Exploring Effects of Algorithm Aversion and False Causality Fallacy on How Students Act on Algorithmic Prediction of Pass/Fail Status

Abstract

Artificial Intelligence (AI), especially its subset machine learning, is attracting more and more educational practitioners' attention as algorithmic prediction has been shown to be powerful in other fields. Recently, in the field of education, it has been a hot topic that AI can be applied to address the problem of student dropout by accurately identifying academically at-risk students. This leads incorporating predictive modelling into student support services (e.g., early warning systems) to be high on the agenda for institutions. However, it is rarely discussed how end-users of school support services will interact with advice from predictive models. The present study argues that two human cognitive biases—algorithm aversion and false causality fallacy—may play a major role in affecting students' perceptions and interpretations of algorithmic outputs, which undermines the effectiveness of AI applications. Based on this perspective, the present study develops a model to predict whether a student will pass or fail a course (Study 1) and conducts an experiment to detect the biases on students' interaction with the model (Study 2) at Shanghai Open University. The experimental results show that students readily rely on algorithmic advice (i.e., no algorithm aversion detected) but are susceptible to the detected effect of false causality fallacy.

Practitioner Notes

What is already known about this topic

- The dropout rate of students from online courses is significantly higher than that of students from traditional education courses.
- Predictive modelling can identify students who are at risk academically to help decrease dropout rates.

What this paper adds

- A generalizable process framework was constructed and a model predicting student pass/fail status in a course was developed with an accuracy rate of 93.36% and a recall rate of 86.57%.
- 10 key factors correlated to student pass/fail status were identified by the model.
- Students' judgements on predicting their success in a course can be informed by the model, suggesting optimism about the potential for usage of algorithmic advice.
- Students may misinterpret algorithmic advice as causal information, indicating a potential risk of AI applications in the field.

Implications for practice and/or policy

- The generalizable process of predictive modelling in the present study can inform producing algorithmic advice on at-risk student identification at institutions.
- Given students' reliance on algorithmic judgment, institutions are encouraged to incorporate predictive modelling into schools' early alert systems to increase adherence to the systems.
- Due to the effect of false causality fallacy, predictive modelling results presented to students should juxtapose to a statement about no assurance of predictive variables' causal effects as an educational nudge.

Introduction

Online platforms have become a mainstream vehicle for education. In recent years, with the rise in popularity of Internet, more and more educational institutions go "campusless" (Thor & Scarafiotti, 2004), and the number of online learning courses offered by educational institutions is also growing at an alarming rate (Levy, 2007). Online education has great value, not only in reaching all kinds of students, but also in attracting them in ways that traditional classrooms cannot compare (Austin, 2009; Brower & Klay, 2000). Although online education is popular, the dropout rate of students in online courses (including MOOCs and SPOCs) has been shown significantly higher than that in traditional education (Lykourentzou et al., 2009), which draws the attention of researchers.

One approach to reducing students' online learning dropout rates is to accurately and timely identify academically at-risk students (who will fail a course) since course failure is correlated to the risk of dropping out (McKee & Caldarella, 2016; Kennelly & Monrad, 2007). Once the students are identified, they can be timely alerted to adjust themselves to the courses. Instructors can also better prepare for meeting the students' specific needs and taking appropriate actions to reduce their likelihood of failing and dropping out (Lykourentzou et al., 2009). Therefore, it is crucial to predict student performance and identify academically at-risk students early (before their dropping out). Since AI-produced advice through predictive modelling has been shown powerful in prediction, more and more educational practitioners are calling for the application of AI to accurately identify academically at-risk students as one of the core components of an early alert system (Arnold & Pistilli, 2012; Kim et al., 2014; Hone & Said, 2016; Lee et al., 2013). In academia, predicting students' performance using algorithms thus has been a hot topic (Arnold & Pistilli, 2012; Baker et al., 2015; Elbadrawy et al., 2015; Jokhan et al., 2018; Macfadyen et al., 2010; Marbouti et al., 2016; Zacharis & Nick, 2015).

Though many studies aimed to enhance the capability of algorithmic prediction, very little research investigated how people interact with the results of algorithmic prediction. Human cognitive biases on the perception and interpretation of AI-produced advice should be of concern since they can affect the effectiveness of investments in AI applications in the field. Some studies have shown that even knowing forecasts made by evidence-based algorithms are more accurate than forecasts made by humans (Silver, 2012; Shaffer et al., 2013), people would still have more faith and trust in human judgements (Dietvorst et al., 2015). This is a human bias termed "algorithm aversion" which may make investments in predictive modelling in the field not cost-effective.

Moreover, even if algorithm aversion is not an issue, there is the other potential bias affecting people's interpretation of and expectation in a predictive model: false causality fallacy. False causality fallacy refers to a human tendency to erroneously identify correlation patterns with causal information (Gambhire & Kshemkalyani, 2000). In a predictive model, the relationship between a dependent variable and predictive variables is a correlation. If false causality fallacy exists in end-users' interpretation of a predictive model—for example, students' interpretation of algorithmic advice on pass/fail status—they may improve predictive variables in the model but still fail at the end. The cognitive bias may make their efforts in the wrong direction (since there may be no guaranteed casual effect of predictive variables), undermining their trust in student support services (e.g., early alert systems).

The human cognitive biases could lead investments in predictive modelling and AI applications at educational institutions to be counterproductive. Therefore, to draw scholars' and practitioners' attention on this underdiscussed issue, the present study will develop a model to predict student pass/fail status in a course. And the following experiment will be designed with the model to investigate algorithm aversion and false causality fallacy on students' (i.e., end-users of support services in schools) interaction with the predictive model.

Literature review

Predictive Modelling in Identifying Academically At-risk Students

Many studies have explored the relationship between log file data of students on learning management systems (LMS) and their academic success (Zacharis & Nick, 2015). For example, Purdue University established Course Signal System in 2007, a learning prediction system based on an algorithm called Student Success Algorithm (SSA). Corresponding studies showed that courses using the course signal system had significantly improved the academic success rate and dropout rate of students (Arnold & Pistilli, 2012).

