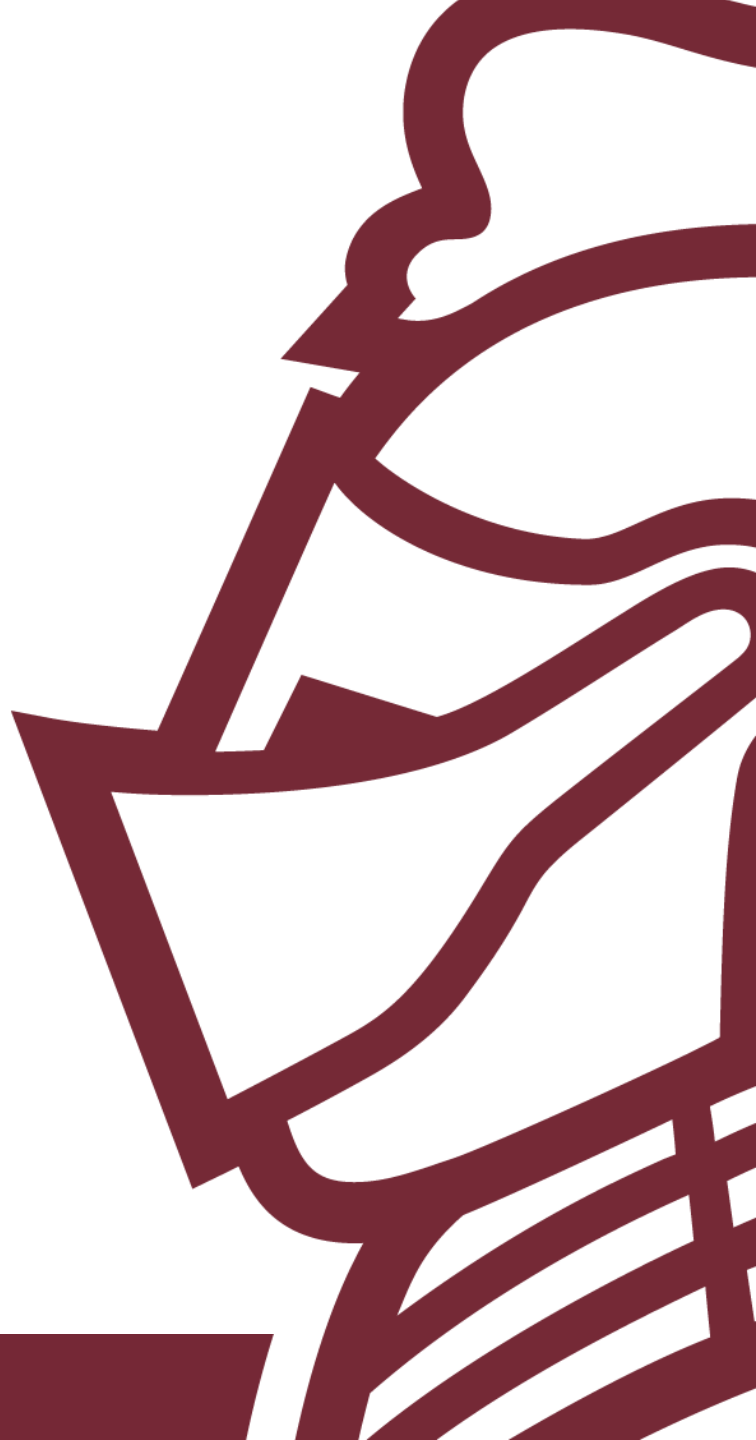
Airline Customer Satisfaction

~

Logistic Regression

By: Robert Pearson



Introduction

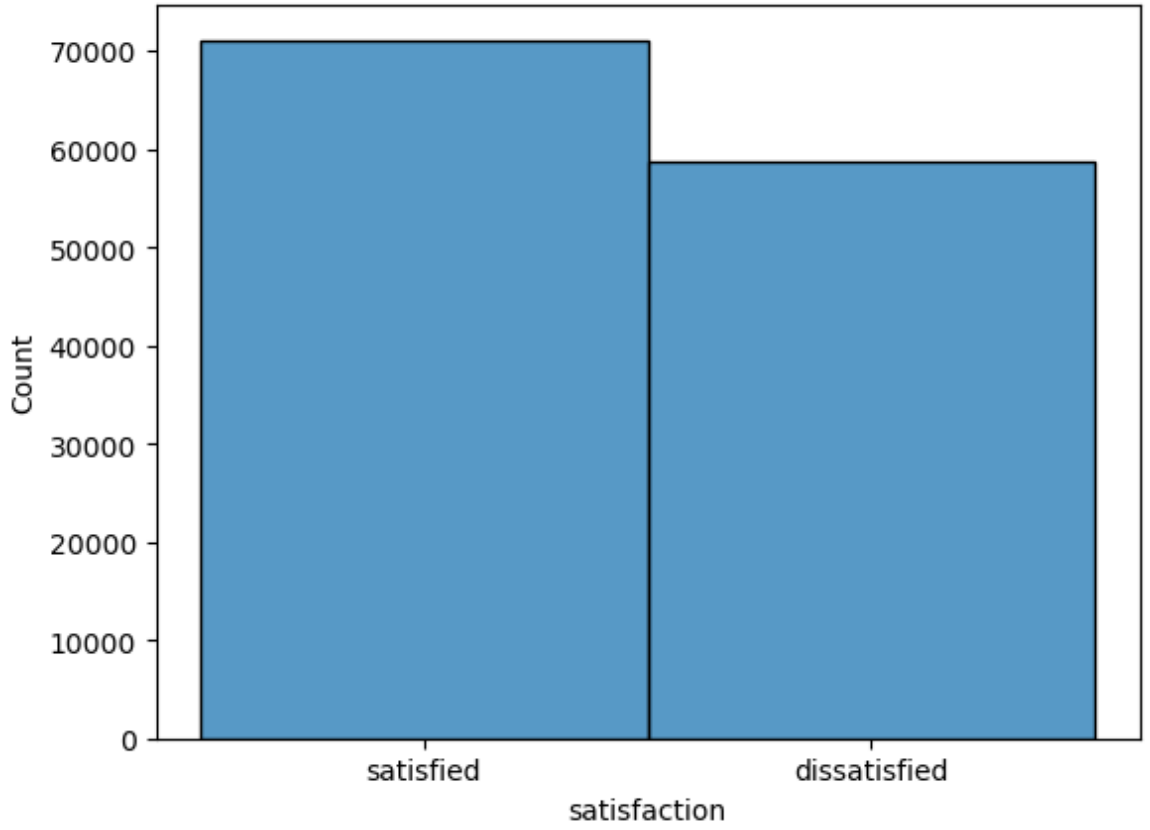
- Goal: Predict if a customer is satisfied with their flight
- Dataset is *Customer Satisfaction in Airline*
  - Survey data taken from an undisclosed airline company.
  - Focused on measuring an airline's passenger satisfaction with their flight.
  - Passengers responded by ranking a number of variables on a scale of 1-5 or responding to certain flight related questions like delay time or class.

