JOURNAL OF STATISTICS EDUCATION
2018, VOL. 26, NO. 1, 31–39
https://doi.org/10.1080/10691898.2018.1426400

OPEN ACCESS

Assessing Instructional Modalities: Individualized Treatment Effects for Personalized Learning

Joshua Beemer[a], Kelly Spoon[a], Juanjuan Fan[b], Jeanne Stronach[c], James P. Frazee[d], Andrew J. Bohonak[e], and Richard A. Levine[f]

[a]Computational Sciences Research Center, San Diego State University, San Diego, CA; [b]Department of Mathematics and Statistics, San Diego State University, San Diego, CA; [c]Analytic Studies and Institutional Research, San Diego State University, San Diego, CA; [d]Instructional Technology Services, San Diego State University, San Diego, CA; [e]Department of Biology, San Diego State University, San Diego, CA; [f]Department of Mathematics and Statistics, San Diego State University, San Diego, CA

ABSTRACT
Estimating the efficacy of different instructional modalities, techniques, and interventions is challenging because teaching style covaries with instructor, and the typical student only takes a course once. We introduce the individualized treatment effect (ITE) from analyses of personalized medicine as a means to quantify individual student performance under different instructional modalities or intervention strategies, despite the fact that each student may experience only one "treatment." The ITE is presented within an ensemble machine learning approach to evaluate student performance, identify factors indicative of student success, and estimate persistence. A key element is the use of a priori student information from institutional records. The methods are motivated and illustrated by a comparison of online and standard face-to-face offerings of an upper division applied statistics course that is a curriculum bottleneck at San Diego State University. The ITE allows us to characterize students that benefit from either the online or the traditional offerings. We find that students in the online class performed at least as well as the traditional lecture class on a common final exam. We discuss the general implications of this analytics framework for assessing pedagogical innovations and intervention strategies, identifying and characterizing at-risk students, and optimizing the individualized student learning environment.

KEYWORDS
Educational data mining; Lasso; Learning analytics; Persistence; Random forest; Traditional lecture classroom

1. Introduction

Against the economic backdrop of workforce demands, universities around the country are striving to meet the needs of their students with fewer resources. For example, the California State University System received 20% less direct state support per student in 2016 than it did in 2007–2008 (California State University 2016). One way universities are trying to address student needs with shrinking resources is to offer large enrollment courses in online formats. Quantitative analysis (service) courses in particular have been identified as settings in which online or partially online (hybrid/blended) instructional modalities may be prudent (see, e.g., Tishkovskaya and Lancaster 2012; or Gibbs 2014, in the context of MOOCs).

Efficacy studies are critical in assessing the impact of these so-called web-learning or e-learning environments on student success and persistence, as well as pedagogical innovations and intervention strategies more broadly. The widespread use of these studies also requires an analytics infrastructure, the use of institutional student information, and data from the course learning management system (Long and Siemens 2011). The resulting student success studies may be used to draft strategic plans, design resource allocation strategies, and inform curriculum redesign and pedagogical refinements. The learning analytics infrastructure is thus required for automated analyses and iterative refinement of pedagogical reforms, especially for at-risk student subgroups.

We have two goals in this article. First, we introduce the concept of individualized treatment effects for characterizing at-risk students to the educational data mining literature. Second, we use the proposed approach to analyze student success in an online offering of the first course of a core applied statistics sequence (*t*-tests/chi-square tests of association/regression/ANOVA), comparing it with standard face-to-face offerings taught by the same instructor. To our knowledge, this is the first empirical study in the statistics education literature of an online, upper division applied statistics course (i.e., beyond first-semester introductory statistics).

1.1. Methodological Contribution: Ensemble Learning for Predicting Individualized Treatment Effects

The machine learning tools we propose for our analytics engine are random forests and lasso (James et al. 2013, chaps. 8 and 6,

CONTACT Richard A. Levine rlevine@mail.sdsu.edu Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182.

respectively). A random forest is a collection of classification and regression trees (CART) that will, through a recursive partitioning algorithm, divide students into homogenous groups relative to the student success outcome measure of interest. The lasso fits a standard linear model, but includes a penalty term in the least squares objective function (a form of regularization) to shrink coefficient estimates toward zero. These two methods are used in concert to compare student success across learning environments, identify inputs important in predicting that success, and quantify individualized treatment effects.

