E-Commerce Customer Churn Prediction

Samantha Lee

Problem Statement

Given historical data on customer interactions and purchase behavior, can we build a model to predict whether a customer will churn?

The goal was to create a binary classification model that distinguishes between customers who are likely to churn and those who are likely to remain active.

Who might care?

What factors might affect E-commerce?

Convenience and Flexibility: E-commerce offers consumers the convenience of shopping from anywhere, at any time, without the constraints of physical store hours or locations. This flexibility appeals to consumers' busy lifestyles and desire for immediate access to products and services.

Product Variety and Discovery: Online marketplaces provide a vast selection of products from around the world, often at competitive prices. Consumers can easily compare products, read reviews, and discover new items that may not be available in local stores.

Personalized Shopping Experiences: E-commerce platforms can collect and analyze user data to personalize shopping experiences, recommending products based on past purchases, browsing history, and preferences. This personalization enhances customer engagement and can lead to increased sales.

Efficient Delivery and Shipping: Reliable and timely delivery of products is crucial for customer satisfaction and repeat business. E-commerce businesses need to have efficient logistics and supply chain management systems to ensure smooth order fulfillment and fast delivery times.

E-COMMERCE SUCCESS FACTORS

Success Factors for an Ecommerce Business

Data Information

Data acquired for period: 2010-2011

This dataset comprises transactions from a UK-based non-store online retailer between January 1, 2010, and September 12, 2011. The company specializes in unique all-occasion gifts and serves a significant customer base of wholesalers.

Number of Records: 406,829

Advanced Feature Engineering

- **Prediction Horizon**:Defined a 30-day prediction horizon for churn predictions.
- **NextOrderDate Column:**Added a 'NextOrderDate' column reflecting the date of each customer's next order.
- **Churn Indicator:**Created a 'Churn' column, indicating whether a customer is likely to churn within the next 30 days based on the time difference between the current order and the subsequent one.
- **TotalPurchase Column:**Introduced a 'TotalPurchase' column representing the total purchase amount for each order.
- **Rolling Sums for Last 30 Days:**Calculated rolling sums for total order amount and quantity in the last 30 days to capture recent customer activity.
- **Additional Features:**Incorporated features like maximum order amount, cumulative maximum, and aggregated metrics to enrich the dataset.
- **Comprehensive Dataset:**The resulting 'customer_churn_data' DataFrame includes essential features such as churn status, total order amount, total order quantity, and maximum order amount for each customer.
- **Foundation for Models:**This advanced feature engineering phase lays the foundation for predictive models specifically aimed at determining customer churn.
- **Model Development and Analysis:**The enriched dataset serves as a basis for subsequent model development and analysis in the context of customer churn prediction.

	CustomerID	Churn	TotalOrderAmountLast30Days	TotalOrderQuantityLast30Days	MaxOrderAmount
0	12346	False	NaN	NaN	1.04
1	12347	True	661.079890	364.252747	12.75
2	12348	True	785.696774	1254.322581	40.00
3	12349	True	454.833288	187.986301	39.95
4	12350	True	170.817647	107.941176	40.00
...	...	...	...	...	...
4367	18280	True	99.965000	22.500000	9.95
4368	18281	True	36.810000	35.571429	16.95
4369	18282	True	95.196154	68.692308	12.75
4370	18283	True	78.791772	53.259259	15.95
4371	18287	True	598.299429	499.257143	8.50

This data snapshot includes customer transactions with purchase date, total amount, quantity ordered in the last 30 days, and maximum amount spent in a single order. The 'Churn' column indicates churn status (True/False). 'NaN' values in 'TotalOrderAmountLast30Days' and 'TotalOrderQuantityLast30Days' suggest insufficient data for some customers in the last 30 days.

Feature Selection Target Definition : Early Churn (7-Day Window)

- **Introduction of Target Variables**:Introduced two target variables: 'Days' and 'Churn1Month.'
- **Capture Customer Behavior:**Targets aim to capture customer behavior within 7-day ('Days') and 30-day ('Churn1Month') windows.
- **Essential for Predictive Models:**These target variables are crucial for training predictive models that can identify customers at risk of churn.
- **Prediction Time Frames :**'Days' and 'Churn1Month' help predict whether a customer will continue engagement within 7-day and 30-day time frames, respectively.
- **Preparation for Modeling:**Setting target variables in this way prepares the data for subsequent modeling and analysis stages.
- **Proactive Churn Identification:**Facilitates proactive identification of customers at risk of churn, enabling businesses to implement measures to retain them.
- **Enhanced Customer Relationship Management:** Contributes to an enhanced customer relationship management strategy by leveraging predictive insights for timely interventions.

