Airline Customer Satisfaction

~

Logistic Regression

By: Robert Pearson

Introduction

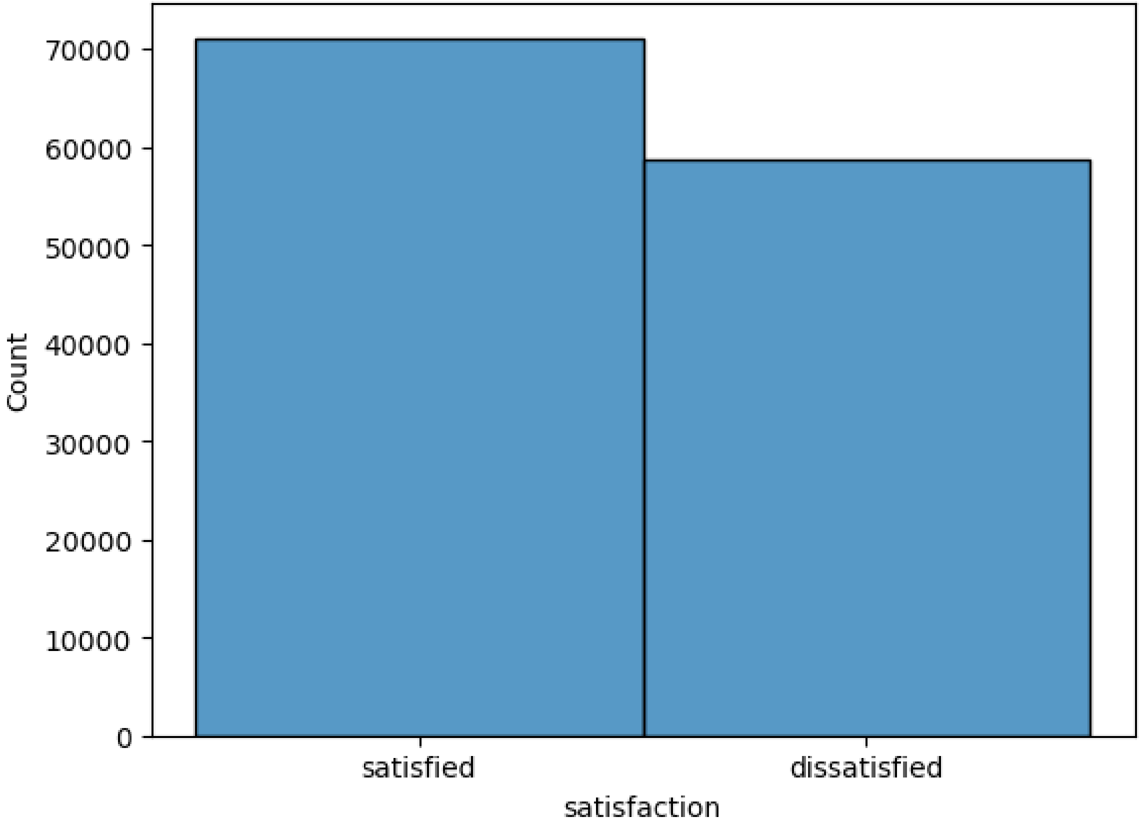
- Dataset is *Customer Satisfaction in Airline*
  - Survey data taken from an undisclosed airline company.
  - Focused on measuring an airline’s passenger satisfaction with their flight.
  - Passengers responded by ranking a number of variables on a scale of 1-5 or responding to certain flight related questions.

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
- 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.



