Predictive Analysis of Customer Churn for SyriaTel Telecom

By: Pacificah Kwamboka Asamba

Project Overview

- Churn prediction helps SyriaTel identify customers likely to leave, enabling early intervention.

Business Understanding

- Customer churn significantly affects telecom revenue. The goal is to proactively detect customers at risk of leaving the service, allowing the business to take retention actions such as improved customer service or promotional offers.
- This project aims to analyze customer behavior and build a predictive machine learning model that can identify customers at risk of churning.

Objective

- The primary objective is to develop a robust and interpretable classification model that can:

- Accurately predict whether a customer is likely to churn.

- Provide insights into the key drivers of churn.

- Support data-driven decision-making for retention initiatives.

Goals

- Understand the distribution and structure of customer-related features.
- - Explore relationships between customer attributes and churn behavior.
- - Engineer relevant features that improve model performance.
- - Train and evaluate multiple classification models using industry-standard metrics.
- - Interpret model results to inform business actions.

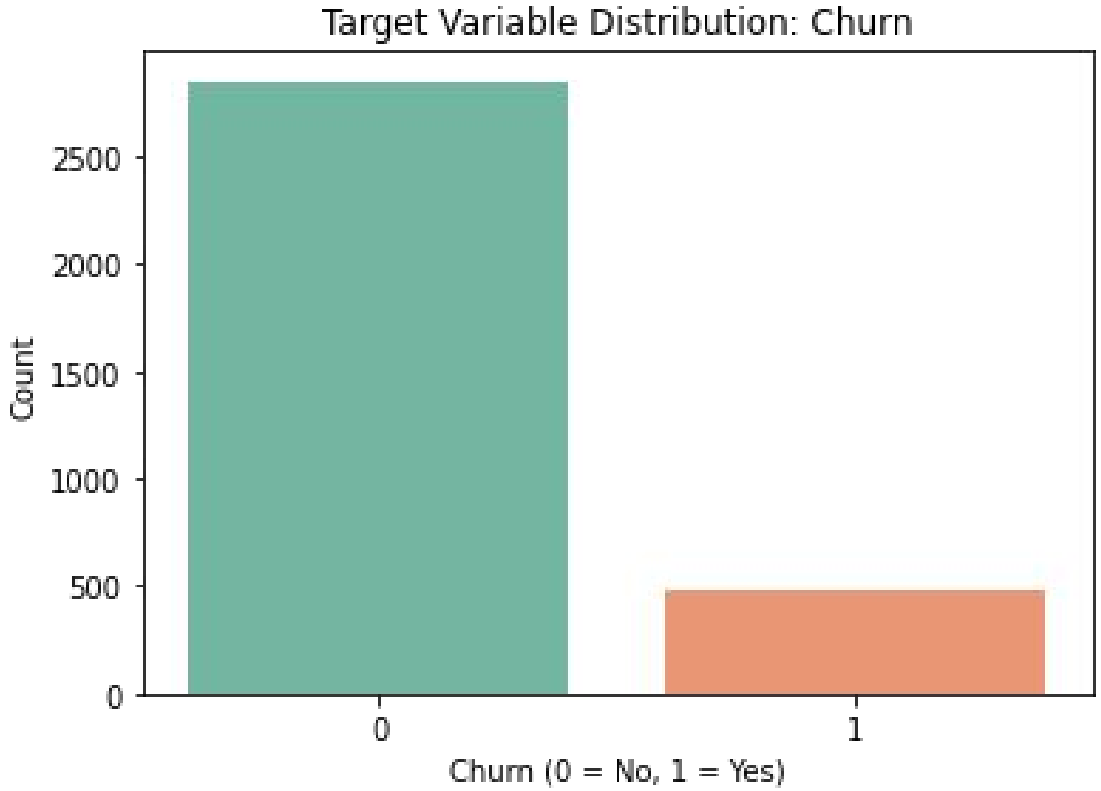
Data Understanding

- The dataset used for the project was obtained from Kaggle(https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset)

Dataset includes 3,333 customer records which include.

