CMPU 4011 Machine Learning for Predictive Analytics
ASSESSMENT 2 – BUILD A CLASSIFIER

This assignment was to develop a classifier for a given set of data and queries. First was the decision of which method would be used for the classifier. The options were The Nearest Neighbour Algorithm, Naive Bayes' Classifier or Linear Regression. Nearest Neighbour was immediately ruled out due to the amount of categorical data which it is not great at handling though can be done with some work. Both Naive Bayes' Classifier and Linear Regression would work for the data but both would need some work to handle both he categorical and continuous data.

Though both could handle the data it was decided to use Linear Regression as the largest number of categorical options in a header was 12 which is a very manageable number for this method. Naive Bayes' could also have been used by binning the continuous data. This was avoided as the data has an odd spread across these features and varying sizes especially across s the balance and paydays table.

One this had been decided the data was imported to begin training the model. The first step was to clean up the data provided. For the most part the data was good but the contact and poutcome columns had many unknows, or no data entries, that would affect the results. Due to this these two columns were dropped as there was more then enough others to compensate. Once this was done the outcome was split off and the training can begin.

For this method each categorical is split into true and false columns for each of its options. This means the months are split into 12 columns with eleven 0's and a single 1 in the column that row originally displayed. This was done for all the categorical values in the data including the outcomes. For each column 1 is true and 0 is false. This data was then used to train the Linear Regression algorithm.

Sklearn's libraries were used to train this model. This data had 40% split off to use as base test data to check the algorithm. Overall, it ended up with a Variance score: 0.0987 whish is very low as the target was 1. Though this if was off for the sake of completion the full testing and queries will be run. The data was graphed for easier viewing before moving into part 2.

For the main query testing the queries were imported and handled much eh same was as the training data but without the output column as that's what we want to find. The contact and poutcome columns were removed and the categorical values were split. Next the same linier regression algorithm will be run. This time however instead of splitting the training data the main data will be used to train and the new imported queries will be used to test. After the training is done sklearns predict function will be used to predict the output column of the queries. This will then be used as the test output to check how accurate these predictions are.

This data received a variance score of 1.0. this means the data classifier is perfect which is very unlikely. This is more than likely due to the incorrect value that the original training and testing showed. This was graphed to show the distribution. One this was done the outputs were exported to a txt file for later viewing.