The individualized treatment effect (ITE) is a concept from the personalized medicine literature (Dorresteijn et al. 2011) to study the impact of, in our case, an online course offering (treatment) and characterize students benefitting from this instructional modality. Of course, students will receive one of either the treatment (take the online offering of the course) or control (take the standard face-to-face offering of the course). ITEs provide a mechanism for predicting the performance difference between treatment and control for each student. These predictions allow for a form of personalized learning by providing instructors and advisors a measure by which they may orient a student toward a given instructional modality (or more generally, suite of intervention strategies and pedagogical approaches). They also facilitate identification of potential at-risk students by suggesting interventions that would maximize the likelihood that the students achieve the course learning outcomes. In Section 2, we detail our proposed machine learning approach.

1.2. Application: Student Success in an Online, Upper Division Statistics Course

The education literature is flush with discussions and debates on the presentation of quantitative courses, notably elementary statistics and pre-calculus, in an online modality. Means et al. (2010) presented a meta-analysis of 41 online learning studies through 2008 (subject matter included computer science, health care, languages, mathematics/statistics, science, and social science). The meta-analysis was performed on effect sizes comparing online and face-to-face course offerings. The authors found that students performed better, on average, in well-conceived online settings than in standard face-to-face instruction environments in terms of learning outcomes. "Well conceived" implies online courses where instructors focused attention on student engagement either by incorporating aspects of face-to-face instruction or creating instructor-directed or collaborative learning environments, as opposed to independent, self-directed instruction. Mills and Raju (2011) presented a review of online statistics instruction from 1999 to 2009. The article draws similar conclusions that performance in online statistics courses can be at least as good as traditional face-to-face lecture styles, but active discussions and interactions among the students and with the instructor are critical in engaging students and facilitating success.

More recently, discipline-specific studies comparing online statistics instruction with a standard face-to-face lecture style have appeared in the literature. Gundlach et al. (2015) found, in one component of the study, that there is no significant difference in student performance between online and standard (though web augmented) face-to-face offerings of a statistical literacy course primarily for liberal arts and health sciences students. Scherrer (2011) considered an online introductory statistics course for industrial engineering technology and Simmons (2014) for business. Each finds that student performance in the online modality was significantly worse than that in a traditional lecture setting. Lu and Lemonde (2013) found that health science students performed at least as well in an online introductory statistics course as compared to a standard face-to-face offering. However, the study concluded that students deemed "academically weak" based on class assignment grades, performed relatively better in a standard face-to-face lecture format. In an introductory statistics course for public health Masters students, de Jong et al. (2013) found that students in an online class perform at least as well as students in a standard face-to-face offering. Each of these later papers considers asynchronous video lectures for the online offerings. Despite the mixed findings, each then stresses the importance of students staying on task, interacting with the instructor, and putting forth effort on par to those in the traditional lecture classroom. The papers also suggest delving into factors and student characteristics that may lead to success in the online statistics classroom. Relative to the contributions of this article then, individualized student management criteria have not been developed and applied, and the focus has been on the introductory statistics course.

We use our proposed learning analytics framework to study an online offering of the first course of a San Diego State University (SDSU) core applied statistics sequence. This course has been identified as a bottleneck course given the popularity of and need for continuing data analysis education, beyond an elementary statistics course, throughout the applied sciences, social sciences, and business. The online course we consider included synchronous video lectures created by the instructor, though archived for future, repeat viewing throughout the course. Additionally, the course was designed within the SDSU Course Design Institute where student engagement is emphasized for primary consideration. The analyses take advantage of data from the SDSU student information database to characterize students in terms of, for example, previous statistics background and general preparation, student educational level, experience with online courses, and a variety of demographics.

Our machine learning approach allows us to address four study goals: (1) evaluate success of the online course implementation compared to previous, comparable standard face-to-face offerings; (2) identify factors most important to predicting success under the instructional modalities; (3) characterize students that benefit from the online offering; and (4) study persistence in the follow-up, second-semester applied statistics course. All analyses were performed in the open source statistical software package R (R Core Team 2017).

In Section 3, we describe our study data. In Section 4, we present the findings from our student success efficacy study.

2. A Learning Analytics Framework: Ensemble Learning and Individualized Treatment Effects

We perform two primary analytics tasks in our applications: (1) identify important predictors of student success and (2) predict the difference in student performance between two instructional modalities. The machine learning tools we choose to perform these tasks are random forests and lasso (James et al. 2013, Chapters 8 and 6, respectively).

Random forests are specifically oriented to provide variable importance rankings. The variable importance rankings are used to identify important inputs in predicting student success as part of assessing the pedagogical innovation. The inputs in our setting are collected from institutional databases that describe student background and performance at SDSU. The dependent variable of student success is course final exam score as a percentage.