	CustomerID	InvoiceDate	Churn7Days	Churn1Month
0	12346	2011-01-18 10:01:00	0	0
1	12346	2011-01-18 10:17:00	0	0
2	12347	2010-12-07 14:57:00	0	0
3	12347	2010-12-07 14:57:00	0	0
4	12347	2010-12-07 14:57:00	0	0
...	...	...	...	...
406824	18287	2011-10-12 10:23:00	0	0
406825	18287	2011-10-12 10:23:00	0	0
406826	18287	2011-10-28 09:29:00	0	0
406827	18287	2011-10-28 09:29:00	0	0
406828	18287	2011-10-28 09:29:00	0	0

[406829 rows x 4 columns]

The output DataFrame have the following columns:

'CustomerID': Identifier for each customer.
'InvoiceDate': Date and time of the customer's order.
'Churn7Days': Binary column indicating whether the customer is predicted to churn within the next 7 days (1 for churn, 0 for no churn).
'Churn1Month': Binary column indicating whether the customer is predicted to churn within the next 30 days (1 for churn, 0 for no churn).

Each row represents an order made by a customer, and the 'Churn7Days' and 'Churn1Month' columns provide early churn predictions based on the defined time windows.

Many entries in the 'Churn7Days' and 'Churn1Month' columns are 0, indicating that the customer is not predicted to churn within the respective time frames.

Feature Engineering

- In the feature engineering phase, created various measures to understand how customers behave. We looked at recent activity, like the number of orders and total spending in the last week (TotalOrdersLast1Week and TotalAmountLast1Week).
- Calculated how much time passed since a customer's last order (TimeSinceLastOrder). For recent behavior, we considered metrics for the last month (TotalOrdersLast1Month and TotalAmountLast1Month).
- Features like MaxAmountLast30Days helped identify big spenders. Other metrics, such as MaxOrderAmount, MinAmountPriorToDate, and MeanAmountPriorToDate, gave insights into historical patterns.
- We also measured customer loyalty and engagement frequency with metrics like AverageOrderDuration, CumulativeTime, and PriorOrderCount. Additional metrics, like AverageOrderAmount and TotalUniqueOrders, provided a comprehensive view of customer behavior, setting the stage for further analysis and predictions.

	CustomerID	InvoiceDate	TotalOrderAmountLast30Days	PreviousStdDevTotalOrderAmount	OrderStdDeviation
0	12346	2011-01-18 10:01:00	NaN	NaN	NaN
1	12346	2011-01-18 10:17:00	NaN	NaN	NaN
2	12347	2010-12-07 14:57:00	25.20	NaN	18.856172
3	12347	2010-12-07 14:57:00	42.20	NaN	18.856172
4	12347	2010-12-07 14:57:00	81.20	NaN	18.856172
...	...	...	...	...	...
406824	18287	2011-10-12 10:23:00	842.92	52.365471	14.206434
406825	18287	2011-10-12 10:23:00	838.12	52.783579	14.206434
406826	18287	2011-10-28 09:29:00	857.68	53.659747	14.206434
406827	18287	2011-10-28 09:29:00	867.04	54.990301	14.206434
406828	18287	2011-10-28 09:29:00	855.28	55.748495	14.206434

[406829 rows x 5 columns]

The final output DataFrame contains a variety of engineered features that provide insights into customer behavior, recency, frequency, and monetary aspects.

