Customer Churn at Syriatel

Phase 3 Data Science Project

Prepared by Paul Kamau

CONTENTS

BUSINESS UNDERSTANDING

DATA UNDERSTANDING

OVERVIEW/ INTRODUCTION

MODEL COMPARISON

CONTENTS

NEXT STEPS

RECOMMENDATIOS

EVALUATION

MODELLING

INTRODUCTION

- Syriatel is a mobile network provider in Syria that was founded in 2000.
- It is one of the two dominant providers in the country, along with MTN Syria.
- Syriatel offers LTE, 3G, and GSM services to its customers, under the brand name Super Surf.
- The project strategically harnesses data analytics and machine learning to augment customer experiences and reduce churn for Syriatel, a leading telecommunications provider in Syria.
- In an intensely competitive market, Syriatel's sustained growth hinges on customer retention and satisfaction.

Challenges

- One of the main challenges we face is the inability to spot the warning signs that a customer is likely to churn.
- This challenge is coupled with the need to understand the complex factors that contribute to a customer's decision to leave.

Proposed Solution

- To tackle these issues, we have initiated a data-driven project that implements advanced techniques to predict and prevent customer churn.

Objectives

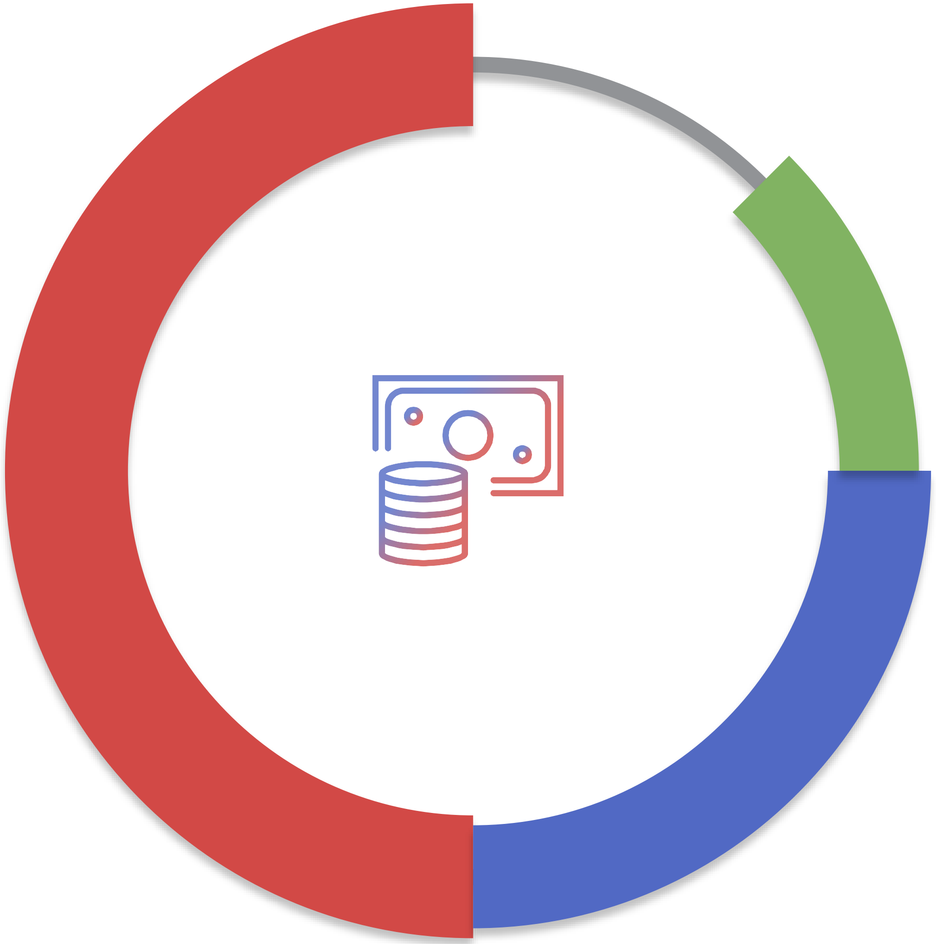
1. Churn Prediction and Mitigation:
By predicting churn, we can take timely action to retain customers.

2. Customer Experience Enhancement:
Understanding our customers allows us to tailor our services to their needs, improving their overall experience with SyriaTel.

3. Strategic Business Decision Making:
Informed by our data analysis, we can make strategic decisions to attract and retain customers, keeping SyriaTel competitive.

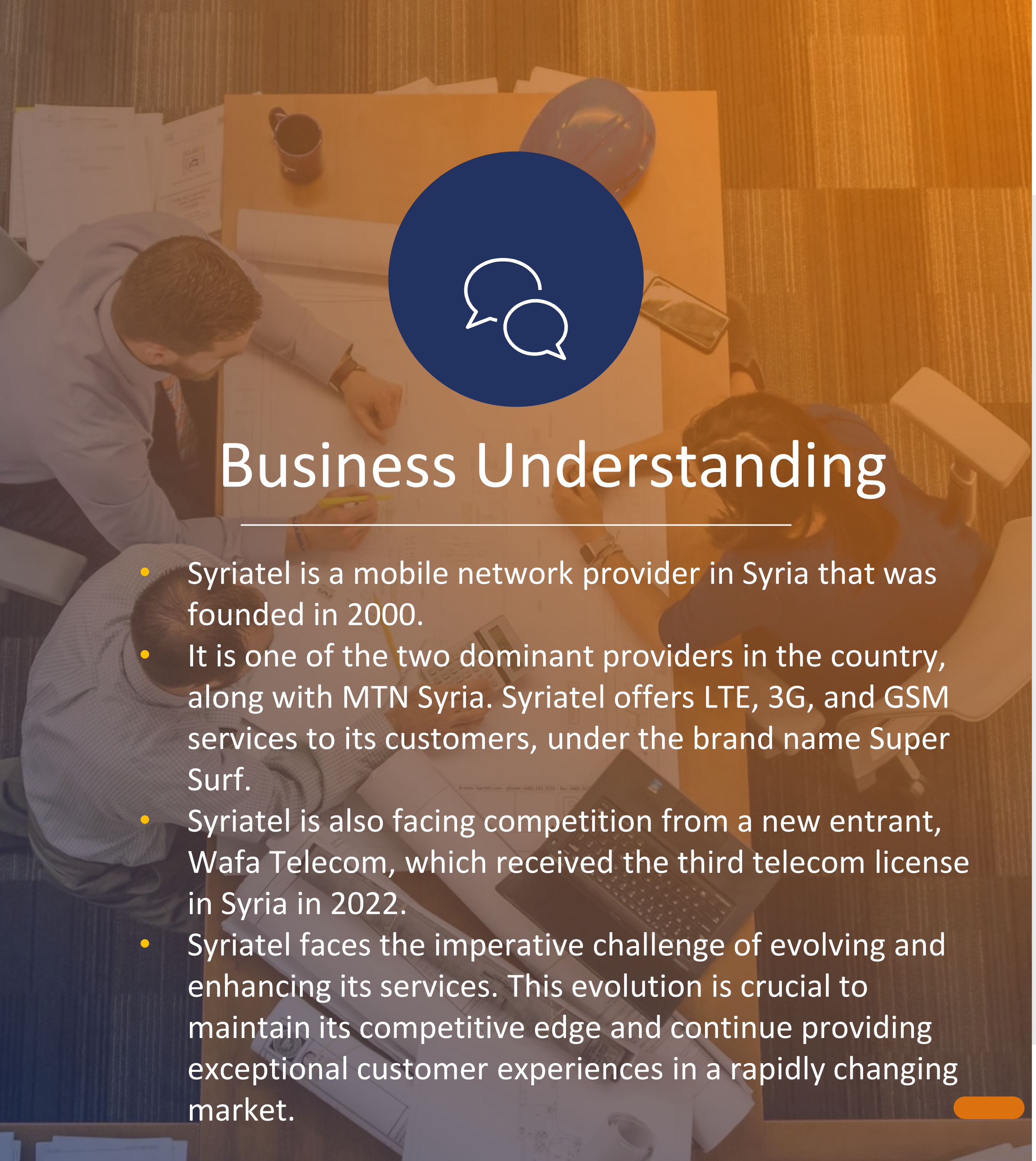
Problem Statement

We lack the ability to identify customers when they're on the cusp of churning (**Churn** is the loss of customers to competition).



I seek to predict likely churners. Our dataset includes 20 variables describing over 3,000 current and churned customers.

Achieving this predictive ability will allow us to examine the data on a rolling basis and quickly implement targeted incentivization.



Business Understanding

- Syriatel is a mobile network provider in Syria that was founded in 2000.
- It is one of the two dominant providers in the country, along with MTN Syria. Syriatel offers LTE, 3G, and GSM services to its customers, under the brand name Super Surf.
- Syriatel is also facing competition from a new entrant, Wafa Telecom, which received the third telecom license in Syria in 2022.
- Syriatel faces the imperative challenge of evolving and enhancing its services. This evolution is crucial to maintain its competitive edge and continue providing exceptional customer experiences in a rapidly changing market.

Data Source

Kaggle: Churn in Telecom's dataset

Link

(https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset/)

Data Description

Variable Descriptions

- **Churn**: Indicates if the customer has stopped doing business with SyriaTel. (False = No churn, True = Churned)
- **State**: The U.S. State of the customer. (Requires one-hot encoding; not ordinal)
- **Account Length**: A smaller number signifies an older account. (Indicative of Customer Lifetime Value)
- **Area Code**: Area code of the customer's phone number.
- **Phone Number**: The customer's phone number.
- **International Plan**: Whether the customer has an international plan. ('yes' or 'no'; binary and thus effectively one-hot encoded)
- **Voice Mail Plan**: Whether the customer subscribes to a voice mail plan. ('yes' or 'no'; as above)
- **Number of Voice Mail Messages**: Total number of voice mail messages left by the customer.
- **Total Day Minutes**: Aggregate of daytime minutes used.
- **Total Day Calls**: Total number of calls made during the day.
- **Total Day Charge**: Total charges incurred for daytime calls.
- **Total Eve Minutes**: Total minutes spent on calls in the evening.
- **Total Eve Calls**: Number of calls made during the evening.
- **Total Eve Charge**: Charges for evening calls.
- **Total Night Minutes**: Total minutes for nighttime calls.
- **Total Night Calls**: Number of calls made at night.
- **Total Night Charge**: Nighttime call charges.
- **Total Intl Minutes**: Cumulative international minutes (covering day, evening, and night).
- **Total Intl Calls**: Total number of international calls (across all time periods).
- **Total Intl Charge**: Total charges for international calls.
- **Customer Service Calls**: Number of calls made to customer service by the customer.

