Nguyen Thanh Dat
2132300562
MIS 395

Final Report

Table of Contents

Github Link

Question 1:

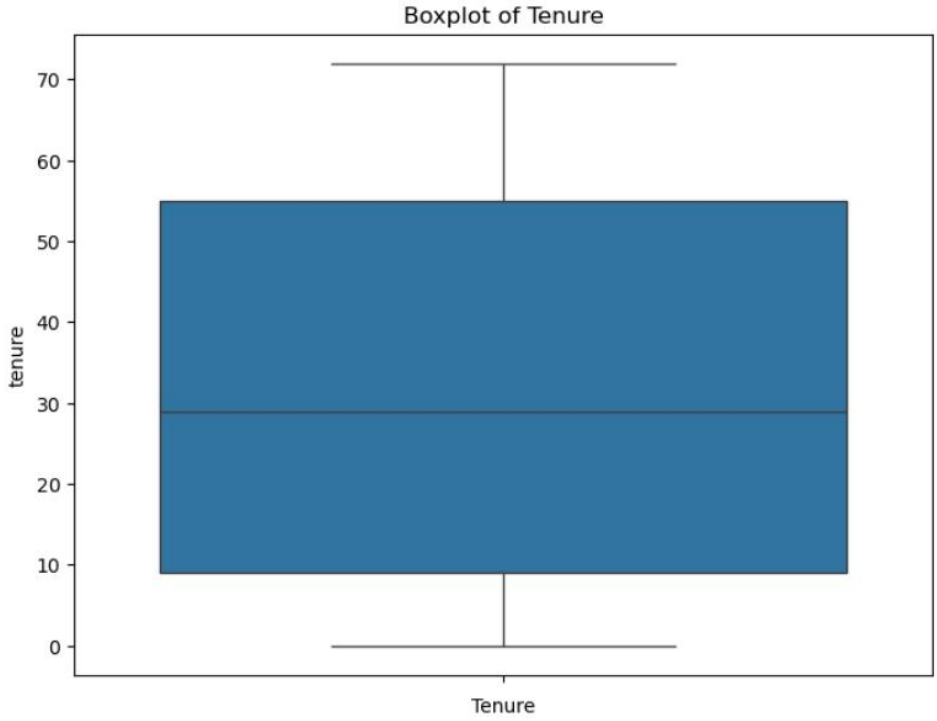
The business problem that can be solved using this dataset is customer churn prediction. By analyzing customer demographics, service usage, and subscription details, the model can predict which customers are likely to leave the service (churn) or stay. This helps businesses identify at-risk customers, optimize retention strategies, and improve customer satisfaction. Reducing churn is crucial for subscription-based businesses, as retaining existing customers is more cost-effective than acquiring new ones, ultimately leading to increased customer lifetime value and sustained revenue growth.

For this problem, classification is the appropriate model type because the target variable, Churn, is a categorical binary outcome, where customers either churn (Yes) or do not churn (No). In classification, the goal is to predict a category or class label based on input features.

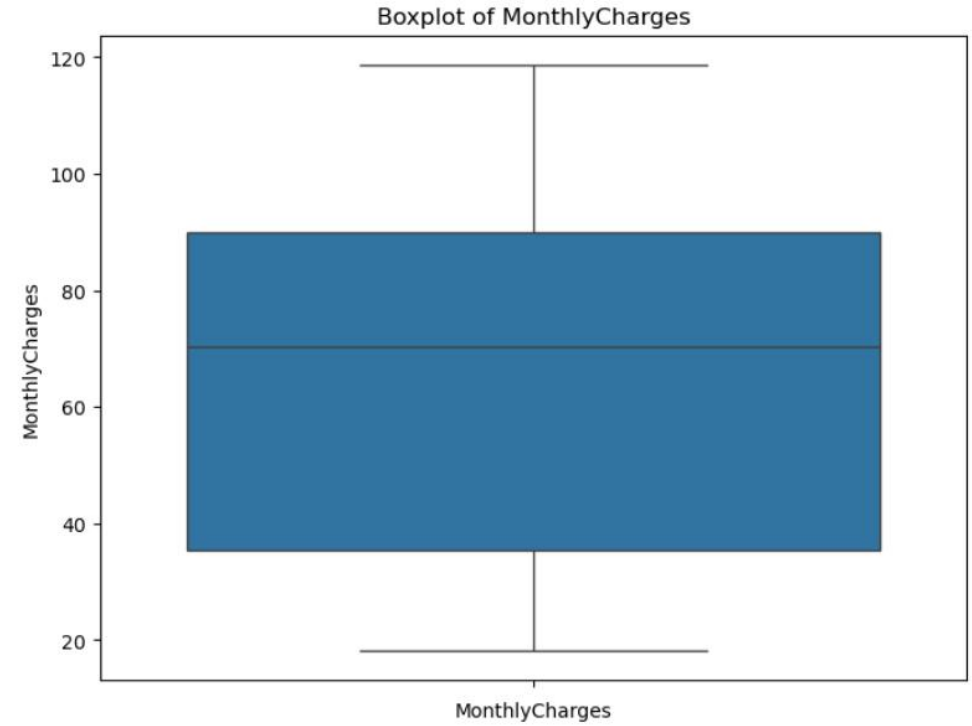
Question 2:

First, I checked for missing values to see if any columns had missing data. The results showed that two columns, PaymentMethod and MultipleLines, had missing values. I filled them using the mode method. However, during further exploration, I noticed that the TotalCharges column also had 11 missing values represented as 'No'. I filled these missing values using the mean method. Second, I found that some columns had incorrect data types. For example, SeniorCitizen was in int64, and TotalCharges was in object format. I converted these columns to the correct data types. Third, I used get_dummies to convert categorical variables into binary values (True/False) to make it easier to run classification models later on. This process ensured that the dataset was clean and ready for model training.