More and more studies used different machine learning techniques to predict student pass/fail status or academic performance that is related to student dropout. Baker et al. (2015) collected student performance data of formative assessment and interactive data on online learning platforms from 4002 online learners to develop a logistic regression model for predicting pass/fail status with an accuracy rate of 56.8%. Elbadrawy et al., (2015) used a multivariate regression model and a single regression model to predict students' performance in courses on Moodle and found the former with a Root Mean Square Error (RMSE) of 0.147 was better than the latter with a RMSE of 0.177. Macfadyen et al. (2010) conducted a regression analysis of students' learning behavior data from 118 online biology courses over three semesters at the University of British Columbia and built a regression model that can predict students who would fail the courses with an 81% accuracy rate. Moreover, some studies also showed usage of predictive modelling to identify important factors. For example, an early alert system at the University of the South Pacific informed by a regression model identified that students' average login times per week and the average completion rate of activities on LMS are important predictive variables for student performance (Jokhan et al., 2018). The model identified high-risk learners with an 60.8% accuracy rate.

Overall, it can be seen that the academic community has conducted in-depth research on using AI techniques such as machine learning algorithms to build predictive models to identify at-risk learners. Since recently reducing student dropout rate has been high on the agenda for educational institutions, especially in online education (Hone & Said, 2016; Lee et al., 2013), a generalizable process framework guiding practitioners to build a predictive model generating accurately algorithmic advice on at-risk student identification is needed.

Challenges to Predictive Modelling Applications: Algorithm Aversion and False Causality Fallacy

Although how accurate AI or machine learning prediction can be is important, how people perceive and interpret algorithmic advice is also consequential. Human cognitive biases may play a role affecting people's interaction with algorithmic judgement. Algorithm aversion refers to a cognitive bias that causes people often not to count on judgements made by algorithms. This widespread phenomenon has been shown in research. Studies by Diab et al. (2011) and Eastwood et al. (2012) showed that laypersons and professionals respectively averse to trusting algorithms, opting instead for the less accurate judgments of humans. Moreover, people are shown to hardly tolerate algorithms' imperfect. When informed that algorithmic predictions objectively outperform human judgements although not 100% accurate, people still prefer to trust the latter. Dietvorst et al. (2015) conducted experiments asking participants to choose between using algorithm-based prediction or manual prediction in motivational tasks. The result showed that when there was no information about the performance of the algorithm, most participants chose to use the algorithm. However, when knowing that the algorithm was still likely to make mistakes with a small probability (i.e., not performing perfectly with 100% accuracy), participants were more likely to choose to use human prediction.

The cost of algorithm aversion is increasingly high for educational institutions as there are more and more investments in the collection, analysis, and exploitation of ever larger quantities of educational data to produce algorithm advice for students, instructors, and administrators (i.e., end-users of AI applications in schools). If these end-users erroneously decline algorithmic advice, the investments in producing algorithmic prediction will be not cost-effective. Moreover, wrong decisions made by solely human predictions would keep costing institutions "as usual" (Grove et al., 2000).

In addition to algorithm aversion, the other bias that may affect people's interaction with algorithms is false causality fallacy. It depicts the phenomenon where people mix correlated events up with a causal relationship (Gambhire & Kshemkalyani, 2000), but correlation alone cannot be used as evidence of causal relationships. This effect may make algorithm prediction mislead people toward making wrong decisions that cost much more. A practical example is that previous epidemiological studies had found that women and children who received a hormone replacement therapy had a lower incidence of Crohn's disease (Lawlor et al., 2004). Therefore, there were lots of investments in hormone replacement therapies. However, the relationship between hormone replacement therapies and Crohn's disease is not a causal relation. The effect of false causality fallacy made the investments wasted (Greenland, 2005).

Unlike the field of health sciences, false causality fallacy used to not be a research topic in the field of education (Tropf & Mandemakers, 2017). However, given the growing use of predictive modelling at educational institutions, more and more members' decision-making will be informed by algorithmic advice (i.e., correlational information) such as informing administrators about potential dropouts (Aulck et al., 2016), helping instructors identify at-risk students at early stages of courses (Jayaprakash et al., 2014), or giving students unbiased understanding of their learning performance (Wilson, 2017). It is worth exploring whether there is the effect of false causality fallacy on such end-users' interpretation of AI-produced advice (Tropf & Mandemakers, 2017).

In summary, investments in predictive modelling at institutions may be much more costly if there is algorithm aversion increasing end-users' distrust of AI-produced results or false causality fallacy making the results misleading. Responding to very little research investigating the issue in the field of education, studies on exploring the effects are needed.

Methodology

The present research was divided into two studies. In Study 1, based on a proposed process framework, a classification model to predict whether a student will pass or fail a course would be developed. The overall goal of Study 1 is to inform educational practitioners about the generalizable process of predictive modelling which can be incorporated into early warning systems at institutions as well as to provide inputs to Study 2. In Study 2, an experiment would be conducted to explore whether there are the effects of algorithm aversion and false causality fallacy on students' interaction with the model developed in Study 1. This study is to identify the potential risk of cognitive biases in end-users on the effectiveness of predictive modelling applications in the field.

Study 1: Prediction of Student Pass/Fail Status Using Classification Models

This study is designed to build a classification model to predict student pass/fail status and identify 10 key factors correlated to students' academic success in a course. Model input data were mainly collected from 2006 learners' LMS log data of an online course "Introduction to Economic Law" in 2017 at Shanghai Open University. The course was a compulsory course for finance-major students and optional to students with other business majors. Data were categorized into three groups: students' basic information, online learning behavior, and academic achievements.