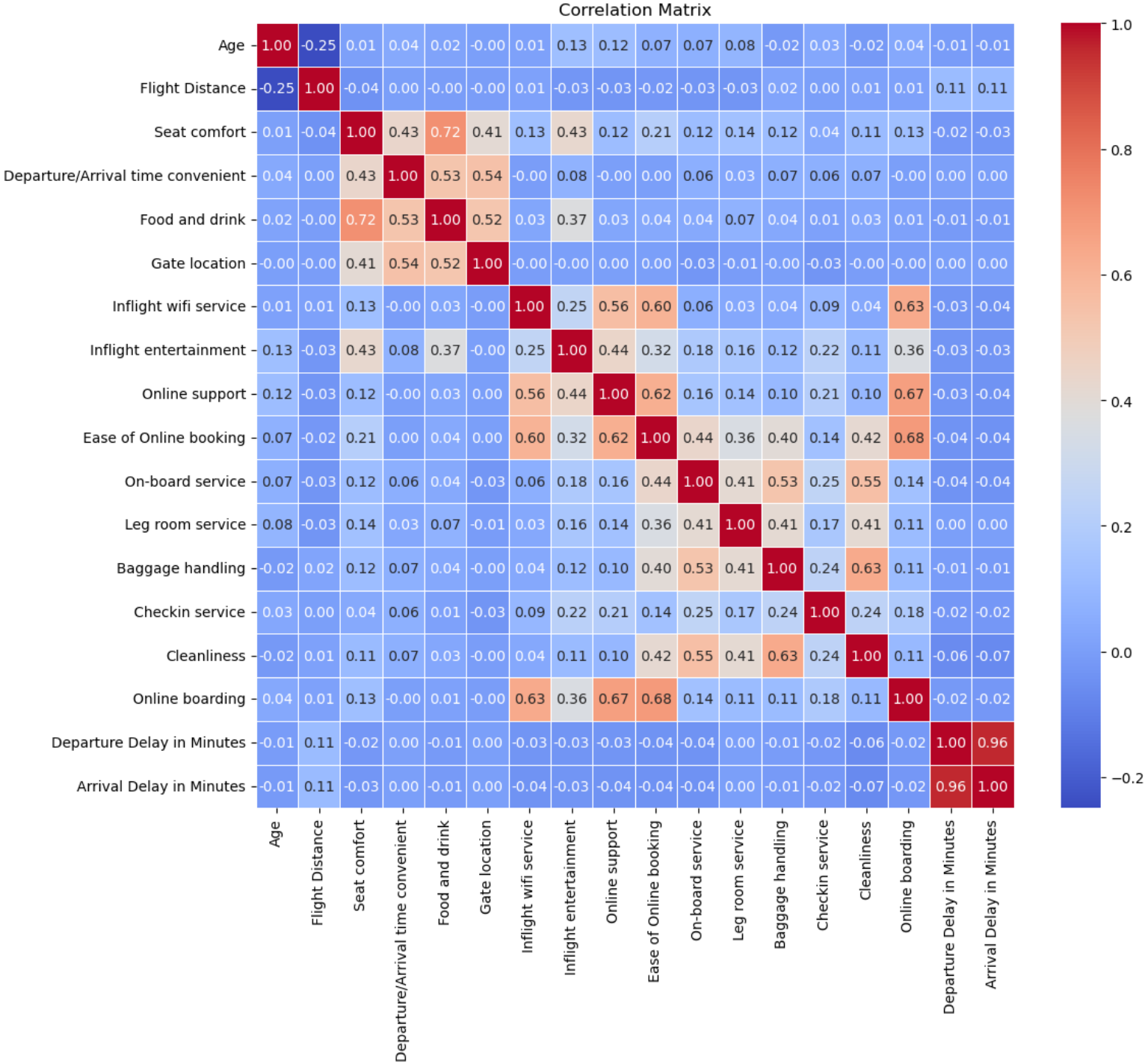
Introduction

- Why is this important?
  - Measuring customer satisfaction with a service is a key component of any business.
  - Airline could have collected the data due to wanting to them wanting to improve the customer’s experience.
  - In an industry, like the airline industry, customer satisfaction plays a big role in customer retention. Airlines want to make customers into lifetime customers and that happens from customers having a good experience on their flight.

Initial Analysis

- 129,880 samples of an airline's customer ratings
- 22 columns of data with a mix of categorical and numerical with ordinal, continuous and nominal data types.
- The dependent variable in this study is satisfaction, this variable's output is either "satisfied" or "dissatisfied."
- Some missing values for the variable "Arrival Delay in Minutes" but was only 393 out of 129,880.