Individualized treatment effects (ITE) are used to quantify individual differences in the outcome with and without treatment, particularly for studies in medicine (see, e.g., Dorresteijn et al. 2011). We are thinking of the online modality as the "treatment" in our study. Students are exposed only to the treatment (online offering) or "control" (standard face-to-face offering). The estimand is then the change in final exam score or class grade from taking the online offering as compared to the standard face-to-face offering of the applied statistics course.

We estimate the ITE for each student as follows. We construct two random forests for the outcome of interest, one for the students in the treatment group and another for the students in the control group. We send the students from the treatment group down the trees in the control group random forest to predict the response if they had in fact been in the control group. The difference between the student's actual outcome under the treatment and the predicted outcome from the control group random forest is the student's individualized treatment effect. We perform an analogous computation for the students in the control group, predicting their response from the treatment group random forest. The individualized treatment effect for this group is the difference between the predicted outcome from the treatment group random forest and the actual outcome under the control.

An analogous analysis may be performed using predictions from a lasso model fit to students in the treatment group and fit to students in the control group. We take an ensemble-learning type approach (Polikar 2006) by averaging the individualized treatment effects output from the random forest and the lasso. We of course may broaden our ensemble to include a suite of machine learners capable of estimating the ITE. However, as we detail in Section 5, in our application, random forest and lasso display the best predictive performance, better even than a more inclusive ensemble learner. Prediction accuracy is measured by randomly selecting a portion of the data to construct the learners (training dataset) and then predicting student performance on the portion of data not used for training (testing dataset); see James et al. (2013, sec. 2.2).

Before delving into the student success study, let us motivate our machine learning approach within the context of our application. The outcome of interest is score on a common course final exam score. In the analysis, we randomly split the data into 70% training, 30% testing. The prediction error analysis based on the test sample finds that in predicting final exam score, the lasso and random forest present root mean squared errors (RMSE) of 0.18 and 0.16, respectively. Random forest out-performs the lasso, each predicting final exam score within 20 percentage points. The scale of this RMSE corresponds to about a letter grade in final exam score. This seemingly moderate RMSE is partly a consequence of the relatively small sample size (57 online section students; 157 traditional section students).

In general, random forest and lasso are potentially complex models for performing statistical inference. Course level analytics within the framework considered in this article may be performed by any number of learners in the data science toolbox including ordinary linear regression (OLS). However, in the study data considered herein, random forest and lasso outperformed not only linear regression, but a suite of individual learners and an ensemble learner in the flavor of Alpaydin (2009, Chapter 17). In particular, compared to random forest and lasso, this ensemble learner displayed a 11% higher RMSE for predicting final exam score. Larger increases were found for other individual learners, including linear regression which displayed a 22% higher RMSE. The random forest and lasso are thus well suited for predicting individualized treatment effects in our student success study.

3. Study Data

In this section, we further motivate and then describe the data from our student success efficacy study: a comparison of online and standard face-to-face offerings of an upper division applied statistics course.

3.1. SDSU Stat 350A: Statistical Methods I

San Diego State University (SDSU) Instructional Technology Services (ITS) offers a Course Design Institute (CDI) whereby, through a competitive process, instructors are chosen to work as a group with ITS personnel to develop an online course. The semester long institute entails weekly meetings during which the instructors are trained in state-of-the-art instructional technology for online course offerings and discuss issues outlined in the California State University Quality Online Learning and Teaching rubric (QOLT), specifically in creating an engaging online learning environment, creating and assessing student learning outcomes, and developing course materials. The instructors are expected to offer a fully online version of a course in their field in the summer following the institute to a class of at least 50 students. The institute thus includes one-on-one sessions and workshops with ITS course design experts as the instructor prepares course materials and experiments with instructional technology tools and universal design concepts.

One goal of the CDI is to create successful, large enrollment online courses to alleviate the impact of bottleneck courses on 4-year graduation rates and student retention, particularly problematic as a consequence of the severe budget crises in recent years. The first course of our core applied statistics sequence, Stat 350A, presents such a bottleneck course. Stat 350A: Statistical Methods I is a junior-level course at SDSU. The pre-requisite is an elementary statistics course covering the basics of statistical inference and design through simple linear regression and correlation. The course is the first semester of a two-semester sequence. The first few weeks in our 15-week course are spent reviewing inferential concepts, ensuring students have a strong foundation in performing, interpreting, and communicating results from hypothesis tests and interval estimation. The course quickly moves into two-sample inference and basic categorical data analysis. By the end of the year, students are familiar with multiple linear regression, experimental design (factorial, block, split-plot, Latin squares, and repeated measures designs) and corresponding ANOVA techniques, including contrasts, and multiple hypothesis testing procedures. The course text is *An Introduction to Statistical Methods and Data Analysis* (Ott and Longnecker 2008, 6th ed.) and the data analysis software package is Minitab.

