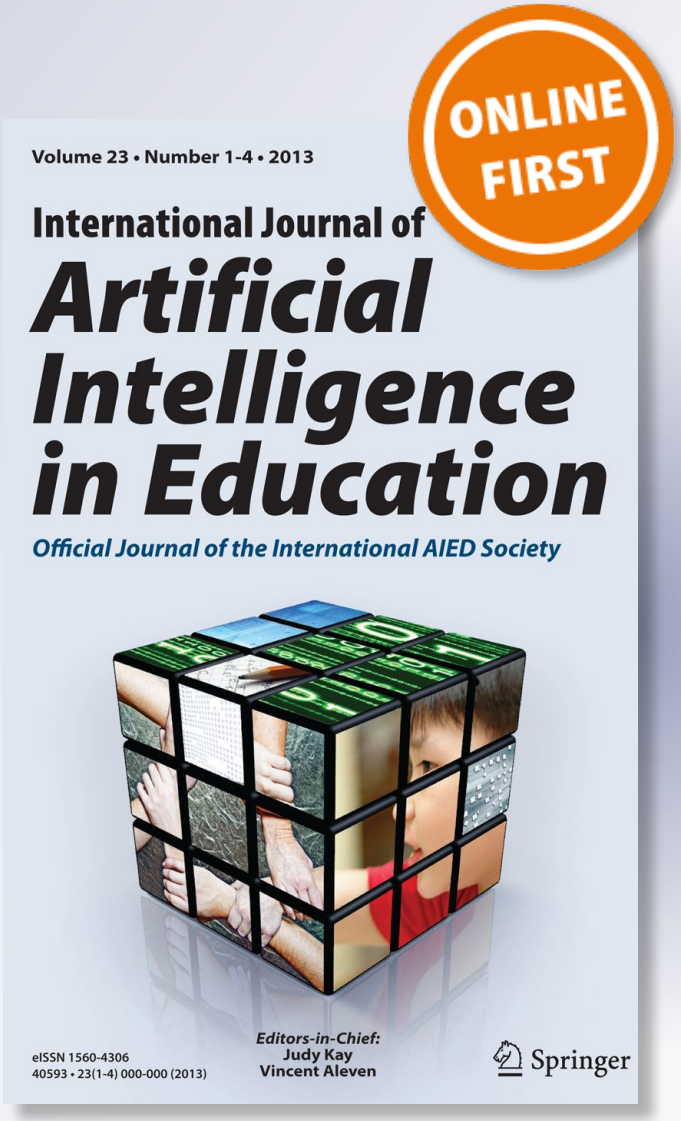
Ensemble Learning for Estimating Individualized Treatment Effects in Student Success Studies

Joshua Beemer, Kelly Spoon, Lingjun He, Juanjuan Fan & Richard A. Levine

International Journal of Artificial Intelligence in Education
Official Journal of the International AIED Society

ISSN 1560-4292

Int J Artif Intell Educ
DOI 10.1007/s40593-017-0148-x



Springer

Int J Artif Intell Educ
DOI 10.1007/s40593-017-0148-x

ARTICLE

Ensemble Learning for Estimating Individualized Treatment Effects in Student Success Studies

Joshua Beemer[1] · Kelly Spoon[1] · Lingjun He[2] · Juanjuan Fan[3] · Richard A. Levine[2,3]

Abstract Student success efficacy studies are aimed at assessing instructional practices and learning environments by evaluating the success of and characterizing student subgroups that may benefit from such modalities. We propose an ensemble learning approach to perform these analytics tasks with specific focus on estimating individualized treatment effects (ITE). ITE are a measure from the personalized medicine literature that can, for each student, quantify the impact of the intervention strategy on student performance, even though the given student either did or did not experience this intervention (i.e., is either in the treatment group or in the control group). We illustrate our learning analytics methods in the study of a supplemental instruction component for a large enrollment introductory statistics course recognized as a curriculum bottleneck at San Diego State University. As part of this application, we show how the ensemble estimate of the ITE may be used to assess

✉ Richard A. Levine
rlevine@mail.sdsu.edu

Joshua Beemer
joshbeemer@hotmail.com

Kelly Spoon
kellyspoon@gmail.com

Lingjun He
lingjun.he@mail.sdsu.edu

Juanjuan Fan
jjfan@mail.sdsu.edu

[1] Computational Science Research Center, San Diego State University, San Diego, CA, USA

[2] Analytic Studies and Institutional Research, San Diego State University, San Diego, CA, USA

[3] Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA

Published online: 25 May 2017

the pedagogical reform (supplemental instruction), advise students into supplemental instruction at the beginning of the course, and quantify the impact of the supplemental instruction component on at-risk subgroups.

Keywords Educational data mining · Personalized learning · Machine learning · Regularized regression · Supplemental instruction

Introduction

In striving to improve graduation rates and reduce achievement gaps, Universities have experimented with a suite of instructional practices and learning environments (for example, see the 2015 issue 2 of *Peer Review* from the Association of American Colleges & Universities). Broadly speaking, these strategies foster student success and engagement through common and collaborative intellectual experiences, student research and internships, and study abroad experiences (Kuh 2008) as well as supplemental instruction and instructional technologies (for example see Dawson et al. 2014; Henrie et al. 2015). An analytics goal is identifying at-risk students that will benefit from one or more of these intervention strategies and, early in their college careers, advise these students accordingly. On the flip side, we also must evaluate each instructional practice and each learning environment on at-risk subgroups for purposes of strategic planning, resource allocation, and program development.

We propose an ensemble learning approach to estimate individualized treatment effects (ITE) to characterize at-risk students and assess student success and retention under intervention strategies. ITE were introduced in the personalized medicine literature (Dorresteijn et al. 2011) to quantify the difference in an outcome of interest between treatment and control for any subject, whether they experience only the treatment or only the control modality. In our setting, ITE allow us to predict the performance difference between experiencing an intervention strategy or not for each student. We may use these predictions to

- evaluate the success of an intervention;
- characterize student subgroups that may benefit from an intervention;
- evaluate the impact of an intervention on at-risk subgroups;
- quantify the impact of an intervention on individual students; and
- provide an early warning system to advise students into an intervention.

An ensemble learning approach provides a natural analytics environment within which to leverage the wealth of data on students from student information system databases and learning management systems to estimate ITE and student success measures. The student success measures may include categorical outcomes, such as non-repeatable grade in a course (e.g., C or better), graduation success, and retention, or continuous outcomes such as course grade (e.g., on a four-point GPA scale), final exam score, and time to graduation. In a learning analytics setting, a set of base learners are trained and then used to predict the student success outcome of interest for each student. A meta-learner combines these base learner predictions for each student. Moon et al. (2007) proves that in the case of classification, an ensemble average

(meta-learner) over a suite of classifiers (base learners) will improve accuracy over a single classifier (single base learner). The case is not as clear cut in a regression context, though Moreira et al. (2012) presents a number of approaches to create an ensemble with improved prediction accuracy.

