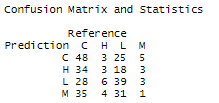
LeTicia Cancel

DATA622 Homework #1

October 28, 2022

The two datasets chosen for this assignment are the 1,000 Sales Records and the 100,000 Sales Records. I opted for smaller datasets, but not the smallest, for this assignment because I want enough rows to make a significant analysis while also not having too much where my machine is not able to handle the data processing. This data has a mix of categorical and numeric variables. The numeric variables related to cost, profit, and revenue. The categorical variables are related to location, item type, and priority. Since this is an exploratory exercise we do not have a specific question or business problem to solve so instead we will choose one of the categorical variables to be our predictor, Order Priority.

The chose Decision Tree to create the model for both datasets but was only successful with one dataset. I used an 80/20 split for the train/test data and examined the results using a confusion matrix. There are 4 possible outputs so it produced a 4 x 4 confusion matrix (table below). The results of the prediction do not look very promising. The value 'C' was correctly predicted 48 times or 59% and was falsely predicted as 'L' 25 times or 30%. Most of the predictions were for 'C' and 'L' and most were incorrect. The overall accuracy of this model is very low at 31.8%



After reviewing the results I am not sure if this is the right model to use for this data or if the model is appropriate and the error is in my data preparation. From my initial review of the data it didn't seem like much cleaning was needed.

When I attempt the same model for the second larger dataset I get an error. The data for the larger file is prepared exactly the same way as the smaller file. When I attempt to build the decision tree it only produces one node so there is nothing to plot. This data needs to be worked some more so it can produce a working decision tree model.