Target Variable

- **Churn**: if the customer has churned (true or false)

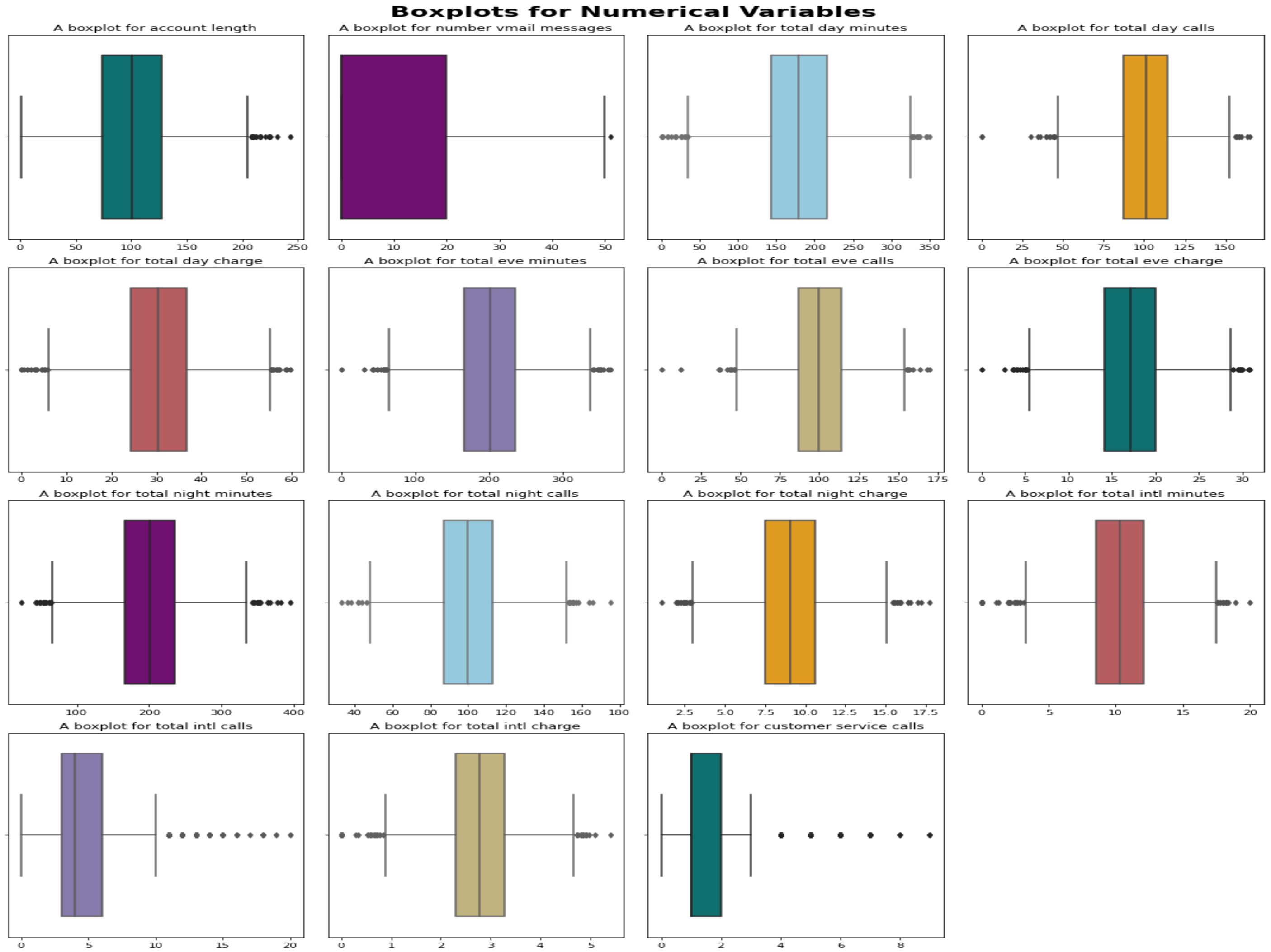
Data Analysis

- My exploration commences with a detailed univariate analysis, scrutinizing each variable in isolation to gauge its individual characteristics and distribution. This foundational step is critical for establishing a baseline understanding of the dataset's intrinsic properties, essential for informed hypothesis formulation and subsequent multivariate analyses.

Observations

- The analysis of the numerical variables provides insightful observations:
- **Account Length**: A notable presence of outliers indicates that a few accounts have unusually lon durations.
- **Number Vmail Messages**: While the majority of customers tend to have a minimal number of voicemail messages, a small group stands out with a considerably higher frequency of messages
- **Total Day Calls**: The data shows a few exceptional cases with extremely low or high numbers of daily calls, deviating from the common trend.
- **Total Intl Minutes**: While shorter international calls are common among most customers, there exists a distinct segment making significantly longer calls.
- **Total Intl Calls**: Generally, customers make a limited number of international calls, but there are exceptions where this number is notably higher.
- **Customer Service Calls**: While the trend leans towards one or two service calls per customer, there are exceptions with an unusually high number of service calls.
- **Total Day Charge**: The distribution appears tightly grouped, suggesting that day charges are fairl consistent among customers, with a few exceptions on the higher end.
- **Total Eve Minutes and Charges**: There's a moderate spread in the evening minutes and charge with a small number of outliers, indicating that while most customers have similar usage patterns i the evening, a few have significantly higher usage.
- **Total Night Minutes and Calls**: The spread is quite similar to that of evening usage, but with few outliers, which could suggest more uniform behavior in nighttime usage among customers.
- **Total Intl Calls and Charges**: The international calls and charges show a relatively tight distribution, yet there are outliers on both the lower and upper ends, highlighting that international communication is not uniform across customers.
- **Customer Service Calls**: This variable shows a right-skewed distribution, with most customers needing only a few service calls, but there's a noticeable amount of customers with many more calls than average, which could indicate issues or high engagement with customer service.
- **Total Night Charge**: Similar to the night minutes, the charges are also consistently distributed wit a few customers experiencing higher charges, which may correspond to the outliers observed in the night minutes.

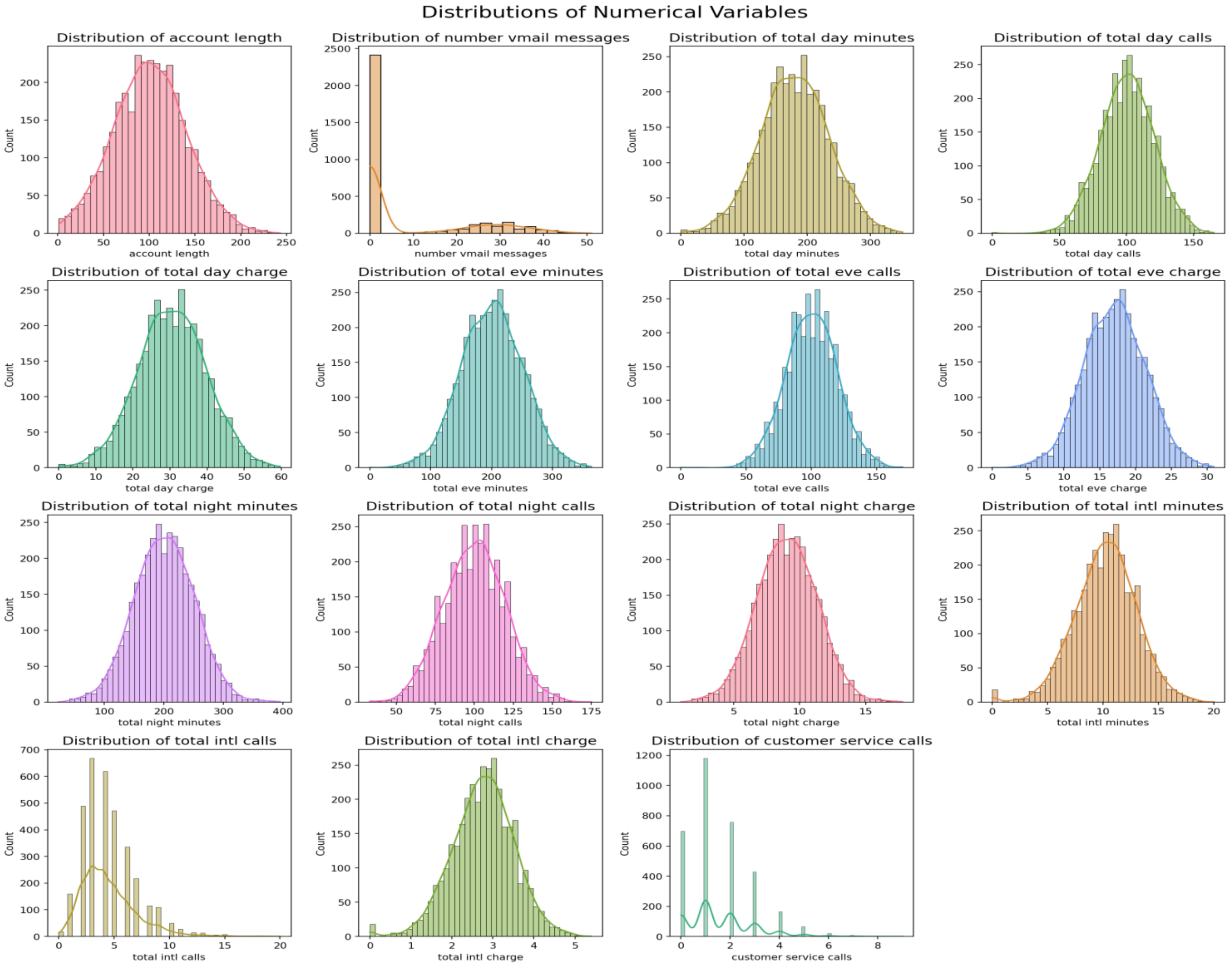
Numerical Variables



Data Spread Overview

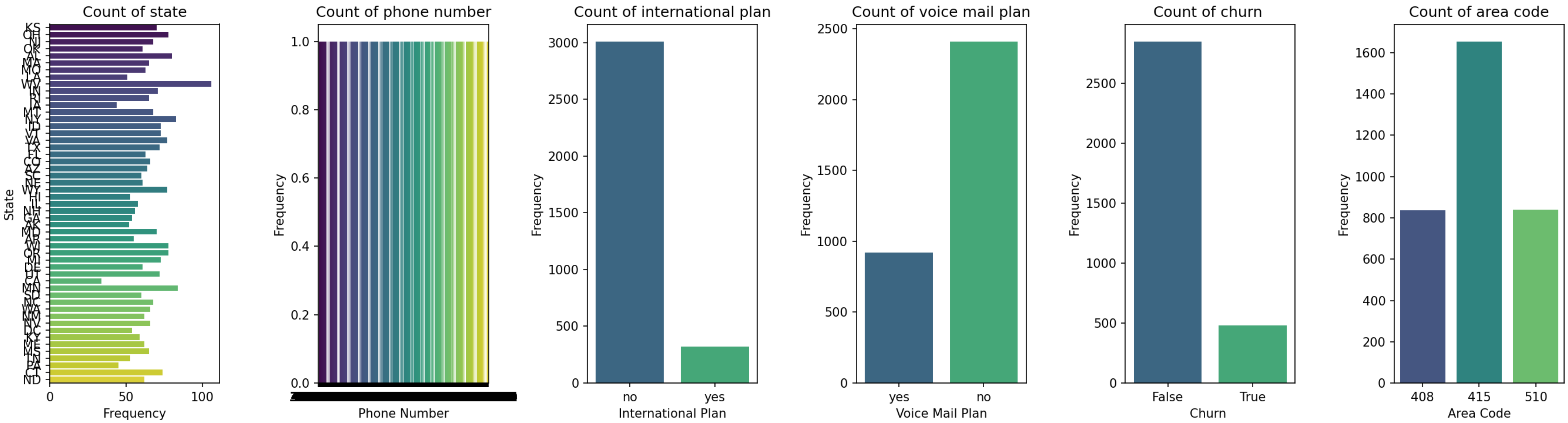
- To gain insight into our dataset, we'll examine the dispersion/distribution of our data points.
- Then we do Exploration of Categorical Variables:
- Then we conduct Bi-Variate Relationship Exploration to understand how they interact with one another.
- Then We do a Multi- collinearity Adjustment where only the most descriptive attributes are retained, ensuring the dataset remains lean and meaningful for predictive modeling.
- **NB: the visualizations are provided on the slides below.**

Distribution of the data
(numerical data)