Stat 350A enrollment has been steadily rising over the last decade, in fact more than tripling in size from the 2007 offering. The course is recognized as a bottleneck for statistics and computer science majors as we are able to offer only one 60-seat section in fall semesters. The summer offering of an online version of Stat 350A was motivated as providing students a second option during the year to take the course and keep them on track to a four-year graduation. At SDSU, summer instructors' salaries are covered completely by tuition, thus requiring only a minimum enrollment for a course offering. This situation is contrary to academic year offerings, where the number of sections of statistics courses offered is limited by statistics faculty teaching load and a very small lecturer budget.

Stat 350A is a required course for our statistics major and minor programs and our computer science major. The course is also popular among students throughout the College of Sciences with a smattering of students from quantitatively-oriented majors around campus (e.g., business, nursing, public health, and sociology).

3.2. Descriptives: About the Audience

In this study, we compared the Summer 2013 premiere offering of an online Stat 350A taught by Professor Juanjuan Fan with four traditional section offerings of the course, also by Professor Fan, in Falls 2007, 2008, 2009, and 2012. The online offering entailed synchronous lectures (up to 100 min) presented through *Blackboard Collaborate* four days a week for a six week session. The students were assessed through eight online, multiple choice quizzes as well as two "midterm" exams and a final exam, all online. To prevent cheating, questions were blocked and randomly chosen, question order randomly selected, answers were presented in random order, and students could not go back after answering a question. Though students could hire someone to take the quizzes for them, since the course included a number of these lower-stakes assessments, we suspect this did not occur.

The students in the online offering were thus assessed twice per week. On each test, the exam questions and multiple choice options were randomly ordered. The learning management system also did not allow backtracking: students could not go back to a previous question once an answer was submitted. The standard face-to-face offering entailed two 75-min lectures per week in the classroom. Assessments included homework assignments once per-week turned in during class, two in-class "midterm" exams, and an in-class final exam. The multiple choice portions of the final exam analyzed in this study were common between the offerings. The final exam was administered in a classroom on campus in the standard face-to-face offering. Exam questions and multiple choice options were randomly ordered. All students then had to answer the same set of questions, but adjacent students (front, back, and side) in the exam room did not have identical exam papers. Also, we note that the summer and semester offerings were identical in length (hours) and content.

The Summer 2013 offering enrolled 57 students, while the four traditional sections enrolled a total of 157 students. Table 1 presents inputs from the student information database for this study along with descriptive summaries. Most variables are self-explanatory, though the table caption details on three inputs (pre-major, admission basis, and EOP) which require further clarification. The variable 'Total # online course units' is a proxy for experience with online courses. We also have access to SAT and ACT scores and high school GPA. However, the database was missing upward of 50% of the SAT scores and over 50% of the ACT scores and high school GPA. We thus chose not to include these inputs in the analysis, relying on last statistics course taken and the grade in that course as a measure of not only statistical competency, but also student academic performance. Table 1 also summarizes the percentage of under-represented minorities (URM) in the dataset. We note though that the subsequent analyses use an ethnicity input that categorizes students into finer ethnicity categories beyond URM.

Figure 1presents the distribution of previous statistics courses taken. Stat 119 is the SDSU elementary statistics course for business students. The course enrolls about 50% pre-business majors and otherwise nonscience students primarily using the course as a general math elective. Stat 250 is the SDSU elementary statistics course for scientists. In comparison to Stat 119, Stat 250 includes a data analysis computing component, covers probability distributions and experimental design more deeply, and delves into power analyses. The online course enrolled a slightly larger number of students having taken AP Statistics as their last Statistics

Table 1. Study inputs from the SDSU student information database.

