A Data-Driven Approach to Improve Customer Retention

Project Overview:

This project aims to develop a classification model that will predict customer churn for SyriaTel, a telecommunications company. I have chosen to follow the CRISP-DM method to complete this project. It will involve six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The project purposes to provide insights into the patterns and factors influencing customer churn, and also develop a predictive model to assist in reducing customer attrition.

Business Understanding:

- SyriaTel is the major stakeholder for this project. They are interested in reducing customer churn. By helping them predict customer churn, they can take proactive measures to ensure maximum customer retentions and profit maximization. The project majorly focuses on identifying patterns that facilitate to customer churn and providing recommendations on how to mitigate this.

.Problem Definition

- Objective: Develop a predictive model to determine whether a customer will churn (binary classification: Yes/No) based on customer usage patterns, interaction with the company, and plan features.

- Outcome: Provide actionable insights to SyriaTel to reduce customer churn by identifying high-risk customers and enabling targeted retention strategies.

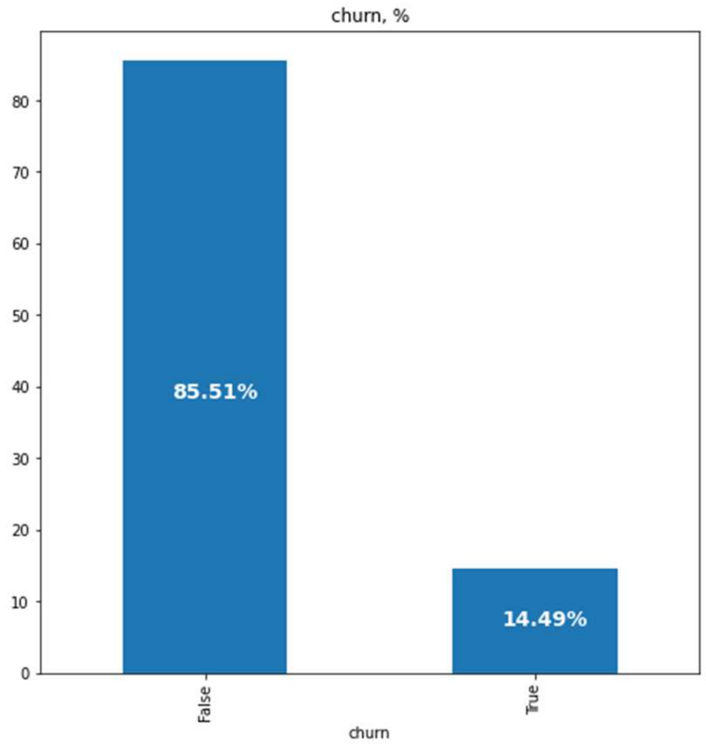
Data Understanding

- The dataset has 3333 rows and 21 columns and has no null values or duplicates. Therefore we do not need to impute any missing values or drop any duplicated values in this case. Among the 21 columns five of them are categorical in nature; 'state', 'phone number', 'international plan', 'voice mail plan','churn'. Churn which is our target variable in the data set is of boolean data type. Thus, we will make it binary later when building our models.

- Some of the columns based on domain knowledge are not actually good predictors and thus dropping them before fitiing into our models will be good. For example, the phone number variable has nothing to do with customer chruning the company.

- Most values in the dataset are numerical in nature. The summary statistics provides a brief overview of the dataset and the range of values observed in each numerical column.

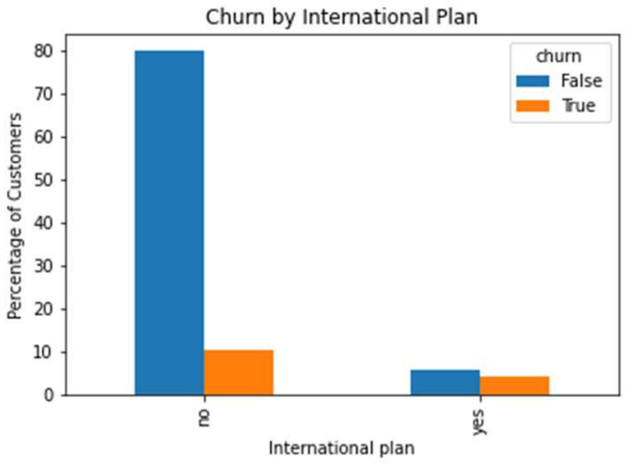
Distribution of churn



The analysis of the churn variable reveals that 85.51% of customers do not churn, while 14.49% of customers churn from the company.

