**Homework 2**

Write-up

**Question 1**

The cereal dataset is given to us to build a predictive model for nutritional rating. After getting the dataset, I downloaded some necessary packages and loaded the data. The cereal data contains 16 columns and 77 rows. I preprocessed the data by checking to see if there is missing value as well as eliminating the name column in the data because it is irrelevant for predicting nutritional rating.

A picture containing logo

Description automatically generated

My first step is to divide the data into test and training. I decide to use 80% for training and 20% for test. Afterwards, I check whether the test and training has been successfully split by checking the row of train and test and they indeed at together into the total cereal dataset’s row number.

A picture containing table

Description automatically generated

The next step is to fit a linear model using the train dataset. After fitting the train dataset, I notice the variables highly correlated with the rating are calories, protein, fat, sodium, fiber, carbo, sugars, potassium, and vitamins. My next step would be to see if the MSE value is high or low for this dataset compare with the test dataset.

Table

Description automatically generated

Using the equation provided in the textbook, I found that the MSE is 1.56e-13, therefore the error is very low.

Text, letter

Description automatically generated

I decide to perform using forward subset selection and notice that base on forward subset selection method, sugars, fiber and sodium are the top three variables that affect the rating. I then use plotting and try to visualize the minimum error which is the lowest point in both Cp and BIC (9). But in order to get an exact number, I use the “min” function to determine the optional point, which is 9 for BIC and 10 for Cp for forward subset selection.

Diagram, shape, rectangle

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Different than forward subset selection which taken into account of variables one at a time, exhaustive search taken into account of all the variables at the same time, and the result is same as the forward selection.

Diagram, shape, rectangle

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Base on my findings above, I conclude that we could use exhaustive search in predicting the optimal point which could be either 9 or 10. And the predictive accuracy is very accurate base on the model above because the error rate is very close to zero. Overall, I found the top three variables that close correlated with the rating are sugars, fiber, and sodium. I would suggest to use

**Question2**

The zipcode data are normalized handwritten digits scanned from envelopes by the USPS. The images have been size normalized and became 16x16 grayscale images. There are 7291 training observations and 2007 test observations.

After load the train and test dataset into data frame and download some require packages, the first thing I did is to get a full picture of the dataset. Then, I took out only 2 and 3 from v1 and put them under as train\_new. The reason for this is to train and test the KNN only for the filtered dataset.

Graphical user interface, text, application

Description automatically generated

In order to show the training and test error for k= 1,3,5,7,9,11,13,15, I will perform a KNN for k = 1,3,5,7,9,11,13,15 for the filtered dataset to get training and test error. The result I got back are down below. These result means that when k=1 the training error is 0 whereas the test error is 2.47, when k=3 base on the nearest three neighbors, the training error is .5% and test error is 3.02%. From this table we can conclude that as K increases the error also increases but the test error is always higher than training error this might be the case due to we are comparing with the training dataset.

|  |  |  |
| --- | --- | --- |
|  | Training error | Test error |
| K=1 | 0 | 2.47 |
| K=3 | 0.5 | 3.02 |
| K=5 | 0.576 | 3.02 |
| K=7 | 0.648 | 3.29 |
| K=9 | 0.936 | 3.57 |
| K=11 | 0.86 | 3.57 |
| K=13 | 0.86 | 3.84 |
| K=15 | 0.936 | 3.84 |

Graphical user interface, text, application

Description automatically generated

I use simple linear regression because we want a binary classification. What I did first is changing the variables in V1 into 1 and 0 instead of 2 and 3. Then I use the new training data and all the variables in the dataset to predict V1 using linear model. And the same for test dataset. The result shows that the training dataset has an error of .5% and test dataset has a error of 4.12%.

Graphical user interface, text, application

Description automatically generated

lr <- lm(V1 ~ ., data = train\_lr)

Text

Description automatically generated

In conclusion, I notice that using knn is a better predictor than linear regression because the error is smaller especially when knn=1-3.

Question 3

The College dataset in ISLR contains 777observations on 18 variables include number of applications accepted, number of new students enrolled, etc. we could find that information using ?College. In order to find the rest error of the training set using least squares method, I first fit a linear model with Apps as the linear model. Then I use the test dataset to predict, the test error came 1754670.

Text, application

Description automatically generated