Article

Supporting Students’ Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics

Gomathy Ramaswami *, Teo Susnjak and Anuradha Mathrani

School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand
* Correspondence: g.ramaswami@massey.ac.nz

Abstract: Learning Analytics (LA) refers to the use of students’ interaction data within educational environments for enhancing teaching and learning environments. To date, the major focus in LA has been on descriptive and predictive analytics. Nevertheless, prescriptive analytics is now seen as a future area of development. Prescriptive analytics is the next step towards increasing LA maturity, leading to proactive decision-making for improving students’ performance. This aims to provide data-driven suggestions to students who are at risk of non-completions or other sub-optimal outcomes. These suggestions are based on what-if modeling, which leverages machine learning to model what the minimal changes to the students’ behavioral and performance patterns would be required to realize a more desirable outcome. The results of the what-if modeling lead to precise suggestions that can be converted into evidence-based advice to students. All existing studies in the educational domain have, until now, predicted students’ performance and have not undertaken further steps that either explain the predictive decisions or explore the generation of prescriptive modeling. Our proposed method extends much of the work performed in this field to date. Firstly, we demonstrate the use of model explainability using anchors to provide reasons and reasoning behind predictive models to enable the transparency of predictive models. Secondly, we show how prescriptive analytics based on what-if counterfactuals can be used to automate student feedback through prescriptive analytics.

Keywords: machine learning; anchors; counterfactuals; explainable machine learning

1. Introduction

Higher education institutions nowadays are driven by a complex collection of information and educational technologies. This has brought about a new phenomenon: the datafication of learning. Each data point, once aggregated and analyzed, may hold the potential to discover impactful new insights into students’ academic profiles and learning behaviors [1]. Learning Analytics (LA) involves the use of a wide selection of institutional data and analytic techniques, ranging from mere descriptive, to predictive, and more recently, prescriptive. Various stakeholders, such as administrators, teaching staff, and students, are now increasingly enabled to act and respond to outputs from data-driven analytics.

Descriptive analytics is the simplest form of data analysis and involves the process of using current and historical data to identify past behaviors and trends from students’ data. On the other hand, predictive analytics leverages Machine Learning (ML) algorithms to analyze students’ data and generate models that can make forecasts about the likelihood of some phenomena occurring, or inform stakeholders about consequential learning behaviors in the context of LA. While immensely useful, the shortcoming of many of these algorithms is that they produce black-box models which do not offer
transparency or insight into the mechanics of these predictions, nor do they provide human-centric explanations of their outputs [2]. This downside often results in distrust in such technologies and a considerable body of work is currently devoted to developing tools that infuse predictive models with explainability of outputs [3]. Meanwhile, prescriptive analytics adds a further layer of sophistication and is arguably the only component of the analytics suite capable of offering actionable insights by modeling what-if scenarios, or counterfactuals. In the context of LA, it is the data-driven evidence generated from counterfactuals that provide automated recommendations to students (and to student advisers) about which adjustments to learning behaviors are most likely to result in improved learning outcomes.

Recent research studies have indicated that effective feedback with specific recommendations on the possible courses of action can lead to self-regulated learning behaviors [4–7]. However, current research in predictive LA has largely been devoid of transparency which can convey to the users what data inputs drive the models towards their conclusions and, crucially, how a model has reasoned in producing the output for a given user. Furthermore, existing research is sparse on automated prescriptive feedback that draws upon data-driven methods to provide advisable behavioral adjustments to students, which may result in more positive outcomes. Meeting this gap has been recognized as an evidence-based pathway for triggering student reflections for maximizing their learning outcomes and ensuring course completions [6].

Taking these two gaps into account, the motivation of this paper is to demonstrate how the potential of predictive analytics can be maximized by communicating the reasoning of how the model arrived at its conclusions. In doing so, we show how prescriptive analytics can be integrated to generate data-driven advice that students can relate to; thus, enabling effectual learning changes that result in increasing the probability of realizing successful outcomes. This integration of different analytics paradigms is the first of its kind to be embedded into a Learning Analytics Dashboard (LAD) currently being piloted at a higher education institution. This study’s main contribution is the utilization of approaches that encompass a more complete spectrum of analytics technologies for the LA domain, together with its implementation and demonstration.