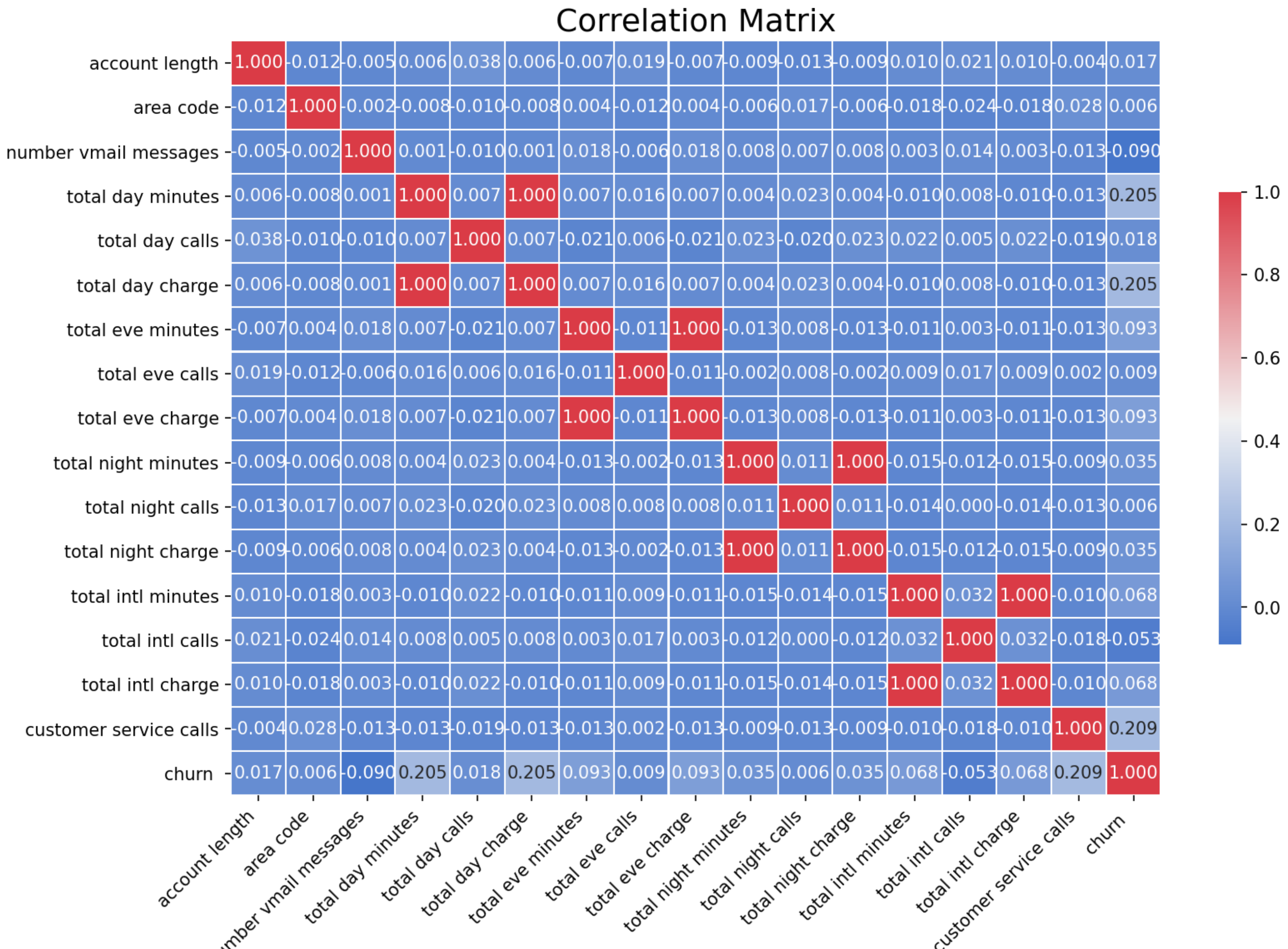


Distributions of Numerical Variables

Distribution of the data *(categorical data)*



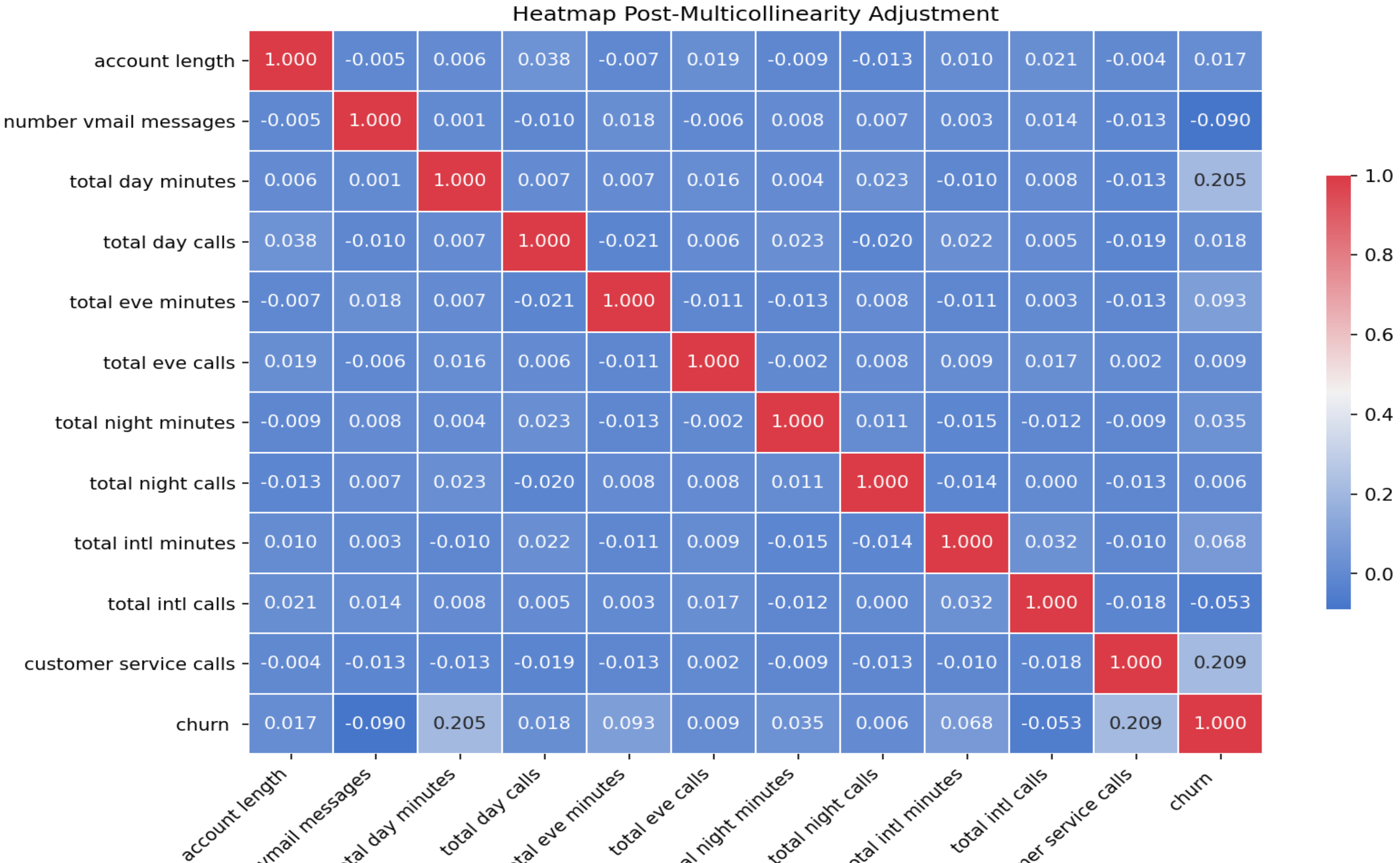
Bi-Variate Relationship Exploration



Correlation Matrix

	account length	area code	umber vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	ustomer service calls	churn
account length	1.000	-0.012	-0.005	0.006	0.038	0.006	-0.007	0.019	-0.007	-0.009	-0.013	-0.009	0.010	0.021	0.010	-0.004	0.017
area code	-0.012	1.000	-0.002	-0.008	-0.010	-0.008	0.004	-0.012	0.004	-0.006	0.017	-0.006	-0.018	-0.024	-0.018	0.028	0.006
number vmail messages	-0.005	-0.002	1.000	0.001	-0.010	0.001	0.018	-0.006	0.018	0.008	0.007	0.008	0.003	0.014	0.003	-0.013	-0.090
total day minutes	0.006	-0.008	0.001	1.000	0.007	1.000	0.007	0.016	0.007	0.004	0.023	0.004	-0.010	0.008	-0.010	-0.013	0.205
total day calls	0.038	-0.010	-0.010	0.007	1.000	0.007	-0.021	0.006	-0.021	0.023	-0.020	0.023	0.022	0.005	0.022	-0.019	0.018
total day charge	0.006	-0.008	0.001	1.000	0.007	1.000	0.007	0.016	0.007	0.004	0.023	0.004	-0.010	0.008	-0.010	-0.013	0.205
total eve minutes	-0.007	0.004	0.018	0.007	-0.021	0.007	1.000	-0.011	1.000	-0.013	0.008	-0.013	-0.011	0.003	-0.011	-0.013	0.093
total eve calls	0.019	-0.012	-0.006	0.016	0.006	0.016	-0.011	1.000	-0.011	-0.002	0.008	-0.002	0.009	0.017	0.009	0.002	0.009
total eve charge	-0.007	0.004	0.018	0.007	-0.021	0.007	1.000	-0.011	1.000	-0.013	0.008	-0.013	-0.011	0.003	-0.011	-0.013	0.093
total night minutes	-0.009	-0.006	0.008	0.004	0.023	0.004	-0.013	-0.002	-0.013	1.000	0.011	1.000	-0.015	-0.012	-0.015	-0.009	0.035
total night calls	-0.013	0.017	0.007	0.023	-0.020	0.023	0.008	0.008	0.008	0.011	1.000	0.011	-0.014	0.000	-0.014	-0.013	0.006
total night charge	-0.009	-0.006	0.008	0.004	0.023	0.004	-0.013	-0.002	-0.013	1.000	0.011	1.000	-0.015	-0.012	-0.015	-0.009	0.035
total intl minutes	0.010	-0.018	0.003	-0.010	0.022	-0.010	-0.011	0.009	-0.011	-0.015	-0.014	-0.015	1.000	0.032	1.000	-0.010	0.068
total intl calls	0.021	-0.024	0.014	0.008	0.005	0.008	0.003	0.017	0.003	-0.012	0.000	-0.012	0.032	1.000	0.032	-0.018	-0.053
total intl charge	0.010	-0.018	0.003	-0.010	0.022	-0.010	-0.011	0.009	-0.011	-0.015	-0.014	-0.015	1.000	0.032	1.000	-0.010	0.068
customer service calls	-0.004	0.028	-0.013	-0.013	-0.019	-0.013	-0.013	0.002	-0.013	-0.009	-0.013	-0.009	-0.010	-0.018	-0.010	1.000	0.209
churn	0.017	0.006	-0.090	0.205	0.018	0.205	0.093	0.009	0.093	0.035	0.006	0.035	0.068	-0.053	0.068	0.209	1.000

