A SYRIATELCUSTOMER CHURN ANALYSIS

BUSINESS PROBLEM

Syria Tel is losing revenue due to preventable customer churn. To combat this, we aim to proactively identify at-risk subscribers using behavioral patterns and spending tiers, enabling targeted retention strategies to curb attrition.

PROJECT OBJECTIVES

Identify the current churn rate.

+ Identify What patterns precede churn.

+ Identify the best retention strategies..

+ Identify which service erodes more revenue to Churn.

DATA UNDERSTANDING

Dataset Summary

The dataset contains 3,333 customer records and 21 features. There are no missing values or duplicate entries, indicating that the data is clean. The features include a mix of categorical, numerical, and Boolean variables. The target variable is churn, which is a binary outcome indicating whether a customer has churned (True) or not (False).

Why Class Imbalance Matters

A model may achieve high accuracy by predicting most customers as “not churned”.
We'll focus on Precision, Recall, F1-Score, and ROC-AUC for evaluation.

DATA CLEANING & PREPARATION

Before modeling, we applied several key steps to prepare the data for analysis:

- Dropped Uninformative Columns:
  phone number: a unique identifier with no predictive value.
  state: high cardinality leading to model complexity and potential overfitting.
  area code: low correlation with churn and limited behavioral insight.

- Categorical Encoding:
  Converted international plan and voice mail plan from 'yes'/'no' to 1/0.

- Target Variable Transformation:
  Encoded churn (True/False) to 1/0 using Label Encoder.

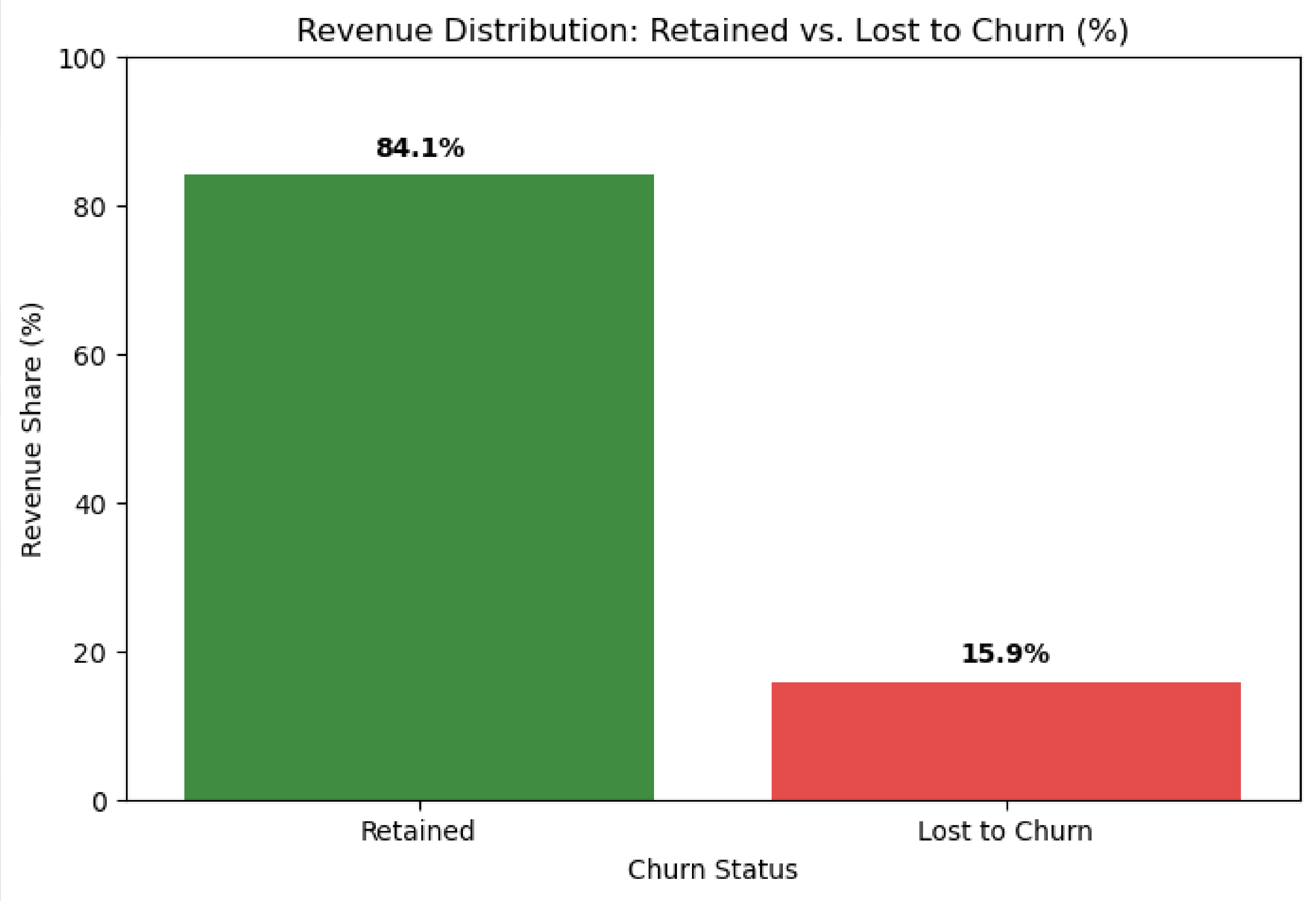
- Column Name Standardization:
  Lowercased all names.
  Trimmed whitespaces.
  Replaced spaces with underscores.

