Statistics and Machine Learning 1

Coursework: EDA & Regression

Student ID: 14141925

1. Brief description of the data

The “MavenRail” dataset is a set of travel information that documents passenger train trips and whether the passengers have asked for refunds. This can be useful for identifying patients who may request a refund.

This dataset consists of 13 columns and 31645 items. The 13 columns are summarized below.

- Payment.Method: The way the ticket was paid for (such as "Contactless," "Credit Card").
- Railcard: The type of railcard that is used to indicate eligibility for discounts (e.g., "Adult," "None").
- Ticket.Class: The ticket's class (such as "Standard").
- Ticket.Type: The kind of ticket that was bought (for example, "Advance").
- Price: The price of the ticket is priced in pounds (£).
- Departure.Station and Arrival.Station: Stations for departure and arrival.
- Departure, Scheduled.Arrival, Actual.Arrival: Timestamps for departure and scheduled/actual arrival times in DD/MM/YYYY HH:mm format.
- Journey.Status: The status of the journey, e.g., "On Time" or "Delayed."
- Reason.for.Delay: Describes the reason for any delay, where applicable (e.g., "Signal Failure").
- Refund.Request: Indicates if a refund was requested (e.g., "Yes" or "No").

Only the data type of “Price” is “int64”, and the rest of the 12 variables' data types are all “object”. The statistical description indicates that the mean value of “Price” is 23.435 and the standard deviation of “Price” is 29.99.

There are 20911 missing values in Railcard, 1880 missing values in

"Actual.Arrival", 3 missing values in "Departure", 4 missing values in "Scheduled.Arrival", and 27479 missing values in "Reason.for.Delay".

2. Exploratory data analysis

The correlation heatmap shows the correlation between all variables, which data type is float and integer. Figure 1 indicates that the correlation between "Journey.Status" and "Reason.for.Delay" is high, indicating a strong positive correlation between these two variables.

