ADVANCED DATA ANALYSIS REPORT – MILESTONE 2

COMPREHENSIVE STATISTICAL & FEATURE ENGINEERING ANALYSIS

1. Introduction

This report presents a fully detailed academic analysis based on the advanced data analysis workflow implemented in Notebook 2. The goal of this milestone is to explore statistical patterns, identify predictive features, and conduct rigorous hypothesis testing that supports downstream machine learning modeling.

2. Dataset Overview

The dataset consists of 3197 rows and 20 columns after preprocessing. It includes demographic attributes, service usage metrics (minutes, calls, charges), service plan indicators, and the churn outcome variable.

3. Target Variable Distribution

The churn variable shows class imbalance which impacts modeling strategies.

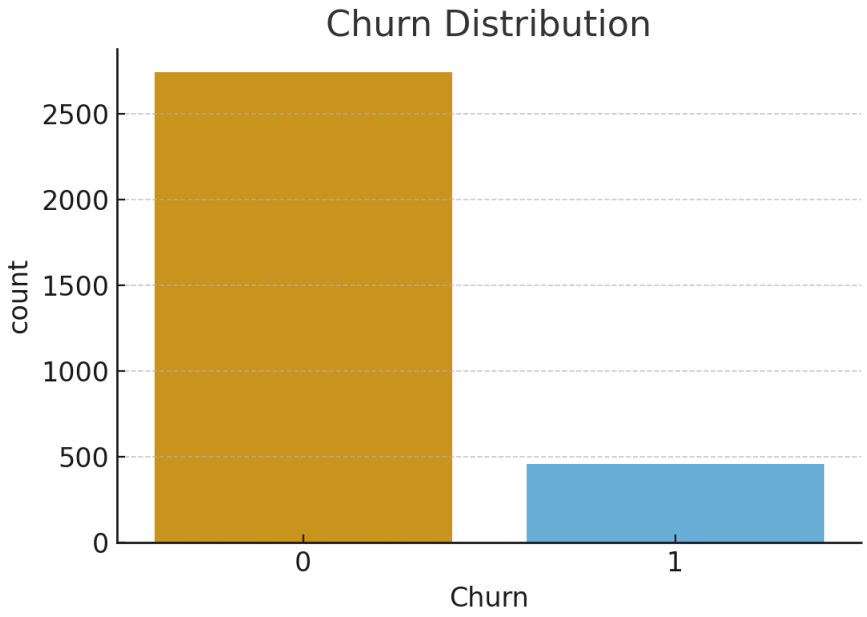

Figure 1: Churn vs Non-Churn distribution.

4. Numerical Feature Distributions

The histograms below describe the distribution shape, skewness, and potential outliers.