A generalizable process framework of developing a predictive model to identify at-risk students is constructed in Figure 1. As shown, the input data would firstly be processed by feature engineering to generate new sample features. All features would be verified and cleaned, and features with more missing data would be eliminated. Moreover, since unbalanced samples have a great impact on classification algorithms for pass/fail status prediction (e.g., support vector machine (SVM), logistic regression, AdaBoost and random forest algorithms used in the present study), samples would be balanced to make sure that the sample size of each of the two labels (i.e., pass/fail) is basically the same. Afterwards, the data set would be divided into training data sets and testing data sets for cross-validation to choose a predictive model. And important features for student pass/fail status prediction would be identified by the model. Finally, the model selected and 10 key features identified would be used in Study 2.

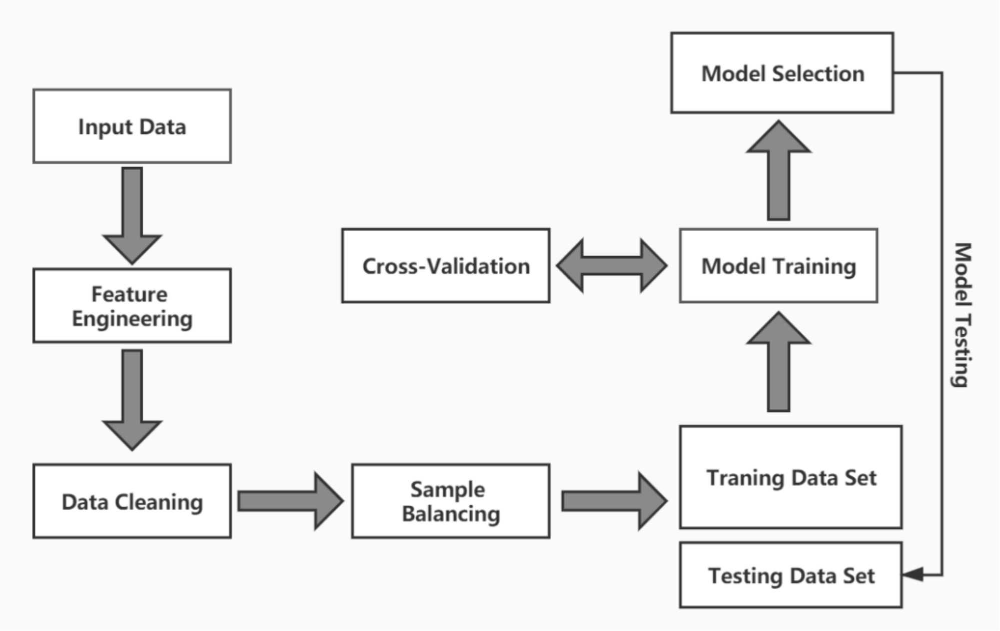


Figure 1. Process Framework of Developing a Predictive Model to Identify At-risk Students

Study 2: Detection of Algorithm Aversion and False Causality Fallacy

Based on outcomes of Study 1 (i.e., a selected model and 10 key factors it identified), the following study would be designed and conducted to examine whether there are the effects of algorithm aversion and false causality fallacy on students' perception of a predictive model. Student participants were recruited through a university-wise email invitation at Shanghai Open University and 348 students consented to participate in the study.

In the study (see Figure 2), the participants were asked to rank the 10 key factors identified in Study 1 in three different scenarios (i.e., three tasks). In the first task, the participants were asked to simply rank the 10 factors in terms of the importance for "predicting" student pass/fail status in a course by their own judgements (as baseline prediction). The order of the features listed for participants to rank was randomized in tasks. In the second task, the rank result by the random forest model built in Study 1 was presented to the participants for ranking the 10 key factors again. In the last task, though the same model-predicted result was presented, the participants would be asked to rank the 10 factors in terms of the importance for "intervening" in order to success in a course.

Afterwards, three point-based rank distributions from Task 1, Task 2, and Task 3 would be developed. The factor ranked first by a participant would be weighted with 10 points, the one ranked second would be weighted with 9 points, and so on. Gathering the point results of the 10 factors from each individual in a task would form a rank distribution. The rank distributions would be compared to examine the effects of algorithm aversion and false causality fallacy, and two hypotheses are proposed below:

$H_0$1: There is no significant difference between rank distributions in Task 1 and Task 2.

(This indicates that the rank distribution in Task 2 will not be informed by the algorithmic result presented, which means it will remain the same as the baseline prediction. Therefore, the statistical null hypothesis above is equivalent to the research hypothesis that the effect of algorithm aversion exists.)

$H_0$2: There is no significant difference between rank distributions in Task 2 and Task 3.

(The indicates that the rank distribution in Task 3 will be the same as that in Task 2, which means participants think predictive factors and causal factors are the same. Therefore, the statistical null hypothesis above is equivalent to the research hypothesis that the effect of false causality fallacy exists.)

