On the Explanation of AI-based Student Success Prediction

Farzana Afrin, Margaret Hamilton, and Charles Thevathyan

School of Computing Technologies, RMIT University
Melbourne, VIC 3000, Australia
s3862196@student.rmit.edu.au, margaret.hamilton@rmit.edu.au,
charles.thevathyan@rmit.edu.au

Abstract. Student success prediction is one of the many applications of artificial intelligence (AI) which helps educators identify the students requiring tailored support. The intelligent algorithms used for this task consider various factors to make accurate decisions. However, the decisions produced by these models often become ineffective due to lack of explainability and trust. To fill this gap, this paper employs several machine learning models on a real-world dataset to predict students' learning outcomes from their social media usage. By leveraging the SHapley Additive exPlanations (SHAP) to investigate the model outcomes, we conduct a critical analysis of the model outcomes. We found several sensitive features were considered important by these models which can lead to questions of trust and fairness regarding the use of such features. Our findings were further evaluated by a real-world user study.

Keywords: Student success prediction · Machine learning Explainability · User study.

1 Introduction and Background

The recent developments of AI in education have shown potential to make informed decisions. Student success is one of the important metrics applied to the performance of education service providers. Moreover, by predicting student success, the educators can come up with actionable plans ahead of time. Eventually, this could contribute to improving the overall student experience. Machine learning algorithms have been successful in predicting student success. In general, machine learning algorithms incorporate data from various sources to model student success at different stages of their academic journey [15, 7, 1, 9, 5].

With the proliferation of AI-based technologies, concerns have been raised about the incorporation of sensitive predictors in the automated decision-making process. This may influence making unfair decisions regarding student success. [2, 10]. The lack of explainability of the deployed models is one of the main reasons for this challenge, which eventually results in trust issues among the system users. Therefore, the AI systems need to be transparent and the users must

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_34

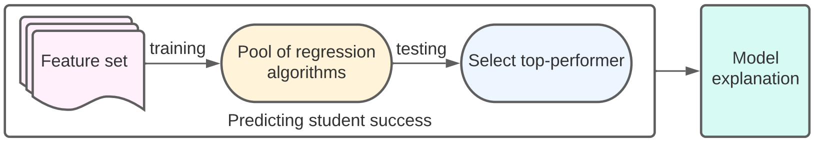

Fig. 1. An overarching view of the methodology

understand the extent to which they can trust the outcomes of the underlying machine learning technologies [12].

The need for explainable AI is gaining momentum in many industries [4, 11]. There are many directions of research targeting explainable AI. Some research is directed towards finding effective ways to explain the model outcomes such as local, global and counterfactual [16, 13, 3]. However, there is limited research on their applicability in educational settings. In this paper, we investigate this issue by formulating a student success prediction problem by utilising data related to students' social media usage and their demographics. We also investigate how such a model evaluates the features to be considered for making predictions. In particular, the contributions of this paper are as follows:

- Prediction of student success from their social media usage. In addition we provide a SHAP-based explanation of the prediction outcomes.
- Critical analysis of explainability for AI-based student success prediction. We also evaluate our findings through a real-world user study.

We discuss our methodological approach in Section 2 which is followed by the experiments and discussion of results in Sections 3 and 4 respectively. Finally, the paper concludes in Section 5 leaving some directions for future studies.

2 Methodology

Our approach contains two main segments: i) Student success predictions and ii) Explaining the predictions. An overview of the methodology is illustrated in Figure 1.

2.1 Student Success Prediction

For student success prediction, we train a collection of state-of-the-art regression algorithms to predict the student success (i.e., final exam marks). For training, we utilise a set of features representing information about students' social media usage times, demographic and background. We formalize our student success prediction problem as follows:

Let, $M_{fe} = \{M_{s1}, M_{s2}, ..., M_{sn}\}$ be the set of final marks of n students. In the final dataset, each instance x is described by a d-dimensional vector of attributes R^d related to the usage time of different social media by students, segregated by

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_34

the purpose of use, students' demographic information, and a final exam mark. If $f(.)$ is the success (i.e., final exam mark) prediction function for an unknown instance with d-features, $f(.)$ predicts $\hat{M}(x_q)$ such as $f(x_q) : R^d \rightarrow \hat{M}(x_q)$ where $\hat{M}(x_q)$ is the predicted final mark for a query instance x_q.

