Chapter 3: Predictive Modelling in Teaching and Learning

Christopher Brooks,[1] Craig Thompson,[2]

[1] School of Information, University of Michigan, Ann Arbor, USA
[2] Centre for Teaching, Learning and Technology, University of British Columbia, Vancouver, Canada

DOI: 10.18608/hla22.003

ABSTRACT

This chapter describes the process, practice, and challenges of using predictive modelling in teaching and learning. In both the fields of Educational Data Mining (EDM) and Learning Analytics (LAK) predictive modelling has become a core practice of researchers, largely with a focus on predicting student success as operationalized by academic achievement. In this paper we aim to provide a general overview of considerations when performing and applying predictive modelling, the steps which an educational data scientist must consider when engaging in the process, and a brief overview of the most popular techniques in the field.

Keywords: Predictive modelling, machine learning, educational data mining (EDM), feature selection, model evaluation

Predictive analytics are a group of techniques used to make *inferences* about uncertain future events. In the educational domain, one may be interested in predicting a measurement of learning (e.g. student academic success, or skill acquisition), teaching (e.g. the impact of a given instructional style or specific instructor on an individual) or other proxy metrics of value for organizations (e.g. predictions of retention or course registration). Predictive analytics in education is a well established area of research, and several commercial products now incorporate predictive analytics in the learning content management system (e.g. D2L[1], Starfish Retention Solutions[2], Ellucian[3], and Blackboard[4]). Furthermore, specialized companies (e.g. Blue Canary[5], now a part of Blackboard learning, Civitas Learning[6]) now operate to provide predictive analytics consulting and products for higher education.

1 INTRODUCTION TO PREDICTIVE MODELLING

In this chapter, we aim to introduce the terms and workflow related to predictive modelling, with a particular emphasis on how these techniques are being applied in teaching and learning. While a full review of the literature is beyond the scope of this chapter, we encourage readers to consider the conference proceedings and journals associated with the Society for Learning Analytics and Research (SoLAR)[7], the International Educational Data Mining Society[8] (IEDMS), and the International Artificial Intelligence in Education Society[9] (IAIED) for more examples of applied educational predictive modelling.

It is useful to distinguish *predictive modelling* from *explanatory modelling*. In explanatory modelling, the goal is to use all available evidence to provide a explanation for a given outcome. For instance, observations of age, gender, and socioeconomic status of a learner population might be used in a regression model to explain how they contribute to a given student achievement result. The intent of these explanations is generally to test causal hypotheses (versus correlative alone, described well by [26]). In predictive modelling, the purpose of the activity is to create a model that will predict the values (or *class* if the prediction does not deal with numeric data) of new data based on observations. Unlike explanatory modelling, predictive modelling is based on the assumption that a set of known data (referred to as *training instances* in data mining literature) can be used to predict the value or class of new data based on observed variables (referred to as *features* in predictive modelling literature). Thus the principle difference between explanatory modelling and predictive modelling is with the application of the model to future events, where explanatory modelling does not aim to make any claims about the future, while predictive modelling does.

More casually, explanatory modelling and predictive modelling often have a number of pragmatic differences when applied to educational data. Explanatory modelling is

[1] http://www.d2l.com/
[2] http://www.starfishsolutions.com/
[3] http://www.ellucian.com/
[4] http://www.blackboard.com/
[5] http://bluecanarydata.com/
[6] http://www.civitaslearning.com/
[7] https://www.solaresearch.org/
[8] http://educationaldatamining.org/
[9] https://iaied.org/about/

a *post-hoc* and reflective activity aimed at testing an understanding of a phenomena. Predictive modelling is an *in situ* activity intended to make systems responsive to changes in the underlying data. It is possible to apply both forms of modelling to technology in higher education. For instance, [24] describe a student-success system built on explanatory models, while [9] describe an approach based upon predictive modelling. While both methods intend to inform the design of intervention systems, the former does so by building software based on theory developed during the review of explanatory models by experts, while the latter does so using data collected from historical log files (in this case, *clickstream* data).

The largest methodological difference between the two modelling approaches is in how they address the issue of evaluation. In explanatory modelling, all of the data collected from a sample (e.g. students enrolled in a given course) is used to describe a population more generally (e.g. all students who could or might enroll in a given course). The issues related to generalizability are largely based on sampling techniques. Ensuring the sample represents the general population by reducing selection bias, often through random or stratified sampling, and determining the amount of power needed to ensure an appropriate sample, through an analysis of population size and levels of error the investigator is willing to accept. In a predictive model a *hold out* dataset is used to evaluate the suitability of a model for prediction, and to protect against the *overfitting* of models to data being used for training. There are several different strategies for producing hold out datasets, including k-fold cross validation, leave-one-out cross validation, randomized subsampling, and application-specific strategies.