Heatmap Post-Multicollinearity Adjustment



	account length	vmail messages	tal day minutes	total day calls	tal eve minutes	total eve calls	al night minutes	total night calls	otal intl minutes	total intl calls	er service calls	churn
account length	1.000	-0.005	0.006	0.038	-0.007	0.019	-0.009	-0.013	0.010	0.021	-0.004	0.017
number vmail messages	-0.005	1.000	0.001	-0.010	0.018	-0.006	0.008	0.007	0.003	0.014	-0.013	-0.090
total day minutes	0.006	0.001	1.000	0.007	0.007	0.016	0.004	0.023	-0.010	0.008	-0.013	0.205
total day calls	0.038	-0.010	0.007	1.000	-0.021	0.006	0.023	-0.020	0.022	0.005	-0.019	0.018
total eve minutes	-0.007	0.018	0.007	-0.021	1.000	-0.011	-0.013	0.008	-0.011	0.003	-0.013	0.093
total eve calls	0.019	-0.006	0.016	0.006	-0.011	1.000	-0.002	0.008	0.009	0.017	0.002	0.009
total night minutes	-0.009	0.008	0.004	0.023	-0.013	-0.002	1.000	0.011	-0.015	-0.012	-0.009	0.035
total night calls	-0.013	0.007	0.023	-0.020	0.008	0.008	0.011	1.000	-0.014	0.000	-0.013	0.006
total intl minutes	0.010	0.003	-0.010	0.022	-0.011	0.009	-0.015	-0.014	1.000	0.032	-0.010	0.068
total intl calls	0.021	0.014	0.008	0.005	0.003	0.017	-0.012	0.000	0.032	1.000	-0.018	-0.053
customer service calls	-0.004	-0.013	-0.013	-0.019	-0.013	0.002	-0.009	-0.013	-0.010	-0.018	1.000	0.209
churn	0.017	-0.090	0.205	0.018	0.093	0.009	0.035	0.006	0.068	-0.053	0.209	1.000

Modeling

- I have developed classification models that can predict customer churn with a high degree of accuracy.
- These models help us identify at-risk customers so we can act before we lose them.
- The models are:
  A. **Baseline model-(Dummy)**
  B. **Logistic Regression Model**
  C. **Decision Tree classifier**
  D. **XGBClassifier Model**

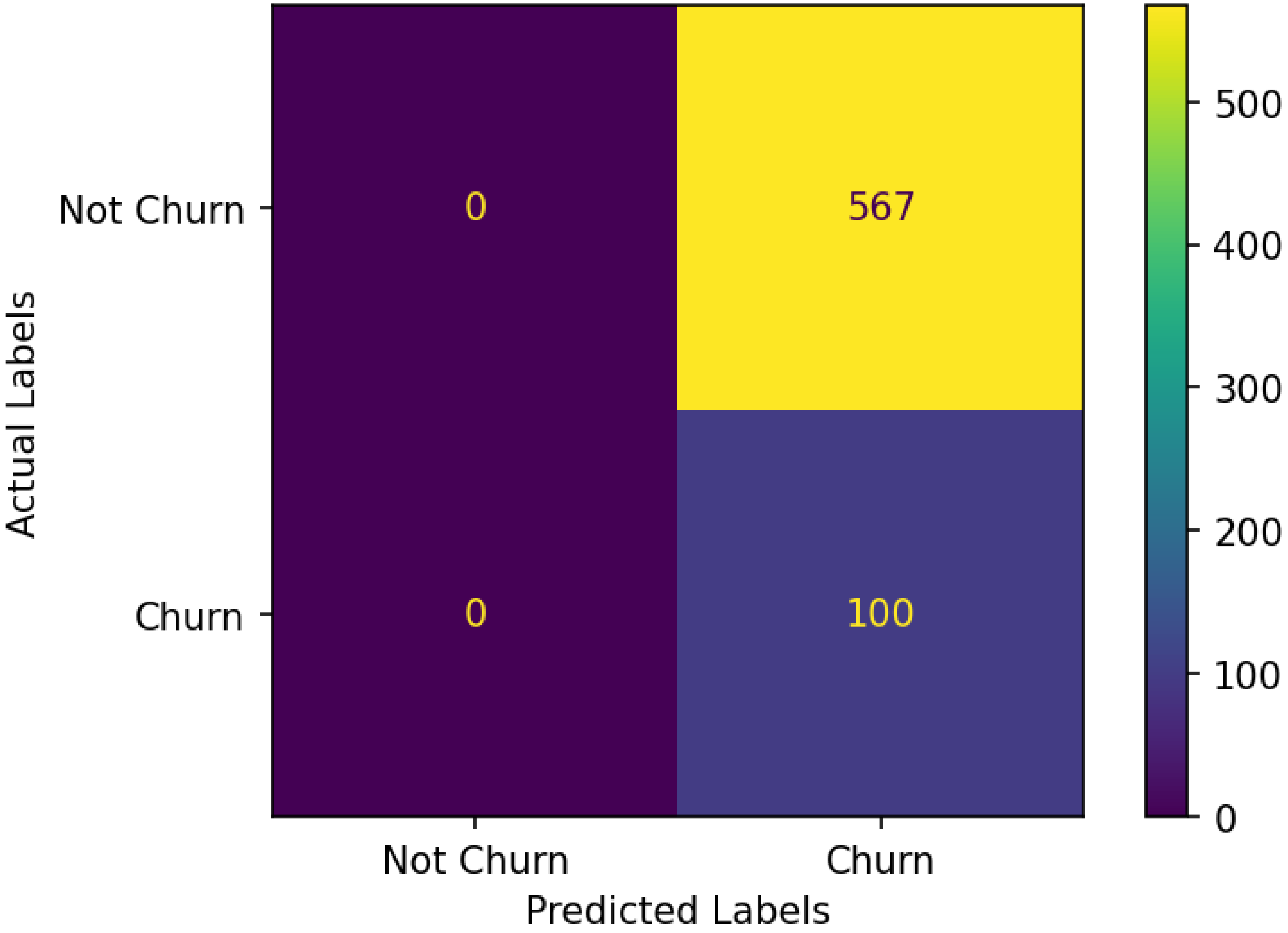
Modeling Cont'd

- Performance
  - Tested 3 main predictive models:
  - Logistic Regression: A basic statistical model
  - Decision Tree: A flowchart-like model to classify churn
  - XGBoost: An advanced ensemble model combining multiple decision trees.
- The XGBoost model emerged as the top performer based on accuracy, ability to detect churn, and consistency across training and validation data.

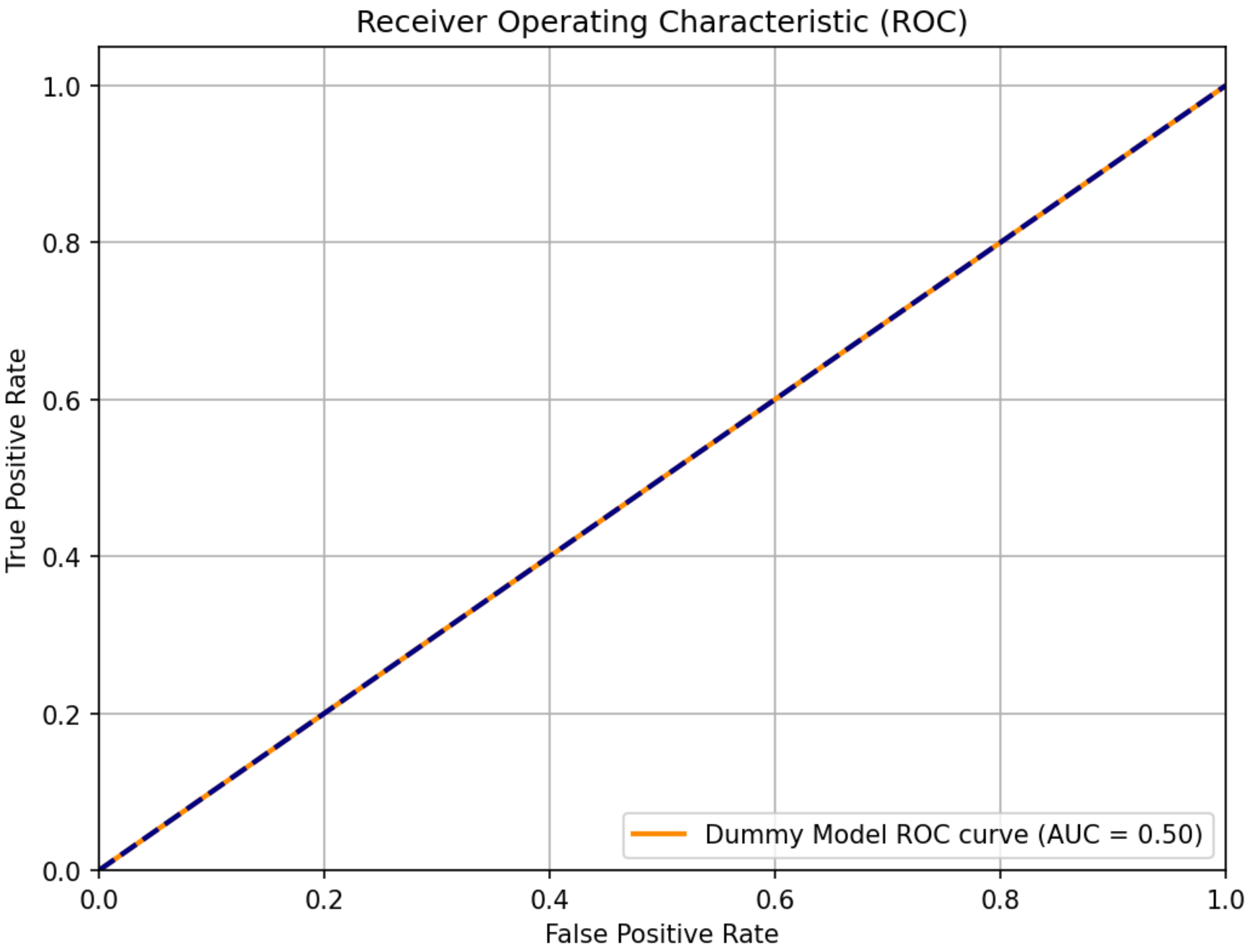
Modeling Cont’d

- Some key metrics:
  - Accuracy: Percentage correct predictions overall- Recall: Percentage of actual churns it identified
  - Low error rates: Very few customers incorrectly predicted to churn
- We specifically focused on maximizing recall, meaning correctly flagging customers who will churn. This enables the retention team to target preventive campaigns.
- Even if we incorrectly predict some customers will churn, retention incentives still provide business value.
- With an accuracy rate over 95% and ability to identify 76% of customers who actually churned, the XGBoost model balances predictive performance with alignment to core business goals of reducing customer