To our knowledge, the educational literature contains only a few student success studies that take advantage of ensemble learning. Pardos et al. (2011a, b) considers ensemble methods to combine latent student knowledge predictions from multiple models of data within a tutoring system. Kotsiantis et al. (2010) considers the use of ensemble methods to predict student success in distance learning using three different techniques (WINNOW, naive Bayes, and 1-nearest neighbour). Cortez and Silva (2008) compares several ensemble techniques to assess student performance under binary, multi-class factor, and continuous responses. Though no ensemble learning approach is applied, Jayaprakash et al. (2014) evaluates a set of base learners (logistic, SVM, decision trees, and naive Bayes) for predicting academic risk. To our knowledge, the education literature contains only two applications of estimating treatment effects, in the context of digital learning environments without random assignment. Beck and Mostow (2008) apply learning curve analysis using nonlinear regression to estimate individual student learning in acquiring reading skills. Pardos et al. (2011a, b) apply Bayesian knowledge tracing to study the effectiveness of tutorial help in a math tutoring system.

In "Analytics Methods", we detail our ensemble learning approach and computation of ITE. In "Application: Impact of Supplemental Instruction Section on Student Success in Introductory Statistics", we step through the applications of ITE in student success studies. For purposes of illustration, we evaluate the success of a supplemental instruction course introduced in a San Diego State University (SDSU) large enrollment introductory statistics course. We stress that the ensemble learning approach we propose is modular. In our application we fix the set of base learners we consider. However, as a function of application ease, computational cost/complexity, or prediction performance, any base learner may be used as part of the ensemble. Analogously, we introduce stacked generalization (Alpaydin 2010, Chapter 17) to construct the meta-learner. Again, this component of the ensemble learner is modular, allowing flexibility in choice of meta-learner for combining the base learner predictions. In "Discussion", we provide a concluding summary, limitations of our proposed approach, and recommendations for future research.

Analytics Methods

In this section we detail the ensemble learning approach for student success study analytics applications. We then detail predicting individualized treatment effects (ITE) with the ensemble learner.

Ensemble Learning

Ensemble learning entails combining predictions from a set of base learners. Intuitively, the ensemble balances base learners that over fit and under fit the data, with

an aim of improving overall prediction accuracy. An ensemble learner will see the greatest gain in predictive performance when combining diverse predictions, that is, base learner predictions that are not highly correlated. A basic ensemble learner is a weighted sum of the predictions from each base learner, weights by minimizing an objective criterion such as mean squared error, likelihood, or entropy (Alpaydin 2010, Chapter 17). We focus on stacked generalization (Wolpert 1992) to combine the base learners. The particular form we use is a variation of stacked regression introduced to the statistics literature by Brieman (1996) and LeBlanc and Tibshirani (1996).

Algorithm 1 presents pseudocode for our proposed ensemble learner. The algorithm requires three data subsets created within nested cross-validation loops. Suppose we have n students in our data set. Let the size of the validation set be denoted by K_E. The first cross-validation loop (steps 1–3) randomly divides the data into n/K_E subsets of students of size K_E. For example, in leave-one-out cross-validation, $K_E = 1$; in ten-fold cross-validation, $K_E = n/10$; etc. In each cycle of the first cross-validation loop, we put aside the K_E students for that subset as a *validation set*. The remaining $n_T = (n - K_E)$ students we call the *ensemble training set*.

Algorithm 1 Ensemble learner: stacked generalization

1. `Randomly` partition the data into subsets of size K_E.
2. `Fix` cross-valdiation counter $cv = 1$. (Note that $cv \in \{1, \ldots, n/K_E\}$.)
3. `Label` the subset of K_E students in data partition cv as the validation set and the remaining $n_E = (n - K_E)$ as the ensemble training set.
4. `Choose` L base learners for constructing the ensemble learner.
5. `Randomly partition` the ensemble training set into subsets of size K.
6. For each partition,
   - `Label` the subset of K students as the test set and the remaining $n_E - K$ students as the training set;
   - `Fit` each of the L base learners to the training set;
   - `Obtain` a prediction for each student in the test set from each fitted base learner.

   end-loop over each K-fold cross validation partition.
7. `Regress` true outcome on the predictions from each base learner: L predictions for each student as inputs into the regression model on n_E students.
8. `Fit` each of the L base learners to the ensemble training set.
9. `Obtain` a prediction for each student in the validation set from each fitted base learner from step 8.
10. `Combine` the predictions from step 9 using the regression coefficient estimates from step 7 as weights in the linear combination.
11. `Increment` cv by one.
12. `Repeat` steps 3–10 until $cv > n/K_E$.

We perform K-fold cross-validation on the ensemble training set. That is, we randomly partition the ensemble training set into n_E/K subsets of K students each (step 5).

In each cycle of this cross-validation loop (step 6), we put aside the K students for that subset as a *testing set*. We then train each base learner chosen on the *training set* of $n_E - K$ students left. Again, this training may be performed using leave-one-out cross validation by setting $K = 1$. The trained base learners are then used to predict the outcome of interest for each of the K students in the testing set. At the end of this loop (steps 5–6), we have a prediction for each of the n_E students in the ensemble training set from each base learner.

The meta-learner entails a regression (step 7) of the true outcome on the predictions from each base learner for the n_E students in the ensemble training set. The regression coefficients represent the weights for combining the base learners into an ensemble prediction. Breiman (1996) notes that the base learner predictions may be highly correlated leading to challenges if linear regression (via ordinary least squares, OLS) is used as the meta-learner. Ridge regression (James et al. 2013, Chapter 6), a common approach in the presence of multi-collinearity, is suggested. As an extension, Reid and Grudic (2009) proposes regularization which also allows for lasso or elastic net (James et al. 2013, Chapter 6) regression techniques. These latter methods provide an alternative method of shrinkage estimation that may select a weight of zero (sparse model) for a base learner. In our application, we find ridge regression sufficient for estimating ITE. However, regularization provides options for stacked generalization to avoid overfitting and improve predictive performance.

The so-called validation set contains students left out of the process for constructing the meta-learner. We thus may use the meta-learner to make predictions for each of the students in the validation set at the conclusion of the outer cross-validation loop (over cv). First, each base learner is trained on the ensemble training set (set 8). A set of predictions is then made for each student in the validation via each base learner (step 9). We thus will have L predictions for each of the n/K_E students in the validation set. These L predictions are combined using the meta-learner (step 10). We thus come out of the outer cross-validation loop with predictions for each student in the data set, predictions made in groups of n/K_E.

