AIX2006 Assignment – Regression Modeling and Evaluation

Team 13
Misong Kim 22200070
Hee Han 22100784
Yoobin Park 21900296

Team Distribution

- Yoobin Park(21900296): writing code, model evaluation
- Hee Han (22100784): writing report (explanation of model evaluation concept), model evaluation
- Misong Kim (22200070): writing report (explanation of MLP model concept), model evaluation

Develop a regression model for analyzing the input-out mapping of the given dataset.

1. **Create an MLP-based regression model using the provided dataset and explain the regression model structure in detail. You may need to come up with by yourself how to divide the dataset.**
    - **The basic concepts of MLP model structure**

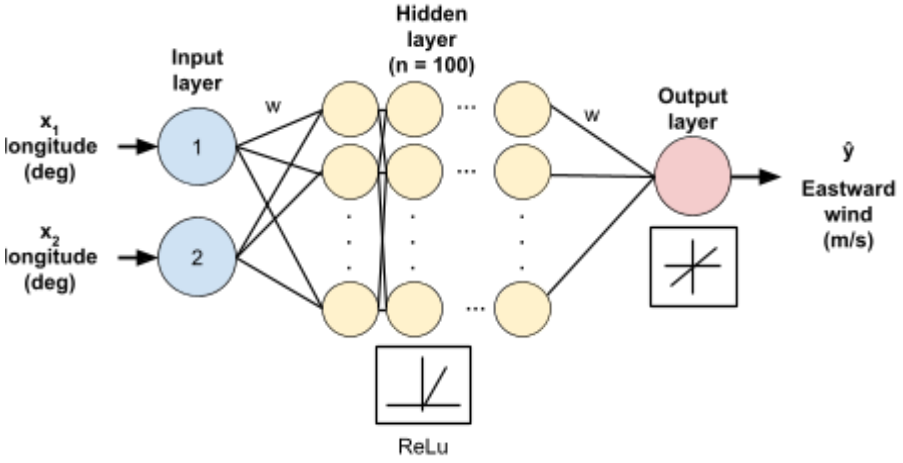
        MLP is a neural network model that reflects the organization of the human brain and consists of input layer, hidden layer, and output layer. Each layer is composed of multiple nodes and determines the activation of each neuron according to the input value.

        The input layer is responsible for receiving the input data, and the hidden layer performs calculations to find the optimal weight parameters through repetition. In the output layer, prediction is made based on the information transmitted from the input layer, and the classified output is compared with the observed output to calculate an error.

        In artificial neural networks, weights between layers are adjusted to minimize overall errors through learning. In the case of MLP, a

backpropagation algorithm that compares output values and actual values and adjusts weights in the reverse direction is mainly used.

- **The models generated by the team (in detail)**



In order to create an MLP-based regression model with the given data, we first decided to scale the data in order to make sure that datasets are on the same scale. If the values of the features are closer to each other, it is likely to train the algorithm well and faster, and it can also produce high accuracy of the model.

Then, we divided our dataset into training dataset and testing dataset, in the ratio of 8:2 respectively. (This ratio of 8:2 is widely known to produce best results based on many empirical studies.) We have also set random_state to random number 42 for reproducibility.

For the activation function of hidden layers, we chose the ‘ReLu” function. ReLu function is a simple function that returns 0 if the input signal is negative and returns the input signal itself if the input signal is positive. Since ‘relu’ avoids vanishing gradient problems and works faster than other activation functions (i.e. sigmoid function) we implied it into our regression model. The activation function for the output layer is linear activation function, which is a default activation function of MLP regressor library. We set the ‘random_state’ to 1 for reproducibility.

