Analysis on passengers' satisfaction level using Logistic Regression

Advanced Probability

2022-2023

Group members:

Ek Vong Panharith	e20200877
Hon Ratana	e20201053
Sao Samarth	e20200084
Pheng Sothea	e20201734
Kry Senghort	e20200706
Tang Piseth	e20201634

Submission Date: 14^{th} June, 2023

Professors/Instructors:

Course section:	PHAUK SOKKHEY
TD:	NHIM MALAI

Contents

Abstract

Background: The airline industry is a competitive industry where customer satisfaction is critical to success. Airlines need to understand the factors that affect customer satisfaction in order to improve their services and attract more customers. This will be done by collecting data from the Airline Passenger Satisfaction dataset kaggle airline dataset and using ordinal logistic regression to analyze the data.

Objectives: The key objective for the project is to identify the most important factors that influence customer satisfaction and to develop strategies to improve their service in order to increase customer satisfaction.

Methodology: Our project pipeline is generic since it follows a general machine learning project. We start off with data pre-processing(handling missing values and outliers) and then go on to feature engineering, which we focused much on feature selection based on a couple of statistical learning(such as χ^2 significance test) and machine learning algorithms(such as decision tree and random forest) and then we use different ML algorithms to train the data set on and evaluate.

Results: At the end, we found out at **random forest classifier** is the best giving the AUC score of 99 consistently with **decision tree classifier** following from behind with a high AUC score of 95.

Conclusions/future work: We believe that there are still much to work on. Such as automating the pre-processing steps by making a pipeline to speed up the project, giving more emphasis on statistical work on finding confounding variables, more analysis on dispersion, correlation analysis, and trying out more feature selection methods to really identify the individual effects of each feature and last but not least we can definitely try out more ML algorithms such as **Ada Gradient Boost** and so on.

Keywords: airline, Kaggle, random forest classifier, customer satisfaction

1 Logistic Regression

Regression analysis presents the association between a response variable and one or more explanatory variables. It is often the situation that the outcome variable is discrete, assuming two or more potential values. BLRA represents a special condition of linear regression analysis LRA used when the response is binary not continuous, and the explanatory variables are quantitative or qualitative variables. It was first suggested in the 1970s to overcome difficulties of ordinary least squares OLS regression in treating binary outcomes. Logistic regression LR uses the theory of binomial probability which represents having only two values to predict: that probability (p) is 1 instead of 0, i.e. the event belongs to one group instead of the other. LR presents the best fitting function depending on the maximum likelihood ML approach, which maximizes the distinguishing probability of the observed data into the suitable category given the coefficients of regression.

1.1 Assumptions of Logistic Regression

- Linearity of independent variables and log-odds
- No strongly influential outliers
- Independence of observations
- Sufficiently large sample size

1.2 Failures of the assumptions of linear regression model

We will give empirical evidence for why OLS model is not suitable for our data set.

- Simple linear regression is one quantitative variable predicting another quantitative variable
- Multiple LR is still simple LR with many independent variables
- Nonlinear regression is a couple of quantitative variables but the data is curvillinear

As such, it is clear that our next and natural decision is logistic regression.

2 Data Preprocessing

2.1 Data Description

```
Data columns (total 24 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   id                                 129880 non-null  int64
 1   Gender                             129880 non-null  object
 2   Customer_Type                      129880 non-null  object
 3   Age                                129880 non-null  int64
 4   Type_of_Travel                     129880 non-null  object
 5   Class                              129880 non-null  object
 6   Flight_Distance                    129880 non-null  int64
 7   Inflight_wifi_service              129880 non-null  int64
 8   Departure/Arrival_time_convenient  129880 non-null  int64
 9   Ease_of_Online_booking             129880 non-null  int64
 10  Gate_location                      129880 non-null  int64
 11  Food_and_drink                     129880 non-null  int64
 12  Online_boarding                    129880 non-null  int64
```

```
 13  Seat_comfort                       129880 non-null  int64
 14  Inflight_entertainment             129880 non-null  int64
 15  On-board_service                   129880 non-null  int64
 16  Leg_room_service                   129880 non-null  int64
 17  Baggage_handling                   129880 non-null  int64
 18  Checkin_service                    129880 non-null  int64
 19  Inflight_service                   129880 non-null  int64
...
 22  Arrival_Delay_in_Minutes           129487 non-null  float64
 23  satisfaction                       129880 non-null  object
dtypes: float64(1), int64(18), object(5)
memory usage: 23.8+ MB
```

```
The literature review will provide an overview of the factors that have
been found to affect customer satisfaction in the airline industry:
```