With nested cross-validation loops, Algorithm 1 appears computationally costly for large data sets. However, the outer cross-validation loop (steps 1–3; validation set) is easily performed in parallel on say a cluster computer. The inner cross-validation loop (steps 5–6; ensemble training set) may also be performed in parallel upon identification of the validation set.

Individualized Treatment Effects

In student success studies, we wish to quantify the difference in outcome under an intervention and under a control regime (typically no intervention). Of course the student will typically either experience the intervention or not. A crossover type design or randomized controlled experiment is typically not an option for studying, for example, instructional practices and learning environments. We can apply the ensemble learning algorithm predictions of "Ensemble Learning" to compute an individualized treatment effect for each student. Algorithm 2 presents the pseudocode.

Algorithm 2 Individualized Treatment Effects (ITE)

1. `Separate` data into treatment group and control group
2. `Train` the ensemble learner of Algorithm 1 on the treatment group
3. `Train` the ensemble learner of Algorithm 1 on the control group
4. `Obtain` a "under treatment" prediction for control group subjects using the treatment group trained learner from step 2
5. `Obtain` a "no-treatment" prediction for treatment group subjects using the control group trained learner from step 3
6. `Compute` ITE for the control group as the difference of the predicted outcome from step 4 and the observed outcome
7. `Compute` ITE for the treatment group as the difference of the observed outcome and the predicted outcome from step 5

In a given study we will have a set of students that receive the "treatment" (experience the intervention strategy) and a set of students that receive the "control" (do not experience the intervention strategy). We may train an ensemble learner on each group separately using Algorithm 1. We then predict the "no-treatment" outcome for the treatment group students using the ensemble learner trained on the control group. The individualized treatment effect for these treatment group students is the difference of the observed outcome under treatment and the predicted outcome under control (step 6). Analogously, we predict the "under treatment" outcome for the control group students using the ensemble learner trained on the treatment group. The individualized treatment effect for these control group students is the difference of the predicted outcome under treatment and the observed outcome under control (step 7). Overall, the ensemble learner is serving as a best guess of the outcome for the treatment (control) group students if they had experienced the control (treatment). Note that the ITE here are formulated as (outcome under treatment) minus (outcome under control). Thus the treatment group ITE are (observed-predicted) and the control group ITE are (predicted-observed).

Application: Impact of Supplemental Instruction Section on Student Success in Introductory Statistics

The California State University (CSU) Chancellor's Office has recently instituted the "Promising Practices for Course Redesign" program aimed at improving student success in bottleneck, typically large enrollment courses (http://courseredesign.csuprojects.org/wp/). Introductory Statistics was identified by CSU as one such bottleneck course, affecting STEM, business, and quantitatively-oriented non-STEM majors. Of particular concern are repeatable grades (at CSU these are grades of C- or worse as well as a withdrawal, W) which in turn potentially lead to lower (STEM) retention/persistence rates, decreased graduation rates, and increased time to graduation.

The San Diego State University (SDSU) Introductory Statistics course under consideration here (STAT 119) enrolls on the order of 1200 students per semester with DFW rate around 30%. DFW denotes grades of D (1.0 on a four-point grade scale), F (failing grade; 0.0 on a four-point grade scale), and withdrawal from the course. To combat this high DFW rate, one arm of our course redesign project introduced supplemental instruction sections to the course. Each section enrolls 20–30 students and meets twice per week for one hour each. The sections are lead by Statistics graduate student teaching assistants (TA) trained prior to the semester for developing an active problem solving environment in the classroom (Savery 2006). Rather than students watching TAs solve problems, the sessions entail students working through problems related to the topics of the week. The TAs circulate around the room answering questions and facilitate group/class discussions of common conceptual difficulties. Due to caps in general elective units for major programs, this supplemental instruction section is selected voluntarily by STAT 119 students for one additional credit unit.

The supplemental instruction section differs from the UMKC model of Supplemental Instruction (SI; often capitalized to note this particular implementation) originally proposed in 1973 (Martin and Arendale 1993). In particular, though students in our study volunteer into the supplemental instruction section, they must enroll in a one-unit course STAT 119A. Furthermore, STAT 119 students who perform below a 70% on an algebra assessment the first week of classes are strongly encouraged to enroll in the supplemental instruction section. Typical SI implementations use "near-peers", namely students who recently took and succeeded in, above a chosen grade threshold, the given course. The STAT 119A instructors are statistics graduate students. That said, the STAT 119A instructors are trained using the SI peer-assisted learning model, facilitate topic content and study skill discussions much like traditional SI sessions, and are regularly evaluated by a course coordinator (akin to an SI supervisor). See Dawson et al. (2014) for discussion of deviations from the traditional SI model in practice.

We consider the initial Fall 2013 offering of supplemental instruction in STAT 119 enrolling 17% of students in the course. We consider three student success outcomes: final exam score (on a scale of 0 to 300), final grade in the course (on a four-point GPA scale), and non-repeatable grade indicator (binary response of whether a student received a grade of 'C' or better). Table 1 presents descriptive summaries of the STAT 119 and STAT 119A students over a number of key variables in this study. Variables that are not self-explanatory: admission basis identifies a student as a first-time freshman or transfer student; first-generation college identifies a student as being the first in the immediate family to attend college; quiz 0 is an algebra assessment made at the beginning of the semester; and the AP indicators present whether a student took AP Calculus and AP Statistics in high school. STAT 119 is offered in a standard lecture format and in a hybrid modality, where the two class meetings each week are divided into one live lecture and one synchronous, but archived, online lecture. Table 1 thus reports the percentage of students enrolled in the hybrid offering and average number of online units for each group. The complete set of inputs for our model are presented later in Tables 3 and 4.