# of hidden layer	max_iter	R squared value
10	500	0.42
50	500	0.8
100	500	0.89
100	1000	0.94
100	1500	0.96
100	3000	0.97
100	10000	0.97

Lastly, for the number of the hidden layers and the maximum number of iteration, our model first resulted in the R_squared value of 0.8 started with the number of hidden layers of 10 and max_iter of 500. Since the R-squared value is the first checkpoint to evaluate the model, we increased the number of hidden layers at first and increased maximum number of iteration until the R-squared value does not show significant increment and close to 1. Finally, we set the number of hidden layer and max_iter to 100 and 3,000, respectively.

Model Evaluation & Discussion

2. **Evaluate the model performance using the given dataset. Specifically, you should provide the R-square value, the actual by predicted plot, the residual by predicted plot, and the model representation error value. Discuss with your teammates the results of the model evaluation. (Demonstrate that the model is accurate to the evaluation metric)**

In order to evaluate the model performance using the given dataset, the model was tested by giving x values (longitude and latitude) in the test dataset as input and the resulting predicted values were compared to the y values (Eastward wind speed) of the test dataset.

Following 4 steps are done for evaluation:

1) **R-squared value**

R-squared value is used to check the goodness of fit of a regression and it lies between 0 and 1, and the value close to 1 means the model perfectly fits the data. Our model resulted the R-squared value of 0.97. An R-squared value over 0.9 is known as a good R-squared value in a regression model since it indicates that 90% of the model fits the data and the differences between predicted value and actual value are small. Therefore we can say that our model is a good model in terms of R-squared value.

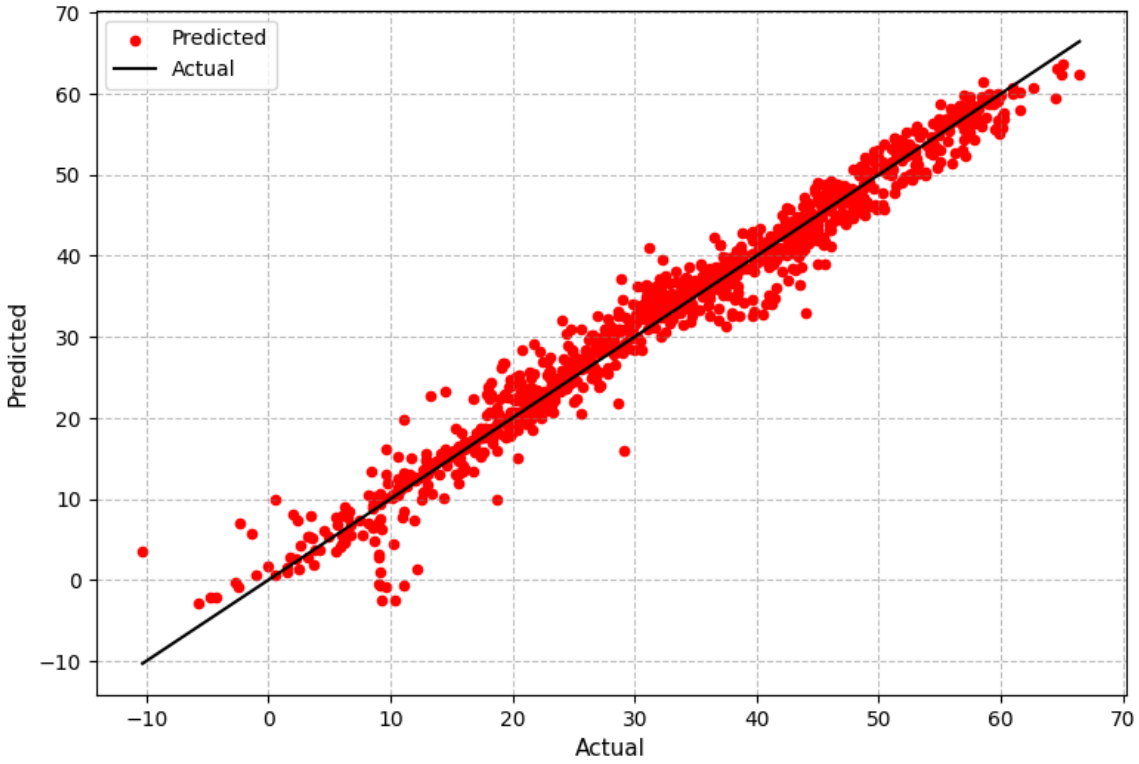
```
[26] # Compute the R-squared value of testing data
     print(f"R-squared value (test): {round(r2_score(Y_test, test_prediction_Y), 2)}")

R-squared value (test): 0.97
```

