REPORT HOMEWORK – 2

Question 1:

Diagram

Description automatically generated(20pointsModifiedExercise14.4inESL)  
Cluster the marketing data of Table 14.1 (ESL) using a classification tree. This data is in the ISLR package, and also available on UB learns.Specifically, generate a reference sample of the same size of the training set. This can be done in a couple of ways, e.g., (i) sample uniformly for each variable, or (ii) by randomly permuting the values within each variable independently. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability. Compare the results to the results near Table 14.1 (ESL), which were derived using PRIM.

Text

Description automatically generated with medium confidence

We can observe from the above that these features can predict to do a classification. To double check this we can predict the model on the training set. Therefore, confirming our assumption that they do not have any predictive power. The terminal node had percentage of 20% with a class 1 probability of 72%. Comparing the results with PRIM we notice that the values as 0.08 when the household were more than equal to 3 and similarity in the trends can be observed in case of 0.25 when the household values were lesser than 3 and language was lower than 2.

Question 2:

1. Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

The choices I made in grouping categories was done taking the quartile, mean, minimum and maximum values of the data (summary) into account .

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

1. Visualize the data using the itemFrequencyPlot in the “arules” package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).

Parameters :

support = 0.02 -> itemFrequencyPlot

Apriori Algorithm with the following parameters was using :

Chart

Description automatically generated support = 0.02, confidence = 0.8

1. A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by “dis”). What can you advise on this matter through the mining of association rules?

From the above results we can suggest that the student should be choosing a housing that is away from his work as the crime rate is seen to be decreasing.

C) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by “dis”). What can you advise on this matter through the mining of association rules?

A picture containing text, receipt, screenshot

Description automatically generated

From the above observed data it can be said that the student should be choosing a house away from work as the crime rate decrease and also an area that has low pupil teacher ratio.

1. A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

A picture containing table

Description automatically generated

From the above observed data it can be said that the family should move to a small residence where it has low proportion of non-retail business acres per town and also the medv is expensive. Also need to look at the average tax. We can observe the combinations in the lhs column in the above attached output to advise the family on what decisions to make.

**Extra Credit**: Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?

Text, table

Description automatically generated

Regression model is used when you want to predict a continuous dependent variable from a number of independent variables.