Table 1 Summary statistics on a number of key variables for students enrolled in the STAT 119: Introductory Statistics course as a whole, the subset of STAT 119 students who enrolled in the STAT 119A supplemental instruction course, and the subset of STAT 119 students who did not enroll in the STAT 119A supplemental instruction course

	STAT 119 (n = 1032)	Enrolled in STAT 119A (n = 169)	Not enrolled in STAT 119 A (n = 863)
Gender (female)	48%	64%	45%
EOP	14%	22%	13%
Live in dorm	52%	32%	56%
Age	19.5 (2.3)	19.7 (2.3)	19.4 (2.3)
Low income	33%	40%	32%
Pell eligible	31%	38%	29%
Level (Freshman, Soph, Junior, Senior)	74%, 12%, 9%, 5%	63%, 20%, 12%, 5%	77%, 10%, 8%, 5%
Admission basis	92% FTF	92% FTF	92% FTF
First-Gen College	18%	22%	17%
Online units	1.6 (3.5)	2.5 (4.0)	1.4 (3.4)
Hybrid class	78%	79%	77%
SAT Math	553 (81)	519 (82)	561 (79)
SAT verbal	509 (99)	496 (97)	512 (99)
HS GPA	3.47 (0.46)	3.44 (0.45)	3.48 (0.46)
Took Calculus? (AP)	38% (23%)	36% (17%)	38% (24%)
Took Statistics? (AP)	32% (14%)	30% (9%)	32% (15%)
Quiz 0: Score	0.75 (0.24)	0.72 (0.23)	0.75 (0.24)
Quiz 0: Time (min.)	27.9 (11.3)	28.4 (11.4)	27.8 (11.3)
HW 1: Score	0.95 (0.17)	0.96 (0.15)	0.94 (0.17)
HW 1: Time (min.)	80.9 (56.7)	80.1 (51.1)	81.2 (57.8)
Final exam	0.67 (0.24)	0.71 (0.19)	0.66 (0.25)
% Pass course	74%	83%	72%

Categorical variables are summarized as percentages for the category names (e.g., "Gender (female)" shows 48% of the STAT 119 students are female). Continuous variables are summarized through the average for that group with standard deviation in parentheses

We use this study data to illustrate ensemble learning for performing predictive and learning analytics in student success studies as follows:

- Does the supplemental instruction section work? Quantify the impact of the supplemental instruction section on course success.
- On whom does the supplemental instruction section work? Identify characteristics of students benefitting from the supplemental instruction section.
- By how much does the supplemental instruction section work? Quantify the impact the supplemental instruction section has on student success for individuals and for at-risk subgroups.

We note that though our illustration is specifically for this supplemental instruction section, the analytics work up may be used generally to study any intervention strategy or pedagogical innovation for evaluating outcomes in student success studies.

In the remainder of this section, we first evaluate the ensemble learner for predicting student success. We then step through a series of analyses afforded by the ensemble learner prediction of individualized treatment effects within a student success study. All analyses were performed in the open source statistical software package *R* (R Core Team 2016) environment. The ensemble learner of Algorithm 1 is performed using leave-one-out cross-validation for the validation set ($K_E = 1$) and ten-fold cross-validation for the ensemble training-testing ($K = 10$). The base learners used are linear regression (or logistic regression depending on the outcome), lasso regression, classification and regression trees (CART), bagging, boosting, random forest, naive Bayes, linear discriminant analysis, support vector machines, and k-nearest neighbors. We refer the reader to James et al. (2013) for details on these methods. Ridge regression is used to combine the base learners (step 7 of the Algorithm 1).

Ensemble Learning Performance Evaluation

Table 2 presents the root mean squared error (RMSE) for predicting final exam score (out of 300) and course grade (four-point scale) by the ensemble learner and the individual learners that make up the ensemble. As mentioned in "Introduction", high correlations are a cause for concern as ensemble learners present the greatest gains when combining individual learners that show diversity in predictions. As Table 2 shows, despite the correlated predictions presented in Fig. 1, the ensemble learner out-performs any single learner for both outcomes.

Table 2 also compares the accuracy of the ensemble learner and individual learners in predicting a non-repeatable grade in the course ('C' or better binary response). The classification ensemble was found by averaging the predicted probabilities from each learner and obtaining an 'optimal' threshold of 0.77 using the `OptimalCutpoints` R package (Lopez-Raton et al. 2014) for predicting a binary response. Under this threshold, the ensemble learner has the highest classification success. Figure 2 presents ROC curves for the ensemble learner and the individual learners. Table 2 presents, in the last column, the area under the curve (AUC) for each of these ROC curves. With respect to this ROC comparison, the ensemble learner out-performs the individual learners.

Success of the Intervention

In our study, the average individualized treatment effect for final exam score was 9.3 with a standard error of 1.48. The average individualized treatment effect for final course grade was 0.45 with a standard error of 0.03. These average ITE are both significantly greater than zero ($p < 0.0001$). Enrolling in a supplemental instruction section not only leads to a moderate improvement in final exam score (on the 300 point scale), but it leads to an increase of almost a half a grade point, on average, in

Table 2 Ensemble learning performance with respect to final exam score (out of 300), course grade, and non-repeatable grade ('C' or better grade)

Method	Final exam score	Method	Course grade	Method	'C' or better grade	
	RMSE		RMSE		Accuracy	AUC
Ensemble	45.3	Ensemble	0.887	Ensemble	80.52%	0.82
Random forest	45.5	Random forest	0.893	LASSO	79.94%	0.80
SVM	45.7	LASSO	0.899	SVM	79.07%	0.77
Boosting	45.9	Linear	0.910	LDA	79.09%	0.79
LASSO	46.0	SVM	0.920	Random forest	78.78%	0.79
Linear reg.	46.6	Bagging	0.927	Boosting	78.49%	0.71
K-nearest neighbor	46.7	Boosting	0.928	K-nearest neighbor	78.49%	0.76
Bagging	46.7	K-nearest neighbor	0.934	Bagging	77.91%	0.78
				Naive Bayes	77.03%	0.77

Root mean squared error (RMSE) is the measure of performance for the former two outputs, accuracy and area under the ROC curve (ROC) for the latter output. The ROC curves appear in Fig. 2

final course grade. (The course is graded on a four-point scale where 4.0 = A, 3.0 = B, 2.0 = C, 1.0 = D, and 0.0 = F.)

The ITE may be used, say at the beginning of a course, to flag students that may benefit from the supplemental instruction course. We may characterize these at-risk students through demographic and educational markers. To this end, the ITE were split into two subgroups: the top 25% and a comparison group (centered around 0). These subgroups were then analyzed to identify average characteristics of students that benefited the most and were not affected by the recitation course respectively; see Tables 3 and 4. The inputs on these two tables are self-explanatory for the most part, however a handful require further documentation (see description of Table 1 as well):

Math Level:	highest math class completed (algebra, pre-calculus, calculus, . . .)
Participation:	score on iClicker questions in Week 2
Quiz 0:	beginning of semester algebra assessment
Calc Level:	applied calculus, calculus 1, calculus 2, calculus 3
Learning Community:	specialized dorm
Compact Scholar:	program partnership with a local school district
First-Gen Some College:	no college degrees in the family
Admission Basis:	first-time freshman or transfer student.