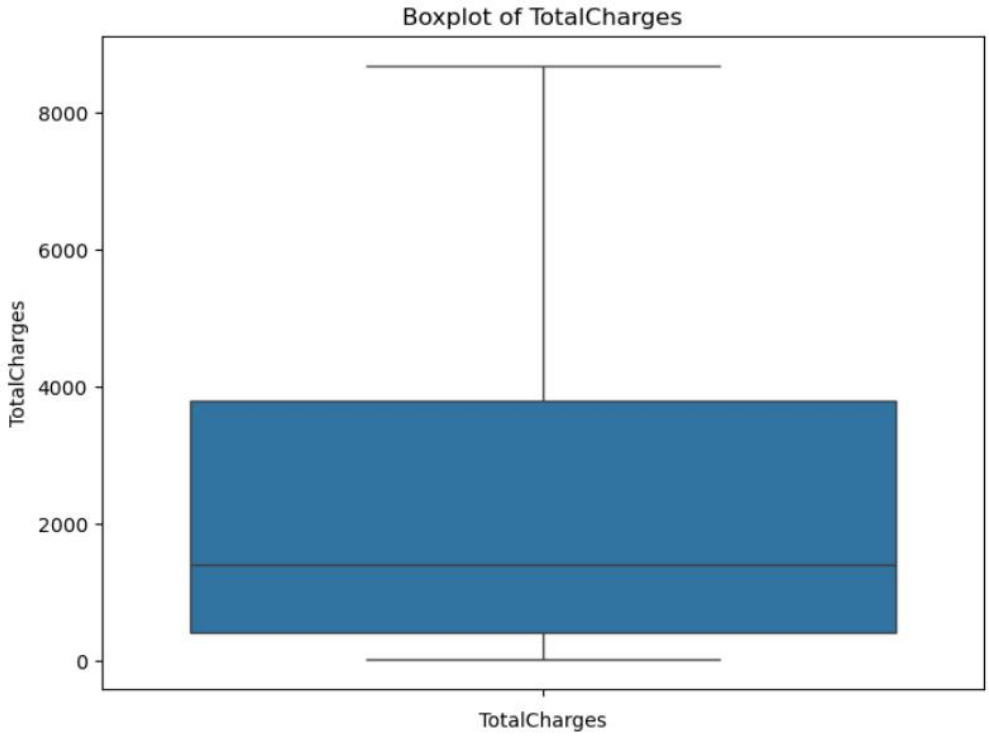
Question 3:



The boxplot of Tenure reveals that the median tenure of customers is approximately 30 months, indicating that the typical customer has been with the service for about 2.5 years. The interquartile range (IQR) spans from 10 to 50 months, showing that most customers have been with the service either for a short duration (less than a year) or a long time (over 2 years), with fewer customers in the middle range. The boxplot is relatively symmetric, suggesting a balanced distribution of tenure without clear skewness. There are no extreme outliers, and the whiskers extend up to 70 months, indicating a small number of customers with much longer tenures. Overall, the data suggests that the majority of customers have moderate tenures, with few outliers at the higher end.

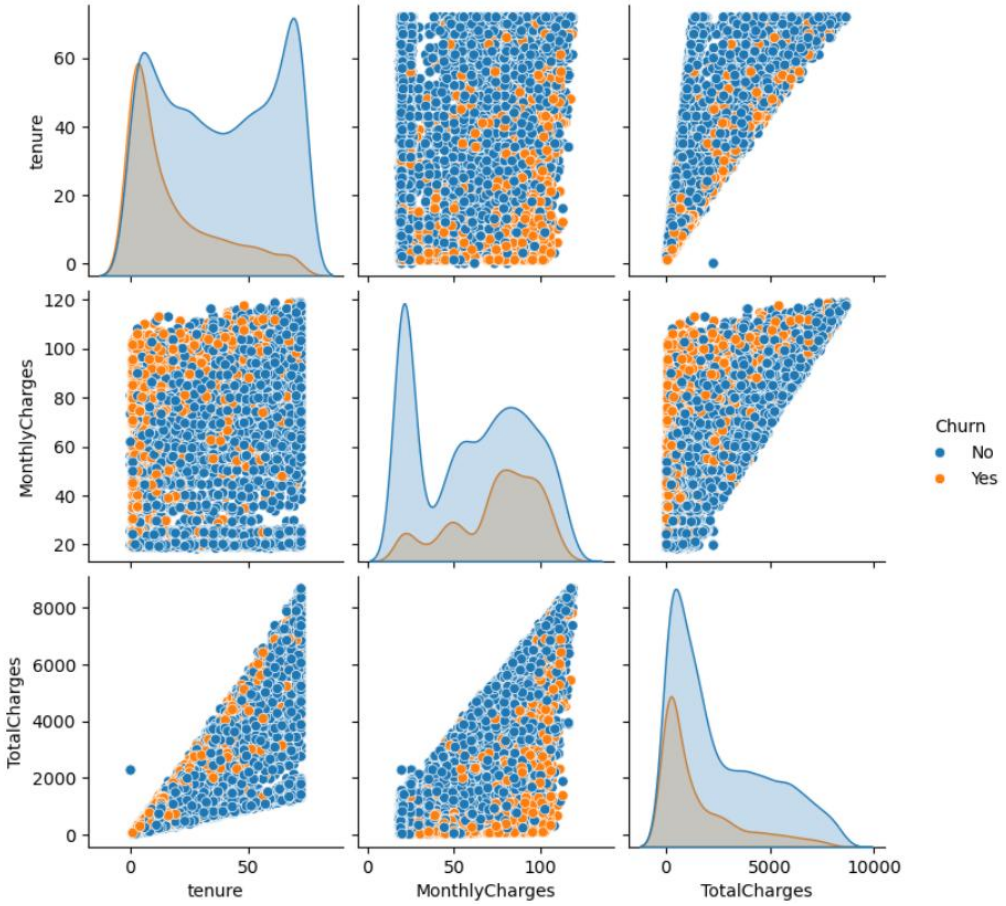


The boxplot of MonthlyCharges shows that the median monthly charge for customers is around $75, with most customers having charges between $60 and $90. The interquartile range (IQR) is fairly narrow, indicating that most customers' monthly charges are concentrated within this range. The whiskers extend from around $40 to $110, suggesting that there are some customers with significantly lower or higher monthly charges. However, there are no clear outliers in the data, as the whiskers do not extend beyond extreme values. This distribution indicates that the charges are relatively consistent across the majority of customers, with only a few customers having much lower or higher charges.