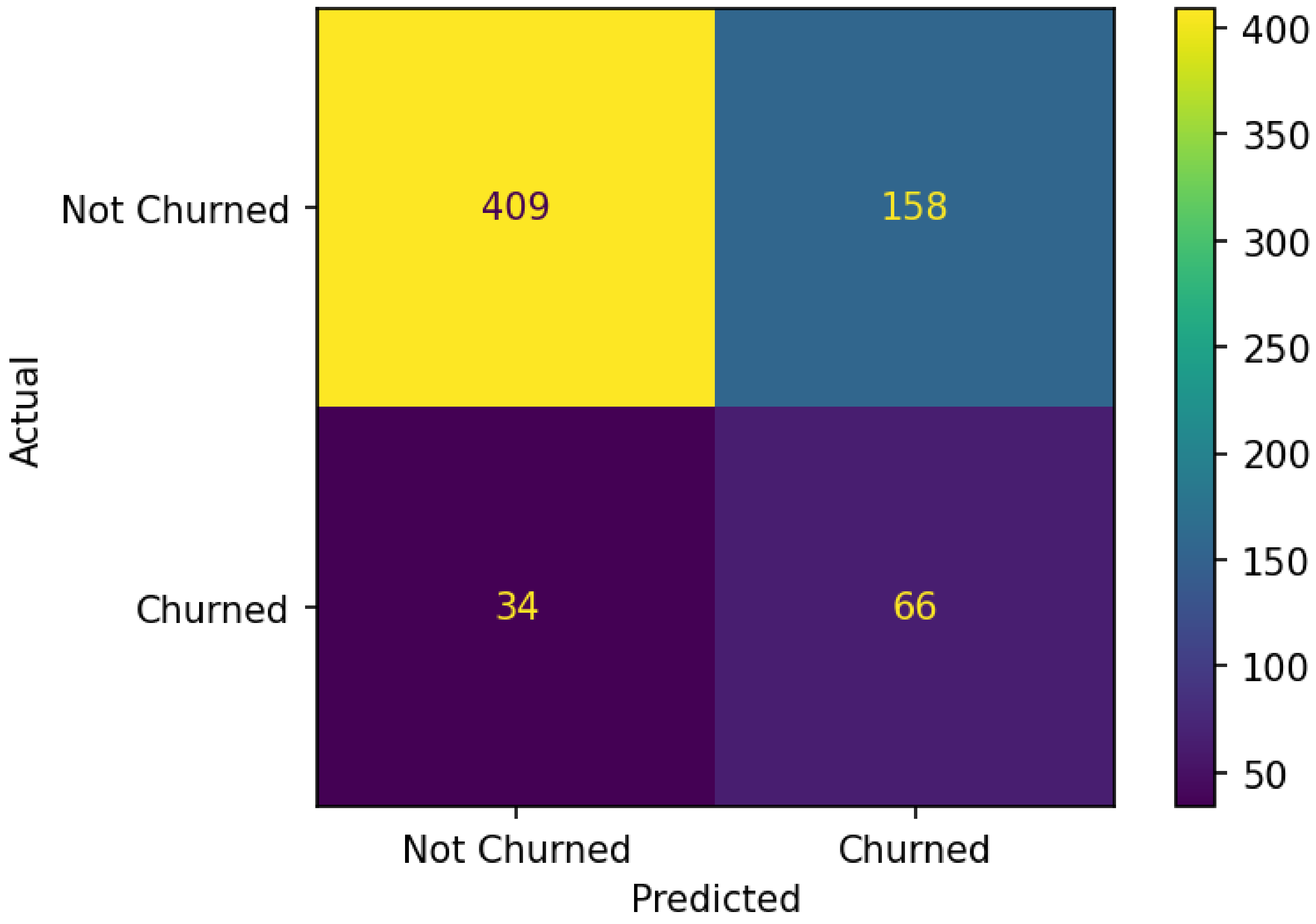
Confusion Matrix – Baseline Model



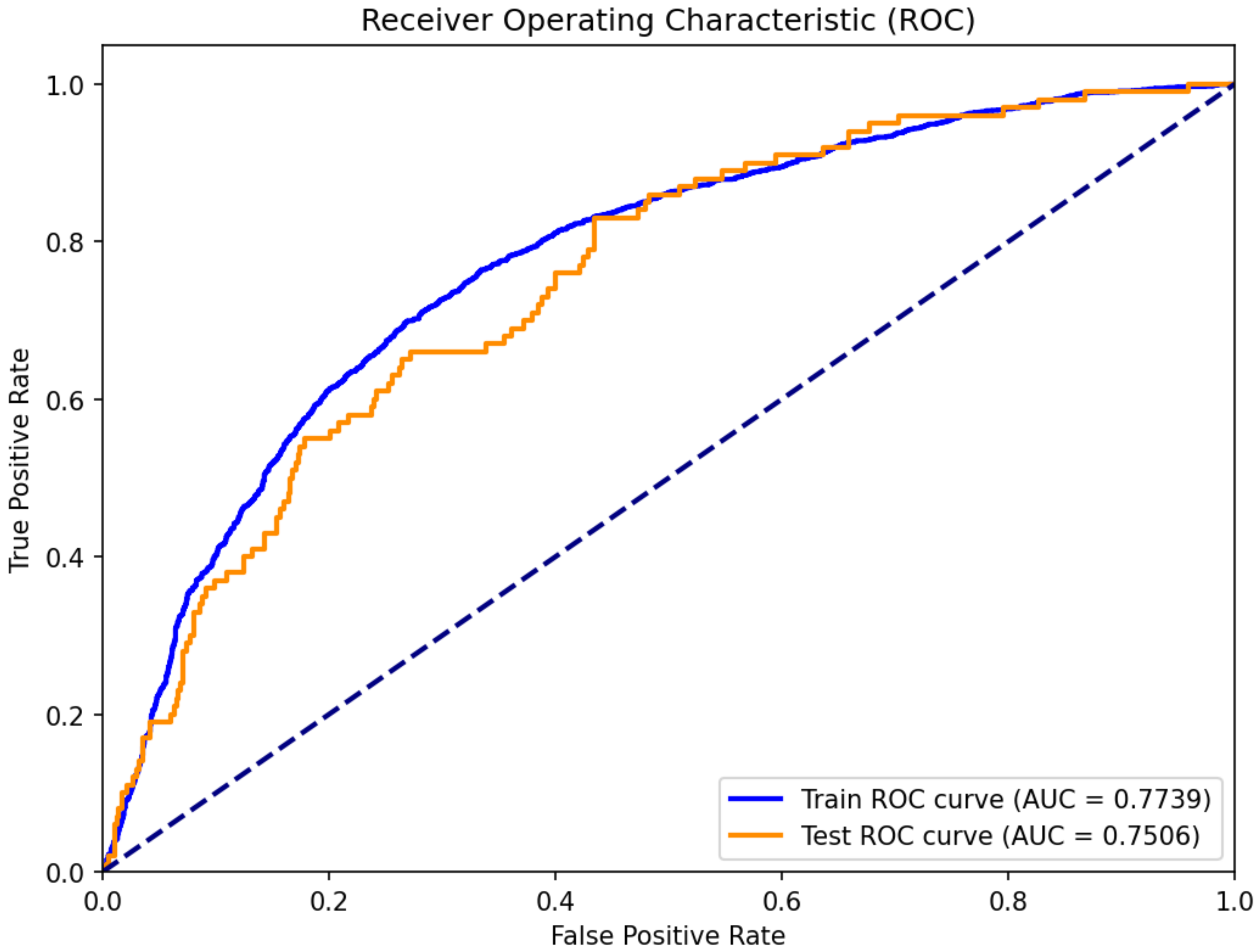
ROC– Baseline Model



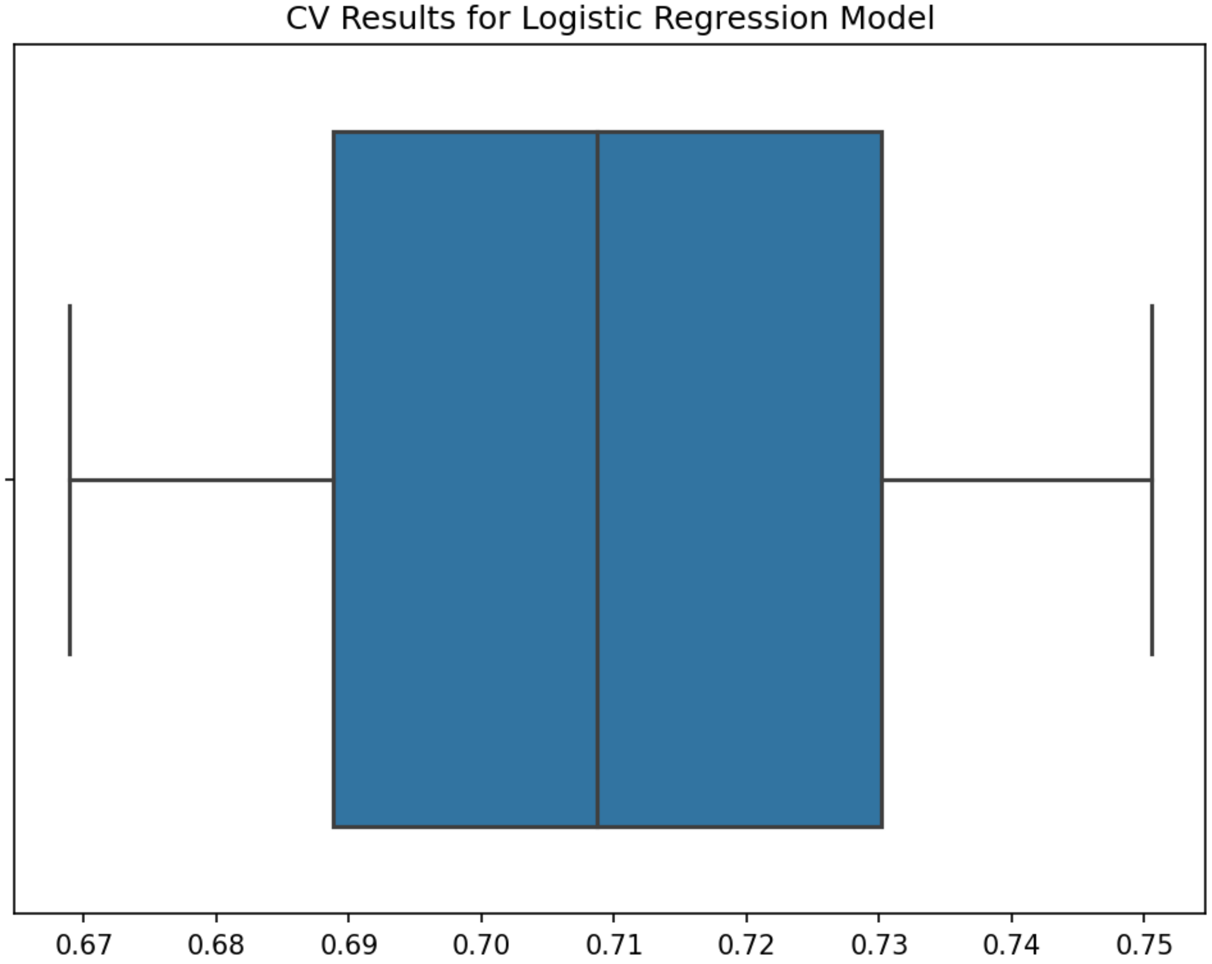
Confusion Matrix – Logistic Model