1. **Gender**: Gender of the passengers (Female, Male)
2. **Customer Type**: The customer type (Loyal customer, disloyal customer)
3. **Age**: The actual age of the passengers
4. **Type of Travel**: Purpose of the flight of the passengers (Personal Travel, Business Travel)
5. **Class**: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6. **Flight distance**: The flight distance of this journey
7. **Inflight wifi service**: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
8. **Departure/Arrival time convenient**: Satisfaction level of Departure/Arrival time convenient
9. **Ease of Online booking**: Satisfaction level of online booking
10. **Gate location**: Satisfaction level of Gate location
11. **Food and drink**: Satisfaction level of Food and drink
12. **Online boarding**: Satisfaction level of online boarding
13. **Seat comfort**: Satisfaction level of Seat comfort
14. **Inflight entertainment**: Satisfaction level of inflight entertainment

15. **On-board service**: Satisfaction level of On-board service
16. **Leg room service**: Satisfaction level of Leg room service
17. **Baggage handling**: Satisfaction level of baggage handling
18. **Check-in service**: Satisfaction level of Check-in service
19. **Inflight service**: Satisfaction level of inflight service
20. **Cleanliness**: Satisfaction level of Cleanliness
21. **Departure Delay in Minutes**: Minutes delayed when departure
22. **Arrival Delay in Minutes**: Minutes delayed when Arrival
23. **Satisfaction**: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

3 Missing values

There are not many missing values in our data set. Only 0.03 % of the data in the Arrival_Delay_in_Minutes column were missing.

```
id 0.0 % missing values
Gender 0.0 % missing values
Customer_Type 0.0 % missing values
Age 0.0 % missing values
Type_of_Travel 0.0 % missing values
Class 0.0 % missing values
Flight_Distance 0.0 % missing values
Inflight_wifi_service 0.0 % missing values
Departure/Arrival_time_convenient 0.0 % missing values
Ease_of_Online_booking 0.0 % missing values
Gate_location 0.0 % missing values
Food_and_drink 0.0 % missing values
Online_boarding 0.0 % missing values
Seat_comfort 0.0 % missing values
Inflight_entertainment 0.0 % missing values
On-board_service 0.0 % missing values
Leg_room_service 0.0 % missing values
Baggage_handling 0.0 % missing values
Checkin_service 0.0 % missing values
Inflight_service 0.0 % missing values
Cleanliness 0.0 % missing values
```

```
Departure_Delay_in_Minutes 0.0 % missing values
Arrival_Delay_in_Minutes 0.003 % missing values
satisfaction 0.0 % missing values
```

Since there are not many missing values, removing them completely will not affect our data set or further interpretation of our models.