- Final Dataset:
  3,333 rows, 20 features.
  No missing or duplicate values.

EXPLORATORY DATA ANALYSIS

Target Variable Distribution

Bar plot showing churn distribution with percentages (84.1% not churned, 15.9% churned)



Key Points:

Significant class imbalance: 15.9% of customers have churned.

Imbalance necessitates resampling techniques during model training to avoid model bias.

CUSTOMER BEHAVIOR INSIGHTS (UNIVARIATE & BIVARIATE ANALYSIS)

Key Usage Patterns (Univariate)

- Most features show balanced (normal) distributions.
- Voicemail usage is low many users have zero messages.
- International and customer service calls are right-skewed , most users make few, but some make many.
- Area codes cluster into two main groups.

Churn vs. Behavior (Bivariate)

- International Plan → Higher churn
- Voice Mail Plan → Lower churn
- More customer service calls → More churn
- More voicemail messages → Less churn

Usage & Charges

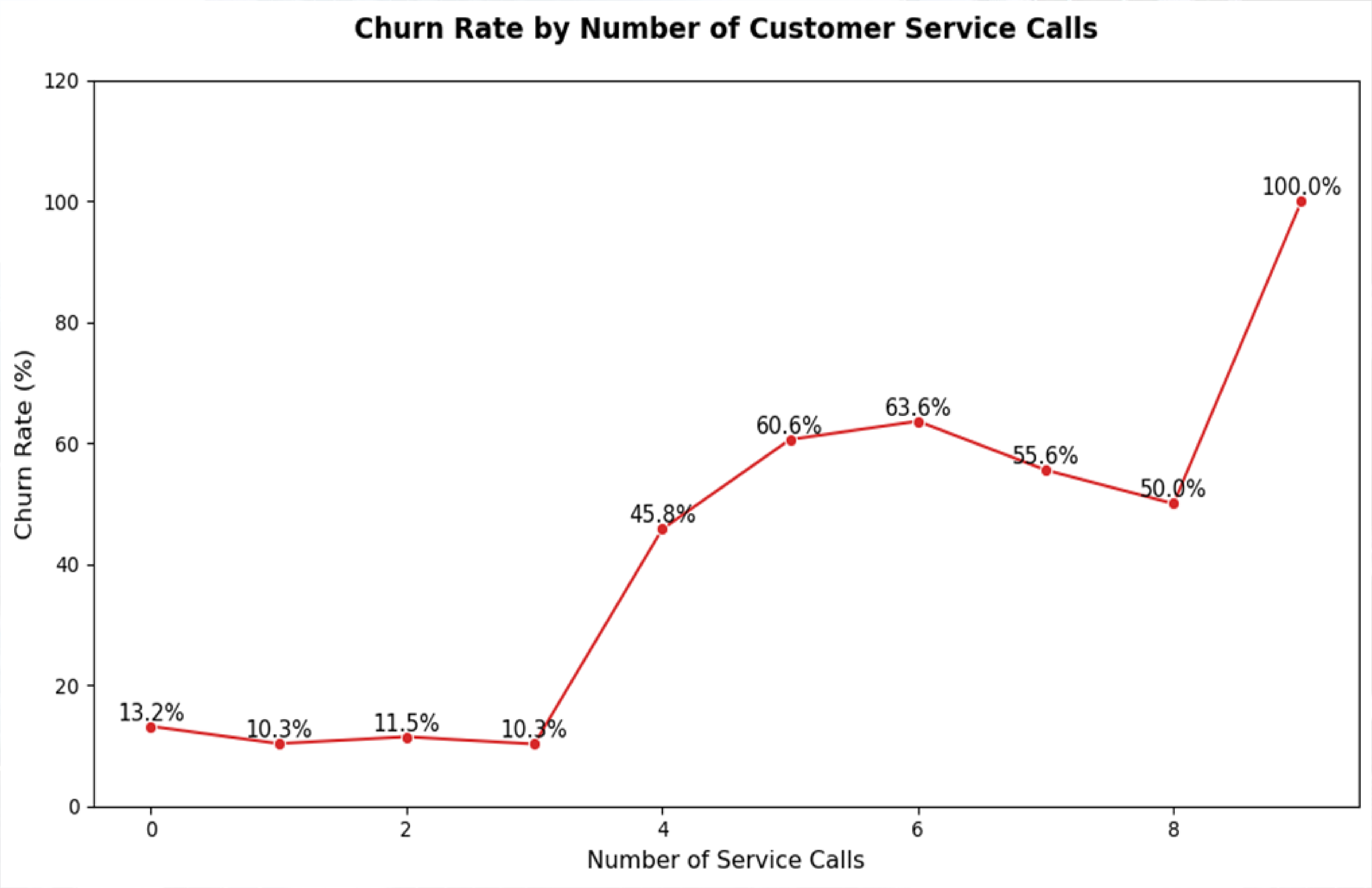
- Churners tend to have higher usage (day, evening, night, international) and higher charges.
- Call counts don't differ much, but charges do — usage intensity matters.
- Account length is slightly shorter for churners, but not a strong factor.