Correlation Matrix
Age 1.00 -0.25 0.01 0.04 0.02 -0.00 0.01 0.13 0.12 0.07 0.07 0.08 -0.02 0.03 -0.02 0.04 -0.01 -0.01
Flight Distance -0.25 1.00 -0.04 0.00 -0.00 -0.00 0.01 -0.03 -0.03 -0.02 -0.03 -0.03 0.02 0.00 0.01 0.01 0.11 0.11
Seat comfort 0.01 -0.04 1.00 0.43 0.72 0.41 0.13 0.43 0.12 0.21 0.12 0.14 0.12 0.04 0.11 0.13 -0.02 -0.03
Departure/Arrival time convenient 0.04 0.00 0.43 1.00 0.53 0.54 -0.00 0.08 -0.00 0.00 0.06 0.03 0.07 0.06 0.07 -0.00 0.00 0.00
Food and drink 0.02 -0.00 0.72 0.53 1.00 0.52 0.03 0.37 0.03 0.04 0.04 0.07 0.04 0.01 0.03 0.01 -0.01 -0.01
Gate location -0.00 -0.00 0.41 0.54 0.52 1.00 -0.00 -0.00 0.00 0.00 -0.03 -0.01 -0.00 -0.03 -0.00 -0.00 0.00 0.00
Inflight wifi service 0.01 0.01 0.13 -0.00 0.03 -0.00 1.00 0.25 0.56 0.60 0.06 0.03 0.04 0.09 0.04 0.63 -0.03 -0.04
Inflight entertainment 0.13 -0.03 0.43 0.08 0.37 -0.00 0.25 1.00 0.44 0.32 0.18 0.16 0.12 0.22 0.11 0.36 -0.03 -0.03
Online support 0.12 -0.03 0.12 -0.00 0.03 0.00 0.56 0.44 1.00 0.62 0.16 0.14 0.10 0.21 0.10 0.67 -0.03 -0.04
Ease of Online booking 0.07 -0.02 0.21 0.00 0.04 0.00 0.60 0.32 0.62 1.00 0.44 0.36 0.40 0.14 0.42 0.68 -0.04 -0.04
On-board service 0.07 -0.03 0.12 0.06 0.04 -0.03 0.06 0.18 0.16 0.44 1.00 0.41 0.53 0.25 0.55 0.14 -0.04 -0.04
Leg room service 0.08 -0.03 0.14 0.03 0.07 -0.01 0.03 0.16 0.14 0.36 0.41 1.00 0.41 0.17 0.41 0.11 0.00 0.00
Baggage handling -0.02 0.02 0.12 0.07 0.04 -0.00 0.04 0.12 0.10 0.40 0.53 0.41 1.00 0.24 0.63 0.11 -0.01 -0.01
Checkin service 0.03 0.00 0.04 0.06 0.01 -0.03 0.09 0.22 0.21 0.14 0.25 0.17 0.24 1.00 0.24 0.18 -0.02 -0.02
Cleanliness -0.02 0.01 0.11 0.07 0.03 -0.00 0.04 0.11 0.10 0.42 0.55 0.41 0.63 0.24 1.00 0.11 -0.06 -0.07
Online boarding 0.04 0.01 0.13 -0.00 0.01 -0.00 0.63 0.36 0.67 0.68 0.14 0.11 0.11 0.18 0.11 1.00 -0.02 -0.02
Departure Delay in Minutes -0.01 0.11 -0.02 0.00 -0.01 0.00 -0.03 -0.03 -0.03 -0.04 -0.04 0.00 -0.01 -0.02 -0.06 -0.02 1.00 0.96
Arrival Delay in Minutes -0.01 0.11 -0.03 0.00 -0.01 0.00 -0.04 -0.03 -0.04 -0.04 -0.04 0.00 -0.01 -0.02 -0.07 -0.02 0.96 1.00
Age
Flight Distance
Seat comfort
Departure/Arrival time convenient
Food and drink
Gate location
Inflight wifi service
Inflight entertainment
Online support
Ease of Online booking
On-board service
Leg room service
Baggage handling
Checkin service
Cleanliness
Online boarding
Departure Delay in Minutes
Arrival Delay in Minutes
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Data Preparation

- Data appeared to be mostly cleaned up already.
- Missing values were imputed with mean value.
- Most of the data was numerical, but the variables “Customer Type”, “Age”, and “Class” required one-hot encoding.