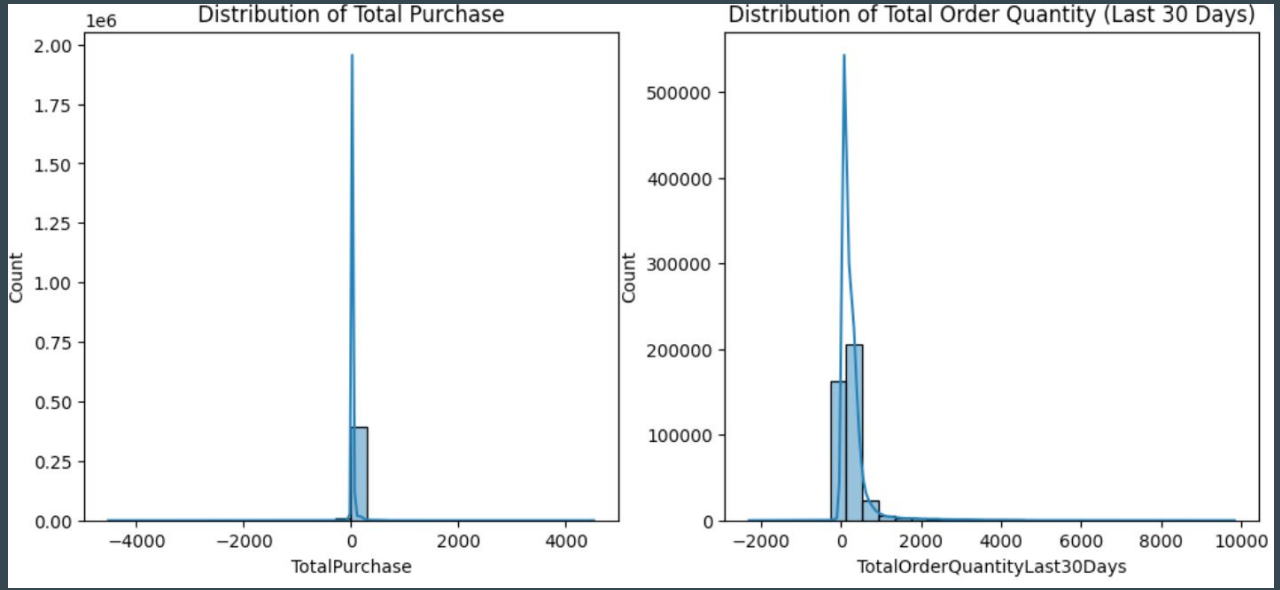
Each row represents a customer order, and the columns include the original features as well as the newly engineered features.

These features can be used for predicting customer churn, identifying high-value customers, and tailoring marketing and retention strategies. The dataset is now enriched with valuable metrics that capture the dynamics of customer spending and engagement over time.

Data Distribution: Histograms of 'TotalPurchase' and 'TotalOrderQuantityLast30Days'

Distribution of total purchase is left skewed, implies that most customers have lower total purchase amounts, but there are some customers who contribute to a longer right tail with higher purchase amounts.

Distribution of total order quantity (last 30 days) is right skewed. implies that there might be a few instances where customers placed a substantially higher number of orders in the last 30 days compared to the majority.

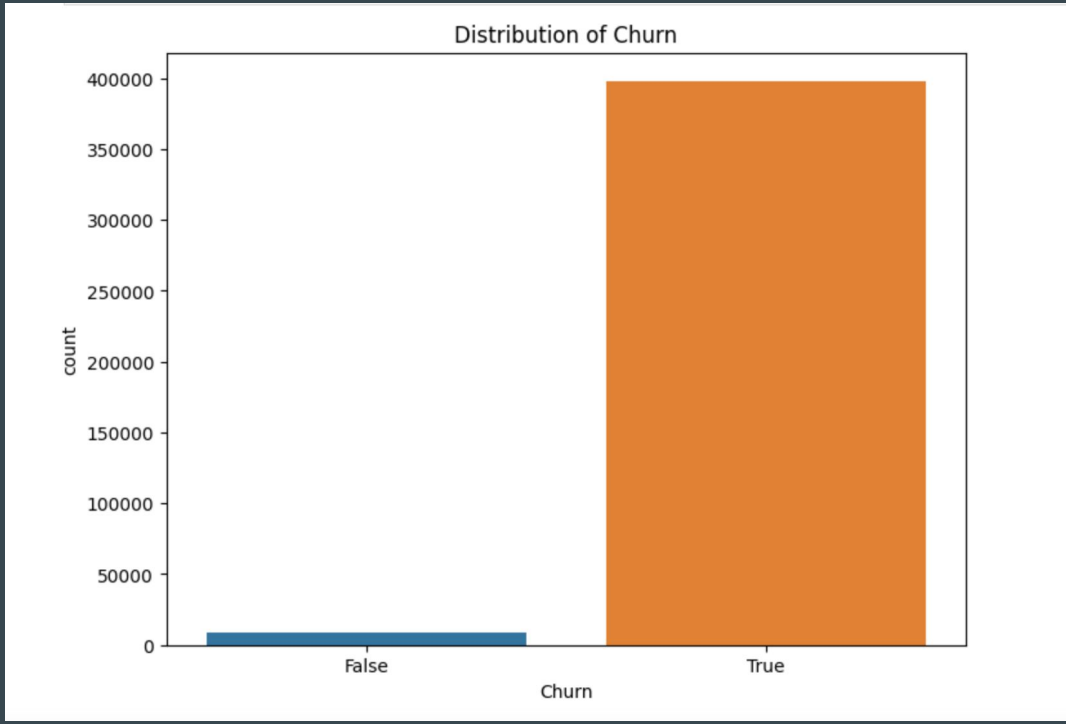


Countplot: Distribution of churned and non-churned customers

The countplot shows that the majority of customers do not churn, while few of the customers do churn.