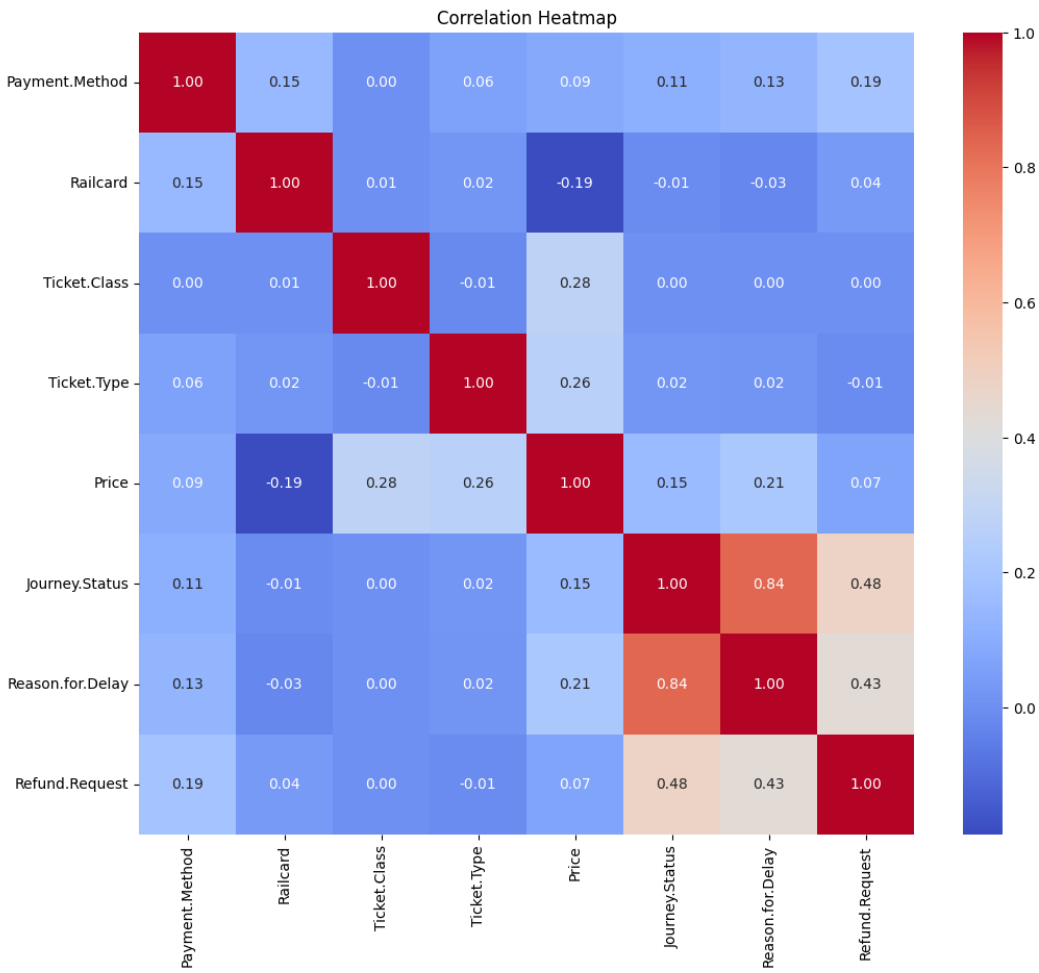

Figure 1

Figure 2 is about price distribution. It indicates that most of the train ticket prices lie between £0 to £50.

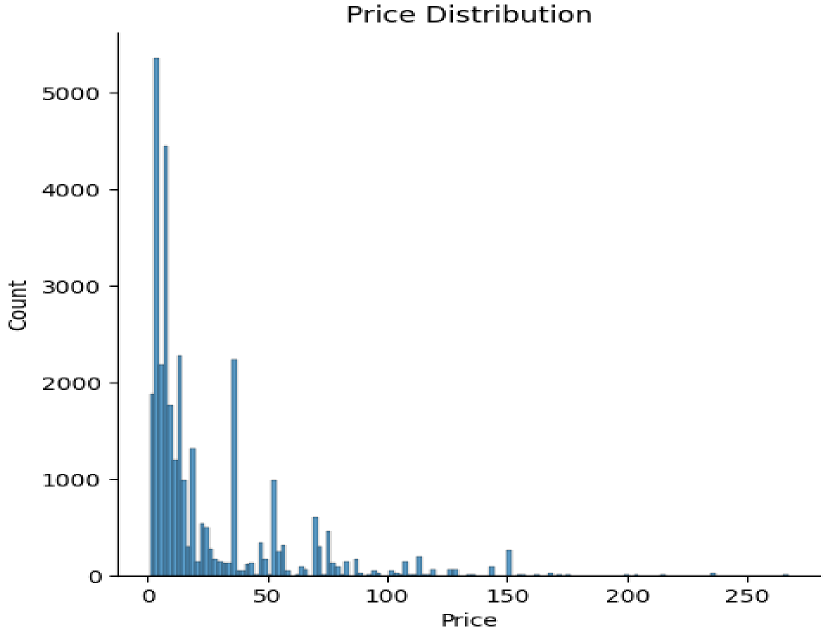

Figure 2

Figure 3, which is the Railcard Distribution plot, indicates that most of the passengers do not have Railcard. If the passenger owns a railcard, most of them own an Adult Railcard.

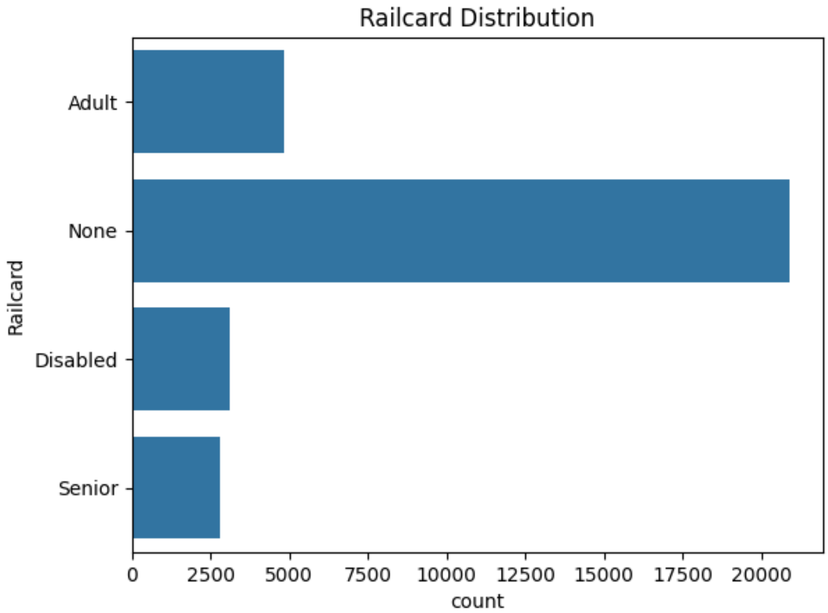

Figure 3

Figure 4 shows that most of the passengers choose to pay with a credit card. However, the number of passengers who pay by debit card is more likely to request a refund.