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

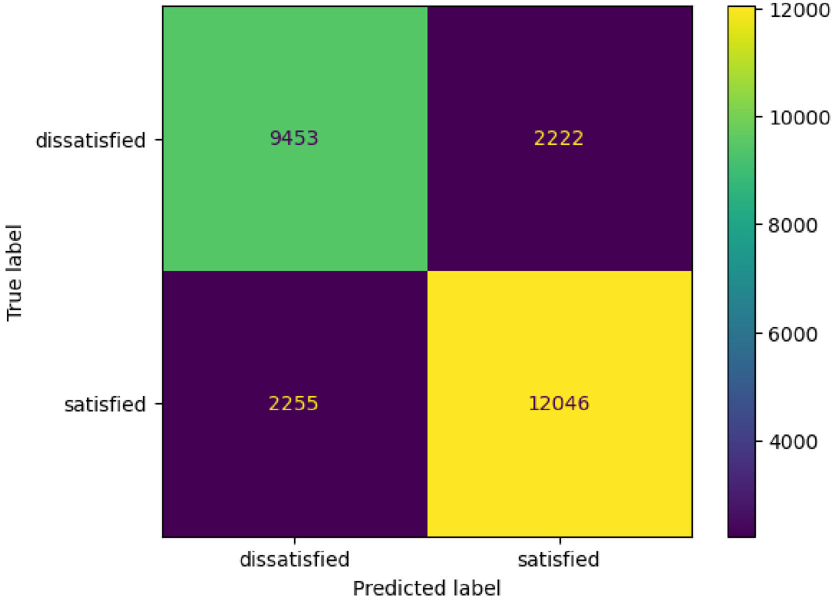
- The selected variables are “*Seat comfort*”, “*Departure/Arrival time convenient*”, “*Food and drink*”, “*Gate location*”, “*Inflight wifi service*”, “*Inflight entertainment*”, “*Online support*”, “*Ease of Online booking*”, “*On-board service*”, “*Leg room service*”, “*Baggage handling*”, “*Check-in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

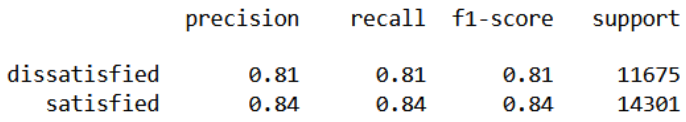
Experiment Number	Parameters
1	All twenty-one (21) variables with 80/20 split for train, and test
2	All twenty-one (21) variables with 70/30 split for train, and test
3	Selected variables with 80/20 split for train, and test
4	Selected variables with 70/30 split for train, and test

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

- Model with twenty-one features and a 70/30 split has the highest true positives. Showing that the model correctly identifies the most satisfied customers.
- The model with twenty-one features and an 80/20 split has the lowest false negatives, showing that it incorrectly predicts the fewest dissatisfied customers as satisfied.

80/20

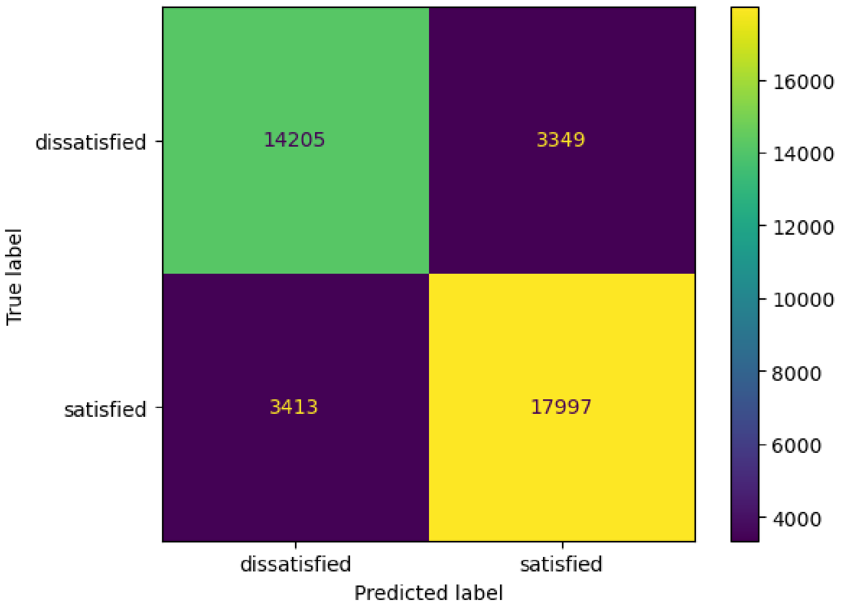


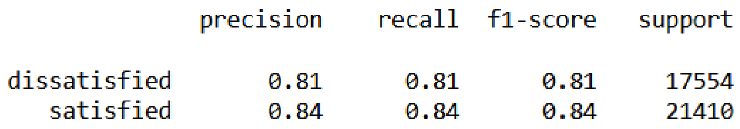


```
              precision    recall  f1-score   support

dissatisfied       0.81      0.81      0.81     11675
   satisfied       0.84      0.84      0.84     14301
```