2.2 Explaining the Predictions

In this step, we consider the top-performing regression algorithm based on produced error, and then investigate the model outputs in terms of global feature-importance along with the relationship between each attribute and the target variable (i.e., final exam mark). A detailed description is given the in following subsections.

3 Experimental Setup

3.1 The Dataset

We utilise a dataset published by [14] which contains information about social media usage and final marks obtained of 505 students (221 males and 284 females) in a course. The dataset was collected from a large metropolitan Australian university, across three teaching session in 2017-2018.The subject, considered to build this dataset, is compulsory for students enrolled across business, commerce, law, engineering, science and information technology disciplines.

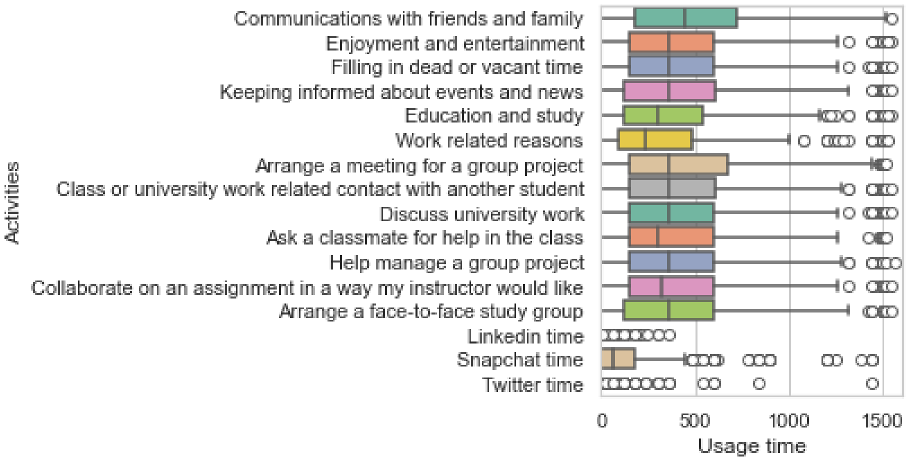

Fig. 2. Time distributions of LinkedIn, Snapchat, Twitter and different types of Facebook usage. Note that the breakdown of Facebook usage time was originally derived by multiplying the total usage time per day (in minutes) with the extent and likelihood students indicate for different reasons for Facebook usage.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_34

In particular, the dataset contains the usage times of Facebook, LinkedIn, Snapchat, and Twitter are logged in this dataset. The Facebook usage times are further decomposed into several purpose of usage. In addition demographic and background information including 'Age', 'Gender', 'WAM'(weighted average mark which is similar to grade point average) of the participant students are also provided. A box-plot illustrating the time distribution of LinkedIn, Snapchat, Twitter and different types of Facebook usage is given in Figure 2. We can see some flier points go past the upper whiskers. We also find a few instances of LinkedIn and Twitter usage times which are scattered in nature.

3.2 Prediction Results

We setup the experiment as per the formulation discussed in Section 2.1. We employ a pool of regression algorithms implemented in scikit-learn [8] including Linear, Random Forest, GBM, LGBM, and XGBM. We randomly split 80 % of the data from training these models and the rest are used for testing. The prediction outcomes are evaluated against two metrics: mean squared error (MSE) and root mean squared error (RMSE). As shown in Table 1, the performance of all the regressors is similar, however, the Random Forest produces slightly less errors than others in the pool.

Table 1. Prediction error by different models

Regression model	**MSE**	**RMSE**
GBM	0.025	0.012
LGBM	0.026	0.013
XGBM	0.026	0.013
Linear	0.025	0.013
Random Forest	**0.024**	**0.011**

3.3 Prediction Explanation

To investigate the model outcomes, we employ SHapley Additive exPlanations (SHAP)[1] which is a game theoretic approach to explain the output of any machine learning model. Figure 3 shows the global feature-importance (i.e., overall contribution of each feature in the model outcome) in terms of mean SHAP value. We can see that WAM, which is equivalent to grade point average, has highest predictive power which is followed by Age, Gender and Snapchat time.