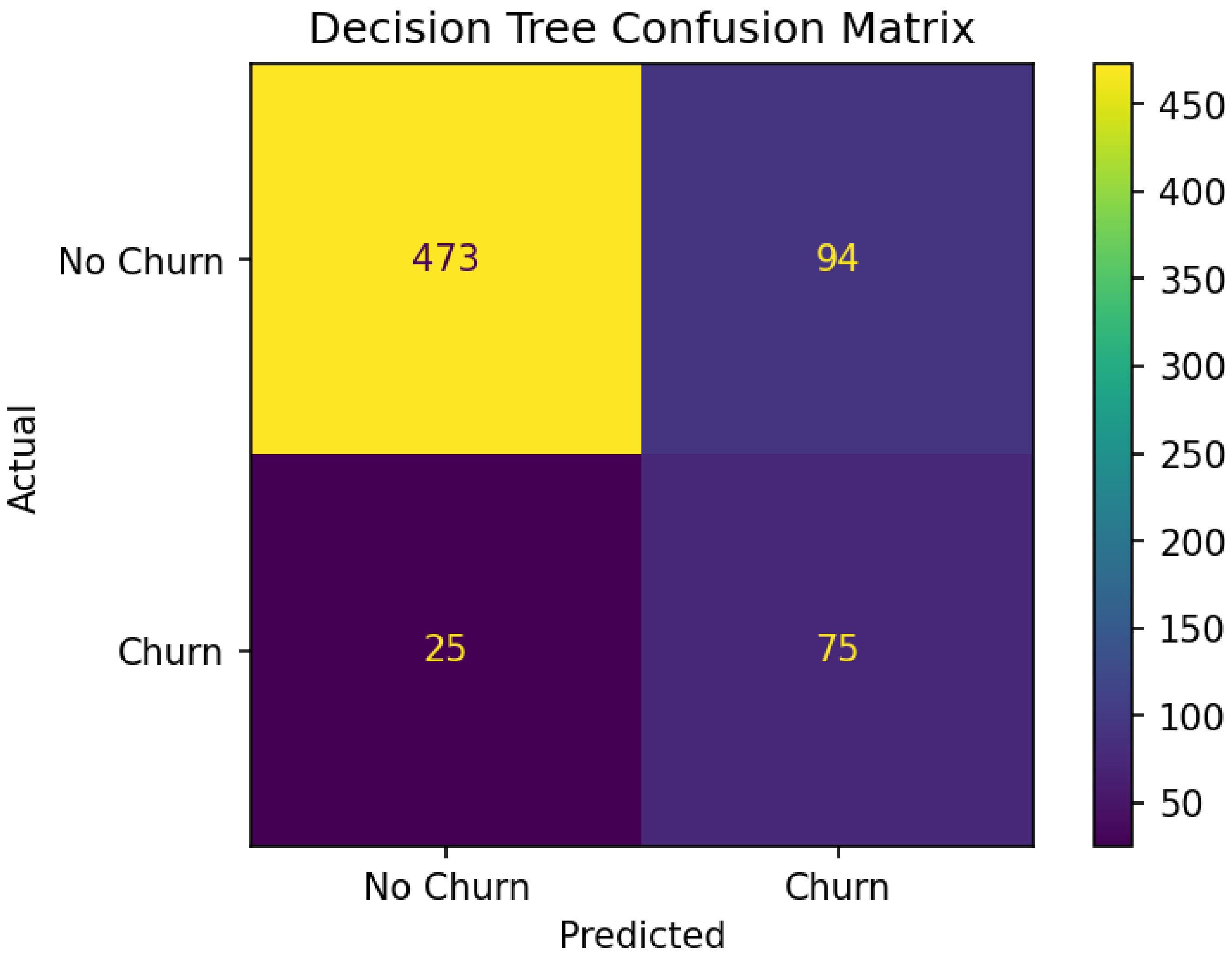
ROC– Logistic Model



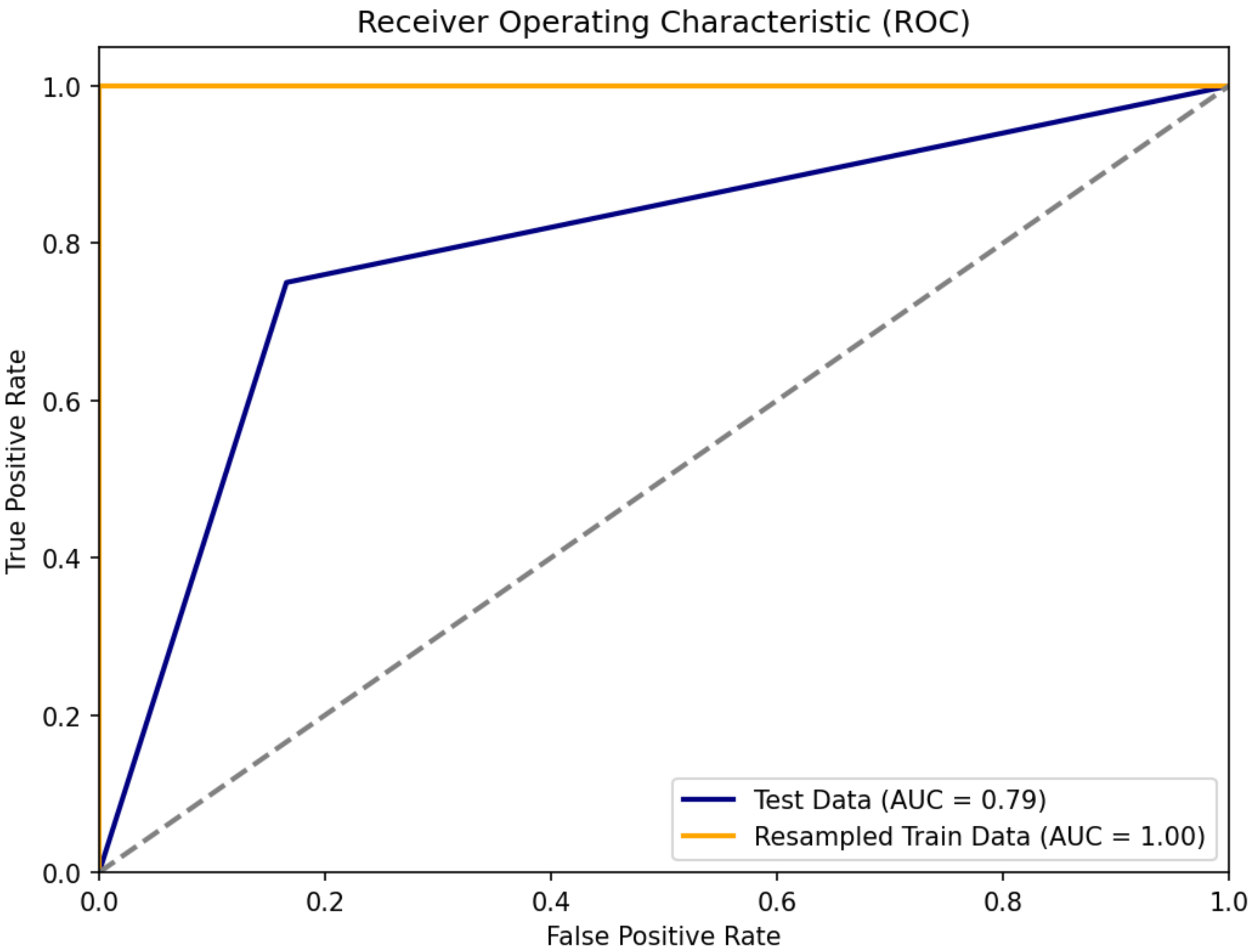
CV Results– Logistic Model



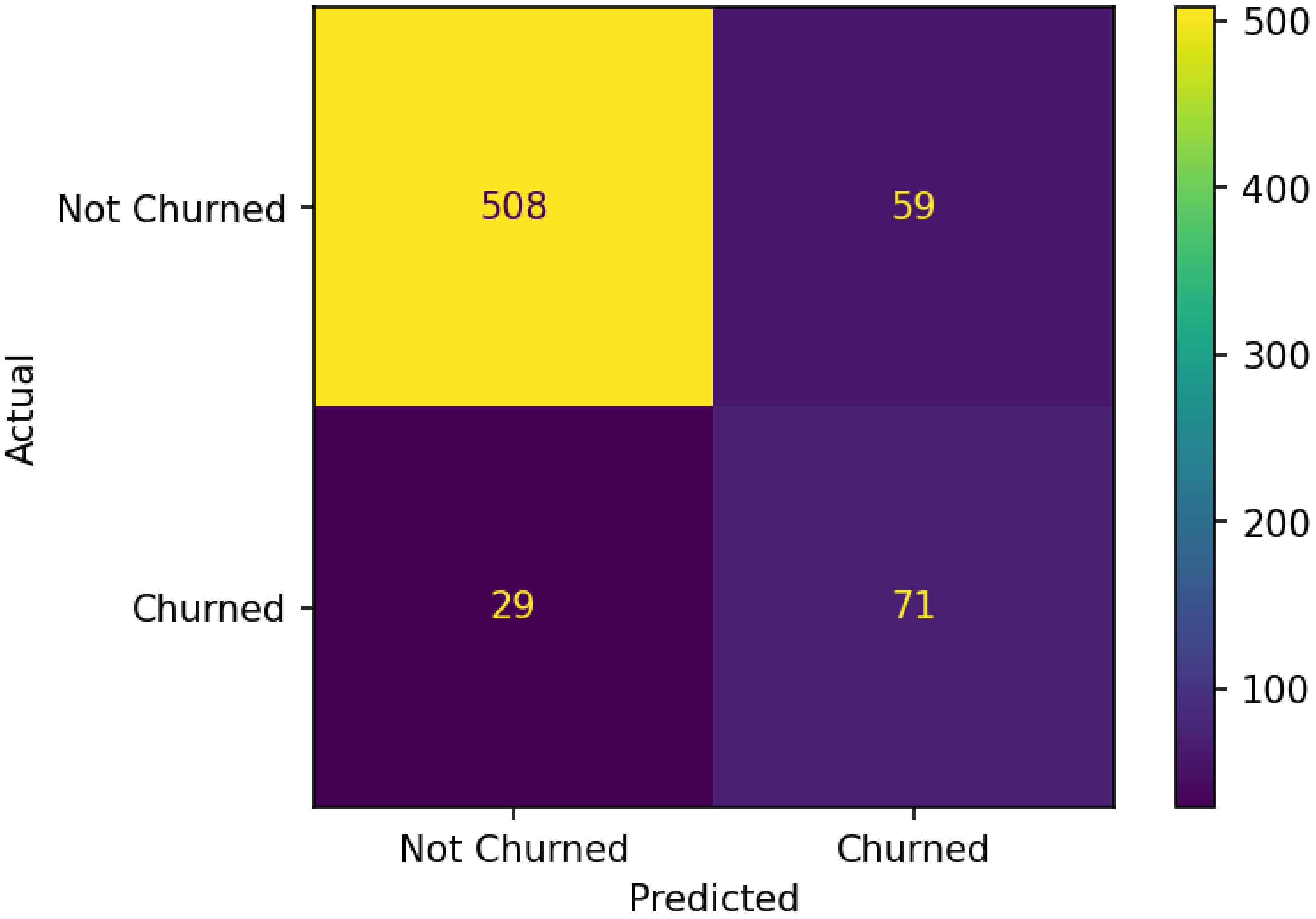
Confusion Matrix – Decision Tree Classifier



