Student_exam_dataset

Notebook link: ML_Student_Exam.ipynb

Brief description of each dataset and tasks

- ***Description***: This dataset is about student exam final scores based on factors. The data includes information such as: Sleep Hours, Previous Scores, Hour Studied,...
- ***Tasks***: Our task is to build a model to predict the final score based on provided information.

Summary of model architectures and training strategies

a. Model architecture:

- The model architectures I used were 1 **Relu** layer, 1 **Dropout** layer, and 1 **linear** layer.

➔ The reason why I used this model architecture is that:

- RELU: Because it is fast and safe
- Dropout: As mentioned in class, Dropout might make the learning process more efficient by creating more difficulties for the model
- Linear: Because our label is a real number

b. Training Strategies:

- My approach was to clean all the data, followed by splitting the train and the test set. Then I did the preprocessing process before actually training the model, and finally ended with validating and testing the model. Along the way, I did add EarlyStopping and ReduceLearningRateOnPlateau to make sure the learning process was 'safe'. But this time, instead of using accuracy as metric, I used mse and rmse

Comparative analysis of performance and feature importance

a. Analysis of performance:

- The model stopped at epoch 267, with loss: 41.6757 - mae: 5.0311 - rmse: 6.4318 - val_loss: 12.1180 - val_mae: 2.9072 - val_rmse: 3.4811 - learning_rate: 5.0000e-04
- Test MAE: 2.92 points
- Test RMSE: 3.62 points
- Test R2: 0.752

➔ Not a good accuracy for such a small dataset like this, so there might be some problems. Moreover, it seems to be sensitive to outliers (not a non-sleep disorder).

b. Feature Importance:

Feature Importance Table (based on permutation importance):

	Feature	Importance (mean)	Importance (std)	Importance (proportion)
0	regular_activity_hours	-1.611957	2.518326	-0.014680
1	attendance_percent	7.760543	3.057186	0.070675
2	sleep_hours	13.162247	4.429964	0.119869
3	previous_scores	16.636848	2.454280	0.151512
4	hours_studied	73.857986	10.938229	0.672625

- Similar to what we find in our EDA, hours_studied is the most important factor. But once again, the newly created feature, regular_activity_hours is not so significant.

Insights into what you discovered in your experiments

- The importance of choosing the correct batch size. At first, I defaultly chose a batch size of 32, which makes my result much worse compared to when I use a smaller batch size.
- Feature Engineering or Extraction is not a must in deep learning. This time, one more time, confirm this statement for me.
- Different situations need different approaches. This time, I had to use metrics as mae and rmse instead of accuracy, since I was working with real numbers.