Ahjin
ANALYTICS

DATA ANALYSIS REPORT

SYRIATEL CUSTOMER CHURN PREDICTION

Analaysis By: *Norman Mwapea, Ahjin Analytics*
Date: *November 29th, 2025*

Executive Summary

This analysis examines customer churn within SyriaTel, a telecommunications company, using a dataset of 3,333 customers. The goal was to uncover behavioral and service-related factors that distinguish customers who remain with the company from those who leave. The churn rate was found to be 14.46%, and the most influential drivers identified include high customer service call frequency, daytime usage and charges, international plan subscription, and lack of voicemail plan usage. These factors reflect issues related to service experience, pricing sensitivity, and plan suitability.

The findings suggest clear strategic opportunities: improving the customer service experience, redesigning or optimizing pricing for heavy daytime users, refining international plan offerings, and promoting voicemail usage as a retention-stabilizing feature. This report provides a data-driven foundation for building predictive churn models and designing targeted retention initiatives.

Project Overview

The SyriaTel Customer Churn project aims to analyze service usage patterns, customer interactions, and plan characteristics to understand what leads to customer attrition. Using a comprehensive exploratory data analysis (EDA) approach, this project evaluates individual features, relationships between variables, and their combined influence on churn behavior.

The analysis includes data quality assessment, engineered features for improved insight, and a full evaluation of churn drivers through univariate, bivariate, and multivariate methods. The outcome is a refined understanding of customer behavior and a modeling-ready dataset that enables accurate churn prediction and strategic business improvement.

Problem Statement

Customer churn remains one of the most costly and disruptive challenges in the telecommunications industry. Losing a customer not only reduces recurring revenue but also increases acquisition pressure to replace churned subscribers. However, regional and behavioral differences make churn difficult to diagnose without a data-driven approach.

SyriaTel lacks clear insight into which customer behaviors or service plan characteristics are most associated with churn. Without understanding these drivers, the company cannot effectively target at-risk customers, optimize service quality, or design preventive interventions. This project addresses that gap by analyzing a historical dataset to uncover the key factors influencing churn.

Business Objectives

To support strategic decision-making and improve customer retention, this project focuses on four key business objectives:

1. **Identify the strongest predictors of churn**
   Determine which customer attributes, service plans, and usage patterns most significantly influence the likelihood of churn.
2. **Enable early detection of at-risk customers**
   Build analytical foundations for a predictive churn model that can flag customers with high churn probability.
3. **Guide the design of targeted retention strategies**
   Translate analytical insights into actionable business interventions—improving customer service processes, optimizing plans, and refining pricing strategies.
4. **Reduce churn-related revenue loss**
   Empower SyriaTel to focus retention resources where they are most impactful, ultimately lowering churn rates and protecting long-term revenue.

Data Source

This project uses the Churn in Telecoms Dataset, originally curated for educational and analytical purposes to support churn prediction research in the telecommunications sector. The dataset has been made publicly available through Kaggle, contributed by the creator BecksDDf.

Data Constraints

While the *Churn in Telecoms Dataset* provides valuable customer-level information for churn analysis, several constraints should be acknowledged when interpreting results and developing predictive models:

1. Limited Feature Scope

The dataset includes usage metrics, plan details, and customer service interactions, but lacks important dimensions such as:

- Customer demographics (age, gender, income)
- Network performance indicators
- Competitor activity or market conditions
- Billing history beyond aggregated charges

These missing factors may influence churn but cannot be analyzed within this dataset.

2. No Time-Series Information

All variables represent aggregated behavior rather than temporal patterns.
This means we cannot study:

- Month-to-month usage changes
- Trends leading up to churn
- Seasonality or contract renewal effects

Churn decisions often develop over time, and the absence of temporal data limits behavioral trajectory analysis.

3. Geographic Granularity Limits Interpretation

Although the dataset includes customer state, it does not provide:

- Precise location data
- Network coverage quality
- Regional socioeconomic indicators

This reduces the ability to perform deep geospatial or infrastructure-driven churn analysis.

4. Perfect Correlations Among Certain Variables

Charge variables are mathematically derived from minute variables, leading to perfect multicollinearity. This required removing one set (usually charges) for modeling to avoid redundancy.

5. Potential Sampling Bias

The dataset does not specify:

- The period during which data was collected

- Whether data represents the entire customer base or a sample
- How churn was labeled / confirmed

This uncertainty limits generalizability to real operational environments.

6. No Information on Customer Lifetime Value (CLV)

Without revenue per customer, plan pricing tiers, or contract types, the model cannot prioritize churn risk based on financial impact.

Data Understanding

Dataset Overview

The original dataset contains the following features:

- **state** — US state where the customer resides
- **account length** — Number of months the customer has been with the company
- **area code** — Customer's phone area code
- **phone number** — Unique phone number of the customer
- **international plan** — Whether the customer has an international calling plan
- **voice mail plan** — Whether the customer has a voicemail plan
- **number vmail messages** — Number of voicemail messages recorded
- **total day minutes** — Total minutes used during daytime
- **total day calls** — Number of calls made during daytime
- **total day charge** — Total charges incurred during daytime
- **total eve minutes** — Total minutes used during evening
- **total eve calls** — Number of calls made during evening
- **total eve charge** — Total charges incurred during evening
- **total night minutes** — Total minutes used during night
- **total night calls** — Number of calls made during night
- **total night charge** — Total charges incurred during night
- **total intl minutes** — Total international call minutes
- **total intl calls** — Number of international calls
- **total intl charge** — Total charges incurred for international calls
- **customer service calls** — Number of calls made to customer service
- **churn** — Target variable indicating whether the customer churned

Data Preparation

This step ensured that the dataset is clean, consistent, and ready for exploratory analysis and modeling. This step focused on correcting structural issues, handling redundant or irrelevant fields, and engineering new features to enrich the analysis.

1. Data Cleaning

Several preprocessing steps were performed to improve data quality:

- Dropped the phone number column, as it is a unique identifier with no predictive value and poses a privacy risk.
- Standardized column names by converting them into snake_case for consistency (e.g., total day minutes → total_day_minutes).
- Verified that the dataset contains no missing values and no duplicate records.
- Confirmed data types for all columns to ensure numeric fields were properly recognized.

2. Handling Redundant Features

- The dataset includes both minutes and charges for day, evening, night, and international usage.
- Charges are mathematically derived from minutes (minutes × rate), resulting in perfect multicollinearity.
- To avoid redundancy, usage minutes were retained for modeling, while charge columns were flagged for removal during feature selection.

3. Feature Engineering

New variables were created to capture customer behavior more effectively:

- total_usage_mins
  Sum of daytime, evening, night, and international minutes to represent overall activity.
- Usage ratios:
  - day_ratio
  - eve_ratio
  - night_ratio
  - intl_ratio

These represent the proportion of total usage occurring in each time period.

