Hw2 Streamlit App Report

zhouyanl@andrew.cmu.edu

1. Motivation and Goals of the Task

With the boom of globalization and economics, taking airflight is becoming more and more common in people's daily life. As a result, people begin to demand higher quality of air travels. To meet customers' needs and provide better service, big airline companies are trying to carry out user study to learn what features influence people's choice of satisfaction to this airflight. This task is designed to help these companies to better understand customers' preferences and the goal is to extract features from passenger satisfaction survey to predict if the customer is satisfied with their airflight.

2. Dataset Introduction and Hypothesis Generation

In this task, I use a dataset from Kaggle (https://www.kaggle.com/johnddddddd/customer-satisfaction), which contains 129880 rows and 24 columns. An introduction to each column and its value is listed below:

Satisfaction: Airline satisfaction level. (Satisfaction, neutral or dissatisfaction)

Age: The actual age of the passengers.

Gender: Gender of the passengers. (Female, Male)

Type of Travel: Purpose of the flight of the passengers. (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Customer Type: The customer type. (Loyal customer, disloyal customer)

Flight distance: The flight distance of this journey.

Inflight wifi service: Satisfaction level of the inflight wifi service. (0:Not Applicable;1-5)

Ease of Online booking: Satisfaction level of online booking.

Inflight service: Satisfaction level of inflight service.

Online boarding: Satisfaction level of online boarding.

Inflight entertainment: Satisfaction level of inflight entertainment.

Food and drink: Satisfaction level of Food and drink.

Seat comfort: Satisfaction level of Seat comfort.

On-board service: Satisfaction level of On-board service.

Leg room service: Satisfaction level of Leg room service.

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient.

Baggage handling: Satisfaction level of baggage handling.

Gate location: Satisfaction level of Gate location.

Cleanliness: Satisfaction level of Cleanliness.

Check-in service: Satisfaction level of Check-in service.

Departure Delay in Minutes: Minutes delayed when departure.

Arrival Delay in Minutes: Minutes delayed when Arrival.

Flight cancelled: Whether the Flight cancelled or not. (Yes, No)

Flight time in minutes: Minutes of Flight takes.

After analysis each feature, I proposed the hypotheses that these nine features are most likely to influence customers' satisfaction with the flight:

- Customer Type: Loyal Customers are more likely to be satisfied with the flight.
- Class: The higher the travel class is, the more likely the customers will be satisfied.
- Flight Distance: The longer the flight distance is, the less likely the customers will be satisfied.
- Seat Comfort: The higher the passengers rank the comfort, the more likely the customers will be satisfied.
- Inflight Wifi Service: The higher the passengers rank the service, the more likely the customers will be satisfied.
- Inflight Entertainment: The higher ranking the passengers give, the more likely the customers will be satisfied.

- Cleanliness: The higher ranking the passengers give, the more likely the customers will be satisfied.
- Departure/Arrival Delay in minutes: More time wasted on delay in departure or arrival will probably lead to customers' dissatisfaction.

3. Data Pre-processing and Feature Extraction

After setting up our task and hypotheses, it's time to preprocess the data and explore it. Let's first get an overview of the imported data sets by using head() and describe() :

	id	satisfaction_v2	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service	Inflight entertainment	Online support
0	11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco	265	0	0	0	2	2	4	2
1	110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business	2464	0	0	0	3	0	2	2
2	103199	satisfied	Female	Loyal Customer	15	Personal Travel	Eco	2138	0	0	0	3	2	0	2
3	47462	satisfied	Female	Loyal Customer	60	Personal Travel	Eco	623	0	0	0	3	3	4	3
4	120011	satisfied	Female	Loyal Customer	70	Personal Travel	Eco	354	0	0	0	3	4	3	4

	id	Age	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service	Inflight entertainment	12
count	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	
mean	64940.500000	39.427957	1981.409055	2.838597	2.990645	2.851994	2.990422	3.249130	3.383477	
std	37493.270818	15.119360	1027.115606	1.392983	1.527224	1.443729	1.305970	1.318818	1.346059	
min	1.000000	7.000000	50.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	32470.750000	27.000000	1359.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	
50%	64940.500000	40.000000	1925.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.000000	
75%	97410.250000	51.000000	2544.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	
max	129880.000000	85.000000	6951.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	

