Regression modeling and Evaluation

Student Name: Eunji Hwang

Student Number: 22100809

Majors: Social Welfare & Data Science

June 16, 2024

Introduction

This project aims to develop a regression model that accurately predicts the velocity of Eastward winds based on geographical inputs: latitude and longitude. To achieve this, we plan to utilize a Multi-Layer Perceptron (MLP), a type of artificial neural network renowned for its ability to model complex nonlinear relationships effectively. This approach is informed by the theoretical knowledge acquired in class, ensuring a robust foundation in both the conceptual and practical aspects of machine learning. Moreover, we will conduct a comprehensive evaluation using a test dataset to verify the predictive accuracy and reliability of our model, thereby confirming its applicability in real-world scenarios.

Project Deliverables:

1. ***Create an MLP-based regression model (Input: Latitude and Longitude, Output: Eastward wind) 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.***

a) Data Preparation

The given data set consists of thousands of records about latitude, longitude, and the corresponding eastern wind speed. First, make sure that there are no missing values in the data: (data.isnull().sum().sum()). Next, longitude and latitude data were converted into floating-point format (float32) to facilitate uniformity and faster calculation.

After that, to create a regression model, the feature (longitude and latitude) and the target variable (east wind m/s) were separated and stored as x and y variables, respectively. Since geographic coordinates and wind speed scales vary, scaling was performed by applying standardization using StandardScaler. This scale improved the stability and performance of the neural network during training by adjusting the data so that the mean was 0 and the standard deviation was 1.

b) Data Splitting

To perform data segmentation, train_test_split in the sklearn.model_selection library was used. In this case, it was designated as test_size=0.2, random_state=42, which 80% is assigned for training and 20% is assigned for testing. This ensures that the model has a significant amount of data to learn while maintaining a significant portion for an unbiased evaluation of the model's performance. Therefore, when 20% is extracted, testing 989 at a total size of 4947 is sufficient to obtain the amount of data. In addition, the random_state parameter is set to check whether segmentation is reproducible, which facilitates consistent evaluation and debugging by generating the same random segmentation each time the code is executed.

c) MLP Model Structure

The MLP model is defined using PyTorch's neural network module (nn.Module). The model comprises:

i. **Input Layer**: Receives the scaled latitude and longitude features. The dimensionality corresponds to the number of features, which is 2.

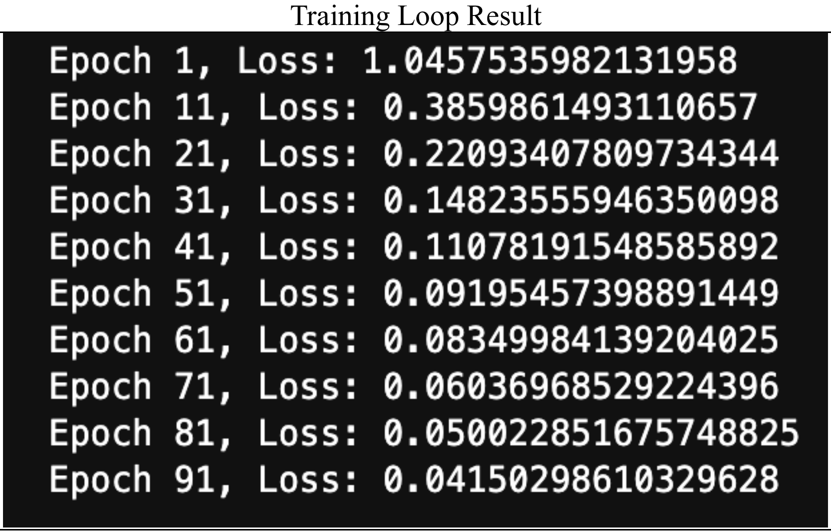
ii. **Hidden Layers**: Consists of two hidden layers. The first hidden layer has 64 neurons, and the second hidden layer has 128 neurons. Each neuron in these layers applies a linear transformation followed by a non-linear activation function, ReLU (Rectified Linear Unit), which introduces non-linearity to the model, allowing it to learn more complex patterns.

iii. **Output Layer**: The final layer of the network is the output layer, which consists of a single neuron. This neuron outputs the predicted value of the eastward wind speed

d) Training Process

The model training is conducted over 100 epochs. In each epoch: first, the model performs forward propagation, computing the wind speed predictions from the input features. Then, the loss is calculated using Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual wind speeds. Thus, backpropagation is executed to update the model's weights, minimizing the loss function using the Adam optimizer, a popular choice for its adaptive learning rate capabilities.