Models

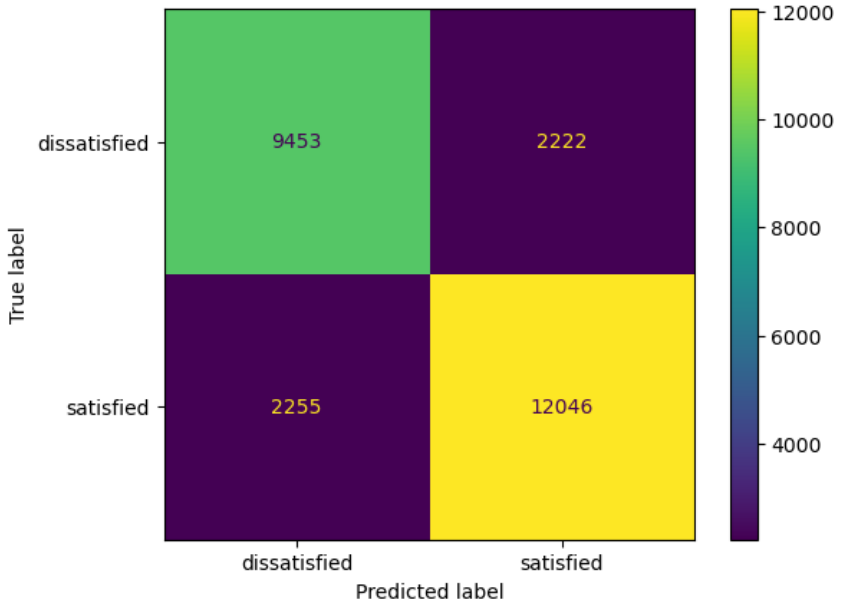
- The selected variables are “*Seat comfort*”, “*Departure/Arrival time convenient*”, “*Food and drink*”, “*Gate location*”, “*Inflight wifi service*”, “*Inflight entertainment*”, “*Online support*”, “*Ease of Online booking*”, “*On-board service*”, “*Leg room service*”, “*Baggage handling*”, “*Check-in service*”, “*Cleanliness*”, and “*Online boarding*”.

Experiment Number	Parameters
1	All twenty-one (21) variables with 80/20 split for train, and test
2	All twenty-one (21) variables with 70/30 split for train, and test
3	Selected variables with 80/20 split for train, and test
4	Selected variables with 70/30 split for train, and test

Results - All 21 variables with 80/20 split and 70/30 split

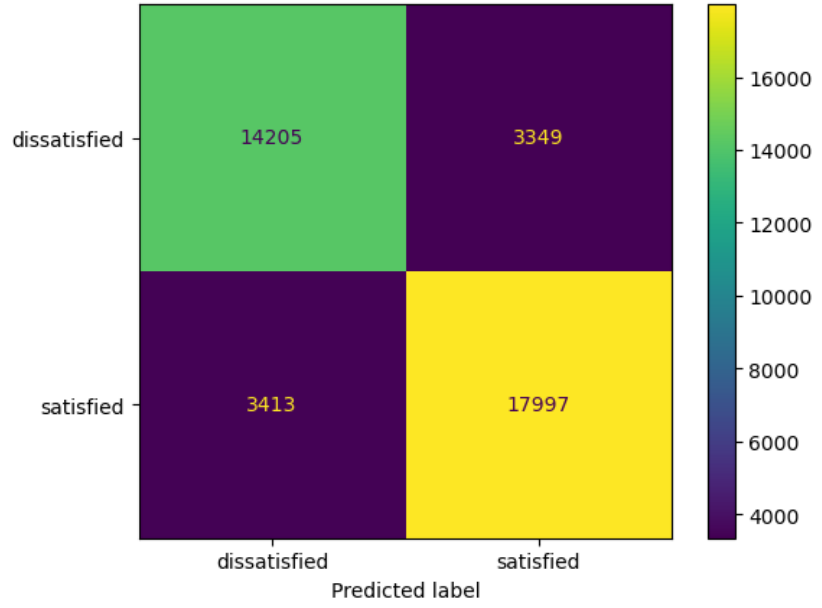
- Model with a 70/30 split has the highest true positives.
- The model with an 80/20 split has the lowest false negatives.
- Both models preformed similarly with difference between true positive, false positive, and true negatives being minimal.

80/20



	precision	recall	f1-score	support
dissatisfied	0.81	0.81	0.81	11675
satisfied	0.84	0.84	0.84	14301