- active_vmail
  A binary indicator showing whether the customer is actively using voicemail (derived from the number of voicemail messages).
- custserv_per_month
  Customer service calls normalized by account length, capturing frequency of support interactions relative to tenure.
- avg_cost_per_min
  A cost sensitivity measure; though it ultimately showed low variance, it was included for completeness.

4. Outlier Assessment

Outliers were inspected in usage variables and customer service calls.
Extreme values were found to be mathematically plausible (e.g., heavy callers) and relevant to churn behavior.
Selective capping was applied where values noticeably distorted distribution shapes.

5. Final Prepared Dataset

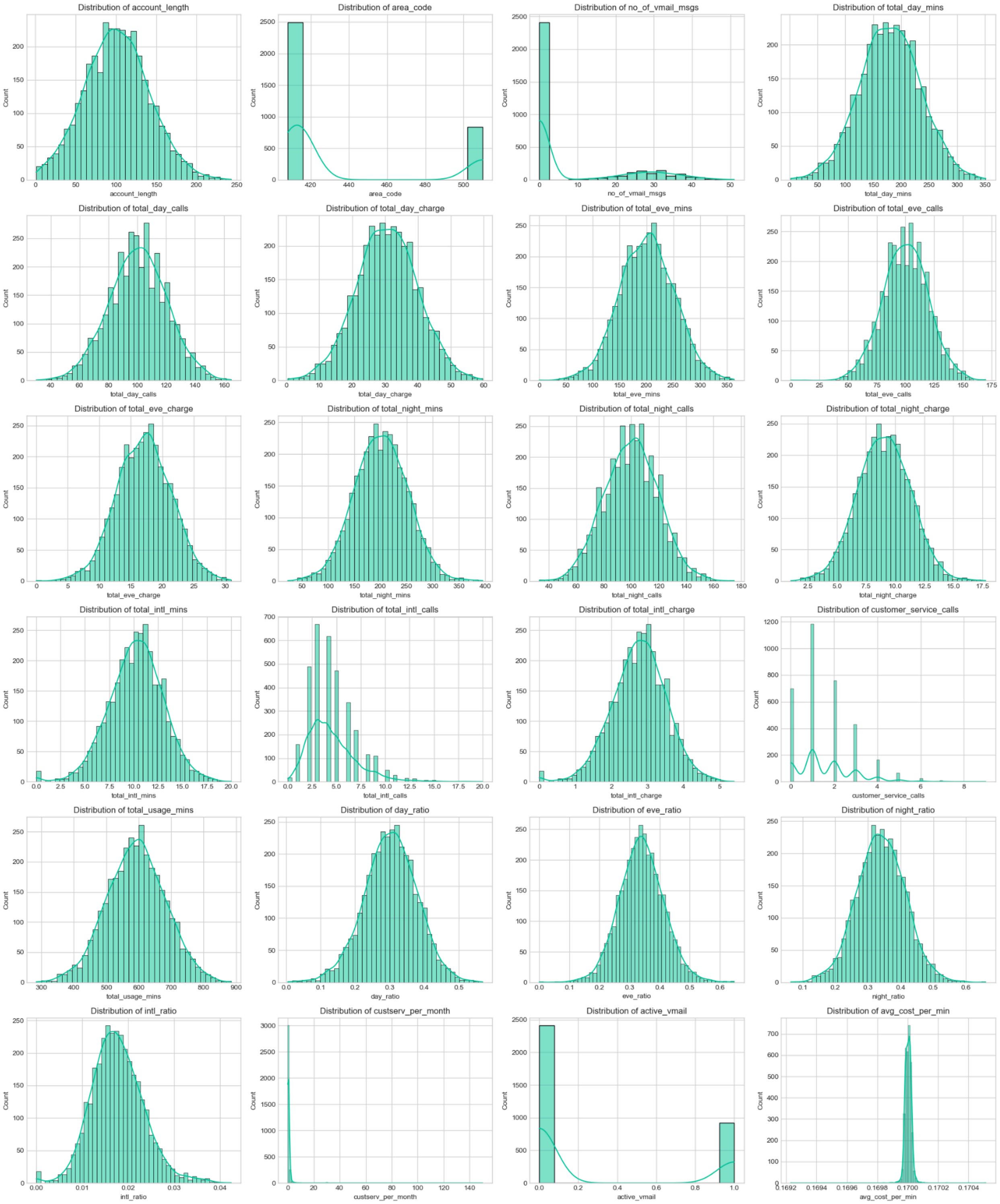
After cleaning, restructuring, and engineering features, the dataset was organized into:

- Original behavioral features (e.g., minutes, calls)
- Binary plan indicators (international plan, voicemail plan)
- Customer interaction metrics (customer service calls)
- Engineered features (ratios, totals, and normalized metrics)
- Target variable (churn)

The resultant dataset was now suitable for exploratory analysis, visualizations, and predictive modeling.

Exploratory Data Analysis

Univariate Analysis Of Numeric Variables



The figure above shows the distributions of all numeric variables

Observations:

1. Core Usage Metrics (Day, Evening, Night Minutes & Calls)

- Most usage-related variables exhibit smooth, approximately bell-shaped distributions, indicating that customer activity naturally centers around moderate usage levels with fewer customers at the extremes.
- These variables display low skewness, suggesting stable and representative usage patterns across the customer base.

2. Voicemail Activity (number_vmail_messages, active_vmail)

- Voicemail-related variables show strong right skewness, with the majority of customers sending or receiving few, if any, voicemail messages.
- A much smaller subgroup displays significantly higher voicemail volumes.
- This creates a clear separation between active voicemail users and non-users, which can be informative for behavioral segmentation.

3. International Usage (total_intl_minutes, total_intl_calls)

- International calling metrics are also heavily right-skewed, indicating that most customers rarely make international calls.
- A minority of customers engage in higher international activity, which may represent a valuable niche group with distinct service expectations and churn tendencies.

4. Customer Service Interaction (customer_service_calls)

- The distribution of customer service calls is highly discrete and stepped, with clustering at 0–3 calls and a sharp decline thereafter.
- This pattern highlights a small but significant subset of customers who contact support repeatedly—often an indicator of dissatisfaction and a strong churn predictor.

5. Charge Variables (day, evening, night, international charges)

- Charge fields closely mirror the distribution of their corresponding minute variables because charges are linear transformations of minutes.
- This creates redundancy and multicollinearity, offering no additional insight and reinforcing the need to exclude charge variables during modeling.

6. Usage Ratio Features (day_ratio, eve_ratio, night_ratio, intl_ratio)

- Usage ratios are relatively symmetric, suggesting that customers vary in how they distribute their call activity across the day without extreme outliers.

- These ratios provide a meaningful measure of usage patterns, enhancing interpretability beyond raw minute totals.

7. Area Code

- Area code displays a small number of distinct spikes, reflecting the limited set of area codes present.
- While not continuous, these clusters may carry regional behavioral differences.