The boxplot of TotalCharges indicates that the majority of customers have cumulative charges between $1,500 and $2,500, with the median around $2,000. The interquartile range (IQR) is fairly compact, suggesting that most customers' total charges are within a narrow range. The whiskers extend from $0 to

approximately $4,500, showing that there are a few customers with much lower or higher total charges. However, no extreme outliers are visible beyond the whiskers, indicating that the distribution is relatively normal, with a few high-end exceptions. Most customers tend to have moderate total charges, with the distribution being more concentrated in the middle range.



- Tenure: Customers who churn (orange dots) are generally those with shorter tenure, as observed in the upper-left part of the plot. The majority of non-churning customers (blue dots) tend to have a longer tenure, with many having more than 30 months with the service. The distribution of tenure also shows a bimodal pattern, indicating two distinct customer groups.
- MonthlyCharges: Churned customers appear to have higher MonthlyCharges than non-churning customers on average, especially in the scatter plot and density plot on the top-right. However, the distribution is more spread out for both groups. This suggests that while higher monthly charges are associated with churn, it's not the sole factor.
- TotalCharges: There is a clear positive correlation between TotalCharges and both MonthlyCharges and tenure, as seen in the lower-left scatter plot. Customers with higher TotalCharges are generally those with longer tenures and higher MonthlyCharges. The churned customers (orange dots) have lower TotalCharges, reinforcing the idea that customers who leave the service tend to have shorter relationships and pay lower overall charges.

Question 4:

1. What type of model did you build?

I built a classification model to predict whether a customer will churn (leave the service) or not churn. Since the target variable Churn is binary (Yes or No), classification is the appropriate model type.

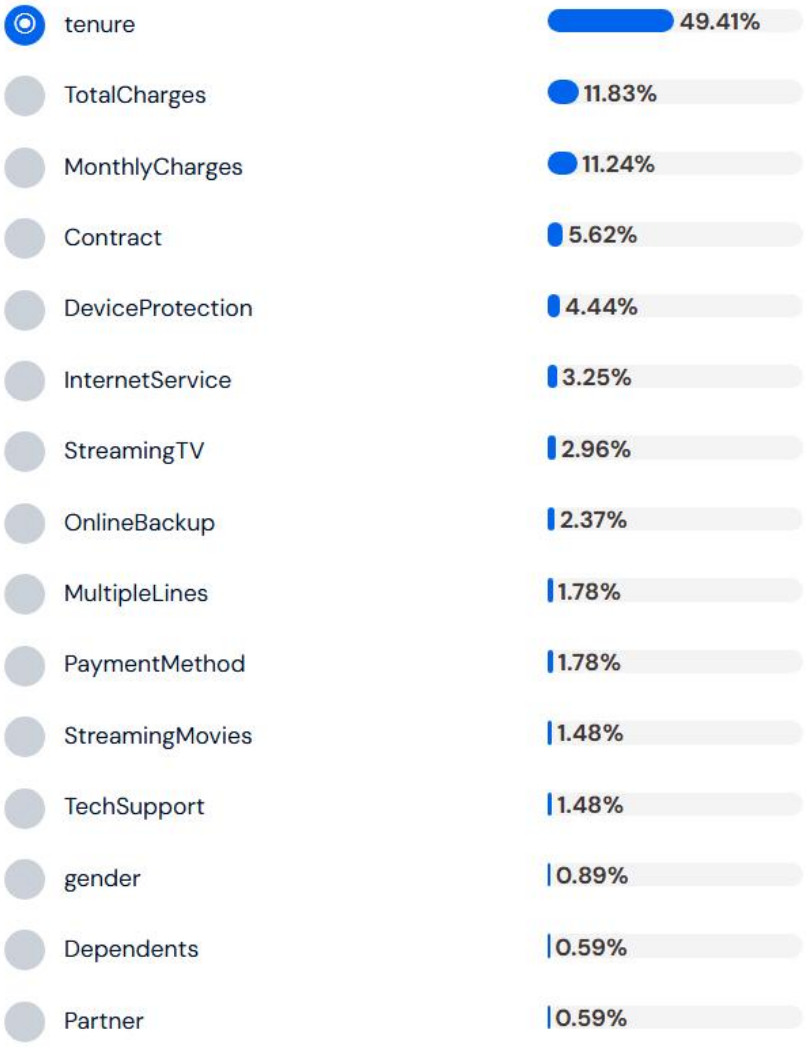
2. What algorithm did you use?

I used the Random Forest algorithm, which is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and avoid overfitting. It's particularly useful for handling a mix of numerical and categorical data, as well as for its robustness in feature selection.

3. Which features did you include in the model, and why?

I included all features except for customerID, which was excluded because it is a unique identifier and doesn't provide predictive power. The other features such as tenure, TotalCharges, MonthlyCharges, Contract, and DeviceProtection were included because they provide valuable insights into customer behavior and directly correlate with the likelihood of a customer churning. The features cover both demographic information (like gender and Partner) and service-related data (like InternetService and StreamingTV), making them critical for building an effective model.