Tables 3 and 4 suggest that students that may benefit the most from the supplemental instruction course have weaker educational preparation (significantly lower SAT

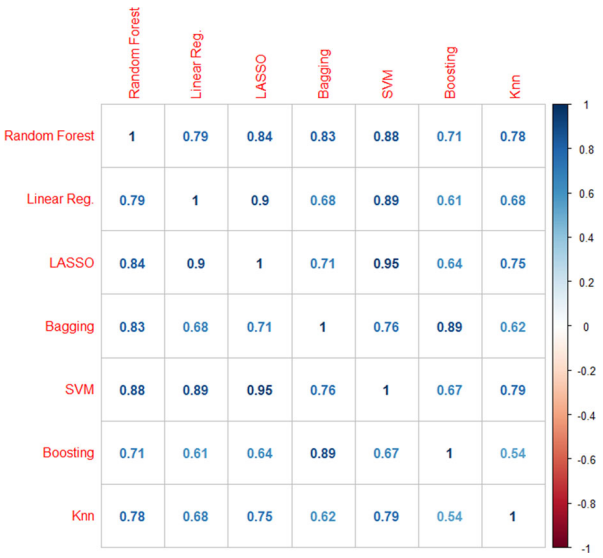


Fig. 1 Correlation matrix plot for individual learners

Math, HS GPA, math level, and previous experience with calculus and statistics), and performed worse at the beginning part of the course (lower clicker score, quiz 0 grade, and performance on homework 1). Furthermore, those students are significantly more likely to be EOP, low income, first-generation, commuter, and/or part-time students. These findings could be considered further as early at-risk indicators for success in the course.

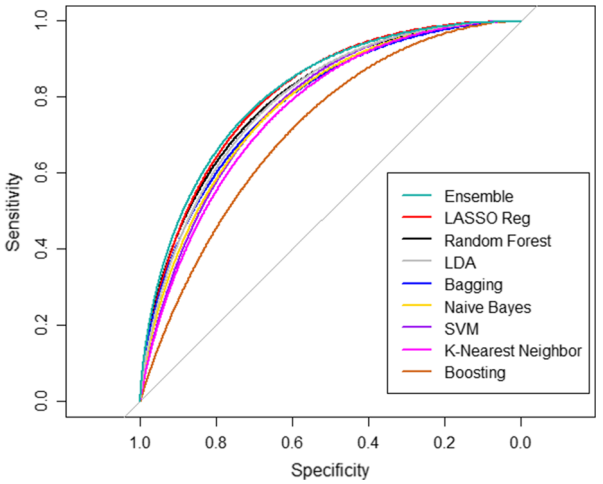


Fig. 2 ROC curves comparing the ensemble learner with each of the individual learners from Table 2

Table 3 *Final exam outcome:* Summaries for students falling in the top 25% in ITE and students falling in a similarly sized comparison group with ITE of zero on average

Continuous inputs	Top 25%	Comp group	p-value
Homework 1, days late	1.60 (7.79)	0.02 (0.19)	0.00
Online units	2.48 (4.24)	1.26 (3.20)	0.00
Calc level	0.32 (0.48)	0.61 (0.73)	0.00
Participation week 2	0.70 (0.45)	0.92 (0.26)	0.00
Quiz 0, grade	0.65 (0.28)	0.83 (0.16)	0.00
SAT Math	516.83 (75.25)	599.67 (78.85)	0.00
Math level	4.48 (1.41)	5.16 (1.57)	0.00
SAT verb	479.83 (99.73)	538.04 (94.94)	0.00
HS GPA	3.3 (0.49)	3.66 (0.35)	0.00
Homework 1, grade	0.90 (0.24)	0.99 (0.04)	0.00
Term units attempted	14.07 (2.16)	14.39 (1.93)	0.09
Homework 1, time in minutes	84.13 (54.39)	78.21 (50.14)	0.22
Quiz 0, time in minutes	28.54 (13.93)	27.24 (9.00)	0.23
Age	19.99 (2.86)	19.13 (1.74)	0.00
HS grad year	2011.48 (2.83)	2012.31 (1.70)	0.00
Final exam	172.3 (55.44)	244.74 (46.61)	
Treatment effect	75.72 (33.37)	0.01 (9.18)	
Categorical inputs	Top 25%	Comp Group	p-value
Compact scholar	12%	3%	0.00
First-generation college	25%	10%	0.00
EOP student	21%	9%	0.00
Part-time student	51%	23%	0.00
Learning community	14%	28%	0.00
Live in dorm	35%	67%	0.00
Stat AP	10%	19%	0.01
Low income	41%	22%	0.00
Pell indicator	37%	20%	0.00
First-generation some college	44%	24%	0.00
Gender (female)	50%	45%	0.27
Location of last Math course			0.00
SDSU	25%	14%	
HS	60%	81%	
TRANS	15%	5%	
Calc AP			0.00
0	83%	65%	
1	16%	28%	
2	1%	7%	
College			0.07
Business	20%	12%	
Sciences	47%	54%	

Table 3 (continued)

Liberal Arts	33%	34%	
Admin basis			0.00
FTF	69%	88%	
LD	3%	5%	
UD	28%	7%	

The p-values are from significance tests between the two groups on each input. The top part of the table considers continuous-valued inputs, presenting the mean value and standard deviation in parentheses for each. The bottom part of the table considers categorical inputs. Except for the multi-category inputs, the features are ordered according to percent difference between the top 25% and comparison groups

Subgroup Analysis

Student success efficacy studies include not only broad-sweeping evaluations of an intervention for students in general, but focus on the impact of the intervention on pre-defined at-risk subgroups. As an example, the STAT 119 class enrolled 32 students from an underrepresented minority (URM) group. For sake of masking, we do not identify the specific URM group. These 32 students displayed ITE significantly greater than zero ($p = 0.007$; the average individualized treatment effect in this group is $+24$ with a standard deviation of 51). The students scored a 205 out of 300 (68%) on the final exam (standard deviation 44).

Table 5 presents characteristics of the 32 students compared to the other 1030 students enrolled in the course. Of note, this URM group contained significantly more EOP, low income, Pell-eligible, transfer, commuter students.

Impact of the Intervention Strategy

The previous two sub-sections considered the impact of the supplemental instruction course on groups of students. The ITE may be used as a form of personalized learning, for each individual student determining if they may benefit from a given intervention strategy. As an illustration, we characterize students who would be predicted to improve their course performance by a letter grade if they had enrolled in the supplemental instruction course. These example students are based on actual students from the STAT 119 class. However, for the sake of confidentiality, we used the STAT 119 students to identify key inputs and then fabricated this group of students for illustration purposes. That said, we also are presenting the summary statistics with a qualifier, rather than exact values, so there is no chance specific students may be identified. All of these students are envisioned to enroll in the hybrid course section.

– *$F \rightarrow C$ student, predicted treatment effect of 130*. Female, Freshman, Pell grant, Kinesiology, commuter student with low SAT math and verbal scores. Has experience with online courses, but no previous statistics or calculus courses. Scored above 85% on quiz 0, and handed HW 1 in on time scoring above 95%. Final exam score of 37%.