With these comparisons made, the remainder of this chapter will focus on how predictive modelling is being used in the domain of teaching and learning, and provide an overview of how researchers engage in predictive modelling process.

THE PREDICTIVE MODELLING WORKFLOW

Problem Identification

In the domain of teaching and learning, predictive modelling tends to sit within a larger action-oriented educational policy and technology context, where institutions use these models to react to student needs in real-time. The intent of the predictive modeling activity is to set up a scenario which would accurately describe the outcomes of a given student assuming no new intervention. For instance, one might use a predictive model to determine when a given individual is likely to complete their academic degree. Applying this model to individual students will provide insight into when they might complete their degrees *assuming no intervention strategy is employed*. Thus, while it is important for a predictive model to generate accurate scenarios, these models are not generally deployed without an intervention or remediation strategy in mind.

Strong candidate problems for a successful predictive modelling approach are those in which there are quantifiable characteristics of the subject being modeled, a clear outcome of interest, the ability to intervene in situ, and a large set of data. Most importantly, there must be a recurring need, such as a class being offered year after year, where the historical data collected about learners (the *training set*) is expected to capture patterns and relationships that will hold true of future learners (the *testing set*).

Conversely, there are several factors that make predictive modelling more difficult, or less appropriate. For example, both sparse and noisy data present challenges when trying to create accurate predictive models. Data sparsity, or missing data, can occur for a variety of reasons, such as students choosing not to provide optional information. Noisy data occurs when a measurement fails to accurately capture the intended data, such as determining a student's location from their IP address when some students are using virtual private networks (proxies used to circumvent region restrictions, a not uncommon practice in countries such as China). Finally, in some domains, inferences produced by predictive models may be at odds with ethical or equitable practice, such as using models of student at-risk predictions to limit the admissions of said students (exemplified in [27]). Lastly, domains where the types of data available change are not well suited to predictive modelling. For example, if a course undergoes significant redesign, shifting coursework from a single term-paper to weekly quizzes, it would be difficult to make predictions about end of term course grades based on term work, as the data about the training and testing populations are no longer directly comparable.

Data Collection

In predictive modelling, historical data is used to generate models of relationships between features. One of the first activities for a researcher is to identify the outcome variable (e.g. grade or achievement level) as well as the suspected correlates of this variable (e.g. gender, ethnicity, access to given resources). Given the situational nature of the modelling activity, it is important to choose only those correlates which can be known at or before the time in which an intervention might be employed. For instance, a midterm examination grade might be predictive of a final grade in the course, but if the intent is to intervene before the midterm, this data value should be left out of the modelling activity.

In time-based modelling activities, such as the prediction of a student final grade, it is common for multiple models to be created (e.g [8]), each corresponding to a different time period and set of observed variables. For instance, one might generate predictive models to be applied each week of the course, incorporating into each model the results of all weekly quizzes, student demographics, and the amount of engagement the students have had with different digital resources to date in the course.

While *state-based* data, such as data about demographics (e.g. gender, ethnicity), relationships (e.g. course enroll-

ments), psychological measures (e.g. grit [14] and aptitude tests) and performance (e.g. standardized test scores, grade point averages), are important for educational predictive models, it is the recent rise of big *event-driven* data collections that has been a particularly powerful enabler of predictive models (see [2] for a deeper discussion). Event-data is largely student activity-based, and is derived from the learning technologies that students interact with, such as learning content management systems, discussion forums, active learning technologies, and video-based instructional tools. This data is large and complex (often on the order of millions of database rows for a single course), and requires significant effort to convert into meaningful features for machine learning. At the same time, while we observe this growth of event-based data we caution that it is not universally more suitable for the generation of predictive models, and the quality and breadth of the data available may depend highly on other factors such as modality of education. For instance, in large online courses such as MOOCs, event-based data is rich because the learning activity is highly instrumented with data collection *and* there is a lack of socioeconomic state-based data describing learners. However, in many higher education residential courses the state-based data is rich (e.g. learner demographic and previous performance measures, such as standardized tests) and the learning technologies are often used shallowly (e.g. as file repositories for lecture material).

A second taxonomic dichotomy exists when considering whether the data was *self-reported* (e.g. a psychological survey) or *observed* (e.g. grades, click-stream log files, or eye tracking measurements). While in recent years predictive models in the field of learning analytics have emphasized the latter, the field of education and educational psychology has explored heavily the former, and instruments to measure psychological states including motivation, aptitude, disposition, and other forms of self-regulation are commonly used.