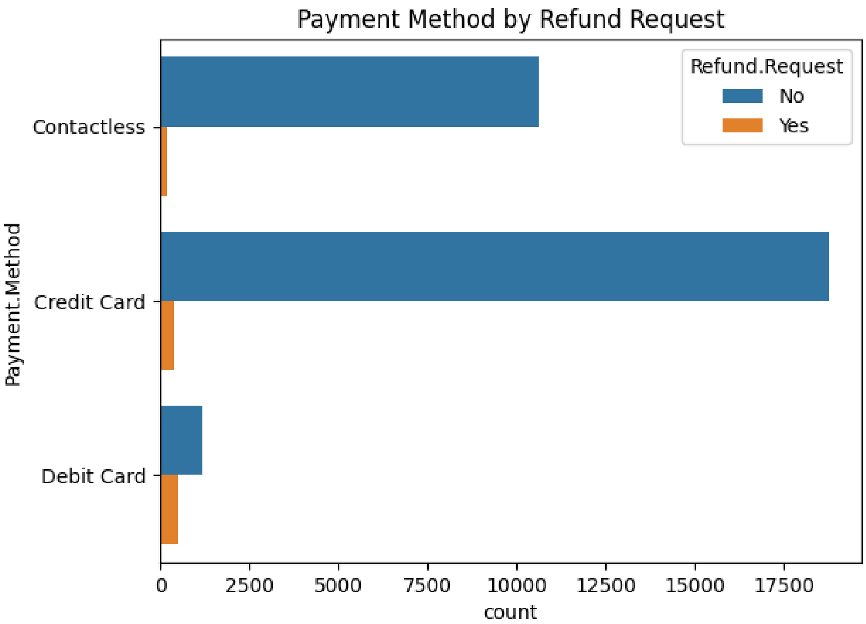

Figure 4

Figure 5 shows that most of the passengers' ticket type is Advance, and most of the passengers who request a refund bought Advance tickets.

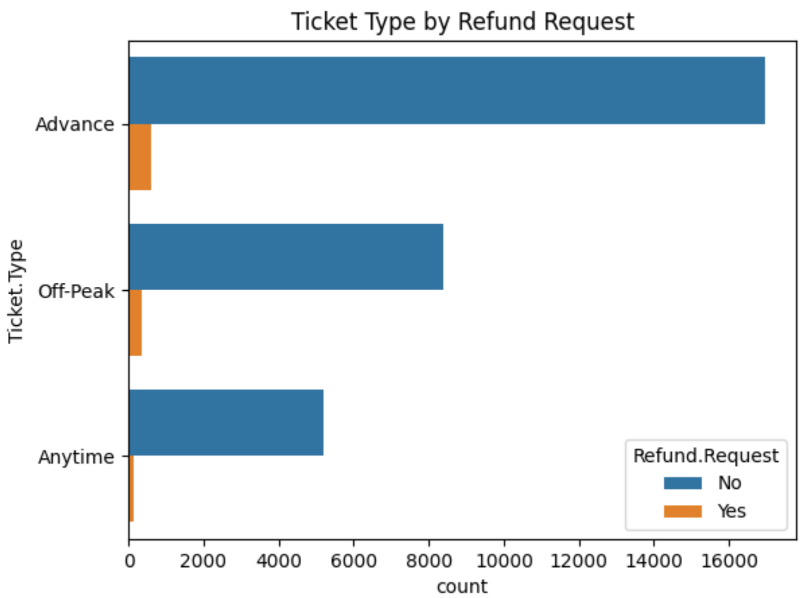

Figure 5

Table 1 and Figure 6 show the distribution of reasons for delay in refund requests. Among the causes of delays, weather factors account for the highest proportion. However, when requesting refunds, technical issues were cited as the highest number of reasons for delays. Over 50 percent of the passengers requested a refund when they faced delays due to technical issues.

