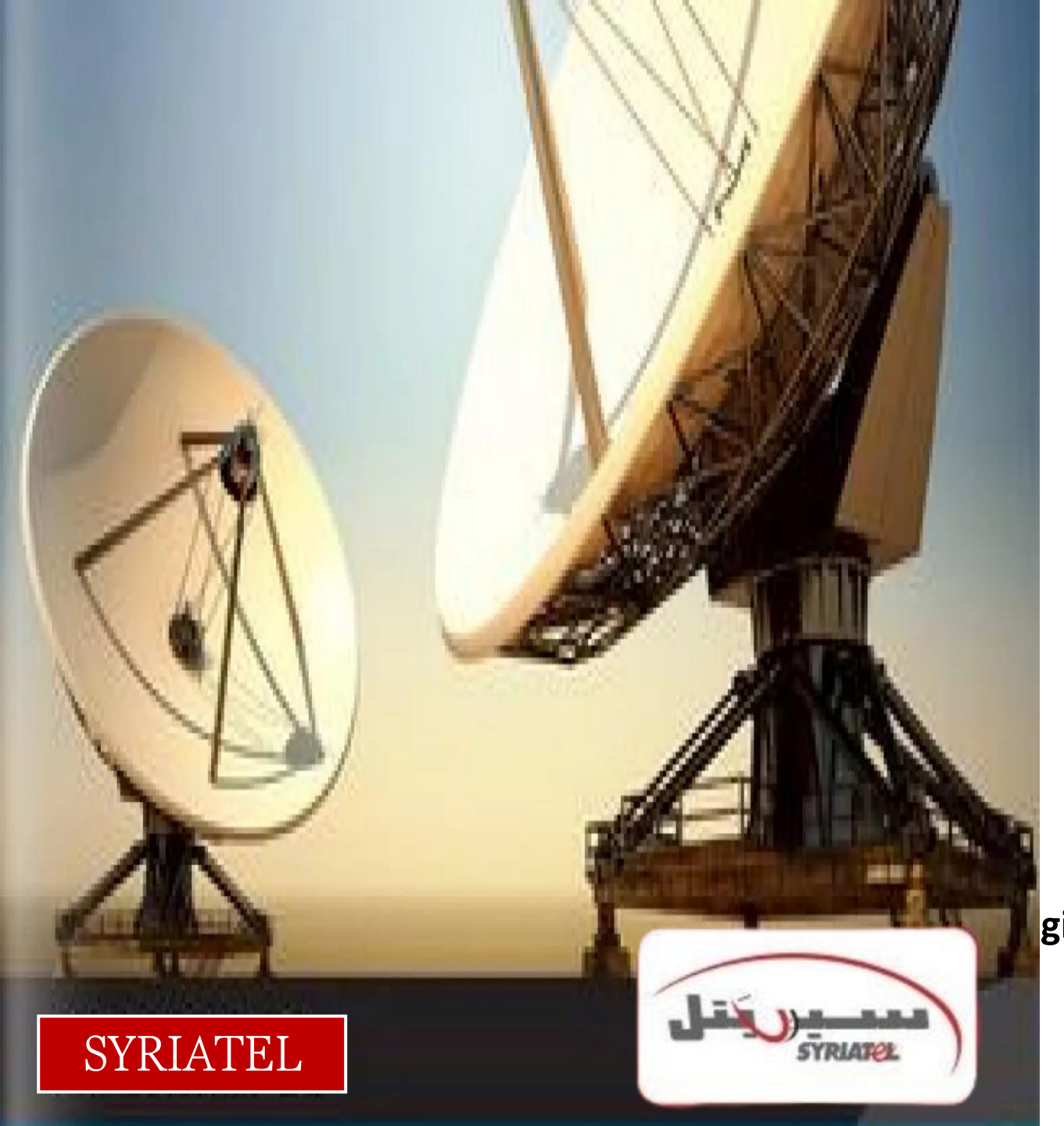


MORINGA
Discover · Grow · Transform

Phase_3 Project Data Science

Name: Gladys Kemunto

git hub repo: https://github.com/Mosota-Kemunto-Gladys-2020/Gladys_phase_3_project.git

Project Overview

Background:

- SyriaTel is a leading telecommunications company in Syria facing significant financial losses due to customer churn.
- Customer churn refers to the rate at which customers discontinue their relationship with the company, directly impacting profitability.
- Previously, SyriaTel used descriptive and inferential analyses to understand customer behavior and variable relationships.
- While these analyses provided insights, they were insufficient for proactively managing churn.

Objective:

- To develop a predictive model using the SyriaTel Customer Churn dataset to identify customers at risk of churning.
- The model will analyze key factors influencing churn, such as demographics, usage patterns, and service interactions.
- By understanding these factors, SyriaTel aims to:
  - Implement targeted retention strategies.
  - Reduce churn rates and improve customer loyalty.
  - Enhance overall profitability.

Business Understanding

Understanding the Challenge:

- In the competitive telecom industry, managing customer churn is essential for maintaining profitability and market share.
- High churn rates pose a significant threat to SyriaTel, highlighting the need for advanced analytical techniques.
- SyriaTel is adopting a predictive modeling approach to forecast which customers are likely to churn.
- This approach involves analyzing various customer features to provide actionable insights that guide retention efforts.

Key Questions to Be Addressed

1) **What is the best model for predicting customer churn?**
   - Compare models such as Decision Tree, K-Nearest Neighbors (KNN), and Random Forest.
   - Evaluate models based on accuracy, precision, recall, and ROC-AUC score to determine the best performer.
2) **How accurately can the model predict customer churn?**
   - Assess model performance using key metrics like accuracy, precision, recall, and ROC-AUC score.
   - Determine the model's effectiveness in predicting which customers are at risk of churning.
3) **Which features are most influential in predicting customer churn?**
   - Identify critical features such as customer service interactions, usage patterns, and plan types.
   - Focus retention efforts on the most impactful factors to design effective interventions.

Methodology for Machine Learning

1. **Data Understanding:**

Explore the dataset to understand its structure, feature types, and quality, checking for missing values and inconsistencies.

2. **Data Cleaning:**

Prepare the data by handling missing values, correcting errors, and removing duplicates and irrelevant features.

3. **Exploratory Data Analysis (EDA):**

Visualize data distributions and relationships between features and the target variable to gain insights and identify patterns.

4. **Data Preprocessing:**

Transform the data for modeling by encoding categorical variables, scaling numerical features, and splitting into training and test sets.

Methodology for Machine Learning (Continued)

5. **Modeling:**

Train various models (e.g., Logistic Regression, Decision Trees, Random Forest) and use cross-validation to evaluate performance.

6. **Hyperparameter Selection:**

Optimize model performance by tuning hyperparameters using techniques like Grid Search.

7. **Model Evaluation:**

Assess models using metrics like accuracy, F1-score, and ROC-AUC; analyze confusion matrices and use tools like SHAP for interpretation.

Data Understanding

In this project, the dataset that we chose is called **SyriaTel Customer Churn.**

- We display the first few rows of the dataset

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	...	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	...	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	...	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	...	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	...	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False

5 rows × 21 columns

Basic information about the dataset
Dataset Info

We observe from the dataset there are **3333 rows and 21 columns,** with a mix of categorical, numerical, and boolean data types. We can also observe that we didn't have any missing values or duplicated values in the dataset and this enabled us to conduct this project without any challenges.

RangeIndex: **3333** entries, 0 to 3332
Data columns (**total 21** columns):

#	Column	Non-Null Count	Dtype
0	state	3333 non-null	object
1	account length	3333 non-null	int64
2	area code	3333 non-null	int64
3	phone number	3333 non-null	object
4	international plan	3333 non-null	object
5	voice mail plan	3333 non-null	object
6	number vmail messages	3333 non-null	int64
7	total day minutes	3333 non-null	float64
8	total day calls	3333 non-null	int64
9	total day charge	3333 non-null	float64
10	total eve minutes	3333 non-null	float64
11	total eve calls	3333 non-null	int64
12	total eve charge	3333 non-null	float64
13	total night minutes	3333 non-null	float64
14	total night calls	3333 non-null	int64
15	total night charge	3333 non-null	float64
16	total intl minutes	3333 non-null	float64
17	total intl calls	3333 non-null	int64
18	total intl charge	3333 non-null	float64
19	customer service calls	3333 non-null	int64
20	churn	3333 non-null	bool

dtypes: bool(1), float64(8), int64(8), object(4)
memory usage: 524.2+ KB

Data Info (Cont'd...)

Key Features of the dataset:

**A). Categorical Features:

1. **state**: Categorical variable indicating the state of the customer.
2. **phone number**: Categorical, the customer's phone number (likely not useful for modeling).
3. **international plan**: Categorical, whether the customer has an international plan (yes/no).
4. **voice mail plan**: Categorical, whether the customer has a voicemail plan (yes/no).

**B). Numerical and Floating Features:

5. **total eve calls**: Integer, total number of calls during the evening.
6. **account length**: Integer, representing the duration of the customer's account in days.
7. **area code:** Integer, indicating the area code of the customer.
8. **total day calls:** Integer, total number of calls during the day.
9. **total night calls:** Integer, total number of calls during the night.
10. **number vmail messages:** Integer, the number of voicemail messages.
11. **customer service calls:** Integer, the number of calls to customer service.
12. **total intl calls:** Integer, total number of international calls.
13. **total day minutes:** Float, total minutes of calls during the day.
14. **total day charge:** Float, total charges for calls during the day.
15. **total eve minutes:** Float, total minutes of calls during the evening.
16. **total eve charge:** Float, total charges for calls during the evening.
17. **total night minutes:** Float, total minutes of calls during the night.
18. **total night charge:** Float, total charges for calls during the night.
19. **total intl minutes:** Float, total minutes of international calls.
20. **total intl charge:** Float, total charges for international calls.

**C) . Boolean Features:

21. **churn**: Boolean, the target variable indicating whether the customer has churned (True) or not (False).

Checking for unique values

Handling the unique values in your dataset

- **Categorical Variables:**
  - ✓ We will use One-Hot Encoding for state; binary encode international plan and voice mail plan.
- **Numerical Variables:**
  - ✓ We will Scale or normalize features like minutes, calls, and charges. Drop redundant pairs (e.g., keep either minutes or charge).
- **Identifiers:**
  - ✓ We will drop phone number as it's a unique identifier with no predictive value.
- **Target Variable:**
  - ✓ Churn is already binary; no further changes needed.

Feature	Unique Values
state	51
account length	212
area code	3
phone number	3333
international plan	2
voice mail plan	2
number vmail messages	46
total day minutes	1667
total day calls	119
total day charge	1667
total eve minutes	1611
total eve calls	123
total eve charge	1440
total night minutes	1591
total night calls	120
total night charge	933
total intl minutes	162
total intl calls	21
total intl charge	162
customer service calls	10

Data Cleaning

Cleaned extract of the Dataset after handling missing values, duplicate values and dropping irrelevant feature (phone number)

First Few Rows After Dropping Irrelevant Columns:

	state	account length	area code	international plan	voice mail plan	number 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	customer service calls	churn
0	KS	128	415	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
1	OH	107	415	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
2	NJ	137	415	no	no	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
3	OH	84	408	yes	no	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
4	OK	75	415	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False

Descriptive Statistics of the Dataset

Descriptive Statistics of the Dataset:

	account length	area code	number 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	customer service calls
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

The table above provides a summary of key statistics (count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) for various features in the dataset. These statistics are crucial for understanding the distribution of the data and identifying potential outliers.