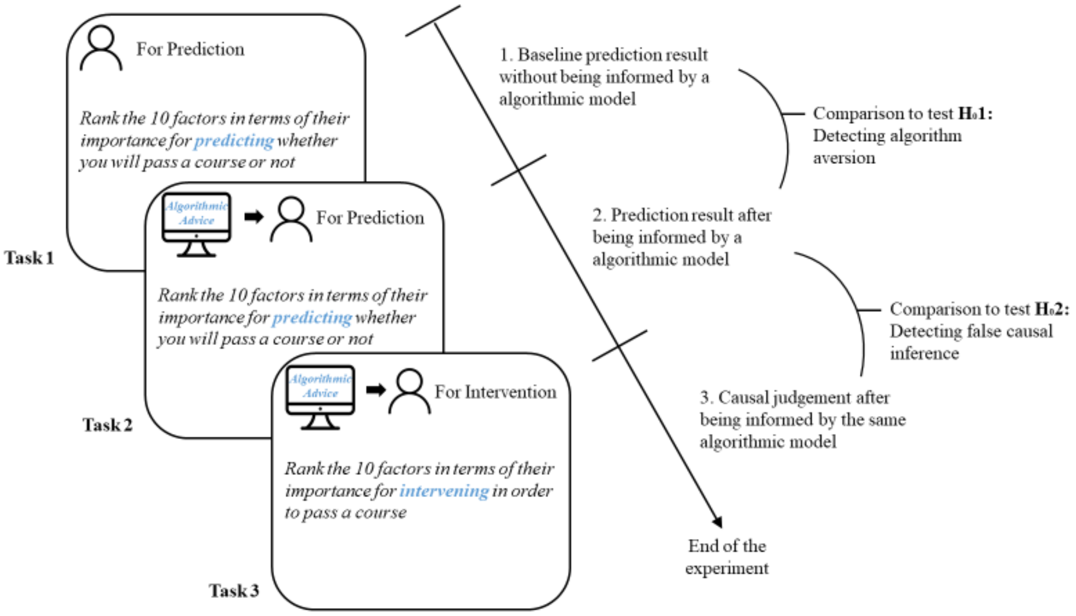

For Prediction
Rank the 10 factors in terms of their importance for predicting whether you will pass a course or not
Task 1
Algorithmic Advice
For Prediction
Rank the 10 factors in terms of their importance for predicting whether you will pass a course or not
Task 2
Algorithmic Advice
For Intervention
Rank the 10 factors in terms of their importance for intervening in order to pass a course
Task 3
1. Baseline prediction result without being informed by a algorithmic model
Comparison to test $H_0$1: Detecting algorithm aversion
2. Prediction result after being informed by a algorithmic model
Comparison to test $H_0$2: Detecting false causal inference
3. Causal judgement after being informed by the same algorithmic model
End of the experiment

Figure 2. Experiment Design and Process to Detect Effects of Algorithm Aversion and False Causality Fallacy

Data Analysis and Results

Study 1: Prediction of Student Pass/Fail Status Using Classification Models

There were 28 variables selected as input features (Table 1); meaningless variables about basic information of students such as user-ID were removed. Students were divided into two categories of pass and fail labels by the standard of 60 out of 100 points (i.e., grade greater than or equal to 60 is labelled with “pass,” vice versa). The number of students whose grades were unqualified was 205, accounting for 10.2% of the total number of students, which means the sample was unbalanced. Sample balancing to improve the performance of the model was needed. The whole data set was randomly divided into the training set and test set at the ratio of 8:2 (i.e., 1404 in the training set and 602 in the testing set). In the training set, the number of students who passed was 1266 and that of students who failed was 138. For sample balancing, SMOTE technology was used to resample the number of students who failed in the training set and increase it to 1266.

Table 1. 28 Selected Features

Feature Name	Description
Gender	Student gender

Performance of Study	Scores on daily learning performance
Group Performance	Grades for group projects
Phase Test Scores	Test average score at each stage of the course
Written Work Scores	Grades for essay assignments submitted
Scores of Assignments	Grades of submitted assignments
Bonus Points	Additional points earned by students
Time on Watching Live Streaming Courses	The total time spent participating in live courses
Formative Assessment Achievement	The average of all quiz scores
Duration for Completion of Assignments	The average time taken to complete assignments
Total Time on Course Browsing	The total time spent browsing course content
Times of Sign in	The times of signing in to the LMS/online platform
Course Practice Scores	Grades for course practice students participate in
Number of Assignments Completed	The number of assignments completed
Number of Participated Forum Activities	The number of participated activities on the forum
Number of Threads	The number of created threads on the forum
Number of Live Posts	The number of posts in live chatroom
Number of Participated Non Real Time Activities	The times of non-synchronous and non-real-time activities students participate in
Number of Course Practice	The times of course practice students participate in
Number of Posts	The number of posts on the forum
Average Test Score	The average of all test scores
Length of Time Spent on Course Video Watching	The total time spent watching instructional videos
Course Assignment Completion Rate	The ratio of the number of completed assignment to the total number of assignments
Activity Participation Rate	The ratio of the number of participated activities to that of all learning activities
Accuracy of Assignments	The average of accuracy rates of individual assignments
Number of Times Course Browsing	The times of browsing web pages on the platform/LMS
Learning Pace	Average time spent completing all tasks at a stage of the course
Activeness on Forum	Average time spent responding to a discussion on the forum

With the balanced training data, four classification algorithms (SVM, logistic regression, AdaBoost and random forest) were trained and evaluated through 3-fold cross-validation (given using grid search with cross-validation, smaller k is considered more time-efficient for computing). The results in terms of accuracy, recall, F1, AUC for the models are shown in Table 2. Though the random forest model was second to the AdaBoost model in terms of accuracy and F1 value, its recall rate (which is important for imbalanced classification cases) was the highest, reaching to 86.57%. Therefore, the random forest model was selected to test.

Table 2. Comparison of the Four Models

Algorithm	*Accuracy*	*Recall*	*F1*	*AUC*
SVM	0.8920	0.8208	0.6285	0.8609
Logistic Regression	0.8920	0.8358	0.6327	0.9386
AdaBoost	0.9451	0.8059	0.7659	0.9753
Random Forest	0.9336	0.8657	0.7435	0.9673

On the testing data (i.e. 602 students among which 535 students passed and 67 students failed), the selected random forest model was tested and the confusion matrix results are shown in Table 3. The ROC curves (using 3-folders) are shown with the AUC values in Figure 3. The x-axis represents False Positive Rate (FPR), and the y-axis represents True Positive Rate (TPR). The solid line is the mean ROC curve under which the area by integration can be represented by an AUC value, and the AUC of the selected random forest algorithm reaches to 0.97. Afterwards, an importance analysis was carried out in order to rank the input factors of the model in terms of their importance of predicting student success in a course. The result is shown in Table 4.