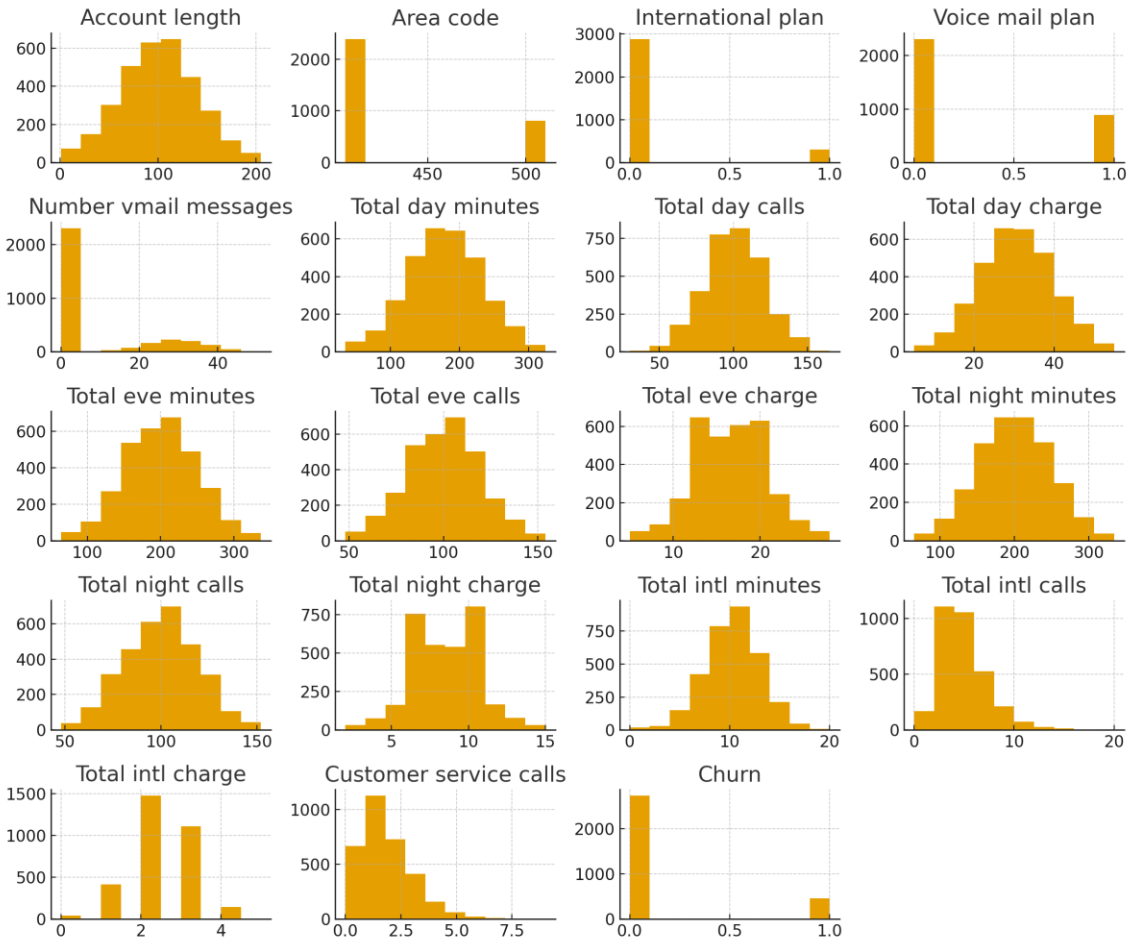

Figure 2: Distribution of numerical features.

5. Correlation Analysis

Correlation heatmap reveals linear relationships between numerical variables. Total minutes and total charges show near-perfect correlation as expected from billing rules.