This study poses the following research questions:

- How can explainable ML approaches be leveraged to provide effective and human-understandable reasoning behind their conclusions about a student’s academic performance?
- How can automatic data-driven feedback and actionable recommendations be derived?

The remainder of this paper consists of five sections. Related work provides an overview of the existing approaches, followed by an analytics workflow that outlines the steps used for processing data in this study’s context. In methods, the techniques used for building predictive and prescriptive models are presented and subsequently described in the results section. The discussion provides perspectives on the findings and directly addresses the research questions, followed by a brief summary of the paper’s contributions in the conclusion section.

2. Related Work

This section presents recent work undertaken in Educational Data Mining (EDM) research streams for the purposes of predicting student academic performance and for the identification of factors that affect their performance. An overview of emerging ML approaches with their limitations and drawbacks is elaborated upon next.

2.1. eXplainable Artificial Intelligence (XAI)

XAI research in the context of ML aims to look inside black-box predictive models for extracting information or explanations as to why a model has come to a given
conclusion. In addition to providing tools that can build more trust and accountability, XAI assists with debugging and identifying bias in ML.

Predicting students’ learning behaviors and their final outcomes is regarded as one of the most important tasks in the EDM field. Prior studies have proposed several data mining approaches to predict academic performance, which would then be followed up by notifying the instructor about who the at-risk students are, which in turn opens opportunities for the instructor to intervene and provide the student with learning pathways for improving their performance. It is common for students’ previous academic performance, demographics, and LMS usage-related features to be used for this predictive task; however, all previous studies’ focus has been merely on either conveying to students their at-risk status of non-completion in a course [8–11] or in delivering the probability of attaining specific grades [12].

None of these predictive LA applications adequately explain to users how the models work, or how their predictions were generated; rather, they simply provide the predicted outcomes as feedback to users. Although such feedback might be helpful to some extent, it does not provide meaningful personalized insights or actionable information about the reasons behind the predictions. This lack of interpretability and explainability of the models makes it difficult for end-users to trust the system with which they are interacting. Moreover, when a model produces unexpected or erroneous output, the lack of trust often results in increased skepticism and possibly even a rejection on the part of the end-user [6]. These errors may have negative side effects as well, such as in instances when certain actions or decisions affecting others are taken, which might be based on false premises arising from misclassifications [13]. Thus, it has become a sine qua non to investigate how the inference mechanism or the decisions of ML models can be made transparent to humans so that Artificial Intelligence (AI) can become more acceptable to users from different application domains [14]. This has led to the establishment of a relatively new research field, namely eXplainable AI. The XAI domain researches and develops tools and techniques that enable the interpretability of autonomous decision-making systems through outputs in the form of simplified textual and visual representations that can be understood by human users [8].

2.2. Post Hoc Explanations of Machine Learning Models

Extracting explanations of ML models after they are induced is crucial for users (or students in this context) to apprehend so that they can act on algorithmic predictions in an informed manner. Moreover, explanations of an ML model’s predictions have found many uses, including understanding which features are most important, debugging the models, and increasing trust through transparency [5]. In addition to furthering the societal acceptance of the recommendations that are based upon algorithmic decisions, they also provide alternate suggestions for overcoming unfavorable predictions.

Meanwhile, when it comes to prescriptive analytics, several studies have leveraged different forms of messaging and recommendation techniques to enhance their prescriptive capabilities (e.g., [8,9,14]). However, none of the existing studies in EDM employed prescriptive modeling algorithms for generating automated personalized recommendations to inform students of specific behavioral changes [6]. In a recent study [6], we implemented an interpretability component and operationalized a LA dashboard that provided post hoc explanations of the prescribed recommendations that were tailored for individual students. We believe this to be the first published state-of-the-art LA dashboard to have incorporated both automated prescriptive analytics as well as transparent predictive models in a student learning context.