Of pragmatic consideration to the educational researcher is obtaining access to event data and creating the necessary features required for the predictive modelling process. The issue of access is highly context-specific, and depends on institutional policies and processes, as well as governmental restrictions (such as FERPA in the United States). One solution is to conduct research using previously established publicly available datasets, such as the Open University Learning Analytics Dataset[22], or the MITx and HarvardX Dataverse[17]. Alternatively, some institutions, such as the University of Michigan, have created standardized and streamlined access procedures for institutional data assets to enable their faculty members to conduct learning analytics research grounded in their unique institutional context.[10]

Classification and Regression

In statistical modelling there are generally four types of data considered: categorical, ordinal, interval, and ratio. Each type of data differs with respect to the kinds of relationships, and thus mathematical operations, which can be derived from individual elements. In practice, ordinal variables are often treated as categorical, and interval and ratio are considered as numeric. Categorical values may be binary (such predicting whether a student will pass or fail a course) or multivalued (such as predicting which of a given set of possible practice questions would be most appropriate for a student). Two distinct classes of algorithms exist for these applications; *classification algorithms* are used to predict categorical labels, while *regression algorithms* are used to predict numeric labels.

Feature Engineering

The raw event data available to researchers is rarely suitable for direct use in the fitting of a predictive model. Instead, it is often transformed through the process of feature engineering (a research field unto itself) into *candidate* features. As one example, timestamped resource access logs may be used to compute "time on task" sessions [21]. When using free-form text from essays or discussion posts, it is common to transform the raw data into more compact representations, including vectorized "bag of words" (e.g. through word2vec [25]), or other computational linguistic measures (e.g. [13]. Lastly, a range of network measures can be applied to quantify the social network characteristics of individual learners, such as their number of connected peers, their centrality in a larger network, or even embeddings within a larger network context (e.g. [15, 19]).

Feature Selection

In order to build and apply a predictive model, features which correlate with the value to predict need to be selected. When choosing what data to collect, the practitioner should err on the side of collecting more information rather than less, as it may be difficult or impossible to add additional data later, but removing information is typically much easier. Ideally, there would be some single feature that perfectly correlates with the chosen outcome prediction. However, this rarely occurs in practice. Some learning algorithms make use of all available attributes to make predictions, whether they are highly informative or not, whereas others apply some form of variable selection to eliminate the uninformative attributes from the model.

Depending on the algorithm used to build a predictive model, it can be beneficial to examine the correlation between features, and either remove highly correlated attributes (the *multicollinearity* problem in regression analyses), or apply a transformation to the features to eliminate the correlation. Applying a learning algorithm that naively assumes independence of the attributes can result in predictions with an over-emphasis on the repeated or correlated features. For instance, if one is trying to predict the grade of a student in a class and uses an attribute of both attendance in-class on a given day as well as whether a student asked a question on a given day, it is important for the researcher to acknowledge that the two features are not independent (e.g. a student could not ask a question

[10] See, for instance, `https://enrollment.umich.edu/data-research/learning-analytics-data-architecture-larc`

if they were not in attendance). In practice, the dependencies between features is often ignored, but it is important to note that some techniques used to clean and manipulate data may rely upon an assumption of independence.[11] By determining an informative subset of the features, one can reduce the computational complexity of the predictive model, reduce data storage and collection requirements, and aid in simplifying predictive models for explanation.

Missing values in a data set may be dealt with in several ways, and the approach used depends on whether the data is missing because it is unknown or because it is not applicable. The simplest approach is to either remove the attributes (columns) or instances (rows) that have missing values. There are drawbacks to both of these techniques. For example, in domains where the total amount of data is quite small, the impact of removing even a small portion of the data set can be significant, especially if the removal of some data exacerbates an existing class imbalance in the data set. Likewise, if all of the attributes have a small hand full of missing values, then attribute removal will remove all of the data, which would not be useful. Instead of deleting rows or columns with missing data, one can also infer the missing values from the other known data. One approach is to replace missing values with a 'default' value, such as the mean of the known values. A second approach is to fill in missing values in records by finding other similar records in the data set, and copying the missing values from their records.

The impact of missing data is heavily tied to the choice of learning algorithm. Some algorithms, such as the Naïve Bayes classifier can make predictions even when some attributes are unknown; the missing attributes are simply not used in making a prediction. The nearest neighbour classifier relies on computing the distance between two data points, and in some implementations the assumption is made that the distance between a known value and a missing value is largest possible distance for that attribute. Finally, when the C4.5 decision tree algorithm encounters a test on an instance with a missing value, the instance is divided into fractional parts which are propagated down the tree and are used for a weighted voting. In short: missing data is an important consideration which both regularly occurs and is handled differently depending upon the machine learning method and toolkit employed.