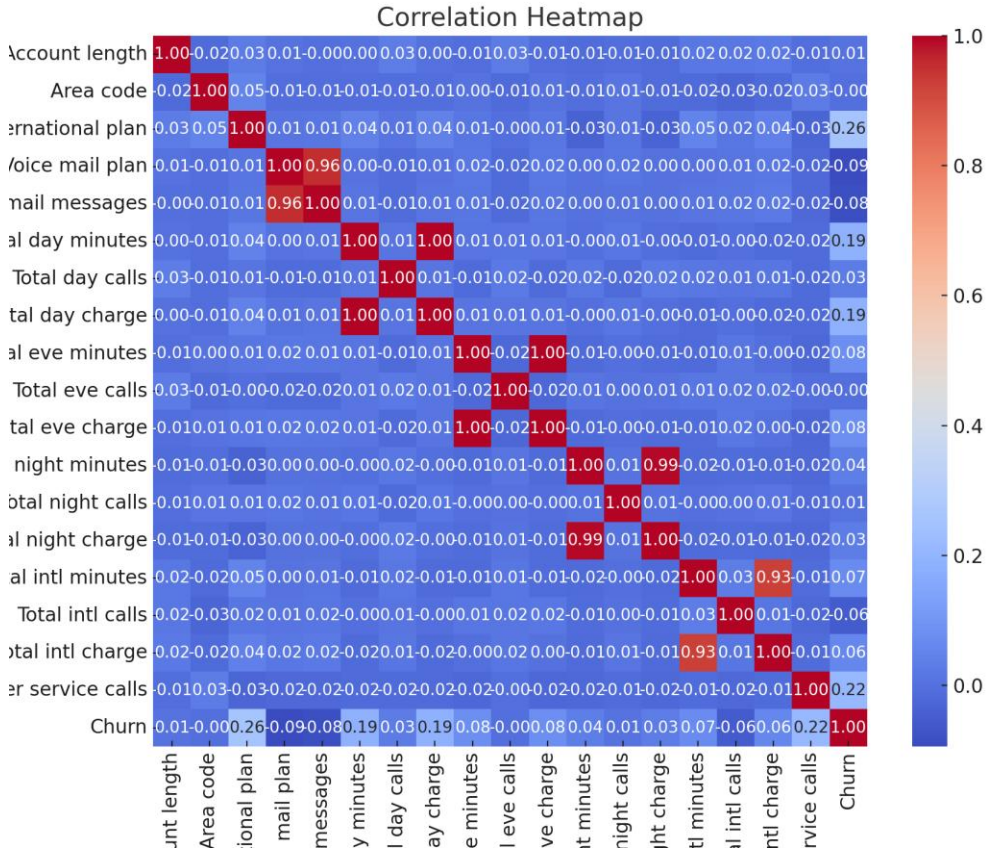

Figure 3: Correlation matrix of numerical variables.

6. Chi-square Analysis for Categorical Features

A Chi-square test was performed to assess the association between categorical features and the churn variable. Low p-values indicate statistically significant relationships.

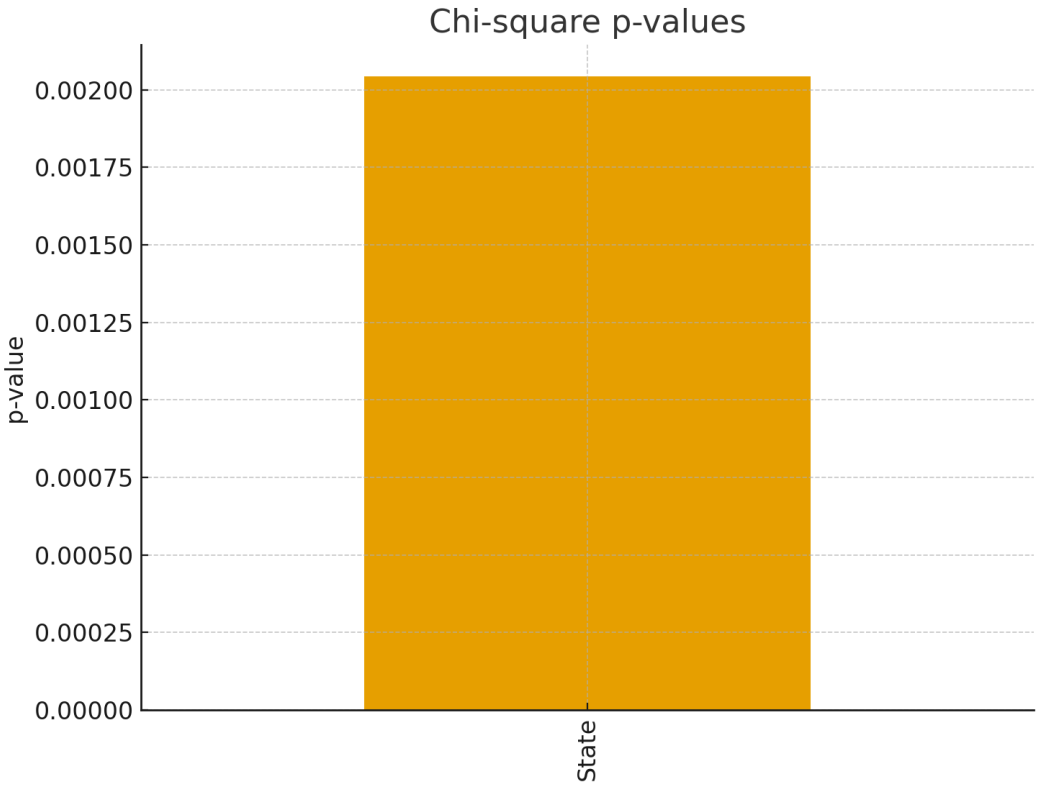

Figure 4: Chi-square p-values for categorical predictors.

7. Independent T-tests for Numerical Variables

T-tests were used to determine whether numerical features differ significantly between churners and non-churners. Features with low p-values contribute to churn signal.