3. Analytics Workflow

This section presents the conventional steps involved in the predictive analytics process flow, to which we add two further steps which introduce both model transparency and prescriptive analytics.
3.1. Standard Predictive Analytics Workflow

The general workflow for predictive analytics for generating ML models involves acquiring relevant datasets as the first step (refer to Figure 1). Next, the pre-processing of the raw (unclean) data is performed to prepare this data, resulting in reliable datasets to enable subsequent analyses. In this step, usually, the total number of features (or variables) is reduced and only the most relevant or complete features are retained. In the data transformation step, additional features are also engineered as derivates of the raw data to produce richer descriptors that are more correlated with the target (dependent) variable.

Predictive models are subsequently induced based on the selected features in conjunction with a suite of ML algorithm implementations. The resulting predictive models are then evaluated using various evaluation measures for their generalizability. In the case of this study, a range of metrics was used, and finally, the best predictive models were selected based on these metrics.

Figure 1. Analytics workflow followed in this study.

3.2. Model Explainability

Once the predictive model has been generated, we propose the introduction of a new step in the workflow, which generates an explanation of how a given model has arrived at a particular prediction for a specific student. When using black-box models, there are technologies available that generate proxy models which emulate the behavior of the underlying models, since the purpose is to simplify and present the overall model behavior in human-understandable terms [15]. Technologies such as Anchors [16], SHAP [17], and LIME [18] can typically be leveraged to generate these simplified explanations.

Anchors are model-agnostic, i.e., they can be applied to any prediction model without requiring knowledge about the internals of the original model. Anchors uses a
perturbation-based strategy to generate easy-to-understand explanations for predictions by emulating the behavior of the original, which are then expressed as IF-THEN rules. In order for them to be meaningful, the Anchors need to have a high precision and coverage score. Precision is the proportion of data points in the region defined by the anchor that have the same class as the data point being explained, while coverage describes how many data points an anchor’s decision rule applies to [19].

3.3. Prescriptive Analytics

The final proposed step involves generating data-driven prescriptive feedback. In this step, the desired target outcome is selected for a given student (i.e., a successful outcome for a student whose prediction is otherwise), and an ML tool is applied which computes the smallest set of changes that would need to take place amongst the selected features to achieve the desired predicted output [20]. Counterfactuals are chosen for generating the automated prescriptive output in this study.

4. Methods

4.1. Dataset and Features

The datasets used in this study were extracted from courses offered at an Australian higher education institution involving a blended learning environment. These comprised data describing learning activities from Moodle (i.e., the learning management system (LMS)), various assessment grades from the Student Management System (SMS), and the demographic and pre-academic data from the Enrolment Management System (EMS). The raw data from the LMS log files were processed to engineer features used for building various analytics components within the predictive and prescriptive models.

During this process, individual student behavioral attributes describing learning patterns were extracted from the raw data and converted into a format that relativized each student’s value with respect to the average patterns of the student’s cohort. Z-score standardization was used to achieve this. This transformation technique had the effect of producing features that were generic and thus applicable across different courses and cohorts. The final features used for model development are shown in Table 1. These consist of assignment score, assignment deviation score, LMS engagement score, LMS deviation score, prior course scores, resources viewed, citizenship, and demographic information such as age and the gender of the students.

For the model training, we defined two categories of students, namely low-risk and high-risk. The threshold used for assigning these labels to the students depended on the students’ final course mark. Students with a course mark of 60% or less were labeled as high-risk, while the remainder were considered low risk. These categories alluded to the risk level of failing a course.