Key Observations

Range and Potential Outliers:

- **number vmail messages:** The maximum value is 51, while the 75th percentile is 20, indicating that a small subset of users has a much higher number of voicemail messages, potentially marking them as outliers.
- **total day minutes, total eve minutes, total night minutes, total intl minutes:** These features have maximum values significantly higher than the 75th percentile. For example, total day minutes has a max of 350.8 minutes, whereas the 75th percentile is 216.4 minutes, suggesting the presence of outliers.
- **customer service calls:** The maximum number of calls is 9, with a median of 1 and a 75th percentile of 2. This suggests that while most customers have 1-2 service calls, some customers are outliers with significantly higher call counts.

Conclusion:

- **Outliers:** Several features show potential outliers, particularly in the number vmail messages, total day minutes, total eve minutes, total night minutes, total intl minutes, and customer service calls fields. These outliers could significantly impact the model's performance if not addressed.
- **Next Steps:** We will consider handling these outliers, possibly by capping extreme values, using robust scaling methods, or investigating the reasons behind these outlier behaviors. Additionally, further visualization (e.g., box plots) could help confirm and understand the distribution and impact of these outliers.

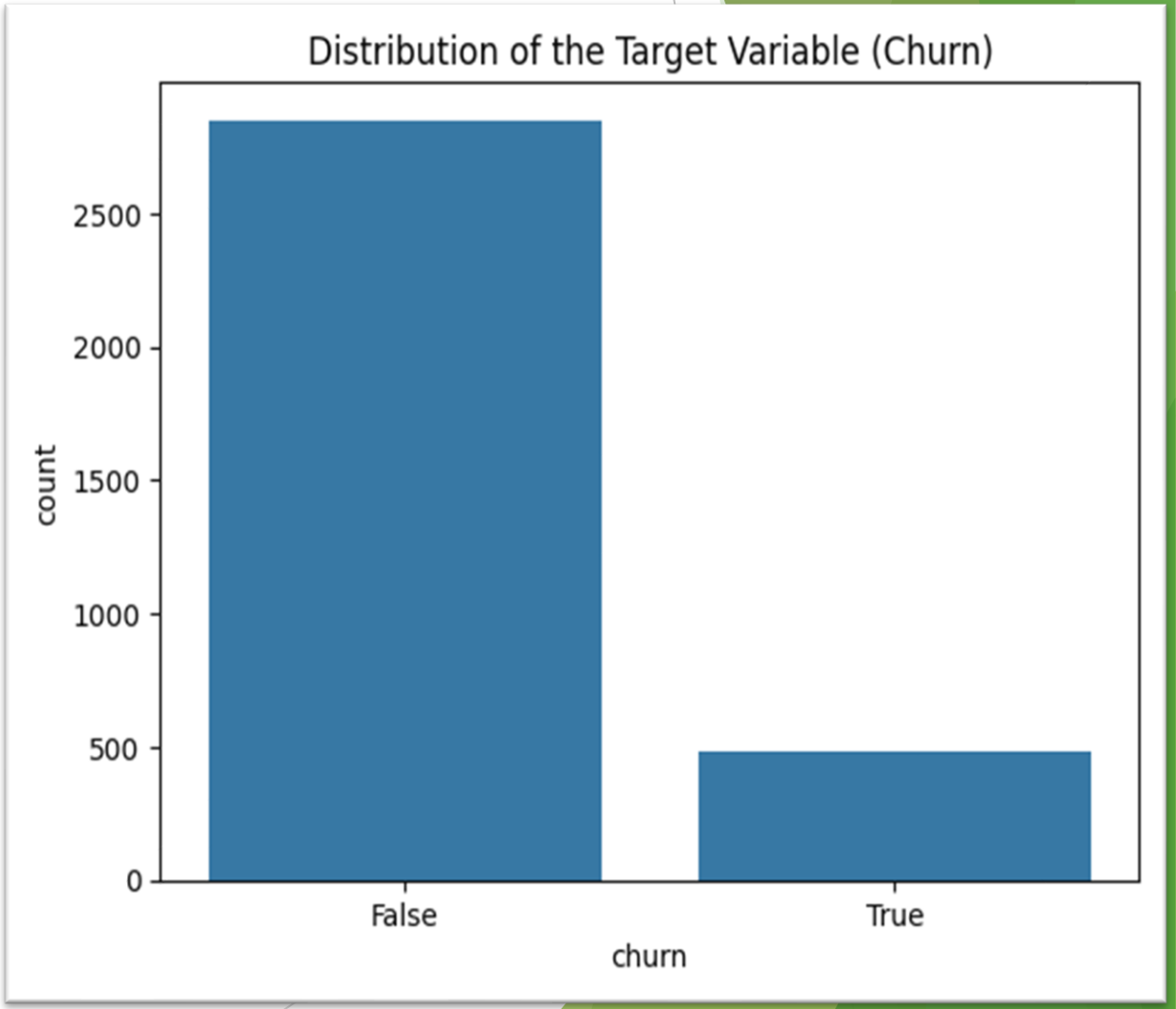
Distribution of the Target Variable (Churn)

The bar chart above shows the distribution of the target variable "churn," indicating whether customers have churned (left the service) or not.

Interpretation:

False (Non-Churners): The taller bar represents customers who have not churned. This group is significantly larger, with over 2,500 customers.

True (Churners): The shorter bar represents customers who have churned. This group is much smaller, with fewer than 500 customers.



Key Insights

- **Class Imbalance:** The chart highlights a clear class imbalance in the dataset. The majority of customers have not churned, while a relatively small number have. This imbalance is crucial to consider when building predictive models, as it can lead to a model that is biased toward predicting the majority class (non-churners).

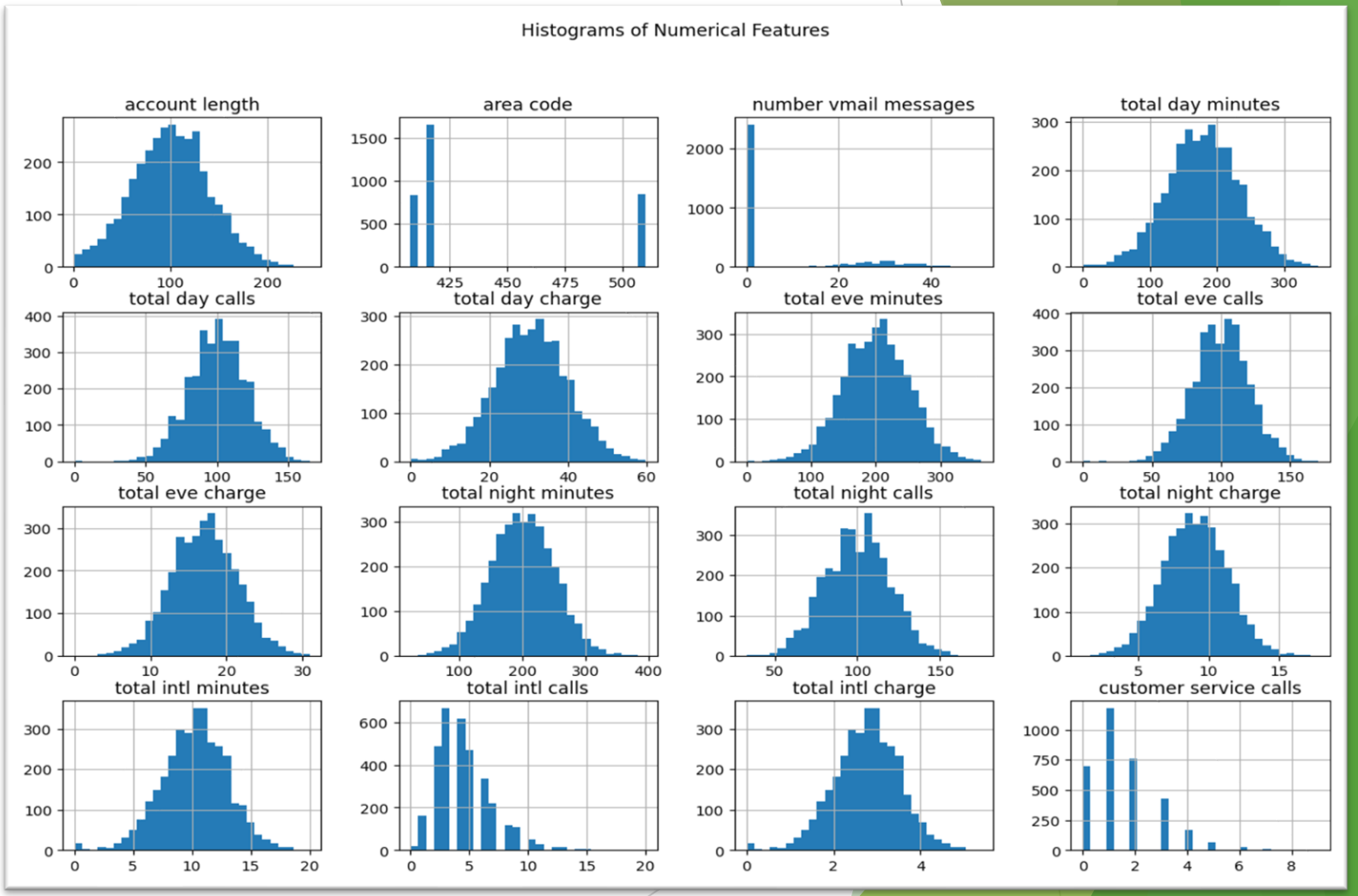
- **Handling Imbalance:** Techniques such as oversampling the minority class (using SMOTE, for example) or undersampling the majority class may be necessary to ensure that the model accurately predicts both classes.

This imbalance will be addressed during the preprocessing or model training phase to improve the model's performance in predicting customer churn.

