Methods for predicting student performance on a linear regression model and neural network

Pierfrancesco Diella
Antonio Rana
Politecnico di BARI. DMMM department.
Bari, Italy

Abstract

Predicting student performance is essential to improve the effectiveness of one's study method. In this report, predictive models based on linear regression and neural network are developed and compared to predict college student performance based on a range of socio-demographic, academic, and behavioral variables. Using a dataset of college students, including information such as age, gender, parents' education level, and study habits, we trained and tested two models to predict students' final grade. During the development process of both models, normalization and regularization techniques were examined and used to optimize the models performance. The results show that both models exhibit good predictive ability, with the neural network showing slightly higher performance than linear regression. This study aims to provide a valuable contribution to research on student performance analysis, offering a basis for future insights into optimizing predictive models in the context of education.

CCS Concepts

• Computing methodologies → Classification and regression trees; **Neural networks**; **• Social and professional topics → Information systems education**.

Keywords

Prediction, Student, Performance, Multi linear regression, Neural network, Big data analytics

1 Introduction

Education is a crucial stage in the development of individuals and preparing for professional success. In recent years, interest in analyzing academic data and predicting student performance has grown significantly, as academic institutions and researchers seek to identify key factors that influence college student success. The availability of large academic datasets provides an unprecedented opportunity to apply advanced big data analytics techniques in order to better understand the educational process and improve teaching practices.

In this study we aim to develop and compare predictive models able to predict the performance of university students using linear regression techniques and neural networks, evaluating the effectiveness of the proposed models in predicting the final grade of students.

The objectives of this study are:

(1) To develop a linear regression model to predict the final grade of students.
(2) Implement a neural network to predict college students' performance.
(3) Compare the performance of linear regression and neural network models in order to try to determine which model is more effective.

The entire code of this study is available in the appendix at the link provided

2 Related work

Student performance analysis is a topic of growing interest to academic institutions and researchers. A review of predictive data mining techniques, used to analyze the factors that influence student performance in higher education, highlighted the importance of advanced data analysis techniques in providing valuable insights to improve the effectiveness of higher education. (Abu, Mostafa, & Khaled, 2023)

Other studies have demonstrated the effectiveness of predictive models even on small datasets, as in the case of this analysis, suggesting that even with limited data it is possible to achieve significant results in predicting student performance. (Zohair, 2023)

In addition, the crucial role of students' self-esteem in influencing their academic performance in higher education was highlighted.(Dinther, Dochyb, & Segersc, 2023)

Through the combination of these studies and available online resources, it aims to deepen the understanding of the factors that influence the success of university students and to develop predictive models to predict their academic performance.

3 Method description

3.1 Dataset

The dataset used in this study was "Higher Education Students Performance Evaluation", (Yilmaz, 2023), that contains data collected through surveys in 2019 from students of the Faculty of Engineering and the Faculty of Education. The dataset is structured in such a way as to include several variables that can influence the performance of university students.

A first operation of visualization of the attributes and understanding of them, also guided by the useful information made available by the author of the dataset, suggested the variable "Grade" as a candidate to be the dependent variable of the models.

Subsequently, an analysis of the distributions of the other variables also did not reveal any particular common distributions. So, the next step was to look for a correlation of the independent variables with the "Grade" feature. The following results (see Tab.1) did confirm a correlation, although not too obvious in some cases.

Variable	Correlation
GRADE	1
GENDER	0.335533
CUML_GPA	0.315493
EXP_GPA	0.248588
READ_FREQ	0.195617
WORK	0.167445
LIKES_DISCUSS	0.146547
COURSE ID	0.142166
CLASSROOM	0.110617
HS_TYPE	0.104821
LISTENS	0.085137
#_SIBLINGS	0.084470
PREP_EXAM	0.073783
KIDS	0.066379
MOTHER_EDU	0.066318
FATHER_EDU	0.063504
NOTES	0.044862
SCHOLARSHIP	0.023963
LIVING	0.023683
PREP_STUDY	0.014638
READ_FREQ_SCI	0.003189
MOTHER_JOB	-0.030747
STUDY_HRS	-0.033065
FATHER_JOB	-0.044268
PARTNER	-0.051778
ACTIVITY	-0.062993
AGE	-0.095251
ATTEND	-0.139564
TRANSPORT	-0.156289
SALARY	-0.166352
ATTEND_DEPT	-0.184763
IMPACT	-0.203273

Table 1: Correlation degree of GRADE variable.

3.2 Proposed system

3.2.1 Multi-Linear Regression model

In this work, a linear regression model has been implemented in order to predict the "Grade" variable as a function of all other independent variables.

The first function defined is the Cost Function, which is calculated using the mean squared error between the model predictions and the actual output values. Next, the gradient of the cost function with respect to the weights W and the intercept b was defined. In the end, the gradient descent function was constructed to optimize the last function up to an absolute minimum of the cost function.

Due to the different distributions and scales of the variables, it has been proposed to use an L2 normalization. However, the results were not as satisfactory as expected, both because of the cost achieved in the order of 10^{-3} which was still too high, and because of an obvious problem of overfitting due to the excessive number of parameters compared to the number of observations. Therefore, the choice fell on the use of a built-in "lambda" regularization term, as follows in the cost function and also in the gradient.