Table 3. Result of Forecast Confusion Matrix

	Pass (by prediction)	*Fail (by prediction)*
Actual pass	504	31
Actual fail	9	58

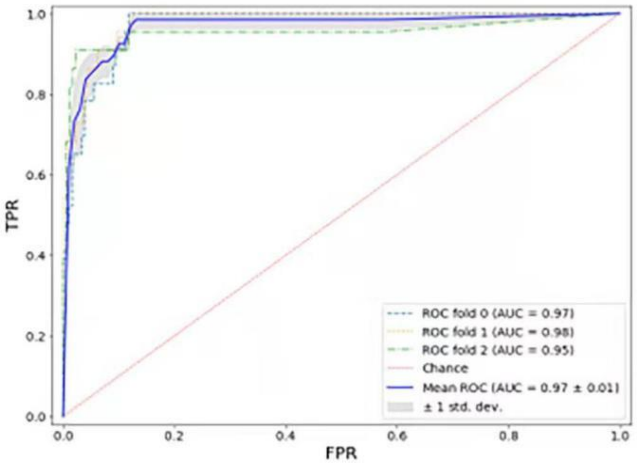


Figure 3. ROC Curves of the Random Forest Model

Table 4. Top 10 Important Factors for Prediction by the Random Forest Model

Feature Name	Description
Formative Assessment Achievement (FAA)	The average of all quiz scores
Number of Assignments Completed (NAC)	The number of assignments completed
Phase Test Scores (PTS)	Test average score at each stage of the course
Duration for Completion of Assignments (DA)	The average time taken to complete assignments
Performance of Study (PS)	Scores on daily learning performance
Scores of Assignments (SA)	Grades of submitted assignments
Number of Times Course Browsing (ntCB)	The times of browsing web pages on the platform/LMS
Total Time on Course Browsing (ttCB)	The total time spent browsing course content
Accuracy of Assignments (AA)	The average of accuracy rates of individual assignments
Length of Time Spent on Course Video Watching (LCV)	The total time spent watching instructional videos

Study 2: Detection of Algorithm Aversion and False Causality Fallacy

The rank results by the predictive model and students in Task 1, Task 2, and Task 3 are shown in Table 5. Before further statistical hypothesis tests on distribution comparison, in terms of algorithm aversion, the table shows that the rank result in Task 2 (i.e., prediction with the model presented) is more alike to the model predicted result than that in Task 1 (i.e., baseline prediction). This may indicate that participants' judgements were informed by the model. In terms of false causality fallacy, the rank result in Task

2 is the same as that in Task 3. This may point out that participants confused causality with correlation and thought interventions on predictive factors can have causal impacts.

Table 5. Rank Results of the Model, Task 1, Task 2, and Task 3

Factor **Rank**	*FAA*	*NAC*	*PTS*	*DA*	*PS*	*SA*	*ntCB*	*ttCB*	*AA*	*LCV*
Ranked by Model	1	2	3	4	5	6	7	8	9	10
Ranked in Task 1	1	5	4	8	3	2	7	9	6	10
Ranked in Task 2	1	3	2	6	5	4	7	9	8	10
Ranked in Task 3	1	3	2	6	5	4	7	9	8	10

The three point-based rank distributions from Task 1, Task 2, and Task 3 are shown in Figure 4. To compare them, a chi-squared test was conducted to investigate whether rank distributions of the tasks differed from each another. The chi-square results are shown in Table 6 and Table 7. Comparing the rank distribution of Task 1 with that of Task 2, the *p*-value of the chi-square test is less than 0.05, rejecting the null hypothesis for H_01 that the effect of algorithm aversion exists. Participants' prediction of academic risk is found to be informed by the model. Therefore, the algorithm aversion effect does not exist in student perception of algorithmic results. On the other hand, comparing the rank distribution of Task 2 with that of Task 3, the *p*-value of the chi-square test is .299 which is higher than 0.05. This fails to reject the null hypothesis for H_02 that the effect of false causality fallacy exists and shows false causality fallacy does lead students to confuse predictive information with causal information.

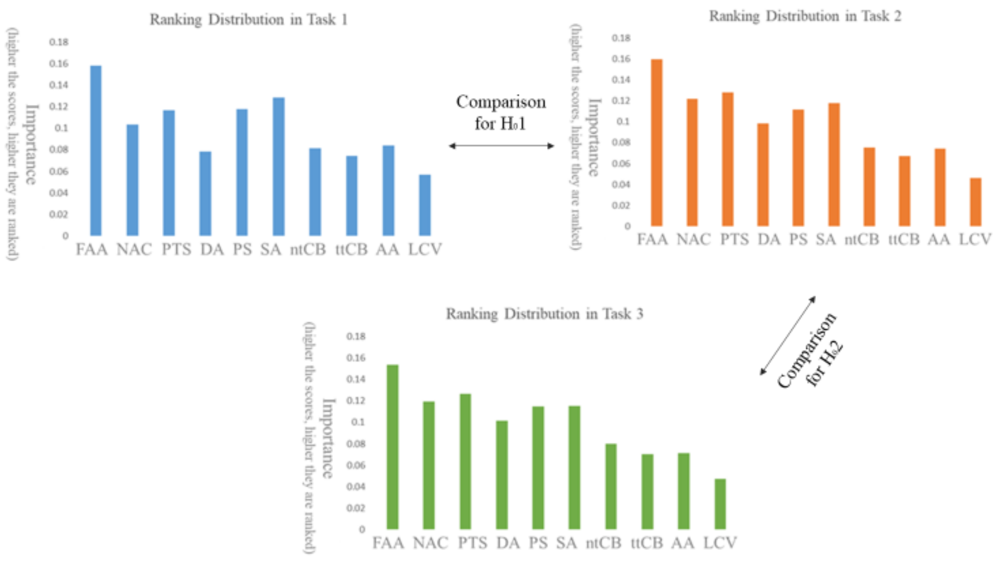


Figure 4. Rank Distributions of Task 1 (Blue), Task 2 (Orange), and Task 3 (Green)

