Shay Snyder

CSCI 4957 Data Analytics

Dr. Husari

Exercise 11: Classification

The goal of my project was to predict whether a person’s income will exceed $50,000 per year. To achieve this goal, I used the provided ‘adultsData.csv’ dataset, decision trees, and the Naïve Bayes model. First, I had to split the data into two distinct groups, test data and train data. The train data will be used to train the models while the test data will be used to evaluate each model’s performance. I used stratified random sampling, via the createDataPartition() function in the caret library, in order to get an even distribution of people in both sets of data. The resulting datasets, training and test, accounted for 80% and 20% of the original dataset. Below is a screenshot listing the dimensions of all dataset used throughout this process.

A screenshot of a cell phone

Description automatically generated

Using the rpart and rpart.plots libraries, enabled me to train a decision tree model based upon the train data. Decision trees are a tree-like model of decisions / questions and their possible outcomes in the context of various problems. This model is built with the goal of determining what ‘questions’ are most effective to determine if an individual is likely to make more than $50,000 per year. To evaluate the performance of said model, I used a confusion matrix via the e1071 library and our test data. Below is the resulting decision tree and confusion matrix.

A screenshot of a cell phone

Description automatically generated

A screenshot of text

Description automatically generated

Using the naivebays library, enabled me to train a Naïve Bayes model based upon the train data. The Bayes classifier is based upon Bayes’ theorem that assumes independence between predictors. This method is quite easy to implement but loses accuracy due to the assumption of attribute independence. To evaluate the performance of said model, I used a confusion matrix via the e1071 library and our test data. Below is the resulting confusion matrix.

A screenshot of text

Description automatically generated

As for comparing the two models, the decision tree based model (Model 1) made less mistakes than the Bayes model (Model 2). Model 1 made 5,478 correct predictions (4617 correct predictions of income <= $50,000 per year income & 784 correct predictions of income > $50,000 per year income) while Model 2 made 5,378 correct predictions (4694 correct predictions of income <= $50,000 per year income & 761 correct predictions of income > $50,000 per year income). From these calculations, we can see that the decision tree-based model (Model 1) performed slightly better than the Bayes model (Model 2) as Model 1 made more correct predictions than Model 2. (I noticed that my confusion matrices varied slightly from the video. I would assume those discrepancies come from differing sampling methods. If I am incorrect, please let me know ☺)