CIS 662 HW5 Report: Regression (Both Linear and Logistic

**PART 1: Linear Regression and Comparison with NN:**

Dataset selected is **71-80.csv**

Data has also been split in the same way as the previous homework 80:20 to compare the results.

Linear regression modelhas certain **assumptions such as:**

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No auto-correlation
* Homoscedasticity

The assumption and requirement of the model is that the data should be normally distributed. Hence, I have applied **MinMaxScaler** as it normalised the data to a common range. Also, since the unit of all the values is common, I wanted to preserve the original range of the values.

Linear Regression uses MSE for loss computation by default.

I have used **MSE**, **MAE** and **MAPE** for analysis and better understanding of the model fit to the data.

Also, the corresponding outputs for both Linear Regression and Neural Network are attached for better analysis.

Formulas for the above can be found in the code file.

**OUTPUT:**

**Model:**

**A computer code with blue background

Description automatically generated**

**MSE for Linear Regression:**

**A screenshot of a computer program

Description automatically generated**

**MAE and MAPE for Linear Regression:**

**A screenshot of a computer error

Description automatically generated**

**Screenshot for Linear Regression output:**

**A screenshot of a computer screen

Description automatically generated**

**Screenshot for NN output:**

**A screenshot of a computer screen

Description automatically generated**

**Comparison with NN:**

**MSE:**

|  |  |
| --- | --- |
| **Neural Network Model** | **Linear Regression** |
| 669.15 | 2219.51 |

**MAE:**

|  |  |
| --- | --- |
| **Neural Network Model** | **Linear Regression** |
| 22.18 | 35.78 |

**MAPE:**

|  |  |
| --- | --- |
| **Neural Network Model** | **Linear Regression** |
| 42.74 | 27.72 |

**CONCLUSION:**

As attached in the screenshots above, NN outperforms Linear Regression as linear Regression is a simple model and can draw a simple best fit line between the data points. However, it cannot capture complex patterns within data as a neural network that contains hidden neurons.

MSE and MAE is lower in NN, but MAPE is higher in neural networks as they capture high **variance** in data and are susceptible to capturing outliers.

A screenshot of a computer

Description automatically generated

Here, we can see the predicted value is 21 but the actual value is 3, and the abs value is 18. Linear Regression doesn’t capture such variance and patterns as they can capture high bias and low variance in data, hence they do not capture outlier patterns and thus, MAE and MSE are higher in Linear Regression models.

**PART 2: Logistic Regression and its comparison with NN:**

Like Linear Regression, I have used normalization to preserve the original characteristics and values as all the feature values are of the same scale.

As the data doesn’t contain categorical values, as done in HW5, I have used binning to categorize the ratio between columns 2022 and 2021 into 3 distinct categories such as {Low, Medium, High}.

Citations columns cit\_2017 to cit\_2022 are considered as feature columns (X features) and column **category** is considered as Y column. (Column to be classified)

Since we are doing multi-label classification, I enabled the multi-class feature available for Logistic Regression.

Also, I have used a classification report and AUC-ROC diagram to compare the results of both logistic regression and neural networks.

**OUTPUT:**

**Logistic Regression Model:**

**A screen shot of a computer program

Description automatically generated**

**Screenshot for Logistic Regression classification results:**

**A screenshot of a computer screen

Description automatically generated**

**NN Classification Report Output:**

**A screenshot of a computer screen

Description automatically generated**

**Logistic Regression Classification Report Output:**

**A screenshot of a computer

Description automatically generated**

**ROC-AUC of NN verses Logistic Regression**

**A graph of a function

Description automatically generated with medium confidence** **A graph of a function

Description automatically generated**

**CONCLUSION:**

As attached in the above screenshots, we can observe that Neural Networks perform better and provide better classification results (F1 score) than Logistic Regression.

Categorizing continuous numerical data using logic of obtaining ratio between 2022/2021 can lead to add a bit of **non-linearity** within the data which **is captured better by neural networks**. Also, neural networks use softmax as an activation function, however for **logistic Regression sigmoid activation function** is used where each output can be interpreted as the probability of the input belonging to a particular class and the most probable value is considered.

As you can see in the screenshots for ROC-AUC for both NN and Logistic Regression, NN performs better with a **higher value AUC value** for all the 3 classes as opposed to that of Logistic Regression.