Table 4 *Course grade outcome:* Summaries for students falling in the top 25% in ITE and students falling in a similarly sized comparison group with ITE of zero on average

Continuous inputs	Top 25%	Comp group	p-value
Homework 1, days late	0.85 (6.66)	0.22 (1.93)	0.16
Calc level	0.38 (0.53)	0.51 (0.68)	0.02
Online units	1.79 (3.79)	1.51 (3.43)	0.39
Participation week 2	0.76 (0.42)	0.86 (0.34)	0.00
Math level	4.62 (1.45)	5.02 (1.57)	0.00
SAT Math	535.88 (80.41)	567.08 (81.06)	0.00
Quiz 0, grade	0.73 (0.24)	0.77 (0.2)	0.04
Homework 1, time in minutes	81.51 (49.52)	85.6 (59.7)	0.41
Homework 1, grade	0.94 (0.18)	0.97 (0.12)	0.06
SAT verb	502.79 (104.47)	518.12 (97.71)	0.10
HS GPA	3.43 (0.47)	3.53 (0.48)	0.02
Quiz 0, time in minutes	28.94 (11.84)	28.17 (10.14)	0.44
Term units att.	14.07 (2.16)	14.39 (1.93)	0.09
Age	19.61 (2.65)	19.23 (1.84)	0.07
HS grad year	2011.84 (2.63)	2012.21 (1.78)	0.07
Final grade	1.71 (1.23)	3.21 (0.90)	
Treatment effect	1.09 (0.49)	−1.17 (0.37)	
Categorical inputs	**Top 25%**	**Comp Group**	**p-value**
Stat AP	10%	20%	0.01
Learning community	15%	25%	0.01
Compact scholar	9%	6%	0.29
Part-time student	38%	28%	0.03
EOP student	17%	13%	0.37
First-generation college	22%	17%	0.21
First-generation some college	39%	31%	0.08
Live in dorm	48%	59%	0.01
Low income	35%	29%	0.17
Pell indicator	32%	28%	0.42
Gender (female)	48%	51%	0.52
Location of last Math course			0.00
SDSU	20%	12%	
HS	70%	80%	
TRANS	10%	7%	
Calc AP			0.15
0	77%	73%	
1	21%	21%	
2	2%	5%	
College			0.19
Business	16%	12%	
Sciences	56%	52%	

Table 4 (continued)

Liberal Arts	28%	35%	
Admin basis			0.01
FTF	72%	82%	
LD	4%	5%	
UD	24%	13%	

The p-values are from significance tests between the two groups on each input. The top part of the table considers continuous-valued inputs, presenting the mean value and standard deviation in parentheses for each. The bottom part of the table considers categorical inputs. Except for the multi-category inputs, the features are ordered according to percent difference between the top 25% and comparison groups

- *$F \rightarrow C$ student, predicted treatment effect of 111*. Male, URM, Freshman, Finance, commuter student with HS GPA below 3.0. Did not take Statistics nor Calculus, scored below 70% on quiz 0, but scored 97% on HW 1. Final exam score 54%.
- *$C \rightarrow B$ student, predicted treatment effect of 51*. Female, International, first generation, EOP, Freshman Management student living on campus, with HS GPA

Table 5 Average student characteristics for the 32 students from an underrepresented minority group in STAT 119

	URM group	Remainder of class
Gender (female)	53%	48%
EOP	56%	14%
Live in dorm	38%	51%
Age	19.1 (0.9)	19.5 (2.3)
Low income	72%	33%
Pell eligible	69%	31%
Level (Freshman, Soph, Junior, Senior)	65%, 28%, 6%, 0%	74%, 12%, 9%, 5%
Admission basis	47% FTF	92% FTF
First-generation college	19%	18%
Online units	2.25 (2.95)	1.61 (3.48)
Hybrid class	81%	78%
SAT Math	477 (80)	553 (81)
SAT verbal	481 (90)	509 (99)
HS GPA	3.53 (0.32)	3.47 (0.46)
Took calculus? (AP)	34% (28%)	40% (26%)
Took statistics? (AP)	31% (16%)	33% (14%)

Parenthetical values are standard deviations except in the last two rows which report the proportion of students taking AP Calculus and AP Statistics. The admission basis row presents percentage of students admitted as first time freshman (FTF; not transfer students)

below 3.5 and low SAT math and verbal scores. Took AP Calculus; but scored below 55% on quiz 0 and below 75% on HW 1. Final exam score of 62%.

- *C → B student, predicted treatment effect of 45*. Male, first-generation, Freshman, Undeclared, commuter student with HS GPA below 3.2 and no experience with online courses. Scored above 90% on quiz 0 and 100% on HW 1. Final exam score of 61%.
- *C → B student, predicted treatment effect of 39*. Female, URM, Pell grant, Senior, International Security and Conflict Resolution, commuter student with HS GPA below 3.3. No previous statistics nor calculus courses; scored below 50% on quiz 0 and above 90% on HW 1. Final exam score of 73%.
- *B → A- student, predicted treatment effect of 13*. Male, International, first generation, Pell grant, Sophomore, Marketing, commuter student with low SAT verbal score but high SAT math score. No previous statistics nor calculus courses; scored above 80% on quiz 0 and 100% on HW 1. Final exam score of 78%.
- *B → A student, predicted treatment effect of 22*. Female, URM, first generation, EOP, Pell grant, Freshman International Business student living on campus with low SAT math and verbal scores. No previous statistics nor calculus courses; scored above 70% on quiz 0 but did not submit HW 1. Final exam score of 85%.
- *B → A student, predicted treatment effect of 38*. Female, Pell grant, Freshman, Economics student living on campus with solid SAT math and verbal scores. Had calculus, but no previous statistics course; did not take quiz 0 nor submit HW 1. Final exam score of 90%.

Discussion

We propose using an ensemble learning approach to make predictions in student success studies of intervention strategies such as instructional practices and learning environments. In our application evaluating success of a supplemental instruction session in a large enrollment introductory statistics course, we found that the ensemble learner out-performed the base learners in classifying a repeatable grade (C- or worse) and predicting final exam score. In this application, base learner predictions were highly correlated, limiting the predictive performance of the ensemble learner. Applications with more diverse predictions will show markedly better performance by the ensemble learner. We also introduced the concept of individualized treatment effect to evaluate an intervention strategy in student subgroups, identify at-risk students that may benefit from the intervention strategy, and quantify the impact of an intervention to advise individual students into that intervention. As part of the illustration, we presented a set of "example students" that provides further insight on characteristics to be considered when developing early warning systems for student success.