Methods for Building Predictive Models

After collecting a data set and performing attribute selection a predictive model can be built from historical data. In the most general terms, the purpose of a predictive model is to make a label prediction, given some related known information. This section will briefly introduce several such methods for building predictive models. A fundamental assumption of predictive modelling is that the relationships that exist in the data gathered in the past will still exist in the future. However, this assumption may not hold up in practice. For example, it may be the case that (according to the historical data collected) a student's grade in *Introductory Calculus* is highly correlated with their likelihood of completing a degree within 4 years. But, if the instructor of the course, the pedagogical technique employed, or the degree programs requiring the course change, this course may no longer be as predictive of degree completion as was originally thought. The practitioner should always consider whether patterns discovered in historical data should be expected to be present in future data.

A number of different algorithms exist for building predictive models. With educational data, it is common to see models built using methods such as:

1. **Linear Regression**, which is used to predict a numeric label from a linear combination of features.
2. **Logistic Regression**, which is used to predict the odds of two or more labels, allowing for categorical predictions.
3. **Nearest Neighbours Classifiers**, which use only the most similar data points in the training data set to determine the appropriate predicted labels for new data.
4. **Decision Trees (e.g. C4.5 algorithm)**, which are repeated partitions of the data based on a series of single attribute "tests". Each test is algorithmically chosen to maximize the purity of the classifications in each partition.
5. **Naïve Bayes Classifiers**, which assume statistical independence of each of the features given the classification, and provide probabilistic interpretations of classifications.
6. **Bayesian Networks**, where graphical models are often manually constructed and provide probabilistic interpretations of classifications.
7. **Support Vector Machines**, which make use of a high dimensional data projection in order to find a hyperplane of greatest separation between the various classes.
8. **Neural Networks**, which are biologically inspired algorithms that propagate data input through a series of sparsely interconnected layers of computational nodes (neurons) to produce a label. While neural networks have been the subject of research for more than 70 years, the area has received renewed interest (and commercial success) due to the advances of *Deep Learning*.
9. **Ensemble Methods**, which use a voting pool of either homogeneous or heterogeneous classifiers. Two prominent techniques are bootstrap aggregating, in which several predictive models are built from random sub-samples of the data set, and boosting, in which successive predictive models are designed to account for the misclassifications of the prior models.

Most of these methods, and their underlying software im-

[11] The authors share an anecdote of an analysis that has fallen prey to the issue of assuming independence of attributes when using resampling techniques to boost certain classes of data when applying the synthetic minority over-sampling technique [10]. In that case, missing data with respect to city and province resulted in a dataset containing geographically impossible combinations, reducing the effectiveness of the attributes and lowering the accuracy of the model.

plementations, have tunable parameters that change the way the algorithm works depending upon expectations of the dataset. For instance, when building decision trees, a researcher might set a minimum leaf size or maximum depth of tree parameter used in order to ensure some level of generalizability.

While R and Python are the two most commonly used programming languages for predictive modelling in the field[12], there are numerous specialized software libraries available for the building of predictive models in these and many other programming languages. Choosing the right package depends highly on the individual researchers experiences, the desired classification or regression approach, and the amount of data and data cleaning that needs to be done. While a comprehensive discussion and comparison of these platforms is out of the scope of this chapter, the authors will suggest that the freely available and open-source package Weka [16] is a an excellent starting point for those who are interested in predictive modelling but have little or no prior programming experience. Weka provides implementations of a number of the previously mentioned modelling methods, does not require programming knowledge to use, and has associated educational materials including a textbook [33] and series of free online courses [32].

While the breadth of techniques covered within a given software package have led to it being commonplace for researchers (including educational data scientists) to publish tables of classification accuracies for a number of different methods, the authors caution against this. Once a given technique has shown promise, time is better spent reflecting on the fundamental assumptions of classifiers (e.g. with respect to missing data or data set imbalance), exploring ensembles of classifiers, or in tuning parameters of particular methods being employed. Unless the intent of the research activity is to specifically compare two (or more) statistical modelling approaches, educational data scientists are better off tying their findings to new or existing theoretical constructs, leading to a deepening of understanding of a given phenomena. Sharing data and analysis scripts in an open science fashion provides better opportunity for small technique iterations than cluttering a publication with tables of (often) impenetrable and uninteresting measurements.