4 Outliers

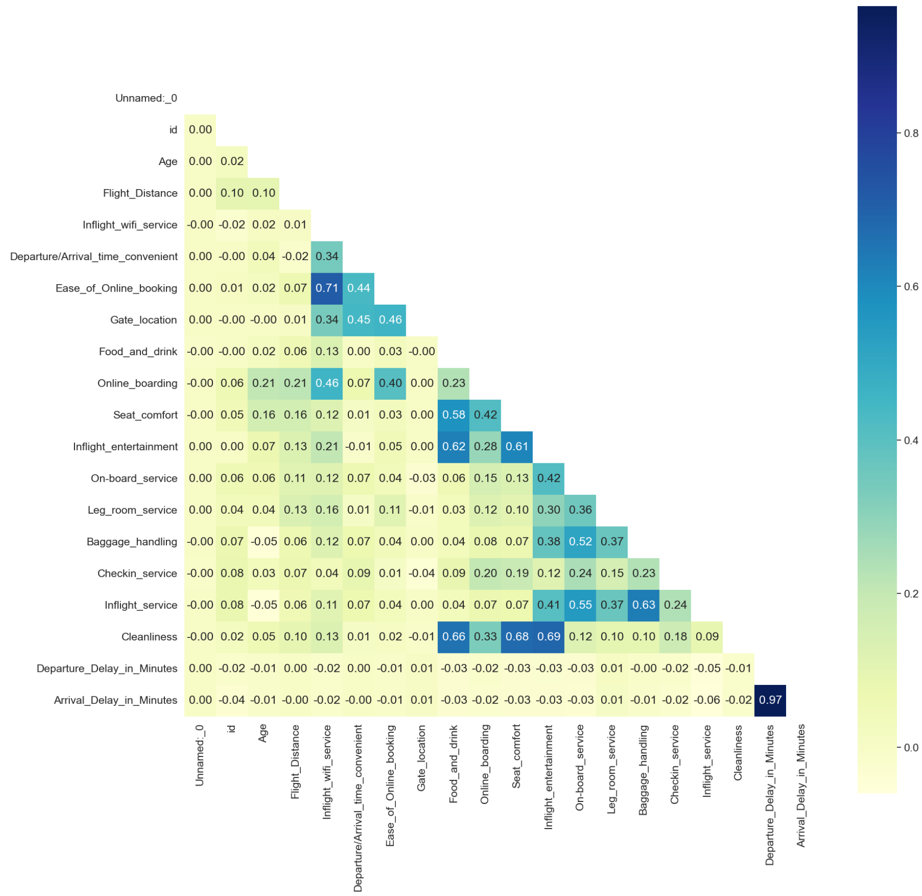
```
{'Unnamed:_0': 0.0,
'id': 0.0,
'Age': 0.0,
'Flight_Distance': 2.198182938096705,
'Inflight_wifi_service': 0.0,
'Departure/Arrival_time_convenient': 0.0,
'Ease_of_Online_booking': 0.0,
'Gate_location': 0.0,
'Food_and_drink': 0.0,
'Online_boarding': 0.0,
'Seat_comfort': 0.0,
'Inflight_entertainment': 0.0,
'On-board_service': 0.0,
'Leg_room_service': 0.0,
'Baggage_handling': 0.0,
'Checkin_service': 12.402987372959656,
'Inflight_service': 0.0,
'Cleanliness': 0.0,
'Departure_Delay_in_Minutes': 13.9344009855251,
'Arrival_Delay_in_Minutes': 13.467816445950106}
```

We chose to remove outliers as we have an abundance of data to spare.

5 Exploratory Data Analysis

5.1 Correlation Analysis

Inflight_wifi_service is very correlated with Ease_of_Online_booking. The ratings for Cleanliness are also correlated to the ratings of Food_and_drink, Seat_Comfort, and Inflight_entertainment. But the two features that are highly correlated are the Departure_Delay_in_Minutes and the Arrival_Delay_in_Minutes, which is very obvious, logically speaking. And on the right we filter down to 4 numerical features instead.