Regional Insights (State-level)

- States like New Jersey, California, and Texas show above-average churn rates.
- Geography may influence churn due to competition, infrastructure, or regional preferences.
- These insights help in targeted business strategy, even if state is dropped in modeling.

RELATIONSHIP BETWEEN CUSTOMER SERVICE CALLS CHARGES AND CHURN RATE

- 0 to 3 service calls: 11.3% churn rate
- 4 service calls: 45.8% churn rate
- 5 service calls: 60.6% churn rate
- 6 service calls: 63.6% churn rate
- 7 service calls: 55.6% churn rate
- 8 service calls: 50% churn rate
- 9 and above service calls: 100% churn rate



CHURN RATE BY SERVICE PLAN

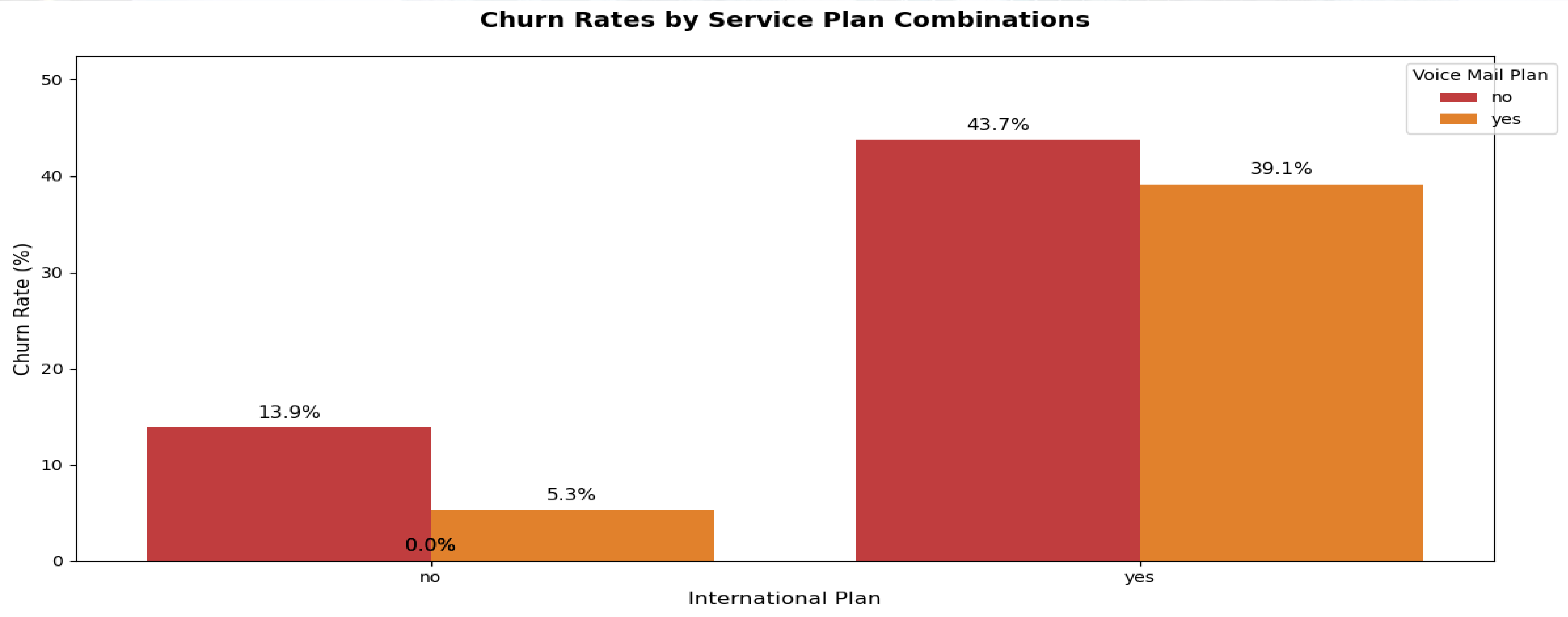
We looked at how churn differs for two customer plans:

International Plan

Customers with an international plan churn more often than those without.
This may point to lower satisfaction among international plan users.

Voice Mail Plan

Customers with a voice mail plan churn less than those without.
This could mean that value-added services help improve retention.



DATA PREPARATION FOR MODELING

Before building our churn prediction model, we refined the dataset to improve accuracy and prevent technical issues:

Final Data Cleanup

- Dropped non-informative features:
  phone_number (just an ID)
  state, area code (low or misleading predictive value)
- Standardized data:
  Encoded international_plan and voice_mail_plan as 1/0
  Encoded target churn as 1/0
  Cleaned column names for consistency

Addressing Multicollinearity

To avoid model confusion from overlapping data, we removed highly correlated features using VIF (Variance Inflation Factor):

- Dropped 7 Features:

Charges like total_day_charge, total_charge etc were redundant with minute values hence we dropped them
number_vmail_messages overlaps heavily with voice_mail_plan hence dropped as well
charge_bin — derived from total_charge, adds no new value hence had to be dropped

Remaining 13 features are clean, independent, and ready for modeling.
Balanced the Target Variable
Original churn rate: ~14.5%

After using SMOTE, churn vs. no-churn is now 50/50 in the training set.

OUR FINAL MODEL: Advanced Random Forest

METRIC	SCORE	INTERPRETATION
Recall	77%	Identifies 7/10 churns
Precision	81%	2 in 10 alarms are false alarms
ROC-AUC	93.1%	Strong separation between churns and loyal customers.

