Cleaning:

To clean the data, I had to first add column headers so that I could identify the different columns. I then used head(), describe(), and info() to look at data types, summaries, and a few lines of the data. I converted all columns into integer variables, with an integer representing the categories. I also made the class a binary variable, with unacceptable as 0 and anything else as 1.

Looking at counts:

After cleaning and initial analysis of the data, I looked at the counts of each variable. They are all evenly distributed, with an equal number of each category belonging to each predictor, besides class. With class, only 29% belonged to the acceptable class. Those graphs are below, and in the zipped folder.
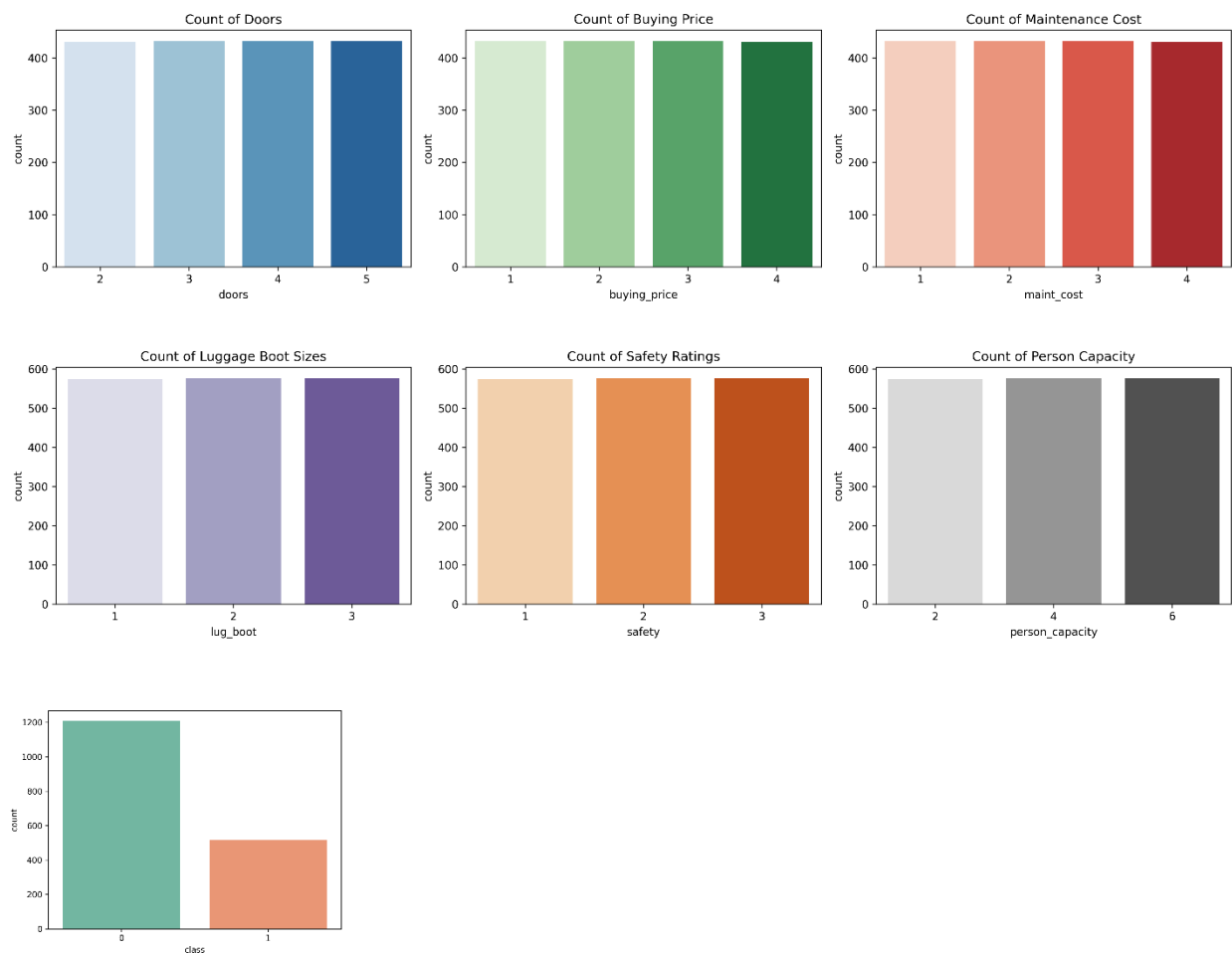
Model accuracy, Confusion Matrix, Precision, Recall, F1:

The final model's test accuracy was ~.977. The training accuracy was just slightly above. The confusion matrix is as follows,

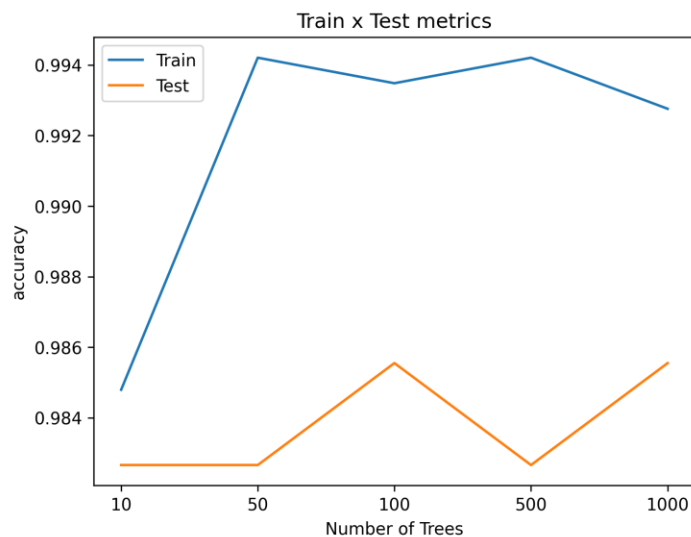| 232	7
1	106

From this confusion matrix, the precision, recall, and F1 score are as follows

Precision: 0.9906542056074766

Recall: 0.9380530973451328

F1 Score: 0.9636363636363636

Validation curve based on number of trees:

Train x Test metrics

Summary:

The model's accuracy was fantastic, being very close to 100%. This means that not much could be done to improve the model. The most challenging part of this project was deciding which model parameters to use, and deciding how to replace the data with usable labels.