Mark Hession

A00239393

Data Mining & Machine Learning Module Assignment

Table of Contents

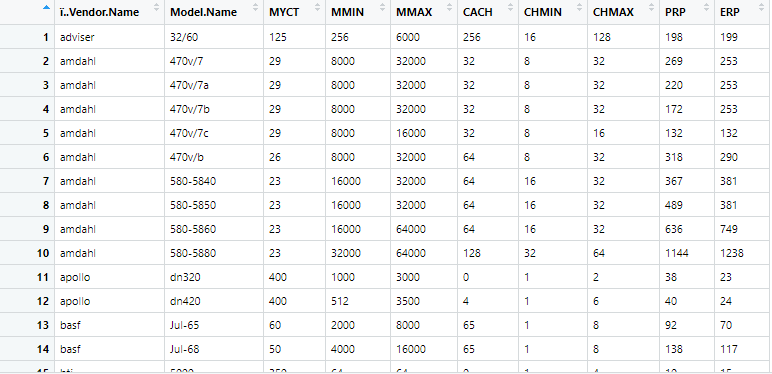
Regression:

1.1

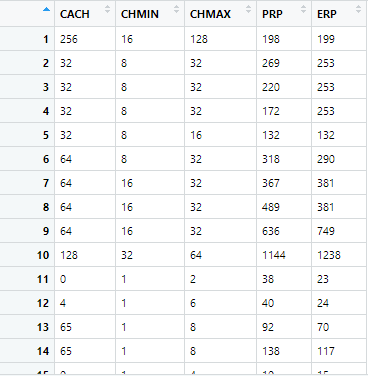
This dataset is about relative CPU performance data, described in terms of its cycle time, memory size, etc. The data we will be using from the data set is CACH: cache memory in kilobytes (integer)CHMIN: minimum channels in units (integer), CHMAX: maximum channels in units (integer), PRP: published relative performance (integer), ERP: estimated relative performance from the original article (integer). We will be predicting the ERP of a system based on its CACH/CHMIN/CHMAX/PRP.

1.2

This is my dataset prior to any alterations.



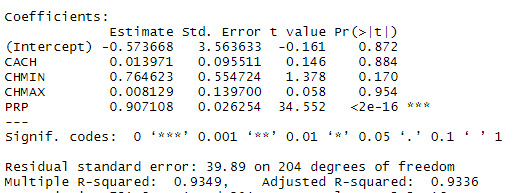
This is my dataset after it is altered to suit the regression process better.

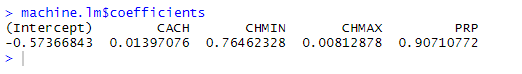


1.3

1.4

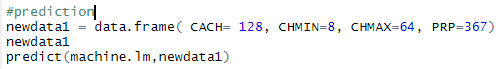




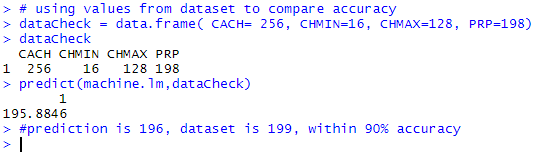
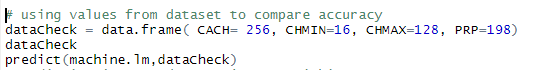
From our model generation, where we will try to predict the ERP, we can see our R-Squared value is 0.93, giving us a 93% accuracy for our prediction, which we will test later.

1.5

We can see that based on our new data frame inserts of newdata1, we have a predicted ERP of 341.



When we test our accuracy of our regression, we can take values from our original dataset and see how the ERP compares to our prediction.



Our dataset had an ERP of 199, our predicted ERP for the same values was 196, which is within the 93% accuracy bounds from before.

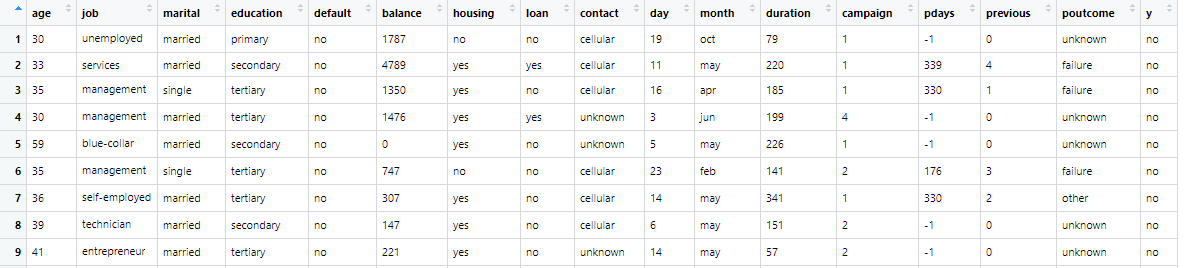
1.6

From our results based on our model, we can see when passing in the values CACH, CHMIN, CHMAX, PRP, we have a 93% accurate result, as proved when we inserted data from our dataset and compared it to our prediction result. For our dataset, to be able to predict the ERP based on these values is beneficial to the consumer as it gives the Estimated Real Performance of a CPU, based off the CPUs cache information, and you can compare it to the PRP of the CPU. This gives consumers an accurate estimate of what type of performance they can expect for the CPU they will buy.