Training Loop Result



```
Epoch 1, Loss: 1.0457535982131958
Epoch 11, Loss: 0.3859861493110657
Epoch 21, Loss: 0.22093407809734344
Epoch 31, Loss: 0.14823555946350098
Epoch 41, Loss: 0.11078191548585892
Epoch 51, Loss: 0.09195457398891449
Epoch 61, Loss: 0.08349984139204025
Epoch 71, Loss: 0.06036968529224396
Epoch 81, Loss: 0.050022851675748825
Epoch 91, Loss: 0.04150298610329628
```

Based on Training Loop Result graph, each Epoch keeps repeated, the Losses are getting decreased. Loss values are continuously decreasing during the training process. This indicates that the model is learning from data for each epoch and gradually reducing prediction errors. In other words, loss reduction generally means that the network is adapting well to the training data. Also, as the loss initially decreases rapidly, the model captures the basic pattern at a high speed. This suggests that the network structure and learning rate setting are effective.

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.***

Calculated Test MSE	0.0372
Calculated R-squared	0.9620

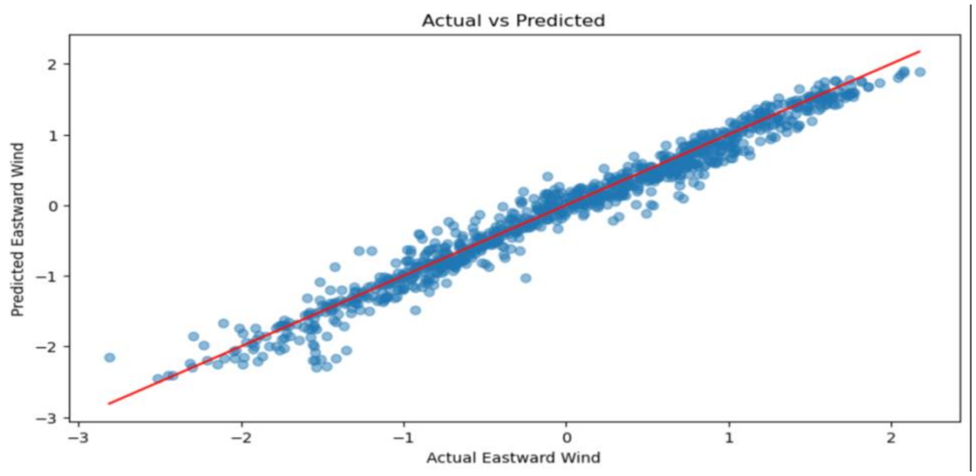
a) Mean Squared Error (MSE)

MSE is a common metric used to measure the mean square of error, i.e. the mean square difference between the estimated and actual values. In this case, the test set calculates the MSE between the predicted and actual wind speeds. In this case, calculation result was 0.0372. Thus, a low MSE means a smaller deviation between the predicted and actual values, suggesting that the model is very accurate in its prediction for the test data.

b) R-Squared (R^2)