Evaluating a model

In order to assess the quality of a predictive model, a test data set with known labels is required. The predictions made by the model on the test set can be compared to the known true labels of the test set in order to assess the model. A wide variety of measures are available to compare the similarity of the known true labels and the predicted labels. Some examples include prediction accuracy (the raw fraction of test instances correctly classified), precision, and recall.

Often, when approaching a predictive modelling problem, only one omnibus set of data is available for building a predictive model. While it may be tempting to reuse this same data set as a test set to assess model quality, the performance of the predictive model will typically be significantly higher on this data set than would be seen on a novel data set (due to the model *overfitting* the training data set). Instead, it is common practice to "hold out" some fraction of the data set and use it solely as a test set to assess model quality.

The most simple approach is to set aside half of the data, and reserve it for testing. However, there are two drawbacks to this approach. First, by reserving half of the data for testing, the predictive model will only be able to make use of half of the data for model fitting. Generally speaking, model accuracy increases as the amount of available data increases. Thus, training using only half of the available data may result in predictive models with poorer performance than if all the data had been used. Second, our assessment of model quality will only be based on predictions made for half of the available data. Generally, increasing the number of instances in the test set will increase the reliability of the results. Instead of simply dividing the data into training and testing partitions, it is common to use a process of k-fold cross validation in which the data set is partitioned at random into k segments. k distinct predictive models are constructed, with each model training on all but one of the segments, and testing on the single held out segment. The test results are then pooled from all k test segments, and a generalized assessment of prediction quality can be performed. The important benefits of k-fold cross validation are that every available data point can be used as part of the test set, no single data point is ever used in both the training set and test set of the same classifier at the same time, and the training sets used are nearly as large as all of the available data.

An important consideration when putting predictive modeling into practice is the similarity between the data used for training the model and the data available when predictions need to be made. Often in the educational domain, predictive models are constructed using data from one or more time periods (e.g. semesters or years), and then applied to student data from the next time period. If the features used to construct the predictive model include factors such as students' grades on individual assignments, then the accuracy of the model will depend on how similar the assignments are from one year to the next. To get an accurate assessment of model performance, it is important to assess the model in the same manner as will be used in situ: to build the predictive model using data available from one year, and then construct a testing set consisting of data from the following year, instead of dividing data from a single year into training and testing sets.

PREDICTIVE ANALYTICS IN PRACTICE

Predictive analytics are being used within the field of teaching and learning for many purposes, with one significant body of work aimed at identifying students who

[12] see for example the number of workshops and tutorials introducing new researchers and practitioners to these tools at recent LAK and LASI events

are at risk in their academic programming. For instance, [1] describe the use of predictive models to determine whether students will graduate from secondary school on time, demonstrating how the accuracy of predictions changes as students advance from primary school through into secondary school. Predicted outcomes vary widely, and might include a specific summative grade or grade distribution for a student or class of achievement [9] in a course. Baker et al. [7] describe a method which predicts a formative achievement for a student based on their previous interactions with an intelligent tutoring system. In lower-risk and semi-formal settings such as Massive Open Online Courses (MOOCs), the chance that a learner might disengage from the learning activity mid-course is another heavily studied outcome [34, 28].

Beyond performance measures, predictive models have been used in teaching and learning to detect learners who are engaging in off-task behavior [35, 5] such as "gaming the system" in order to answer questions correctly without learning [6]. Psychological constructs such as affective and and emotional state have also been modeled with predictive models [11, 30], using a variety of underlying data as features, such as textual discourse or facial characteristics. More examples of some of the ways predictive modelling has been used in Educational Data Mining in particular can be found in [20].

At the same time, there are both warnings and criticism of the creation of predictive models for education which focus on the issue of deployment. Writing in [18], Andrew Ho reminds the reader that "...the purpose of education is not prediction but learning". He goes on, writing:

> In short, we want educational predictions to be wrong. If our predictive model can tell that a student is going to drop out, we want that to be true in the absence of intervention, but if the student does in fact drop out, then that should be seen as a failure of the system. A predictive model should be part of a prediction-and-response system that a) makes predictions that would be accurate in the absence of a response and b) enables a response that renders the prediction incorrect. In a good prediction-and-response system, all predictions would ultimately be negatively biased.