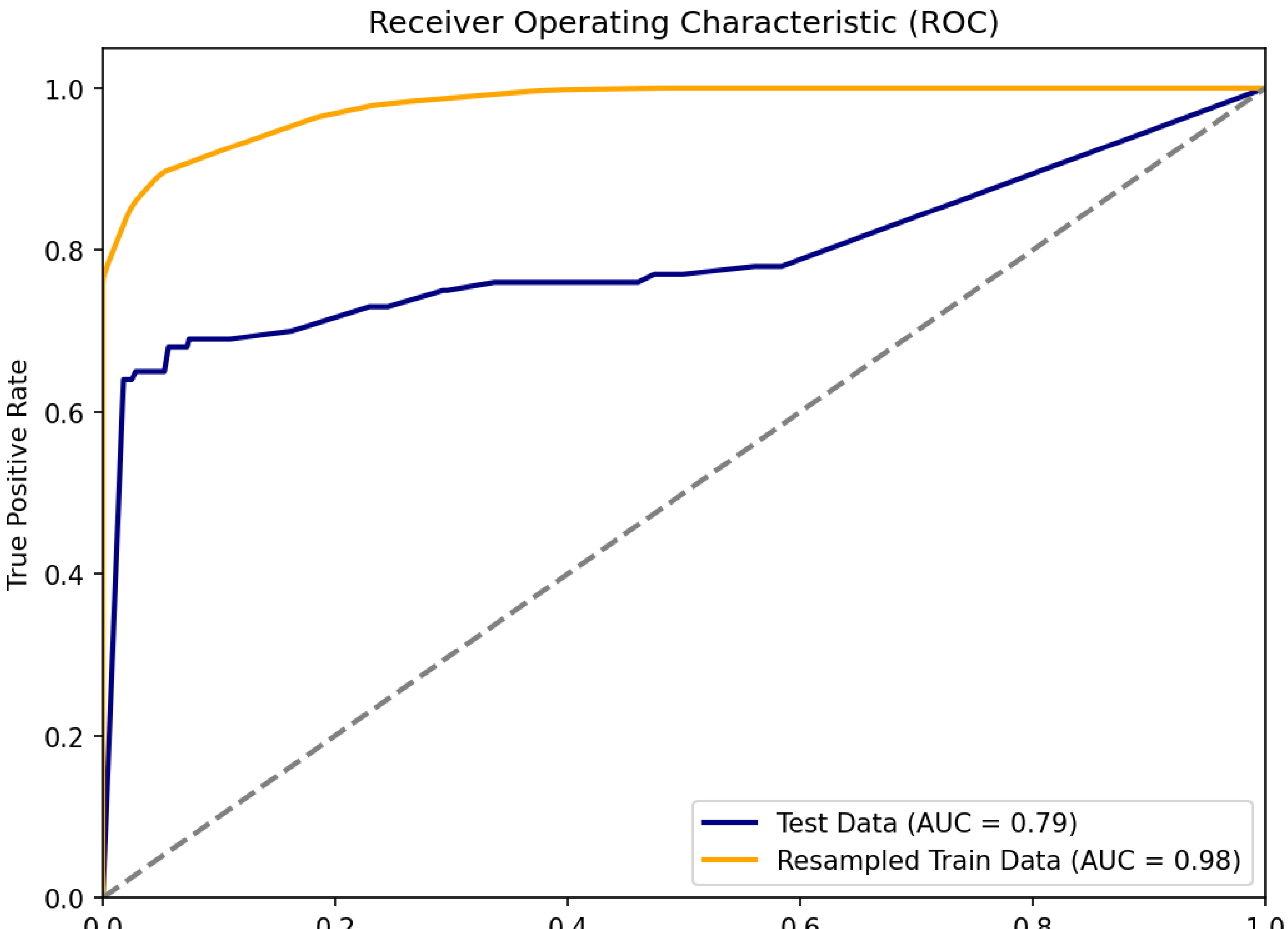
ROC– Decision Tree Classifier



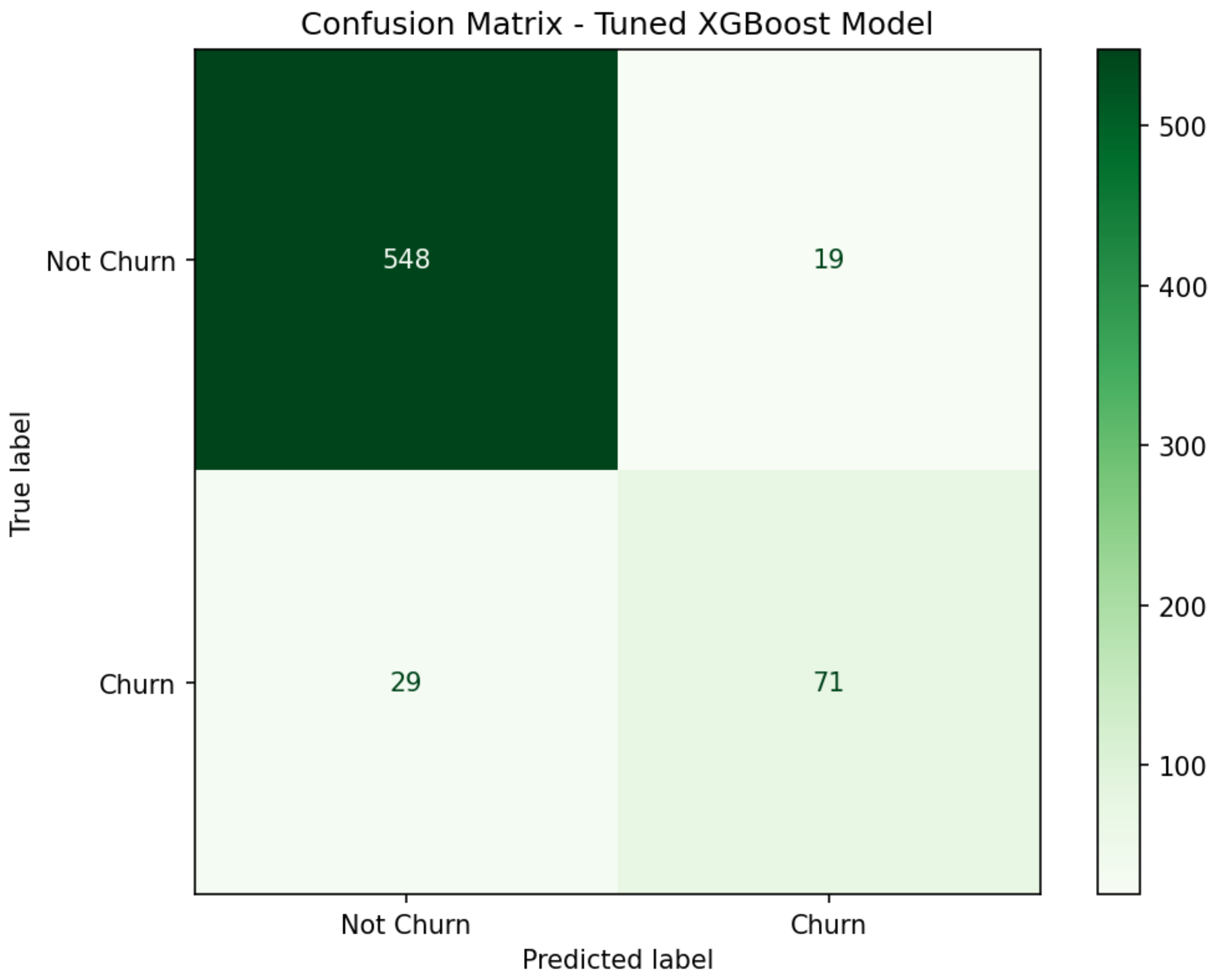
Confusion Matrix – Decision Tree Classifier (Tuned)



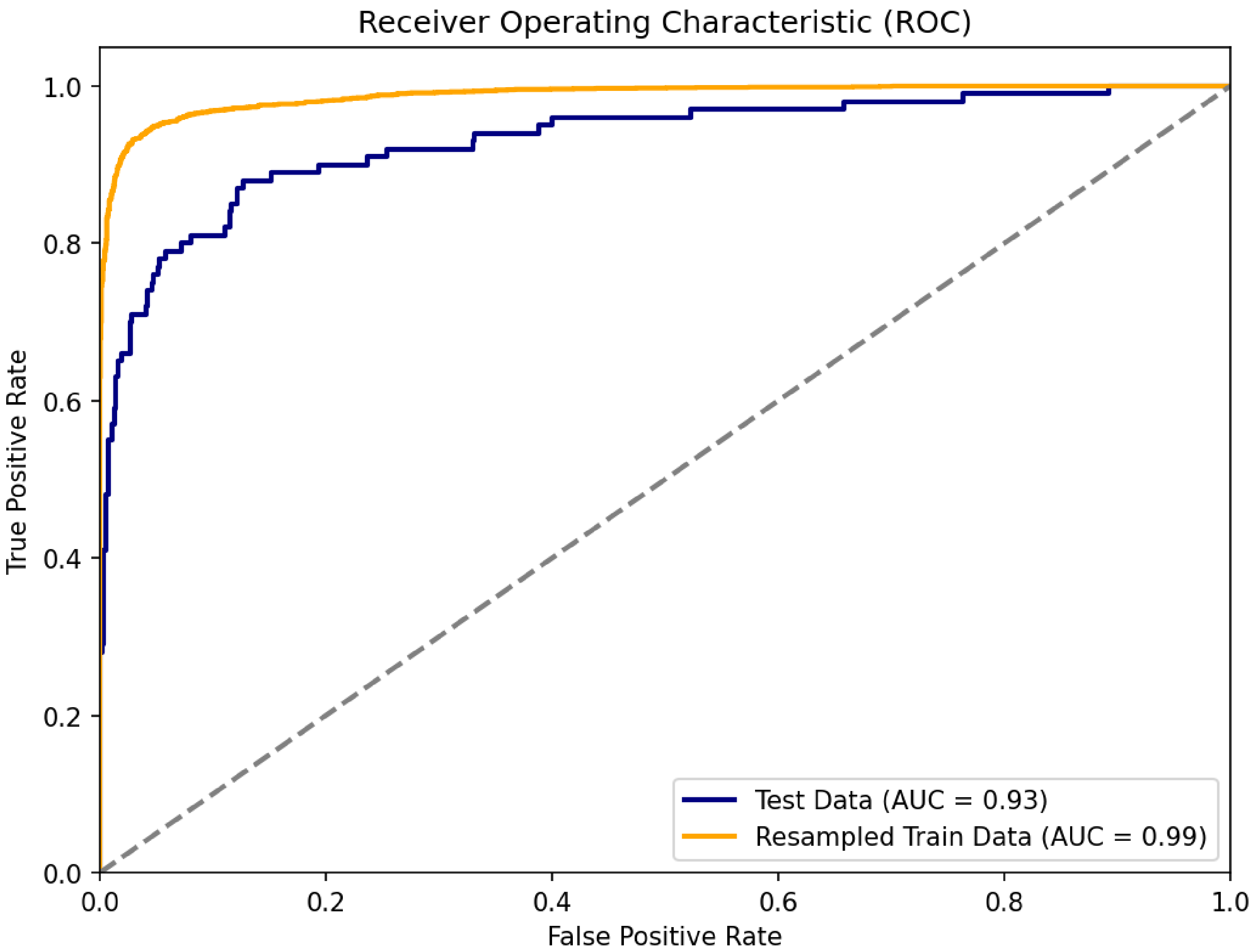
ROC – Decision Tree Classifier (Tuned)