2) The actual by predicted plot

The actual by predicted plot shows the actual value for the prediction. The closer the data points are to the perfect fit line, the more likely the regression equation is sufficiently modeling the behavior of the given data. The model in good fit has data points in 95% confidence intervals that are plotted very closely to the perfect fit line and the predicted values are randomly scattered along the line.

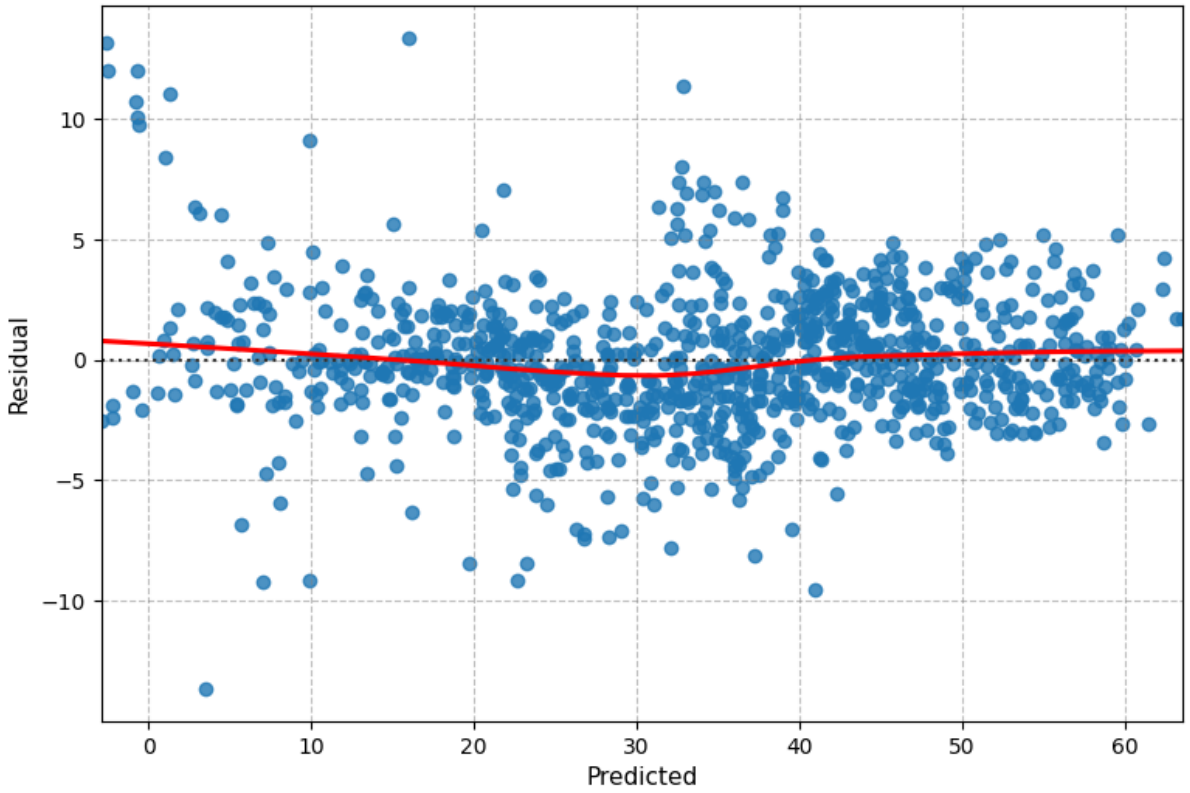
In our actual by predicted plot, x-axis indicates the actual value and the y-axis indicates the predicted value. The black solid line indicates the perfect fit line that represents the trend when the predicted values are equal to the predicted values. Also, each dot of the scatter plot indicates the predicted value of the model. As you can see, our model's predicted values are well aligned along the diagonal line (perfect fit line).