The relatively low churn rate suggests that the company is doing a good job of retaining its customers.

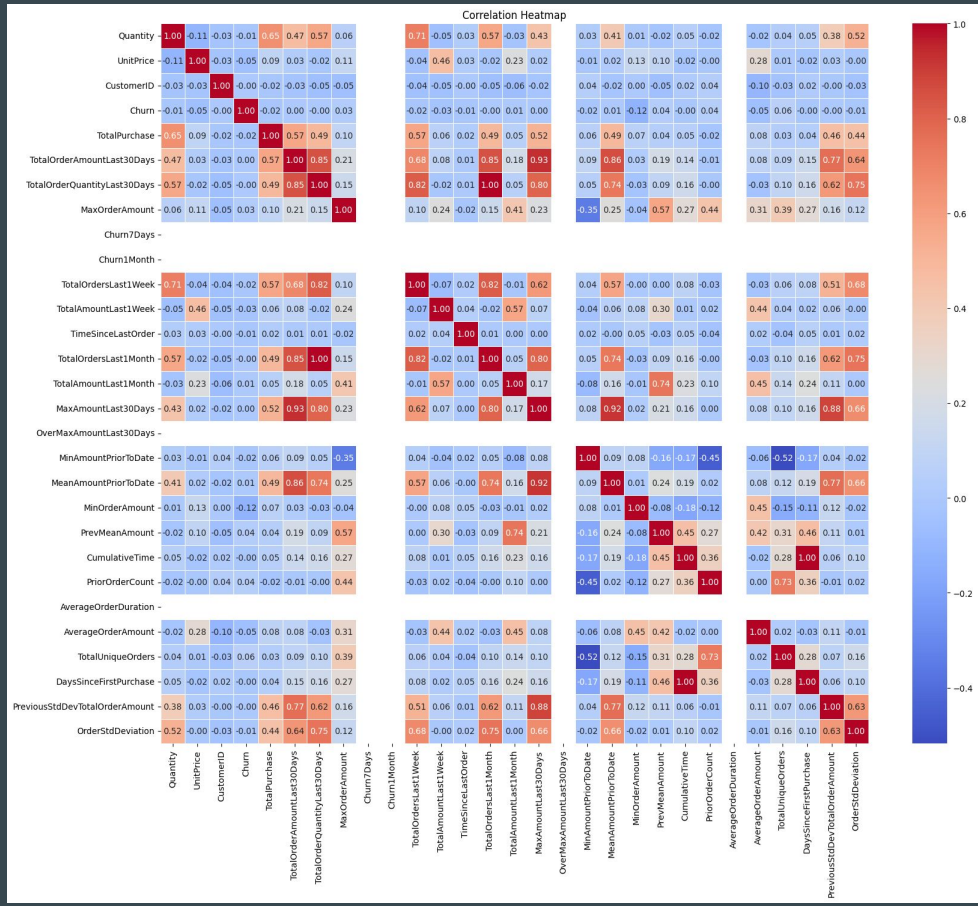
However, there is still room for improvement, very few of the customers are churning.



Correlation Heat Map

Findings:

- **TotalPurchase vs. Quantity (0.65):**
  Positive correlation indicates that as the quantity of items purchased increases, the total purchase amount also tends to increase.
- **TotalOrderAmountLast30Days vs. TotalOrderQuantityLast30Days (0.85):**
  Strong positive correlation suggests that as the quantity of orders increases, the total order amount for the last 30 days also increases.
- **TotalOrdersLast1Week vs. TotalOrderQuantityLast30Days (0.82):**
  Another strong positive correlation, indicating that the total order quantity in the last 30 days is related to the total orders in the last week.
- **TotalAmountLast1Week vs. UnitPrice (0.46):**
  Positive correlation suggests that the total amount spent in the last week is related to the unit price of items.
- **Churn vs. MinOrderAmount (-0.12):**
  Negative correlation indicates that as the minimum order amount decreases, the likelihood of churn slightly increases.