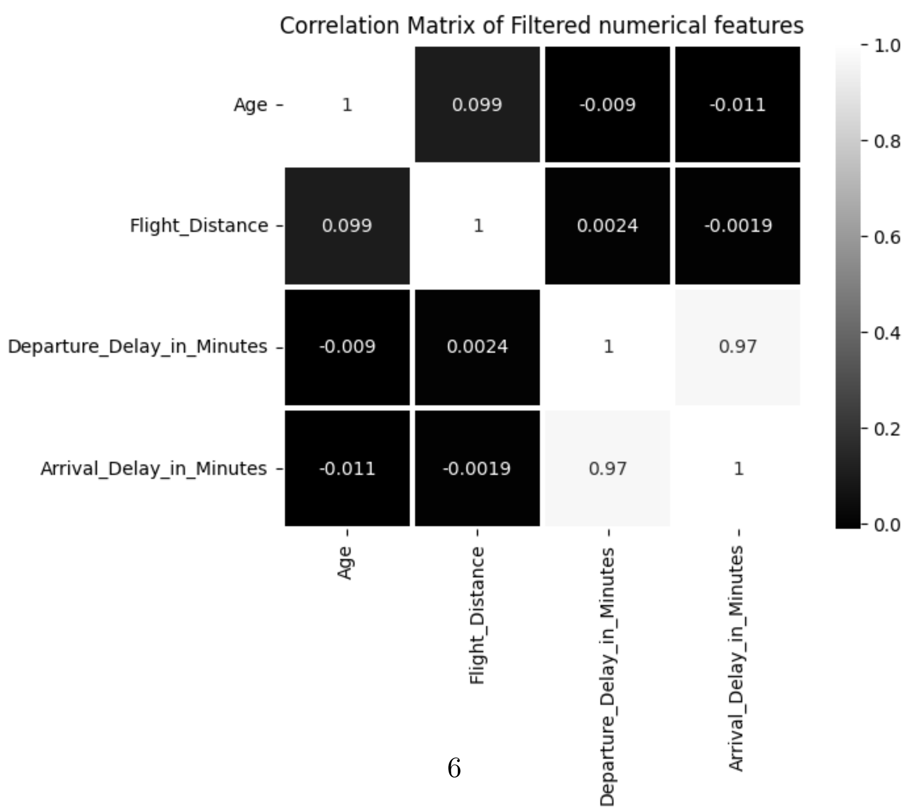


Figure 1: Example Image

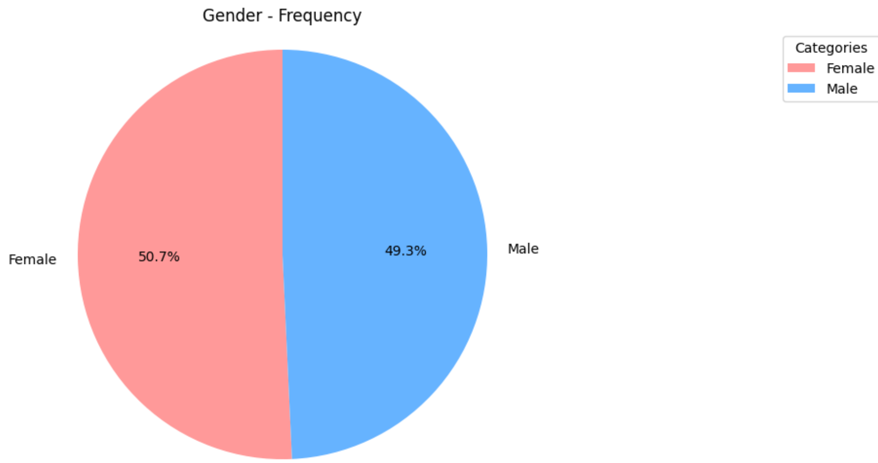


Figure 2: Example Image

Inflight_wifi_service is very correlated with Ease_of_Online_booking. The ratings for Cleanliness are also correlated to the ratings of Food_and_drink, Seat_Comfort, and Inflight_entertainment. But the two features that are highly correlated are the Departure_Delay_in_Minutes and the Arrival_Delay_in_Minutes, which is very obvious, logically speaking. And we will see later on that most of our machine learning algorithms will handle this multicollinearity by applying regularized methods.

5.2 Univariate Analysis

In this section, we will plot each feature with respect to its frequency. Pie charts have been selected as the plotting method. Here are some figures.