The cost formula (1), in the final analysis, is:

$$J(W,b) = \frac{1}{2m}\sum_{i=1}^{m}\left(f_{W,b}\left(x^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n} W_j^2 \qquad (1)$$

Later, to find the best lambda term, they compared it to the cost achieved in about 100 runs of the model, and a lambda of 0.3 seems to be the most efficient.

The models were trained on the training set (70% of dataset) using the descending gradient algorithm, iteratively, updating the model weights to reduce the cost function. After, the model was used to make predictions on the test set and evaluated using mean square error (MSE).

3.2.2 Neaural Network model

To further explore predicting student performance, a neural network-based model was implemented. Below is a detailed description of this model.

The model was built using the Keras library, with a sequential approach. A dense (fully connected) layer with 10 neurons and ReLU function as the activation function was added. Next, an output layer with 1 neuron and linear function activation was added. The L2 technique, already mentioned above, was applied to limit overfitting.

The model was compiled using the Root of Mean Square Error (RMSE) as a loss function and the optimizer "Adam", with a learning rate of 0.01. This configuration provides the necessary guidance to the model on how to optimize and minimize error during training.

Subsequently, the model was trained for a total of 50,000 epochs

4 Result and discussion

In this phase of evaluation of the implemented models, the performance achieved was carefully analyzed, through different metrics and evaluation approaches.

As for the linear regression model, in order to evaluate its ability to make accurate predictions, predicted values were calculated with the parameters obtained by the ultimately developed model. These were first evaluated through the mean square error (MSE) compared to the actual values of the test set and this provides an indication of the average discrepancy between the model predictions and the actual values of the target: the value of this metric turned out to be about 2.98. This is hardly an astounding result, but it is still about half of what was achieved by training the same model without regularization. The value of the cost, ultimately using (1) is around 10^{-7}. Next, these results were visualized through a graph (see Fig.1) showing the actual values of the target, the model predictions, and the error lines between the actual values and the predictions, to better highlight this discrepancy between actual and predicted values.

In addition, the results of the neural network model were analysed. After training the model and making the predictions, the performance was evaluated by comparing the predictions with the actual values of the target, in particular the first predictions were compared with the expected values, as shown on Tab.2.

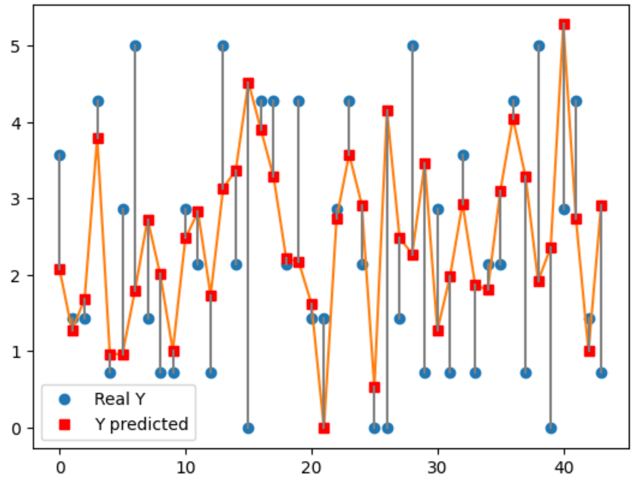

Figure 1: Error between real Y and predicted Y

Predicted Grade	Expected Grade
1.52	1.00
2.68	1.00
1.23	1.00
1.49	1.00
1.71	1.00
2.45	2.00
3.85	5.00
1.01	2.00
5.33	5.00
0.03	0.00
2.33	2.00
0.29	0.00
0.21	0.00
1.70	1.00
2.11	2.00
1.86	2.00
2.98	1.00
2.32	2.00
2.23	2.00
3.13	3.00
1.65	1.00
1.06	1.00
2.56	3.00
0.98	1.00
3.20	2.00

Table 2: Comparison of expected and actual results.

5 Conclusion

Through the analysis of the data and the implementation of these models, considering the results obtained and the evaluations made on them, the following conclusions can be reached.

The implemented linear regression and neural network models proved to be effective in predicting student performance. Both models were able to capture complex relationships between input variables and students' final grade. However, the neural network model seems to perform better, especially considering that increasing the number of epochs, compared with increasing the number of iterations in regression training, would lead to a faster improvement in performance.

In addiction, the strong practical applicability of these models is evident: The results obtained can be used to support targeted decisions and interventions within the study. And at this point the possible future developments of this study are also gathered: if we focus in a second analysis on the factors that influence student performance, this more than anything else could be used to develop personalized support strategies. Therefore, this work is a solid basis for further research in this field, as well as a valuable contribution to understanding of the dynamics that influence the performance of university students.

A Code

The Python code to developing the described analyses is available at the following link, with also the starting dataset.

https://github.com/PierDiella/Big-Data-Analytics

B Bibliography

(1) Abu, S. A., Mostafa, A.E., & Khaled, S. (2023). Factors Affecting Students' Performance in Higher Education: a Systematic Review of Predictive Data Mining Techniques. Springer Nature.

(2) Dinther, M. V., Dochyb, F., & Segersc, M. (2023). Factors affecting students' self-efficacy in higher education.

(3) Yilmaz, & Sekeroglu. (2023, July 10). From kaggle.com: https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation

(4) Zohair, L. M. (2023). Prediction of Student's performance by modelling small dataset size.