Exploring Counterfactual Explanations for Predicting Student Success

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. Artificial Intelligence in Education (AIED) offers numerous applications, including student success prediction, which assists educators in identifying the customized support required to improve a student's performance in a course. To make accurate decisions, intelligent algorithms utilized for this task take into account various factors related to student success. Despite their effectiveness, decisions produced by these models can be rendered ineffective by a lack of explainability and trust. Earlier research has endeavored to address these difficulties by employing overarching explainability methods like examining feature significance and dependency analysis. Nevertheless, these approaches fall short of meeting the unique necessities of individual students when it comes to determining the causal effect of distinct features. This paper addresses the aforementioned gap by employing multiple machine learning models on a real-world dataset that includes information on various social media usage purposes and usage times of students, to predict whether they will pass or fail their respective courses. By utilizing Diverse Counterfactual Explanations (DiCE), we conduct a thorough analysis of the model outcomes. Our findings indicate that several social media usage scenarios, if altered, could enable students who would have otherwise received a failing grade to attain a passing grade. Furthermore, we conducted a user study among a group of educators to gather their viewpoints on the use of counterfactuals in explaining the prediction of student success through artificial intelligence.

Keywords: Student success prediction · counterfactual explanation · social media usage behaviour.

1 Introduction and Background

Machine learning (ML) is a type of artificial intelligence (AI) that can learn and make predictions based on patterns in data. In the context of predicting student success, ML algorithms can be trained on historical data to identify factors that are associated with student success, such as demographic information, academic performance, and behavioral data [9]. These algorithms can then be used to

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

make predictions about which students are likely to succeed or struggle in the future [2, 3].

One of the challenges with using ML in the context of student success prediction is that the models used can be quite complex and difficult to interpret. This means that it can be hard to understand why the model is making certain predictions, which can be problematic for educators and other stakeholders who need to make decisions based on the predictions.

Therefore, it is important to build explainable ML models for student success prediction [1]. Explainability refers to the ability to understand how a model arrived at its predictions. By incorporating explainability into ML models for student success prediction, educators and other stakeholders can better understand the factors that are driving the predictions, which can help them make more informed decisions. This can ultimately lead to more effective interventions and supports for students who may be at risk of struggling, which can improve their chances of success. Additionally, improved explainability can help build trust in the use of ML by ensuring that stakeholders understand the basis for the predictions being made [7]. This can be achieved by using techniques such as feature importance analysis, which can help identify which features in the data were most important in making the prediction. Other techniques that can be used to improve explainability include decision tree analysis, which can provide a visual representation of the decision-making process, and local interpretable model-agnostic explanations (LIME), which can provide explanations for individual predictions. Another approach known as SHapley Additive exPlanations (SHAP) which can be used for both local and global explanations. However, these methods are often inadequate in fulfilling the specific needs of individual students in determining the cause-and-effect relationship of specific features associated with their success.

To address this gap, first we employ a set of state-of-the-art ML models to predict student success by leveraging their social media usage behaviour. Second, we employ a counterfactual approach which allows us to simulate different hypothetical scenarios enhancing the explainability of student success prediction outcomes. In our context of predicting student success, we investigate how changing the value of certain factors may have impacted student success outcomes. In particular, we investigate *"what factors needed to be changed if a failing student had to pass in the course?"*. Additionally, we investigate the practicability of adopting counterfactual explanation in student success prediction. To achieve this, we conduct a real-world user-study comprising educators of a tertiary institution. Specifically, we conduct a short survey to understand their viewpoint on using counterfactual explanations in their decision making process for the given scenario and beyond. The contributions of this paper are as follows.

- Predict student success (in terms of pass or fail) in a course by leveraging students social media usage behaviour. In addition, we provide a counterfactual analysis of the prediction outcomes.
- Conduct a user study to understand the practicability of employing counterfactual explanations in student success prediction.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

The organization of the paper includes the modelling and evaluation approaches of student success prediction in Section 2, which is followed by a counterfactual generation and analysis in Section 3. We report the findings of user-study in Section 4. The paper concludes in Section 5.

2 Predicting student success

2.1 Dataset and pre-processing

In our study, we made use of an open dataset from [8], which comprised data on the social media activity and final marks of 505 students (221 males and 284 females) who took a compulsory course across multiple disciplines including business, commerce, law, engineering, science, and information technology. The dataset was gathered from a large metropolitan Australian university over the course of three teaching sessions between 2017 and 2018. The dataset we used logs information on the usage times of Facebook, LinkedIn, Snapchat, and Twitter. Notably, the Facebook usage times are further broken down into various purposes of usage including 'Communications with friends and family', 'Enjoyment and entertainment', 'Filling in dead or vacant time', 'Keeping informed about events and news', 'Education and study', 'Work related reasons', 'Arrange a meeting for a group project', 'Class or university work related contact with another student', 'Discuss university work', 'Ask a classmate for help in the class', 'Help manage a group project', 'Collaborate on an assignment in a way my instructor would like', 'Arrange a face-to-face study group'. Note that the breakdown of Facebook usage time is calculated by multiplying the total usage time per day (in minutes) with the extent and likelihood students indicate for different reasons for Facebook usage which is denoted as proxy times in [8]. Additionally, the dataset also provides demographic and background information on the participating students, including their age, gender, and WAM (weighted average mark, which is akin to grade point average). As part of our data pre-processing, we categorized all final exam marks into two groups: pass and fail. Specifically, any marks of 50 or higher were classified as pass, while marks below 50 were labeled as fail. The final dataset is almost balanced where the number of pass and fail labels are 261 and 244 students respectively.