We further investigate the relationships (i.e., positive and negative) between all the predictors and the target variable. We plot these relationships in Figure 4, where red and blue dots indicate a higher and lower features values respectively. We found that a higher value of WAM, age, gender (1-male, 0-female) results

[1] https://shap.readthedocs.io/en/latest/index.html

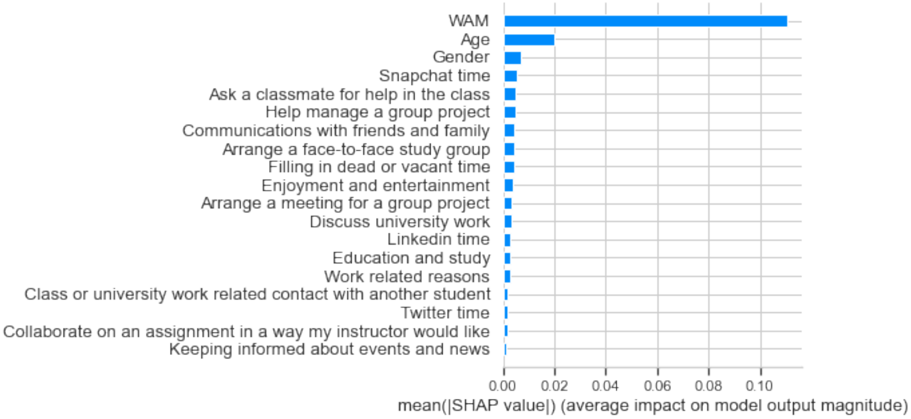

Fig. 3. Global feature-importance

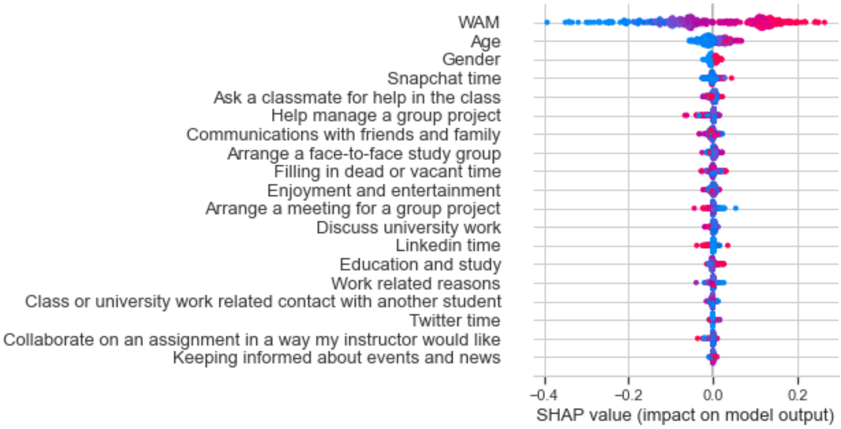

Fig. 4. Relationship between predictors and the target variable. Note: red and blue colors indicated high and low feature values respectively

in a higher SHAP value, which influences the prediction outcome positively. Next, we investigate how students perceive this proposition as previous research highlighted the sensitiveness of demographics and background related features.

4 Discussion and Analysis

In practice, the consideration of WAM, age, and gender in the modelling of student success may cause trust issues as these values are meant to be private and influence making an unfair decision. We further examine the perceived fairness through a real-world user study as discussed below.

DOI: 10.1007/978-3-031-08754-7_34

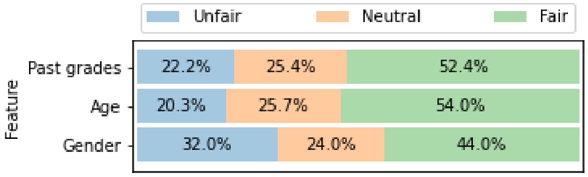

Fig. 5. Proportion of user perceptions regarding gender, age and past grades.