The application of our approach found that students enrolling in the supplemental instruction course performed significantly better than students who did not enroll with respect to final exam score and course grade in a large enrollment introductory statistics course. These results align with findings in the SI effectiveness review

article of Dawson et al. (2014). Of particular note, Dawson et al. (2014) summarize a study by Hodges and White (2001) where students were either mandated to attend SI sessions (25% of students) or voluntarily attend SI sessions (25% of students). While both groups of SI attendees performed significantly better with respect to DFW-rate than non-attendees, the group mandated to attend SI performed better than the group attending SI voluntarily. Our implementation required students to enroll in, and regularly attend supplemental instruction sessions. A beginning of semester algebra assessment was also used to strongly encourage students. While not a mandate, our model follows more closely to the mandatory attendance of Hodges and White (2001). Of course none of the SI effectiveness studies provide in-depth subgroup analysis nor individual assessments as we are able to perform through individualized treatment effects.

Limitations

In our application, the ensemble learner presented the best predictive performance across each outcome measure considered. Furthermore, no single learner came out on top across all of the outputs. See Table 2 for performance details. However, the ensemble learner out-performed the best single learner by less than one percentage point in accuracy and 0.02 in ROC AUC for predicting a C or better grade, and less than 0.01 in RMSE for predicting course grade. A user may thus decide to employ a single learner such as LASSO, which performed close to best across the board. From both a computational complexity and interpretable machine learning perspective, LASSO is less computationally expensive (i.e., faster to fit) and allows for the relationship between inputs and the output to be explained through coefficient estimates. That said, given a library of base learners, ensemble learning approaches can be made computationally efficient for producing predictions. Furthermore, as mentioned in "Ensemble Learning Performance Evaluation", applications displaying less correlation between the base learner predictions, and perhaps larger sample size given the cross-validation steps required, will realize stronger ensemble predictions. In this sense, our application may provide a level of worst-case scenario for ensemble learning ITE estimates in education analytics.

The non-randomized treatment assignment in observational studies may lend to selection bias from imbalance between treatment and control subjects relative to an unobserved confounder. If this important confounder is not collected or excluded from modeling, treatment effects will thus not be sufficiently adjusted. Treatment randomization overcomes this challenge by balancing subjects with respect to all variables/characteristics except the treatment assignment. However randomized controlled trials are often not an option in education studies. Model-based adjustments of confounders, as performed by the base learners in this paper, adjust treatment effects for covariates. Ensemble learning approaches, by combining predictions over a set of single learners, may improve predictive performance (Poliker 2006) so that the confounder adjustments are potentially less model-dependent. In a situation where a randomized trial is not an option, no approach can adjust for important, unobserved confounders. This emphasizes the importance of study design in observational studies and pursuit of an approach like ours that is less model-dependent.

In our study, the inputs to the model consisted of all data available in the SDSU student information database. These variables encapsulate student demographics, educational background, academic (particularly mathematics) preparation, student performance metrics, and SDSU program involvement. Though we believe this set of covariates captures the primary suite of confounders, our study does not include direct measures of student attitudes towards statistics (the course topic under study), social and academic behavior, nor student motivation. Such measures would need be self-reported, that is, collected through standardized survey instruments. These student characteristics may be unobserved confounders that may perhaps bias our ITE estimates.

Ensemble learning methods run the risk of trading off interpretability for predictive performance. In many applications, an interpretable machine learning framework is critical to practical use. That said, the flexibility in choice of base learner and meta-learner allows the user to potentially strike a desired balance (see e.g., Otte 2013). We will say more on this point in the next paragraph.

Recommendations for Further Study

In our application, we selected a specific suite of base learners to combine for the ensemble prediction. But of course at that stage of the algorithm the meta-learner needs know only the number of base learners and predictions from each learner. The choice and number of base learners is at the discretion of the user. Choice of meta-learner is also at the discretion of the user. We chose ridge regression for two primary reasons. First, Reid and Grudic (2009) suggest regularized regression in stacked generalization, in fact finding that ridge regression performed best in their experiments. Second, regularized regression provides for an interpretable machine learning framework through an optimal weighting of base learners, with respect to the regression model as a meta-learner.

Poliker (2006) and Moreira et al. (2012) present meta-learner options as part of their surveys of ensemble learning approaches for classification and for regression respectively. We will not present an exhaustive list of alternatives for the meta-learner here, but mention two promising options we are currently pursuing. Merz and Pazzani (1999) suggests applying principal components regression for overcoming multi-collinearity issues in correlated base learner predictions. Friedman and Popescu (2003) proposes an importance sampling learning ensemble (ISLE) for combining base learner predictions. The models are chosen through a Monte Carlo sampling scheme and the model weights are chosen by a regularized regression scheme. Friedman and Popescu (2008) presents ISLE as a unifying ensemble framework by thinking of the base learners as rules derived from the data. The correct decision analysis for combining these rules will improve prediction accuracy and, more importantly, aid interpretation. Akdemir et al. (2013) extends this rule ensembles approach by using soft rules (e.g., converting hard binary decision rules from a decision tree into smooth decision functions via logistic regression).

The study in this paper serves as a first illustration of ensemble learning for estimating individualized treatment effects in student success efficacy studies. Generalizability is a critical component to putting these machine learning approaches into

educational data mining practice. Our current work not only considers alternative implementations for predictive performance improvement in our ensemble learning framework, but testing and evaluating the effectiveness of the methods across a suite of educational data sets.

We find the open source statistical software environment *R* (R Core Team 2016) ideal for our educational data mining tasks. Though we coded our own ensemble learner, we note here that a number of *R* packages exist to perform ensemble learning. The package `Rminer` (Coretz 2015) presents a suite of 14 classification and 15 regression methods. The package `caret` (Kuhn 2008) presents a training/tuning environment for a set of 23 machine learning methods in *R*. We may present an ensemble learning wrapper around the output from these *R* packages. The package `subsemble` (LeDell et al. 2015; Sapp et al. 2014) presents a subset ensemble prediction method on a set of up to 30 machine learning methods. Subsemble is a variant of the Super Learner prediction method of van der Laan et al. (2007), which is implemented in the `H2Oensemble` (LeDell 2015) and `SuperLearner` (Polley et al. 2016) packages.

On the front of student success efficacy studies, course (student) performance, as measured by instructor-created measures of student learning in our study, provides one avenue for evaluating an intervention. Statistical reasoning, student attitudes and beliefs, and student evaluation surveys provide important alternative angles for assessing the effectiveness of an intervention on learning (Gundlach et al. 2015). The Statistics Education field has validated a number of concept inventories and standardized assessment instruments which we plan to incorporate into future studies of reforms in the Statistics classroom.