5.3 Feature Selection

We used 8 feature selection methods: **chi**$_{s}quaredtest, Wrappermethod(randomforest), decisiontreeandpermu$

1. **chi_square = ['Customer_Type', 'Type_of_Travel', 'Class', 'Inflight_wifi_service', 'Online_boarding', 'Seat_comfort', 'Inflight_entertainment', 'On_board_service', 'Leg_room_service', 'Cleanliness']**

2. **Decision Tree = ['Online_boarding', 'Inflight_wifi_service', 'Type_of_Travel']**

3. **Random Forest = ['Online_boarding', 'Inflight_wifi_service', 'Type_of_Travel', 'Class', 'Inflight_entertainment', 'Seat_comfort', 'Flight_Distance', 'Customer_Type', 'Ease_of_Online_booking', 'On_board_service']**

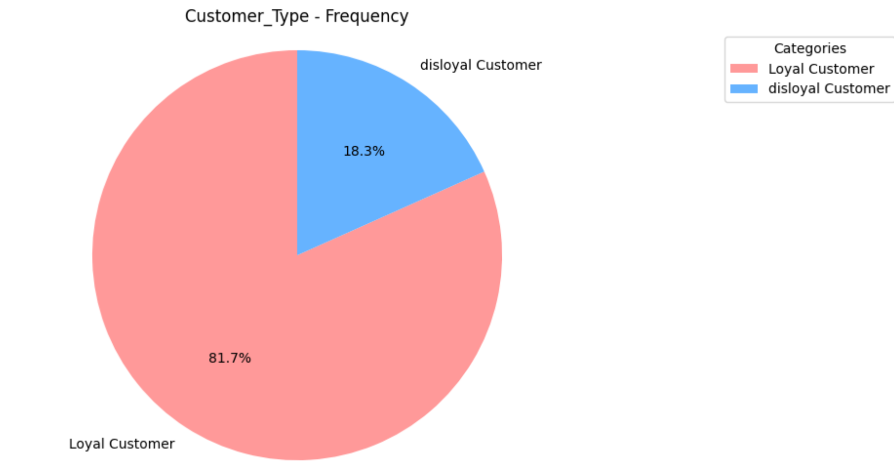


Figure 3: Example Image

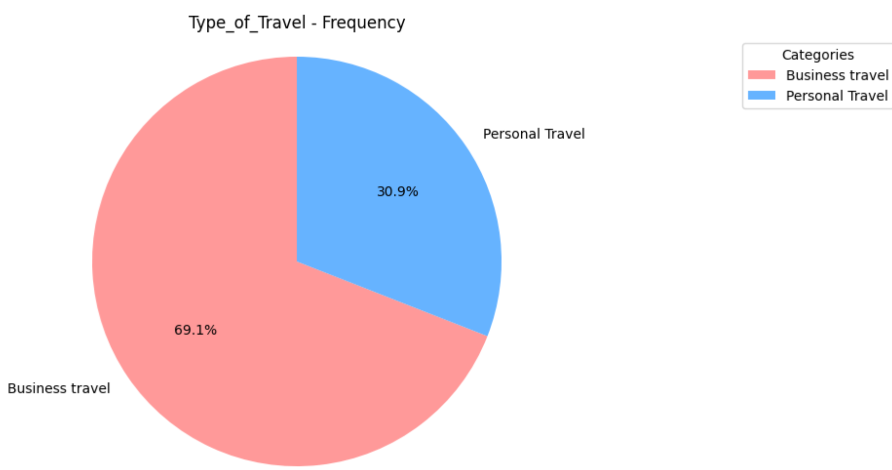


Figure 4: Example Image

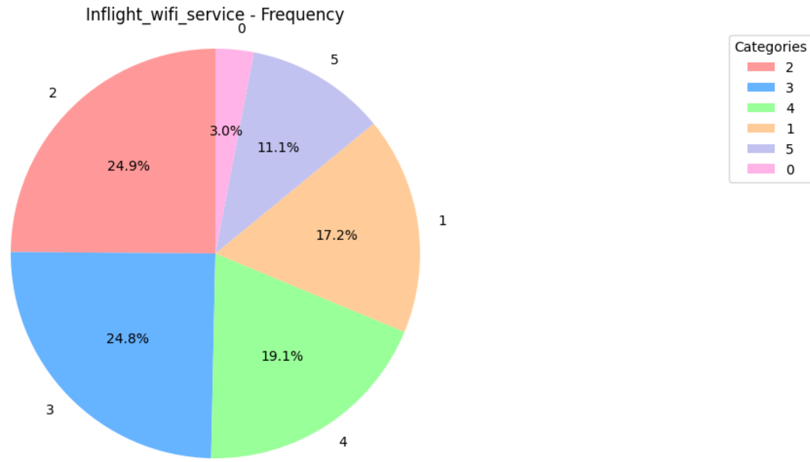


Figure 5: Example Image

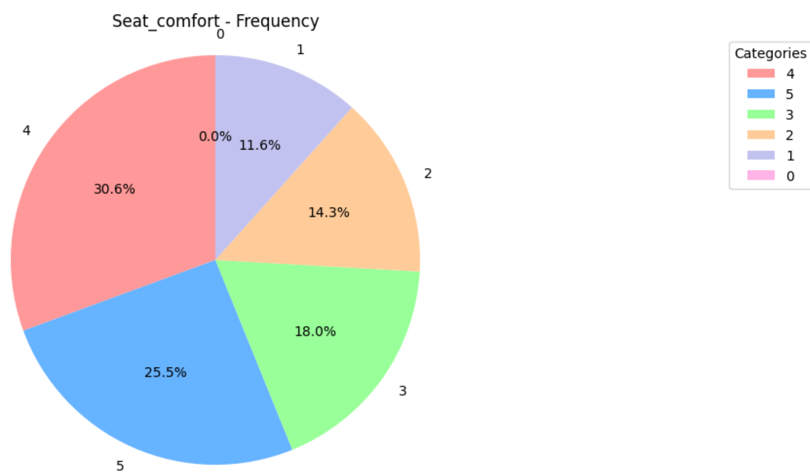


Figure 6: Example Image

