Chapter 5: Predictive Modelling in Teaching and Learning

Christopher Brooks, Craig Thompson

[1] School of Information, University of Michigan, US
[2] Department of Computer Science, University of Saskatchewan, Canada

ABSTRACT

This article 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 (LA) predictive modelling has become a core practice of researchers, largely with a focus on predicting student success as operationalized by academic achievement. In this chapter, we provide a general overview of considerations when using predictive modelling, the steps that 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 modeling, 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 administrations (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] Civitas Learning[6]) now provide predictive analytics consulting and products for higher education.

In this chapter, we 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) and the International Educational Data Mining Society (IEDMS) for more examples of applied educational predictive modelling.

First, it is important to distinguish predictive modelling from explanatory modelling.[7] In explanatory modelling, the goal is to use all available evidence to provide an 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 be causal (versus correlative alone), though results presented using these approaches often eschew experimental studies and rely on theoretical interpretation to imply causation (as described well by Shmueli, 2010).

In predictive modelling, the purpose 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

[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] Shmueli (2010) notes a third form of modelling, descriptive modelling, which is similar to explanatory modelling but in which there are no claims of causation. In the higher education literature, we would suggest that causation is often implied, and the majority of descriptive analyses are actually intended to be used as causal evidence to influence decision making.

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 principal 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 a post-hoc and reflective activity aimed at generating an understanding of a phenomenon. 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, Lonn and Teasley (2014) describe a student-success system built on explanatory models, while Brooks, Thompson, and Teasley (2015) 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 generalizability. 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 the predictive modelling process.

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 modelling activity is to set up a scenario that 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 modelled, 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 ordered year after year, where the historical data on learners (the training set) is indicative of future learners (the testing set).

Conversely, several factors 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 capture the intended data accurately, 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 Stripling et al., 2016).

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 available 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., Barber & Sharkey, 2012), each corresponding to a different time period and set of observed variables. For instance, one might generate predictive models for each week of the course, incorporating into each model the results of weekly quizzes, student demographics, and the amount of engagement the students have had with respect digital resources to date in the course.

While state-based data, such as data about demographics (e.g., gender, ethnicity), relationships (e.g., course enrollments), psychological measures (e.g., grit, as in Duckworth, Peterson, Matthews, & Kelly, 2007, 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 Alhadad et al., 2015 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 in the order of millions of database rows for a single course), and requires significant effort to convert into meaningful features for machine learning.

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). The issue of converting complex data (as is the case with event-based data) into features suitable for predictive modelling is referred to as *feature engineering*, and is a broad area of research itself.

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 as 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 values, while regression algorithms are used to predict numeric values.

Feature Selection

In order to build and apply a predictive model, features that correlate with the value to predict must be created. 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 if they were not in attendance). In practice, the dependencies between features are often ignored, but it is important to note that some techniques used to clean and manipulate data may rely upon an assumption of independence.[8] 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.

[8] The authors share an anecdote of an analysis that fell prey to the dangers of assuming independence of attributes when using resampling techniques to boost certain classes of data when applying the synthetic minority over-sampling technique (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). 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.

Missing values in a dataset may be dealt with in several ways, and the approach used depends on whether data is missing because it is unknown or because it is not applicable. The simplest approach either is to 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 dataset can be significant, especially if the removal of some data exacerbates an existing class imbalance. Likewise, if all attributes have a small handful 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 "normal" 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 dataset, 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 the 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 that are propagated down the tree and used for a weighted voting. In short, missing data is an important consideration that both regularly occurs and is handled differently depending upon the machine learning method and toolkit employed.

Methods for Building Predictive Models

After collecting a dataset 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 prediction of some unknown quantity or attribute, 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. However, if there is a change in the instructor of the course, the pedagogical technique employed, or the degree programs requiring the course, 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 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 these:

1. **Linear Regression** predicts a continuous numeric output from a linear combination of attributes.
2. **Logistic Regression** predicts the odds of two or more outcomes, allowing for categorical predictions.
3. **Nearest Neighbours Classifiers** use only the closest labelled data points in the training dataset to determine the appropriate predicted labels for new data.
4. **Decision Trees** (e.g., C4.5 algorithm) are repeated partitions of the data based on a series of single attribute "tests." Each test is chosen algorithmically to maximize the purity of the classifications in each partition.
5. **Naïve Bayes Classifiers** assume the statistical independence of each attribute given the classification, and provide probabilistic interpretations of classifications.
6. **Bayesian Networks** feature manually constructed graphical models and provide probabilistic interpretations of classifications.
7. **Support Vector Machines** use a high dimensional data projection in order to find a hyperplane of greatest separation between the various classes.
8. **Neural Networks** are biologically inspired algorithms that propagate data input through a series of sparsely interconnected layers of computational nodes (neurons) to produce an output. Increased interest has been shown in neural network approaches under the label of *deep learning*.
9. **Ensemble Methods** 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 dataset, 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 implementations, 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.

Numerous software packages are available for the building of predictive modelling, and choosing the right package depends highly on the researcher's experience, the desired classification or regression approach, and the amount of data and data cleaning required. While a comprehensive discussion of these platforms is outside the scope of this chapter, the freely available and open-source package Weka (Hall et al., 2009) 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 (Witten, Frank, & Hall, 2011) and series of free online courses (Witten, 2016).

While the breadth of techniques covered within a given software package has 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 dataset imbalance), exploring ensembles of classifiers, or tuning the parameters of particular methods being employed. Unless the intent of the research activity is to compare two statistical modelling approaches specifically, educational data scientists are better off tying their findings to new or existing theoretical constructs, leading to a deepening of understanding of a given phenomenon. 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) uninteresting precision and recall values.

