Optimizing Student Success Predictions using Artificial Intelligence

Sameeksha Vashishtha

08/20/2024

Inspirit Ai 1:1 Mentorship Program

Abstract

The education field and literacy rates remain popular subjects of discussion within both the research community and the AI field, as people continually seek ways to promote learning and enhance literacy. This work aims to examine the extent of how predictions of students' success in school can be improved by using AI models and neural networks with respect to student performance, and consequently study more effective resource allocation among students. We concentrate on exploiting artificial neural networks to predict student results; using information about academic history, socio-economic status, attendance rates and engagement levels etc., which inform the allocation of resources like educational materials, teachers, technology devices, & mentors accordingly. We tested different models, including Linear Regression, Logistic Regression and Decision Decision Trees, and neural networks with dense layers. The study finds that using advanced predictive models can greatly increase the types of educational support we offer, reduce disparities in student performance, and address issues revolving around how and where resource allocations may be most equitable. Therefore, we propose a strand of research to mitigate educational resource imbalances and better support low-attainment students with targeted interventions informed by predictive analytics.

1. Introduction

From an educational data science perspective, tackling discrepancies in student performance and resource allocation are perhaps the key drivers of inequality within education as a whole. In this research, we determine if it is possible to train a neural network model in order for school districts to accurately predict student outcomes and allocate limited resources more effectively towards low-performing / under-resourced students. This research seeks to weaponize the new predictive wealth of analytics technologies on the ongoing problem of educational inequity. Achievement gaps are created as a result of the inequitable distribution of resources, which leaves students needing more resources, with less help and support. In this study, we try to address these problems with a data driven approach by developing a neural network model to predict student success.

The study mobilizes supervised regression analysis to predict how students will do on quantitative metrics of student performance. These ranges of variables include academic background, socio-economic status, attendance patterns and levels of engagement. The outputs of the model will be performance predictions and advice on resource allocation, related to allocating resources, such as funding, technology devices for a classroom experiment, or opportunity assignments such as projects, internships proposals submissions, and mentorship programs. This is to leverage these predictions in decision-making processes within academic institutions.

This research seeks to increase the efficiency and effectiveness of support systems for under-performing, under-resourced students by incorporating predictive insight into resource allocation strategies. Superior and consistent outcomes are essential if we want to move the needle on narrowing achievement gaps, and making thereby better use of our resources to impact students who need access the most. Remedying current inequalities further creates a more just

and healthy culture of learning within the educational system by utilizing research-proven interventions.

2. Background

Artificial intelligence, machine learning specifically, has already been transforming education and re-modelling the way we teach and learn. Recent studies signify how further research can improve education. Nevertheless, there are still some limitations.

Research has found that making education interactive and personalized could be done naturally, with the help of adaptive learning [2]. These technologies help drive engagement by embracing personalized learning as well as predictive data through which we can identify learners who are most likely to fail, and at the same time they make assessment management easy. Challenges could range from data privacy concerns to how to make sure all students have the same access technology, and thus not widen a digital divide being observed across geography around the globe.

Another study found that an ML-based resource-matching model can achieve educationally efficient resource allocation [3]. It had good recollection and coverage fees which confirmed its capability to handle lengthy-tail issues, mainly facts sparsity. However, its overall performance is statistics satisfactory based and the generalizability of this technique in diverse academic contexts have no longer been explored.

A systematic literature evaluation regarding system learning techniques of predicting pupil performance demonstrates how these algorithms are beneficial [4]. An evaluation of the literature unearths behavioral targeting strategies that are effective at threat identification when analyzing information from a couple of assets to enhance outcomes. However, given the extensive range of research and methodologies employed in this evaluation, restricted evidence derived from these effects can be generalizable throughout a variety of populations.

To summarize, AI and systems gaining knowledge are particularly powerful equipment for determining a problematic aspect of schooling from personalized catering to powerful useful resource allocation and predicting performance. Overcoming the hurdles related to data privacy, getting right of entry to, and high-quality will play a function in making sure they meet their genuine well worth.