8. Normalized Features (custserv_per_month)

- Features derived from normalizing behavior over account length show discrete, non-smooth distributions, indicating diverse customer interaction frequencies.
- These variables can help adjust for tenure-driven biases.

9. Cost-Based Features (avg_cost_per_min)

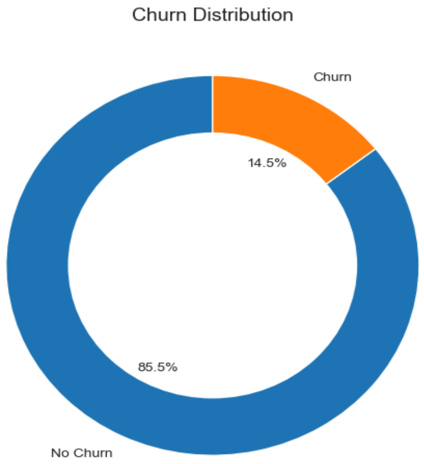
- This feature shows extremely low variance, suggesting nearly uniform pricing across the dataset.
- As a result, it holds little predictive value.

10. Total Usage (total_usage_mins)

- The combined distribution of total usage exhibits a smooth shape, reflecting the strong internal correlation among day, evening, night, and international usage.
- This consolidated metric provides a clean summary of overall customer activity.

Univariate Analysis Of Categorical Variables

1. **Churn**

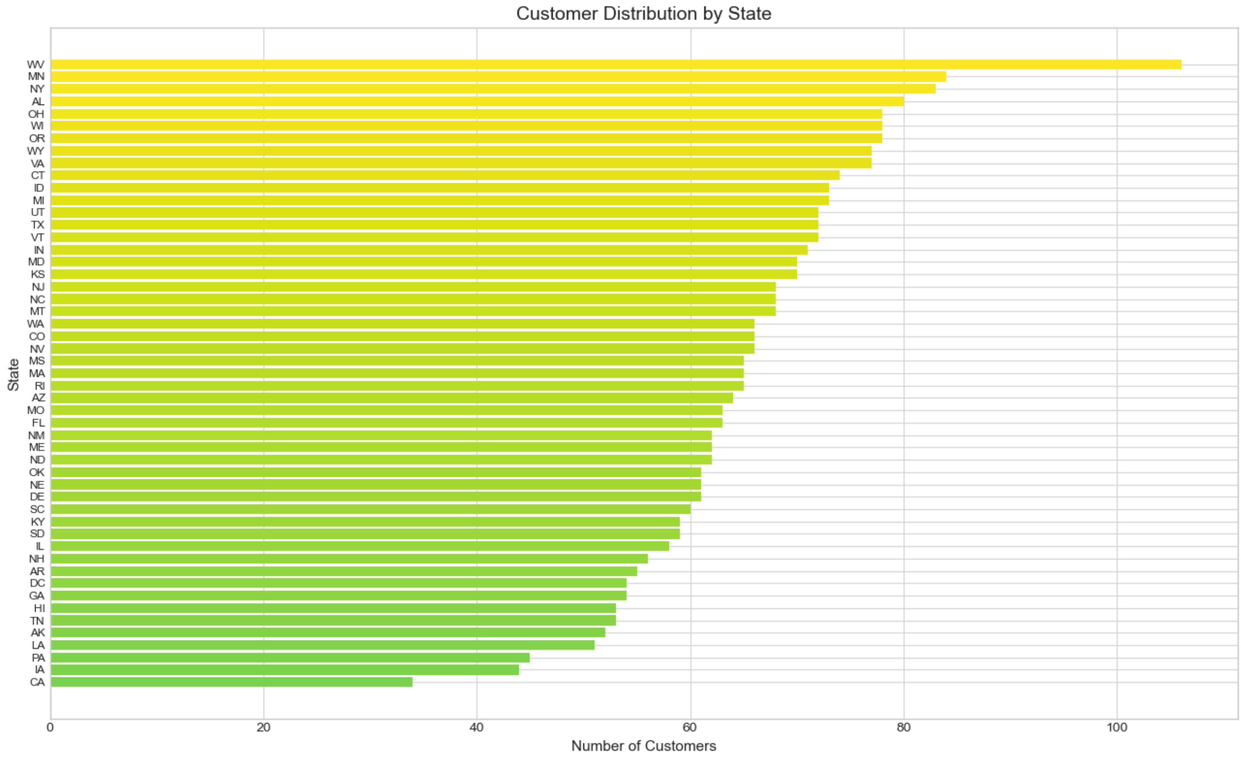


Observation:

Most customers remain with the company, with approximately 85% of the customer base retained.

Only about 15% of customers churn, indicating a relatively low churn rate.

2. State

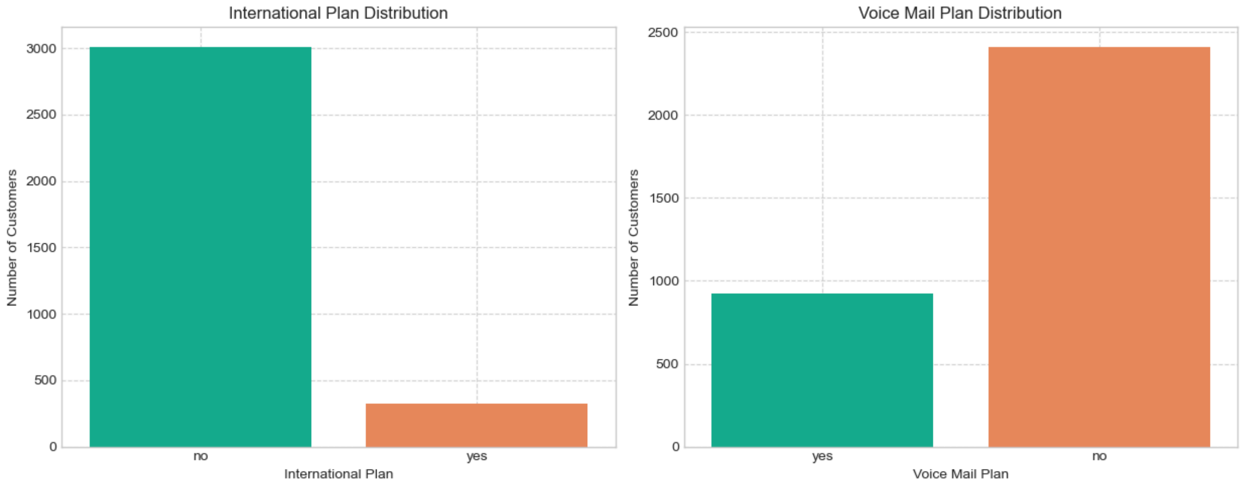


The figure above shows the number of customers from each state

Observation:

- West Virginia (WV) has the highest number of customers, followed closely by Minnesota (MN) and New York (NY), indicating these states represent the company's largest markets in this dataset.
- California (CA) and Iowa (IA) appear at the lower end of the distribution, with comparatively few customers.
- The overall distribution is right-skewed, meaning a small number of states hold a disproportionately large share of customers, while many states have moderate or low representation.