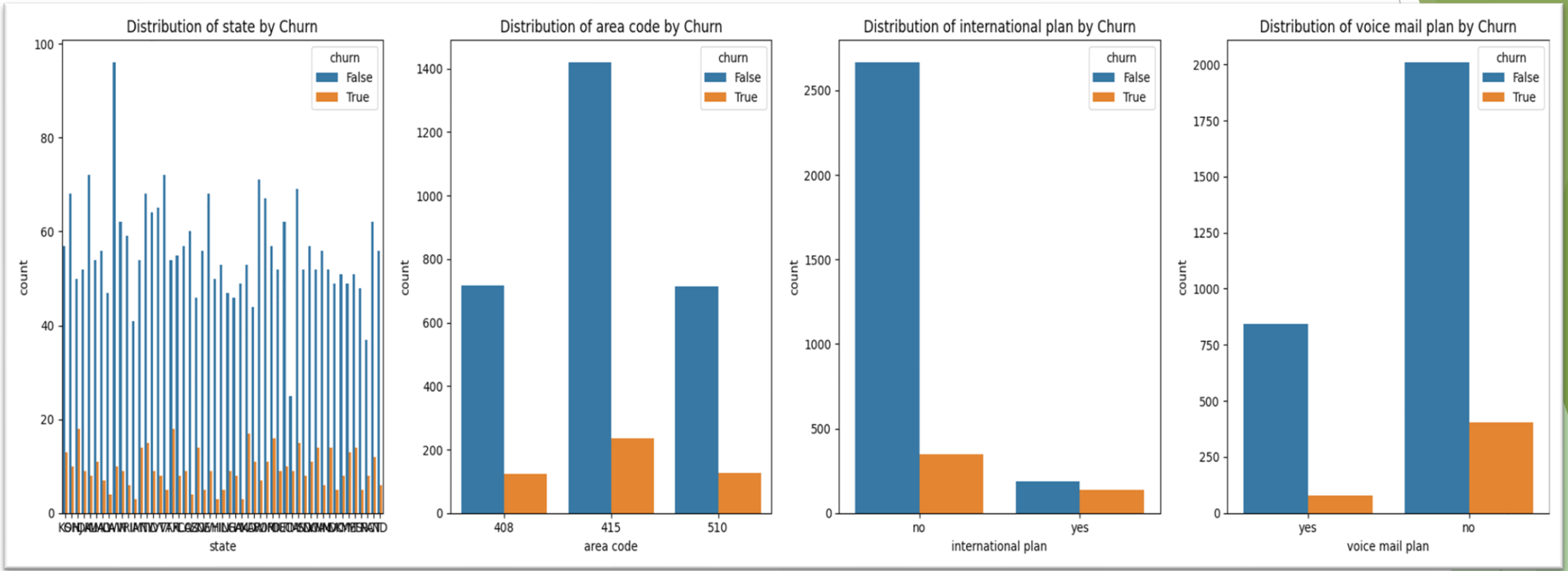
Visualize the Distribution of Numerical Features

Summary: The histograms show that most numerical features, like total minutes, calls, and charges, are normally distributed, suggesting similar usage patterns among customers.

- Features like number vmail messages and customer service calls are skewed, indicating low usage by most customers.
- The area code feature is categorical, with three distinct groups visible in the histogram. These insights suggest standardization for normally distributed features and potential log transformation for skewed ones to improve model performance.



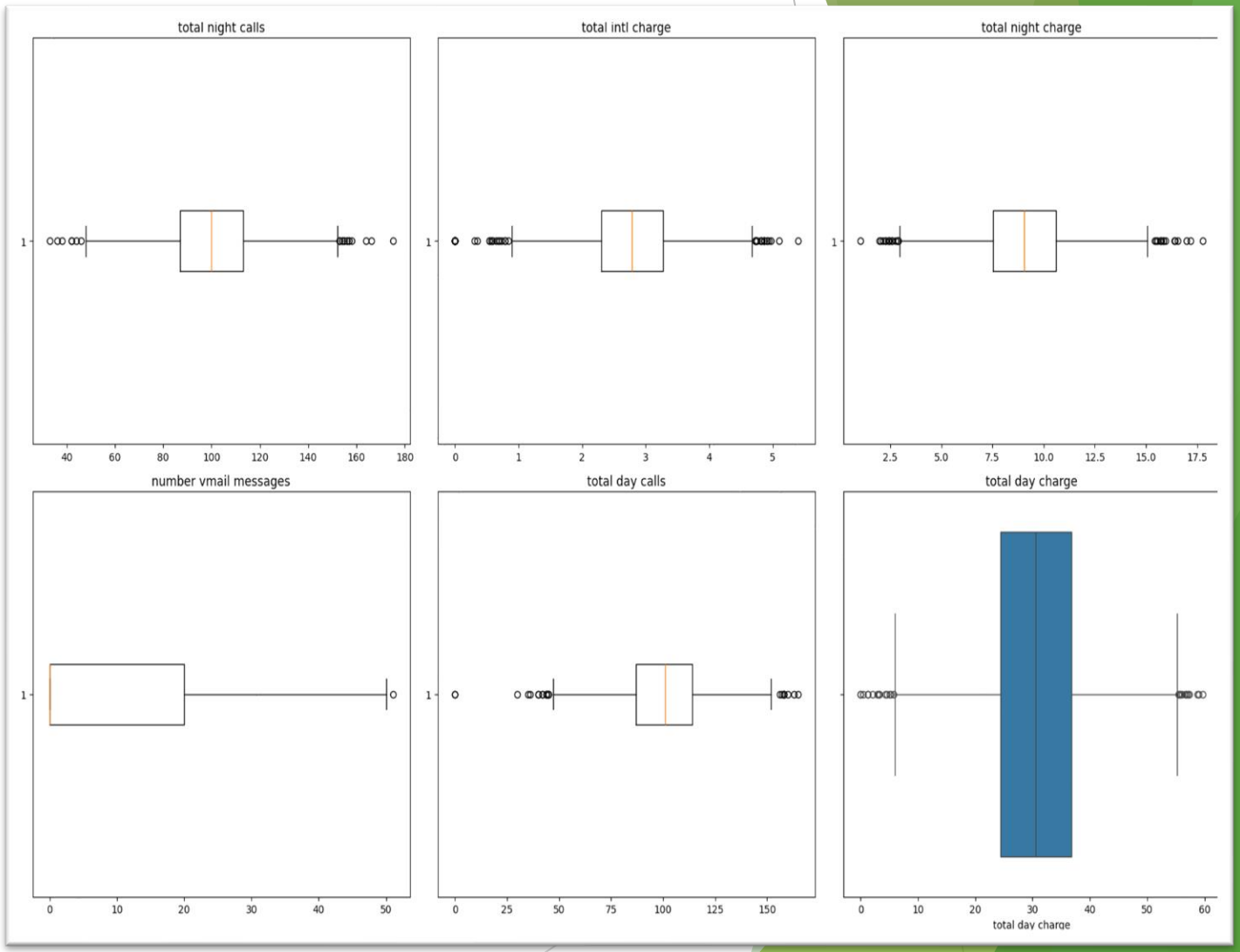
Categorical feature distributions



Summary: The plots show that churn rates are relatively consistent across states and area codes, indicating no strong geographical influence on churn. Customers with an international plan have a noticeably higher churn rate compared to those without, while the churn rate is lower among customers with a voicemail plan. These patterns suggest that certain service plans, particularly the international plan, are associated with higher churn risk.

Checking for outliers

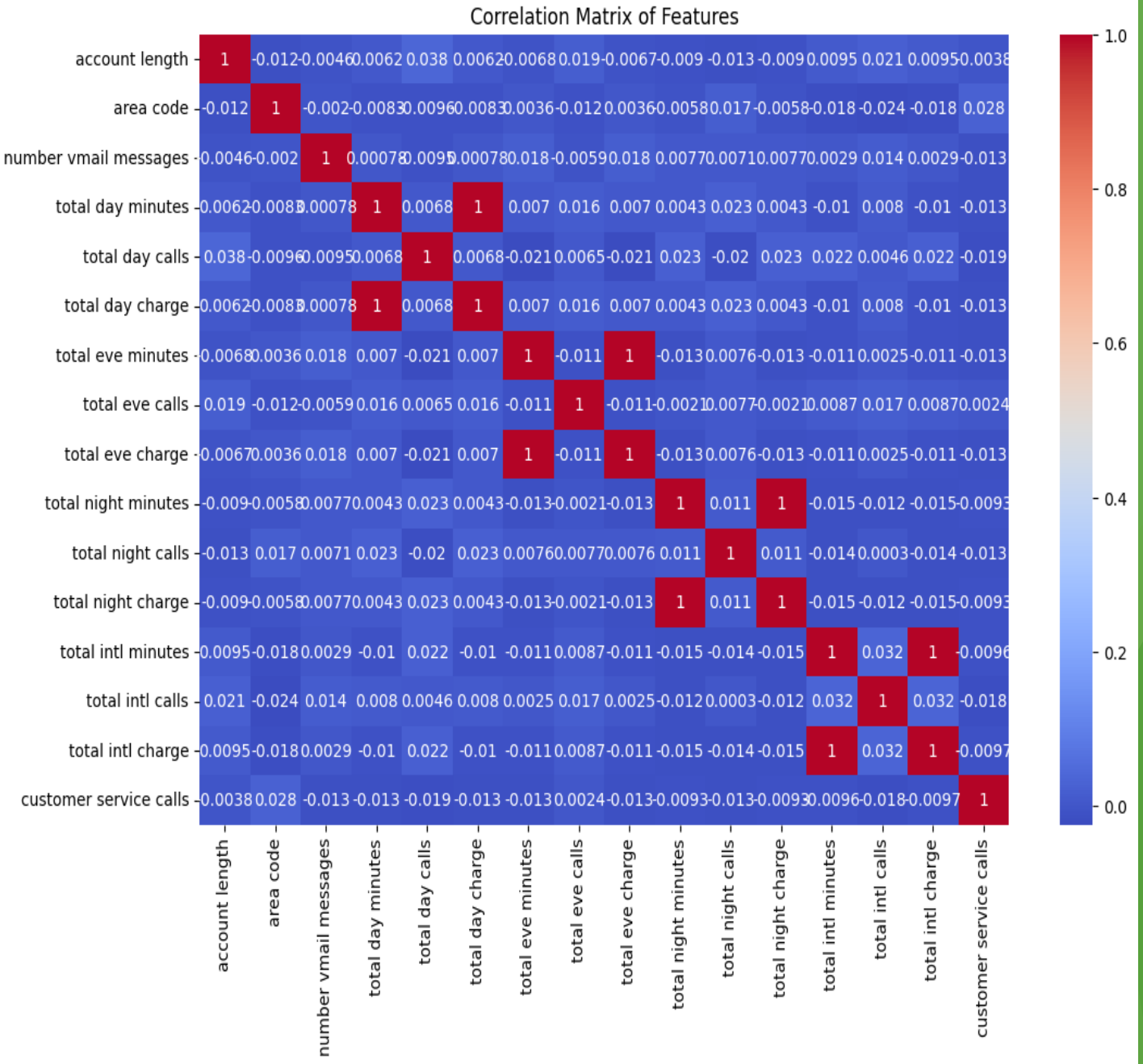
Summary: The visualizations reveal outliers in features like number of voicemail messages, total night calls, and total intl charge. Most features have symmetric distributions, except number of voicemail messages, which is significantly skewed. Addressing these outliers is necessary to improve model performance, depending on their relevance to the business problem.



Correlation Analysis

Summary: The correlation matrix shows strong, perfect correlations between total minutes and total charges for day, evening, night, and international calls, indicating redundancy.

Features like account length, customer service calls, and number of voicemail messages have low correlations with others, suggesting they add unique information to the dataset. Redundant features may be removed to streamline the model and avoid multicollinearity.