Threshold Tuning Insight

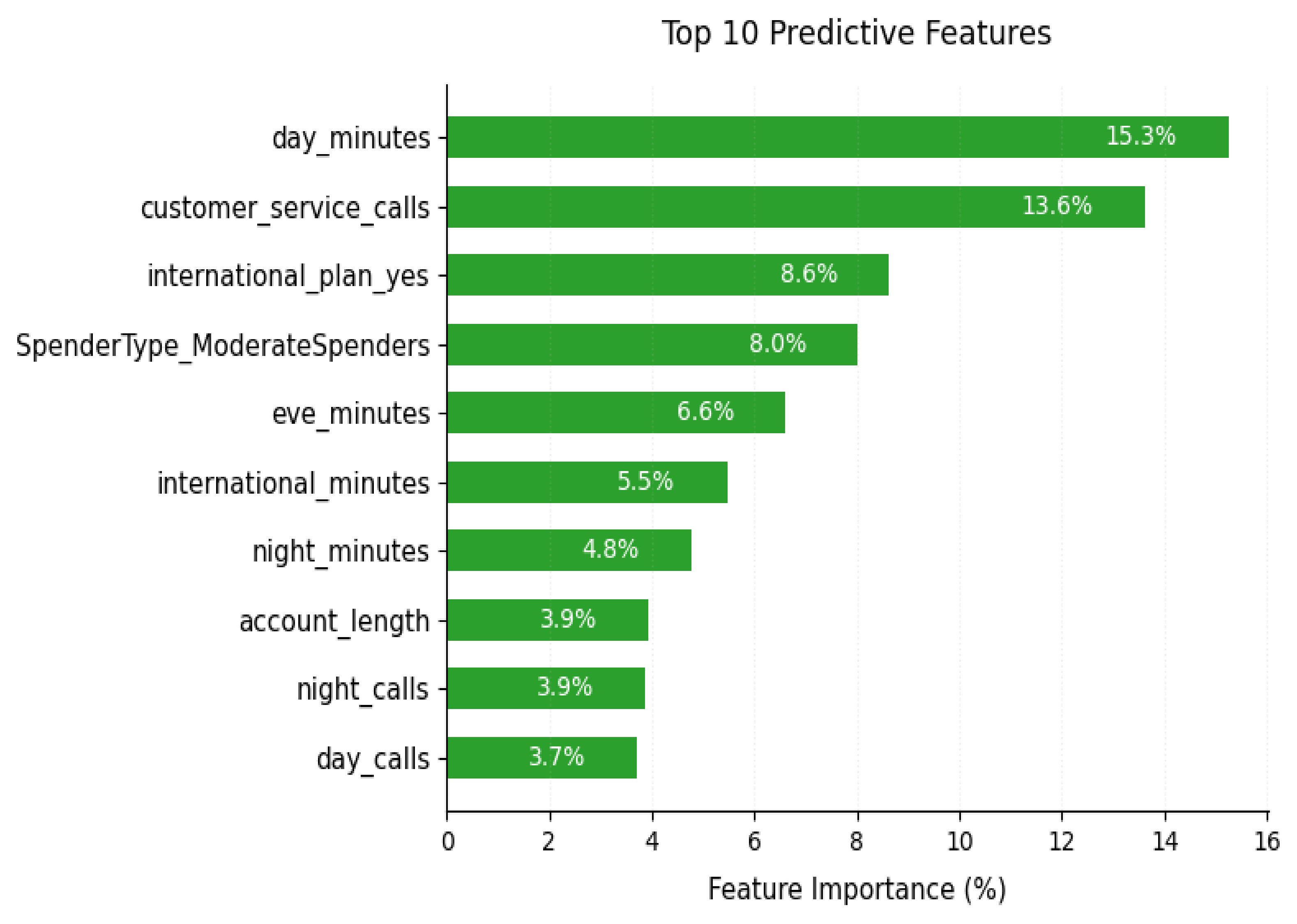
By adjusting our prediction threshold to 23%, we identified 5% more churners while only slightly increasing false alerts.

FEATURE IMPORTANCE

What Actually Predicts Churn?

Our Advanced Random Forest model ranked features by importance. Top contributors included:

- Total day minutes (15.3%)
- Customer service calls (13.6%)
- International plan (8.6%)
- Spender type(8.0%(
- Evening minutes (6.6%)



RECOMMENDATIONS

-Run localized campaigns in high-churn states (Washington, Texas) with deeper investigation into region specific issues like service quality, network coverage or billing concerns.

- Develop onboarding programs for new customers and loyalty program for long term customers to reduce early drop-offs and late disengagement.
- Proactively monitor customers with 3+ support calls and prioritize them for resolution. Train customer service team to resolve issues on the first contact to prevent frustration.
- Reassess the value proposition of the international plan. This could involve improving call quality, reducing costs, or bundling with other perks to increase satisfaction.
- Introduce spending caps or usage notifications for customers who pay more especially during daytime & International calls to help manage expectations and reduce bill shock as these users are more likely to churn.
- Have a higher budget for areas with a high churn rate for marketing.

LIMITATIONS

- SMOTE may cause overfitting, affecting real-world performance.
- Model scope was limited to three algorithms without deeper tuning.
- Some churn patterns lack context, needing more customer behavior data.
- Geographic churn trends weren't deeply explored, missing regional insights.
- Call data lacked quality indicators, limiting support-related analysis.

THANK YOU

For any inquiries, please contact me at:

gedimwas@gmail.com

Username: Gerald Mwangi