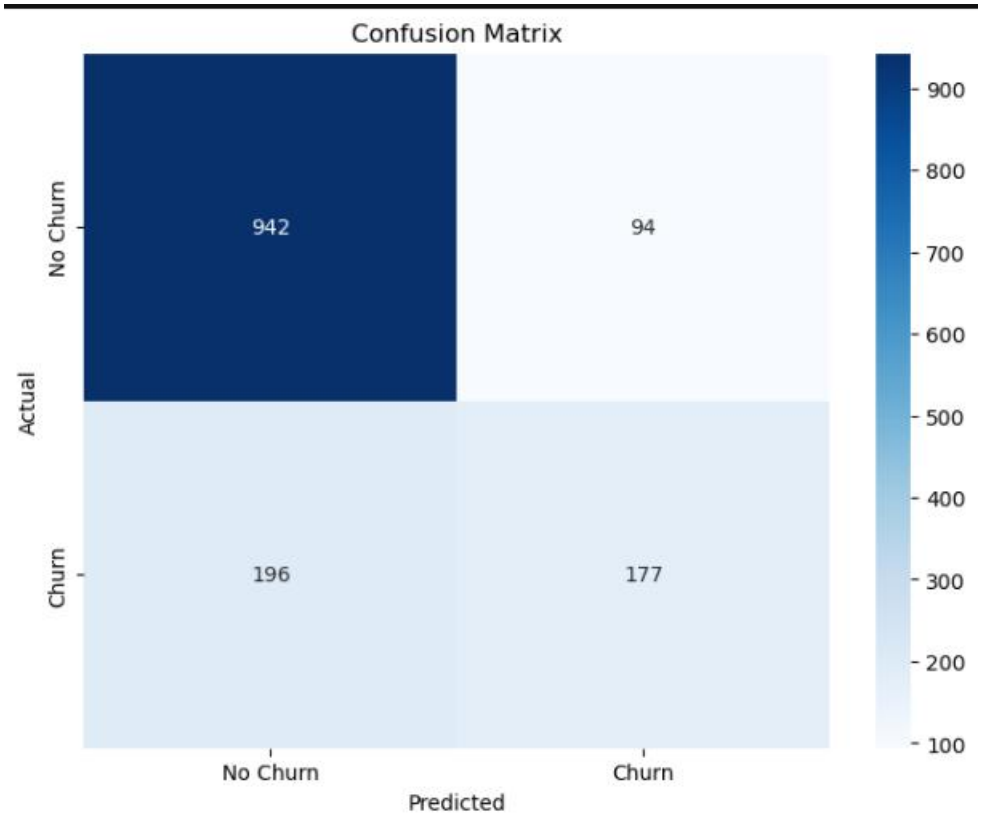
4. What are key predictions or patterns your model reveals?



Based on the importance of features, the key prediction revealed by the model is that tenure is the most significant factor in predicting churn, accounting for 49.41% of the model's predictive power. This suggests that customers who have been with the service for a shorter period are more likely to churn. Other important features include TotalCharges and MonthlyCharges, which contribute 11.83% and 11.24% to the prediction, respectively. These findings indicate that customers with higher charges or who have been with the service for a longer time are less likely to churn. The least important feature for prediction is PhoneService, contributing only 0.59% to the model's performance. This pattern highlights the importance of customer tenure and service charges in predicting churn.

Question 5:

RandomForest Classifier:

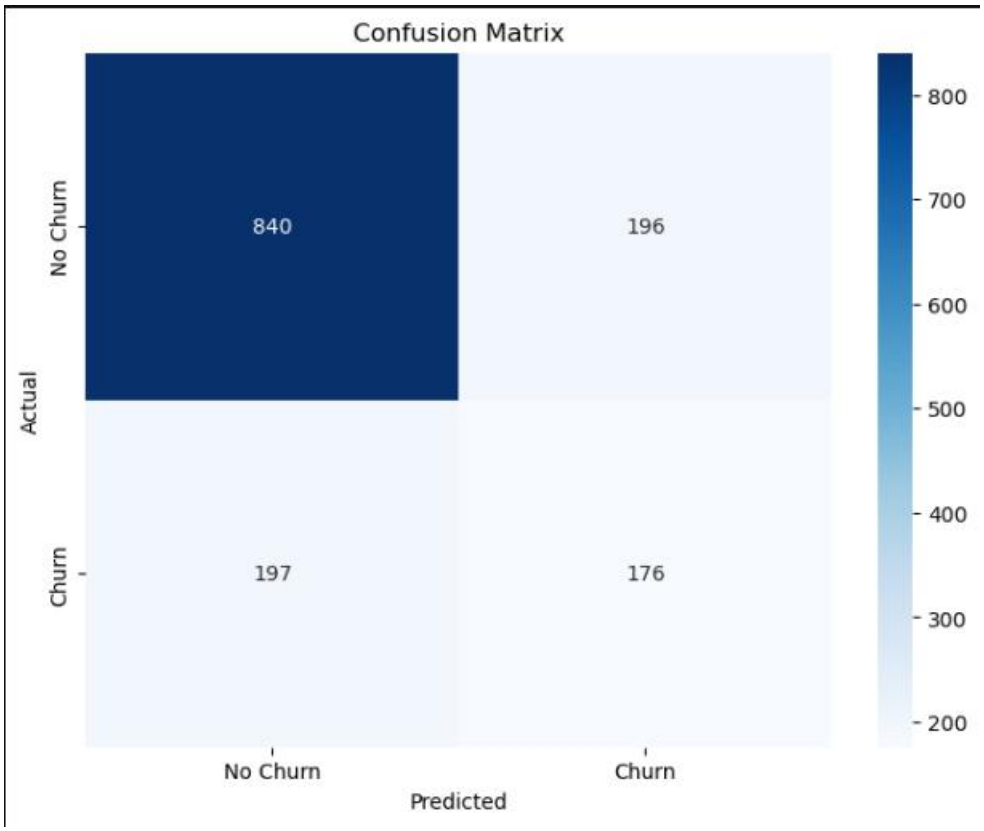


	precision	recall	f1-score	support
No	0.83	0.91	0.87	1036
Yes	0.65	0.47	0.55	373
accuracy			0.79	1409
macro avg	0.74	0.69	0.71	1409
weighted avg	0.78	0.79	0.78	1409

The confusion matrix and classification metrics provide a comprehensive evaluation of the Random Forest model for predicting churn. The model performs well in identifying customers who do not churn, with True Negatives (942) and a high Recall of 0.91, meaning it correctly identified 91% of customers who stayed. However, the model struggles with predicting customers who churn, as indicated by the False Positives (94) and a low Recall for Churn of 0.47, meaning it missed more than half of the actual churned customers. The Precision for No Churn is 0.83, showing the model is accurate when predicting customers who stay, while the Precision for Churn is 0.65, indicating that when the model predicts churn, it is correct 65% of the time. Overall, the model has an accuracy of 0.79, meaning it correctly predicts the outcome 79% of the time. The F1-Score for No Churn is strong at 0.87, but the F1-Score for Churn is much lower at 0.55. This suggests

that the model has a bias towards predicting No Churn and requires further improvement to better predict customer churn.