Inputs	Descriptives for face-to-face offering	Descriptives for online offering
Online/Traditional indicator	157 students	57 students
Semester enrolled	Fall 07, 08, 09, 12	Summer 2013
Last Stat course taken	See Figure 1	See Figure 1
Grade in last Stat course	3.2 (0.8)	3.0 (0.9)
SDSU GPA	3.0 (0.6)	3.0 (0.6)
Total GPA	3.1 (0.5)	3.1 (0.5)
Student Level	See Figure 1	See Figure 1
Major	See Figure 1	See Figure 1
Pre-major indicator		
Admission basis (first-time-freshman)	55%	47%
Units attempted that semester	12.5 (3.3)	7.1 (4.1)
Units earned that semester	12.1 (3.3)	6.9 (3.7)
Total units attempted	110.5 (38.4)	114.5 (36.5)
Total units earned	110 (31)	109 (33)
Total # online course units previously	2.3 (median 0)	3.6 (median 3)
Age	24.4 (6.0)	25.1 (6.8)
Gender (female)	30%	30%
Ethnicity	22% URM	28% URM
Low income	37%	0%
Equal opportunity program (EOP)	18%	12%
1st generation college student	13%	14%
County resided		
Previous institution		

NOTE: SDSU places students in a pre-major prior to completing lower division general elective requirements as well as pre-requisites for a given major (pre-major indicator). Admission basis categorizes students as entering SDSU as first-time freshmen (FTF) or transfer students and as California residents, out-of-state, or international students. The CSU equal opportunity program (EOP) is designed to improve access and retention to historically low-income and educationally disadvantaged students. For continuous inputs, the average value is reported with standard deviation in parentheses unless otherwise stated.

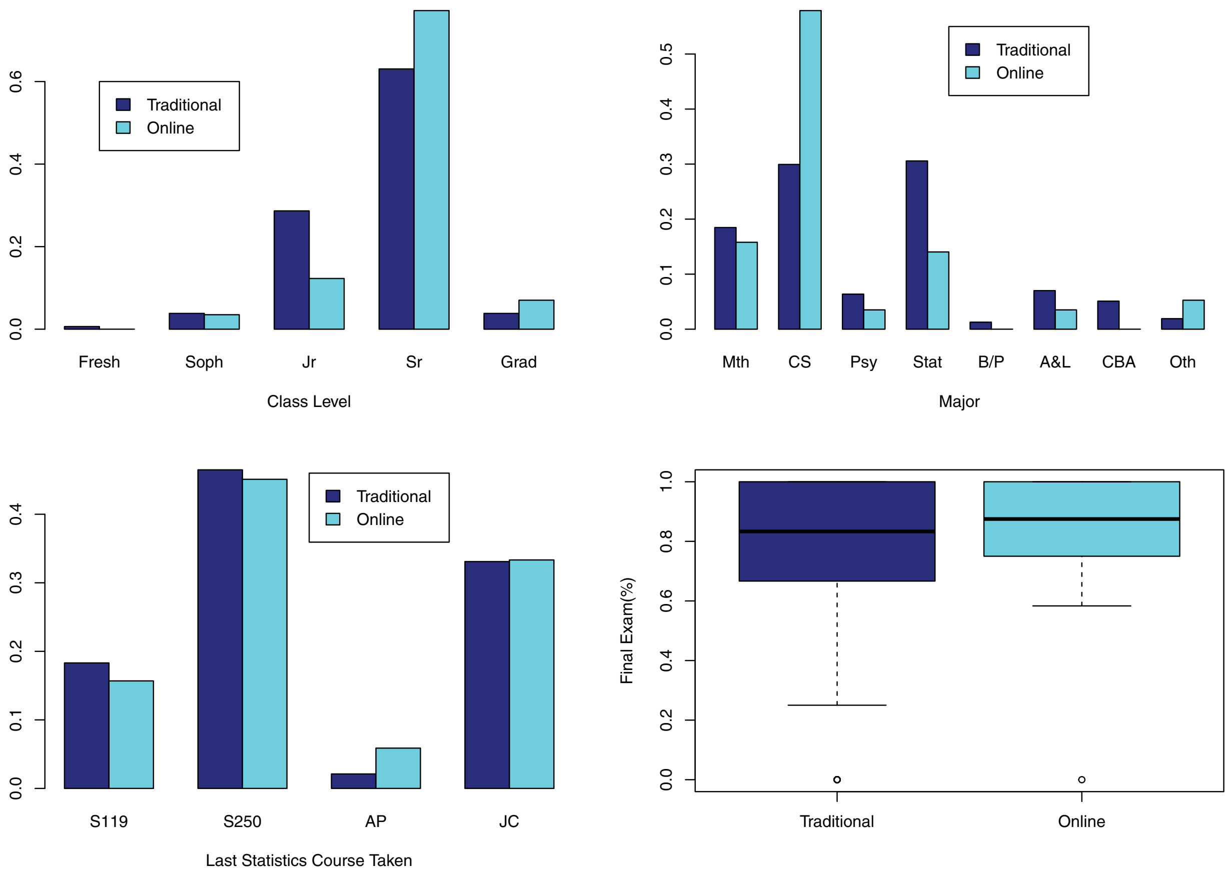

Figure 1. The bar charts present relative frequencies of class level, majors, and previous statistics course taken. The side-by-side boxplot presents final exam scores as a percentage. All graphs are stratified by enrollment in the online or traditional course. The major categories on the second bar chart of the top row are math, computer science (CS), psychology (Psy), statistics (Stat), biology/physics (B/P), arts & letters (A&L), business (CBA), and other. On the bar chart in the bottom row, JC denotes junior or community college.

course, and a correspondingly lower number having enrolled in the SDSU elementary statistics courses. The summer online offering enrolled a larger percentage of computer science majors and seniors. The traditional offerings enrolled a larger percentage of statistics majors.