Refund.Request	No	Yes
Reason.for.Delay		
Signal Failure	753	215
Staff	320	79
Staffing	228	179
Technical Issue	319	387
Traffic	193	121
Weather	1239	133

Table 1: Reason for Delay by Refund Request

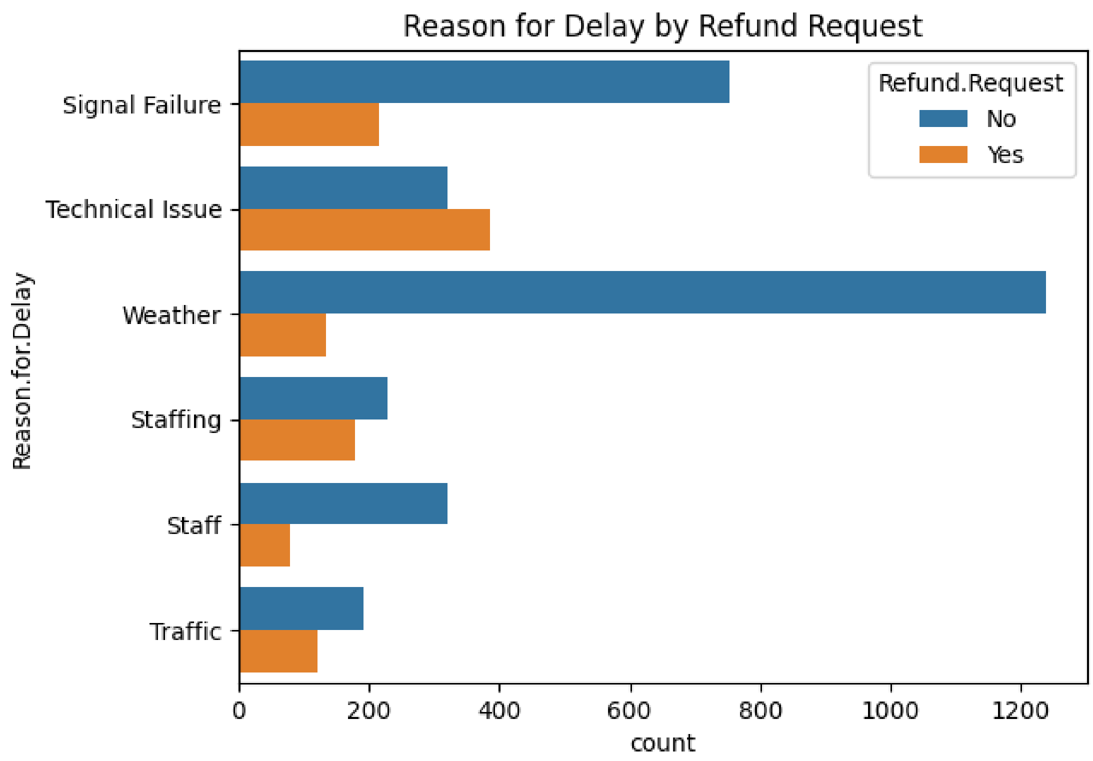

Figure 6

3. Add “DelayInMinutes” column

	Payment.Method	Railcard	Ticket.Class	Ticket.Type	Price	Departure.Station	Arrival.Station	Departure	Scheduled.Arrival	Actual.Arrival	Journey.Status	Reason.for.Delay	Refund.Request	DelayInMinutes
0	Contactless	Adult	Standard	Advance	43	London Paddington	Liverpool Lime Street	2024-01-01 11:00:00	2024-01-01 13:30:00	2024-01-01 13:30:00	On Time	NaN	No	NaN
1	Credit Card	Adult	Standard	Advance	23	London Kings Cross	York	2024-01-01 09:45:00	2024-01-01 11:35:00	2024-01-01 11:40:00	Delayed	Signal Failure	No	5.0
2	Credit Card	None	Standard	Advance	3	Liverpool Lime Street	Manchester Piccadilly	2024-01-02 18:15:00	2024-01-02 18:45:00	2024-01-02 18:45:00	On Time	NaN	No	NaN
3	Credit Card	None	Standard	Advance	13	London Paddington	Reading	2024-01-01 21:30:00	2024-01-01 22:30:00	2024-01-01 22:30:00	On Time	NaN	No	NaN
4	Contactless	None	Standard	Advance	76	Liverpool Lime Street	London Euston	2024-01-01 16:45:00	2024-01-01 19:00:00	2024-01-01 19:00:00	On Time	NaN	No	NaN

Table 2: Added new column “DelayInMinutes”