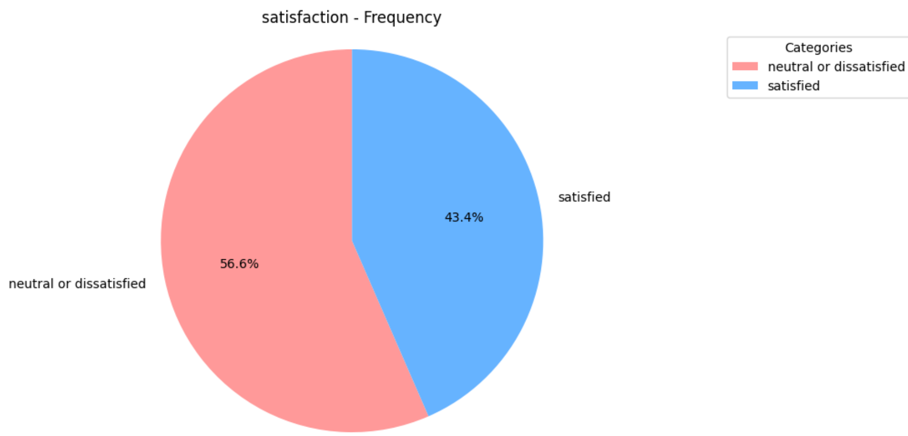


Figure 7: Example Image

4. **Permutation = ['Inflight_wifi_service', 'Type_of_Travel', 'Customer_Type', 'Online_boarding', 'Class', 'Checkin_service', 'Seat_comfort', 'Baggage_handling', 'Inflight_service', 'Cleanliness']**

5. **RFE = ['Customer_Type', 'Type_of_Travel', 'Class', 'Inflight_wifi_service', 'Ease_of_Online_booking', 'Online_boarding', 'On-board_service', 'Leg_room_service', 'Checkin_service', 'Cleanliness']**

6. **Backward Selection = ['Customer_Type', 'Type_of_Travel', 'Class', 'Inflight_wifi_service', 'Ease_of_Online_booking', 'Online_boarding', 'On-board_service', 'Leg_room_service', 'Checkin_service', 'Cleanliness']**

7. **Forward Selection = ['Customer_Type', 'Type_of_Travel', 'Inflight_wifi_service', 'Gate_location', 'Online_boarding', 'Inflight_entertainment', 'On-board_service', 'Leg_room_service', 'Checkin_service', 'Cleanliness']**