2.2 Modelling approach and Evaluation

The idea of student success prediction task is to identify whether a student will pass or not in a course by leveraging a set of features representing corresponding information about students' social media usage behaviour, demographic and background. A formal definition of our prediction task can be given as follows:

Let's say we have a set $G_{fe} = \{P, F\}$ that includes two possible final exam grades for n students in their course. In the final dataset, each instance x is a d-dimensional vector of attributes from R^d, which contains information about the students' usage of various social media platforms, demographic details, and

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

Table 1. Prediction performance by different models

Classification models	**Accuracy**	F_1**-score**
GBM classifier	0.74	0.74
LGBM classifier	0.75	0.75
XGBM classifier	0.72	0.72
Logistic regression classifier	0.66	0.66
Random Forest classifier	0.71	0.71
SVC	0.55	0.54

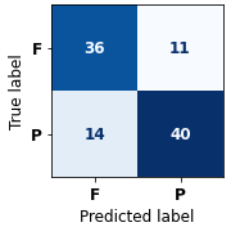

Fig. 1. Confusion matrix showing prediction performance of LGBM.

background information discussed in section 2.1. If we have a prediction function $g(.)$ that can forecast the success (final exam grade) of a student using their d attributes, we can represent $g(.)$ as $g(x_q) : R^d \rightarrow \hat{G}(x_q)$, where $\hat{G}(x_q) \in G_{fe}$ is the predicted final grade for student x_q whose grade was previously unknown.

For our student success prediction experiment, we employ a set of classification models implemented in scikit-learn [5] including Support Vector Classifier (SVC), Random Forest, Logistic regression and three variants of boosting techniques including Gradient Boosting classifier (GBM), Light Gradient Boosting classifier(LGBM), and Extreme Gradient Boosting classifier(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: accuracy and F_1 score. As shown in Table 1, the LGBM classifier produces best prediction results in terms of both accuracy and F_1 score. A confusion matrix detailing prediction outcomes of LGBM is given in Fig. 1.

3 Counterfactual generation and analysis

For counterfactual generation, we leverage Diverse Counterfactual Explanations (DiCE) for Machine Learning Classifiers [4]. A primary goal of DiCE is to provide explanations for the predictions of ML-based systems that inform decision-making in critically important areas such as finance, healthcare, and education. DiCE approaches the discovery of explanations essentially as an optimization problem, similar to how adversarial examples are identified. As we generate explanations, we require perturbations that change the output of a machine learning model while at the same time being varied and practical to implement. Hence,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

the DiCE approach facilitates the generation of counterfactual explanations that can be customized for diversity and resemblance to the original input.

Moreover, DiCE allows for the imposition of basic feature constraints to ensure that the counterfactual instances are plausible. In our experiments, the demographic features such as age, gender, and WAM cannot be varied, since that is not practical. Similarly, we set the number of generated counterfactuals to 3, but determining the optimal number is still a research question.

A snapshot of generated diverse counterfactual set (CFs) for an individual failing student is illustrated in Table 2 (see Appendix A). Note that the generated counterfactual explanations are based on our top-performing black-box model LGBM. The first explanation proposes that decreasing the duration of communication with friends and family, as well as minimizing the arrangement time for face-to-face meetings, while increasing the amount of time devoted to project management, may result in a shift from a failing to a passing grade. The second counterfactual explanation recommends allocating more time to studying and education, while reducing the time spent on seeking assistance from classmates through Facebook or scheduling face-to-face meetings. The third counterfactual explanation suggests a substantial decrease in the time spent filling up dead periods and vacant slots, as well as minimizing collaboration time for assignments via Facebook. Although these explanations can be highly beneficial, the feasibility of implementing some of the recommended strategies may be uncertain. As a result, we carried out a user study to assess the effectiveness of the proposed counterfactual explanations, as well as practitioners' perspectives on the utilization of such explanations in their decision-making processes.

4 User study

Counterfactual explanations can be validated through user studies [6]. In particular, the goal of our user study is to understand the perception of educators towards employing counterfactual explanation in analyzing the student success prediction outcomes. For this purpose, we recruited 18 educators from a tertiary education institution. We provided them with 3 counterfactual explanations corresponding to each individual failing student. Then we asked -

How much do you believe the recommendations provided by DiCE are effective in changing the student success prediction outcome from fail to pass?.