- Account length
- Customer service calls
- Minutes and charges during different times of the day
- International and voice mail plans
- Churn label (target variable)

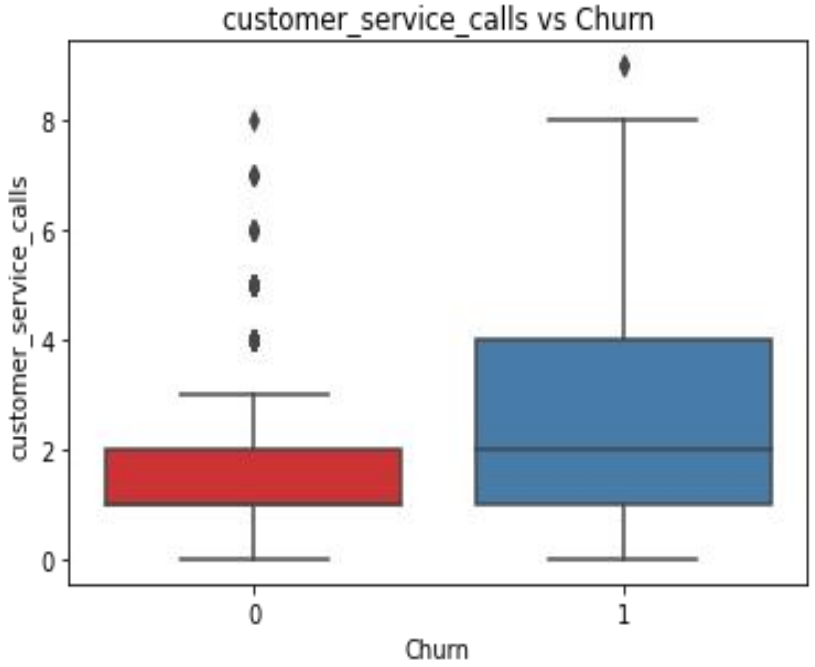
Target variable Distribution



This bar plot show the distribution of the target variable - Churn.
The left bar (0) shows non-churn(those that will not leave) and the right bar(1) shows those customers linkely to churn.

Churn vs Key Features

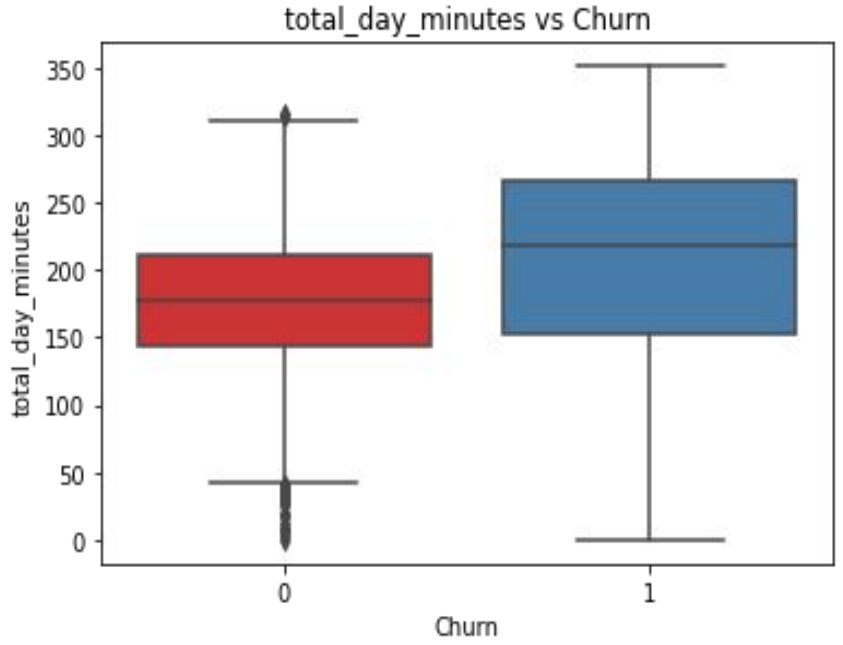
1. **Customer service calls vs churn**



Insight:
Customers who **churned** made **significantly more customer service calls** compared to those who stayed.
This suggests that **frequent complaints or unresolved issues** may be a key churn driver.

Churn vs Key Features cont.