8. **Stepwise Selection = ['Customer_Type', 'Type_of_Travel', 'Class', 'Inflight_wifi_service', 'Departure/Arrival_time_convenient', 'Online_boarding', 'Inflight_entertainment', 'On-board_service', 'Leg_room_service', 'Checkin_service']**

6 Evaluation metrics used

As for our evaluation metrics, we incorporated ROC curve, confusion matrix, and AUC(Area Under Curve). We give much emphasis on the AUC value as it will also include information about the True Positive Rate and True Negative Rate

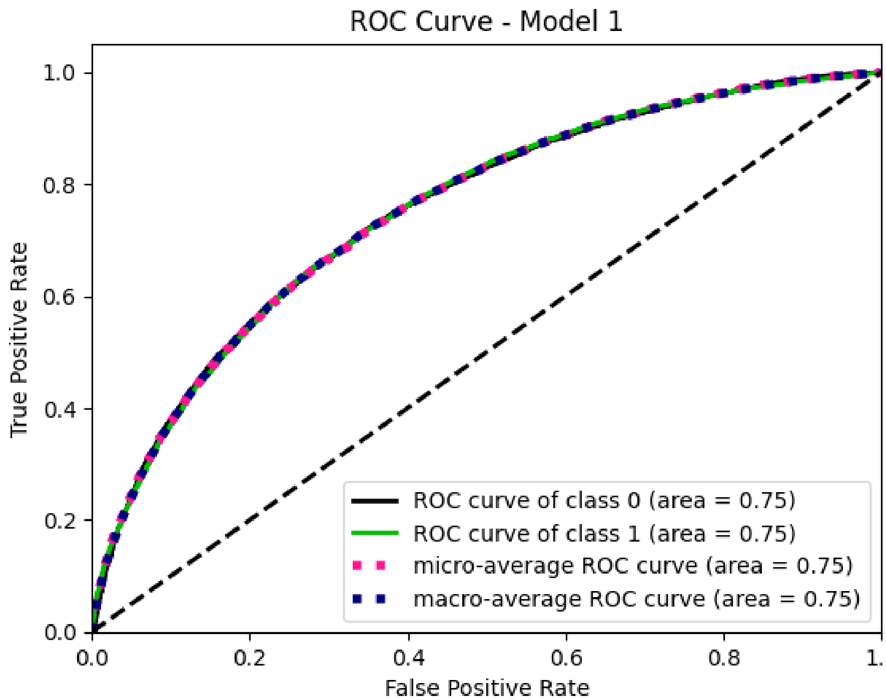


Figure 8: Example Image

7 Results and Discussion

Below are the plots of the ROC curves with the mentioned evaluation metrics of our models:

Model 1: AUC = 0.7506861718066697 Model 2: AUC = 0.9267654577702052 Model 3: AUC = 0.9079386048955072 Model 4: AUC = 0.6412692385246368 Model 5: AUC = 0.9459782563100856 Model 6: AUC = 0.9935930212658386

8 Conclusion

In conclusion, the random forest classifier surpasses all other models in this training instance. And for important factors regarding logistic regression, we can say that ['Customer$_Type','Type_of_Travel','Class','Inflight_wifi_service','Departure/Arrival_time_conve$
$board_service','Leg_room_service','Checkin_service']arethemostimportantfactorsinpredictingairlinecustomers$

