Institute of Technology of Cambodia

Department of Applied Mathematics and Statistic

Group: I3-AMS-03

Subject: Advance Probability

Topic: Analyzing the factors that affect customer satisfaction in the airline industry using ordinal logistic regression

Tang Piseth	e20201634
Kry Senghort	e20200706
Sao Samarth	e20200084
Hon Rathana	e20201053
Pheng Sothea	e20201734
Ek Vongpanharith	e20200087

CONTENTS:

1. LINEAR REGRESSION ANALYSIS
2. FEATURE SELECTION
3. CONFOUNDER
4. MODEL CREATION
5. MODEL SELECTION
6. MODEL EVALUATION
7. CONCLUSION

1. Linear regression analysis

Linearity

satisfaction ~ Departure_Delay_in_Minutes + Arrival_Delay_in_Minutes

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     241.2
Date:                Tue, 23 May 2023   Prob (F-statistic):          2.79e-105
Time:                        23:16:39   Log-Likelihood:                -92900.
No. Observations:              129880   AIC:                         1.858e+05
Df Residuals:                  129877   BIC:                         1.858e+05
Df Model:                           2
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                     0.4454      0.001    301.903      0.000       0.443       0.448
Departure_Delay_in_Minutes    0.0008      0.000      6.455      0.000       0.001       0.001
Arrival_Delay_in_Minutes     -0.0015      0.000    -12.115      0.000      -0.002      -0.001
==============================================================================
Omnibus:                   466876.760   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            21368.532
Skew:                           0.263   Prob(JB):                         0.00
Kurtosis:                       1.084   Cond. No.                         61.8