2. **Total day minutes vs churn**



Insight:
Churned customers tend to have **higher total day call minutes** than non-churners.
This implies that **high-usage customers** may be more likely to leave, possibly due to **cost concerns or unmet expectations**.

Modelling

In this section, I build predictive models to classify whether a customer will churn or not. The modeling workflow follows these key steps:

1. Define the classification problem
2. Split data into features and target
3. Build a scalable ML pipeline
4. Train and evaluate multiple models iteratively
5. Extract predictive insights from the models

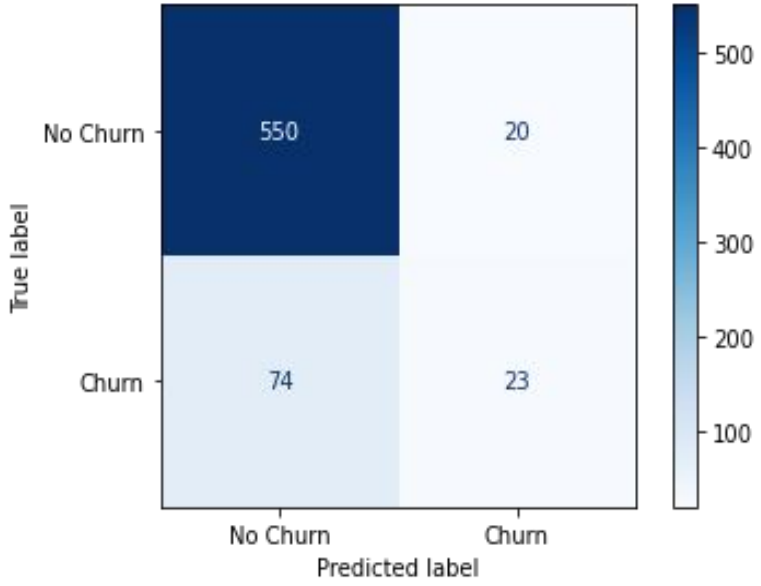
Modelling: Machine Learning

In this section I trained three models:

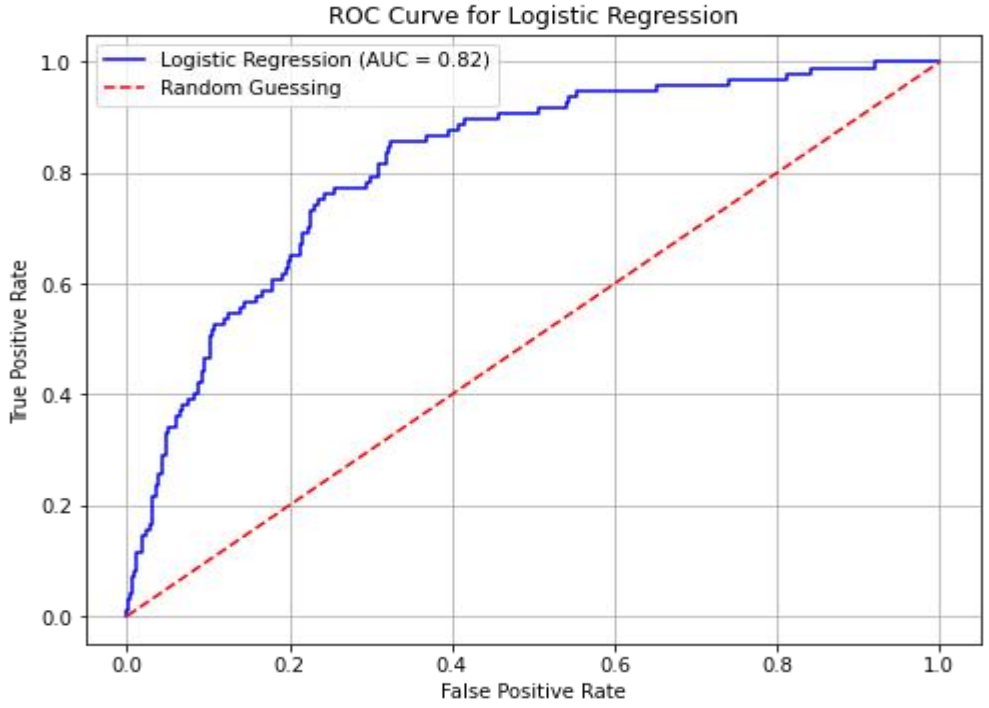
1. Logistic Regression model – This acts as a baseline model on which others will be built
2. Decision Tree Classifier -a **supervised machine learning algorithm** used for classification tasks. It models decisions and their possible consequences as a **tree-like structure** of branches and nodes.
3. Random Forest Classifier - This is an ensemble model that improves prediction using multiple decision trees. Handles non-linear relationships and feature importance.