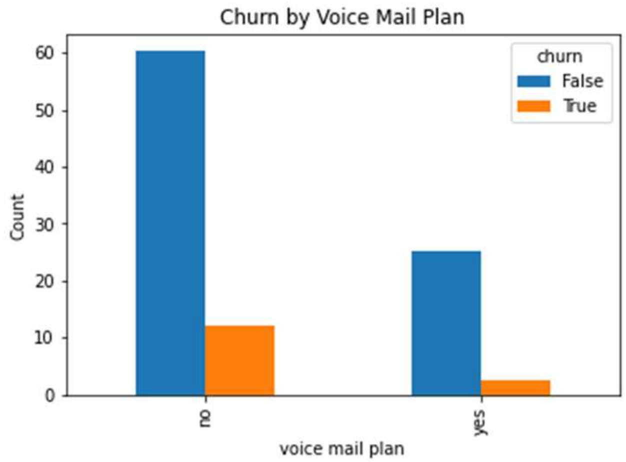
This indicates an imbalance in the distribution of the binary classes. To address this issue and prevent the model from making false predictions, we will need to apply class imbalance treatment techniques.

Churn by International Plan



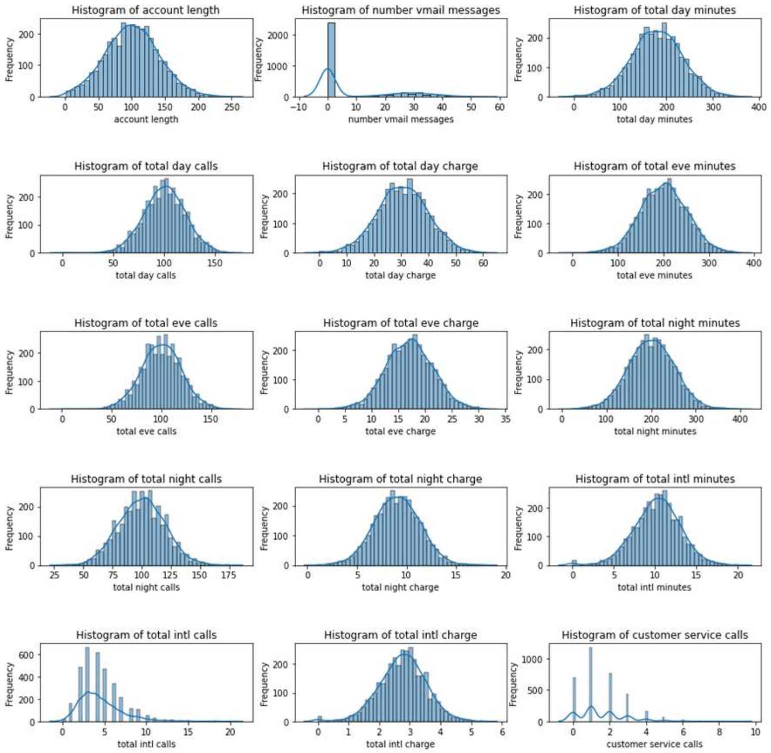
Based on the bargraph above, it is evident that customers without an international plan have a higher percentage in both the 'False' and 'True' categories compared to customers with an international plan. This suggests that having an international plan may be associated with a lower likelihood of churn.