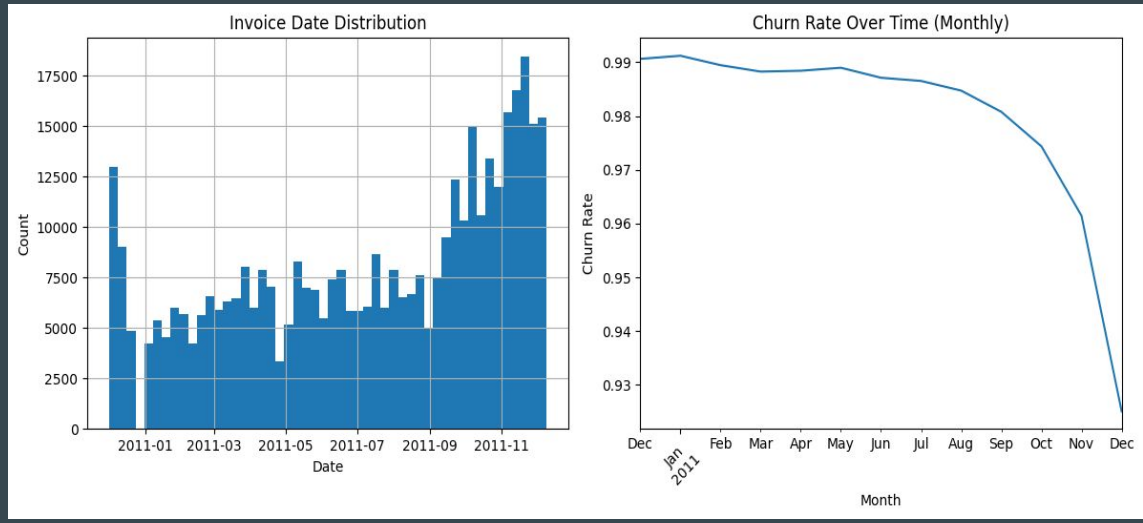
Time-Based analysis

The invoice date distribution graph shows that the number of invoices issued is relatively constant over time. This suggests that the company has a steady stream of new customers throughout the month.

The graph shows that the churn rate has been declining over time. This suggests that the company is becoming better at retaining its customers.

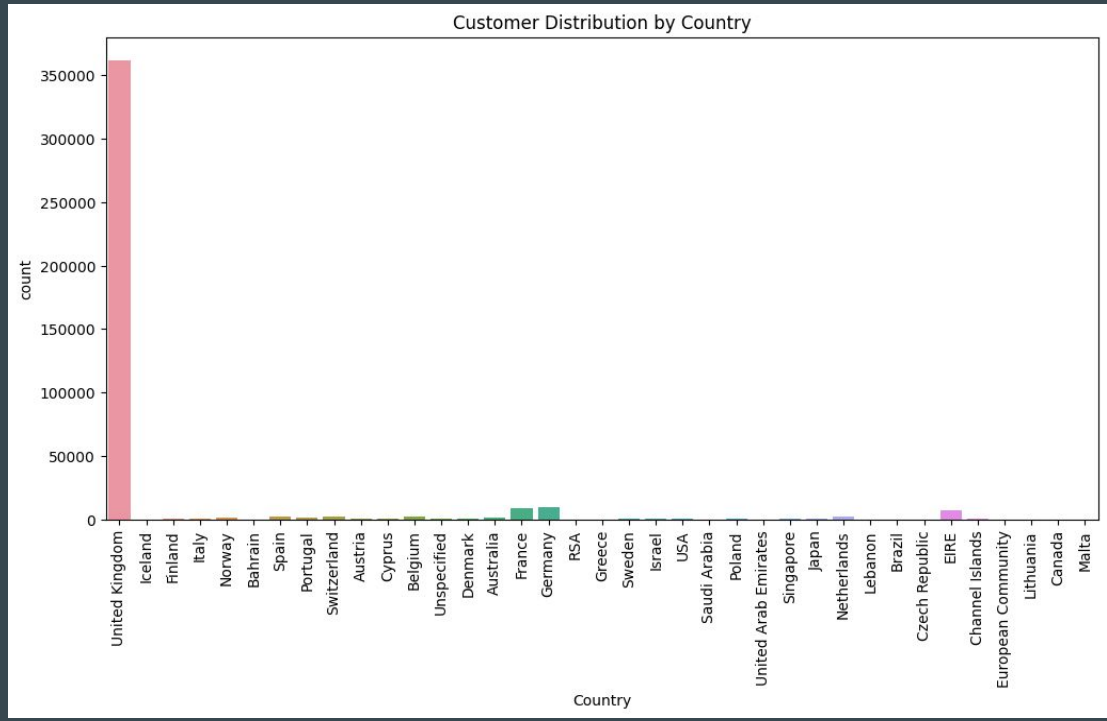
There are a few possible explanations for the declining churn rate.