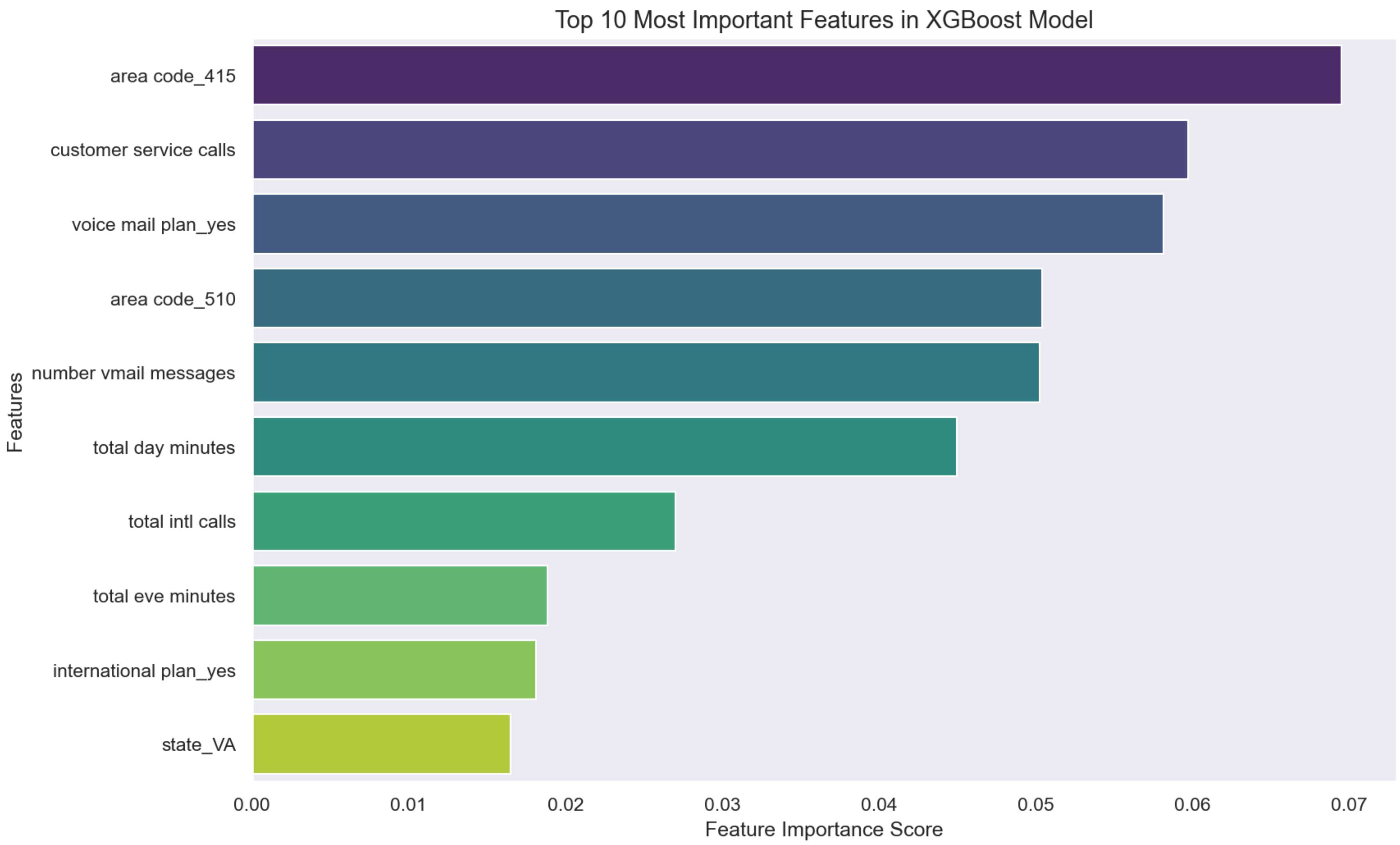
Confusion Matrix – XGBoost Classifier (Tuned)



ROC – XGBoost Classifier



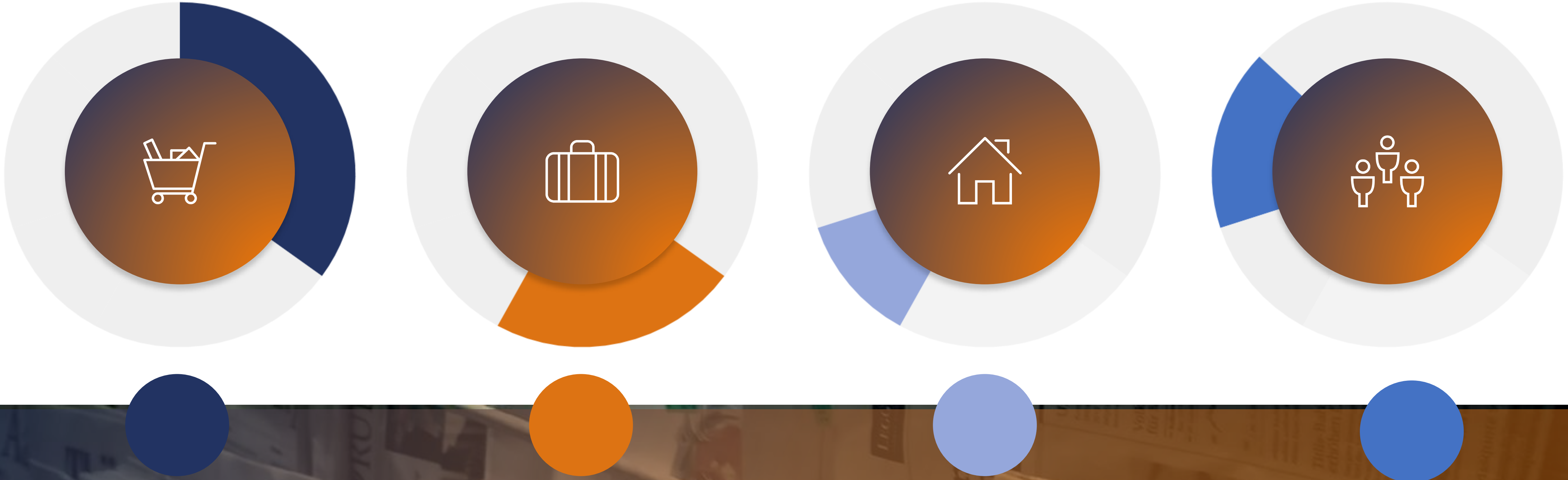
Feature Importance



MODEL PERFORMANCE COMPARISON

Model	Accuracy	ROC AUC	Precision (True)	Recall (True)	F1-Score (True)
Logistic Regression	0.706	0.683	0.287611	0.65	0.398773
Decision Tree (Tuned)	0.818	0.778	0.436364	0.72	0.543396
XGBoost (Tuned)	0.935	0.843	0.835294	0.71	0.767568

RECOMMENDATIONS



Implement XGBoost Model

Integrate optimized XGBoost model into CRM and billing systems
Build real-time dashboard to monitor customer churn risk scores
Create workflow to automatically flag high-risk customers

Improve Customer Service

Expand customer service team and slash wait times
Implement customer surveys and incentivize participation
Identify pain points through analytics of calls, survey verbatims

Offer Retention Incentives

Proactively reach out to high-risk customers with personalized promotions
Tailor plans and discounts based on usage patterns
Focus on high-value customers initially to maximize lifetime value

Adjust Marketing Strategies

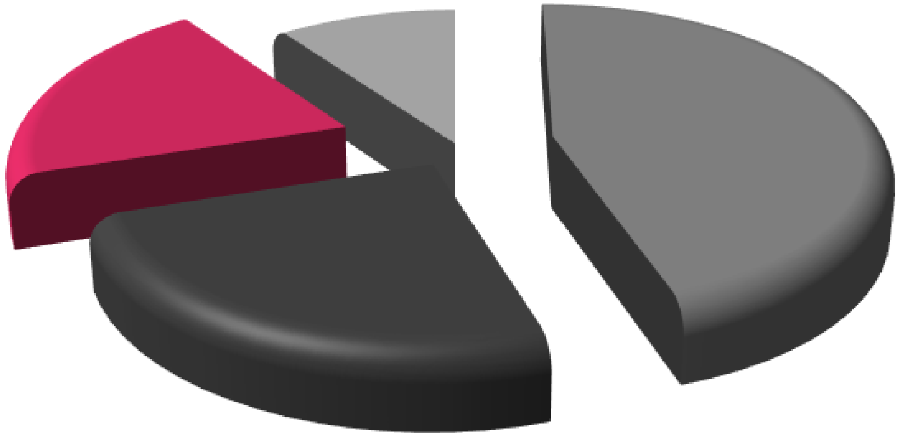
Shift campaigns from acquisition to retention
Target regions and demographics likely to churn
Revise voice/data plan bundles based on model feature importance

NEXT STEPS

Monitor Model Accuracy

- Compute performance metrics on rolling basis
- Establish accuracy thresholds and triggers for retraining

1



3

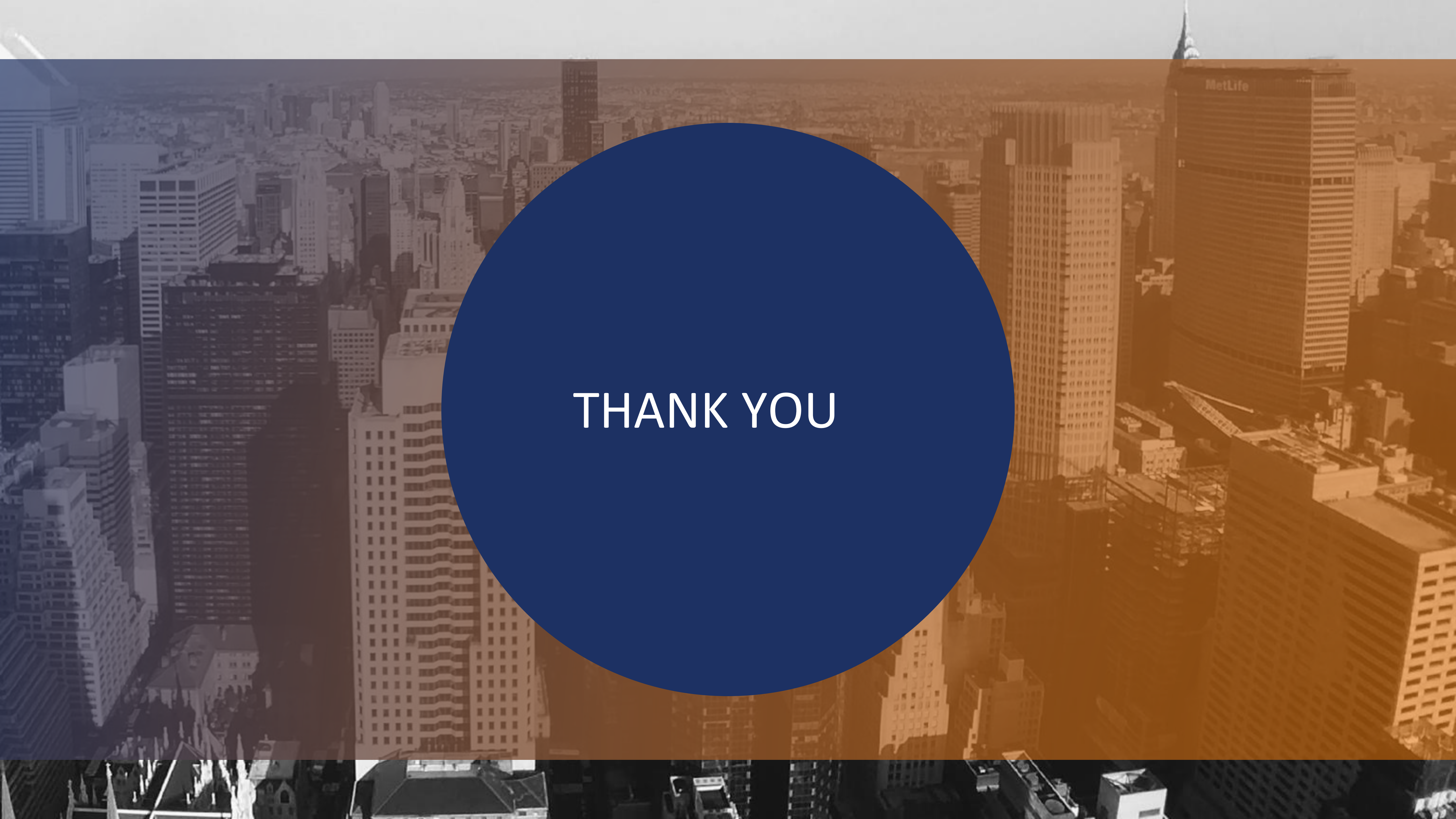
Iterative Enhancements

- Expand model features for greater precision
- Continually tune model hyper-parameters
- Feedback insights into future iterations

Launch Retention Team Pilot

- Start with small high-risk customer cohort
- Measure impact on churn rates and revenue
- Expand program over time based on insights

2



THANK YOU