churn by voice mail plan



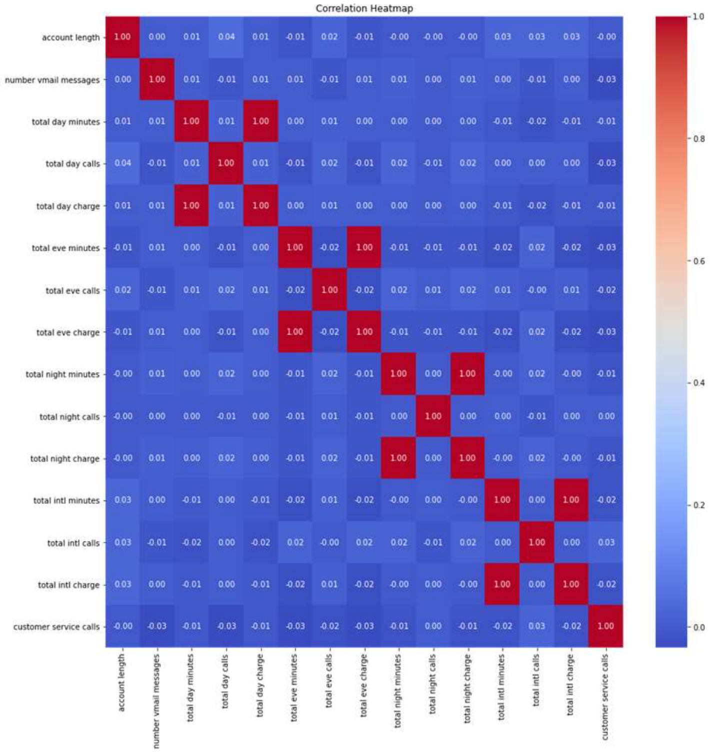
From the graph above, it can be observed that the majority of customers who do not have a voice mail plan are in the 'False' category, while a smaller proportion is in the 'True' category. In addition, customers with a voice mail plan have a higher count in the 'False' category compared to the 'True' category. This may suggest that having a voice mail plan may have some influence on reducing churn,

Distribution of numeric variables



The numerical variables exhibit diverse distributions and ranges, indicating variations in customer behavior and call patterns. While some variables follow approximately normal distributions, others display skewed distributions. This suggests that the variables may require different handling approaches based on their distributions for further analysis and modeling.

correlation matrix



Correlation Heatmap

	account length	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
account length	1.00	0.00	0.01	0.04	0.01	-0.01	0.02	-0.01	-0.00	-0.00	-0.00	0.03	0.03	0.03	-0.00
number vmail messages	0.00	1.00	0.01	-0.01	0.01	0.01	-0.01	0.01	0.01	0.00	0.01	0.00	-0.01	0.00	-0.03
total day minutes	0.01	0.01	1.00	0.01	1.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	-0.02	-0.01	-0.01
total day calls	0.04	-0.01	0.01	1.00	0.01	-0.01	0.02	-0.01	0.02	-0.01	0.02	0.00	0.00	0.00	-0.03
total day charge	0.01	0.01	1.00	0.01	1.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	-0.02	-0.01	-0.01
total eve minutes	-0.01	0.01	0.00	-0.01	0.00	1.00	-0.02	1.00	-0.01	-0.01	-0.01	-0.02	0.02	-0.02	-0.03
total eve calls	0.02	-0.01	0.01	0.02	0.01	-0.02	1.00	-0.02	0.02	0.01	0.02	0.01	-0.00	0.01	-0.02
total eve charge	-0.01	0.01	0.00	-0.01	0.00	1.00	-0.02	1.00	-0.01	-0.01	-0.01	-0.02	0.02	-0.02	-0.03
total night minutes	-0.00	0.01	0.00	0.02	0.00	-0.01	0.02	-0.01	1.00	0.00	1.00	-0.00	0.02	-0.00	-0.01
total night calls	-0.00	0.00	0.00	-0.01	0.00	-0.01	0.01	-0.01	0.00	1.00	0.00	0.00	-0.01	0.00	0.00
total night charge	-0.00	0.01	0.00	0.02	0.00	-0.01	0.02	-0.01	1.00	0.00	1.00	-0.00	0.02	-0.00	-0.01
total intl minutes	0.03	0.00	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.00	0.00	-0.00	1.00	0.00	1.00	-0.02
total intl calls	0.03	-0.01	-0.02	0.00	-0.02	0.02	-0.00	0.02	0.02	-0.01	0.02	0.00	1.00	0.00	0.03
total intl charge	0.03	0.00	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.00	0.00	-0.00	1.00	0.00	1.00	-0.02
customer service calls	-0.00	-0.03	-0.01	-0.03	-0.01	-0.03	-0.02	-0.03	-0.01	0.00	-0.01	-0.02	0.03	-0.02	1.00

From the above correlation matrix, we can observe that most of the variables are not strongly correlated. However, there are some variables that exhibit a perfect correlation. This makes sense since some variables are directly correlated.

