End of Term Assessment

The data used for these models is from the Car Evaluation data set from the UCI machine learning repository. The purpose the of the data is to determine the condition and value of a car based on a number of input variables. This could prove useful for quick estimations of a car or for consumers and new sales people for cars. Using an accurate prediction model will allow these people to compare against a model of known accuracy which can help them in their judgement to the value of a car. As a system for consumers this will allow them to better understanding the value of the car prior to purchasing or attempting to sell a car. This will ensure that so long as the model is accurate that they are receiving a fair price for their vehicle. For a novice sales person the model can act as a quick point of comparison for their estimations ensuring that the sales are in line with company pricing and frees up time from a senior sales person as they will not need to take the time to verify the estimation of the novice sales person.

KNN:

The car evaluation data comes with a mixture of numeric values and text labels based on the input variable. Each of these values had to be converted to numeric input variables to avoid some errors in R. Each of the columns had their text values switched to numeric values with low values being the starting point and increasing in increments on 1. For example the first column “buying” originally had the labels “vhigh”, “high”, “med” and low”. Each of these were switched to 1-4 starting from low and increasing to vhigh. The data was then randomised to remove any patterns that may have occurred due to the order the data was inputted as this could create a model that is accurate for this particular data set but would be significantly less for future data not included in this data set.

The data set was then split into training and test data. The training data set used the first 1500 rows of the data set so that there would be sufficient data to create the model. The testing data contained values 1501-1728 and was used for testing the model. There are two more variables used for the prediction and cross table. The train value was used as the class for the prediction and the test value for the cross table.

The modelling process for KNN took the data as mentioned above and placed it into a prediction model. The only variable the required changing was k. After running the model a number of times the value of k which produced the most accurate result on average was k=4. These values were then placed into a cross table to view the number of errors made in tabular form but also what the incorrect predictions were mislabelled as. This was then placed in an equation to get an accuracy value based on the table from the data.

From the results the random data was producing a model on average with an accuracy lower than that of the original ordered data. This indicates that when new data is introduced especially from a different source that the model will not be as accurate if using the original order of the data.

The model using the original ordering produced a model with a sum value of .89 while the random ordering created a model with on average .5. This shows that using randomised data has a significant impact on the accuracy of the model. The model with random data is not suitable for use as it only offers a 50% chance of estimating the correct value. The model with the original data is accurate but has too many limitations for practical use based on the loss of accuracy from randomising the data.

Decision Tree:

The decision tree model required less preparation before use. The values did not have to be changed to a numeric format. The data was randomised again and the data was split into training and test data using the same ratio as before. The output variable in the training data set had to be converted prior to be used in the model.

The modelling process for the decision tree involved taking the training data and the training label and placing them within a C5.0 function. This created a model which is then placed inside a predict function against the testing data producing a prediction variable. The prediction variable is then placed in a cross table to compare the predicted values against the actual values.

The model produced from the decision tree using the car data proves to be quite accurate using the randomised data. The ordered data created a slightly less accurate model. From the table using the original data cars labelled as “acc” tend to have the most errors in prediction while on both the ordered and randomised data all other output variables tend to be accurate. This may indicate that the model would benefit from more data on cars that have been labelled as acceptable.

Kmeans:

The data from Kmeans had to be prepared first of all like KNN so all text values were converted to numeric values first. The data was then randomised to remove any ordering that may have originally been present in the data.

The model was created using a kmeans function with the randomised data and choosing a number of clusters. The model was then checked for the withinss for a measure of the clusters.

Based on the results from the Kmeans model the output was not impacted in any meaningful way by randomising the data. A model created using 8 clusters showed a significant improvement of a model using 4 clusters however, the results from both displayed a large degree of error.

Based on the 3 techniques used for creating models decision trees proved to be the most suitable way for creating a prediction model using the car evaluation data. Both Kmeans and KNN have a large degree of error indicating that clustering or looking at the nearest values does not work well with this data for evaluation cars. Decision trees allow the model to take branch off based on certain variables. In the case of this data set having a safety rating of low may immediately branch towards an unacc car which could help with the accuracy of the model. In practice using the current data decision trees are the only suitable technique to provide a level of accuracy that could be useful to an end user.