Model 1: Logistic Regression

Confusion Matrix

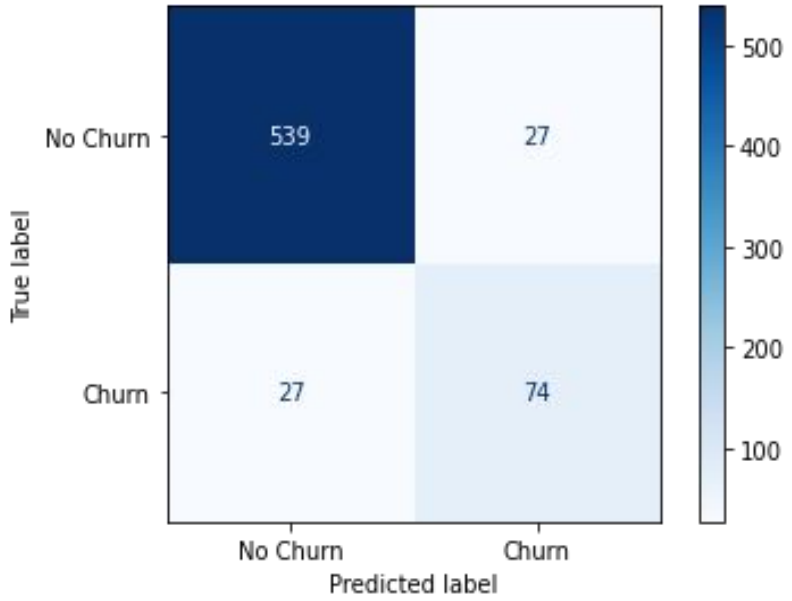


ROC Curve

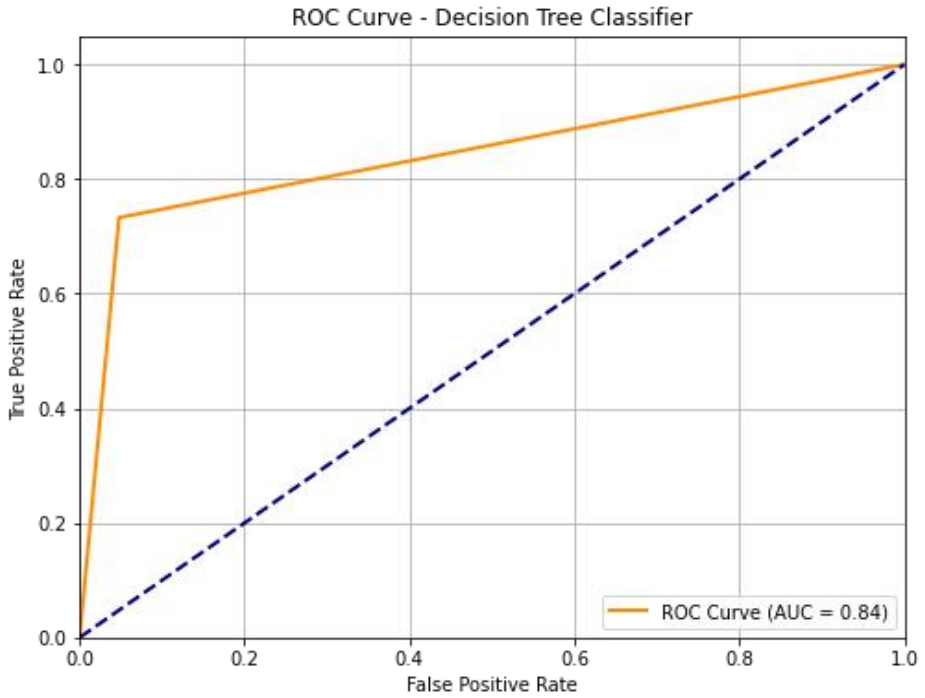


Model 2: Decision Tree Classifier

Confusion Matrix

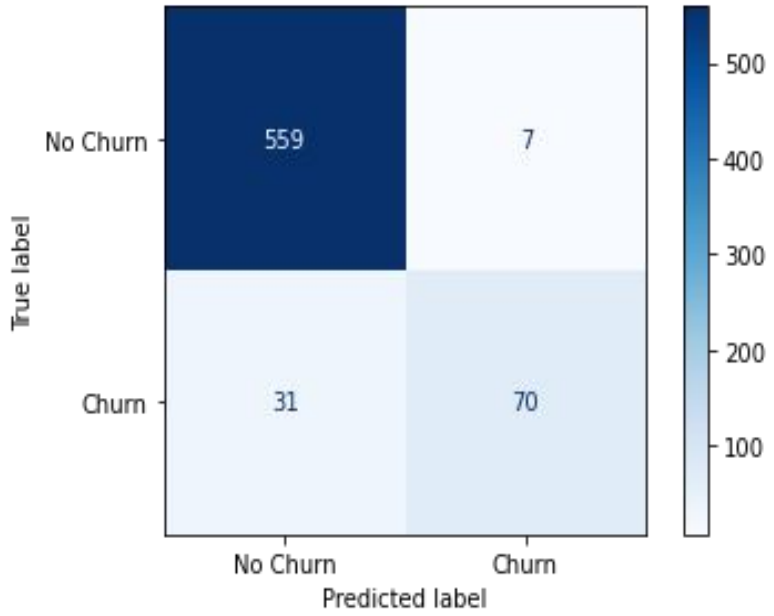