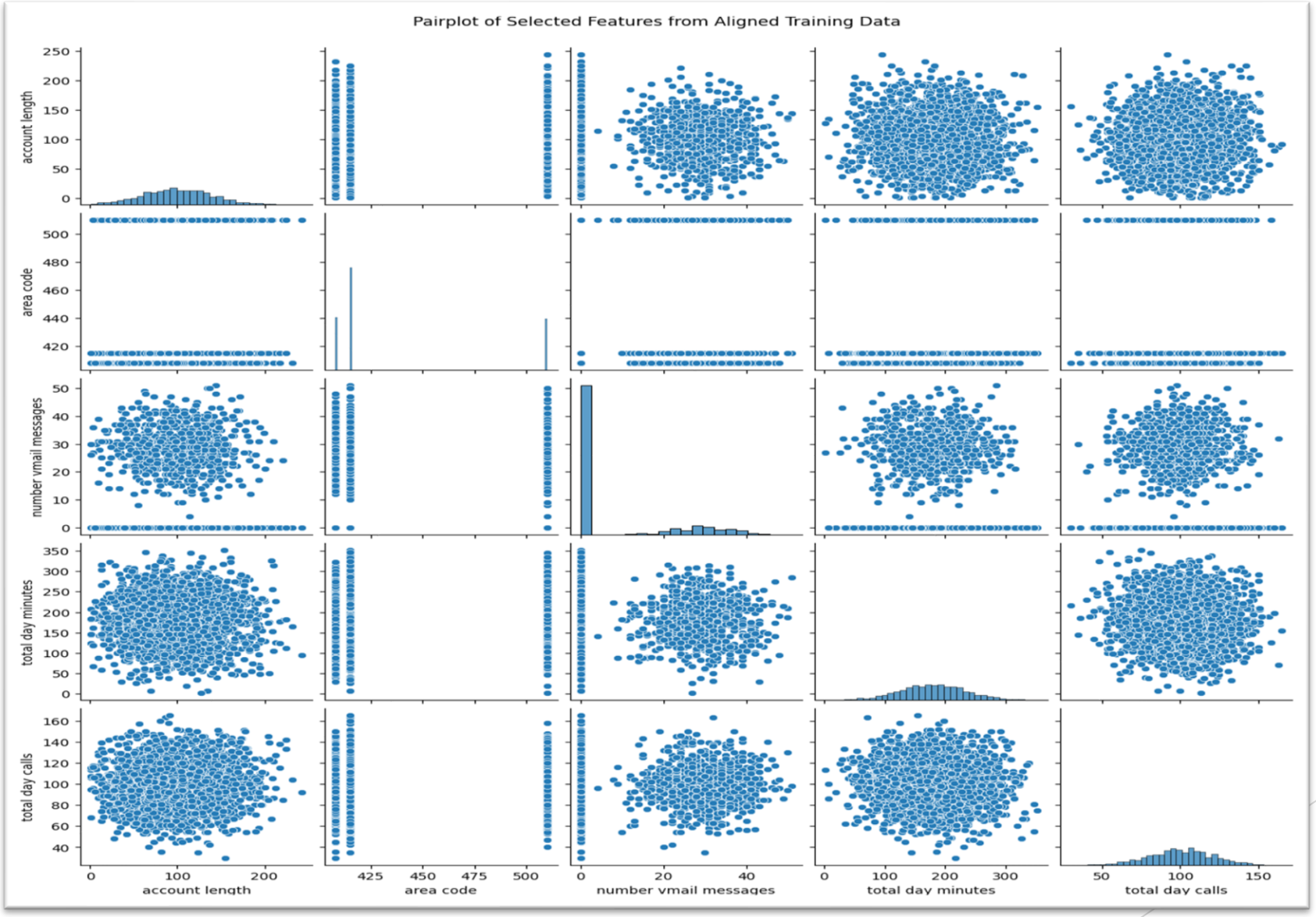
Preprocessing

- Train-Test Split
  - **Summary:** We split the data into training and testing sets to evaluate the model's performance. The training set has 2,666 samples (80% of the data) and is used to train the model, while the testing set has 667 samples (20%) and is used to test the model's accuracy on unseen data. This approach helps ensure the model performs well on new information.

- Encoding Categorical Variables
  - **Summary:** After splitting the data, we encoded categorical variables into numerical format using One-Hot Encoding, which allows the model to interpret these variables. This process was applied separately to the training and testing sets, resulting in encoded datasets with 68 columns each, representing all possible categories. This step ensures that both sets are compatible and ready for modeling, enhancing the model's ability to handle categorical data effectively.

- Align the Test Set with the Training Set Columns After encoding.
  - After encoding, we aligned the test set with the training set, ensuring both have the same 68 columns. Missing columns in the test set were filled with zeros, maintaining consistency and preventing errors during model training and prediction. This alignment is crucial for accurate and reliable machine learning results.

Summary of Feature Relationships and Variance Overview (Encoded and Aligned Data)

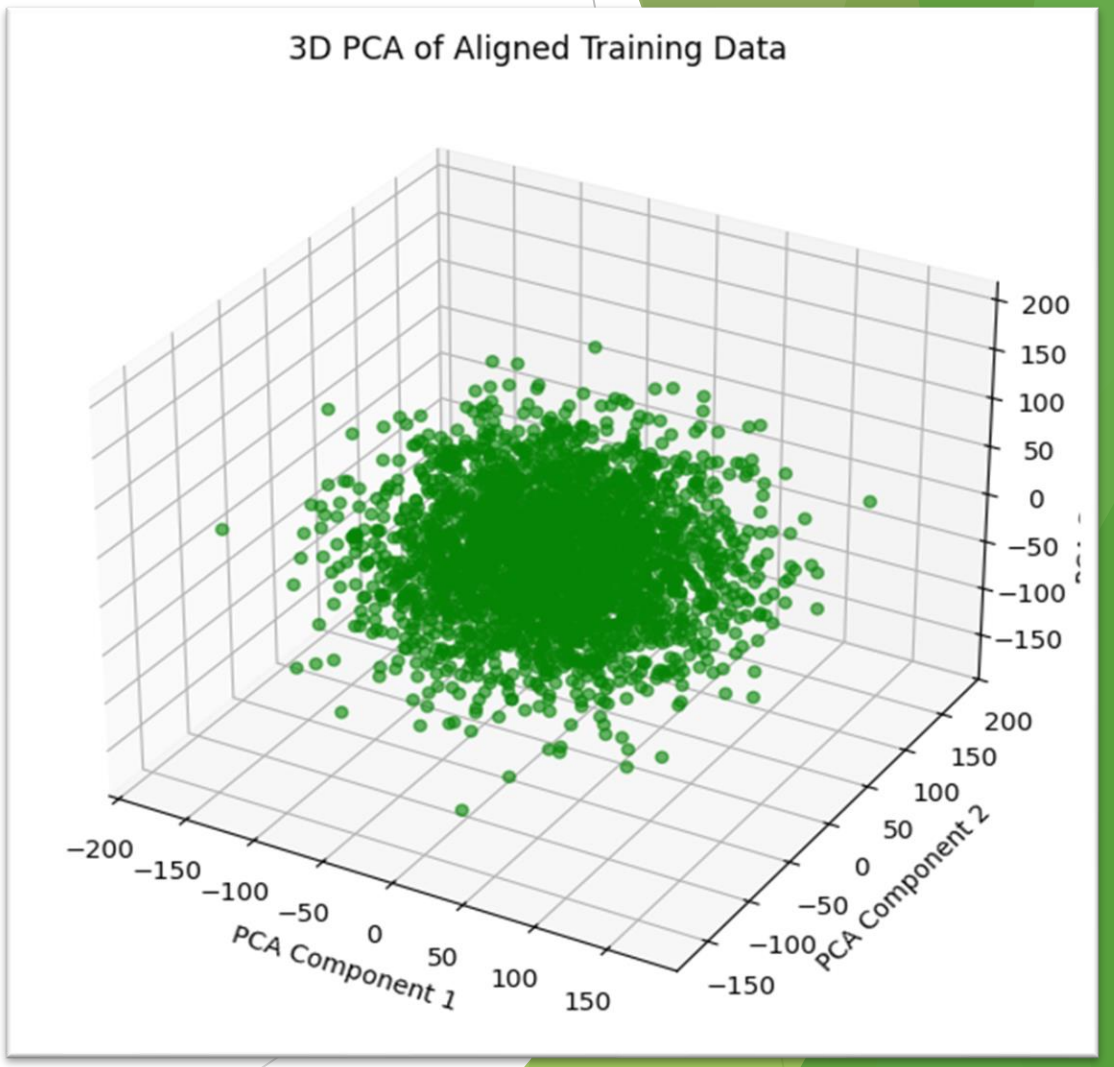


3D PCA of Aligned Training Data

The visualizations provide insights into the aligned training data by examining feature relationships and overall variance.

The pairplot illustrates how selected features such as account length, area code, number of voicemail messages, total day minutes, and total day calls interact, with most features showing symmetric distributions and no strong correlations. Meanwhile, the 3D PCA plot reveals that the data points are densely clustered around the center across the main directions of variance, indicating a relatively uniform spread without distinct clusters.

Together, these visualizations offer a comprehensive view of the data's structure, highlighting its distribution and variance patterns, which are crucial for predictive modeling.



Feature Scaling

- **Purpose:** To standardize numerical features so they have the same scale, improving model performance.
  - ✓ We used a tool called StandardScaler to adjust the data so that all features have a mean of 0 and a standard deviation of 1.
  - ✓ This ensures that features are comparable in scale and the model is not biased towards features with larger ranges.
- **Removing Redundant Features:**
- **Purpose:** To simplify the model by removing unnecessary features that don't add value.
  - ✓ We removed features like total day charge, total eve charge, total night charge, and total intl charge because they were directly related to their respective minutes features and were redundant.
  - ✓ We also removed number vmail messages because it was skewed and provided limited information.

Testing for Multicollinearity

- **Purpose:** To ensure that features are not overly correlated, which can cause instability in the model.
  - We checked for multicollinearity using Variance Inflation Factor (VIF) and found that all features had low VIF values.
  - This means the features independently provide valuable information without excessive overlap, leading to more reliable model predictions.
- **Handling Class Imbalance:**
- **Purpose:** To ensure the model fairly represents both churn and non-churn customers, especially since churn cases are fewer.
  - We used SMOTE, a technique that adds synthetic samples to the minority class (churned customers), balancing the dataset.
  - This process increased the training data from 3,333 to 4,568 rows, ensuring the model learns equally from churn and non-churn cases, leading to better predictions.