Table 1. Feature descriptions.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment score</td>
<td>Assignment score received by a student</td>
</tr>
<tr>
<td>Assignment deviation</td>
<td>The Z-score of the student’s mean assignment score as a deviation from the cohort mean</td>
</tr>
<tr>
<td>score</td>
<td></td>
</tr>
<tr>
<td>LMS engagement score</td>
<td>The count of all activities performed by a student on the Moodle platform.</td>
</tr>
<tr>
<td>LMS deviation score</td>
<td>The engagement score expressed as a Z-score of a student as a deviation from the cohort mean</td>
</tr>
<tr>
<td>Prior course scores</td>
<td>The mean score achieved by a student from across all previous course scores</td>
</tr>
<tr>
<td>Resources viewed</td>
<td>Resources viewed by a student on the Moodle</td>
</tr>
<tr>
<td>Citizenship</td>
<td>The nationality of the student</td>
</tr>
</tbody>
</table>
4.2. Tools

We used a mixture of Python’s scikit-learn library [21] and the separate CatBoost [22] library implementation for classification algorithms covering the predictive analytics component. For the explainability component of our analytics workflow, we used Anchors proxy models, and specifically the Python [21] Anchors library implementation. An anchor explanation is simply a translation of the black box into a set of rule-based outputs based on if-else conditions and is thus intuitive, and easy to comprehend. In order to describe high-level model mechanics, we used the Shapely Additive Explanations (SHAP). SHAP calculates the local feature importance for every observation. The SHAP method constructs an additive interpretation model based on the Shapley value. The Shapley value measures the marginal contribution of each feature to the entire cooperation.

For generating prescriptive analytics outputs leveraging counterfactual explanations, we used Python’s DiCE (Diverse Counterfactual Explanation) [20] implementation in our experiments. DiCE computes a set of diverse counterfactual explanations by finding candidate feature vectors that are close to the query instance but with an opposite prediction [23].

4.3. Machine Learning Algorithms and Evaluation

We employed five ML algorithms for the prediction of learning outcomes (low-risk/high-risk), namely Logistic Regressions (LR) [24], k-Nearest Neighbors (KNN) [25], Random Forest (RF) [26], Naïve Bayes (NB) [27], and CatBoost. The purpose was to compare and evaluate these classifiers thoroughly to determine which one consistently performed better on this dataset. To ensure that the final model (that has been described in detail in our published work [28]) generalizes well, the predictive model’s performance was evaluated using accuracy, precision, recall, F-measure, and area under the curve (AUC).

A modified k-fold cross-validation approach was used to evaluate the models. Our dataset consisted of seven courses and a total of 10 separate deliveries of those courses. The models were trained on nine course deliveries and tested against the remaining hold-out course offering. The process was repeated 10 times with a different combination of training and hold-out courses in order to arrive at our final, aggregated evaluation scores for our models. We can observe from the results (Table 2) that the overall accuracy of all the algorithms has ranged between 65% and 75%, with the result from CatBoost clearly outperforming the other algorithms on these datasets. The accuracies of the bottom three algorithms, namely Naïve Bayes, Random Forest, and Logistic Regression, did not exhibit significantly divergent accuracy results.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>F-Measure</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CatBoost</td>
<td>0.77 ± 0.024</td>
<td>75 ± 2.1</td>
<td>0.87 ± 0.023</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.67 ± 0.025</td>
<td>67 ± 2.4</td>
<td>0.74 ± 0.015</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.67 ± 0.023</td>
<td>68 ± 2.3</td>
<td>0.71 ± 0.034</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.68 ± 0.031</td>
<td>67 ± 3</td>
<td>0.73 ± 0.025</td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>0.71 ± 0.02</td>
<td>71 ± 2.4</td>
<td>0.72 ± 0.022</td>
</tr>
</tbody>
</table>

Figure 2 depicts the F-measure for various classifiers on the hold-out dataset, depicting the scores at different snapshots in time, where the snapshots were defined by the
week number in a given semester at which point the predictions were made. The snapshots were 2, 4, 6, and 8-week time points in a semester. The figure shows that, on average, the predictive accuracy of the models improved as the semester progressed, and more data about the students was gathered. The final accuracy from all the test datasets was displayed to the students as a measure of confidence in the reliability of the underlying model.

Figure 2. Predictive performances of various classifiers.

4.4. Feature Importance

Feature importance analysis is an important concept of machine learning as it helps to estimate which features are making the most impact on a model’s decision-making. Figure 3 shows the feature importance of our model, listing the most important features at the top, with the size of the bars indicating the magnitude of impact. The figure indicates that the top three features in the model’s reasoning all heavily relied on students’ academic performance in a prior course.
In order to extract some additional insights into the mechanics of the models, we use feature dependence plots in Figure 4. A dependence scatter plot explores the interaction effects between different features on the predicted outcome of a model. The $x$-axis denotes the value of a target feature, and the $y$-axis is the corresponding SHAP value for that feature, which relates directly to the effect it has on the final prediction.