However, some of the predicted values showed misalignment approximately at 10. We hypothesized this was because the training data with an actual value of 10 might have different characteristics or patterns compared to other data in the dataset or the model has not been trained on a diverse range of data points around the value of 10 so it struggled to predict the actual values around 10.

3) **The residual by predicted plot**

In the residual by predicted plot, the x-axis indicates the predicted values and the y-axis indicates the residuals between actual values and predicted values (residual = actual - predicted). The red solid line on the scatter plot shows the trend of the residuals. We expected the residuals scattered in a random pattern centered at zero and our model showed well distributed residuals centered at zero. The red solid line was almost aligned to 0 with a little dampen at predicted value of 30, and the mean of residuals was -0.2, very close to zero, which evidenced that our model is a good model.



4) The model representation error value

Lastly, we evaluated the performance of our model through calculating and comparing the value of the Model Fit Error(MFE) and Model Representation Error(MRE). Model Fit Error represents how well the model fits the data points, that is, error between the actual data in the training dataset and the values predicted by the model. Model Representation Error(MRE) represents how well the regression model fits the actual response, that is, the error between the values in the testing dataset and the values predicted by the model. The lower the values of MRE and MFE are, it represents a good model fit. Moreover, in an ideal model, MRE value should be greater than the value of MFE since MFE is a calculation of error with the data points that we used in the model training process.

Both MFE and MRE of our model were evaluated by calculating the Root Mean Squared Error.

Our model resulted in approximately 2.83 in MRE and 2.76 in MFE. Since a good model fit has a greater value of MRE than MFE,

our MRE and MFE values indicate that our model is good.

```
MFE value: 2.7586677385532723
MRE value: 2.826593645508809
Model Evaluation: Good model!
```

5) Conclusion

Collectively, according to the 4 steps of model evaluation, we concluded that our model is in good fit.

3. Citation

1) R-squared value
https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/
2) Interpretation of residual by predicted plothttps://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/
https://datagy.io/seaborn-residplot/
https://stats.stackexchange.com/questions/87793/does-this-residual-plot-look-bad
3) Generating MLP model, Data preprocessing, Visualizing model Evaluation, Residual mean calculation
https://chat.openai.com/
4) visualizing predicted vs residual plot and actual by predicted plot
[Matplotlib Tutorial](https://wikidocs.net/92094)
5) Generating MLP model
[scikitlearn](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html)
6) Calculating residual for predicted vs residual plot.
[Qualtrix.XM](https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/)
7) drawing residual plot
[seaborn.residplot](https://seaborn.pydata.org/generated/seaborn.residplot.html)
8) Evaluation of model, Regression
[Introduction to Machine learning lecture notes by professor Junghyun Kim]

9) ReLu function
[Dive into Deep Learning](https://ko.d2l.ai/chapter_deep-learning-basics/mlp.html)
10) MLP
[ScienceDirect](https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron)
11) 노주만&조홍종, 「Multilayer Perceptron(MLP), Nonlinear Autoregressive exogenous (NARX) 모형을 이용한 계통한계가격(SMP) 예측」, 한국자원공학회지, 제57권 제6호, 2020, p587: MLP
12) <배광수 외>, 「이항로지스틱 회귀모형과 MLP 신경망을 이용한 신호위반사고 위험교차로 예측모형 개발」, 교통안전연구, 제36권, 2017, p.123: MLP
13) Spliting training set and testing set
https://scholarworks.utep.edu/cs_techrep/1209/
14) Data scaling
https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/