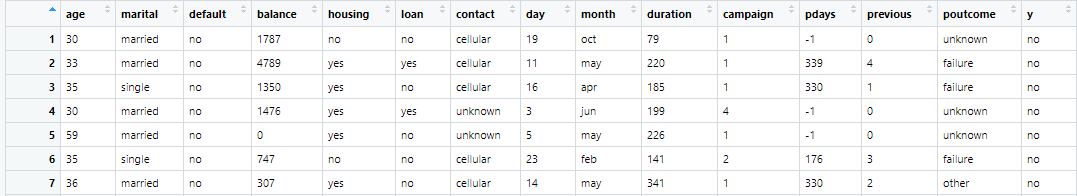
Decision Tree:  
2.1 – For this analysis, I have chosen a dataset which is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The goal is to predict if a client will say yes or no to a term deposit subscription. The dataset includes several details such as the client’s age/job/education along with banking details such as if they have a loan and what type of contact, they have. The decision tree will show the branching paths based on the client’s information and predict how likely they are to say yes or no based on that information.

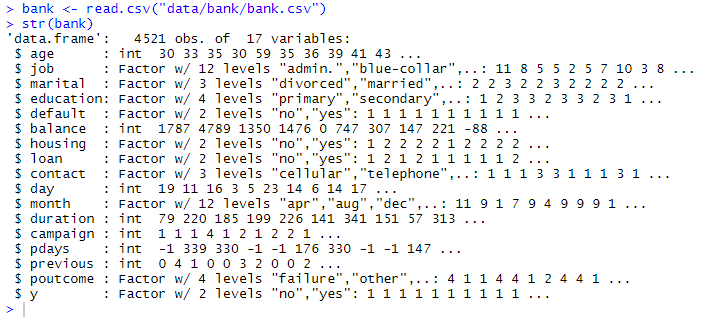
2.2

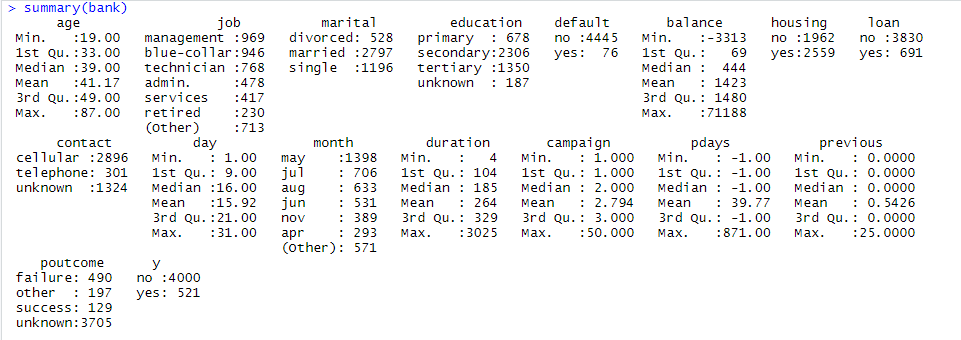
Bank Dataset before alterations



Bank dataset after alterations



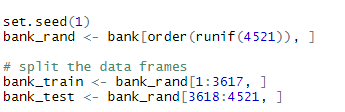




2.3

Splitting the data frames.

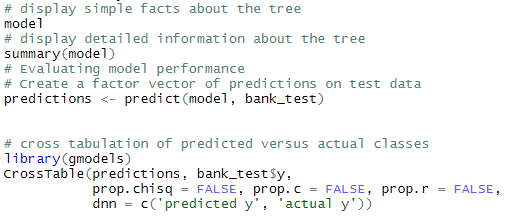
For the training and testing, I split it 80/20, 80% for training and 20% for testing.

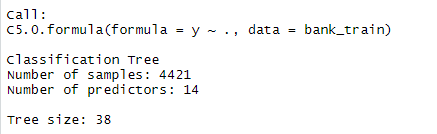




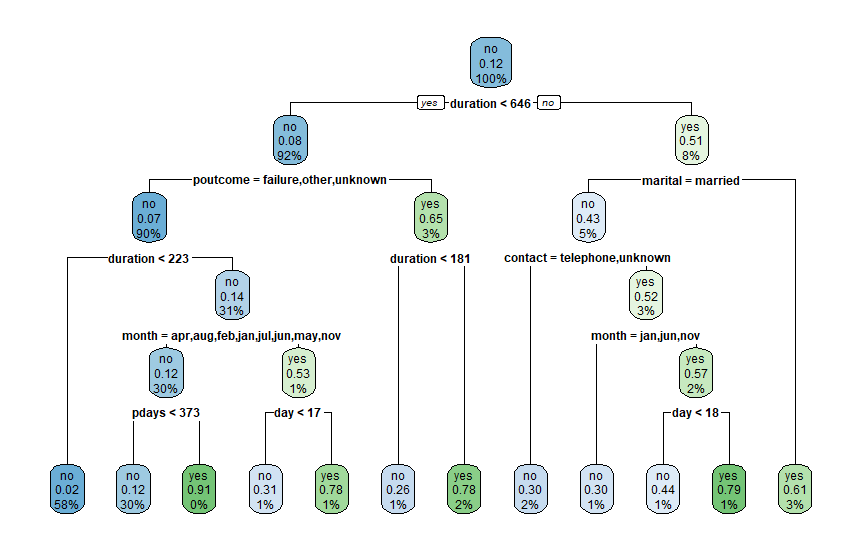
2.4

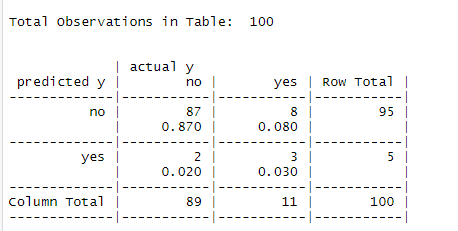
Model





2.5

Predictions



2.6

Evaluation

KNN:

3.1

For my KNN prediction, I used the same dataset as decision trees, to compare the results given for each.