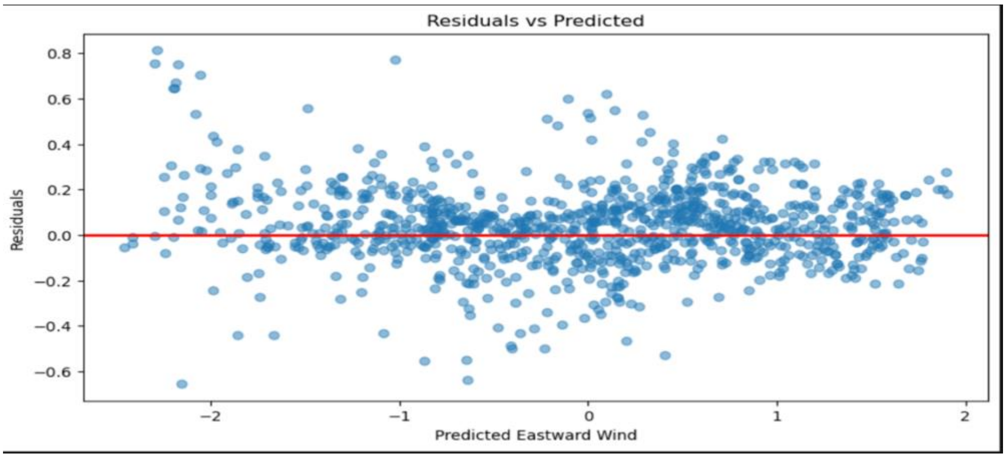
R-square is a statistical measure of the ratio of variance to a dependent variable, described as an independent variable or variable, in a regression model. This means that the fit is high, and a value close to 1 means that it fits better. Here, the value 0.9620 means that the model can account for about 96.20% of the variability in the data, which can be viewed as a very high score. Thus, a high coefficient of determination indicates that the model is predicting an outcome very similar to the actual data.

c) Visualization of model predictions

Actual by Prediction Plot	

This graph inputs latitude and longitude and outputs a prediction for the east wind speed. As can be seen from the range of values shown (-3 to +2), both the actual and predicted wind speed values have been standardized or normalized. In the graph, the x-axis represents the actual east wind speed obtained from the test data set, and the y-axis represents the predicted east wind speed, which is the result calculated from the MLP model. Each point in the figure represents an individual prediction whose location is determined by the actual and predicted values. The red line in the graph ideally indicates where the prediction exactly matches the true value. The closer the blue point (the actual prediction) is to this red line, the more accurate the prediction becomes.

Analysis of the graph suggests a high level of accuracy, as evidenced by the concentration of points along the red line. This alignment indicates that the model's predictions closely mirror the actual values across the entire range of data. Additionally, the even distribution of points across the graph demonstrates the model's consistent performance across varying wind speeds.

Actual by Residual Plot	

Residuals are calculated to subtract the predicted value from the actual value and represent the error of the prediction. In the graph, the x-axis represents the predicted east wind speed, and the y-axis represents the residual of each prediction. A point on the graph represents the residual that is associated with a particular predicted value. The graph allows us to observe whether the residual is patterned over the predicted range. A line drawn in red allows us to distinguish whether the residual is positive or negative, and if it is positive, it means overestimating the residual, and if it is negative, it means underestimating the residual.

Analysis of graph suggests that graph is randomly distributed and not biased to one side. That is, the residuals are randomly distributed around the red line. However, it seems necessary to determine outliers or investigate additionally when several points are plotted at the top or bottom.

3. *Discuss the results of the model evaluation.*

First, for the quantitative evaluation, based on the previously mentioned metrics, the model exhibits a Mean Squared Error (MSE) of 0.0372 and an R-squared value of 0.9620. The low MSE indicates that the model's predictions are close to the actual values, with a minimal average squared error. The high R-squared value suggests that the model explains 96.20% of the variance in eastward wind speeds, underscoring its effectiveness in capturing the dynamics of the dataset.

Second, for visual evaluation, the Actual VS Predicted plot showing that points closely aligned along the diagonal line suggests that the model has a high predictive accuracy. The data points are tightly clustered around this line, which visually confirms the high R-squared value. For Residuals vs. Predicted Plot, showed some points that were significantly off from the zero line, especially at the extremes of predicted values. In this case, presence of significant residuals at the extremes could suggest that the model struggles with extreme values or outliers.

Therefore, to improve the model, considering the model's response to extreme values and outliers, a few approaches can be contemplated. Implementing robust scaling, transforming the target variable (e.g., log transformation), or incorporating normalization could be viable strategies to enhance the model. Additionally, employing cross-validation techniques would provide a more comprehensive evaluation of the model's performance across various subsets of the data.

Reference

- OpenAI. ChatGPT. https://openai.com/chatgpt
- Lecture Notes on Module 7-2-2 Artificial neural network II
- Lecture Notes on Module 7-4-1 Regression model evaluation
- Yeturu, Jahnavi. (2019). Analysis of weather data using various regression algorithms. 117-141.