4. Regression Model of MediumPrice

	Payment.Method	Railcard	Ticket.Class	Ticket.Type	Price	Departure.Station	Arrival.Station	Departure	Scheduled.Arrival	Actual.Arrival	Journey.Status	Reason.for.Delay	Refund.Request	DelayInMinutes	MediumPrice
1	Credit Card	Adult	Standard	Advance	23	London Kings Cross	York	2024-01-01 09:45:00	2024-01-01 11:35:00	2024-01-01 11:40:00	Delayed	Signal Failure	0	5.0	True
8	Credit Card	None	Standard	Advance	37	London Euston	York	2024-01-01 00:00:00	2024-01-01 01:50:00	2024-01-01 02:07:00	Delayed	Signal Failure	0	17.0	False
20	Debit Card	Adult	Standard	Advance	7	Birmingham New Street	Manchester Piccadilly	2024-01-01 11:15:00	2024-01-01 12:35:00	2024-01-01 13:06:00	Delayed	Technical Issue	1	31.0	False
26	Credit Card	Senior	First Class	Advance	34	Oxford	Bristol Temple Meads	2024-01-01 14:15:00	2024-01-01 15:30:00	2024-01-01 15:54:00	Delayed	Signal Failure	1	24.0	False
39	Credit Card	None	Standard	Advance	7	London Euston	Birmingham New Street	2024-01-02 02:15:00	2024-01-02 03:35:00	NaT	Cancelled	Technical Issue	0	NaN	False

Table 3: Added new column “MediumPrice

I used logistic regression, which is a statistical analysis module to predict a binary outcome based on other variables in the data set, to fit the model and to predict whether a passenger will ask for a refund using a single variable "MediumPrice". The model was trained on 70% of the data and tested on the remaining 30%, achieving an accuracy of 0.73 and an AUC of 0.53.

Simple model: use binary predictor “MediumPrice”

$$log\left(\frac{\widehat{p_i}}{1-\widehat{p_i}}\right) = -1.076 + 0.258 \times MediumPrice$$

- The probability of refund if the ticket costs £5:

the odds of a passenger requesting refund if “MediumPrice” = 0:

$$\left(\frac{\widehat{p_i}}{1-\widehat{p_i}}\right) = exp(-1.076) = 0.341$$

the probability of a passenger requesting refund if “MediumPrice” = 0:

$$\widehat{p_i} = \frac{0.341}{1.341} = 0.254$$

- The probability of refund if the ticket costs £25:

the odds of a passenger request refund if “MediumPrice” = 1:

$$\left(\frac{\widehat{p_\iota}}{1-\widehat{p_\iota}}\right) = exp(-0.818) = 0.441$$

the probability of a passenger request refund if “MediumPrice” = 1:

$$\widehat{p_\iota} = \frac{0.441}{1.441} = 0.306$$

The probability of a refund is 0.254 for a £5 ticket and 0.306 for a £25 ticket.

5. Prediction using ToPredict.csv

I added a categorical variable, “MediumPrice,” in both MavenRail.csv and ToPredict.csv: tickets under £10 are coded as 0, tickets between £10 and £30 as 1, and tickets over £30 as 2. Since our target variable, “Refund.Request,” is binary, I used logistic regression to train and predict the data.

Several logistic regression models were fitted using different predictor combinations and test-train splits. Model 1 includes the predictors “Payment.Method,” “Railcard,” “Ticket.Class,” “Ticket.Type,” “Price,” “Journey.Status,” “Reason.for.Delay,” “DelayInMinutes,” and “MediumPrice” to predict “Refund.Request.” The dataset was split with 70% of MavenRail used for training and 30% for testing.

Passengers	Probability to request a refund
1	0.038890
2	0.002538
3	0.446166
4	0.004587
5	0.291547
6	0.082446
7	0.417309
8	0.410066

Table 4: Probability of requesting a refund by Module1

Measure	Value
AUC - ROC Score	0.969647
Accuracy	0.962604
Precision	0.405286
Recall	0.294872
F1 Score	0.341373

Table 5: Measures of Module1

Table 6: AUC-ROC Curve of Module1

Model 2 fits a logistic regression using the predictors "Payment.Method," "Railcard," "Ticket.Class," "Ticket.Type," "Price," "Reason.for.Delay," "DelayInMinutes," and "MediumPrice" to predict "Refund.Request." Here, 70% of the MavenRail data is used for training, and 30% for testing. Due to the high correlation between "Reason.for.Delay" and "Journey.Status," only "Reason.for.Delay" is included in this model.

