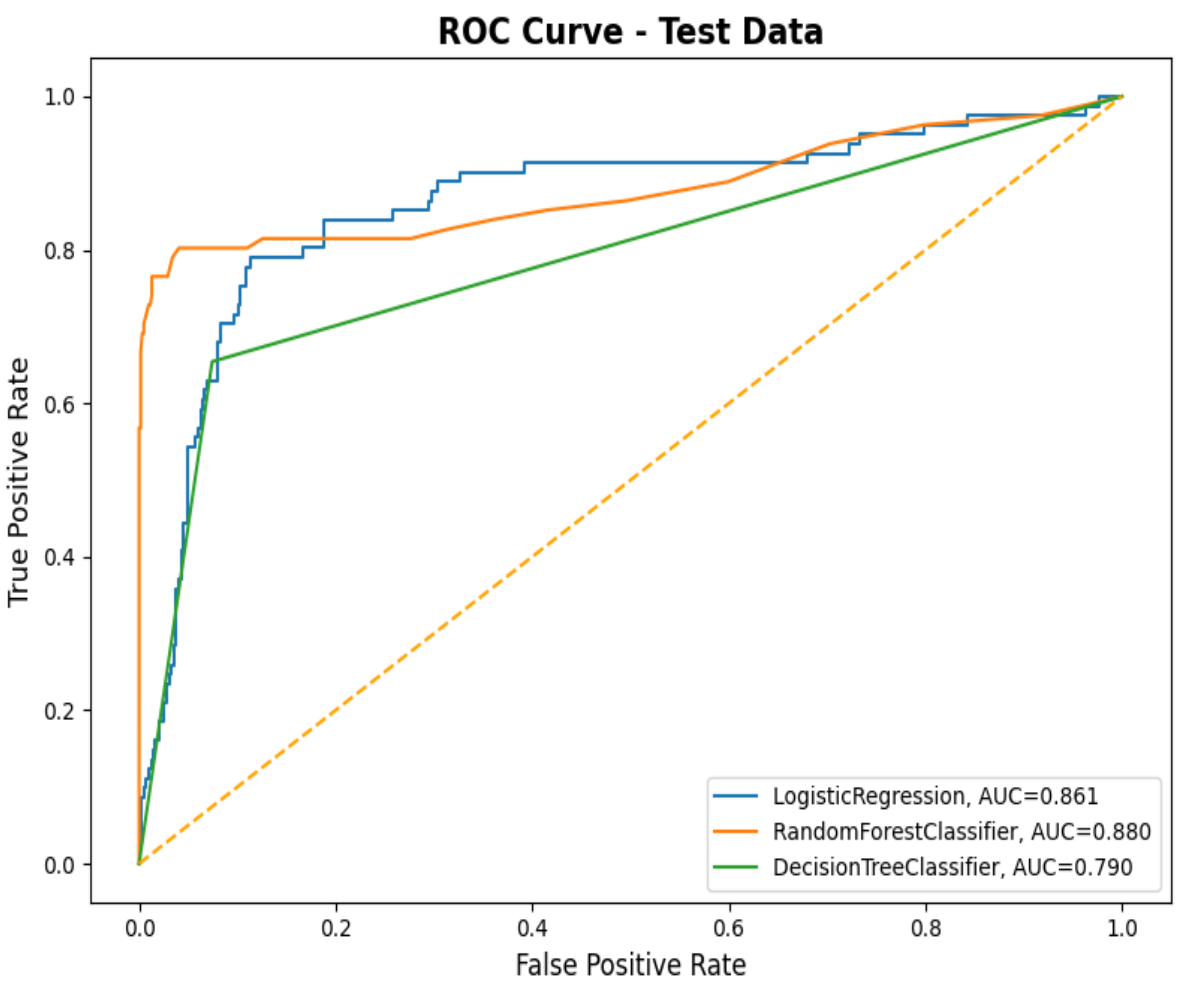
- **Results:**
  - Training Accuracy: 100% (Severe overfitting).
  - Test Accuracy: 90.86%.
  - AUC: 0.814.
- Observations:
  - High precision for "No Churn" (96%).
  - Moderate recall for "Churn" (72%).

ROC Curve Comparison

- **Key Points:**
- Random Forest outperforms Logistic Regression and Decision Tree.
- ROC curves demonstrate trade-offs between models.

Combined ROC curves for all models.