Since our hypotheses only focus on nine features, we extract them and discard the rest columns. After this, we first try to find out if there're missing values in the dataset:

```
missing_values = temp_data.isna().sum()
print(missing_values)
```

```
Customer Type                      0
Class                              0
Flight Distance                    0
Seat Comfort                       0
Inflight Wifi Service              0
Inflight Entertainment             0
Cleanliness                        0
Departure Delay in Minutes         0
Arrival Delay in Minutes         393
Satisfaction                       0
dtype: int64
```

After removing missing values, we try to discard some outliers and map some string values in datasets into numerical:

```
clean_data['Customer Type']= clean_data['Customer Type'].map({'Loyal Customer':1, 'disloyal Customer':0})
clean_data['Class']= clean_data['Class'].map({'Eco':0, 'Eco Plus':1, 'Business':2})
clean_data['Satisfaction']= clean_data['Satisfaction'].map({'satisfied':1, 'neutral or dissatisfied':0})
clean_data.nunique()
clean_data.head()
```

	Customer Type	Class	Flight Distance	Seat Comfort	Inflight Wifi Service	Inflight Entertainment	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
90120	1	2	2719	4	4	4	2	0	0	1
70120	1	2	324	1	1	3	1	73	59	0
67781	1	2	1527	1	5	3	1	0	16	0
18868	1	2	895	1	5	5	4	6	2	1
55550	0	0	1755	3	1	3	2	7	0	0

After feature extraction and data preprocessing, we get a dataset containing 115020 rows and 10 columns.

4. Model Building

Now it’s time to split the dataset into training data and testing data. Here I choose 0.8 as the ratio to split the dataset. And then we need to choose a machine learning model for this task.

Since it's a binary classification task, I believe random forest is a good choice because it's fast and easy to use. In this task I set max_depth = 7 to prevent the model from overfitting after some experiments for tuning.

Then we can get the test error by our training model:

```
pred_label = model.predict(test_data)
accuracy_score(test_label,pred_label)
```

```
0.8847
```

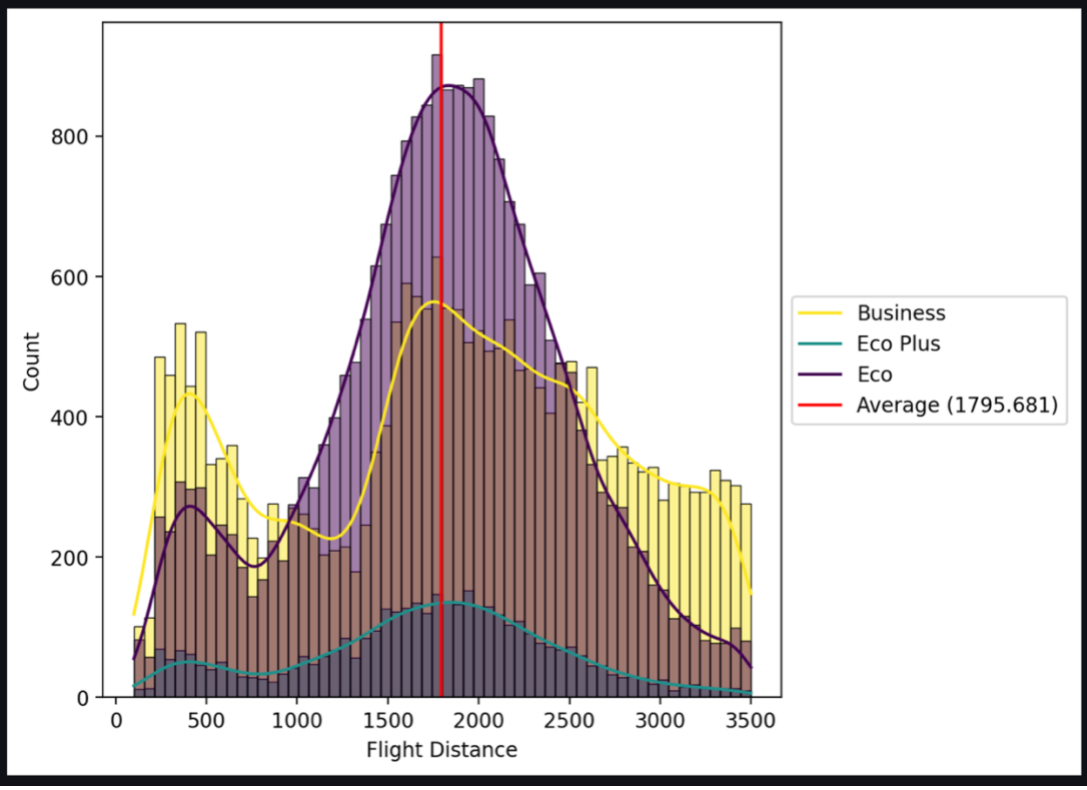
This is a good model for predicting passengers' satisfaction towards the air flight. We save it to file and deploy it later.