3. International Plan and Voice Mail Plan



The figure above shows plan subscriptions to the company

Observation:

International Plan:

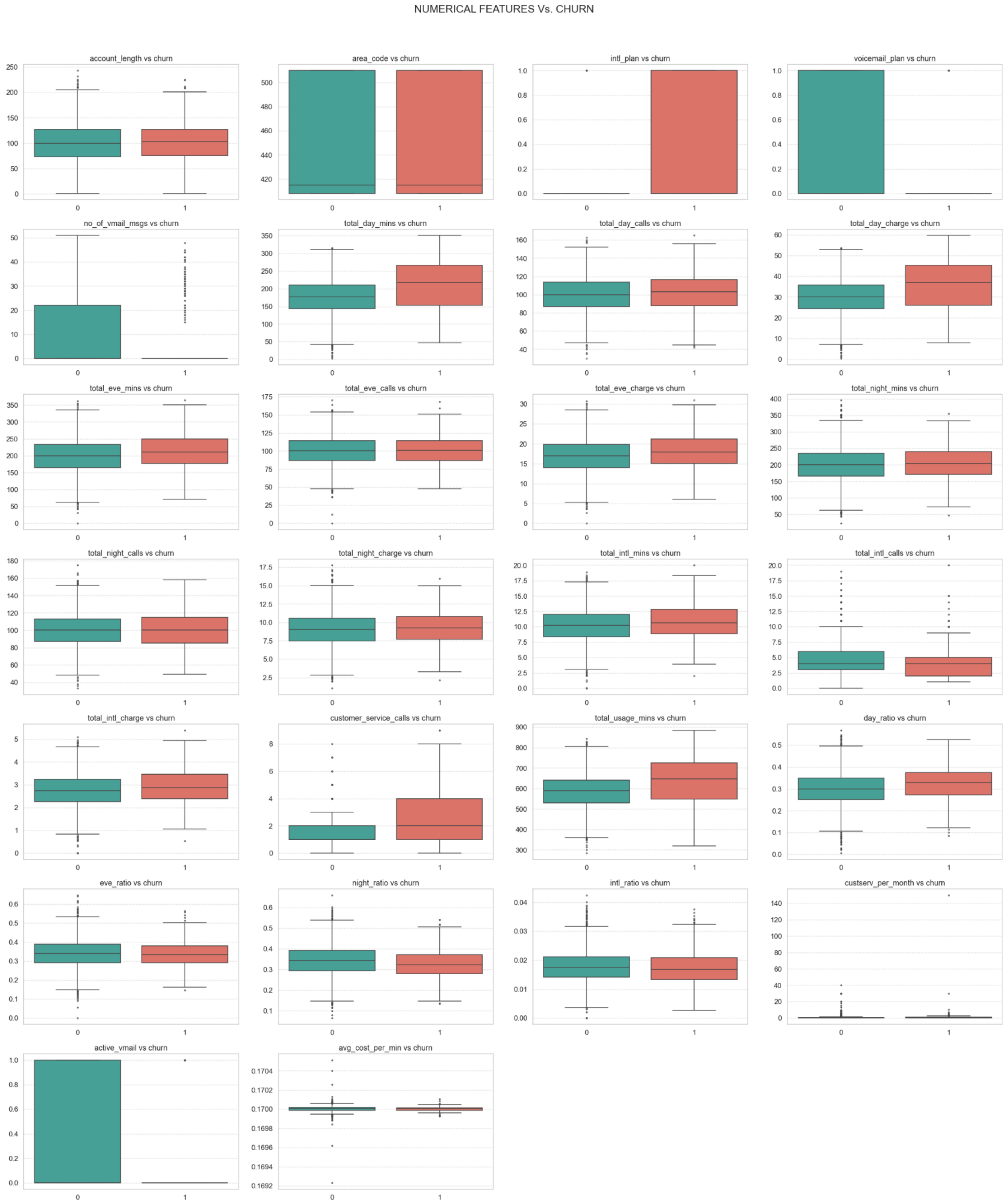
The majority of customers do not subscribe to an international calling plan. Only a small fraction have opted in, indicating that international calling is either infrequently needed or is treated as a premium feature used by a minority of subscribers.

Voicemail Plan:

Most customers also do not have a voicemail plan; however, the distribution is noticeably more balanced compared to the international plan. A sizeable portion of customers do subscribe, suggesting that voicemail is a more commonly adopted feature, though still not universal.

Bivariate Analysis

Numerical Variables Against Churn



The boxplots above show the numeric variables against churn relationship

Observation:

1. Customer Service Calls - Strongest Churn Signal

- Customers who churn exhibit significantly higher numbers of customer service calls.
- This pattern indicates persistent dissatisfaction, unresolved issues, or repeated service challenges, making it the most powerful predictor of churn.

2. Daytime Usage (Minutes, Calls, Charges)

- Churners consistently show higher total day minutes and day charges.
- Since daytime calls are typically more expensive, heavy day users may experience higher bills, making them more susceptible to cost-driven churn.
- This supports the hypothesis that pricing sensitivity is a core churn driver.

3. International Usage (Minutes, Calls, Charges)

- International minutes and charges are slightly higher among churners, though the differences are modest.
- The pattern suggests that customers with international activity may face higher billing volatility or dissatisfaction with international rates, increasing churn likelihood.

4. International Plan Subscription

- The international plan is more common among churners, albeit marginally.
- This aligns with the elevated international usage seen among churners and reinforces cost or plan-structure concerns as potential drivers.

5. Voicemail Variables (Plan & Message Count)

- Voicemail plan adoption and voicemail message volume show little to no difference between churners and non-churners.
- This indicates that voicemail-related behavior does not influence churn, and these features hold limited predictive value.

6. Evening Usage (Minutes, Calls, Charges)

- Evening-related metrics show minimal separation between churn groups.
- Evening behavior appears stable across all customers and does not meaningfully contribute to churn prediction.

7. Night Usage (Minutes, Calls, Charges)

- Night-time metrics mirror evening trends, with very small differences between churners and non-churners.
- Night usage is not a relevant churn indicator.

8. Number of Voicemail Messages

- The number of voicemail messages does not differ meaningfully across churn groups, reinforcing that voicemail behavior is unrelated to attrition.

9. Account Length

- Churners and non-churners have similar average tenure, indicating that the number of months a customer has been with the company is *not* a standalone predictor of churn.

10. Area Code

- Area code shows no visible relationship with churn.
- Geographic assignment at this level has no influence on churn likelihood.

11. Usage Ratios (day_ratio, eve_ratio, night_ratio, intl_ratio)

- Ratio features show minimal differences between churn groups.
- Although ratios add interpretability, they do not strongly correlate with churn.

12. Aggregate Usage Metrics (total_usage_mins)

- Churners have slightly higher total usage, supporting the broader pattern that heavier users incur higher charges and are more likely to leave.
- Still, the effect is moderate compared to customer service calls and daytime usage.