70/30





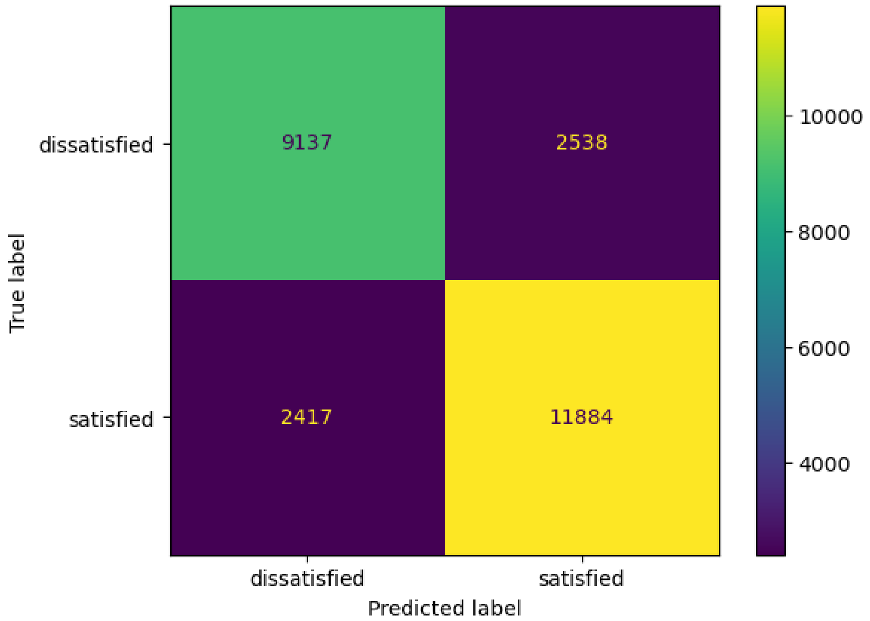
```
              precision    recall  f1-score   support

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.

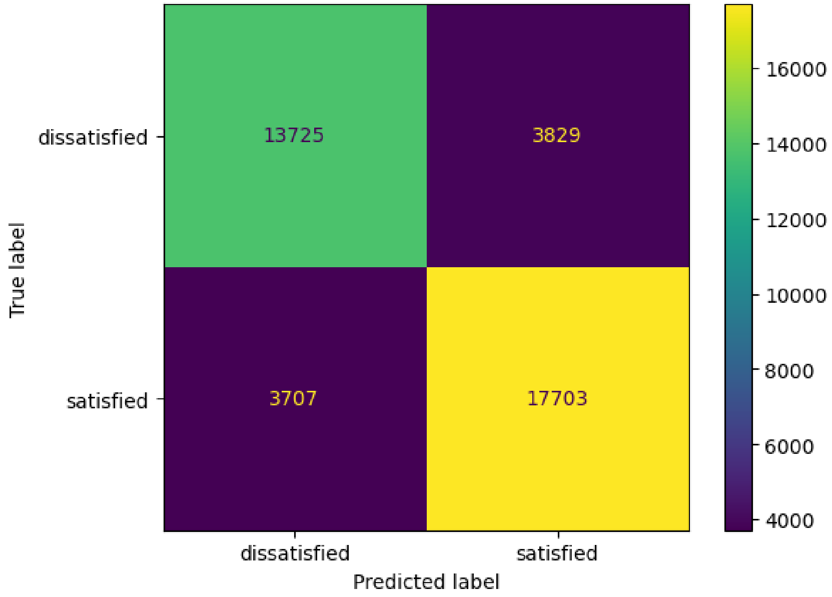
80/20