3. Dataset

The dataset used for this study was taken from the Faculty of Engineering and Faculty of Educational Sciences students at University of California - Irvine [7]. This dataset included 145 samples and it has 31 features which are the personal details of the samples: student id, age, sex, type of high school, scholarship type, additional work, artistic or sporty, partner or not, total salary, transportation type, accommodation type, parents' education, number of siblings, parental status, parents' occupation, weekly study hours, reading frequency (non-scientific books/journals/scientific books/journals), attendance, impact of your projects, preparation to midterm exams 1 and 2, notes, listening in classes, discussion, flip-classroom, GPA course id, and output grade. This dataset includes both numerical and categorical variables and a pre-processed version can be seen in Table 1.

In order to pre-process the dataset, we started by splitting up the data for tables and rearranged features in new manners. A better insight visualization from the data were obtained when we extracted and discretized it in plots on some features distributions or relations. We split the data (80%) as training and (20%) for the test model to train and test. The pre-processing involved creating the feature matrix *X* and target variable *y*, with *X* consisting of all columns except the last one, and *y* being the 'GRADE' column.

INDEX	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	2	2	3	3	1	2	2	1	1	1	1	2	3	1	2	5	3	2	2	1	1	1	1	1	3	2	1	2	1	1
1	2	2	3	3	1	2	2	1	1	1	2	3	2	1	2	1	2	2	2	1	1	1	1	1	3	2	3	2	2	3
2	2	2	2	3	2	2	2	2	4	2	2	2	2	1	2	1	2	1	2	1	1	1	1	1	2	2	1	1	2	2
3	1	1	1	3	1	2	1	2	1	2	1	2	5	1	2	1	3	1	2	1	1	1	1	2	3	2	2	1	3	2
4	2	2	1	3	2	2	1	3	1	4	3	3	2	1	2	4	2	1	1	1	1	1	2	1	2	2	2	1	2	2

Table 1: Pre-processed version of the data. The last column is not considered and variable y is the grade column