Evaluating a Model

In order to assess the quality of a predictive model, a test dataset 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 is 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. While it may be tempting to reuse this same dataset as a test set to assess model quality, the performance of the predictive model will be significantly higher on this dataset than would be seen on a novel dataset (referred to as overfitting the model). Instead, it is common practice to "hold out" some fraction of the dataset and use it solely as a test set to assess model quality.

The simplest approach is to remove 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, 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 would 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 dataset 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 an assessment of model 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 modelling 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. 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 at risk in their academic programming. For instance, Aguiar et al. (2015) 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 (Brooks et al., 2015) in a course. Baker, Gowda, and Corbett (2011) describe a method that 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 (Xing, Chen, Stein, & Marcinkowski, 2016; Taylor, Veeramachaneni, & O'Reilly, 2014).

Beyond performance measures, predictive models have been used in teaching and learning to detect learners who are engaging in off-task behaviour (Xing and Goggins, 2015; Baker, 2007) such as "gaming the system" in order to answer questions correctly without learning (Baker, Corbett, Koedinger, & Wagner, 2004). Psychological constructs such as affective and emotional states have also been predictively modelled (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2007; Wang, Heffernan, & Heffernan, 2015), 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 Koedinger, D'Mello, McLaughlin, Pardos, and Rosé (2015).

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 a number of challenges and opportunities face the learning analytics community when building, validating, and applying predictive models. We identify three areas that could use investment in order to increase the impact that predictive modelling techniques can have:

1. *Supporting non-computer scientists in predictive modelling activities*. The learning analytics field is highly interdisciplinary and educational researchers, psychometricians, cognitive and social psychologists, and policy experts tend to have strong backgrounds in explanatory modelling. Providing support in the application of predictive modelling techniques, whether through the innovation of user-friendly tools or the development of educational resources on predictive modelling, could further diversify the set of educational researchers using these techniques.

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 regarding dropout in massive open online courses, where a number of different authors (e.g., Brooks et al., 2015; Xing et al., 2016; Taylor et al., 2014; Whitehill, Williams, Lopez, Coleman, & Reich, 2015) have all done work 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 and the broader data science community 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 related to feature engineering.

3. Engaging in second order predictive modelling. In the context of learning analytics, we define second order predictive models as those that include historical knowledge as to the effects of and intervention in the model itself. Thus a predictive model that used student interactions with content to determine drop out (for instance) would be an example of first order predictive modelling, while a model that 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.

Despite the multidisciplinary nature of the learning analytics and educational data mining communities, there is still a significant need for bridging understanding between the diverse 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" (Anderson, 2008) 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

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K. L. (2015). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 93–102). New York: ACM.

Alhadad, S., Arnold, K., Baron, J., Bayer, I., Brooks, C., Little, R. R., Rocchio, R. A., Shehata, S., & Whitmer, J. (2015, October 7). The predictive learning analytics revolution: Leveraging learning data for student success. Technical report, EDUCAUSE Center for Analysis and Research.

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. Wired. https://www.wired.com/2008/06/pb-theory/

Baker. R. S. J. d. (2007). Modeling and understanding students' on-task behaviour in intelligent tutoring systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 28 April–3 May 2007, San Jose, CA (pp. 1059–1068). New York: ACM.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). On-task behaviour in the cognitive tutor classroom: When students game the system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '04), 24–29 April 2004, Vienna, Austria (pp. 383–390). New York: ACM.

Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2011). Towards predicting future transfer of learning. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED '11), 28 June–2 July 2011, Auckland, New Zealand (pp. 23–30). Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 259–262). New York: ACM. doi:10.1145/2330601.2330664

Brooks, C., Thompson, C., & Teasley, S. (2015). A time series interaction analysis method for building predictive models of learners using log data. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 126–135). New York: ACM.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2007). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1–2), 45–80.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), 1087–1101.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1), 10–18. doi:10.1145/1656274.1656278.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4), 333–353.

Lonn, S., & Teasley, S. D. (2014). Student explorer: A tool for supporting academic advising at scale. *Proceedings of the 1st ACM Conference on Learning @ Scale* (L@S 2014), 4–5 March 2014, Atlanta, Georgia, USA (pp. 175–176). New York: ACM. doi:10.1145/2556325.2567867

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

Stripling, J., Mangan, K., DeSantis, N., Fernandes, R., Brown, S., Kolowich, S., McGuire, P., & Hendershott, A. (2016, March 2). Uproar at Mount St. Mary's. The Chronicle of Higher Education. http://chronicle.com/specialreport/Uproar-at-Mount-St-Marys/30.

Taylor, C., Veeramachaneni, K., & O'Reilly, U.-M. (2014, August 14). Likely to stop? Predicting stopout in massive open online courses. http://dai.lids.mit.edu/pdf/1408.3382v1.pdf

Wang, Y., Heffernan, N. T., & Heffernan, C. (2015). Towards better affect detectors: Effect of missing skills, class features and common wrong answers. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 31–35). New York: ACM.

Whitehill, J., Williams, J. J., Lopez, G., Coleman, C. A., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. XXX–XXX). International Educational Data Mining Society. http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_112.pdf

Witten, I. H. (2016). Weka courses. The University of Waikato. https://weka.waikato.ac.nz/explorer

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*, 3rd ed. San Francisco, CA: Morgan Kaufmann Publishers.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low-hanging fruit through stacking generalization. *Computers in Human Behavior*, 58, 119–129.

Xing, W., & Goggins, S. (2015). Learning analytics in outer space: A hidden naive Bayes model for automatic students' on-task behaviour detection. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 176–183). New York: ACM.