==============================================================================
```

satisfaction ~ Arrival_Delay_in_Minutes

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     440.6
Date:                Sat, 20 May 2023   Prob (F-statistic):           1.17e-97
Time:                        17:20:38   Log-Likelihood:                -92920.
No. Observations:              129880   AIC:                         1.858e+05
Df Residuals:                  129878   BIC:                         1.859e+05
Df Model:                           1
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                    0.4458      0.001    302.274      0.000       0.443       0.449
Arrival_Delay_in_Minutes    -0.0008   3.57e-05    -20.990      0.000      -0.001      -0.001
==============================================================================
Omnibus:                   466538.303   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            21395.028
Skew:                           0.263   Prob(JB):                         0.00
Kurtosis:                       1.083   Cond. No.                         44.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

satisfaction ~ Departure_Delay_in_Minutes

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     335.2
Date:                Sat, 20 May 2023   Prob (F-statistic):           8.63e-75
Time:                        17:20:40   Log-Likelihood:                -92973.
No. Observations:              129880   AIC:                         1.859e+05
Df Residuals:                  129878   BIC:                         1.860e+05
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                     0.4442      0.001    301.615      0.000       0.441       0.447
Departure_Delay_in_Minutes   -0.0007   3.61e-05    -18.310      0.000      -0.001      -0.001
==============================================================================
Omnibus:                   465598.855   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            21462.111
Skew:                           0.263   Prob(JB):                         0.00
Kurtosis:                       1.079   Cond. No.                         43.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Multicolinearity

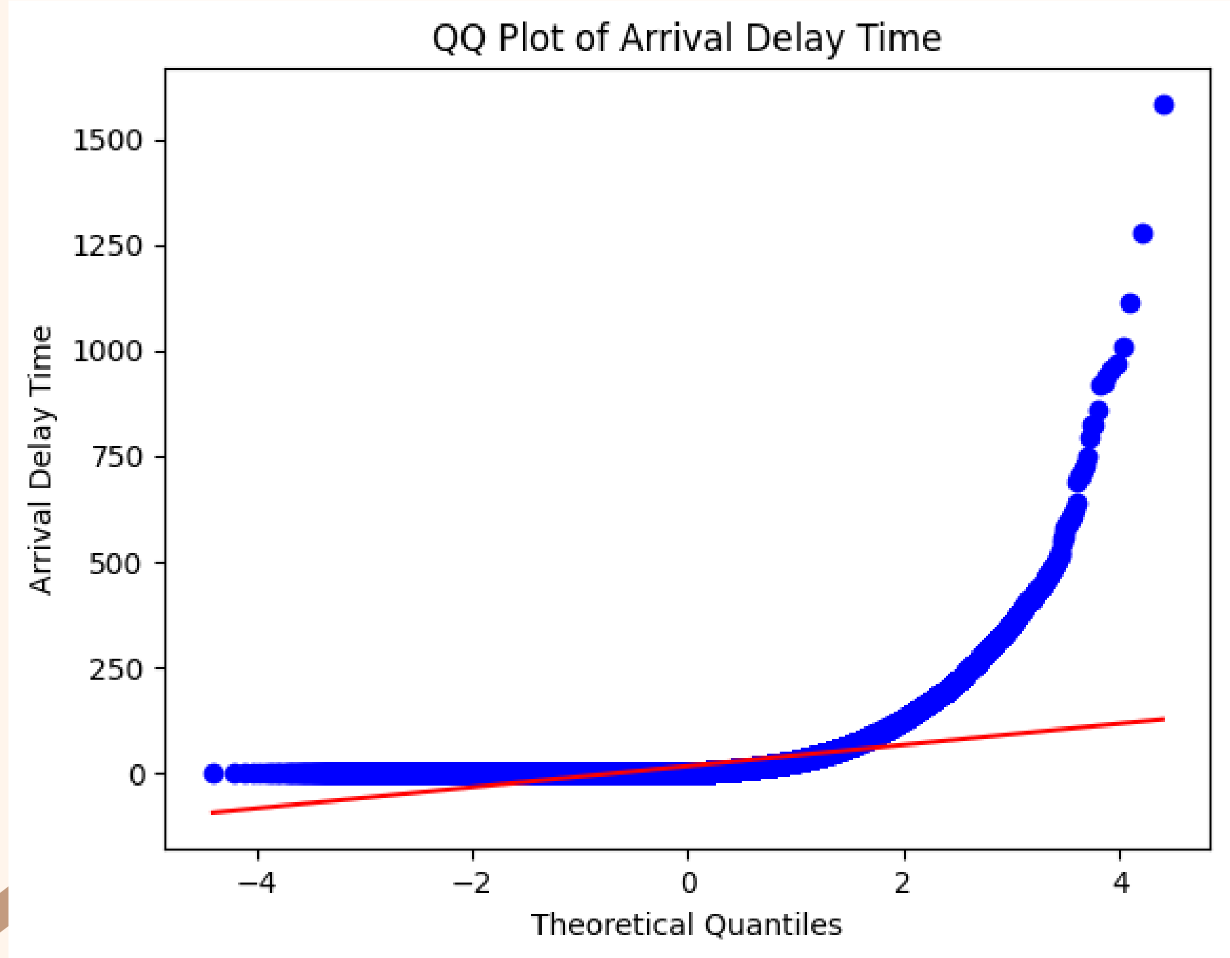
	VIF Factor	features
0	12.565103	Departure_Delay_in_Minutes
1	12.565103	Arrival_Delay_in_Minutes
2	1.154836	Intercept

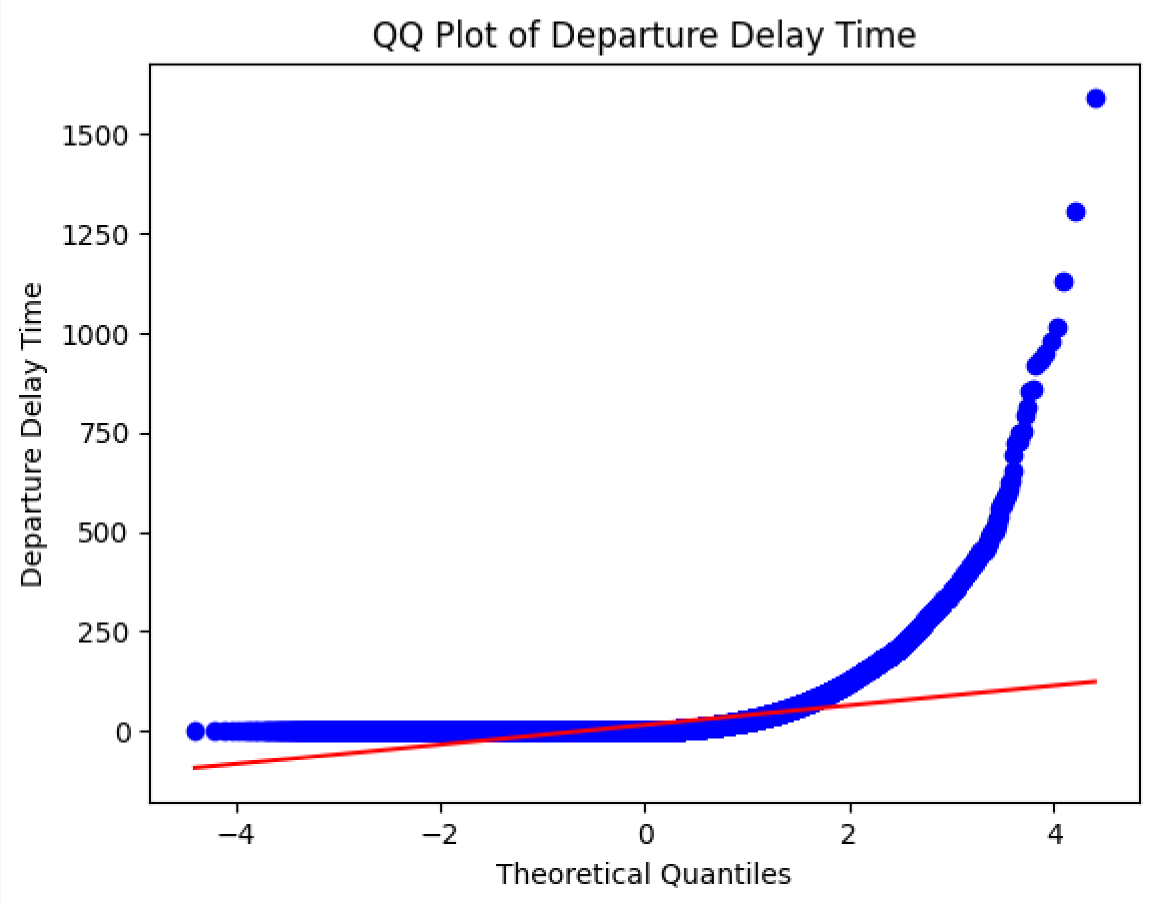
	Departure_Delay_in_Minutes	Arrival_Delay_in_Minutes	Intercept
0	25	18.0	1
1	1	6.0	1
2	0	0.0	1
3	11	9.0	1
4	0	0.0	1
...	...	...	...
129875	0	0.0	1
129876	0	0.0	1
129877	0	0.0	1
129878	0	0.0	1
129879	0	0.0	1

129880 rows × 3 columns

Both 'Departure_Delay_in_Minutes' and 'Arrival_Delay_in_Minutes' have relatively high VIF factors of approximately 12.565103. VIF values above 5 or 10 are often considered indicative of significant multicollinearity

NORMALITY


QQ Plot of Arrival Delay Time
Arrival Delay Time
1500
1250
1000
750
500
250
0
−4
−2
0
2
4
Theoretical Quantiles


QQ Plot of Departure Delay Time
Departure Delay Time
1500
1250
1000
750
500
250
0
−4
−2
0
2
4
Theoretical Quantiles


QQ Plot of Age
Age
100
80
60
40
20
0
−20
−4
−2
0
2
4
Theoretical Quantiles

2. Feature selection

Encoding categorical variables

Label encoding is the simplest method of encoding categorical variables. involves replacing each category with a numerical value.