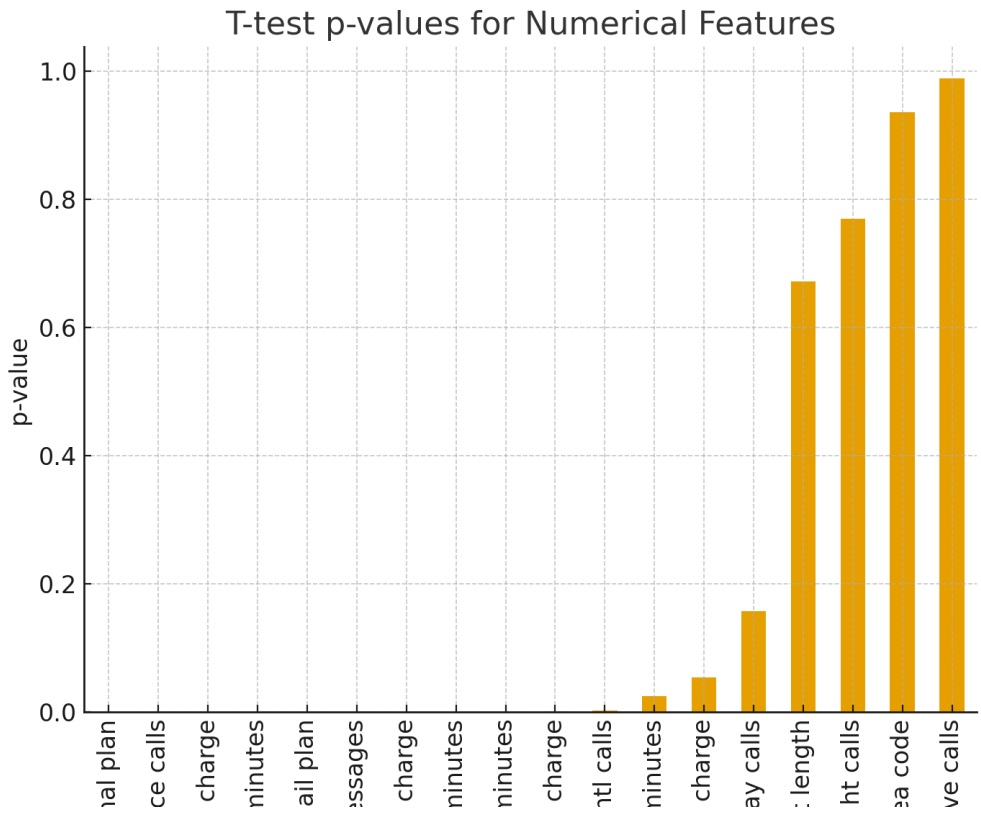

Figure 5: T-test p-values for numerical variables.

8. Feature Importance (Random Forest)

A Random Forest model was trained to evaluate the relative importance of each feature. The top 20 influential features are displayed below.