70/30

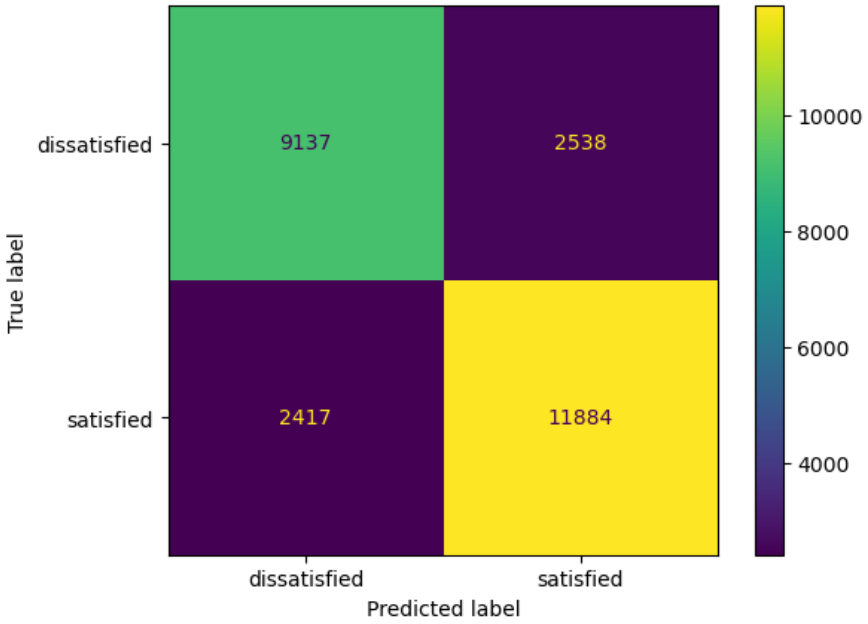


	precision	recall	f1-score	support
dissatisfied	0.81	0.81	0.81	17554
satisfied	0.84	0.84	0.84	21410

Results - Selected variables with 80/20 split and 70/30 split

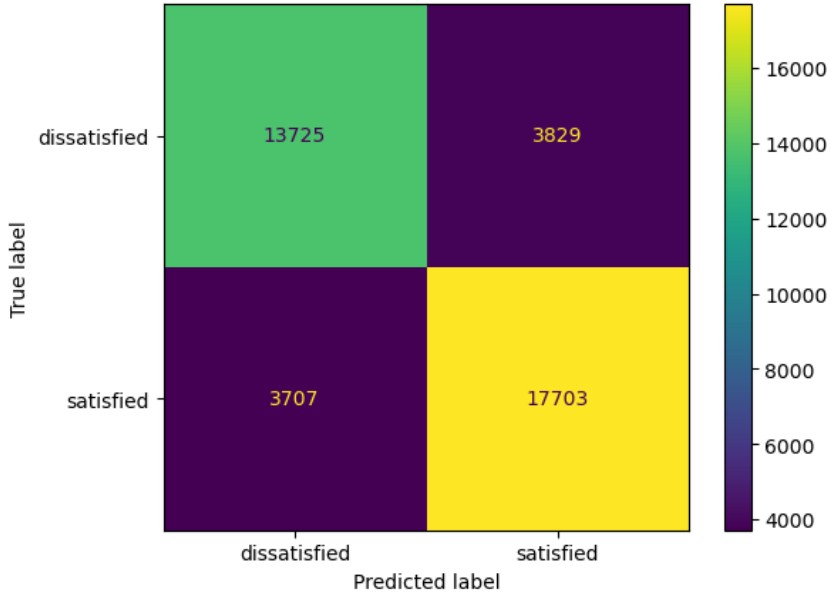
- Both models had a higher false positive count and false negative count compared to their counterparts.
- It misclassified satisfied customers as dissatisfied and vice versa more often compared to the other models.

80/20



	precision	recall	f1-score	support
dissatisfied	0.79	0.78	0.79	11675
satisfied	0.82	0.83	0.83	14301