Although the sections were taught by the same instructor, we are concerned about grade inflation in the course grade in the online course. In particular, the instructor admitted to being more lenient in grading the online class being the first online offering of this course and the instructor's first offering of an online course in her career. The final exam, consisting of common multiple choice questions, allows for a fairer comparison in this respect and thus is our primary outcome measure of student success in this study. Finally, Figure 1 presents the final exam score distribution for the traditional and online sections. The online students seem to perform better, with a larger percentage of scores above 75%, offset by a smaller percentage of scores below 60%. In the following section, we dive deeper into an exploration of this difference.

4. Success Study: Online vs. Standard Face-to-Face Offerings

4.1. *Comparing the Two Modalities*

In this section, we test for a significant difference in student performance between the online and standard face-to-face offerings of Stat 350A. The ultimate goal is to regress final exam outcome on modality controlling for student information database covariates. The number of inputs considered is rather large (see Table 1), complicating a regression model selection process. Random forests take into account interactions between inputs (James et al. 2013, Chap. 8). We thus use the random forest as an initial screening to choose a subset of the most important variables (Hapfelmeier and Ulm 2013). We then perform a traditional AIC-based stepwise model selection procedure on this smaller subset of inputs. The final regression model chosen is used to draw inferences on instructional modality (online vs. standard face-to-face offering).

The student performance metrics (GPA and grade in last Statistics course taken) and educational background measures (total number of units, total number of online units, major, and admission basis, which identifies transfer students and residence status) rose as most important. The final exam score consisted of 14 multiple choice questions from which 12 were counted. A regression on the final exam score found that, after controlling for covariates, students in the online class score 7 percentage points higher (95% confidence interval 1 percentage point to 14 percentage points) on the final exam than students in the traditional sections ($p = 0.03$).

4.2. *Variable Importance Ranking*

In this section, we use the random forest to identify the most important variables in predicting success on the final exam

Table 2. Variable importance ranking (in rank order) for random forest fit of final exam score on inputs for students in the online class (left) and for students in the traditional sections (right).

Online offering		Traditional offering	
Input	Direction of association	Input	Direction of association
GPA	+	GPA	+
Grade in last Stat course	+	Major	
EOP	–	Total # units attempted	–
Admission basis		Total # units earned	+
Major		Grade in last Stat course	+
Total # units attempted	–	Age	–

NOTE: The direction of association identifies either a positive or negative trend for each continuous and binary variable.

amongst the students in the online class and separately among the students in the standard face-to-face class. The random forest procedure ranks variable importance according to percent loss in mean squared error when permuting a given variable in the data (James et al. 2013, Chap. 8.2).

Analogous to the findings in Section 4.1, Table 2 identifies GPA, grade in last Statistics course taken, major, and total number of units as important inputs in predicting final exam score in both instructional modalities. The random forest does not provide us with an effect size on these variables; as we allude to in Section 5, this is an avenue of our current research interests. However, we are able to obtain the direction of association for each important input on the final exam outcome by identifying the trend direction in partial dependence plots (Sec. 8.3 of James et al. 2013). We confirm these trends using a regression model on the variables.

The lasso does not provide a variable importance ranking per se, but it does provide us a form of variable selection by shrinking regression coefficients to zero. Table 3 presents the lasso coefficient estimates. As the table shows, first generation college students, first time freshmen, and Computer Science majors are less successful on the final exam. GPA, experience with online courses, and age are positively related to success on the final exam. The differences in the inputs used by the random forest and lasso procedures motivates combining predictions from these models for individualized treatment effect estimates later. Interestingly, experience in online units appears in both lasso fits, but not as an important variable in the random forest. On the flip side, grade in last Statistics course appears in both random forest fits but not lasso. Furthermore, in the analysis of the online class subset, the indicator of a first generation college student appears with a nonzero coefficient in the lasso, while the equal opportunity program (EOP) indicator is identified as an important variable in the random forest.