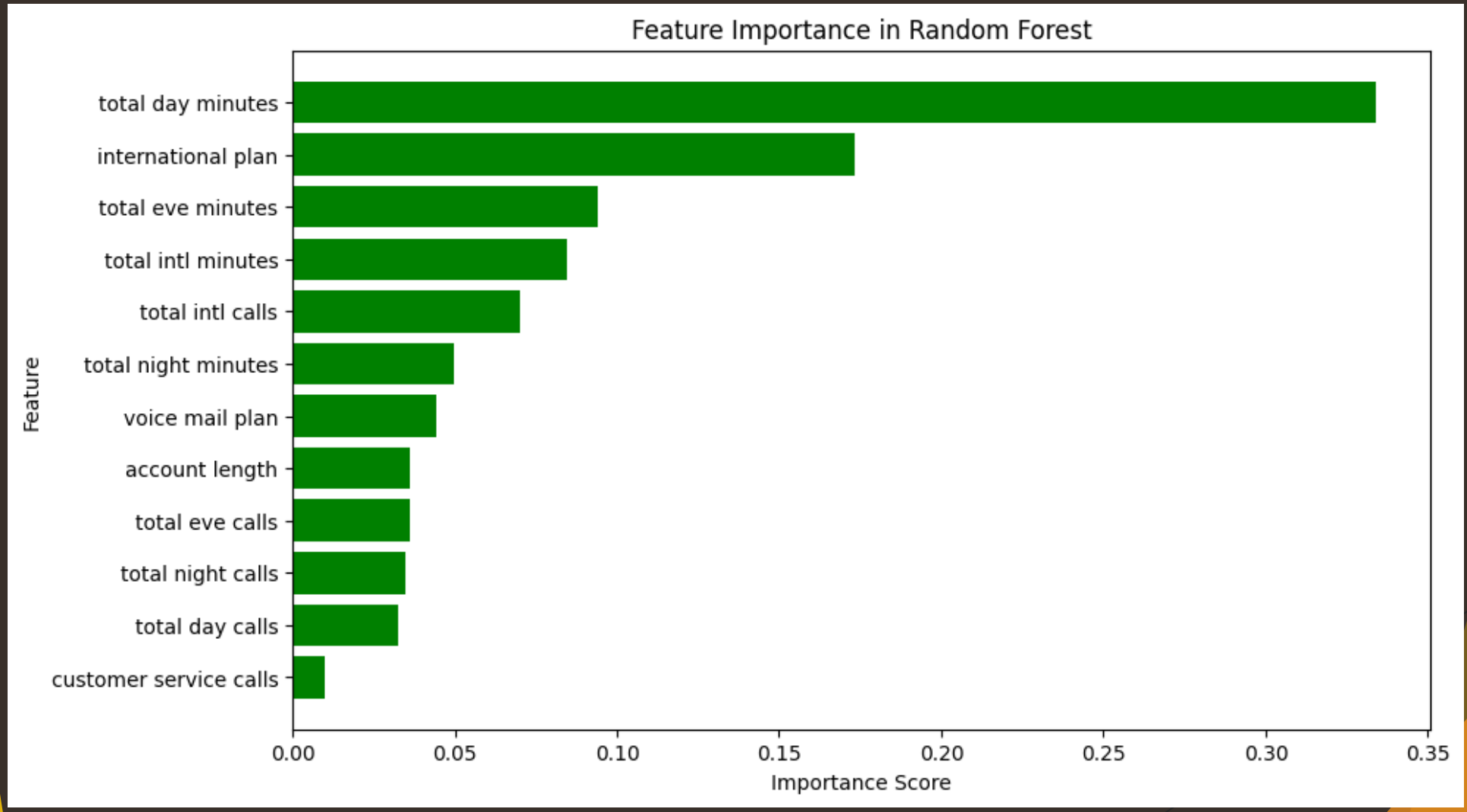




Feature Importance Analysis

- **Most Influential Features:**
  - total day minutes (33.4% importance).
  - international plan (17.3% importance).
  - total eve minutes (9.4% importance).
- **Lesser Influential Features:**
  - customer service calls (0.9% importance).
  - total day calls (3.2% importance).

Feature importance bar chart for RF.



Model Selection

- **Summary**:
  - Random Forest is the best-performing model:
  - Test Accuracy: 94%.
  - AUC: 0.871.
- Handles class imbalance better with SMOTE resampling.

Conclusion

- **Key Takeaways:**
  - Random Forest is the most suitable model for deployment.
  - Focus on features like total day minutes and international plan.
  - Address class imbalance with SMOTE or similar techniques.
- **Next Steps:**
  - Optimize hyperparameters further.
  - Deploy the model in a real-world environment.

Actionable Insights

- **Focus on High-Risk Customers:**
  - Customers with an international plan have a 42% churn rate.
  - Target this group with improved international plan offerings or discounts.
- **Enhance Customer Engagement:**
  - Customers with a voice mail plan churn less (8.6% vs. 16.7% without).
  - Promote voice mail plans to reduce churn rates.
- **Monitor Usage Patterns:**
  - High total day minutes is a strong predictor of churn.
  - Analyze high usage patterns and offer loyalty incentives or tailored plans.