70/30

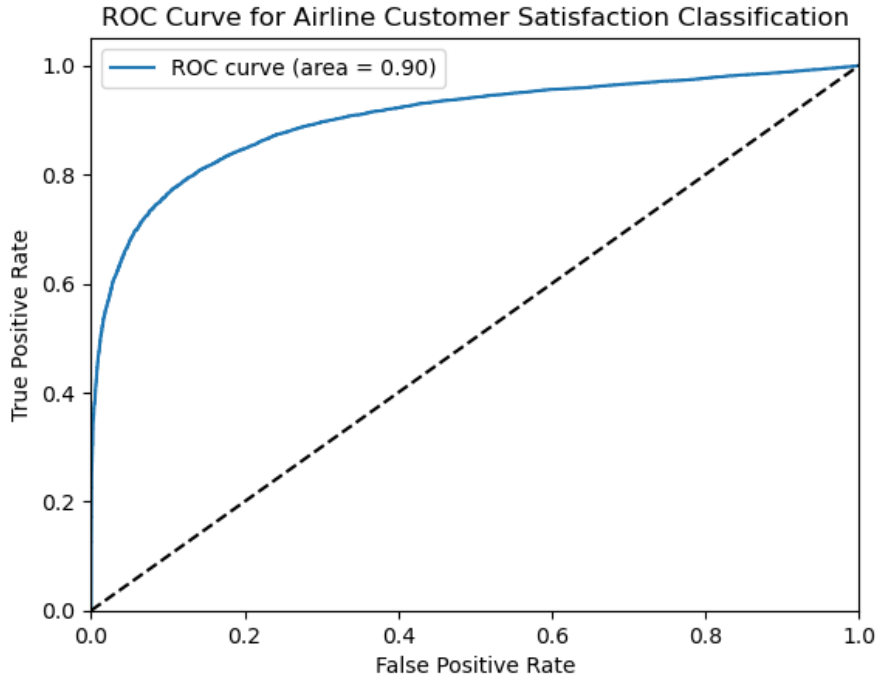


	precision	recall	f1-score	support
dissatisfied	0.79	0.78	0.78	17554
satisfied	0.82	0.83	0.82	21410

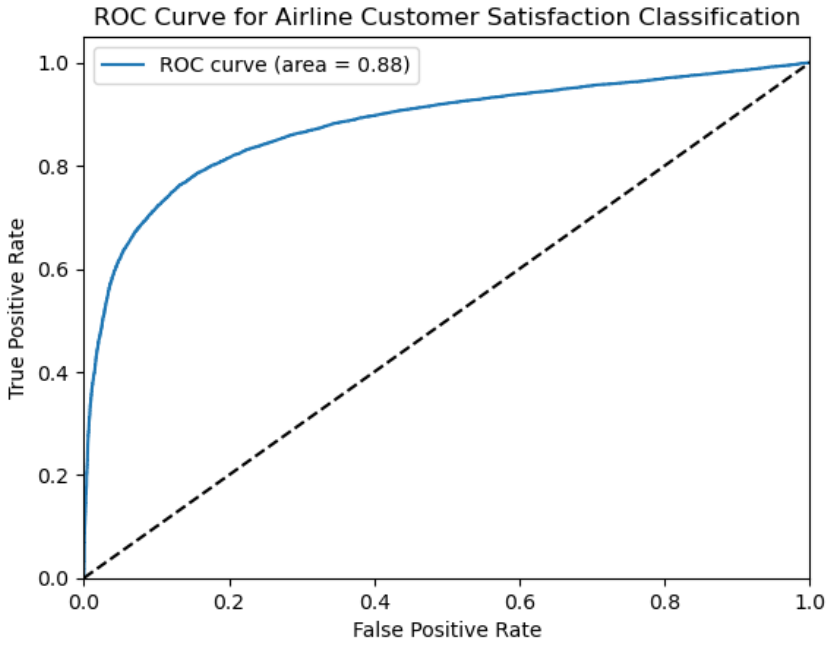
Results

- ROC curve created by plotting the true positive rate against the false positive rate.
- The AUC is the probability that the model will rank a randomly chosen positive example more highly than a randomly chosen negative example.
  - Ranges from 0 to 1, 0.5 = random guessing, and 1 = perfect performance.

- All models can achieve a high true positive rate while keeping the false positive rate relatively low.



Selected variables **21 variables**



Problems Encountered

- Obtaining the data / choosing dataset
- Creating a ROC curve (receiver operating characteristic curve)
- Interpretating the ROC curve and AUC

Future Improvements

- Run more experimenters with different variables selected.
- More in-depth analysis into the relationship between each variables and how it affects a customer being satisfied.
- Test different models

Conclusion

- The models with all twenty-one features only slightly outperformed those with fourteen features.
  - So “Customer Type”, “Age”, “Type of Travel”, “Class”, “Flight Distance”, “Departure Delay in Minutes”, and “Arrival Delay in Minutes” could be somewhat significant.
- The models with twenty-one variables, 70/30 split model better in identifying satisfied customers, the 80/20 split model minimized misclassification of dissatisfied customers.
- The model with all fourteen features and an 70/30 split was the least favorable performer due to lower scores in the classification table.

Dataset

- https://www.kaggle.com/datasets/yakhyojon/customer-satisfaction-in-airline