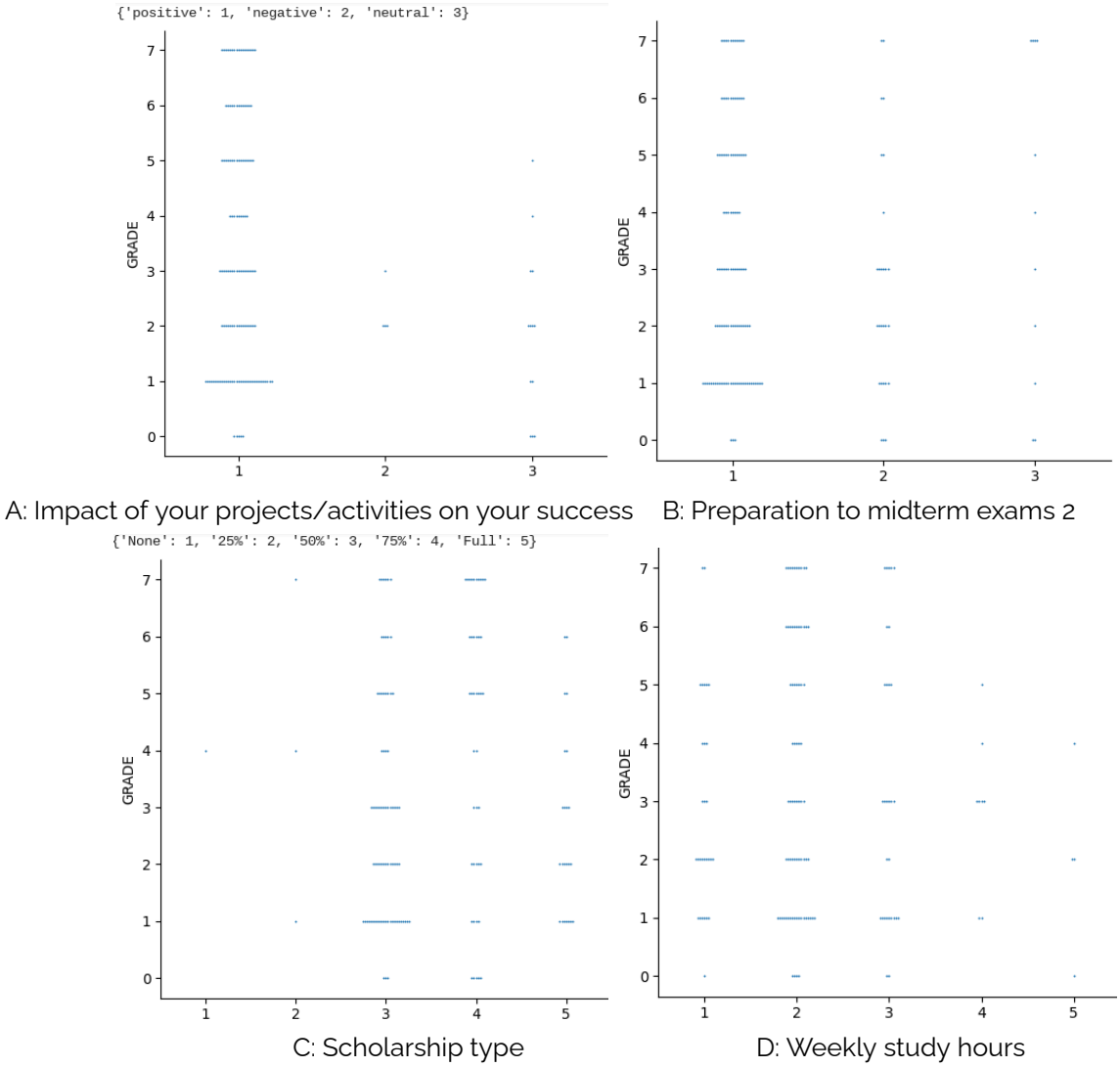


Figure 1: Different examples of swarm plots for different features.

An example, the first plot was the column named "Impact of your projects/activities on your success". From this swarm plot we can understand if their project/activity influenced their overall grade or not (positive: 1, negative: 2, neutral: 3).

4. Methodology / Models

We sought to address the student performance prediction and resource optimization problem with a variety of machine learning algorithms, each having different strengths in adapting the data. In this part, we illustrate models used and some steps taken.

The dataset used features including demographic information, academic background status, socio-economic data and engagement metrics. Categorical variables ("Parental Education" and "Accommodation Type") were one-hot encoded. The dataset was further separated into an 80% training set, and a remaining of 20%, testing the model architecture used.

4.1 Logistic Regression

Logistic regression is a type of statistical model, it sees the past data and provides results in binary itself with simply yes or no. We first pre-prepared the data and then passed it into the 'LogisticRegression' model from sklearn library followed by printing accuracy.

4.2 Linear Regression

Linear Regression is a model in which, depending upon the already known data values, one tries to advance a line equation for an expected value of new data.

4.3 Decision Tree Regressor

Decision tree regressor is a model where the known data is structured in the form of a tree during training and when we need to predict data, it gives continuous data meaning that the output is not just the known data. The criteria we used within the decision tree was 'friedman_mse'. This uses mean squared error with Friedman's improvement score for potential splits.

4.4 Neural Networks

Neural networks used to teach computers how to do things in the same way as a human brain processes information. Deep learning is a subfield of machine-learning theory that uses networked nodes or neurons arranged in layers to function like the cells found in our bodies and brains. We started by creating a network with split hidden layers of size [50, 60, 87]. We constructed a second neural network, with a batch size of 16, and the number of epochs set to EPOCHS. The 'livelossplot' library was used to visualize the loss and accuracy metrics during the training process.

5. Results and Discussion

This research focused on developing and evaluating various machine learning models for predicting student performance based on combined academic history, socio-economic status, attendance rates, as well as engagement levels. Some of the tested models are Logistic Regression, Linear Regression, Neural Networks, and Decision Tree model, which yielded the most accurate results.

5.1 Logistic Regression

For predicting pass or fail for students, a logistic regression model was used, achieving an accuracy of around 31%. This gave the model its low predictive power due to it being too simplistic. This was limited by low accuracy and the difficulty in handling more complex interactions between features.

5.2 Linear Regression

As another baseline, we used Linear Regression to predict the grades of students. It showed weak performance with only 0.06 accuracy, but this was expected as this is a simple method. The output of this model showed how this model can not work with the complex features in the dataset.