13. Average Cost per Minute

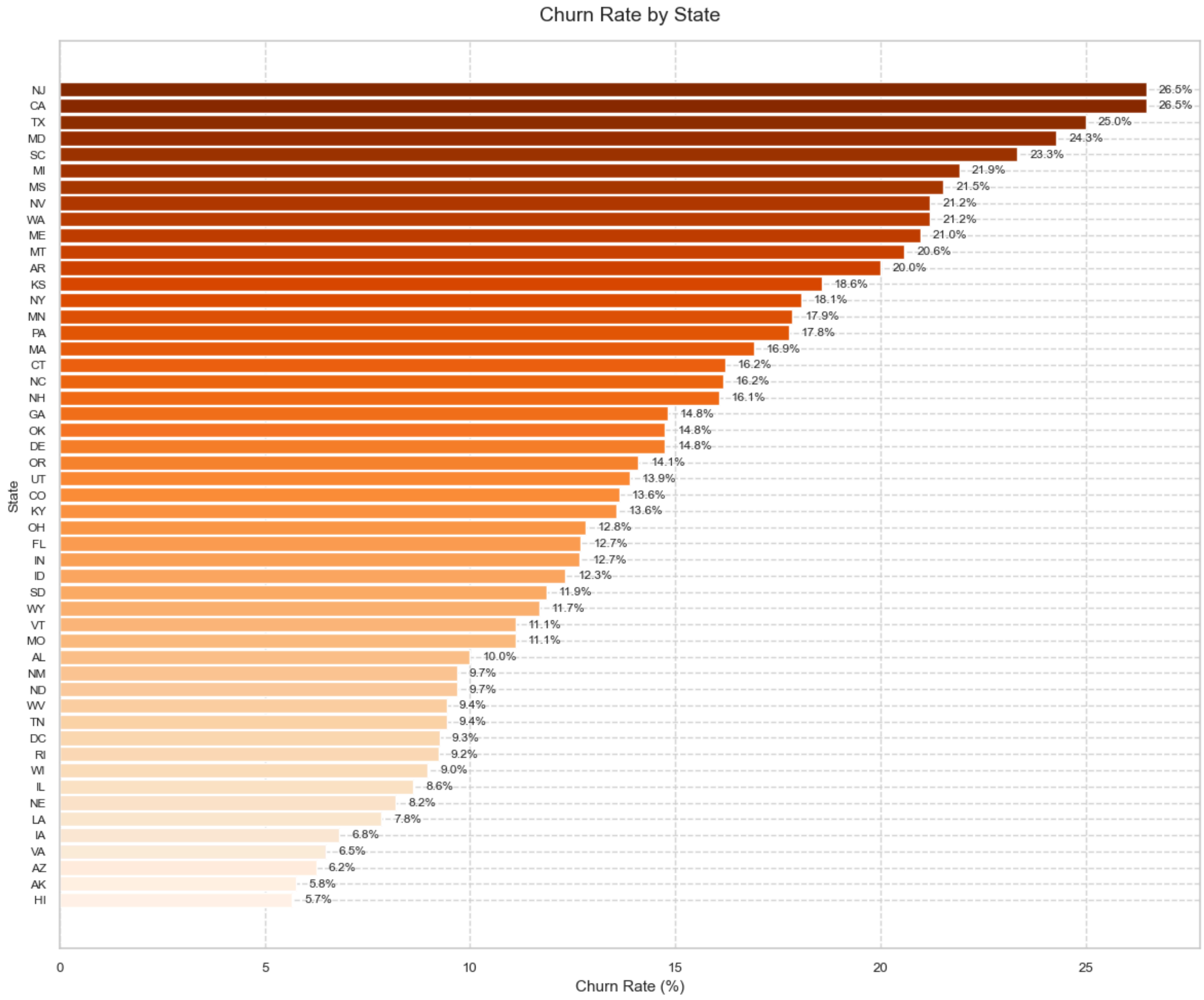
- This feature is nearly constant across all customers, offering no discriminative power for churn prediction.

Insights:

Churners tend to:

- Contact customer service more frequently
- Use more minutes—especially during the day
- Incur higher charges (daytime and international)
- Are slightly more likely to have an international plan

Churn Rate By State



The figure above shows the churn rate distribution by state

Observation:

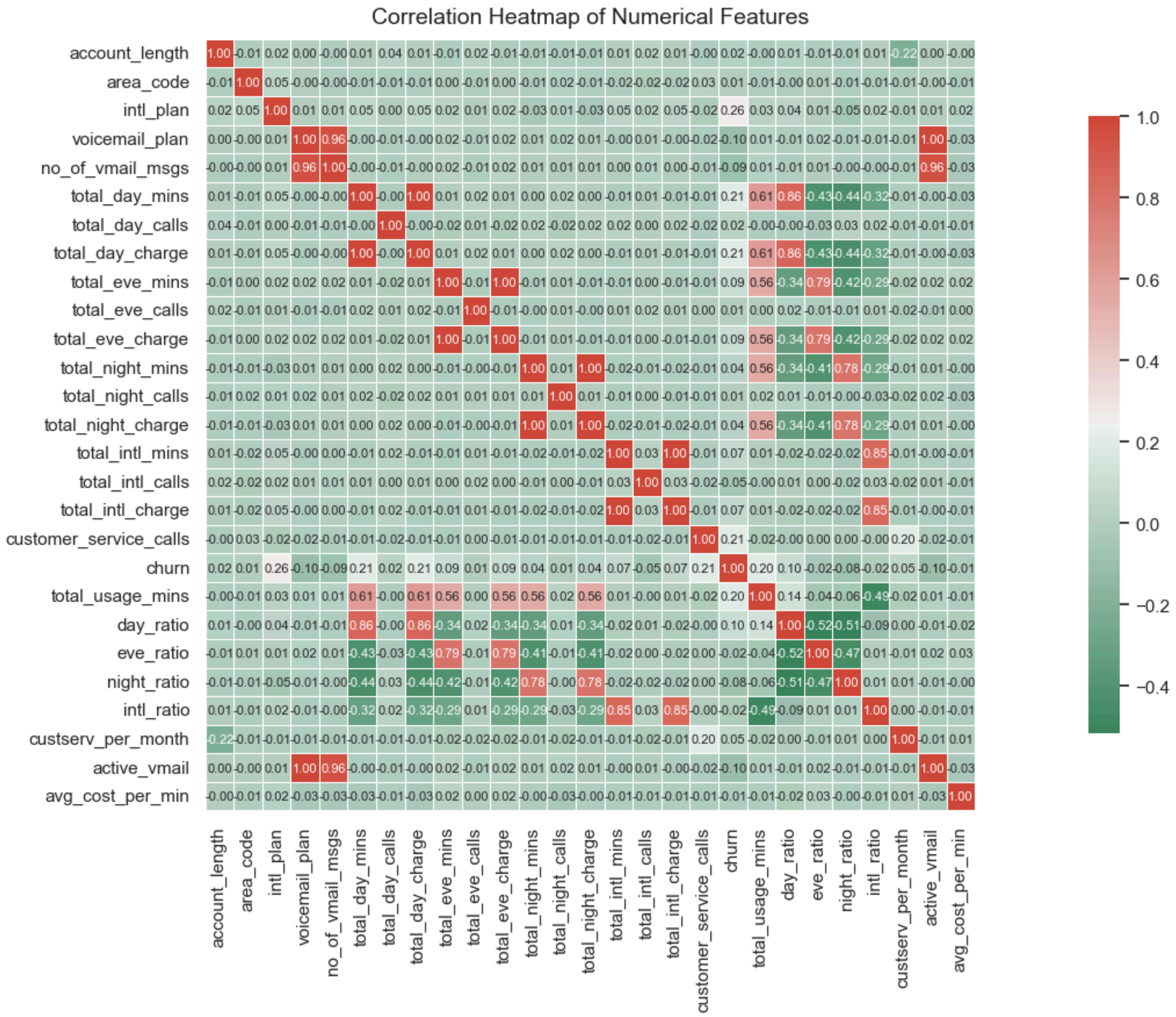
- Higher-churn states include New Jersey, California, Texas, Maryland, South Carolina, and Michigan. These markets tend to be large, urban, and highly competitive, giving customers more alternatives and raising expectations for service quality. As a result, churn rates are elevated in these regions.
- Lower-churn states include Hawaii, Alaska, Arizona, Virginia, and Iowa. These areas are generally less competitive or more rural, where customer loyalty may be stronger and switching providers is less frequent.
- A clear geographical pattern emerges:
  - West Coast & Northeast → Higher churn