	Probability to request a refund
Passengers	
1	0.091808
2	0.010597
3	0.657184
4	0.002472
5	0.055418
6	0.066440
7	0.652812
8	0.037966

Table 7: Probability of requesting a refund by Module2

	Value
Measure	
AUC - ROC Score	0.908976
Accuracy	0.965870
Precision	0.457746
Recall	0.208333
F1 Score	0.286344

Table 8: Measures of Module2

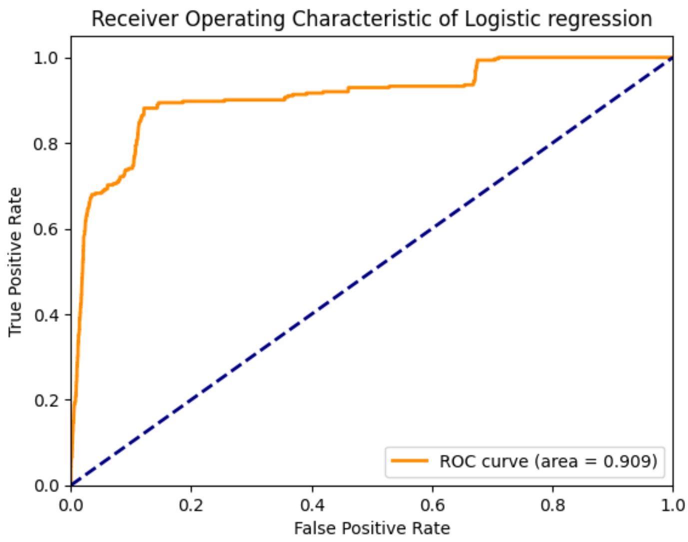

Table 9: AUC-ROC Curve of Module2

Model 3 fits a logistic regression using “Payment.Method,” “Ticket.Type,” “Reason.for.Delay,” and “MediumPrice” as predictors for “Refund.Request.” The data is split with 80% of MavenRail used for training and 20% for testing.

Measure	Value
AUC - ROC Score	0.902888
Accuracy	0.965239
Precision	0.380952
Recall	0.160000
F1 Score	0.225352

Table 10: Probability of requesting a refund by Module3

Passengers	Probability to request a refund
1	0.095821
2	0.012329
3	0.724318
4	0.003702
5	0.052103
6	0.059684
7	0.724318
8	0.033410

Table 11: Measures of Module3

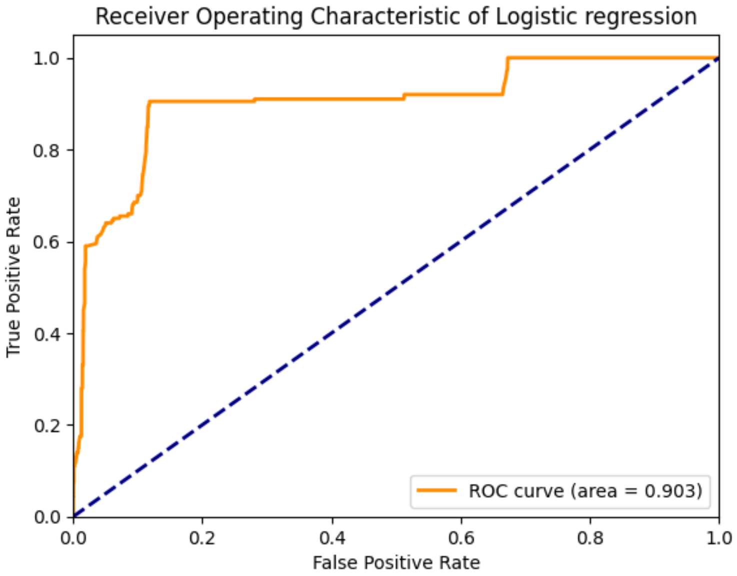

Table 12: AUC-ROC Curve of Module3

The ROC curve shows a binary classifier's performance across different decision thresholds by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The ROC AUC score, the area under this curve, summarizes the model's ability to distinguish between positive and negative examples. A score of 0.5 indicates random guessing, while 1 indicates perfect performance. Therefore, since the AUC-ROC score of Module1 is the highest, which is about 0.970, I chose Module1 as my final module.

6. References

- Understanding Logistic Regression in Python
- **AUC and the ROC Curve in Machine Learning**
- **How to explain the ROC curve and ROC AUC score?**