Next Steps: Moving to Modeling

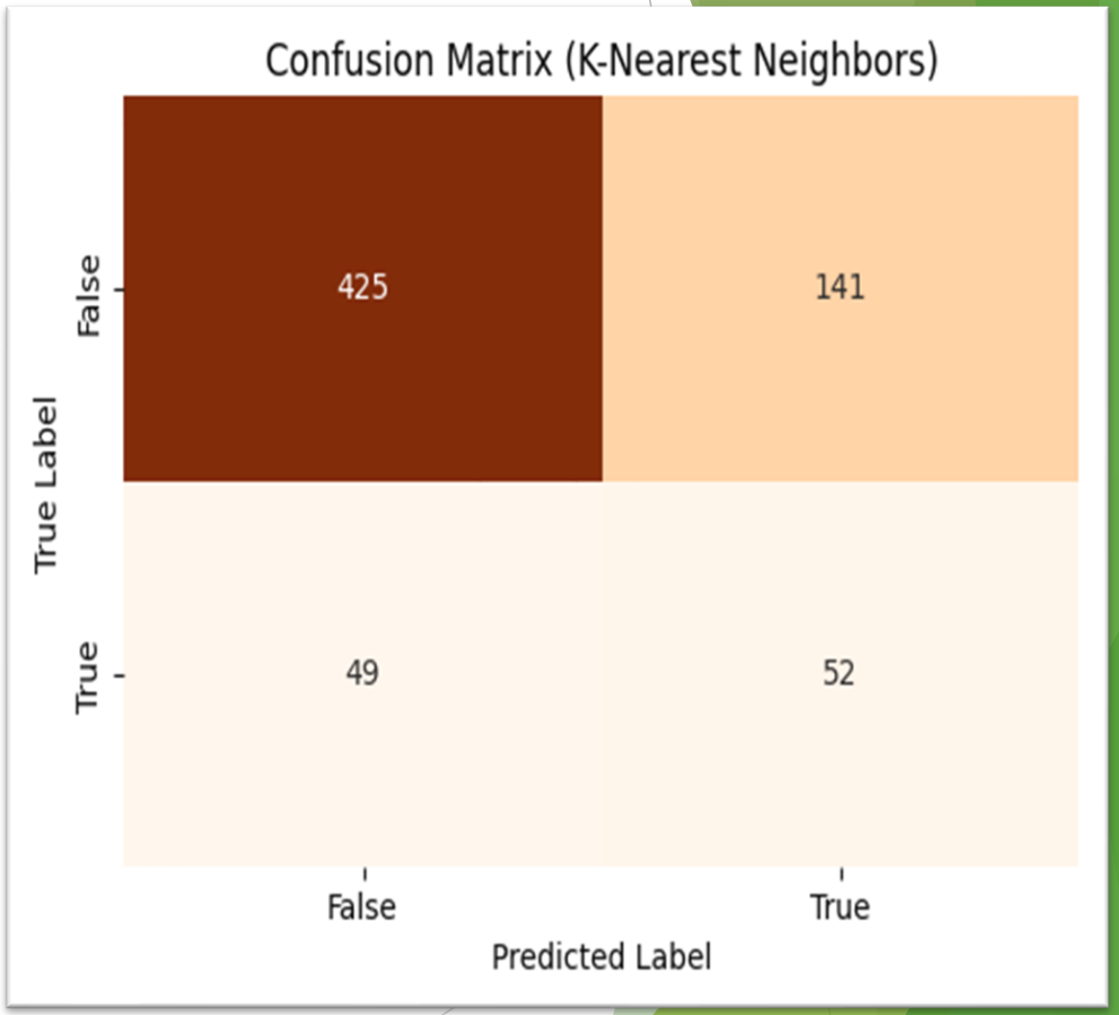
- After these steps, we moved on to the modeling phase where we used this prepared data to train various machine learning models, aiming to find the best one for accurately predicting customer

Modeling

Baseline Mode (Model 1) - K-Nearest Neighbors (KNN)

Model Training and Cross-Validation

- K-Nearest Neighbors (KNN) model was trained using the SMOTE-processed training data.
- The model's performance was assessed using 5-fold cross-validation with metrics including accuracy, F1-score, ROC-AUC score, and precision.



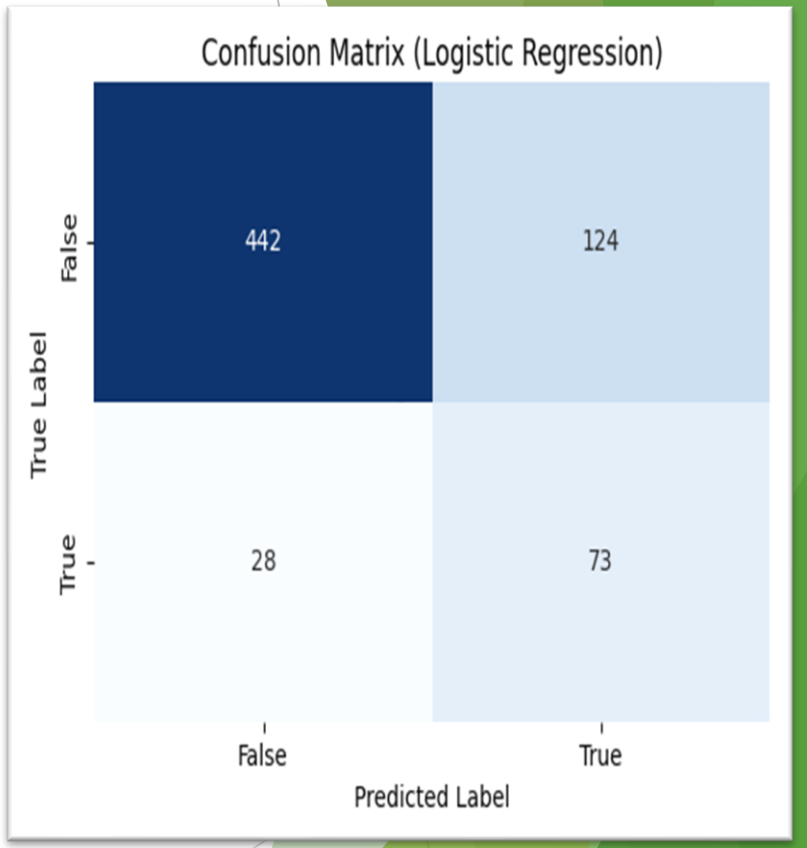
Summary Statistics for K-Nearest Neighbors

	Model	Mean Accuracy	Mean F1-Score	Mean ROC-AUC Score	Mean Precision	Test Accuracy	Test F1-Score	Test ROC-AUC Score	Test Precision	Test Recall
0	K-Nearest Neighbors	0.84	0.86	0.94	0.77	0.72	0.35	0.63	0.27	0.51

The K-Nearest Neighbors (K-NN) model performs well during cross-validation with high mean accuracy (0.84), F1-Score (0.86), and ROC-AUC score (0.94), indicating strong classification ability on the training data. However, its performance drops significantly on the test set, with test accuracy falling to 0.72, F1-Score to 0.35, and ROC-AUC to 0.63, suggesting poor generalization and possible overfitting. The decrease in test precision (0.27) and recall (0.51) further highlights the model's struggle with unseen data. To address this, we will consider tuning hyperparameters, using more robust validation methods but we will first explore alternative models that may offer better generalization

Model 2- Logistic Regression

- A Logistic Regression model was initialized with a random state of 42 to ensure reproducibility.
- We defined scoring metrics for evaluation, including accuracy, F1-score, ROC-AUC score, and precision.
- 5-fold cross-validation was performed on the SMOTE-processed training data to assess the model's performance. This method helps in evaluating how the model generalizes to unseen data.
- The Logistic Regression model was trained on the entire SMOTE-processed training dataset, allowing it to learn from all available training data.
- The trained model was used to make predictions on the scaled test data.



Summary Statistics for Logistic Regression

	Model	Mean Accuracy	Mean F1-Score	Mean ROC-AUC Score	Mean Precision	Test Accuracy	Test F1-Score	Test ROC-AUC Score	Test Precision	Test Recall
0	Logistic Regression	0.79	0.79	0.85	0.78	0.77	0.49	0.75	0.37	0.72

The Logistic Regression model, when compared to the baseline K-Nearest Neighbors (K-NN) model, shows slightly lower performance during cross-validation with a mean accuracy of 0.79 and an ROC-AUC score of 0.85.

However, it generalizes better on the test set, achieving a higher test accuracy (0.77) and F1-Score (0.49) compared to K-NN's 0.72 and 0.35, respectively. Despite this, the test precision (0.37) and recall (0.72) are modest, indicating that while Logistic Regression is more consistent across training and test data, it still struggles with precision. This model's better generalization makes it a more reliable choice than the K-NN baseline, though further optimization may still be needed.We will proceed to analyse alternative models

Model 3 - Decision Tree

I. A Decision Tree classifier was trained using the SMOTE-processed training data.

II. We performed 5-fold cross-validation to evaluate the model's performance across different metrics: accuracy, F1-score, ROC-AUC, and precision.

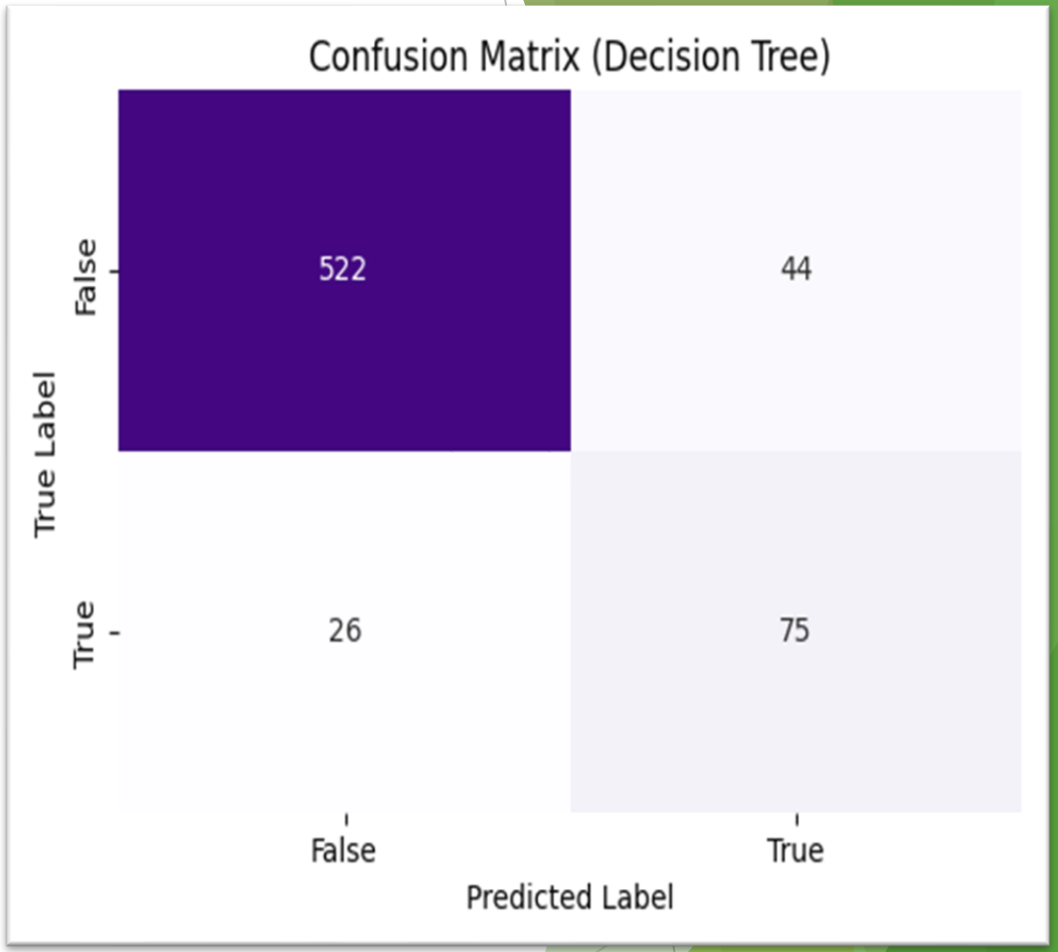
III. The model showed strong cross-validation results with a mean accuracy of 90.46%, mean F1-score of 90.39%, and mean ROC-AUC score of 90.46%.