One possibility is that the company is improving its customer experience. For example, the company may be providing better customer support, offering more personalized products or services, or making it easier for customers to do business with the company. There is still room for improvements.



Customer Distribution by Country

The chart shows the distribution of customers based on their country of origin. The countries with the most customers are represented by larger bars, while those with fewer customers are represented by smaller bars. The United Kingdom has the highest distribution. Followed by Germany, France and EIRE.



Model and Training Validation

The model training and validation phase resulted in an accuracy of 0.9798 and an F1-score of 0.989. These high scores indicate that the model was able to correctly predict churn in almost all cases.

- **Best Threshold Selection:**The model works best when you consider a customer at risk of churn if the predicted probability exceeds 0.45.
- The **confusion matrix** is a table that describes the performance of a classification model. In this case, it breaks down the predictions made by your churn prediction model on the validation dataset.

```
Best Threshold: 0.45
Best F1-Score: 0.9898180341388012
```

```
Validation Accuracy: 0.9798680814453685
Confusion Matrix:
[[   162   2373]
 [    84 119426]]
Classification Report:
              precision    recall  f1-score   support

       False       0.66      0.06      0.12      2535
        True       0.98      1.00      0.99    119510

    accuracy                           0.98    122045
   macro avg       0.82      0.53      0.55    122045
weighted avg       0.97      0.98      0.97    122045
```

- True Positive (TP): 119,426
  - This represents the number of customers correctly predicted as "True" churn cases. In simpler terms, these are customers who were predicted to churn and actually did.
- False Positive (FP): 2,373
  - These are customers predicted to churn but didn't. In other words, the model made an incorrect prediction that they would churn.
- True Negative (TN): 162
  - Customers correctly predicted as not churning. These are customers who were correctly identified as not leaving.
- False Negative (FN): 84
  - These are customers who actually churned but were predicted as not churning. This is where the model missed predicting churn.

Hyperparameter Tuning with Random Undersampling

- **Data Preprocessing:** One-hot encode the categorical column 'Country.'Perform feature engineering on datetime columns to extract relevant information.
- **Random Undersampling (RUS):**Utilized Random UnderSampling to balance the dataset by randomly removing instances from the majority class (non-churn).
- Train-Test Split for RUS Dataset: Split the RUS dataset into training and testing sets.
- **XGBoost Hyperparameter Tuning:**Define a hyperparameter grid for XGBoost model tuning. Used GridSearchCV to find the optimal hyperparameters based on ROC AUC.
- **XGBoost Model Training with Early Stopping**:Trained an XGBoost classifier with early stopping using the best hyperparameters from the grid search.
- **Evaluation Metrics:**Calculated accuracy, confusion matrix, and classification report on the test set.Compute ROC AUC score to assess the model's performance.

Summary of Findings:

- The tuned XGBoost model achieved a ROC AUC score of 0.8712 on the test set.
- Validation accuracy is 87.07%, indicating good overall performance.
- The confusion matrix and classification report provide detailed insights into the model's precision, recall, and F1-score for each class.