ROC Curve

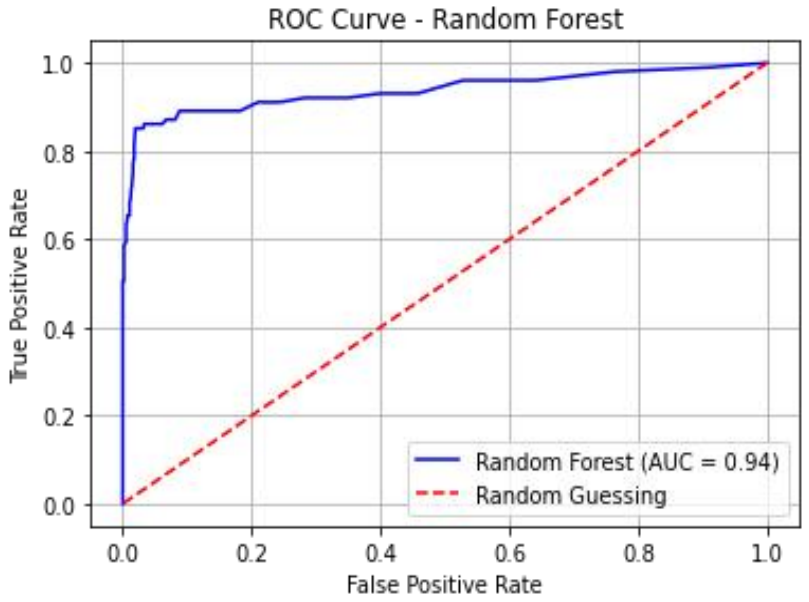


Model 3: Random Forest classifier

Confusion Matrix

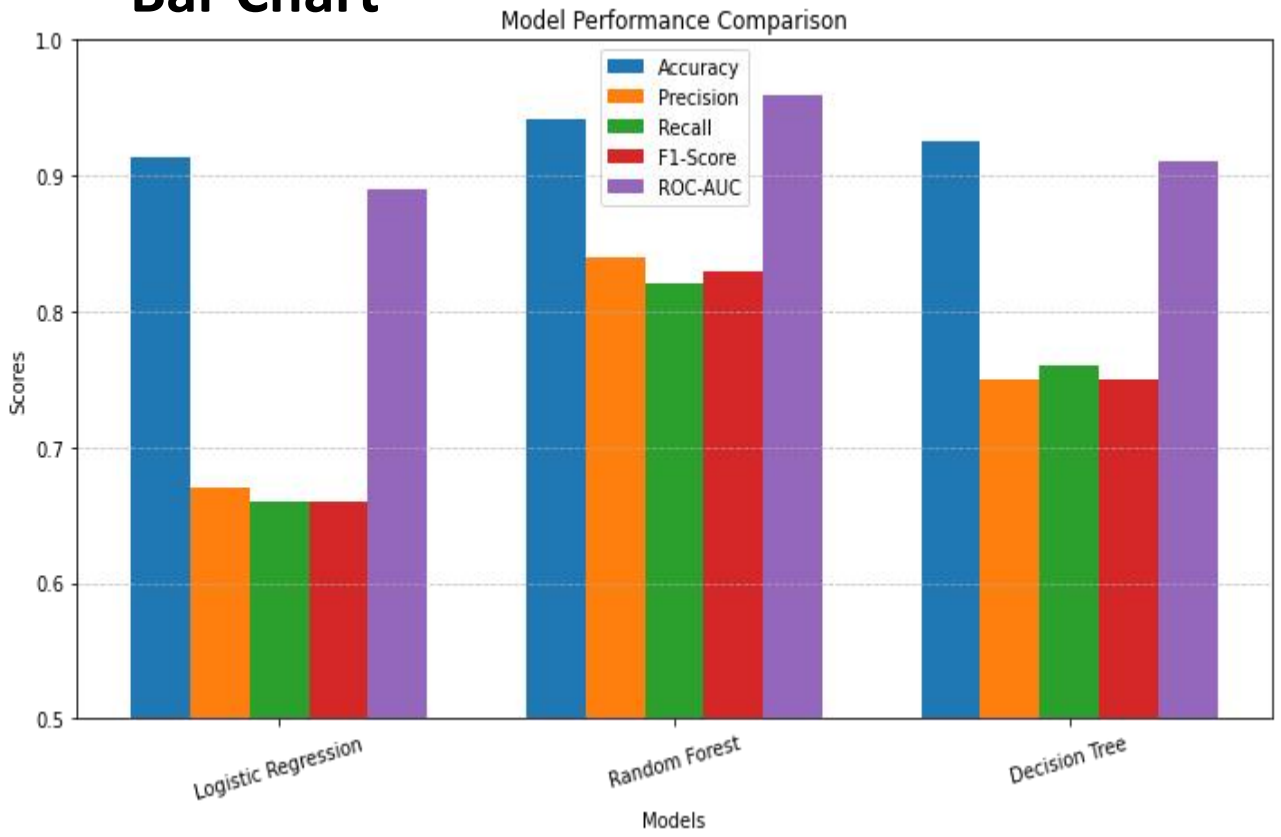


- ROC Curve



Model Evaluation: Comparing the three models

Bar Chart



Bar chart interpretation

Random Forest clearly leads in all metrics.

Decision Tree offers a strong balance of performance and interpretability.

Logistic Regression is decent but lags slightly in recall and F1-score for churn prediction.