As a final comment, our proposed application of individualized treatment effects is not limited to course-level analytics problems of the type considered in this paper. ITE may be applied broadly to learning analytics and academic analytics tasks, in the terminology of Long and Siemens (2011). These problems include evaluation of program/department, institutional, and state/national driven intervention strategies for at-risk subgroups. The ITE approach also allows for flexibility in the array of outcomes to assess in these arenas as well, for example program success, (STEM) program retention, time to graduation, graduation rates, and student engagement. As illustrations of learning analytics applications at different scales, we highlight three:

- At a system level, California State University (CSU) recently proposed a "Graduation Success Initiative" (http://graduate.csuprojects.org/), setting graduation rate goals for each of its 23 campuses. To justify funding from the state legislature, the CSU will need to assess treatment effects relative to the success of programs aimed at achieving these goals, with respect to the impact of the initiative on at-risk subgroups, and for a cost/benefit analysis.
- At an institution level, major concerns at Universities across the country are STEM persistence and closing the achievement gap (President's Council of Advisors on Science and Technology (PCAST) 2012). The SDSU Compact Scholar program, briefly mentioned in "Application: Impact of Supplemental Instruction Section on Student Success in Introductory Statistics", represents a program aimed at improving student engagement and graduation rates to this

end in a local school district. Individualized treatment effect estimates are critical to improving program educational practices and evaluating program students relative to graduation success benchmarks and key learning outcomes.

– At a College or Department level, individualized treatment effects are critical for evaluating, for example, new online degree and certificate programs or advising systems and strategies. Again focus is on time to graduation or time to enter major, graduation success, and post-graduation success measures.

Our experience and expertise lies within higher education, university systems. However, we may envision analogous student success studies in public school (K-12) or community college districts, of (online) tutoring systems, or for continuing education and adult education programs. In each of these settings, individualized treatment effects allow us to evaluate and refine initiatives/programs, assess impact on (at-risk) subgroups, and quantify program impact relative to resource demands.

Acknowledgements This research was supported in part by NSF grant 1633130. Josh Beemer was supported by an NSF S-STEM fellowship and as a graduate research assistant in the SDSU office of Analytics Studies and Institutional Research and office of Instructional Technology Services.

References

Akdemir, D., Heslot, N., & Jannink, J.-L. (2013). Soft rule ensembles for supervised learning. Technical report arXiv:1205.4476v3.

Alpaydin, E. (2010). Introduction to machine learning. In *Adaptive computation and machine learning*, 2nd edn. Cambridge: MIT Press.

Beck, J.E., & Mostow, J. (2008). How who should practice: using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In: *International conference on intelligent tutoring systems* (pp. 353–362). Berlin: Springer.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64.

Coretz, P. (2015). Data mining classification and regression methods. https://cran.r-project.org/package=rminer.

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In *5th future business technology conference* (pp. 5–12).

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: a systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research*, *84*, 609–639.

Dorresteijn, J.A.N., Visseren, F.L.J., Ridker, P.M., Wassink, A.M.J., Paynter, N.P., Steyerberg, W.W., van der Graaf, Y., & Cook, N.R. (2011). Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ*, *343*.

Friedman, J.H., & Popescu, B.E. (2003). Importance sampled learning ensembles. Department of Statistics, Stanford University technical report.

Friedman, J.H., & Popescu, B.E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, *2*, 916–954.

Gundlach, E., Richards, K.A.R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education*, *23*(1), 33 pp.

Henrie, C.R., Halverson, L.R., & Graham, C.R. (2015). Measuring student engagement in technology-mediated learning: a review. *Computers & Education*, *90*, 36–53.

Hodges, R., & White, W.G. (2001). Encouraging high-risk student participation in tutoring and supplemental instruction. *Journal of Developmental Education*, *24*(3), 2–43.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

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

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, 529s–535.

Kuh, G.D. (2008). *High-impact educational practices: what they are, who has access to them, and why they matter*. Washington DC: Association of American Colleges and Universities.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*, 1641–1650.

LeDell, E. (2015). H2O ensemble learning. https://cran.r-project.org/package=h2oensemble.

LeDell, E., Sapp, S., & van der Laan, M. (2015), An ensemble method for combining subset-specific algorithm fits. https://cran.r-project.org/package=subsemble.

Long, P., & Siemens, G. (2011). Penetrating the fog: analytics in learning and education. *EDUCAUSE Review*, 31–40.

Lopez-Raton, M., Rodriguez-Alvarez, M.X., Suarez, C.C., & Sampedro, F.G. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, *61*(8), 1–36.

Martin, D.C., & Arendale, D.R. (1993). *Supplemental instruction: improving first-year student success in high-risk courses,* 2nd edn. Columbia: National Resource for the First Year Experience and Students in Transition, University of South Carolina.

Merz, C.J., & Pazzani, M.J. (1999). A principal components approach to combining regression estimates. *Machine Learning*, *36*, 9–32.

Moon, H., Ahn, H., Kodell, R., Baek, S., Lin, C., & Chen, J. (2007). Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine*, 197–207.

Moreira, J.M., Soares, C., Jorge, A.M., & de Sousa, J.F. (2012). Ensemble approaches for regression: a survey. *ACM Computing Surveys*, *45*, 10:1–10:40.

Otte, C. (2013). Safe and interpretable machine learning: a methodological review. In C. Moewes & A. Nürnberger (Eds.) *Computational intelligence in intelligent data analysis* (pp. 111–122).

Pardos, Z.A., Dailey, M., & Heffernan, N. (2011a). Learning what works in ITS from non-traditional randomized controlled trial data. *International Journal of Artificial Intelligence in Education*, *21*(1), 47–63.

Pardos, Z.A., Gowda, S.M., Baker, R.S.J.D., & Heffernan, N.T. (2011b). The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations*, *13*(2).

Poliker, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*, 21–45.

Polley, E., LeDell, E., & van der Laan, M. (2016). Super learner prediction. https://cran.r-project.org/package=SuperLearner.

President's Council of Advisors on Science and Technology (PCAST) (2012). Engage to excel: producing one million additional college graduates with degrees in STEM. https://www.whitehouse.gov/administration/eop/ostp/pcast/docsreports.

R Core Team. (2016). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Reid, S., & Grudic, G. (2009). Regularized linear models in stacked generalization. In J. A. Benediktsson, J. Kittler, & F. Roli (Eds.) *Proceedings of the 8th international workshop on multiple classifier systems* (pp. 112–121).

Sapp, S., van der Laan, M.J., & Canny, J. (2014). Subsemble: an ensemble method for combining subset-specific algorithm fits. *Journal of Applied Statistics*, *41*, 1247–1259.

Savery, J.S. (2006). Overview of PBL: definitions and distinctions. *Interdisciplinary Journal of Problem-based Learning*, *1*(1), 9–20.

van der Laan, M.J., Polley, E.C., & Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6* (online).

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.