Table 6. Chi-square Result for Testing the Null Hypothesis for H$_0$1

	Value	*df*	*Asymptotic Significance (2-sided)*
Pearson Chi-square	138.484	9	.000
Likelihood Ratio	138.666	9	.000
Linear-by-linear Association	72.836	1	.000
N of Valid Points	38280		

Table 7. Chi-square Result for Testing the Null Hypothesis for H$_0$2

	Value	*df*	*Asymptotic Significance (2-sided)*
Pearson Chi-square	10.673	9	.299
Likelihood Ratio	10.674	9	.299
Linear-by-linear	2.427	1	.119

Association	
N of Valid Points	38280

Discussion

In the present paper, it was found that when presented with the developed random forest model that was 93.36% accurate in identifying academically at-risk students, student participants did trust the model. Their predictions of pass/fail status was found to be informed by the model; the rank result in Task 2 is shown to lean to the model's prediction result (see Table 5 & Figure 5). This finding indicates that students are willing to change their own judgements (i.e., baseline prediction) and take the algorithmic advice to re-rank the 10 features for prediction. This result counters the widespread conclusion that people would avoid algorithms and prefer to go with their gut especially after knowing algorithms are imperfect (i.e., prediction accuracy rates are not 100%) even though they still give high accuracy, obviously beating human judgments (Dawes et al., 1989; Dietvorst et al., 2015; Kleinmuntz, 1990; Kleinmuntz & Schkade, 1993; Tetlock & Gardner, 2015). The result of the present study suggests that students readily rely on algorithmic advice about pass/fail status prediction in a course.

As for practical implications, as educational institutions have been investing in data collection and predictive modelling, the result shows optimism about usage of algorithmic advice for informing students about the risk of failing a course which can be incorporated into an early alert system. Besides, it should be kept in mind that although algorithm aversion was not detected in the present study, it may still exist if participants consider a model not accurate enough. Therefore, it is worth exploring the threshold of accuracy rates that will lead to algorithm aversion for future studies. The second point worth exploring is whether informing participants about precision or recall instead of accuracy of the model will influence the result. It is plausible that the former two may induce ambiguity aversion, a tendency to favor the familiar over the unfamiliar (Fox & Tversky, 1995), since they are not as intuitive as accuracy to students in general—which may further lead to algorithm aversion.

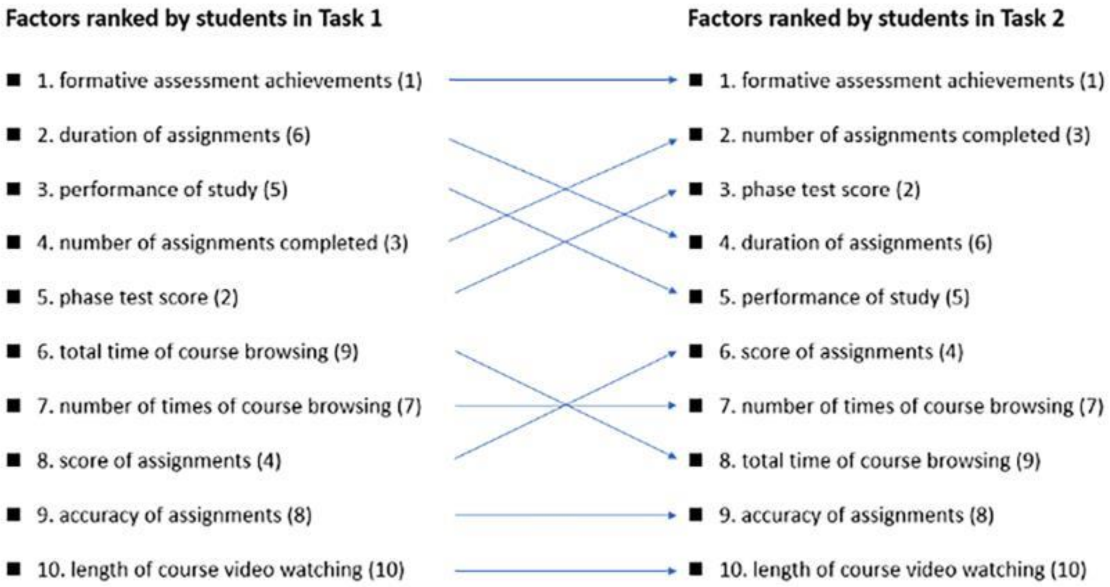
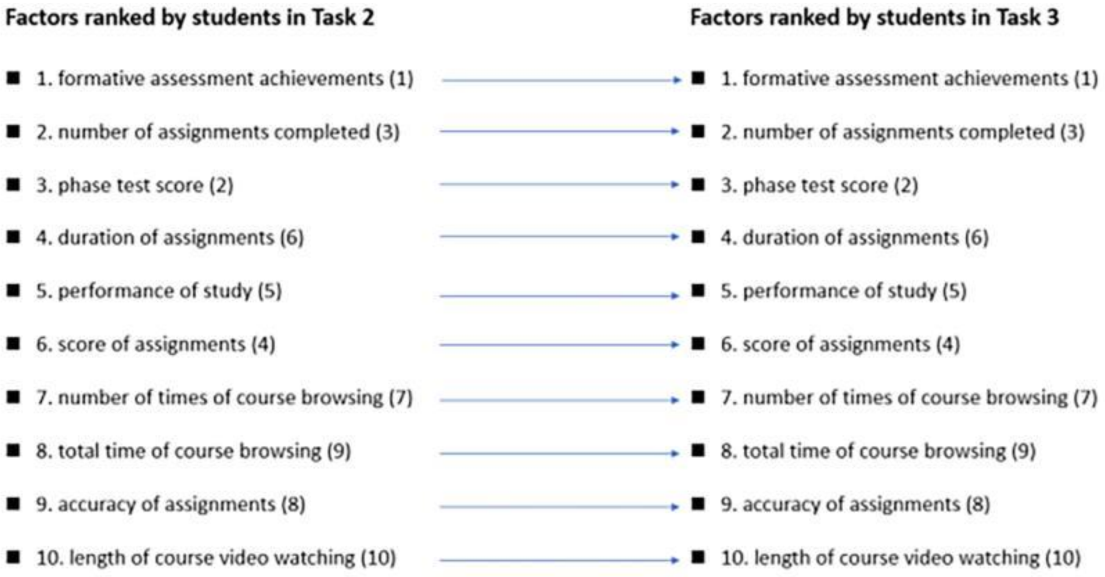