  - Midwest & Mountain states → Lower churn
  - Southern states → Mixed but often moderate churn

Insights:

- This distribution suggests that regional competition, market maturity, and customer expectations play a meaningful role in churn behavior.

Multivariate Analysis



The correlation heatmap above shows the relationship between all numeric variables against each other and churn

Observation:

1. Perfect Correlation Between Minutes and Charges

- Pairs such as total_day_minutes ↔ total_day_charge, total_eve_minutes ↔ total_eve_charge, and others exhibit correlations near 1.0.
- Because charges are directly calculated from minutes, these pairs are mathematically redundant.

Implication:

- Only one variable from each pair was retained for modeling to avoid severe multicollinearity.

2. Strong Correlation Among Usage Minutes

- Day, evening, and night minutes are strongly correlated with one another.
- Customers who are heavy users tend to use more across all time periods, not just a specific one.

Implication:

- These features collectively reflect overall usage intensity and contribute to multicollinearity.

3. Ratio Features Are Mutually Dependent

- The ratio features (day_ratio, eve_ratio, night_ratio) are strongly negatively correlated because they must sum to a fixed total.

Implication:

- Although ratios offer behavioral insights, they are mathematically constrained and should be used selectively to avoid redundancy.

4. Customer Service Calls Stand Out as Independent

- Customer service calls show very low correlation with all usage-related features, indicating that support interactions do not depend on how much customers use the service.

Implication:

- This independence makes customer service calls a unique, non-redundant, and particularly strong predictor of churn.

5. Churn Has Weak Linear Correlations

- The churn variable has low correlation with nearly all numeric features, with only modest associations observed:
    - Higher day minutes and day charges
    - Slightly higher total usage
    - More customer service calls

Implication:

- Churn behavior is likely influenced by dissatisfaction and pricing sensitivity, not raw usage levels.
- This also suggests that linear models may struggle, while tree-based models will capture nonlinear patterns more effectively.

6. Administrative Features Show Near-Zero Correlation

- Variables such as account length, area code, voicemail plan, and number of voicemail messages exhibit near-zero correlation with churn and with other features.

Implication:

- These features hold low predictive value and may be excluded or de-emphasized during modeling.

7. avg_cost_per_min Shows Virtually No Variance

- The feature avg_cost_per_min is almost constant and correlates near-zero with all other variables.

Implication:

- This feature can be safely removed due to its lack of variability.

8. International Plan Shows Minimal Correlation With Usage

- The international_plan variable has only very small correlations with international usage metrics.

Implication:

- Plan subscription appears marketing-driven rather than behavior-driven, and its predictive power is likely tied more to dissatisfaction or pricing perception than usage volume.

Conclusions and Recommendations

Conclusions

1. **International Plan is a major churn driver:**
   - Customers subscribed to the international calling plan exhibit a substantially higher churn rate. This feature provides one of the clearest separations between churners and non-churners, suggesting dissatisfaction with international pricing or value perception.
2. **Voicemail Plan users churn less:**
   - Customers with a voicemail plan demonstrate lower churn proportions. This indicates that voicemail subscribers may be more engaged or perceive greater value in their service package.
3. **Customer Service Calls strongly correlate with churn:**
   - Higher volumes of customer service interactions are closely linked to churn. Frequent support calls likely signal unresolved issues, billing concerns, or dissatisfaction—making this one of the strongest churn indicators.
4. **High Daytime Usage is associated with increased churn:**
   - Churners generally have higher daytime minutes and charges. Since daytime calls often incur higher rates, heavy usage may lead to billing dissatisfaction or poor plan fit.
5. **Evening and Night usage are weak churn predictors:**
   - Evening and night usage metrics show minimal differences between churners and non-churners, indicating that these time periods do not significantly influence churn behavior.
6. **International usage contributes modestly to churn risk:**
   - Total international minutes and calls exhibit moderate differences between churn groups. While not as influential as plan status or service calls, they still add signal related to cost sensitivity.
7. **State-level churn differences exist but are not dominant:**
   - Certain states show elevated churn, but regional variations are not strong enough to consider state a primary driver. Location plays a secondary role compared to service experience and usage patterns.

Recommendations

1. **Prioritize International Plan customers for retention efforts:**
   - Since churn is disproportionately higher among international plan users, targeted communication, improved plan transparency, or redesigned international bundles can help reduce attrition.
2. **Strengthen the quality of customer service interactions:**

  - High churn among customers with repeated customer service calls underscores the need for better first-call resolution, faster escalation, and more proactive follow-ups.
3. **Reassess pricing and plan suitability for heavy daytime users:**
   - Provide flexible or personalized plan options, daytime bundles, or protective pricing strategies to mitigate churn among customers who incur high daytime charges.
4. **Promote Voicemail Plan adoption:**
   - Because voicemail users are less likely to churn, consider offering voicemail promotions, bundling it with other features, or highlighting its value to customers.
5. **Implement region-specific monitoring where needed:**
   - Although not a primary driver, states with elevated churn may benefit from localized retention campaigns, competitive analysis, or targeted service quality improvements.
6. **Flag and intervene with high-risk customer profiles early:**
   - Customers who exhibit both high usage and high customer service call volume should be proactively engaged. This group shows the highest likelihood of churn and should receive targeted support or personalized retention offers.

