COMP237 – Linear Regression Lab

Daniyal Ahmed – 301152472

**Exercise 1 – Uniform Distribution with/without noise**

First we imported the required modules and created the values for x and y using a random uniform distribution for x and the formula y = (12 \* x) – 4.

Text

Description automatically generated

Second we plotted the graph of X vs Y with labels

Chart

Description automatically generated

Then we were to do the same with noise added. To calculate the value for noise using the np.random.normal() function in python, it requires the mean and standard deviation as arguments. Calculate the mean and standard deviation

Text

Description automatically generated

Now we must create a value from noise using a random sample from the gaussian distribution using mean and standard deviation, then create a second set of numbers y2 = ((12 \* x) - 4) + noise.

Text

Description automatically generated

Plot the graph with new equation with noise

Chart

Description automatically generated

Graphical user interface, application, chat or text message

Description automatically generated

**Analysis**:

The slope or angle of the line stayed the same. All that occurred was the values moving. The movement is very minimal however since the value for noise is not large, as can be seen in the last figure.

**Exercise 2 – Linear Regression**

First import the required modules

Text

Description automatically generated

Load data into data frame

Text

Description automatically generated

Explore data and find out column info

Text

Description automatically generatedText

Description automatically generated

Transform the categorical variables to numeric variables using a for loop and get\_dummies function

Text

Description automatically generated

Drop categorical and transaction id columnsText

Description automatically generated

Create function that normalizes data setText

Description automatically generated

Use function on dataset



Print the first 2 records

Text

Description automatically generated

Plot histogram of all variablesGraphical user interface, text

Description automatically generated

Plotted a scatter matrix Text

Description automatically generated

Create data to be used for training and testing data using predictor variables and output variable ‘Total Spend’Text

Description automatically generated

Split into training and testing data

Set random seed to 72 (last 2 digits of student number)



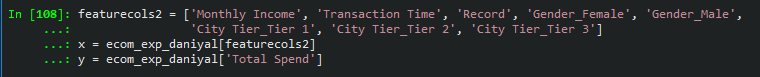
Using the training data, fit data into linear regression modelText

Description automatically generated

Print out the coefficients (weights) Text

Description automatically generated

Print out training score/R-squared value and get R1(first r squared value) = 0.196 = 19.6%

Now we set new values for our x and y from our feature columns, with the value ‘Record’ as well this time

Split the new set of training data into training and testing models. A screenshot of a computer

Description automatically generated with medium confidence

Set random seed to 72



Build another linear regression model with new training set

Text

Description automatically generated

Display the coefficients

Text

Description automatically generated

Display the score/R-squared value and get R2 = 0.919 = 91.9%

**Analysis:**

After creating linear regression models for this data set with and without the ‘Record’ it can be seen that you receive an Rsquared value much closer to 1 or 100% when building the model using the ‘Record’ column. It can be seen that without that column the efficiency of this model is only 19.6% which means that the model does not fit our data very well at all. With the ‘Record’ column included we receive an Rsquared value of almost 100% which means that this model with that column is a much better fit for the data set.

----NOTE----

THESE ARE STILL MY SCREENSHOTS, I SWITCHED TO VSCODE AFTER TAKING SCREENSHOTS FOR LAB ASSIGNMENTS 3 AND 4