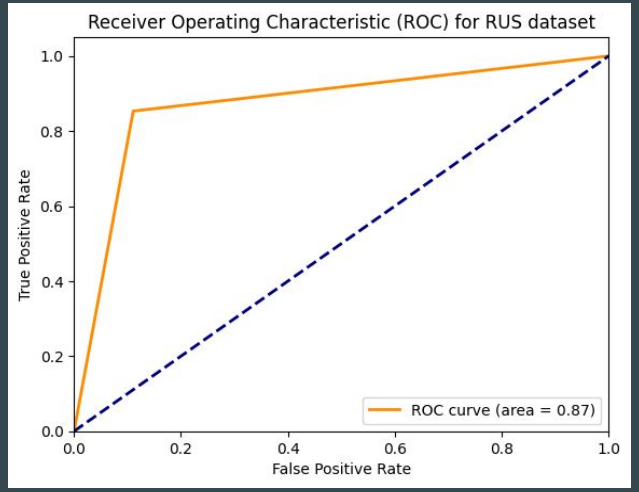
```
ROC AUC: 0.871203420590647
```

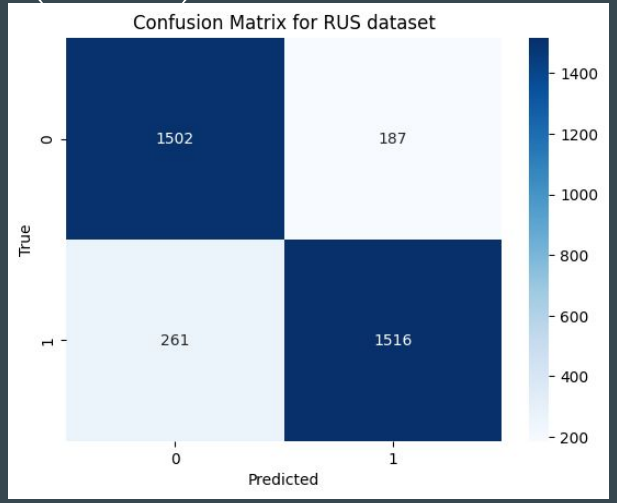
```
Validation Accuracy: 0.8707443739180611
Confusion Matrix:
[[1502  187]
 [ 261 1516]]
Classification Report:
              precision    recall  f1-score   support

       False       0.85      0.89      0.87      1689
        True       0.89      0.85      0.87      1777

    accuracy                           0.87      3466
   macro avg       0.87      0.87      0.87      3466
weighted avg       0.87      0.87      0.87      3466
```

ROC Curve / Confusion Matrix (RUS) Dataset:

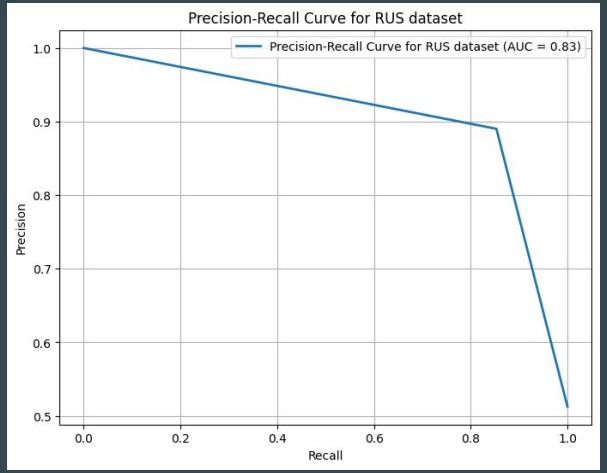


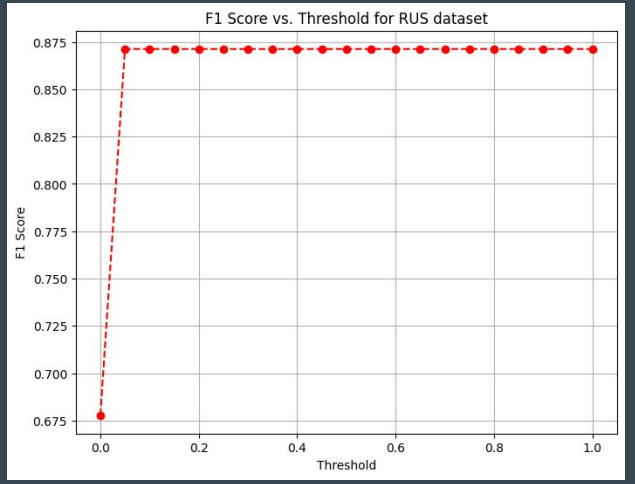


- **ROC Curve Findings:**AUC: The area under the ROC curve is 0.87, reflecting the model's strong ability to differentiate between churn and non-churn instances.
- **Curve Shape:** The curve is closer to the top-left corner, indicating robust model performance.
- **Dashed Line**: Represents a random classifier; The model outperforms randomness.

- **Confusion Matrix:**The confusion matrix breaks down the model's predictions:
- **True Negatives (TN):** 1502 instances where the model correctly predicted non-churn.
- **False Positives (FP):** 187 instances where the model incorrectly predicted churn.
- **False Negatives (FN):** 261 instances where the model incorrectly predicted non-churn.
- **True Positives (TP):** 1516 instances where the model correctly predicted churn.