Model 1.Logistic regression

- **************** LOGISTIC REGRESSION CLASSIFIER MODEL RESULTS ****************
- Accuracy: 0.79857
- Precision: 0.32984
- Recall: 0.82895
- F1 Score: 0.47191

- The Logistics Regression model performance metrics are as follows:

- Accuracy (0.79857): The model accurately predicted 79.86% of all instances. Precision (0.32984): Of the cases predicted as "Churn," 32.98% were correct. Recall (0.82895): The model successfully identified only 82.90% of the actual "Churn" cases. F1 Score (0.47191):
-
-
- Accuracy 80% of all predictions are correct — but this can be misleading with imbalanced classes.

- Precision 33% of customers predicted as churners actually churned (many false positives).

- Recall 83% of actual churners were correctly identified — very good recall.

- F1 Score 0.47 — a balance between precision and recall; still moderate overall performance.

model 2 .Decision Tree

- Best Parameters: {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 10}
- precision recall f1-score support

- No Churn 0.97 0.88 0.92 624
- Churn 0.43 0.75 0.54 76

- accuracy 0.86 700
- macro avg 0.70 0.81 0.73 700
- weighted avg 0.91 0.86 0.88 700

- Precision: Of the predicted class instances, how many were correct.

- For Churn, 43% of the predicted churns were actual churns.

- Recall: Of the actual class instances, how many were captured.

- For Churn, the model caught 75% of actual churn cases.

- F1-score: Harmonic mean of precision and recall (balance between them).

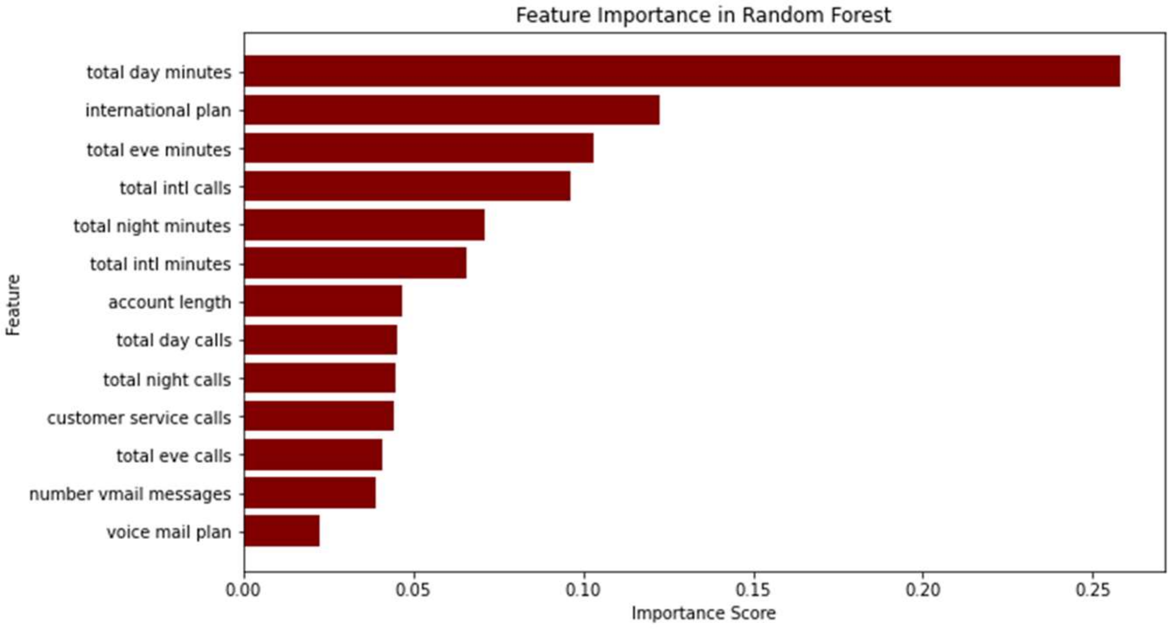
model 3.RANDOM FOREST

- Best Hyperparameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
- **************** TUNED RANDOM FOREST CLASSIFIER RESULTS ****************
- Accuracy: 0.94429
- Precision: 0.76056
- Recall: 0.71053
- F1 Score: 0.73469

- The tuned Random Forest model achieves excellent performance across all metrics, significantly improving the detection of "Churn" compared to earlier models like Logistic Regression. It effectively balances precision and recall, making it reliable for applications where identifying churners accurately is critical for business strategy. The chosen hyperparameters likely enhanced the model's ability to generalize and capture the complexities of the data.