5. Model Deployment with Streamlit

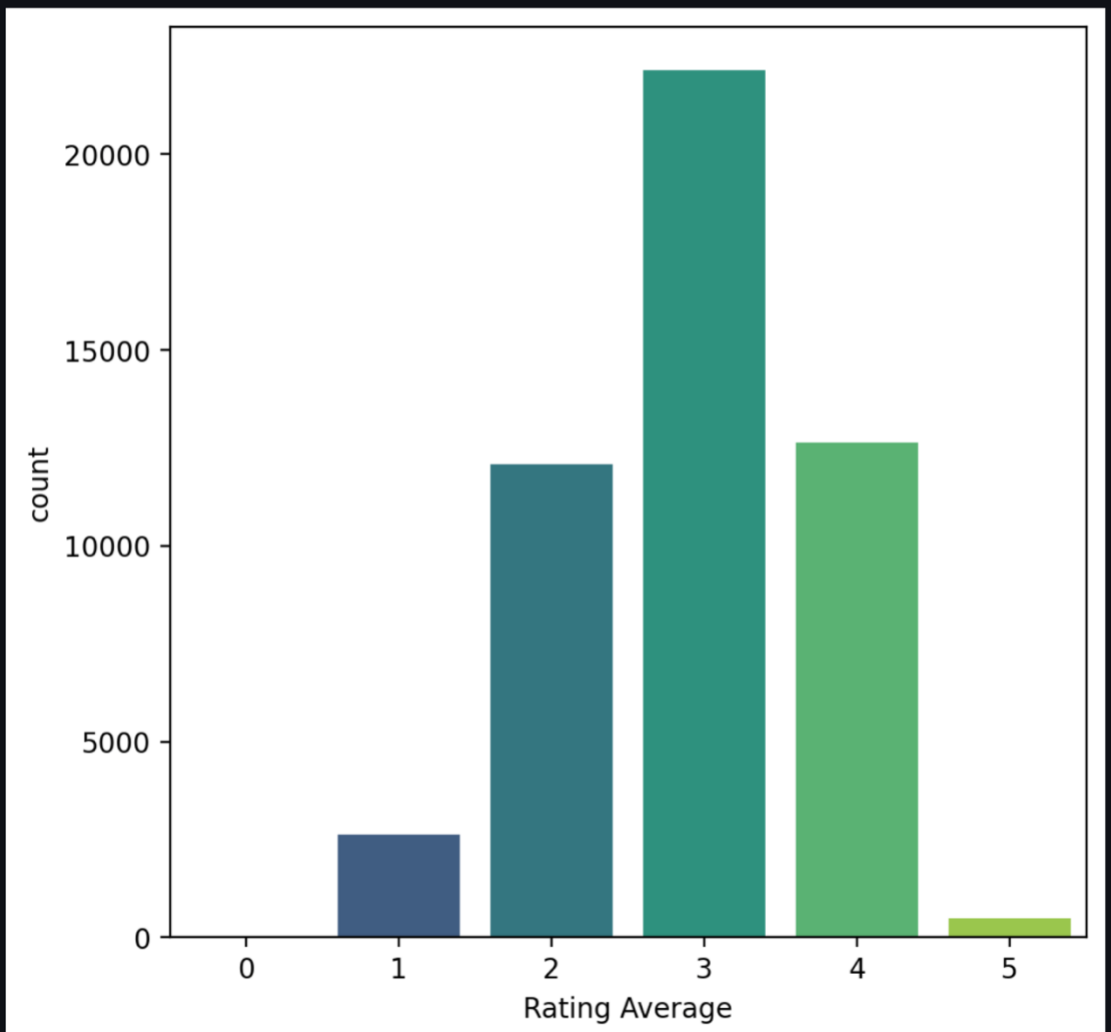
Now it's time to use streamlit to build our app on the website. For the app on website, I included two parts in it:

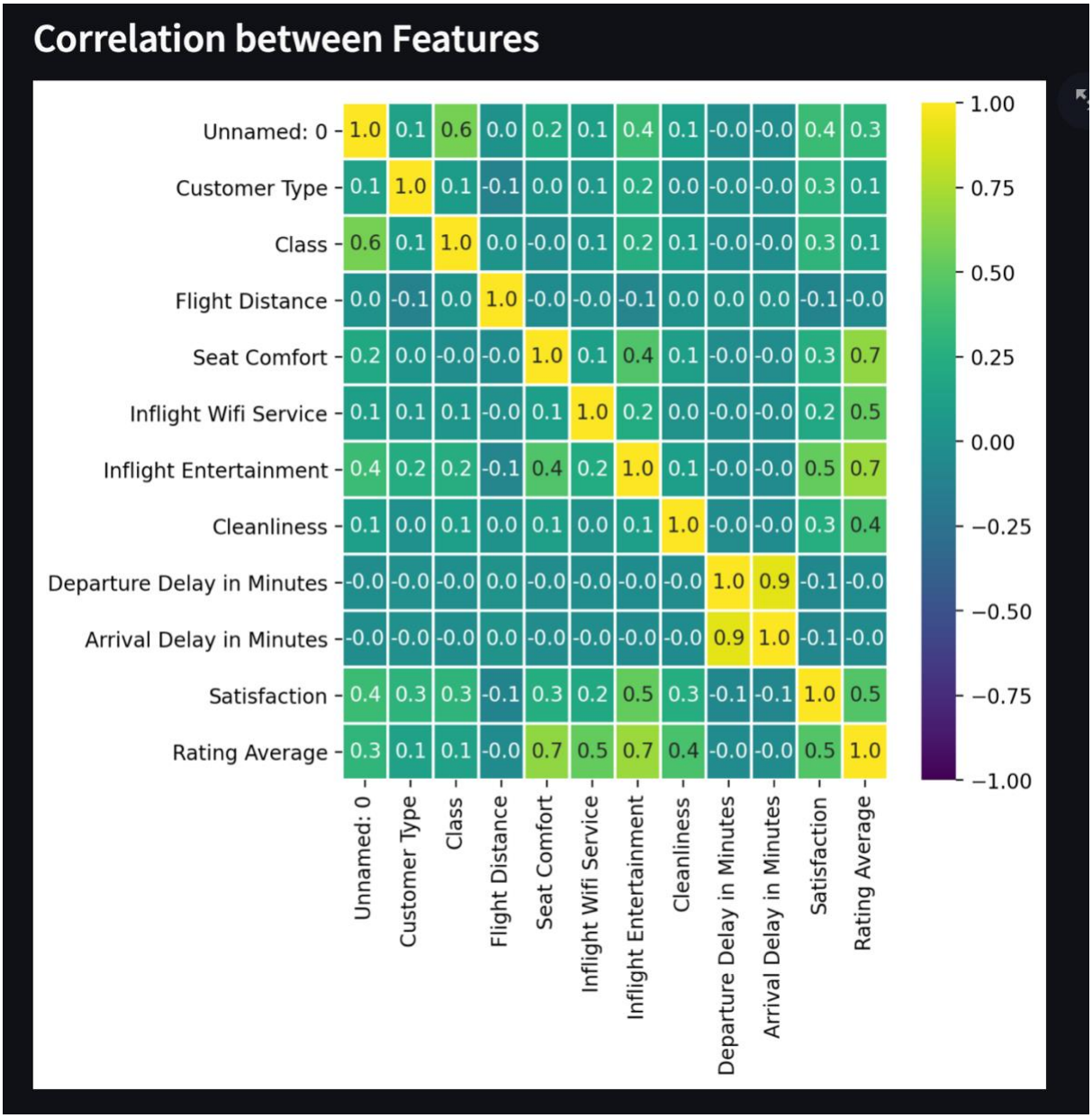
For the first part, I include three plots in it to help users understand the relationship between each features and passengers' rankings and satisfactions. The plots are listed below:

The relationship between Flight Distance and Class

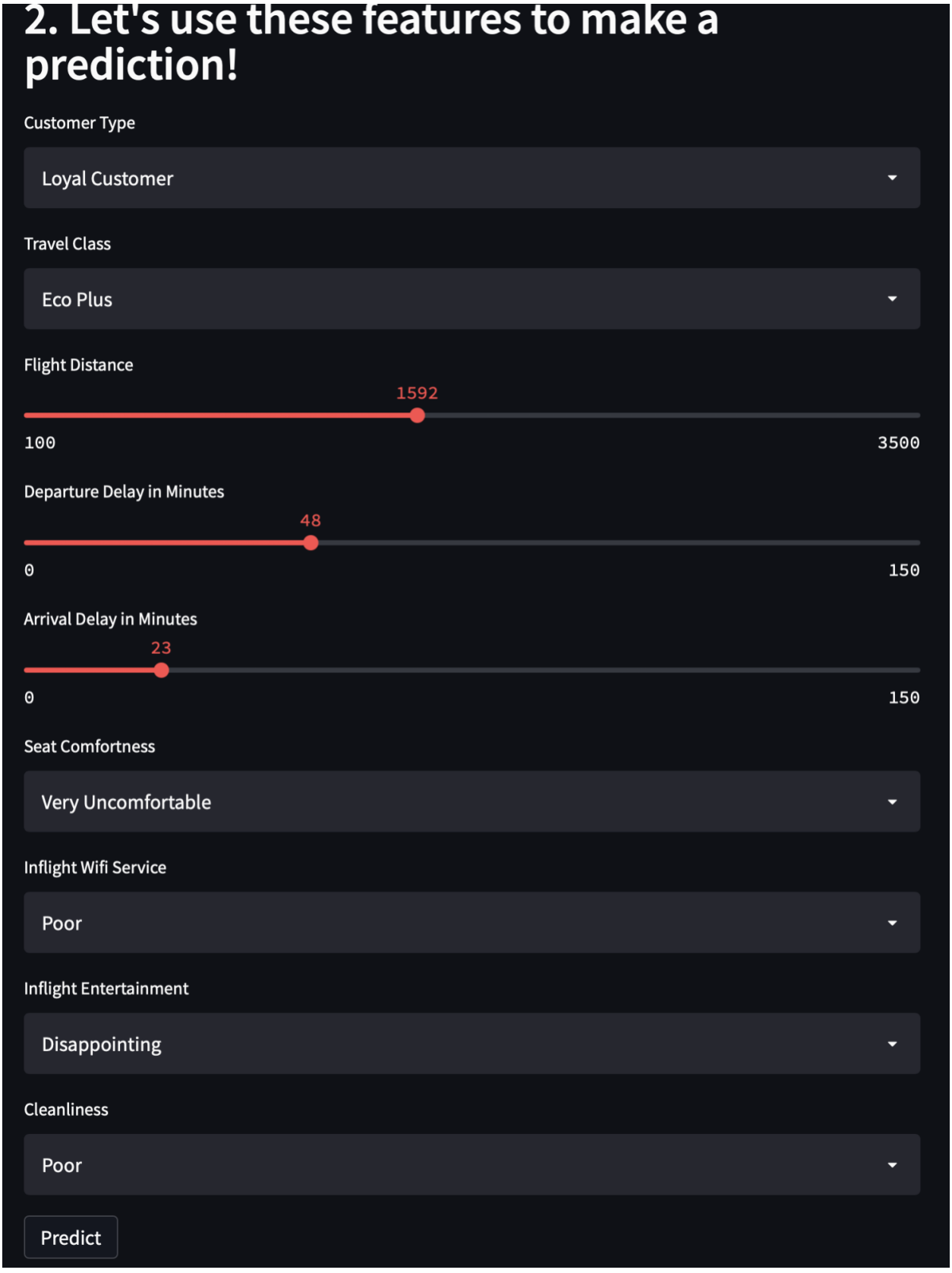


Passenger Average Rating Distribution





For the second part, we ask the users to input values for each feature and then we use these values with our models to generate a prediction.


2. Let's use these features to make a prediction!
Customer Type
Loyal Customer
Travel Class
Eco Plus
Flight Distance
1592
100
3500
Departure Delay in Minutes
48
0
150
Arrival Delay in Minutes
23
0
150
Seat Comfortness
Very Uncomfortable
Inflight Wifi Service
Poor
Inflight Entertainment
Disappointing
Cleanliness
Poor
Predict