4.1 Investigating User Perceptions

We investigate user perceptions about the incorporation of various features in machine learning algorithms to predict student success by adapting a framework presented in [6]. The framework implements a survey where the participant students were given a set of features (predictors) related to demographics, internal evaluation, psychometric and course content, as highlighted in the recent literature that are commonly used to predict student success. Then we ask students whether they perceive a specific feature as fair or not when used to generate a model outcome. We collected and analyzed 1658 completed responses contributed by university students from Australia, Bangladesh and Saudi Arabia. We observed that the fairness perception of a specific features differs across cohorts of participants. Fig. 5, illustrates the proportion of perceived responses for three selected features out of a large pool.

5 Conclusions and Future Work

This paper investigates the explanation of model outcomes in predicting student success in terms of final exam score. We found that the Random Forest regressor was the best performing model in our pool, which considers WAM (previous grade point average), age, gender, snapchat usage time and time spent on asking friends for help as the top-5 features that influence the model outcome. The consideration of sensitive features (age, gender and past results) are an obvious reason that questions the fairness and trustworthiness of the deployed model as it is also verified by our user study consisting of 1658 tertiary students from three different countries.

Future research may consider ways to enhance model outcomes without considering these sensitive features, enabling trust and fairness in automated decision making. Future research also may include the use of counterfactuals which are capable of describing a causal situation (e.g. what values of a subset of features make a specific student perform better). Another direction could be to investigate user perception under various scenarios.

Acknowledgements

Farzana was supported through an Australian Government's RTP Scholarship.

References

1. Afrin, F., Rahaman, M.S., Rahman, M.S., Rahman, M.: Student satisfaction mining in a typical core course of computer science. AJSE **11**(1) (2012)
2. Baker, R.S., Hawn, A.: Algorithmic bias in education. International Journal of Artificial Intelligence in Education pp. 1–41 (2021)
3. Briz-Ponce, L., Juanes-Méndez, J.A., García-Peñalvo, F.J., Pereira, A.: Effects of mobile learning in medical education: a counterfactual evaluation. Journal of medical systems **40**(6), 1–6 (2016)
4. Gade, K., Geyik, S.C., Kenthapadi, K., Mithal, V., Taly, A.: Explainable ai in industry. In: Proceedings of the 25th ACM SIGKDD. p. 3203–3204 (2019)
5. Giunchiglia, F., Zeni, M., Gobbi, E., Bignotti, E., Bison, I.: Mobile social media usage and academic performance. Comp. in Human Behavior **82**, 177–185 (2018)
6. Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P., Weller, A.: Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In: Proceedings of the 2018 world wide web conference. pp. 903–912 (2018)
7. Liu, Z.: A practical guide to robust multimodal machine learning and its application in education. In: Proc. of the Fifteenth WSDM. p. 1646. New York, NY, USA (2022)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
9. Rodrigo, M.M.T., Baker, R.S., Jadud, M.C., Amarra, A.C.M., Dy, T., Espejo-Lahoz, M.B.V., Lim, S.A.L., Pascua, S.A., Sugay, J.O., Tabanao, E.S.: Affective and behavioral predictors of novice programmer achievement. In: Proceedings of the 14th ITiCSE. pp. 156–160 (2009)
10. Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V.M., Gasevic, D., Chen, G.: Assessing algorithmic fairness in automatic classifiers of educational forum posts. In: AIED. pp. 381–394 (2021)
11. Slijepcevic, D., Horst, F., Lapuschkin, S., Horsak, B., Raberger, A.M., Kranzl, A., Samek, W., Breiteneder, C., Schöllhorn, W.I., Zeppelzauer, M.: Explaining machine learning models for clinical gait analysis. ACM TCH **3**(2) (2021)
12. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., van Moorsel, A.: The relationship between trust in ai and trustworthy machine learning technologies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 272–283. FAT* '20, New York, NY, USA (2020)
13. Verma, M., Ganguly, D.: Lirme: Locally interpretable ranking model explanation. In: SIGIR. p. 1281–1284. NY, USA (2019)
14. Wakefield, J., Frawley, J.K.: How does students' general academic achievement moderate the implications of social networking on specific levels of learning performance? Computers & Education **144**, 103694 (2020)
15. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: Evaluating different sources of student data. International Educational Data Mining Society (2020)
16. Zytek, A., Liu, D., Vaithianathan, R., Veeramachaneni, K.: Sibyl: Explaining machine learning models for high-stakes decision making. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA (2021)