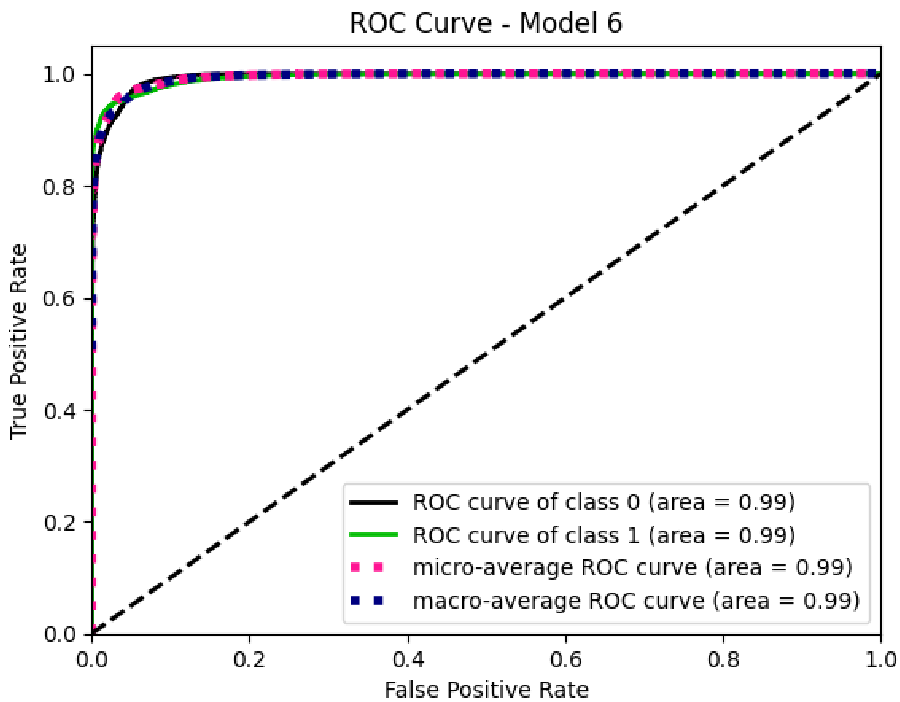


Figure 9: Example Image

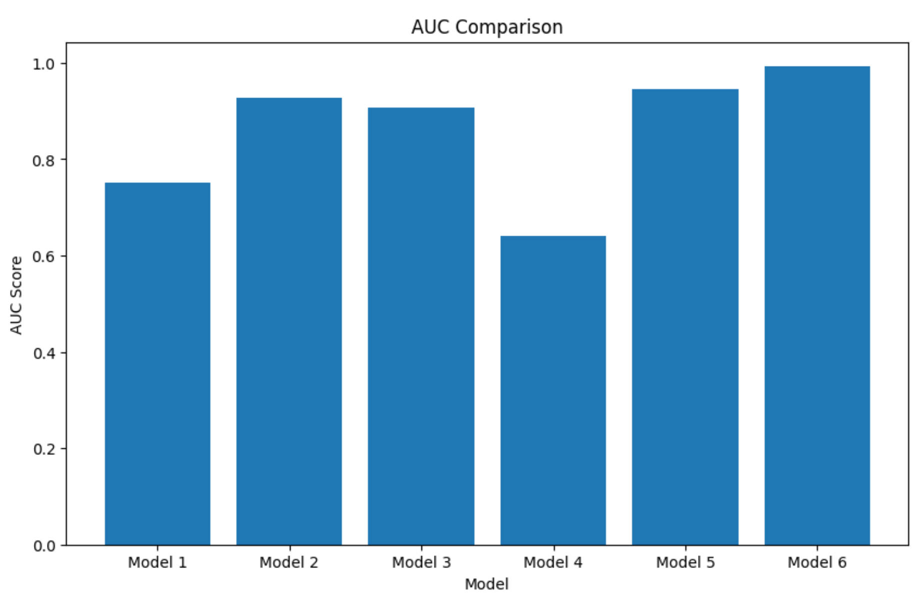


Figure 10: Example Image

9 Further work

We can certainly create pipelines to ease the process and reduce time to rewrite code. We can also tune the hyper-parameter(meta-parameter) of our regularized methods. We can also train more machine learning algorithms such as Ada Gradient Boosting which is a famous and powerful classification model.

References

[1] History of Robots.https://www.roboticsacademy.com.au/history-of-robots/

[2] Development and history, https://www.robotpark.com/History-of-Robotics

[3] Robotnik, https://robotnik.eu/history-of-robots-and-robotics/

[4] A light read from Standford https://cs.stanford.edu/people/eroberts/courses/soco/projects/1998-99/robotics/history.html

[5] Wikipedia Robotics https://en.wikipedia.org/wiki/Robotics