- Hyperparameter tuning marginally improved overall performance, with a higher accuracy, precision, and F1 score compared to the baseline. While recall slightly decreased, the improvement in precision ensures that the tuned model is more reliable and consistent in its predictions. This makes the tuned Random Forest classifier a more robust choice, especially in scenarios prioritizing reduced false positives without sacrificing much recall.

feature importance on churn



- The feature importance scores indicate how much each feature contributes to the Random Forest model's predictions. Here's an explanation of the results:

- # Top Contributors:

1.total day minutes (0.265): This feature has the highest importance, meaning the total minutes a customer spends on daytime calls is the most critical factor in predicting churn.

2.international plan (0.150): Whether a customer has subscribed to an international plan is the second most influential factor, reflecting its impact on churn decisions.

3. total intl calls (0.101): The total number of international calls made is another significant factor, showing its relevance in customer churn behavior.

- # Less Significant Features:

4. Call and account-related features like total day calls (0.048), account length (0.047), and total eve calls (0.043) have lower importance, suggesting they are less predictive of churn compared to the top features.

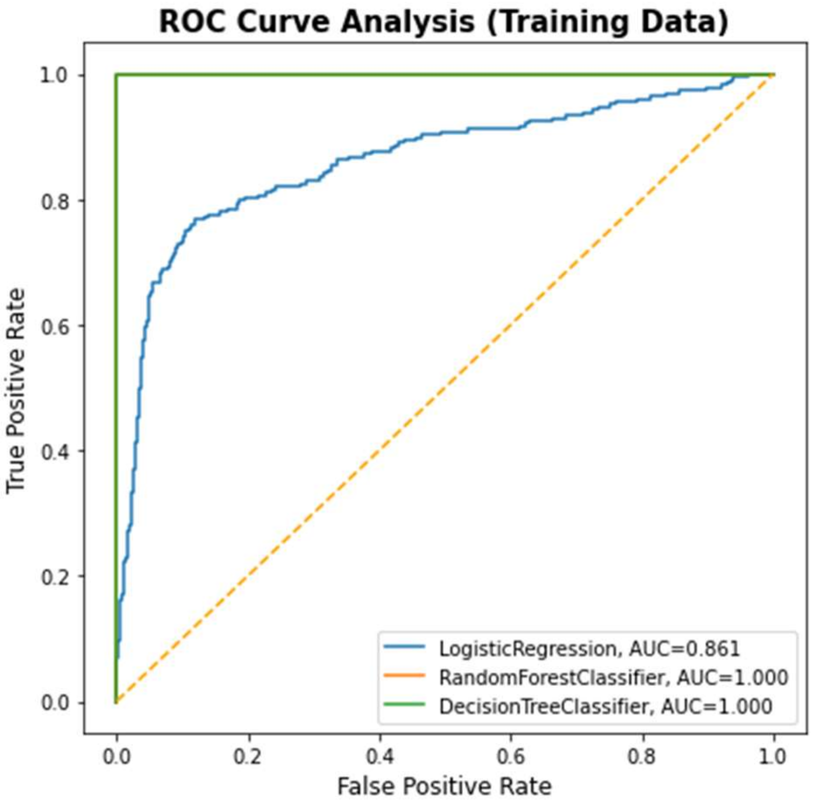
5.customer service calls (0.035): While low, this feature still has some influence, as frequent interactions with customer service might be a signal of dissatisfaction.

6.voice mail plan (0.035): This feature has minimal impact, indicating it is not a major factor in predicting churn.