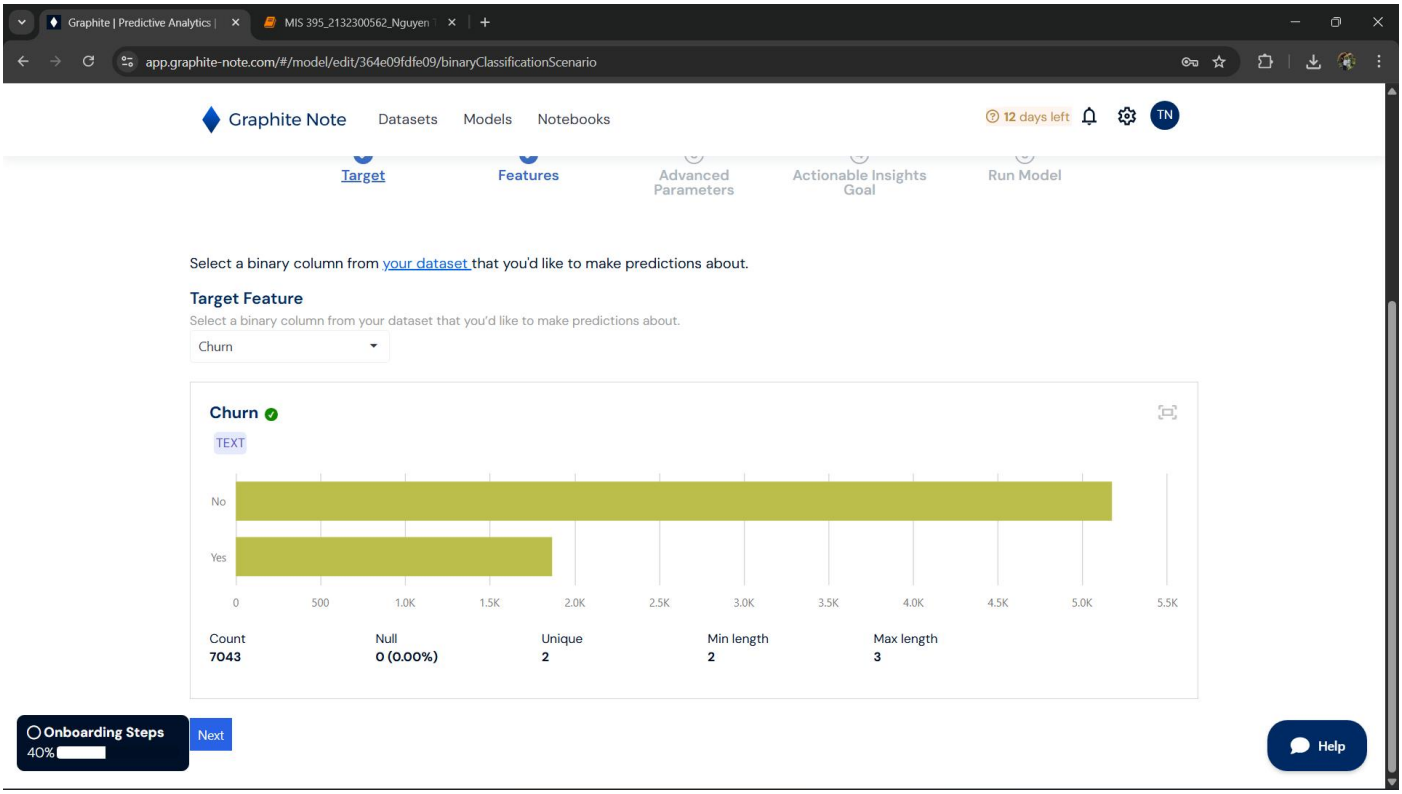
Decision Tree:

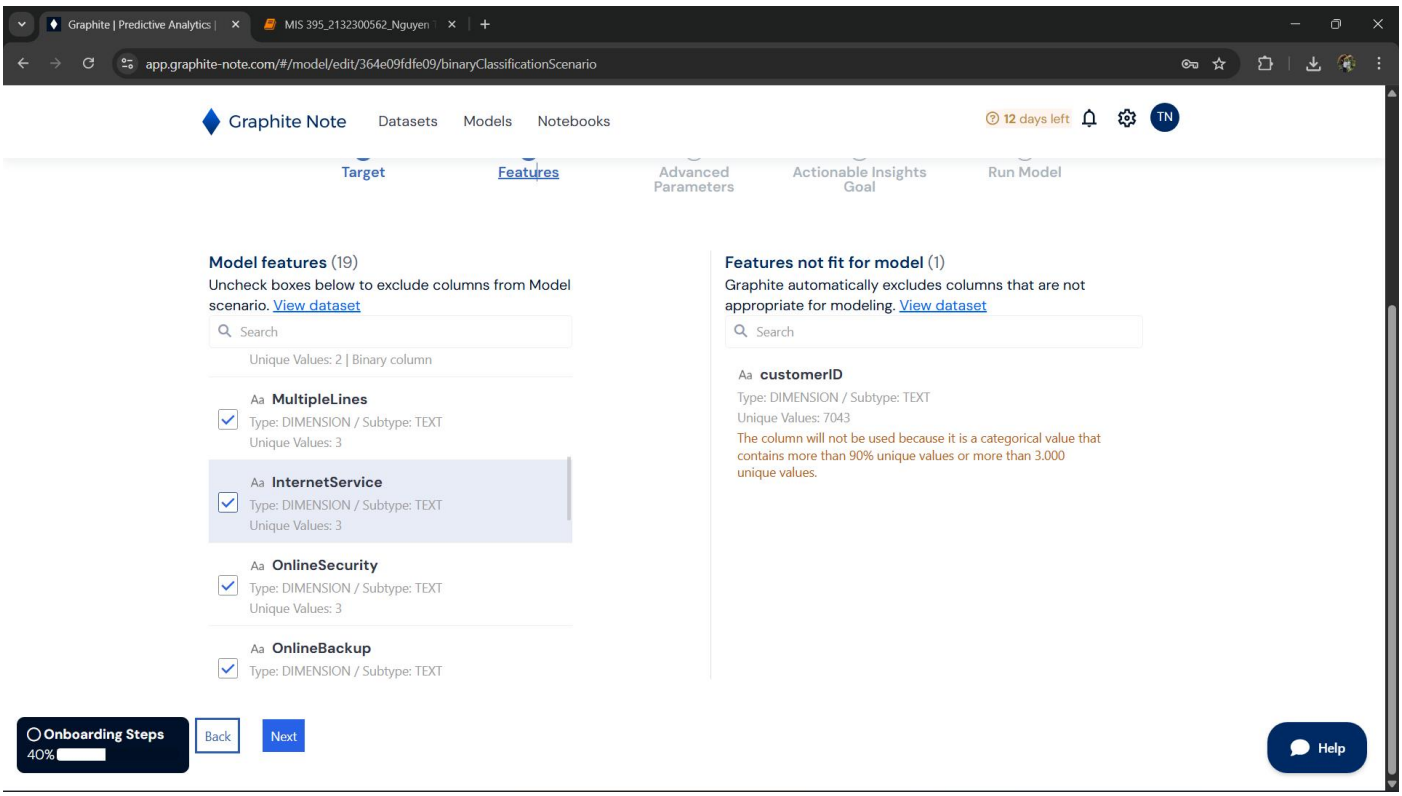


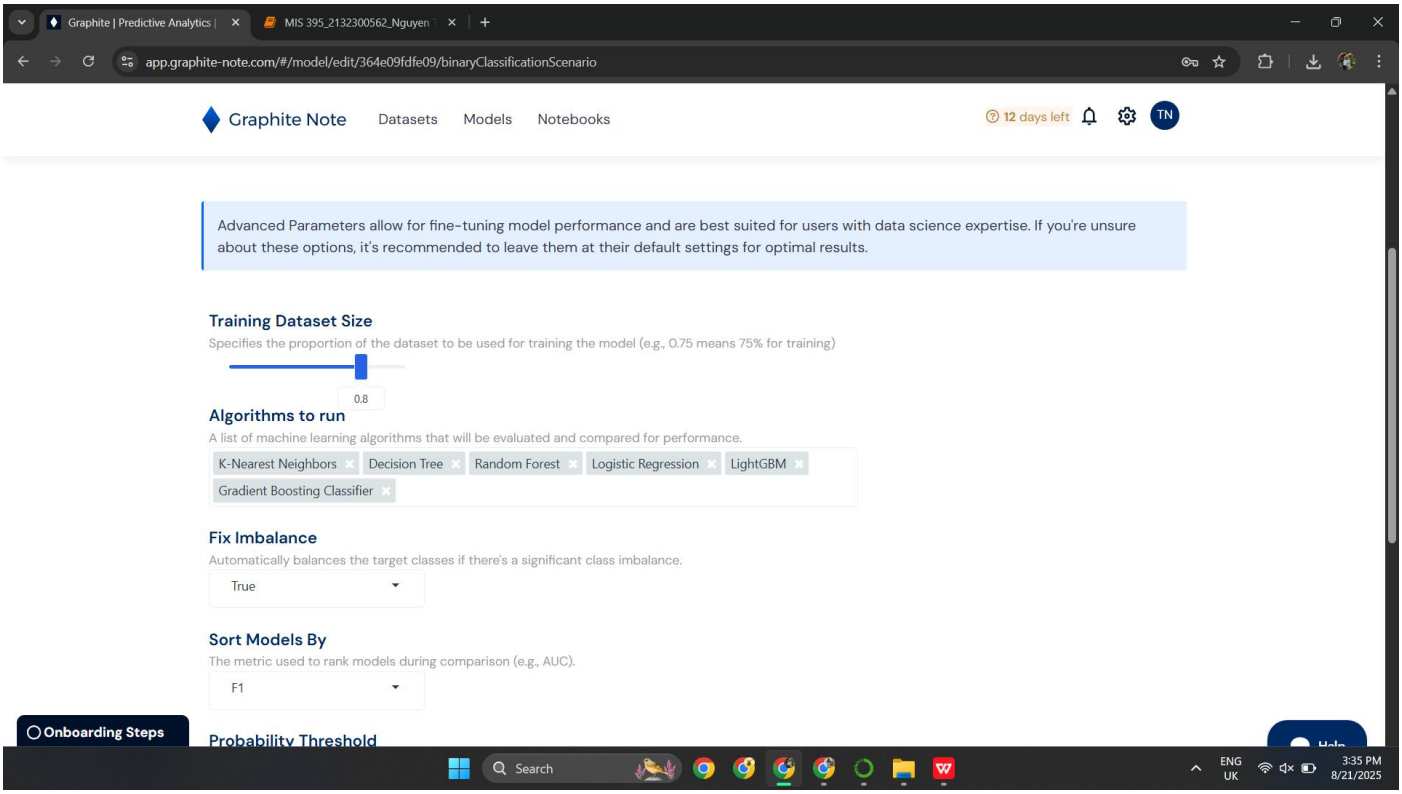
	precision	recall	f1-score	support
No	0.81	0.81	0.81	1036
Yes	0.47	0.47	0.47	373
accuracy			0.72	1409
macro avg	0.64	0.64	0.64	1409
weighted avg	0.72	0.72	0.72	1409