```
Column: Customer_Type
Loyal Customer -> 0
disloyal Customer -> 1

Column: Type_of_Travel
Personal Travel -> 1
Business travel -> 0
```

Method used for feature selection

- **The Chi-Square method**

In feature selection, it measures the dependency between each feature and the target variable using the χ^2 statistic. It selects the k features with the highest χ^2 scores. This method is suitable for categorical features and a categorical target variable.

```
The 10 most important features selected by chi-square test:
Index(['id', 'Age', 'Type_of_Travel', 'Class', 'Flight_Distance',
       'Online_boarding', 'Seat_comfort', 'Inflight_entertainment',
       'Departure_Delay_in_Minutes', 'Arrival_Delay_in_Minutes'],
      dtype='object')
```

• The Wrapper Method

The Wrapper method evaluates the model's performance with different subsets of features. Base on this method used a Random Forest model as the base estimator to perform feature selection and we could find the important features such as:

```
Index(['Online_boarding', 'Inflight_wifi_service', 'Type_of_Travel', 'Class',
       'Inflight_entertainment', 'Seat_comfort', 'Flight_Distance',
       'Customer_Type', 'Ease_of_Online_booking', 'On-board_service'],
      dtype='object')
```

- **Feature Permutation Importance**

Permutation Importance is a technique that measures the impact of shuffling a feature's values and calculates the importance of each feature by evaluating how much the model's performance decreases when the feature's values are randomly permuted. Higher importance scores indicate more influential features.

Weight	Feature
0.1466 ± 0.0008	Inflight_wifi_service
0.1350 ± 0.0016	Type_of_Travel
0.0532 ± 0.0006	Customer_Type
0.0411 ± 0.0011	Online_boarding
0.0339 ± 0.0005	Class
0.0259 ± 0.0004	Checkin_service
0.0186 ± 0.0005	Seat_comfort
0.0179 ± 0.0003	Baggage_handling
0.0158 ± 0.0002	Inflight_service
0.0151 ± 0.0006	Cleanliness
0.0146 ± 0.0007	id
0.0108 ± 0.0004	Age
0.0101 ± 0.0005	On-board_service
0.0084 ± 0.0003	Leg_room_service
0.0082 ± 0.0003	Flight_Distance
0.0074 ± 0.0003	Inflight_entertainment
0.0073 ± 0.0001	Arrival_Delay_in_Minutes
0.0060 ± 0.0002	Ease_of_Online_booking
0.0043 ± 0.0003	Gate_location
0.0040 ± 0.0002	Departure_Delay_in_Minutes
… 3 more …	

From all above results:

- Really Important Featurues:

Type_of_Travel, Inflight_wifi_service, Online_boarding, Seat_comfort

- Important Features:

Class, Flight_Distance, Inflight_entertainment, On-board_service, Leg_room_service, Cleanliness, Checkin_service, Inflight_service, Baggage_handling

- **Recursive Feature Elimination**

RFE is an recursive feature selection method that starts with all features and progressively eliminates the least important features based on their coefficients or importance scores.

It utilizes the logistic regression model to assess the importance of each feature and recursively prunes the least important features until the desired number of features is reached.