Factors ranked by students in Task 1
1. formative assessment achievements (1)
2. duration of assignments (6)
3. performance of study (5)
4. number of assignments completed (3)
5. phase test score (2)
6. total time of course browsing (9)
7. number of times of course browsing (7)
8. score of assignments (4)
9. accuracy of assignments (8)
10. length of course video watching (10)
Factors ranked by students in Task 2
1. formative assessment achievements (1)
2. number of assignments completed (3)
3. phase test score (2)
4. duration of assignments (6)
5. performance of study (5)
6. score of assignments (4)
7. number of times of course browsing (7)
8. total time of course browsing (9)
9. accuracy of assignments (8)
10. length of course video watching (10)

Figure 5. Rank Distribution Shifting from Task 1 to Task 2 Leans to the Model's Ranking Result. (The numbers in the parentheses indicate the ranking in the predictive model; the numbers without parentheses indicate the ranking in the corresponding tasks.)

Moreover, though algorithm aversion was not detected, it is still worth discussing whether there is a threat to the study's internal validity, erroneously leading to the optimistic result that contradicts the majority of studies about the existence of algorithm aversion. The researchers consider the expectancy bias plausibly playing a role in the experiment. In Task 1 and Task 2, participants were asked to conduct the same task (i.e., to rank the 10 factors) and the only difference between the two tasks is whether the model was presented, of which participants could be conscious. This may unintentionally become a hint for or mislead participants to subconsciously use the model as conforming to a rule. To eliminate any potential threat to the internal validity of similar experiments, future works could juxtapose a presented model with a human judgement, between which participants can choose for a forecasting task.

On the other hand, the effect of false causality fallacy was found in the present study. The rank result of Task 3 remained the same as that of Task 2 (see Table 5 & Figure 6) and the rank distributions showed no significant difference through a chi-square test. These findings indicate that students would misinterpret the predicted results (correlation information) as causal information, which may mislead students to make efforts on factors that have no causal effect on their success in a course. This points out the importance of clearly explaining the differences between causality and correlation to students when schools and universities incorporate learning risk prediction models into student support systems or services (e.g., early warning systems) to students. This also holds implications for delivering insights produced by predictive modelling to other school members (e.g., administrative staff). In addition, future studies are suggested to explore whether there is a causal relationship between the success in a course and any of the 10 factors identified in the present study. This can inform instructors about truly casual factors that can be intervened to help students succeed in a course.



Factors ranked by students in Task 2	Factors ranked by students in Task 3
1. formative assessment achievements (1)	1. formative assessment achievements (1)
2. number of assignments completed (3)	2. number of assignments completed (3)
3. phase test score (2)	3. phase test score (2)
4. duration of assignments (6)	4. duration of assignments (6)
5. performance of study (5)	5. performance of study (5)
6. score of assignments (4)	6. score of assignments (4)
7. number of times of course browsing (7)	7. number of times of course browsing (7)
8. total time of course browsing (9)	8. total time of course browsing (9)
9. accuracy of assignments (8)	9. accuracy of assignments (8)
10. length of course video watching (10)	10. length of course video watching (10)

Figure 6. Rank Distribution Remains the Same from Task 2 to Task 3. (The numbers in the parentheses indicate the ranking in the predictive model; the numbers without parentheses indicate the ranking in the corresponding tasks.)

In the present paper, algorithm aversion (undetected) and false causality fallacy (detected) were considered the two potential cognitive biases affecting students' interaction with algorithmic predicted results. However, they may also affect other members at institutions (e.g., faculty or administrative staff) and there may be other human cognitive biases on AI or machine learning applications. For example, algorithm aversion may lead to instructors' erroneous distrust of data outputs. It is worth investigating whether instructors would show a hostile unwillingness to take advice from a model when informed that the model is highly but not 100% accurate. The strong unwillingness to take data-driven advice could make institutional investments in data mining and predictive modelling not cost-effective. Moreover, confirmation bias may be a human cognitive bias on administrators' model preference. Confirmation bias refers to ones' overconfidence in predictions supporting her expectations (Nickerson, 1998). Thus, as administrators are shown to have a desire for a lower dropout rate, it is worth exploring whether administrators would tend to take advice from models predicting a lower dropout rate, even if having a lower accuracy at the same time. If that is the case, confirmation bias may make data mining and predictive modelling just a desired information search tool.

Human cognitive biases on the interpretation of algorithmic predicted results have been an underdiscussed issue. However, they can make AI or machine learning applications in education settings an ineffective or even counterproductive investment—which is worth investigation. The present study does not intend to prevent school leaders from investing in AI applications or to encourage eliminating human involvements in AI applications. Instead, it is to encourage leveraging the deeper understanding of humans with predictive modelling of ever larger quantities of educational data to help members at institutions (e.g., staff, faculty, and students). Its intention is to raise the alarm, drawing scholars' attention on human cognitive biases that may damage AI applications in educational contexts.