[18, p. 36]

In the broadest sense, we agree with this perspective – the intention of an applied predictive model should be to enable better education outcomes for learners, not simply to measure existing education outcomes. At the same time we argue that the issue is nuanced and that there is value in accurate educational predictive modeling both as a field of research and in real-world educational technologies. In the former the argument largely rests on the value of interdisciplinary teams to address the prediction-and-response system; whether tightly or loosely coupled, there is opportunity to the marriage of technical experts (e.g. computer scientists, statisticians, engineers) who might build models to the pedagogical experts (e.g. educational researchers, domain experts, cognitive psychologists) who might design interventions. Without these accurate models the job of building an intervention becomes not only harder to make, but harder to measure the effects of. Of pragmatic concern is the issue of limited resources within education systems. Simply put, most educational predictive models not only tell you who is likely to fail, but also who is likely to succeed, and allow institutions (and researchers) to focus their interventions directly towards specific populations of interest. Narrowing the population of students to whom an intervention is applied allows for more targeted and better resourced interventions. This is of specific value when engaging with educational policy makers who are often asked to resource a breadth of intervention programs and must balance the anticipated outcomes of different approaches. With this nuance explored, we reiterate that the key agreement we share with Ho is that the predictive model is only one half of the prediction-and-response system, and it is important for researchers and practitioners to keep this in mind.

CHALLENGES AND OPPORTUNITIES

Computational and statistical methods for predictive modelling are mature, and over the last decade a number of robust tools have been made available for educational researchers to apply predictive modelling to teaching and learning data. Yet there are a number of challenges and opportunities in this space, and we address a few areas of growth which could use investment from the learning analytics community in order to increase the impact predictive modelling techniques can have. These are:

1. **Supporting non-computer scientists in the educational predictive modelling workflow** Learning analytics is becoming normalized in higher education. Providing support in the interpretation and understanding of predictive modelling techniques, whether it be through the innovation of user-friendly tools or development of educational resources on predictive modelling, could help to assuage fear and uncertainty about algorithmic predictions.
   Related to this, the rise of Master of Data Science programs in recent years has greatly increased the number of highly skilled individuals capable of engaging successfully in predictive modelling. However, *Data Engineering*, the practice of provisioning data suitable for analysis, is a growing challenge. Students engage with a greater variety of learning tools than ever before, which provides an opportunity for incredibly rich analysis. But, these learning tools do not necessarily track comparable log events, retain log data in comparable formats, or have APIs (application programming interfaces) to integrate this data together. Many institutions are now engaged in the creation of learning record stores or data lakes to support the analysis of learning data aggregated across the range of learning tools that students interact with. As the number of technologies students use in their studies continues to grow, the need for data engineers to become a part of the interdisciplinary learning analytics

team is more apparent.

2. **Creating community-led educational data science challenge initiatives.** It is not uncommon for researchers to address the same general theme of work but use slightly different datasets, implementations, and outcomes and, as such, have results that are difficult to compare. This is exemplified in recent predictive modelling research efforts around dropout in massive open online courses, where a number of different authors (e.g. [9, 34, 28, 31]) have done work all with different datasets, outcome variables, and approaches.
   Moving towards a common and clear set of outcomes, open data, and shared implementations in order to compare the efficacy of techniques and the suitability of modelling methods for given problems could be beneficial for the community. This approach has been valuable in similar research fields[13] and the broader data science community[14], and we believe that educational data science challenges could help to disseminate predictive modelling knowledge throughout the educational research community while also providing an opportunity for the development of novel interdisciplinary methods, especially as it relates to feature engineering. Ryan Baker's six problems for the learning analytics community are an example of this community challenge initiative[4].
3. **Engaging in 2nd order predictive modelling.** In the context of learning analytics, we define second order predictive models as those which include historical knowledge as to the effects of and intervention in the model itself. Thus a predictive model which used student interactions with content to determine drop out (for instance) would be a example of first order predictive modelling, while a model which also includes historical data as to the effect of an intervention (such as an email prompt or nudge) would be considered a second order predictive model. Moving towards the modelling of intervention effectiveness is important when multiple interventions are available and personalized learning paths are desired.
4. **Bias in educational predictive models.** A growing concern in the predictive modeling and machine learning community is the potential for models to become biased with respect to their performance for different classes of people. In addition to being addressed within existing scholarly communities, this concern has spawned the creation of new academic conferences focused specifically on issues of bias and fairness (e.g. the ACM Conference on Fairness, Accountability, and Transparency (FAccT)[15]). Within the area of learning analytics specifically there have been a number of works looking at how to measure bias in predictive models [29], the impact of user choice on bias in models [23], and the bias in underlying methods applied in educational models [12]. What is lacking within the field, however, is an understanding of how evidence of bias should influence the use of predictive models in education. For instance, if a model has a bias against a given subpopulation, does that mean the model shouldn't be used at all? How big must the bias be before it is a concern? What subpopulations are important in a given learning context? These thorny sociotechnical issues need further exploration, as the work to date has largely been technical or measurement focused.