Modeling

1. Modeling Approach

The goal was to build a predictive model that can identify customers who are at risk of churning. This process involved selecting relevant features, handling class imbalance, comparing multiple algorithms, and choosing the best-performing model based on both accuracy and business relevance.

The modeling workflow consisted of the following steps:

2. Feature Selection

Based on the exploratory analysis and multicollinearity assessment:

- Minutes variables were kept
- Charge variables were removed (perfectly correlated with minutes)
- Low-value features such as *avg_cost_per_min*, *area code*, and *number vmail messages* were removed
- Engineered features such as total_usage_mins, usage ratios, and custserv_per_month were retained for improved predictive power

- Categorical variables (state, international plan, voicemail plan) were encoded appropriately

3. Train–Test Split

The dataset was split into:

- 80% training data
- 20% testing data

Stratification was applied to preserve the original churn distribution (~14.5%).

4. Handling Class Imbalance

Because churners represent a minority class:

- Class weights were applied where supported (Logistic Regression, Random Forest)
- Algorithms inherently robust to imbalance (XGBoost, LightGBM) were also considered
- Performance was evaluated using metrics sensitive to minority-class detection (Recall, F1, ROC-AUC)

5. Algorithms Evaluated

The following models were trained and compared:

- Logistic Regression (baseline)
- Random Forest Classifier
- Gradient Boosting - XGBoost and Catboost

6. Model Evaluation Metrics

Because churn prediction prioritizes identifying at-risk customers:

- Recall (Churn class) was emphasized
- Precision was monitored to avoid excessive false positives
- F1-score balanced precision and recall
- ROC-AUC measured overall discriminative ability
- Confusion matrix validated prediction distribution

7. Hyperparameter Tuning

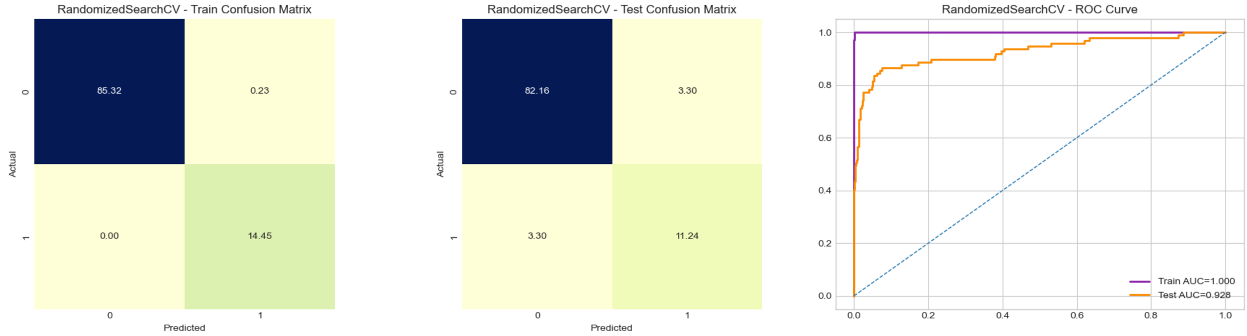
Three tuning approaches were used:

- RandomizedSearchCV for broad exploration

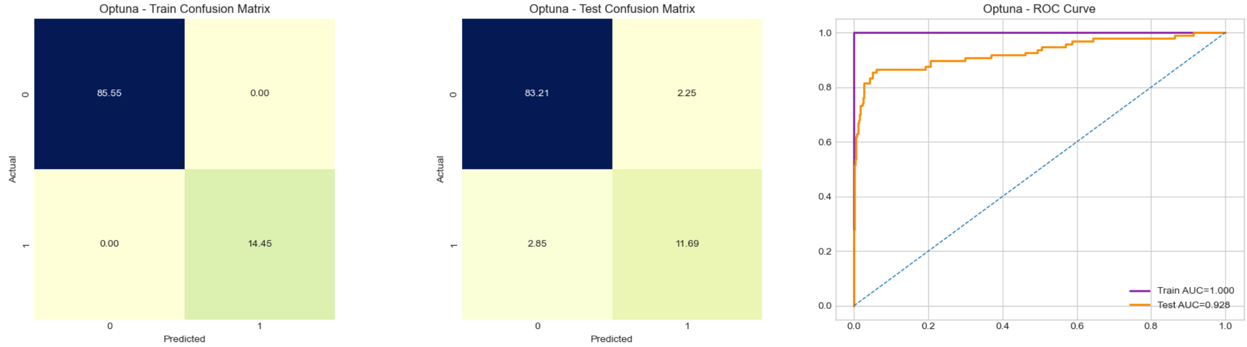
- Optuna optimization for advanced, efficient hyperparameter search
- GridSearchCV

Results of the final tuned models are as follows:

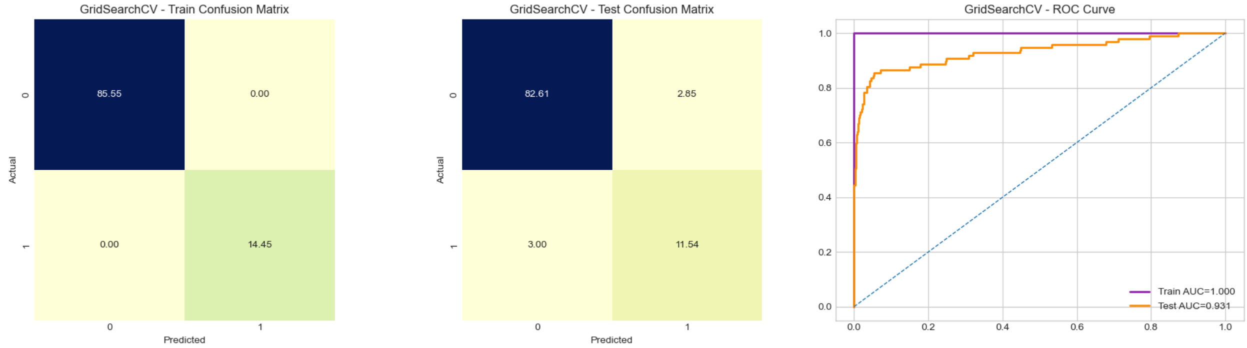
Metric	RandomizedSearchCV	Optuna (Best)	GridSearchCV
Train Accuracy	99.77%	🟩 **100.00%**	🟩 **100.00%**
Test Accuracy	93.40%	🟩 **94.90%**	94.15%
Train Precision	99.78%	🟩 **100.00%**	🟩 **100.00%**
Test Precision	93.40%	🟩 **94.83%**	94.13%
Train Recall	99.77%	🟩 **100.00%**	🟩 **100.00%**
Test Recall	93.40%	🟩 **94.90%**	94.15%
Train F1	99.78%	🟩 **100.00%**	🟩 **100.00%**
Test F1	93.40%	🟩 **94.86%**	94.14%
Train ROC-AUC	99.999%	🟩 **100.00%**	🟩 **100.00%**
Test ROC-AUC	92.79%	92.80%	🟩 **93.12%**