5.3 Decision Tree Regressor

The most significant model was the Decision Tree with an accuracy of 63%.

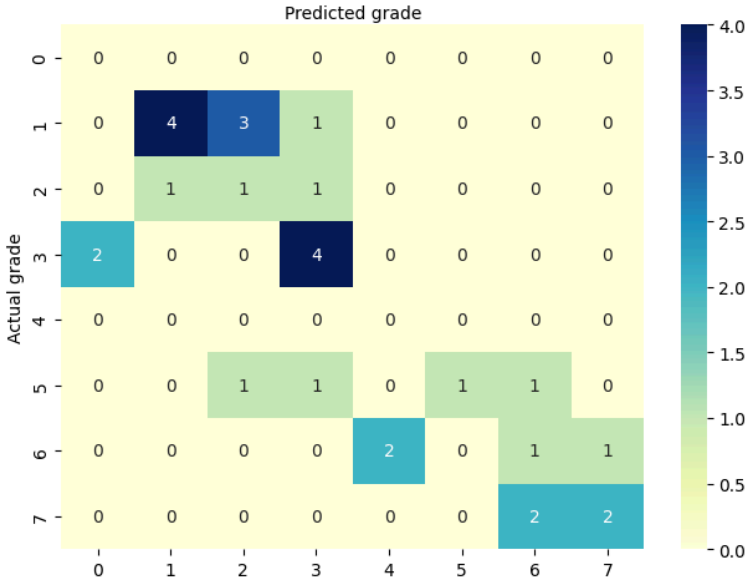


Figure 1

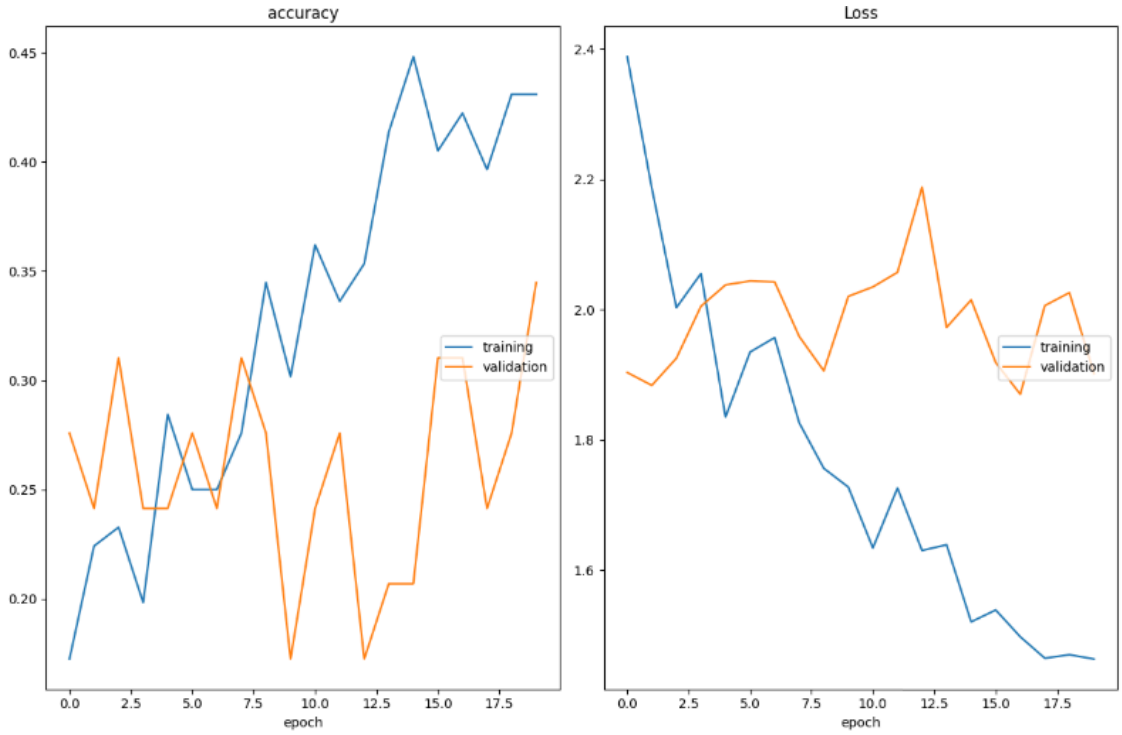
This is the confusion matrix used to visualize the model's performance and interpretability. This shows a high true positive rate, but with some false positives, indicating occasional overprediction of student success. Below are the results of the model.