Precision Recall Curve / F1 score vs Threshold for RUS

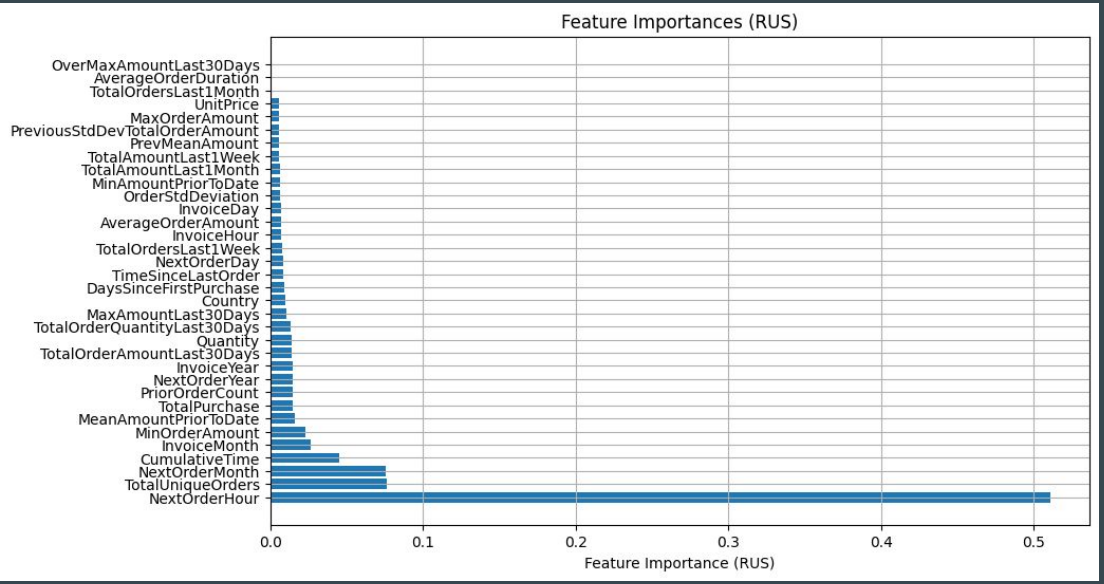




- **Findings from the Curve:**
- Trade-off Analysis: Balance between precision and recall is evident.
- Improving one metric may lead to a decrease in the other.
- High AUC and AP:AUC value of 0.87 and high AP indicate strong model performance.
- **Effective Positive Predictions:**Model can make accurate positive predictions.Maintains a balance with recall.
- **Model Suitability for Churn Prediction:**Curve reinforces the model's effectiveness in predicting customer churn.Provides actionable insights for retention strategies.

- **Threshold vs. F1 Score Plot Summary:**
- **F1 Score vs. Threshold Plot:X-axis (Threshold):** Ranges from 0 to 1 with increments of 0.05 intervals.
- Y-axis (F1 Score): Reflects the F1 score for each threshold.
- **Red Dotted Line:** Represents the F1 score curve.
- **Optimal Threshold:** Identifiable where the F1 score peaks.

Feature Importances of RUS dataset:



- Features contributing more to predicting churn have higher importances.The plot helps identify which features are significant in the model's decision-making.
- Higher importance values indicate that the corresponding feature has a more significant impact on the model's predictions.
- Lower importance values suggest that the feature has less influence on the model's decisions.

The top 4 most important features are:

Rank	Feature
1	NextOrderHour
2	TotalUniqueOrders
3	NextOrderMonth
4	CumulativeTime

Conclusion

The predictive model developed for customer churn demonstrates robust performance and effectiveness.

Model Performance:

- The XGBoost classifier, trained on a Random UnderSampled (RUS) dataset, exhibits high accuracy, with a ROC AUC of 0.87.
- Precision, recall, and F1-score metrics consistently indicate the model's ability to balance true positive and false positive rates.

Feature Insights:

- Top features such as NextOrderHour, TotalUniqueOrders, NextOrderMonth, and CumulativeTime significantly contribute to predicting customer churn.
- These insights provide actionable information for implementing targeted retention strategies.

Business Impact:

- The model serves as a valuable tool for businesses to proactively manage customer churn, allowing for timely intervention and personalized customer engagement.The identified features offer strategic guidance, aiding in the development of effective customer retention initiatives.
- Future Opportunities:Continuous model monitoring and updates are crucial to adapt to evolving business dynamics and changing customer behaviors.
- Exploring advanced modeling techniques and ensemble methods could further enhance predictive accuracy.

Overall Assessment:

- The project successfully addresses the challenge of customer churn prediction, providing a reliable solution for businesses seeking to optimize customer retention strategies.
- The model's effectiveness, backed by robust metrics and actionable insights, positions it as a valuable asset for businesses aiming to sustain customer loyalty and maximize revenue.