3.1 Liner regression

A linear regression model is constructed to compare the difference between using data that is normalised and using normalised data, and to compare the difference between using all the variables and using only the two most relevant variables. In addition, in order to reduce the disturbance caused by randomness, to a more reliable estimate of the model performance. The experiment was run 30 times and the mean and variance of RMSE and R2 were calculated for a more reasonable assessment of the model fit. The results are shown in Table 3.1.1.

Tible 3.1.1 different type of liner models RMSE and R2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Value | Data\_set | all\_unnorm | all\_norm | 2\_feat\_unn | 2\_feat\_norm |
| mean\_rmse | train | 2.183512011 | 2.183512011 | 2.502071386 | 2.502071386 |
| std\_rmse | 0.032705781 | 0.032705781 | 0.039948721 | 0.039948721 |
| mean\_rmse | test | 2.234785573 | 2.234785573 | 2.5213425 | 2.5213425 |
| std\_rmse | 0.051467146 | 0.051467146 | 0.060533765 | 0.060533765 |
| mean\_r2 | train | 0.539888382 | 0.539888382 | 0.395903699 | 0.395903699 |
| std\_r2 | 0.011478221 | 0.011478221 | 0.012943363 | 0.012943363 |
| mean\_r2 | test | 0.520481816 | 0.520481816 | 0.390007884 | 0.390007884 |
| std\_r2 | 0.022482515 | 0.022482515 | 0.01938759 | 0.01938759 |

For all\_unnorm and all\_norm (models using all features, unnormalised and normalised). The RMSE and R² for the training and test sets are almost the same, 2.18 and 2.23 (RMSE) and 0.54 and 0.52 (R²), respectively. The two models are equally effective in explaining the variance of the training and test sets, while their results on normalisation or not are not very different, indicating that normalisation has little effect on the results.

While 2\_feat\_unn and 2\_feat\_norm (using the two most relevant features, unnormalised and normalised): the RMSE is higher for the training set and the test set, 2.50 and 2.52 respectively, and the R² is lower, about 0.39. This indicates that the model performs poorly and the error increases when only two features are used.

So all\_unnorm and all\_norm are the best choice because they have the lowest RMSE and highest R².

The reason why normalisation does not have much effect on the corresponding model for this data is probably because normalisation does not have much effect on linear regression, which is not sensitive to the scaling of the features, and there is not much difference in the scale of the features. Since the RMSE is calculated based on the original value of the target variable y, not on the normalised feature X, the RMSE before and after normalisation usually does not change much, or may not even change at all.

3.3 MLP Neural networks

This paper attempts to find the neural network construction method that is most suitable for predicting this data. Therefore, three mainstream methods for building multilayer perceptron are used, i.e., using TensorFlow library, Kears library and using Sklearn library respectively. At the same time, in order to maximise the prediction effect of each model, we allow different hyperparameters for each model, i.e., we find the most suitable case for the model by using different permutations of the learning rate, the number of hidden layers and the number of neurons. The optimal hyperparameters found for the different models are given in the table 3.3.1:

Tible 3.3.1 hyperparameters of different models

|  |  |  |  |
| --- | --- | --- | --- |
| hyperparameters | Tronsflow | Kears | Sklearn |
| neurons | 128 | 128 | 10, 5 |
| hidden layers | 2 | 1 | 2 |
| learning rate | 0.01 | 0.01 | 0.001 |

Through continuous testing, the most suitable values for the corresponding models were found and the best hyperparameter values were determined for different models.

The different models are run 30 times respectively and the average of the returned RMSE and R2 values are used to determine which training model to use is more suitable for the present model, the results of which are shown in Table 3.2.

Tible 3.3.2 RMSE and R2 of neural networks constructed by different methods

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Value | Data | Tronsflow | | Kears | | Sklearn | |
| set | mean | std | mean | std | mean | std |
| RMSE | train | 2.180065492 | 0.122532131 | 2.203645055 | 0.03445195 | 2.078268605 | 0.032046049 |
| test | 2.224939989 | 0.12533724 | 2.22817351 | 0.055788159 | 2.120992627 | 0.049290777 |
| R2 | train | 0.523608921 | 0.052074454 | 0.531379775 | 0.011368161 | 0.583188681 | 0.009965963 |
| test | 0.540080013 | 0.053893599 | 0.523691322 | 0.013207534 | 0.568274816 | 0.015204515 |

For Transflow, the average RMSE of the training set is 2.180 and the average RMSE of the test set is 2.225. The RMSEs of the training and test sets are close to each other, which means that the model is fitted in a stable way and there is no obvious overfitting or underfitting. Meanwhile, the R² of the training set is 0.524, and the R² of the test set is 0.540, the difference between the R² of the training set and the test set is small, which indicates that the model does not have obvious overfitting phenomenon. However, the R² value is slightly lower, indicating that the fit is not optimal.

For Kears, the RMSE of the training set is 2.203 with a standard deviation of 0.034, and the RMSE of the test set is 2.229 with a standard deviation of 0.055. The RMSEs of the training and test sets are very close to each other, and the standard deviation is small, and the model shows high stability. Besides, the R² of the training set is 0.532 and the R² of the test set is 0.524. The R² value is slightly higher than that of Transflow, which indicates that the fitting effect is better, and the R² values of the training and test sets are close to each other, which means that the model is more stable.

In Sklearn, the RMSE is 2.078 for the training set and 2.121 for the test set. it has the smallest RMSE value, which indicates that it has a lower error than the other two models, which suggests that it has the best fit on both the training and test sets. The training set R² is 0.583 and the test set R² is 0.568. it has the highest R² value which indicates that it fits best.