The participants' feedback was collected using a 5-point Likert scale (ranging from 1 for Strongly Disagree to 5 for Strongly Agree) as summarized in Figure 2 (a). While a majority of the participants agreed with the validity of the generated explanations, some provided negative ratings. Additionally, an optional open text field was provided to gather more detailed responses. It was found that some respondents did not believe certain features (such as communication with friends and family) should be included in the explanation.

We also inquired the participants - *How many counterfactual explanations do you believe should be suggested so that you can implement them effectively and efficiently?*.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

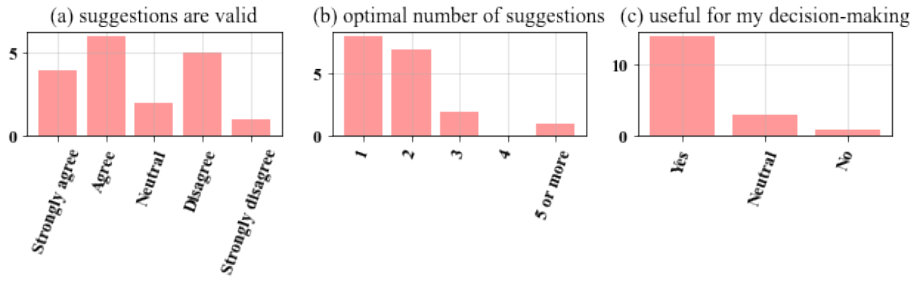

Fig. 2. Summary of user responses. Note: Y-axis denotes frequency of responses

As depicted in Fig. 2 (b), the majority of the participants favored 1-2 suggestions, with only one participant preferring 5 or more. However, this respondent did not indicate any specific reason for the selection.

Furthermore, we asked the participants whether they would be willing to incorporate counterfactual explanations into their decision-making process. As depicted in Figure 2(c), only one participant from our cohort did not find counterfactual explanations to be very useful.

Three participants from our cohort also provided additional insights on what need to be taken into account for the effective utilization of counterfactual explanations in predicting student success. The responses are given below.

"The utilization of counterfactual explanations can offer multiple benefits. Nonetheless, to determine the most effective interventions or supports for enhancing student success, an expert-in-the-loop approach is necessary, where experienced educators assist in finalizing the list of attributes to be taken into account for a particular student."

"Finalizing interventions for students who are going to receive them is of utmost importance. It is essential to comprehend and address any associated risks, such as the possibility of negative outcomes for a student receiving an intervention."

"It is important to recognize that there may be additional factors beyond the dataset that were not taken into account."

5 Conclusion

This paper uses social media usage behavior of university students to predict whether they will succeed or fail their courses. Using Diverse Counterfactual Explanations (DiCE), this study examines how a student who is failing could pass if certain factors were altered. Additionally, the paper includes an 18-educator user-study to evaluate the practicality of DiCE in predicting student success, and to gather educators' opinions on using counterfactual explanations in their decision-making process.

The educators who participated in the study found counterfactual analysis to be a powerful tool for advancing the prediction of student success. They

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

agreed that by simulating different scenarios and considering the impact of various factors, more effective interventions and support systems could be developed to improve a student's chances of success. The educators also emphasized the importance of including appropriate factors in the collected dataset to achieve more reliable prediction outcomes and explanations. One possibility for future research is to conduct a real-world trial of counterfactual explanations. This could involve creating a scenario in which at-risk students receive a specific intervention, and then comparing their outcomes to similar students who did not receive the intervention. By examining the impact of the intervention on student outcomes, educators could gain insights that could inform future interventions and support. Future research could also explore other datasets, with other explanations, which could help by providing educators with a concrete tool for decision-making.

Acknowledgements

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

References

1. Afrin, F., Hamilton, M., Thevathyan, C.: On the explanation of ai-based student success prediction. In: Computational Science–ICCS 2022: 22nd Int. Conference, London, UK, June 21–23, 2022, Proceedings, Part II. pp. 252–258. Springer (2022)
2. 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)
3. 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)
4. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
5. 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)
6. Spreitzer, N., Haned, H., van der Linden, I.: Evaluating the practicality of counterfactual explanations. In: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022 (2022)
7. 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)
8. 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)
9. 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)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44

A Appendix

Table 2. A snapshot of a set of generated counterfactuals with new outcome 1 (pass) for a query instance with original outcome 0 (fail).

Features	**Query Ins.** Outcome:0 #0	**Diverse CFs** Outcome:1 #0	#1	#2
Communications with friends and family	1320	921		
Enjoyment and entertainment	1320			
Filling in dead or vacant time	1320			241
Keeping informed about events and news	1320			
Education and study	1320		1556	
Work related reasons	1320			
Arrange a meeting for a group project	1320			
Class or university work related contact with another student	1320			
Discuss university work	1320			
Ask a classmate for help in the class	990		366	
Help manage a group project	1320	2290		
Collaborate on an assignment in a way my instructor would like	990			497
Arrange a face-to-face study group	1320	613	597	400
Linkedin time	0			
Snapchat time	0			
Twitter time	0			
Age	19			
Gender	1			
WAM	58.67			
FE-Grade	0	1	1	1

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_44