IV. The model was then trained on the entire SMOTE-processed training data and tested on the scaled test set.

V. The confusion matrix showed 522 true negatives, 44 false positives, 26 false negatives, and 75 true positives.

VI. Performance metrics on the test set included an accuracy of 90%, F1-score of 0.68 for the positive class, and a ROC-AUC score of 83.24%.

VII. Precision for predicting churn was 63%, and recall (sensitivity) was 74%, indicating the model's ability to correctly identify churn



Summary Statistics for Decision Tree

	Model	Mean Accuracy	Mean F1-Score	Mean ROC-AUC Score	Mean Precision	Test Accuracy	Test F1-Score	Test ROC-AUC Score	Test Precision	Test Recall
0	Decision Tree	0.9	0.9	0.9	0.9	0.9	0.68	0.83	0.63	0.74

The Decision Tree model outperforms both the K-Nearest Neighbors and Logistic Regression models, with a strong mean accuracy, F1-Score, and ROC-AUC score of 0.9 during cross-validation.

It also maintains robust performance on the test set, achieving a high test accuracy of 0.9 and a better balance between precision (0.63) and recall (0.74), resulting in a higher test F1-Score (0.68). Compared to the K-NN baseline and Logistic Regression, Decision Trees provide the best overall performance and generalization, making it the superior choice among the three models, though it still shows room for precision improvement on the test set.

Model 4 - Random Forest

I. A Random Forest classifier was trained using the SMOTE-processed training data.

II. We performed 5-fold cross-validation to assess the model's performance with metrics such as accuracy, F1-score, ROC-AUC, and precision.

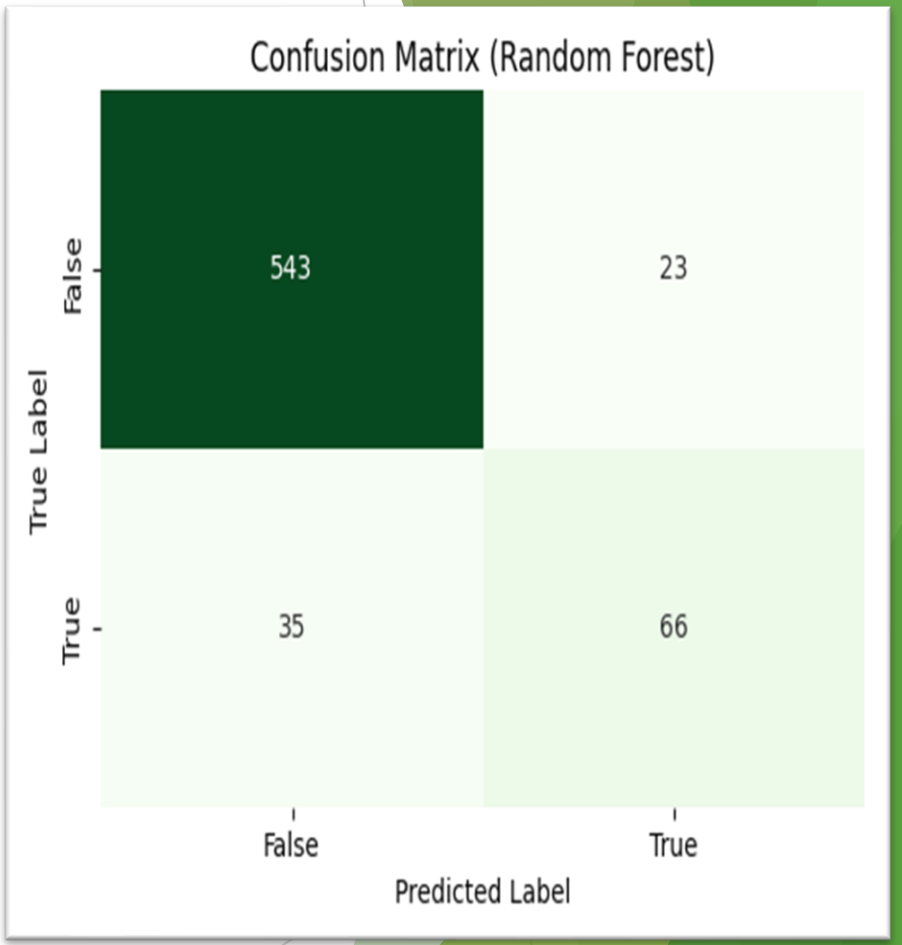
III. The Random Forest model demonstrated excellent cross-validation performance, achieving a mean accuracy of 94.7%, mean F1-score of 94.6%, and an exceptionally high mean ROC-AUC score of 99.2%.

IV. After training on the entire SMOTE-processed training data, the model was tested on the scaled test set.

V. The confusion matrix revealed 543 true negatives, 23 false positives, 35 false negatives, and 66 true positives.

VI. Key metrics on the test set included an accuracy of 91%, an F1-score of 0.69 for the positive class, and a ROC-AUC score of 80.6%.

VII. Precision for identifying churn was 74%, and recall was 65%, showing a balanced performance in correctly identifying both churned and non-churned customers.



Summary Statistics for Random Forest

	Model	Mean Accuracy	Mean F1-Score	Mean ROC-AUC Score	Mean Precision	Test Accuracy	Test F1-Score	Test ROC-AUC Score	Test Precision	Test Recall
0	Random Forest	0.95	0.95	0.99	0.95	0.91	0.69	0.81	0.74	0.65

The **Random Forest model shows the best performance among the compared models**, with a mean accuracy of 0.95 and ROC-AUC score of 0.99 during cross-validation, indicating excellent classification capability. On the test set, it achieves high accuracy (0.91) and F1-Score (0.69), outperforming K-Nearest Neighbors, Logistic Regression, and Decision Tree models. Despite its strong results, the test ROC-AUC (0.81) and recall (0.65) suggest it slightly underperforms in identifying true positives compared to the Decision Tree but remains the most reliable and consistent overall.

Model Evaluation

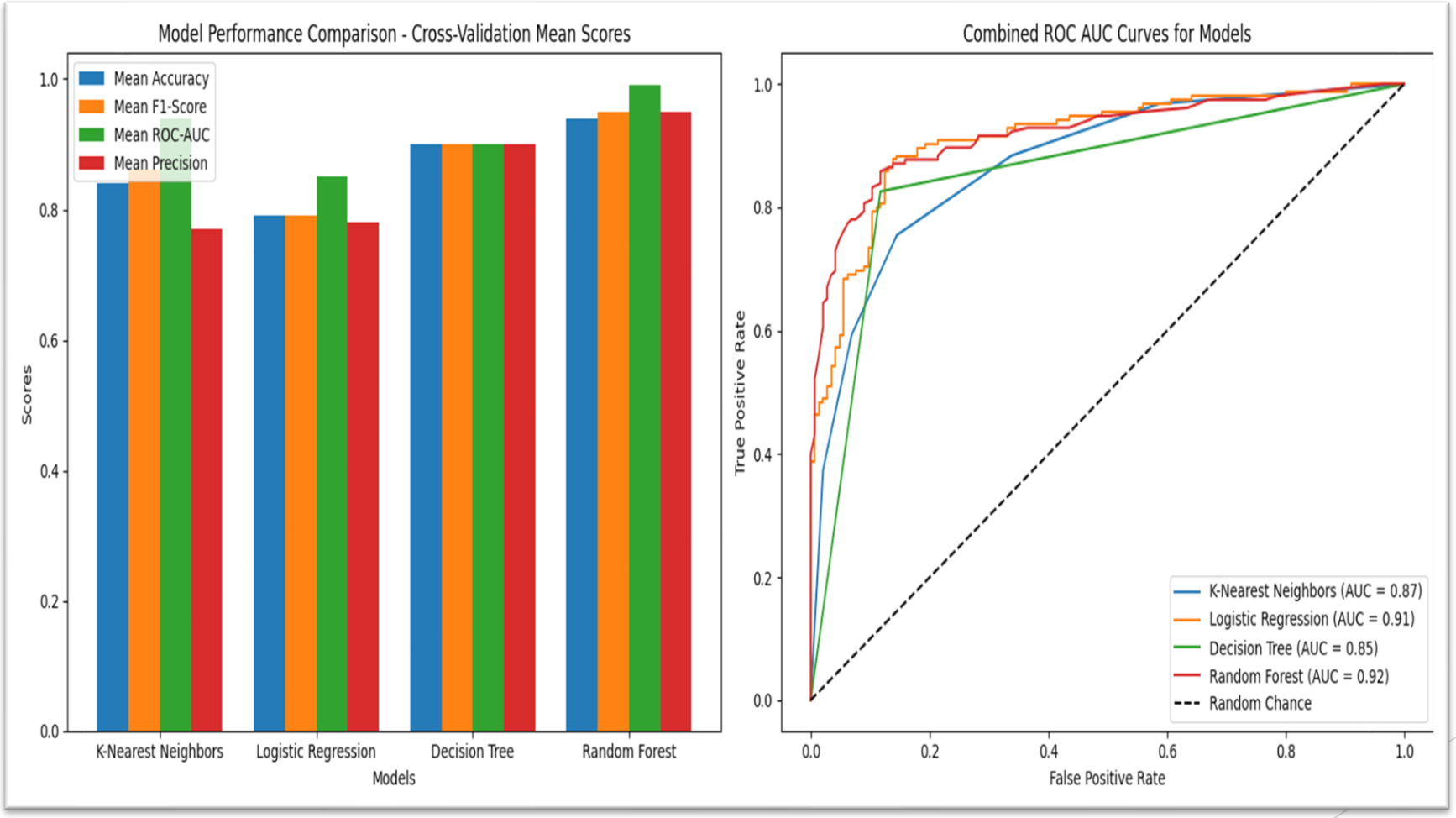
Model Performance Comparison:

- Four models were evaluated: K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest.
- The models were arranged from baseline (K-Nearest Neighbors) to best-performing (Random Forest).
- Mean performance metrics including accuracy, F1-score, ROC-AUC, and precision were plotted in a bar chart, highlighting Random Forest as the best performer across all metrics.