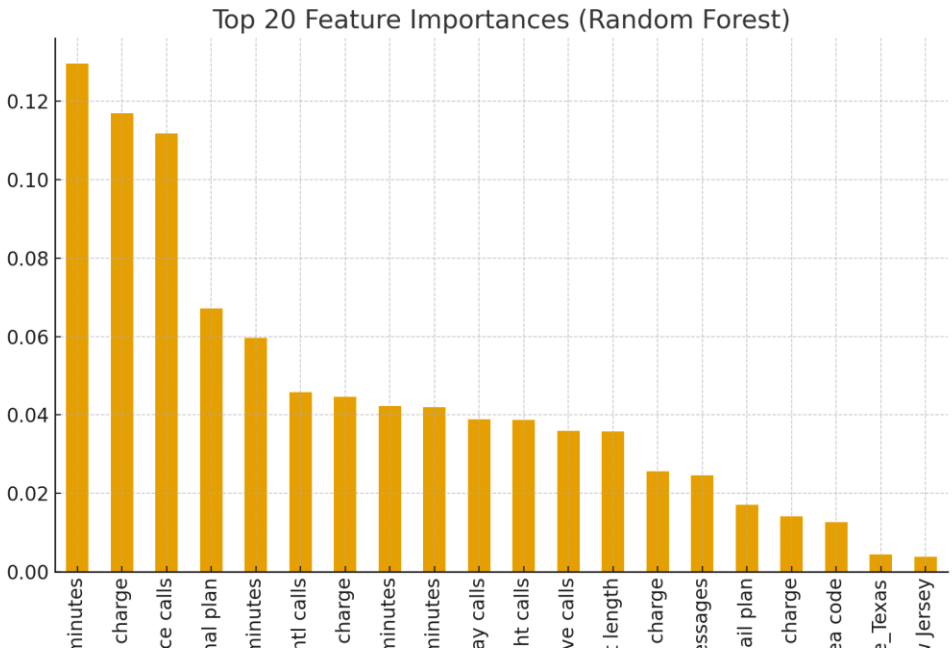

Figure 6: Top predictive features ranked by importance.

9. Conclusion

This analysis confirms that service usage metrics, customer support interactions, and service plans play major roles in churn behavior. Statistical and feature-level insights extracted here will guide the development of predictive machine learning models in the next milestone.

Milestone 2 Detailed Addendum

This addendum expands the analysis presented in the original report by incorporating a full statistical breakdown, complete feature engineering documentation, and a detailed interpretation of feature selection results (RFE).

1. Feature Engineering

- Total minutes: Sum of day, eve, night, and international minutes.
- Total calls: Sum of call counts across all periods.
- Avg minutes per call: Total minutes divided by total calls (with zero-division handled).
- High service calls: Indicator for users calling customer service more than 3 times.
- Tenure category: Derived categorical variable from Account Length.

2. Statistical Significance – Summary

Significant features ($p < 0.05$) include: State, International plan, Voice mail plan, Number vmail messages, Total day minutes/charge, Total eve minutes/charge, Total intl minutes/calls/charge, Customer service calls, Total minutes, Avg minutes per call, High service calls.

Non-significant features include: Account length, Area code, Total day calls, Total eve calls, Total night minutes, Total night calls, Total night charge, Total calls, Tenure category.

3. RFE Feature Selection

Important predictors selected via Logistic Regression RFE include: International plan, Voice mail plan, Number vmail messages, Total day minutes, Total day charge, Total eve charge, Total night minutes, Total night charge, Total intl metrics, Customer service calls, Total national metrics, Avg minutes per call, Cost per minute, High service calls, plan-related indicators.

A convergence warning occurred during RFE, indicating that scaling will be required in the next milestone for improved model stability.

4. Visualization Summary

The Top 10 Feature Importances figure (RFE) highlights the strongest churn predictors.

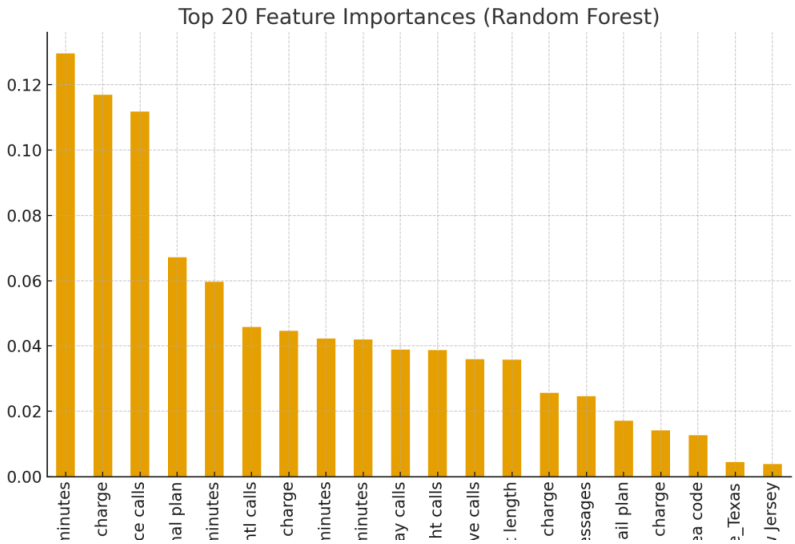

Figure: Top 20 Feature Importances (RFE Model)