```
['Customer_Type',
 'Type_of_Travel',
 'Class',
 'Inflight_wifi_service',
 'Ease_of_Online_booking',
 'Online_boarding',
 'On-board_service',
 'Leg_room_service',
 'Checkin_service',
 'Cleanliness']
```

3. Confounder

We used multi-variate analysis on the suspected features that we think are confounders: "Age, "Type of Travel" and "Class".

Based on the analysis, Age is not a confounder, while both Type_of_Travel and Class can be considered confounders. Therefore, from now on, we will not include Age in our combined features and it's pretty logical that id should also not be in there.

5. Model selection

We tried all iterative selection methods and found out that most of them resulted in more than 15 features.

6. Model evaluation

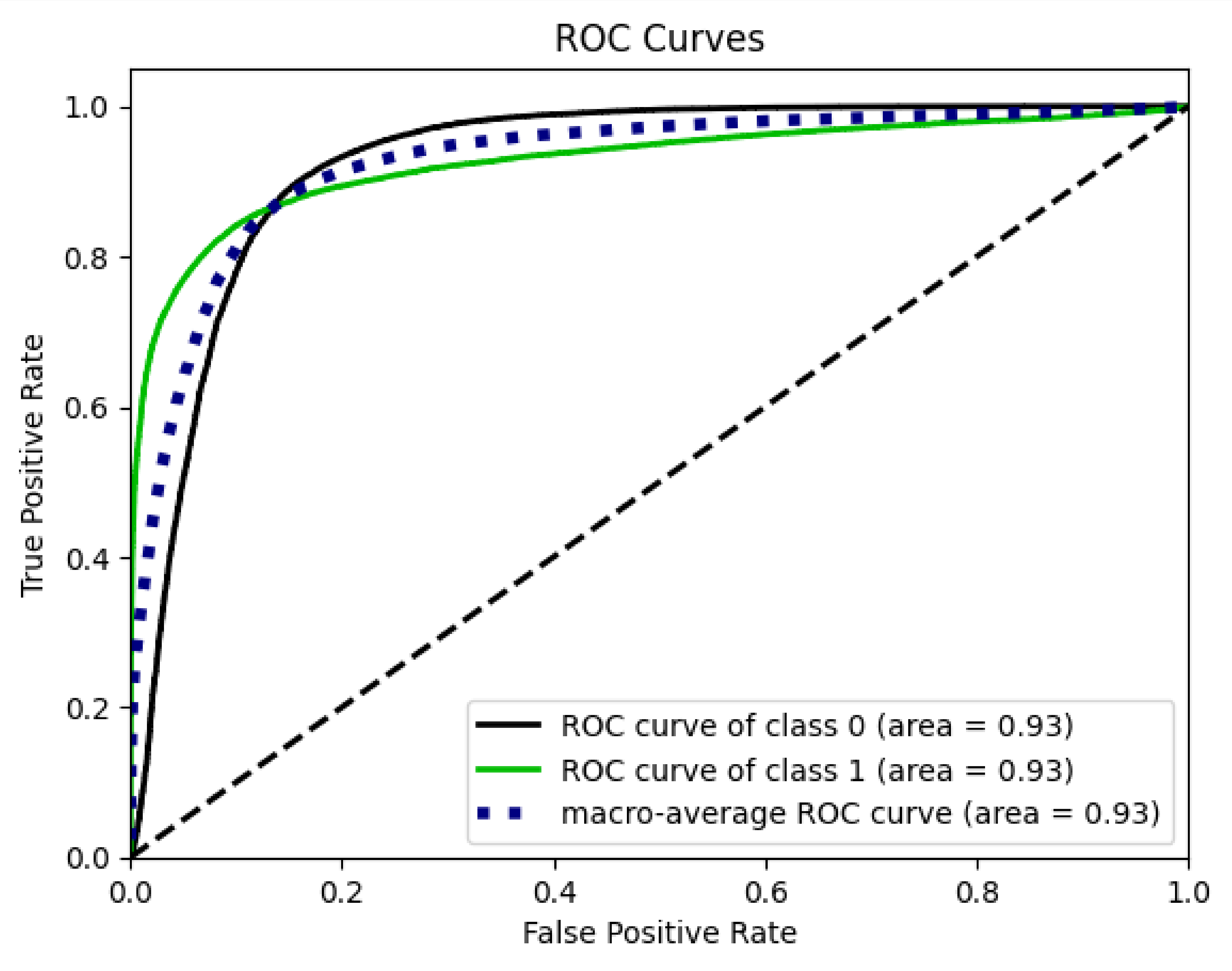
As our evaluation metrics, we used the ROC curves, AUC, and confusion matrix. But we believe that the confusion matrices carry less values and decide to only show the ROC curves.

4. Model creation

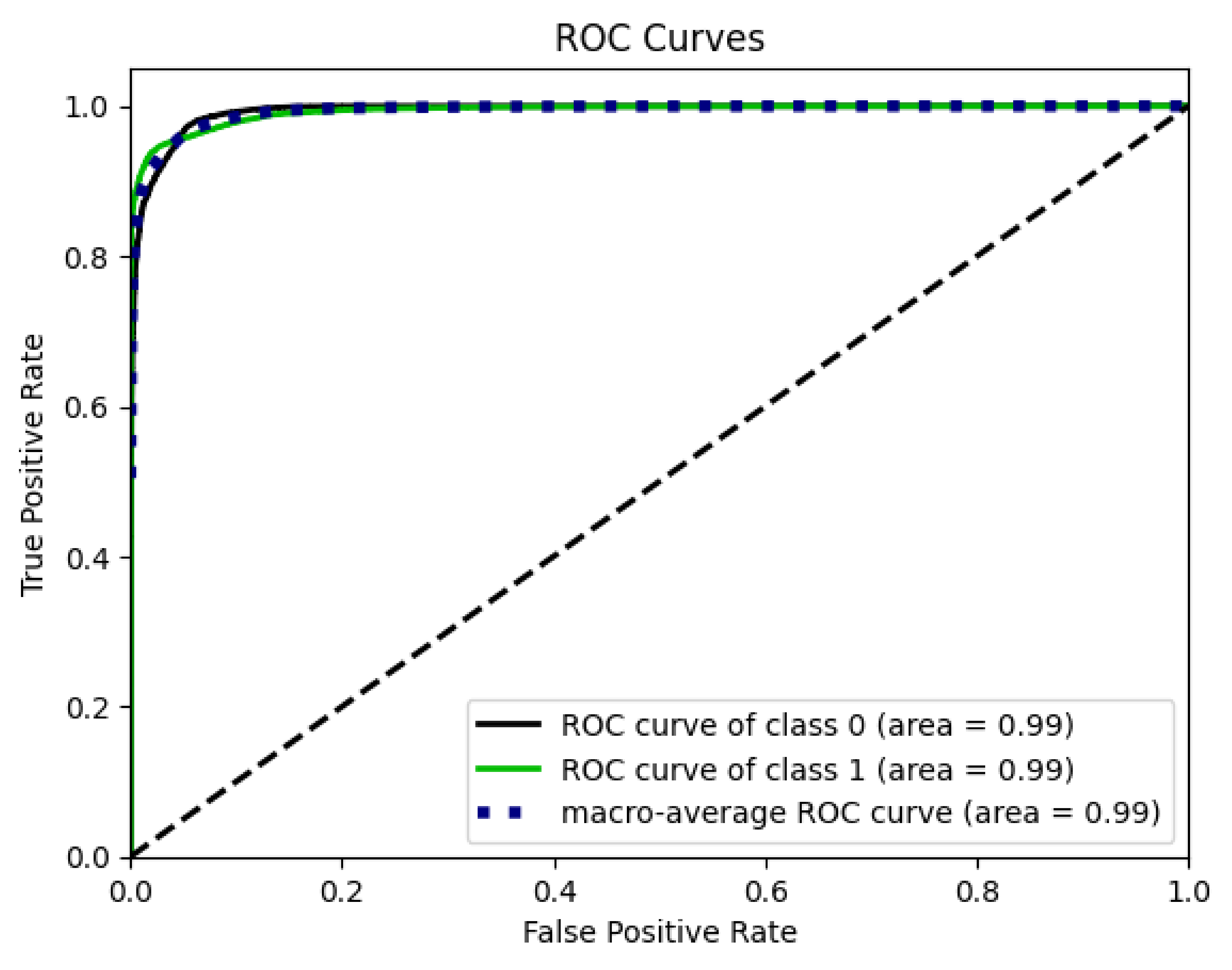
We fit/train only two models:

- **A logistic model(simple one)**
- **Random Forest Classifier**

ROC Curve of Simple Logistic regression



ROC Curve of RFC



7. Conclusion

Based on the ROC curve and AUC score, it's clear that random forest classifier perform better than the fit Logistic Model.