[13] http://www.kdd.org/kdd-cup
[14] http://www.kaggle.com/
[15] https://facctconference.org/

Despite the multi-disciplinary nature of the learning analytics and educational data mining communities, there is still a significant need for bridging understanding between the diverse array of scholars involved. An interesting thematic undercurrent at learning analytics conferences are the (sometimes heated) discussions of the roles of theory and data as drivers of educational research. Have we reached the point of "the end of theory" [3] in educational research? Unlikely, but this question is most salient within the subfield of *predictive modelling in teaching and learning*: while for some researchers the goal is understanding cognition and learning processes, others are interested in predicting future events and success as accurately as possible. With predictive models becoming increasingly complex and incomprehensible by an individual (essentially black boxes), it is important to start discussing more explicitly the goals of research agendas in the field, to better drive methodological choices between explanatory and predictive modelling techniques.

REFERENCES

[1] Everaldo Aguiar, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, and Kecia L Addison. "Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 2015, pp. 93–102.

[2] Sakinah Alhadad, Kimberly Arnold, Josh Baron, Ilana Bayer, Christopher Brooks, Russ R Little, Rose A Rocchio, Shady Shehata, and John Whitmer. *The Predictive Learning Analytics Revolution: Leveraging Learning Data for Student Success*. EDUCAUSE Center for Analysis and Research, 2015.

[3] Chris Anderson. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In: *Wired magazine* 16.07 (June 2008). URL: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

[4] Ryan S. J. d. Baker. "Challenges for the future of educational data mining: The Baker learning analytics prizes". In: *JEDM | Journal of Educational Data Mining* 11.1 (2019), pp. 1–17.

[5] Ryan S. J. d. Baker. "Modeling and understanding students' off-task behavior in intelligent tutoring systems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2007, pp. 1059–1068.

[6] Ryan S. J. d. Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z. Wagner. "Off-task behavior in the cognitive tutor classroom: when students game the system". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 383–390.

[7] Ryan S. J. d. Baker, Sujith M Gowda, and Albert T. Corbett. "Towards Predicting Future Transfer of Learning". In: *Artificial Intelligence in Education*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 23–30.

[8] Rebecca Barber and Mike Sharkey. "Course Correction: Using Analytics to Predict Course Success". In: *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*. LAK '12. event-place: Vancouver, British Columbia, Canada. New York, NY, USA: ACM, 2012, pp. 259–262. ISBN: 978-1-4503-1111-3. DOI: 10.1145/2330601.2330664. URL: http://doi.acm.org/10.1145/2330601.2330664.

[9] Christopher Brooks, Craig Thompson, and Stephanie Teasley. "A Time Series Interaction Analysis Method for Building Predictive Models of Learners Using Log Data". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. LAK '15. New York, NY, USA: ACM, 2015, pp. 126–135.

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* (2002), pp. 321–357.

[11] Sidney D'Mello, Scotty D Craig, Amy Witherspoon, Bethany McDaniel, and Arthur Graesser. "Automatic detection of learner's affect from conversational cues". In: *User Model. User-adapt Interact.* 18.1 (2007). Publisher: Springer Netherlands, pp. 45–80.

[12] Shayan Doroudi and Emma Brunskill. "Fairer but Not Fair Enough On the Equitability of Knowledge Tracing". In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 335–339. ISBN: 978-1-4503-6256-6. DOI: 10.1145/3303772.3303838. URL: https://doi.org/10.1145/3303772.3303838.

[13] Nia Dowell, Tristian Nixon, and Arthur C. Graesser. "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions". In: *Behavior Research Methods* 51 (2019), pp. 1007–1041.

[14] Angela L. Duckworth, Christopher Peterson, Michael D Matthews, and Dennis R Kelly. "Grit: perseverance and passion for long-term goals." In: *Journal of personality and social psychology* 92.6 (2007). Publisher: American Psychological Association, p. 1087.

[15] Josh Gardner and Christopher Brooks. "Coenrollment Networks and Their Relationship to Grades in Undergraduate Education". In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. LAK '18. event-place: Sydney, New South Wales, Australia. New York, NY, USA: Association for Computing Machinery, 2018, pp. 295–304. ISBN: 978-1-4503-6400-3. DOI: 10.1145/3170358.3170373. URL: https://doi.org/10.1145/3170358.3170373.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA Data Mining Software: An Update". In: *SIGKDD Explor. Newsl.* 11.1 (Nov. 2009). Place: New York, NY, USA Publisher: ACM, pp. 10–18. ISSN: 1931-0145. DOI: 10.1145/1656274.1656278. URL: http://doi.acm.org/10.1145/1656274.1656278.