Figure 4a indicates that there is a positive correlation between LMS engagement scores and assignment scores. Two noteworthy patterns emerge from Figure 4a. First, we can see that students who score highly on the assignment scores but exhibit poor engagement levels with the LMS receive strongly negative outcome predictions by the model. We also see there is a threshold of 0.4 for the LMS engagement score, and those that score
above this threshold are positively correlated with the model’s predictions for successful outcomes, which is amplified further for those with higher assignment scores.

Figure 4b denotes feature interaction between the LMS engagement score and the age of the student. It can be seen from the plot as the age of a student increases, the effect on the model predictions for positive outcomes becomes stronger. From approximately age 26 onwards, increases in the student age carry a stronger positive effect on model predictions for positive outcomes until age 40, from which point there do not appear to be any further increasing positive effects. The students who are most at-risk are those in their early twenties, with LMS engagement scores having no clear positive effect on prediction outcomes for this student demographic.

5. Result

This section presents the results of transforming a black-box model prediction into explainable outputs, and it also shows how prescriptive feedback can be generated from what-if analyses. The results are a demonstration taken from two hypothetical students (Students A and B) who were identified as high-risk by the predictive model that was trained as outlined in Figure 1 and described in the methodology section.

5.1. Model Explainability Example using Anchors

Figure 5 shows how the model has reasoned in concluding that Student A and Student B as being at high risk of course non-completion. The output is generated by the Anchors method, which creates an intermediate (proxy) model that emulates the behavior of the actual underlying black-box model and converts it into a human-understandable sequence of predicates. The rule-based classifier can be understood as categorizing a student as high-risk if every predicate evaluates as true. The conversion process from a black box into an intermediate proxy model not only provides insights into exact feature value thresholds that contribute to the eventual predicted outcome but also demonstrates which particular features are important in the classification process. In this example for Student A (Figure 5a), the features assignment scores, and the online engagement with the Moodle online learning environment have influenced the model to classify the student into the high-risk category. The simplified output of the proxy model is able to clearly communicate that both the low assignment scores and the student’s lower engagement score compared with the student’s cohort, as being the key drivers of the prediction while also showing the exact thresholds in those values.

![Prediction explainability rule](image)
However, in the second example for Student B in Figure 5b, we can see that the model explanation deviates from that of the previous example; in contrast, it draws upon two different features in order to explain the high-risk classification for this student. The explanation utility, in this case, uses the student’s prior course scores and the online resources viewed features to arrive at the student’s eventual classification. Herein lies the potential of prescriptive analytics that is able to leverage different combinations of business rules in order to provide tailored prescriptive feedback to the students. Students can thus be better advised on what remedial actions could be taken, which might result in different and more positive outcomes.

5.2. Prescriptive Analytics Example using Counterfactuals

The proxy models in the previous section are particularly useful in offering explanations of their reasoning to relevant stakeholders. Some prescriptive insights can be extracted from them; however, in this section, we demonstrate how with the aid of an additional technique, we can generate more precise and specific prescriptive feedback. To demonstrate this, we use the same hypothetical students as in the previous section as examples. The outputs in Table 3 are generated using counterfactuals that model what-if scenarios by calculating what minimal changes in the existing student’s feature values need to occur in order for an opposite outcome (low-risk in our context) to be predicted [29].
The tables show Student A and Student B values termed as a query instance (together with the original classification outcome) and the generated counterfactual set (desired outcome for the student, which is low risk in this context). The values column in the tables represents the actual and the modeled feature values with their corresponding feature names. Table 3a shows a single counterfactual set being generated for Student A, while Table 3b demonstrates how multiple counterfactual sets can be generated. All the generated counterfactual sets are based on the underlying black-box model and demonstrate minimal changes to the various feature values that would be required to flip the classification of those students from high to low-risk, which can then directly be used for prescriptive feedback. For Student A (Table 3a), the risk level from “high” to “low” would occur if there is an improvement in scores in the upcoming assignment and online engagement with Moodle; however, for the hypothetical Student B (Table 3b), the first pathway suggests that the change in prior course scores along with assignment scores can help the student to change their risk level; albeit, in this context, a change in prior course scores cannot really be actioned for the current course, though it can indicate to the student how they could be classified as lower risk for their subsequent course if they achieve a certain score in their present course. In the second alternative set, it is suggested that the changes in assignment score, resources viewed, and engagement with Moodle can help the corresponding student to fall into the low-risk level.