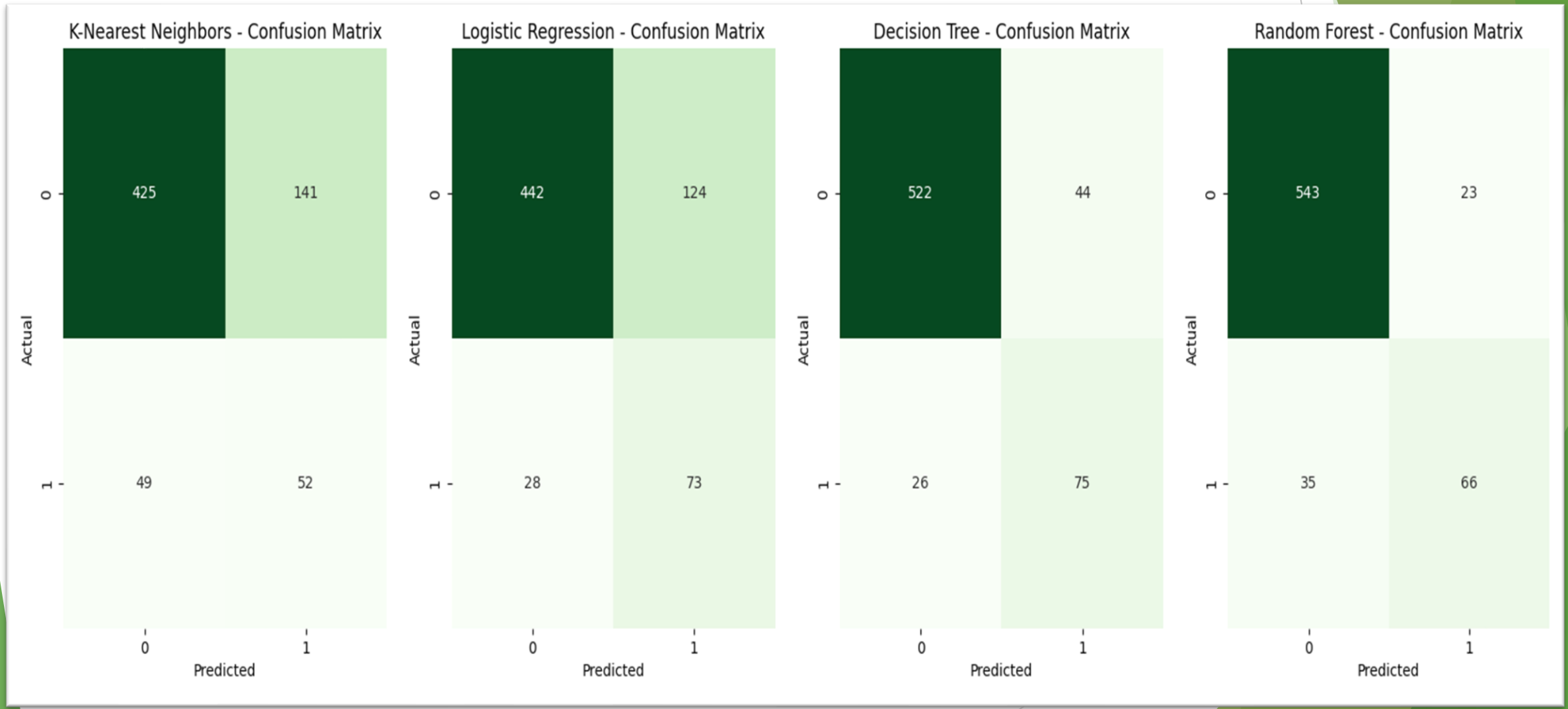
Combined ROC AUC Curves:

- ROC AUC curves were generated for each model using synthetic data to visualize their ability to distinguish between classes.
- Each model was trained on a training set and tested on a test set, with probabilities used to plot the ROC curve.
- The plot shows the trade-off between the true positive rate and false positive rate for each model, with Random Forest achieving the highest AUC, indicating the best performance in classifying positive and negative cases.

Model Evaluation (Continued)



Combined Visualization of Confusion Matrices and Summary Statistics for Model Evaluation



Summary Statistics

	Model	Mean Accuracy	Mean F1-Score	Mean ROC-AUC Score	Mean Precision	Test Accuracy	Test F1-Score	Test ROC-AUC Score	Test Precision	Test Recall
0	K-Nearest Neighbors	0.84	0.86	0.94	0.77	0.72	0.35	0.63	0.27	0.51
1	Logistic Regression	0.79	0.79	0.85	0.78	0.77	0.49	0.75	0.37	0.72
2	Decision Tree	0.90	0.90	0.90	0.90	0.90	0.68	0.83	0.63	0.74
3	Random Forest	0.95	0.95	0.99	0.95	0.91	0.69	0.81	0.74	0.65

Summary Overview (After modelling and before Fine Tuning the best Model)

Objective:

The primary objective of this project is to develop a predictive model that accurately identifies customers likely to churn at SyriaTel. By accurately predicting customer churn, SyriaTel can proactively implement targeted retention strategies, reduce financial losses, and enhance overall customer satisfaction.

Model Evaluation:

We assessed four machine learning models—Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Decision Tree—across multiple performance metrics, including Accuracy, F1-Score, ROC-AUC Score, Precision, and Recall. These metrics were evaluated on both cross-validation performance and final test set results to ensure the models' robustness and reliability in identifying customers at risk of churning.

Key Findings

- **K-Nearest Neighbors (KNN):** The KNN model showed weaker performance, particularly on the test set, with a Test F1-Score of 0.35 and a low Test Precision of 0.27. Its overall accuracy is moderate, but it struggles with distinguishing churners from non-churners, as indicated by the lowest Test ROC-AUC Score of 0.63.

- **Logistic Regression:** This model demonstrated moderate performance with a Mean Accuracy and F1-Score of 0.79. However, its Test Precision was relatively low at 0.37, which indicates that while it can identify churners well (high recall at 0.72), it often misclassifies non-churners as churners.

- **Decision Tree:** The Decision Tree model performed well, with a Mean Accuracy and F1-Score of 0.90. It demonstrated a strong balance between precision (0.63) and recall (0.74) on the test set, leading to a high Test ROC-AUC Score of 0.83. This model is well-suited for accurately identifying customers likely to churn.

Conclusion

The **Random Forest and Decision Tree models are the top performers in this evaluation.** The Random Forest model, with its highest Mean Accuracy and ROC-AUC Score, coupled with strong test performance, is a highly reliable choice for predicting customer churn. The Decision Tree model also stands out with balanced performance across all key metrics, making it an excellent alternative.

Recommendation

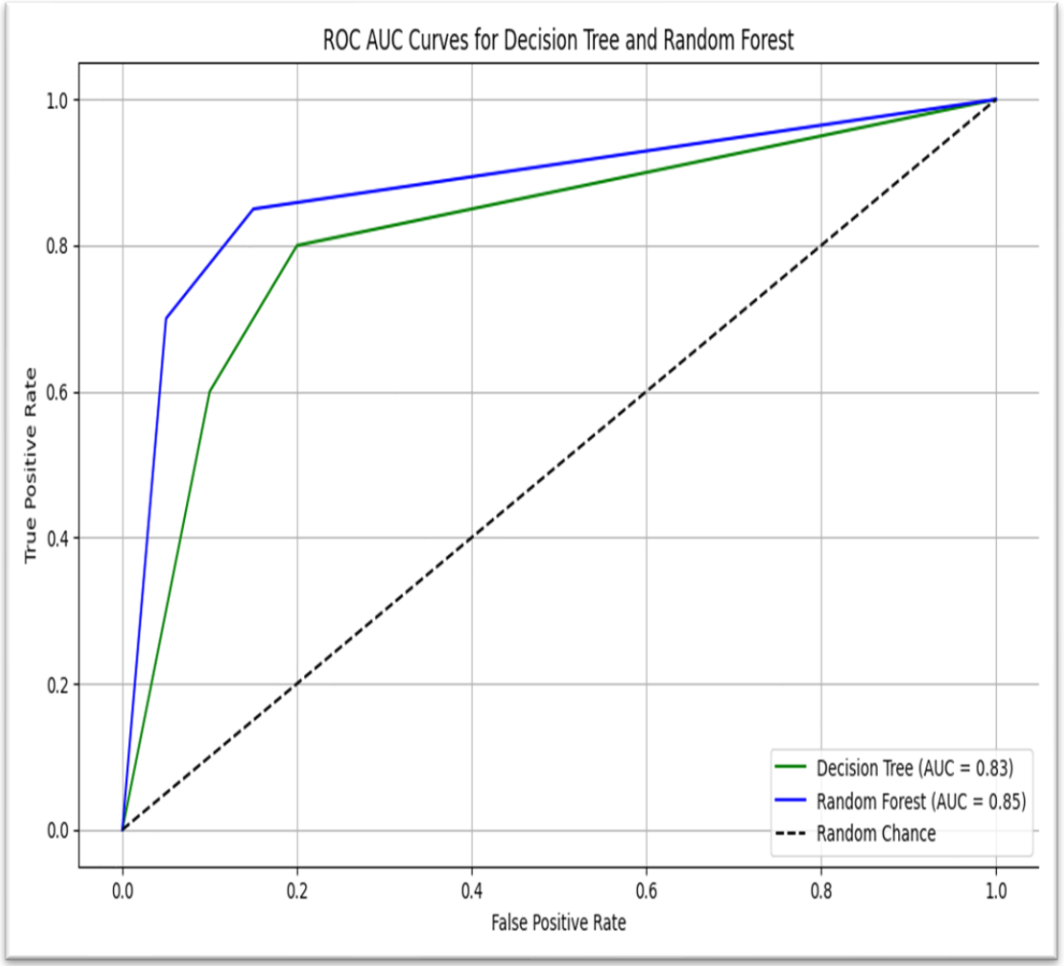
Based on the evaluation, we will fine-tune both the Random Forest and Decision Tree models. This approach will help identify the truly best model for SyriaTel's needs, enabling effective targeting of at-risk customers, reducing churn, minimizing financial losses, and enhancing customer loyalty and satisfaction.