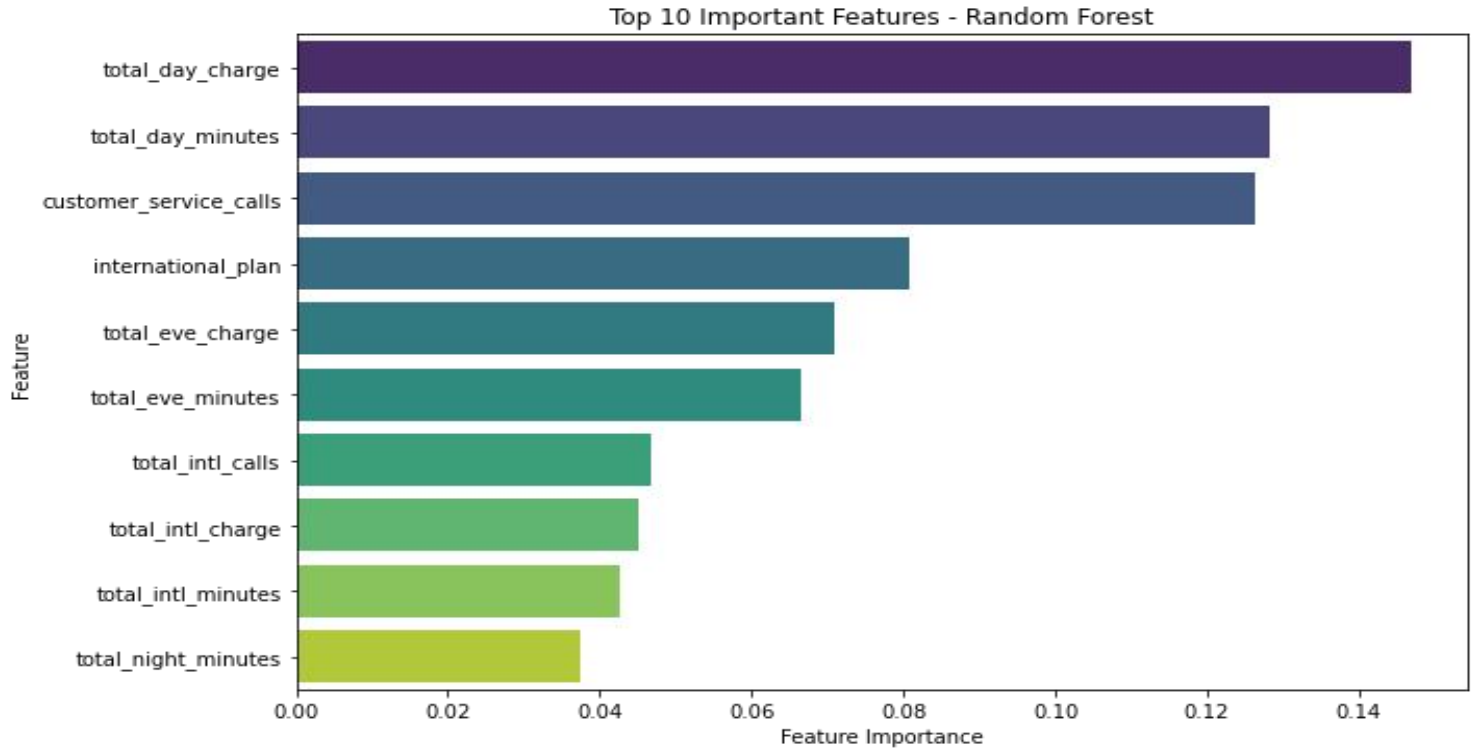
Model Performance Comparison Table

Metric	Logistic Regression	Random Forest	Decision Tree
Accuracy	91.3%	94.1%	92.5%
Precision (Churn)	~0.67	~0.84	0.75
Recall (Churn)	~0.66	~0.82	0.76
F1-Score (Churn)	~0.66	~0.83	0.75
ROC-AUC	~0.89	~0.96	~0.91

Random Forest Outperforms Logistic Regression and Decision Tree.
- Recall (Churn) is the most critical metric — Random Forest recovers nearly 3x more churners.
- AUC of 0.89 means excellent ability to distinguish churn vs. no-churn cases.
- Lower false negatives makes it more trustworthy for retention strategies.

Feature Importance Analysis

Based on the best performing model(Random Forest)



Features that most influence customer churn

1. Total_day_minutes, total_day_charge: Heavy daytime users are at higher risk of churning, possibly due to cost sensitivity or dissatisfaction with service quality.
2. Customer_service_calls: Strongest churn indicator — frequent interactions suggest unresolved issues or complaints.
3. International_plan_yes: Users with international plans churn more, potentially due to high costs or unmet expectations.
4. Evening and international usage features also ranked highly, reinforcing the importance of usage behavior.

Recommendations

1. Target High-Risk Users with Proactive Retention
2. Improve Customer Support Experience
3. Promote Bundled Plans
4. Monitor and Update the Model Regularly
5. Business Integration - Embed the model into CRM to flag risky users.

Next Steps

1. Use SMOTE or class weighting to further address class imbalance
2. Deploy the model for real-time scoring
3. Collaborate with business teams for retention campaigns based on predictions

Questions?

THANK YOU