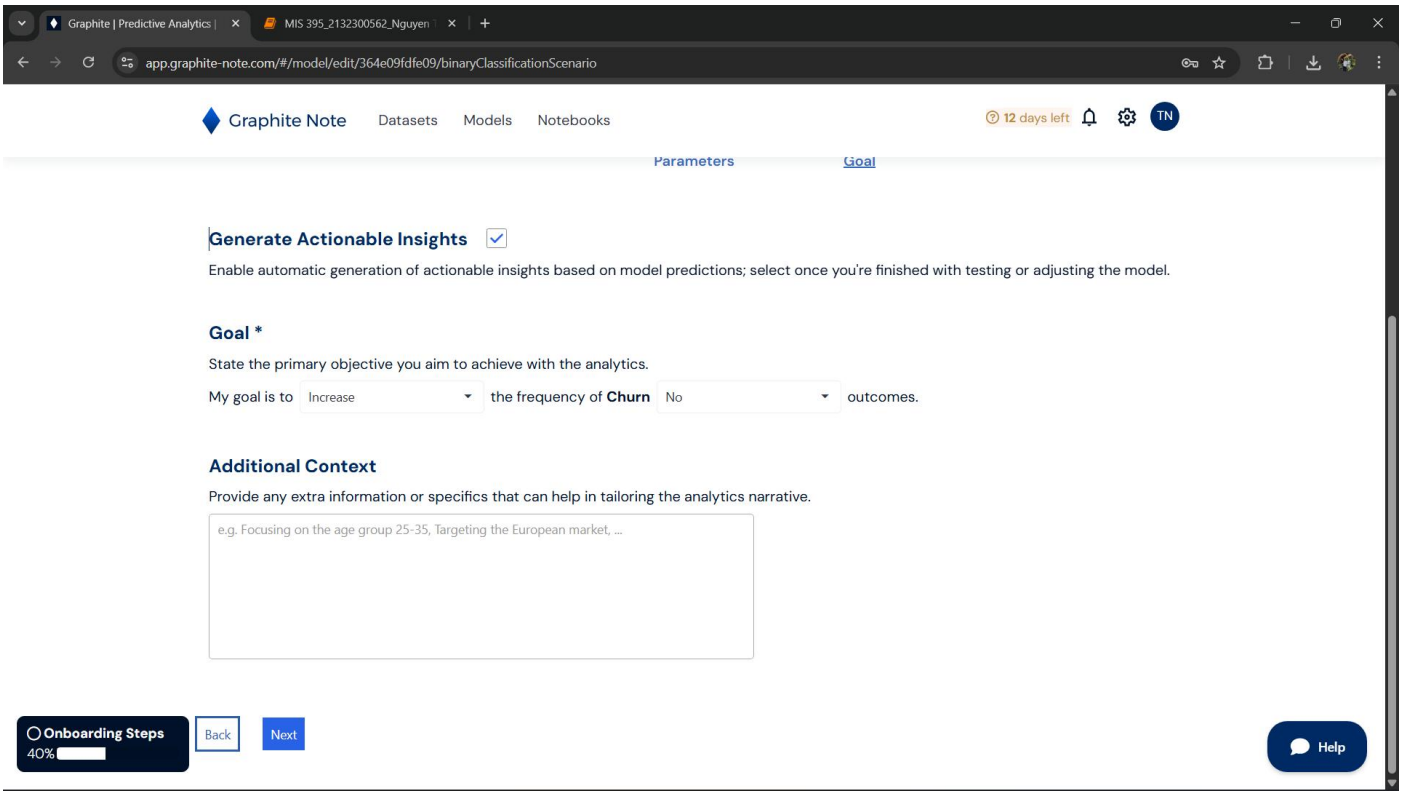
The confusion matrix and classification metrics provide a comprehensive evaluation of the Decision Tree model for predicting churn. The model performs well in identifying customers who do not churn, with True Negatives (840) and a high Recall of 0.81, meaning it correctly identified 81% of customers who stayed. However, the model struggles with predicting customers who churn, as shown by the False Positives (196) and a low Recall for Churn of 0.47, meaning it missed more than half of the actual churned customers. The Precision for No Churn is 0.81, showing the model is accurate when predicting customers who stay, while the Precision for Churn is 0.47, indicating that when the model predicts churn, it is correct 47% of the time. The F1-Score for No Churn is 0.81, but the F1-Score for Churn is much lower at 0.47. Overall, the model has an accuracy of 0.72, meaning it correctly predicts the outcome 72% of the time. The Macro Average is 0.64, and the Weighted Average is 0.72, indicating that the model's performance for Churn needs improvement. This suggests the model may be biased towards predicting No Churn and requires further improvement to better predict customer churn.

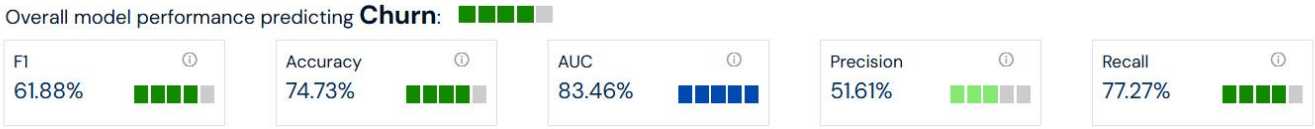
Graphite Note:





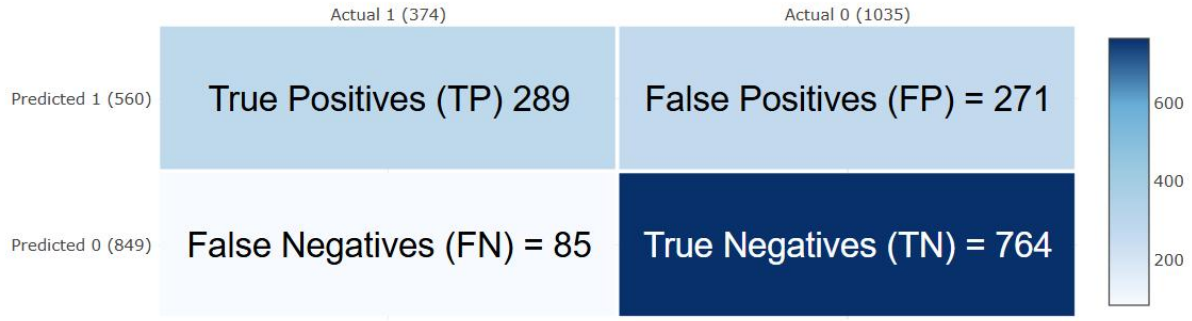




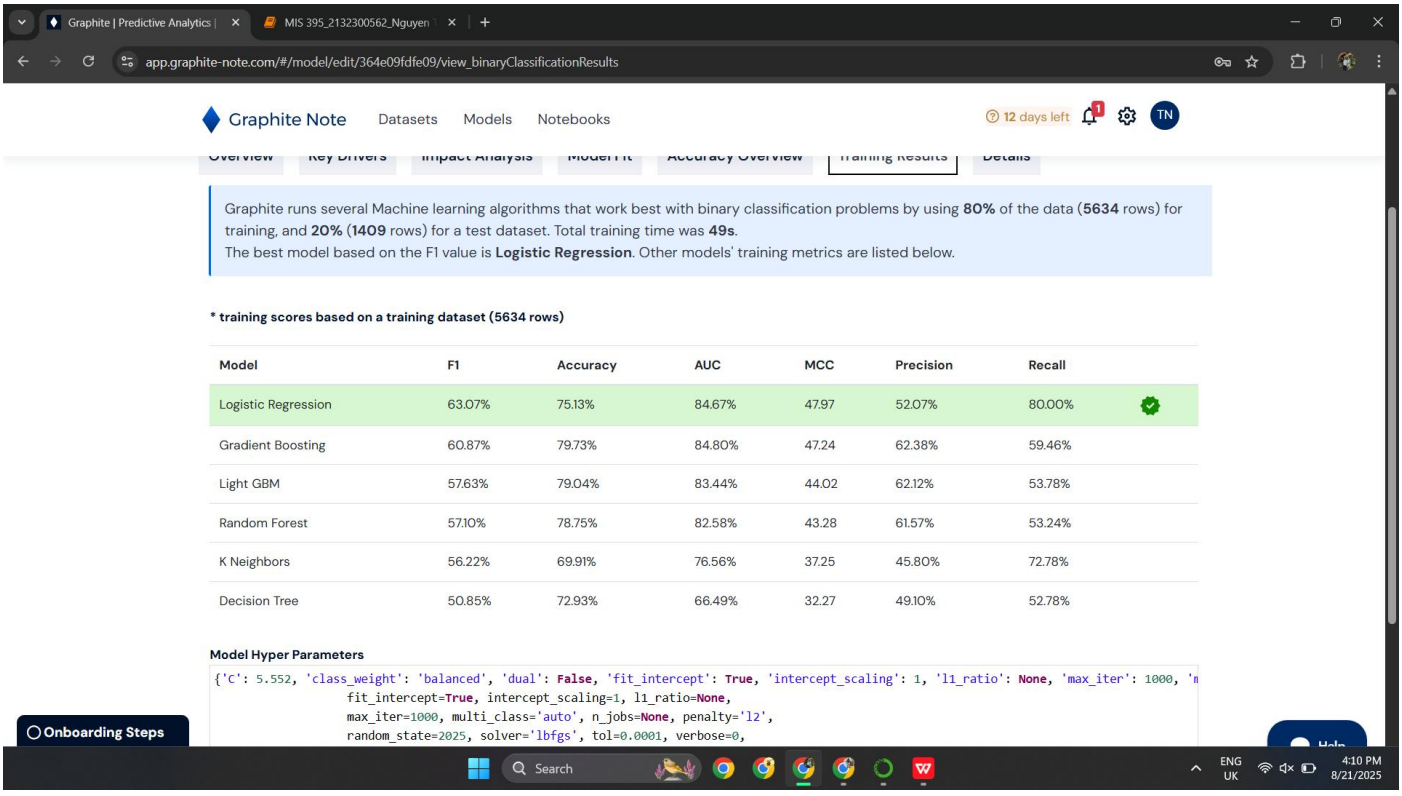


The model's performance in predicting Churn shows a F1-Score of 61.88%, indicating a moderate balance between precision and recall. While the model correctly predicts 74.73% of the instances, suggesting a reasonable accuracy, the AUC score of 83.46% reflects its ability to distinguish between churned and non-churned customers effectively. However, Precision is relatively low at 51.61%, meaning that when the

model predicts churn, it is correct only about half of the time, implying a high number of false positives. On the other hand, the model performs better in Recall with a score of 77.27%, correctly identifying 77% of the churned customers. This indicates that while the model detects a good portion of the churn, it still misses a significant number of cases. Overall, the model could benefit from further improvements in precision and tuning to better predict churn.



The confusion matrix reveals the model's performance in predicting churn. There are 289 True Positives (TP), meaning the model correctly predicted customers who churned. However, the model also made 271 False Positives (FP), incorrectly classifying non-churning customers as churned. Additionally, there are 85 False Negatives (FN), where the model missed predicting customers who actually churned. On the positive side, there are 764 True Negatives (TN), indicating that the model correctly predicted non-churning customers. While the model performs well in identifying No Churn cases, it struggles with accurately predicting Churn, as evidenced by the significant number of false positives and false negatives. This suggests the model needs improvement, particularly in detecting churned customers more accurately.



Insights for businesses:

- **Target High-Risk Customers with Retention Strategies:**

Since the model can identify customers who are likely to churn, businesses should focus on these high-risk customers with targeted retention strategies. Offering personalized incentives, discounts, or improving customer support for these individuals can help reduce churn rates. This could be particularly important for customers with high MonthlyCharges but lower tenure, as they may be more likely to leave if their needs are not addressed.

- **Optimize Customer Experience Based on Key Features:**

Based on the analysis, features like tenure, TotalCharges, and MonthlyCharges are significant drivers of churn. Businesses should prioritize improving the customer experience for long-term customers (those with high tenure) and ensure that customers who are paying higher charges feel they are receiving value for their money. Addressing customer service issues and offering better service packages could help improve retention.

- **Use Model Insights to Fine-Tune Marketing and Sales Strategies:**

The model can help businesses better understand the factors influencing churn, allowing them to adjust their marketing and sales efforts accordingly. For example, businesses can identify customers at risk of churning early and offer them tailored marketing campaigns. Additionally, understanding that certain service features (like DeviceProtection or InternetService) affect churn can help refine product offerings or upselling strategies.