Once the above counterfactual explanations have been induced, they can then be translated into a human-readable format as seen below (Figure 6) in order to provide students with clear, precise, and actionable suggestions on what adjustments to their learning behavior and performance are most likely to assist them in realizing positive outcomes. The translation itself can also be automated with the assistance of natural language processing tools with the confidence that the prescriptive feedback given to students is data-driven and reliable, provided that feasible counterfactual sets are selected.
Figure 6. (a) Counterfactual explanation translation for Student A. (b) Counterfactual explanation translation for Student B.

6. Discussion

ML is at the core of many recent advances in science and technology. Whether humans are directly using ML classifiers as tools, or are deploying models within other products, a vital concern remains: if users do not trust a model or a prediction, they will not use it. This is particularly relevant in settings where users are expected to make decisions based on the outputs of machine learning models. Besides helping to debug ML models, explanations of black-box models improve the interpretability and trustworthiness of algorithmic decisions and enhance human decision-making [17]. We propose the use of anchors to identify the influencing factors in the predictive models to explain to students in an easily understandable way how the model works and how the prediction is generated [30].

Not only should the explanatory model provide customized individual suggestions to users, but they should also prioritize which changes in behavior and which learning strategies the student should adopt that will most likely translate into favorable results. This should take the guesswork out of the equation and help students make good decisions. Expanding the LA models to include prescriptive analytics that integrates counterfactuals and automated customized suggestions to students creates an approach that further supports students and helps them to achieve the best possible learning outcomes.

Our study highlighted the importance of moving beyond black-box approaches and the value of openly explaining to students how the prediction of their outcome or risk status was made, and furthermore providing customized and actionable suggestions
based on real data to help the students work towards a favorable outcome. If students first understand why they have been classified in a certain way and then provided with clear and precise steps they can take to change a potentially unfavorable outcome, they are more likely to make the necessary changes. This study, as far as we know, is the first in the education domain in the field of LA dashboards to combine these approaches.

7. Conclusions

The prediction of academic performance is considered one of the most popular tasks in the field of LA. However, we found from prior studies that a disproportionate amount of attention is given to merely creating predictive models, and an insufficient amount of focus is placed on addressing model interpretability and the explainability of the predictive outputs. The latter lowers the utility of this technology, and over time erodes the trust of users. Furthermore, the current focus of machine learning technologies mostly stops at the predictive analytics step and does not provide stakeholders with the translation to actionable insights that can be derived from more sophisticated approaches which yield prescriptive analytics.

Hence the current study not only aimed to demonstrate how model interpretability and explainability of the individual predictions to students can be embedded into LA systems, but crucially, we demonstrate how this can be taken one step further and converted into counterfactuals, from which prescriptive interventions can be provided to the students. We present a demonstration of how LA systems can be developed which encompass a more comprehensive machine learning approach towards providing support to students which integrates both predictive and prescriptive analytics while maximizing transparency and trust in the underlying data-driven technologies.

The developed system presented in this study has been integrated into a learning analytics dashboard, and a pilot study is currently being conducted with the students to analyze the effectiveness of the dashboard with student final outcomes.

Author Contributions: Conceptualization, G.R., T.S., and A.M.; Writing—original draft, G.R., T.S., and A.M.; Writing—review and editing, G.R., T.S., and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study has received approval from the Institutional Human Ethics Committee to conduct the research.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interests.

References