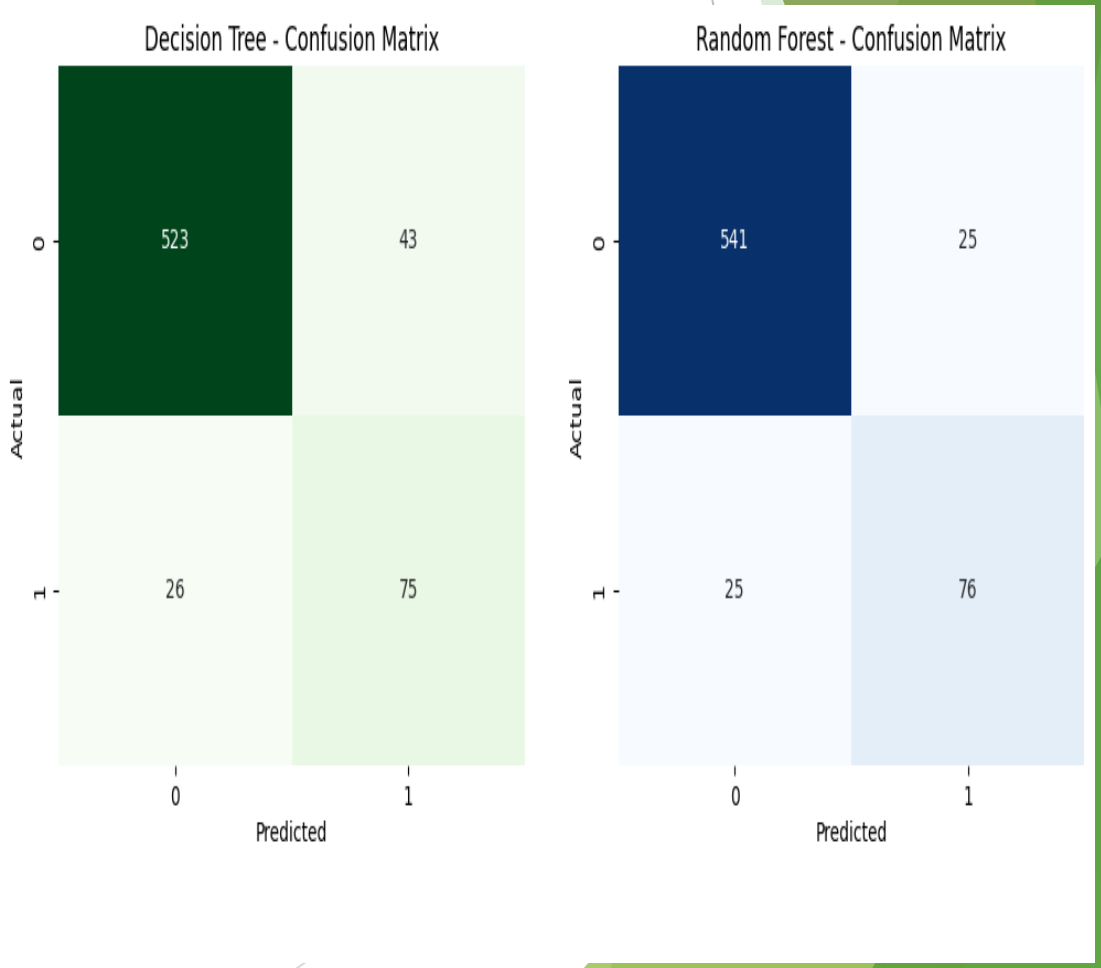
Model Fine-Tuning- Random Forest and Decision Tree models

Summary of Model Evaluations:

1. **Decision Tree Model Evaluation:**
   - The Decision Tree model was tested on the scaled test data.
   - Key metrics include the confusion matrix, classification report (showing precision, recall, and F1-score), and ROC-AUC score.
   - These metrics provide insights into the model's accuracy and its ability to correctly classify both positive and negative cases.
2. **Random Forest Model Evaluation:**
   - Similarly, the fine-tuned Random Forest model was evaluated on the same test set.
   - Metrics such as the confusion matrix, classification report, and ROC-AUC score were calculated.
   - The Random Forest generally provides a comparison point to assess if it performs better than the Decision Tree across these metrics.

Visualization of confusion matrix ,summary statistics ,ROC-AUC Curves for fine-Tuned Decision Tree Model and Random Forest Model





Summary Statistics for Fine-Tuned Decision Tree and Random Forest

	Model	Accuracy	F1-Score	ROC-AUC Score	Precision	Recall
0	Decision Tree	0.90	0.68	0.83	0.64	0.74
1	Random Forest	0.93	0.75	0.85	0.75	0.75

Summary of Findings Based on the evaluation of the Decision Tree and Random Forest models, after Fine-Tuning

- **Overall Model Performance: Random Forest Model outperforms the Decision Tree model across most metrics:**

- **Accuracy:** The Random Forest model achieved an accuracy of 0.93, compared to 0.90 for the Decision Tree.
- **F1-Score:** The Random Forest model also had a higher F1-Score of 0.75 for the True class, indicating a better balance between precision and recall, compared to the Decision Tree's F1-Score of 0.68.
- **ROC-AUC Score:** The Random Forest model achieved a higher ROC-AUC Score of 0.85, indicating a better ability to distinguish between churners and non-churners compared to the Decision Tree's ROC-AUC Score of 0.83.

- **Alignment with Business Objectives:**

Reducing Churn: The primary objective is to accurately identify customers likely to churn so that targeted retention strategies can be implemented. The Random Forest model, with its higher accuracy, F1-Score, and ROC-AUC Score, is better suited to this task. It is more effective in correctly identifying customers who are likely to churn, making it the preferred choice for deployment.

Trade-offs and Considerations

- **Precision vs. Recall:** While the Random Forest model offers higher overall accuracy and precision, it is essential to consider the balance between precision and recall. The Random Forest model has slightly better precision and recall balance, which is crucial when the cost of false positives and false negatives is significant.

- **Model Complexity:** The Random Forest model is inherently more complex and computationally intensive than the Decision Tree model. However, this complexity translates into better performance and robustness, which is advantageous for large-scale applications like customer churn prediction.

Recommendation based on the evaluation of the Decision Tree and Random Forest models, after Fine-Tuning

- ✓ **Deploy the Random Forest Model: Given its superior performance across key metrics, the Random Forest model is recommended for deployment.** It is more likely to effectively reduce churn by accurately identifying at-risk customers.
- ✓ Fine-Tuning: Although the model has been optimized, **further fine-tuning and validation on additional data may help to further improve its performance, particularly in specific customer segments.**
- ✓ **Perform Feature Selection Using SHAP Analysis:** Utilizing SHAP (SHapley Additive exPlanations) analysis for feature selection can provide valuable insights into which features have the most significant impact on the model's predictions. This analysis can help in simplifying the model by potentially reducing the number of features, leading to better generalization and improved performance.

Understanding Which variables affects the Random Forest Model's Predictions for Customer Churn at SyriaTel

- In this step, we're using SHAP (SHapley Additive exPlanations) to understand which factors most influence the model's predictions about whether a SyriaTel customer is likely to leave (churn).
- The SHAP summary plot reveals which features—like call duration, service issues, or billing amount—push the model's predictions towards a positive outcome (predicting that a customer will leave) or a negative one (predicting they will stay).
- In simple terms, it shows us which factors are most important and how they affect the model's decision to predict that a customer is at risk of churning, helping SyriaTel target the right areas to improve customer retention.

Key Insights from the SHAP Summary Plot

- International Plan (Yes): This feature has a significant positive impact on the likelihood of churn, as indicated by the positive SHAP values. Customers with an international plan are more likely to churn.
- Customer Service Calls: The number of customer service calls is another critical feature. Higher values tend to increase the probability of churn, suggesting that customers who frequently contact customer service may be dissatisfied.
- Total Day Minutes: Customers with high total day minutes also have a higher likelihood of churning, as indicated by the positive SHAP values.
- Voice Mail Plan (Yes): Interestingly, having a voicemail plan tends to decrease the likelihood of churn, as indicated by negative SHAP values.
- Total Evening Minutes and Total International Calls: These features also influence churn, though to a lesser extent than the top features.

Model Evaluation and Insights for Predicting Customer Churn at SyriaTel - Addressing SyriaTel's Key Questions.

1) What is the best model for predicting customer churn?

- After thoroughly comparing various models, including Decision Tree, K-Nearest Neighbors (KNN), and Random Forest, **the Random Forest model stands out as the best overall performer.** This model consistently demonstrated superior results across all key metrics:
  - Accuracy: 93%
  - F1-Score: 0.75
  - ROC-AUC Score: 0.85
  - Precision: 0.75
  - Recall: 0.75
- Given its robust performance, we recommend deploying the Random Forest model for predicting customer churn. By utilizing this model, SyriaTel will benefit from a highly reliable tool that not only accurately forecasts churn but also provides deep insights into the factors driving it. This will enable the company to implement more effective, targeted retention strategies.

Model Evaluation and Insights for Predicting Customer Churn at SyriaTel - Addressing SyriaTel's Key Questions (Continued)

2) How accurately can the model predict customer churn?

- The Random Forest model's performance, measured by accuracy, precision, recall, F1-score, and ROC-AUC score, indicates that it can accurately predict customer churn. Specifically:
  - Confusion Matrix: The model correctly identified 541 non-churning customers and 76 churning customers, with minimal false positives and false negatives.
  - Accuracy: The overall accuracy of 93% ensures that SyriaTel can confidently identify at-risk customers, focusing retention efforts where they are most needed.
- This high level of accuracy directly supports the effectiveness of SyriaTel's retention strategies by minimizing errors in identifying customers likely to churn.

Model Evaluation and Insights for Predicting Customer Churn at SyriaTel - Addressing SyriaTel's Key Questions (Continued)

3) Which features are most influential in predicting customer churn?

Insights from SHAP Analysis:

- **International Plan Usage**: Customers with an international plan ("international plan_yes") are more likely to churn, making this feature a critical factor in retention strategies.
- **Customer Service Interactions**: Frequent customer service calls are strong indicators of potential churn, particularly when issues remain unresolved.
- **Total Day Minutes**: Higher usage during daytime ("total day minutes") correlates with increased churn risk, suggesting that heavy users during peak hours might be less satisfied with the service.
- **Voicemail Plan**: Having a voicemail plan ("voice mail plan_yes") is associated with lower churn, indicating that bundling this service with other plans could enhance customer retention.

These insights enable SyriaTel to prioritize its retention efforts effectively. For instance, customers frequently contacting support may benefit from proactive follow-ups or personalized offers, while high-usage customers could be targeted with specialized plans that better meet their needs.

Recommendations

1. **Deploy the Random Forest Model:**
   - Integrate the Random Forest model into SyriaTel's CRM system for real-time churn prediction.
   - Regularly update the model with new data to maintain and improve its predictive accuracy.
2. **Enhance Customer Service:**
   - Implement proactive support strategies to reduce churn among customers with frequent service interactions.
   - Improve feedback mechanisms to identify and address customer pain points early.
3. **Tailor Retention Strategies Based on Feature Importance:**
   - Develop and offer specialized plans for high-usage customers to increase loyalty.
   - Implement loyalty programs with incentives for international plan users and other high-risk segments identified by the model.
4. **Monitor and Optimize:**
   - Continuously monitor the model's predictions and the effectiveness of retention campaigns.
   - Use data-driven insights to refine retention strategies and improve overall customer satisfaction.

Conclusion

By addressing these key questions with a data-driven approach and implementing the recommended strategies, SyriaTel will be well-equipped to reduce customer churn. The deployment of the Random Forest model, coupled with targeted retention strategies based on the most influential features, will not only improve customer satisfaction but also enhance the company's financial performance. This proactive shift from understanding to action will enable SyriaTel to maintain a competitive edge in the telecommunications industry.

WITH THANKS

Gladys Kemunto
9/1/2024