Table 3. Non-zero coefficients from lasso fit of final exam score on inputs for students in the online class (left) and for students in the traditional sections (right).

Online offering		Traditional offering	
Input	Lasso coefficient	Input	Lasso coefficient
GPA	0.10 (0.02)	GPA	0.07 (0.04)
First generation	– 0.04 (0.05)	Term units attempted	0.04 (0.02)
Online units	0.04 (0.02)	Online units	0.05 (0.02)
Major (CS)	– 0.01 (0.04)		
Age	0.02 (0.01)		
First-time freshman	– 0.002 (0.03)		

NOTE: Standard errors in parentheses are obtained from the selectiveInference R package (Tibshirani, et al. 2016).

4.3. Individualized Treatment Effects

In this section, we study individualized treatment effect estimates to compare student performance between the online and standard face-to-face sections of Stat 350A. We also use the ITEs to characterize students benefitting from the online offering.

Table 4 presents the average individualized treatment effects for students enrolled in the online class and for students enrolled in the traditional sections. The ITEs for the online offering are reported as the difference in the observed outcome in the online class minus the predicted outcome if in a traditional section. For example, the random forest predicted ITE in Table 4 reports that students in the online class are predicted to score 14 percentage points worse on average on the final exam if they had enrolled in the traditional section. The ITEs for the standard face-to-face offering are reported as the difference in predicted outcome in the online class minus the observed outcome in a standard section. For example, the lasso predicted ITE in Table 4 reports that students in the standard offering are predicted to score 2 percentage points better on average on the final exam if they had enrolled in the online section.

In Table 5, we present characteristics of the typical student benefitting from the online offering. For this analysis, we separate the top 20% of students in terms of the estimated individualized treatment effect. This group contains students that are predicted to perform better in the online offering (average ITE being 34 percentage points higher on the final exam on average). We also construct a comparison group with the same number of students (20%), but with negative estimated individualized treatment effect. This comparison group contains students that are predicted to perform better in the standard face-to-face offering (average ITE being 14 percentage points higher on the final exam on average). The comparison group has a larger percentage of first time freshman (FTF, 70%), but smaller percentage of upper division transfer students (38%) than the group benefitting from the online offering (44% FTF, 44% upper division transfers). The comparison group also is more than 2 years younger on average and a larger percentage of first generation college students.

4.4. Performance in Follow-Up Stat 350B Course

Stat 350A is the first semester of a two-semester applied statistics sequence. The follow-up course Stat 350B is offered in

Table 4. Average individualized treatment effects for final exam score (percentage points) for students enrolled in the online class and in the standard face-to-face lecture classes.

	Online offering	Standard offering
Random forest	14	1
Lasso	7	2

Table 5. Inputs distinguishing students who benefited from the online offering (ITE Top 20%) in terms of final exam score.

	p-value	ITE Top 20% *n*=43	Comp gp *n*=43
Final Exam			
Term Units Attempted	0.003	9.4	12.2
Term Units Earned	0.005	9.0	11.7
Age	0.11	25.6	23.2
Online Units	0.06	2.2	3.8
First Generation	0.12	30%	49%
Admission Basis	0.04		
First Time Freshman		44%	70%
Upper Division Transfer		44%	38%
Graduate		12%	2%

NOTE: The *p*-values are determined from the appropriate *t*-test comparing the two groups on the given input. The latter two columns present average summaries on each input for the top 20% and comparison groups, respectively.

spring semesters. Stat 350B is a required course in all SDSU Statistics undergraduate programs. Although Stat 350B is not a required course for any other program of study, quantitatively oriented majors, such as business, economics, and psychology, enroll in Stat 350B to round out their statistics training and prepare them for statistical applications in their field.

In the current study cohort, 11 out of 57 students in the online class (19%) and 77 out of 157 students in the traditional sections (49%) completed Stat 350B. We use the same model selection procedure as in Section 4.1 to test the difference between the online and traditional sections in terms of persistence and performance in Stat 350B. We measure persistence by an indicator of whether a student continued on to Stat 350B after successfully completing Stat 350A. We measure performance by the student course grade in Stat 350B. The most important predictors of persistence are major, student level, and total units earned. Class type (online vs. traditional) was not significantly related to enrollment. The distribution of majors plays a significant role in this analysis. Recall from Figure 1 that the online class enrolled a larger percentage of computer science majors than the traditional sections. As seen in Figure 2, these students do not tend to enroll in Stat 350B. The small number of online students (11) continuing through the sequence is a consequence of this observation.