Last but not least, though the model developed in the present paper showed high accuracy and recall rates, the purpose of building the model is to demonstrate the generalizable "process" shown in Figure 1 to practitioners as well as launch Study 2. Therefore, the model itself and the top 10 features it identified were not expected to be generalizable across different settings since the sample size of the present study might be not large enough especially compared with other studies on MOOCs (De Barba et al., 2016; Gardner & Brooks, 2018; Kloft et al., 2014; Moreno-Marcos et al., 2018). Moreover, the model input data were from single online course which may lead to selection bias and the external validity problem—undermining the model's generalizability over diverse subjects. And there was a course prerequisite that all registered students needed to complete an introductory course "Introduction to Law and Practice," which may lead to survivorship bias and the external validity problem as well. For future works, if to build a predictive model with high accuracy rates and high generalizability across different settings is the goal, the distribution, sample size, representativeness of input data should be of concern.

Conclusion

Practitioners are calling for the application of AI in learning risk prediction to accurately identify academically at-risk students. The present study used questionnaire and log data on LMS (28 independent variables as input features) from 2006 students at Shanghai Open University for model building. A random forest model to predict student pass/fail status in a course with an accuracy rate of 93.36% and a recall rate of 86.57% was developed. By the model, 10 most important factors among the 28 input features for prediction were identified.

Moreover, given that very little research investigated how people interact with algorithmic prediction results, the following study was conducted to test student biased perception (i.e., algorithm aversion) and misinterpretation (i.e., false causality fallacy) of AI-produced advice. The results showed the present model with high accuracy (i.e., above 90%) yet imperfect (i.e., not 100% accurate) would not lead to algorithm aversion. This finding suggested that students readily relied on algorithmic advice and supported AI applications in education settings. However, false causality fallacy was detected, which indicated that it is indispensable to provide instructional information clarifying the correlation between the outcome (e.g., pass/fail status) and predictive variables when presenting algorithmic forecasting to students.

Acknowledgments

This paper is supported by China's National General Project sponsored by the China National Office for Education Sciences Planning (Grant No. BCA160053), The Construction and Application of Online Learners' Persona based on Big Data Analysis. We would like to thank Aizhen Fu and Zhimin Pan of East China Normal University and Junjie Ju of Shanghai Normal University for participating in literature review and data collection.

Statements on open data, ethics and conflict of interest

The collected data used in this study are available upon request to the first author of this paper. The data were collected according to the ethical standards of Shanghai Open University. All participants were informed about research purposes and consented to participate in this study. The data had been anonymised before processing. No potential conflict of interest was reported by the authors.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *In Proceedings of the 2nd international* *conference on learning analytics and knowledge. ACM International Conference Proceeding Series,* 267-270.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting Student Dropout in Higher Education. *In 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications,* 16-20.

Austin, E. K. (2009). Limits to technology-based distance education in MPA curricula. *Journal of Public Affairs Education, 15*(2), 161-176.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing early at-risk factors in higher education e-learning courses. *In Proceedings of the 8th International Conference on Educational Data Mining*, 150-155.

Brower, R. S., & Klay, W. E. (2000). Distance learning: Some fundamental questions for public affairs education. *Journal of Public Affairs Education, 6*(4), 215-231.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). *Clinical versus actuarial judgment. Science, 243*(4899), 1668-1674.

De Barba, P. G., Kennedy, G. E., & Ainley, M. D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*(3), 218-231.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment, 19*(2), 209-216.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155-1170.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114-126.

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making, 25*(5), 458-468.

Elbadrawy, A., Studham, R. S., & Karypis, G. (2014). Personalized Multi-Regression Models for Predicting Students' Performance in Course Activities. *UMN CS,* 11-14.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics, 110*(3), 585-603.

Gambhire, P., & Kshemkalyani, A. D. (2000). Reducing false causality in causal message ordering. *In International Conference on High-Performance Computing,* 61-72.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User*

Modeling and User-Adapted Interaction, 28(2), 127-203.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 168*(2), 267-306.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19 -30.

Hone, K. S., & Said, G. R. E. (2016). Exploring the factors affecting mooc retention: a survey study. *Computers & Education, 98,* 157-168.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6-47.

Kennelly, L., & Monrad, M. (2007). *Approaches to dropout prevention: Heeding early warning signs with appropriate interventions*. Washington, DC: National High School Center at the American Institutes for Research.

Kim, C. M., Park, S. W., & Cozart, J. (2014). Affective and motivational factors of learning in online mathematics courses. *British Journal of Educational Technology, 45*(1), 171-185.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*(3), 296-310.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*(4), 221-227.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. *In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs,* 60-65.

Lawlor, D. A., Smith, G. D., & Ebrahim, S. (2004). Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology?. *International Journal of Epidemiology, 33*(3), 464-467.

Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology, 44*(2), 328-337.

Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. *Computers & Education, 48*(2), 185-204.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*(3), 950-965.

Macfadyen, L. P., & Dawson, S. (2010). Mining lms data to develop an "early warning system" for educators: a proof of concept. *Computers & Education, 54*(2), 588-

599.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103,* 1-15.

McKee, M. T., & Caldarella, P. (2016). Middle school predictors of high school performance: A case study of dropout risk indicators. *Education, 136*(4), 515-529.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies, 12*(3), 384-401.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Shaffer, V. A., Probst, C. A. , Merkle, E. C. , Arkes, H. R. , & Medow, M. A. . (2013). Why do patients derogate physicians who use a computer-based diagnostic support system?. *Medical Decision Making, 33*(1), 108-118.

Silver, N. (2012). *The signal and the noise: why so many predictions fail--but some don't*. London: Penguin.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown Publishing Group.

Thor, L. M., & Scarafifiotti, C. (2004). Mainstreaming distance learning into the community college. *Journal of Asynchronous Learning Networks, 8*(1), 16-25.

Tropf, F. C., & Mandemakers, J. J. (2017). Is the association between education and fertility postponement causal? the role of family background factors. *Demography, 54*(1), 71-91.

Wilson, A., Watson, C., Thompson, T. L., Drew, V., & Doyle, S. (2017). Learning analytics: challenges and limitations. *Teaching in Higher Education, 22*(8), 991-1007.

Zacharis, & Nick, Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education, 27,* 44-53.