Metric	Value
Criteria	`friedman_mse`
Accuracy	`63.13%`
Mean Absolute Error	`0.965`
Mean Squared Error	`1.931`
R^2 Score	`0.602`

Table 1

5.4 Neural Networks

Neural networks were used in this study because they have the capability of capturing sophisticated patterns within data. The model was layered deeply. Problematically, it was sometimes overfitted and did not always outperform the Decision Tree. While it could catch fine-grained patterns, it was not always better than simpler models. Future work could scale this up with a larger dataset and some regularization techniques.



6. Conclusions

This study sought to use more advanced AI models such as neural networks, in order to anticipate the performance of students and optimize resource distribution within education. We aimed at creating a machine learning model that could predict academic outcomes based on prior history both academically and socio-economically, attendance & participation levels. Although the results from models such as Linear Regression, Logistic Regression, and Neural Networks were promising, the Decision Trees model was found to be more accurate. When it came to predicting school grades in the finer grain scale from 0–7, the accuracy scored lower results. Reducing the prediction to a simple success vs. fail classification can improve results and interpretive clarity.

In the future, exploring other models and improving the data preprocessing is a good step ahead. A smaller, higher-frequency dataset would improve the resolution of predictions while performing accurately than using this model over a larger dataset. Summarizing grades allows categories to be defined and analyzing other performance metrics areas may provide additional better alerts that can drive an improvement in resource allocation efficiencies. In the future we will need to deal with these issues to expand this research further.

Acknowledgments

I would like to thank my mentor Mr. Jose Reyes for his support and guidance on this research project. I would also like to thank the Faculty of Engineering and Faculty of Educational Sciences at University of California - Irvine for providing the data that fueled my project.

References

[1](n.d.). scikit-learn: Machine learning in Python — scikit-learn 1.5.1 documentation. [online] Available at: https://scikit-learn.org/stable/ [Accessed 13 August 2024].

[2]Applications of AI and Machine Learning in Education. (2023, November 16). Shiksha. [online] Available at: https://www.shiksha.com/online-courses/articles/applications-of-ai-and-machine-learning-in-education/ [Accessed 13 August 2024].

[3]Alshammari, M. M., Ahmed, T. M. Z., Alsaedi, R. A., & Alshamari, H. A. (2023). Development of an intelligent education system: A resource matching model based on machine learning. Smart Innovations, Systems and Technologies, 34(1), 1-10. [online] Available at: https://publications.eai.eu/index.php/sis/article/view/345/259 [Accessed 13 August 2024].

[4]Kotsiantis, S. B. (2021, September 16). A systematic literature review of student performance prediction using machine learning techniques. ERIC. [online] Available at: https://files.eric.ed.gov/fulltext/EJ1314372.pdf [Accessed 13 August 2024].

[5]Ofori, F. (2020, March 26). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. [online] ResearchGate. Available at: https://www.researchgate.net/publication/340209478_Using_Machine_Learning_Algorithms_to_Predict_Students'_Performance_and_Improve_Learning_Outcome_A_Literature_Based_Review [Accessed 13 August 2024].

[6]Pizur, A., Pizur, J., & Chugh, R. (2021). A comparative study of machine learning algorithms for predicting student performance in higher education. Studies in Systems, Decision and Control, 345, 259-270. [online] Available at: https://doi.org/10.1007/978-3-030-60154-1_16 [Accessed 13 August 2024].

[7]Yilmaz, N., & Şekeroğlu, B. (2023, August 14). Higher education students performance evaluation. UCI Machine Learning Repository. [online] Available at: https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation [Accessed 13 August 2024].

Notes:

All experiments, model implementations, and statistical results can be found here https://colab.research.google.com/drive/1-r5sx2pPJuqZTNQd-IUpjO2coz0BLLxe?usp=sharing