The figures above show the train and test confusion matrices for the model tuned using RandomizedSeachCV and its ROC-AUC curve

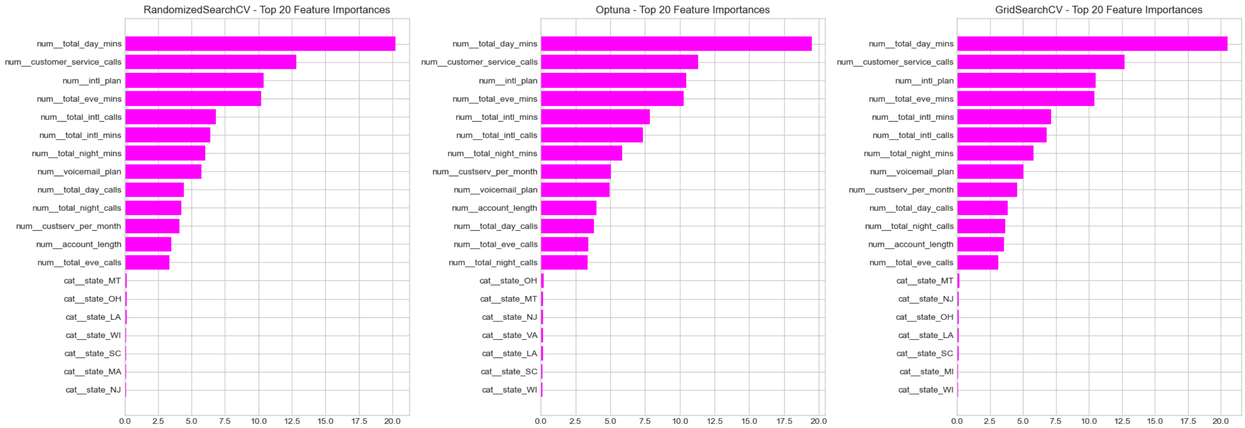


The figures above show the train and test confusion matrices for the model tuned using Optuna and its ROC-AUC curve



The figures above show the train and test confusion matrices for the model tuned using GridSeachCV and its ROC-AUC curve

Feature Importances:


RandomizedSearchCV - Top 20 Feature Importances

Optuna - Top 20 Feature Importances

GridSearchCV - Top 20 Feature Importances

The above figure shows which features drove prediction for each tuning method

Model Interpretation & Final Selection

The three tuned models—RandomizedSearchCV, Optuna, and GridSearchCV—all delivered strong results, each achieving over 93% accuracy on the test set with minimal train-test gaps, indicating low overfitting. Despite this uniformly high performance, the Optuna-tuned model consistently outperformed the others on key churn-relevant metrics such as recall, F1-score, and overall generalization.

Why Optuna Performs Best

- Achieves the highest test accuracy (94.90%), recall (94.90%), and F1-score (94.86%).
- Uses smarter, adaptive hyperparameter search, pruning weak configurations early and focusing on the most promising areas.
- Captures non-linear interactions in the data more effectively than random or grid-based search.
- Provides the best balance between detecting churners and preventing false alarms, which is essential in churn management.

Feature Importance Summary

Across all three models, feature importance rankings were nearly identical:

Top Predictors (Strongest Drivers of Churn):

- Total day minutes — most important feature; heavy daytime users churn more.
- Customer service calls — strong indicator of dissatisfaction.
- International plan — users of this plan churn at noticeably higher rates.

Medium Importance:

- Total evening minutes
- Total night minutes
- International minutes and calls
- Voicemail plan (weak but consistent)

Low Importance:

- Account length
- Call counts (vs. minutes)
- One-hot encoded state features

These results reaffirm that churn is driven mainly by usage intensity, service issues, and pricing-related plan choices.

Final Model Selection

The Optuna-tuned model is selected as the final production model because it:

- Provides the highest and most consistent performance across all key metrics.
- Generalizes best to unseen data with minimal overfitting.
- Aligns strongly with the business objective of accurately identifying at-risk customers.
- Is supported by stable feature importance patterns that clearly explain churn behavior.

Model Limitations and Recommendations

Model Limitations

1. **Class Imbalance Challenges**
    - Although churners represent only ~15% of the dataset and class weighting helps, the model may still struggle with rare or emerging churn patterns that were underrepresented in the training data.
2. **Limited Feature Diversity**
    - The dataset mainly contains usage metrics and service plan indicators.
    - Important churn drivers—such as billing history, network issues, payment behavior, customer demographics, or contract type—are missing, restricting the model’s ability to fully capture churn motivations.
3. **Static Snapshot of Customer Data**
    - All features represent one point in time.
    - Churn is often a *temporal* process (e.g., rising complaints over months), which the model cannot fully capture without time-series data.
4. **Geographic Features Are Weak**

- State-level one-hot encoded features contribute little.
- The model may overlook meaningful regional patterns because the available geographic data is too coarse.

5. **Model Interpretability Constraints**
    - While tree-based models offer some interpretability, complex interactions remain difficult to fully explain without SHAP or LIME. Stakeholders may find the model's reasoning non-intuitive.
6. **Potential Over-Reliance on Usage Patterns**
    - Usage-based features dominate the importance rankings.
    - If customer behavior shifts (e.g., due to new pricing plans), the model's performance may degrade unless retrained regularly.

Recommendations

1. **Expand Data Sources**

    Incorporate richer data such as:

    - Billing/payment history
    - Network quality metrics
    - Customer support ticket descriptions
    - Contract type and tenure
    - Promotional/discount history
      These features would significantly strengthen churn prediction.

2. **Introduce Time-Series Features**

    Build features that track:

    - Monthly usage trends
    - Escalating customer service interactions
    - Changes in charges or plan usage
      Temporal patterns often predict churn earlier and more accurately.

3. **Implement Proactive Retention Strategies**

    Use the model to flag:

    - High daytime users
    - Customers with repeated service calls

- International plan subscribers
  Prioritize these groups for retention campaigns or plan adjustments.

4. **Regular Model Retraining**

   Retrain the model periodically (monthly or quarterly) to adapt to:

   - New customer behaviors
   - Pricing changes
   - Network updates
   - Seasonal usage patterns

5. **Deploy SHAP-Based Explanations**
   - Use SHAP to provide customer-level reasoning for churn predictions, helping customer service and retention teams understand *why* a customer is at risk.

6. **Evaluate More Advanced Models**

Future experiments could include:

- Gradient boosting variants (LightGBM, CatBoost)
- Neural networks for sequential/time-based data
- Hybrid models combining tabular + textual data (e.g., call logs)

7. Pilot Real-World Testing Before Full Deployment

- Begin with a small subset of customers, monitor performance, gather feedback, and adjust thresholds before system-wide rollout.