[17] HarvardX. *HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0*. 2014. DOI: 10.7910/DVN/26147. URL: https://doi.org/10.7910/DVN/26147.

[18] Andrew Ho. "Four variations on a theme of data-intensive research in education". In: *Workshop on Data-Intensive Research in Education*. Arlington, VA: Computing Research Association, 2015, pp. 79–82. URL: http://archive2.cra.org/uploads/documents/events/bigdata/Workshop_Briefing_Book.pdf.

[19] Sreckp Joksimovic, Areti Manataki, Dragan Gasevic, Shane Dawson, Vitomir Kovanovic, and Inés Friss de Kereki. "Translating Network Position into Performance: Importance of Centrality in Different Network Configurations". In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. LAK '16. event-place: Edinburgh, United Kingdom. New York, NY, USA: Association for Computing Machinery, 2016, pp. 314–323. ISBN: 978-1-4503-4190-5. DOI: 10.1145/2883851.2883928. URL: https://doi.org/10.1145/2883851.2883928.

[20] Kenneth Koedinger, Sidney D'Mello, Elizabeth A McLaughlin, Zachary A Pardos, and Carolyn P Rosé. "Data mining and education". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 6.4 (2015). Publisher: Wiley Online Library, pp. 333–353.

[21] Vitomir Kovanovic, Dragan Gašević, Shane Dawson, Srećko Joksimovic, and Ryan S. J. d. Baker. "Does time-on-task estimation matter? Implications on validity of learning analytics findings". In: *Journal of Learning Analytics* 2.3 (2016), pp. 81–110.

[22] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. "Open university learning analytics dataset". In: *Scientific data* 4 (2017). Publisher: Nature Publishing Group, p. 170171.

[23] Warren Li, Christopher Brooks, and Florian Schaub. "The Impact of Student Opt-Out on Educational Predictive Models". In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. LAK19. event-place: Tempe, AZ, USA. New York, NY, USA: Association for Computing Machinery, 2019, pp. 411–420. ISBN: 978-1-4503-6256-6. DOI: 10.1145/3303772.3303809. URL: https://doi.org/10.1145/3303772.3303809.

[24] Steven Lonn and Stephanie D. Teasley. "Student Explorer: A Tool for Supporting Academic Advising at Scale". In: *Proceedings of the First ACM Conference on Learning @ Scale Conference*. L@S '14. event-place: Atlanta, Georgia, USA. New York, NY, USA: ACM, 2014, pp. 175–176. ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2567867. URL: http://doi.acm.org/10.1145/2556325.2567867.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality*. _eprint: 1310.4546. 2013.

[26] Galit Shmueli. "To explain or to predict?" In: *Stat. Sci.* (2010). Publisher: JSTOR, pp. 289–310.

[27] Jack Stripling, Katherine Mangan, Nick DeSantis, Rio Fernandes, Sarah Brown, Steve Kolowich, Patricia McGuire, and Anne Hendershott. "Uproar at Mount St. Mary's". In: *The Chronicle of Higher Education* (Mar. 2016). Publisher: The Chronicle of Higher Education. URL: http://chronicle.com/specialreport/Uproar-at-Mount-St-Marys/30.

[28] Colin Taylor, Kalyan Veeramachaneni, and Una-May O'Reilly. "Likely to stop? Predicting Stopout in Massive Open Online Courses". In: (2014). _eprint: 1408.3382.

[29] Dirk Tempelaar, Bart Rienties, and Quan Nguyen. "Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application". In: *PloS one* 15.6 (2020), e0233977.

[30] Yutao Wang, Neil T Heffernan, and Cristina Heffernan. "Towards better affect detectors: effect of missing skills, class features and common wrong answers". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 2015, pp. 31–35.

[31] Jacob Whitehill, Joseph Jay Williams, Glenn Lopez, Cody Austun Coleman, and Justin Reich. "Beyond prediction: First steps toward automatic intervention in MOOC student stopout". In: *Available at SSRN 2611750* (2015).

[32] Ian H. Witten. *Weka Courses*. 2016. URL: https://weka.waikato.ac.nz/explorer.

[33] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN: 0-12-374856-9 978-0-12-374856-0.

[34] Wanli Xing, Xin Chen, Jared Stein, and Michael Marcinkowski. "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization". In: *Comput. Human Behav.* 58 (May 2016), pp. 119–129.

[35] Wanli Xing and Sean Goggins. "Learning analytics in outer space: a Hidden Naïve Bayes model for automatic student off-task behavior detection". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 2015, pp. 176–183.