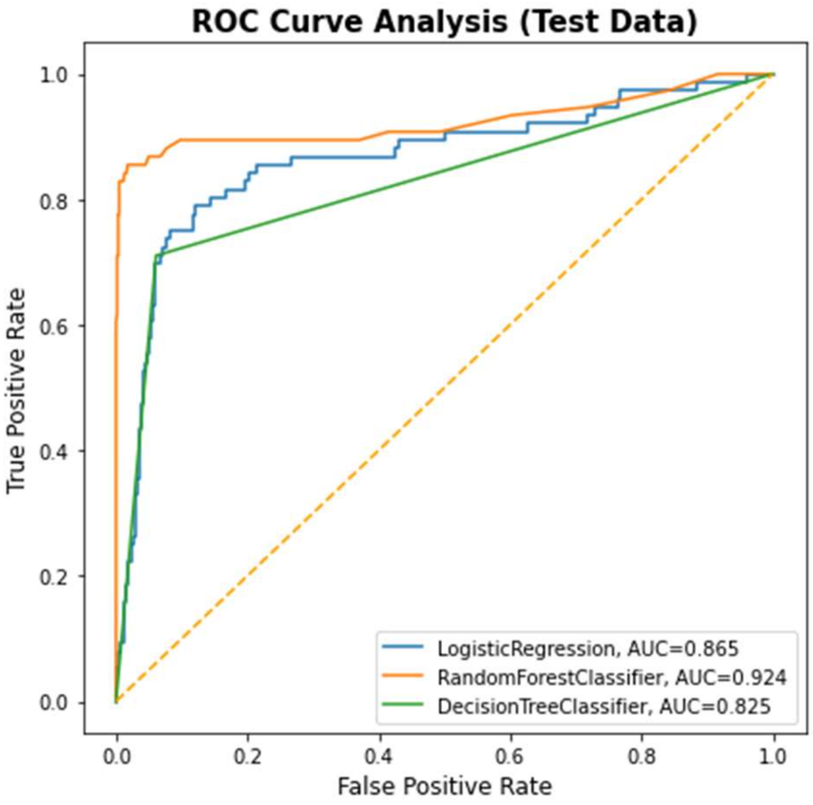
Summary

- The model emphasizes call usage patterns (minutes and international calls) and subscription plans (international plan) as the primary predictors of churn. Features like account length, voice mail plan, and customer service calls have relatively less influence. These insights could guide strategies for churn reduction by focusing on optimizing services related to the most critical features.

Modelling



ROC Curve Analysis (Training Data)



ROC Curve Analysis (Test Data)

COMPARISON

- *************** MODEL COMPARISON RESULTS ***************
- Training Data:
- classifiers auc accuracy
- 0 LogisticRegression 0.86147 0.904149
- 1 RandomForestClassifier 1.00000 1.000000
- 2 DecisionTreeClassifier 1.00000 1.000000

- Best Model on Training Data: RandomForestClassifier (AUC: 1.000, Accuracy: 1.000)

- Test Data:
- classifiers auc accuracy
- 0 LogisticRegression 0.864604 0.898571
- 1 RandomForestClassifier 0.921084 0.961429
- 2 DecisionTreeClassifier 0.810855 0.910000

- Best Model on Test Data: RandomForestClassifier (AUC: 0.921, Accuracy: 0.961)

Final model

- Of the three models (Logistic Regression, Random Forest, and Decision Tree) based on their AUC and accuracy scores for both training and test data we can conclude as follows;

- Training Data Results: Random Forest and Decision Tree models have perfect accuracy and AUC scores of 1.000, suggesting they fit the training data perfectly. However, this could indicate overfitting, as these models may have memorized the training data rather than generalizing well.
- Test Data Results: When evaluated on the test data, the Random Forest classifier stands out with the highest AUC (0.9179) and accuracy (96.28%). It outperforms the other two models, indicating better generalization and performance on unseen data. The Decision Tree model has a relatively high accuracy but a lower AUC, suggesting it may not handle the complexity of the data as well as Random Forest. Logistic Regression also has a lower AUC and accuracy compared to Random Forest.
- Conclusion: The Random Forest classifier is the best model to use, as it achieves the highest accuracy and AUC on both training and test data, with the best generalization capability to unseen data.