```
              precision    recall  f1-score   support

dissatisfied       0.79      0.78      0.79     11675
   satisfied       0.82      0.83      0.83     14301
```

70/30



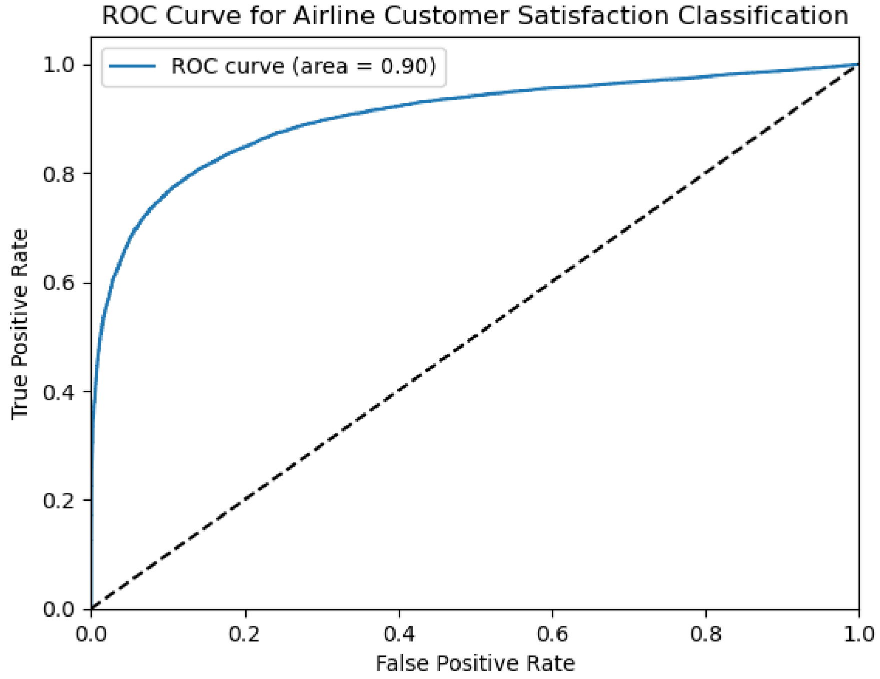
```
              precision    recall  f1-score   support

dissatisfied       0.79      0.78      0.78     17554
   satisfied       0.82      0.83      0.82     21410
```

Results

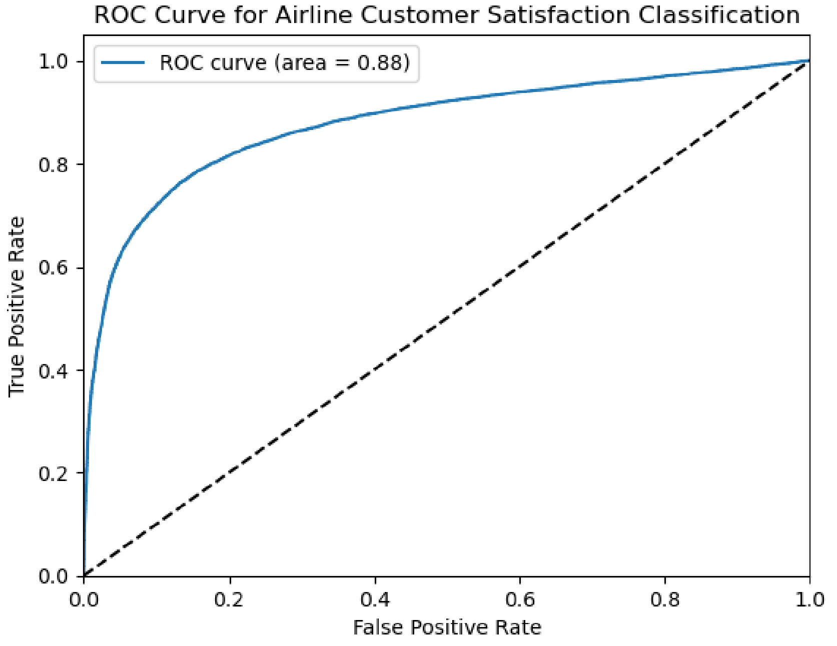
- ROC curve created by plotting the true positive rate against the false positive rate.
- The AUC is the probability that the model will rank a randomly chosen positive example more highly than a randomly chosen negative example.
  - Ranges from 0 to 1, 0.5 = random guessing, and 1 = perfect performance.

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



21 variables

Selected 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
- The models with twenty-one variables, 70/30 split model excelled in identifying satisfied customers, the 80/20 split model minimized misclassification of dissatisfied customers.
- The model with all fourteen features and an 80/20 split was the least favorable performer due to higher false positive and false negative counts.