Figure 2 presents the Stat 350B grade distribution for students having taken Stat 350B, delineated by enrollment in the online class or traditional sections. Given the small sample of online students in the Stat 350B class, we merely performed a Fisher's exact test of association between class modality (online vs. traditional) and grade; we did not control for any other inputs. We focus on course grade rather than final exam score since the students enrolled in Stat 350B in different semesters/years with different instructors. Thus, a common final exam is not available for this analysis. Class modality is not significantly related to Stat 350B course grade ($p = 0.49$).

An analysis of persistence, and even performance, in Stat 350B in this cohort is challenging given that students enrolling in the online course come from majors that typically do not complete the sequence. Additionally, the Spring 2014 Stat 350B instructor was different than the instructor of the online Stat 350A course. Students in the Stat 350A traditional sections that continued on to Stat 350B had the same instructor for both courses. Nonetheless, we did not find a statistically significant difference in Stat 350B final course grades between students from the online class and students in the traditional sections.

5. Discussion

We present an ensemble learning framework, based on random forests and lasso, as an analytics engine for student success

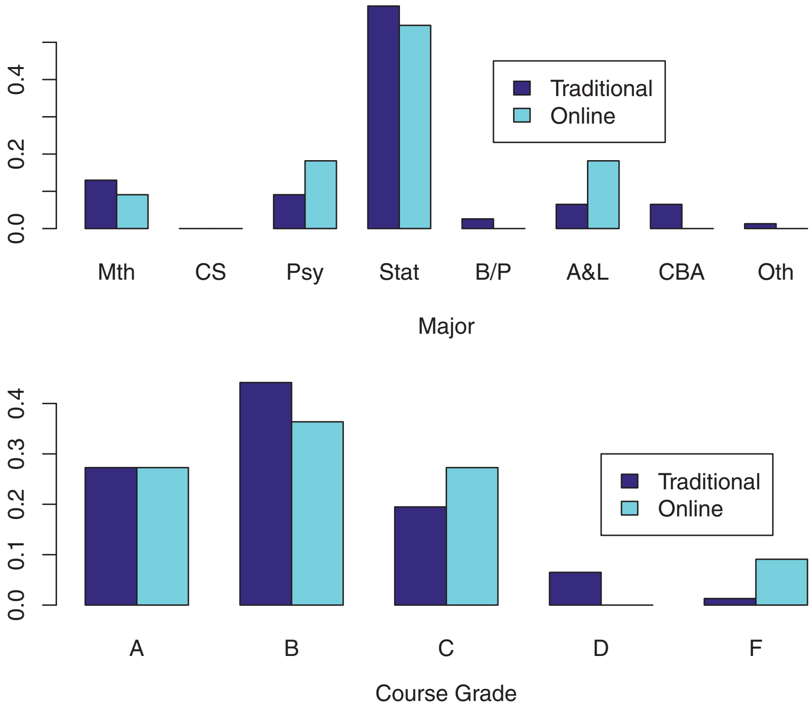

Figure 2. Percentage of majors (top) and grades (bottom) for students from the Stat 350A online class (11 students) and traditional sections (77 students) completing the follow-up course Stat 350B. The bottom graphic displays course grade in Stat 350B.

studies. The method provides an assessment of student performance under pedagogical reforms, identifies factors important in predicting student success, identifies at-risk students, and characterizes students benefitting from the web-based learning environment. As part of the development, we introduce the individualized treatment effect for quantifying student performance under two treatment regiments, where each student is exposed to only one of the treatments.

As an application of the proposed learning analytics approach, we compare online and standard face-to-face offerings of a follow-up to an elementary statistics course: the first course of a two-semester, applied statistics sequence for undergraduates at SDSU (Stat 350A). The study provides further evidence to the literature on the potential success of an engaging, online applied statistics class. In particular, the online class performed at least as well on a common final exam, with the same instructor as the standard face-to-face sections. The most important predictors of success in both modalities were GPA, last statistics course taken, major, and number of University units. We also found that performance in the second semester continuation of the applied statistics sequence, Stat 350B, did not significantly differ between students coming from the online offering of Stat 350A and students coming from a standard face-to-face section.

We do not have measures of statistics concept competencies (CAOS, delMas et al. 2007), student attitudes toward statistics (SATS, Schau 2000), nor student anxiety toward statistics (STARS, Cruise, Cash, and Bolton 1985) as part of the study (see also the analysis of DeVaney 2010 in comparing online and traditional offerings of introductory statistics courses). These instruments were not administered to the students when they were taking the course. Our planned future evaluations of instructional strategies in large enrollment lower division statistics courses, including Stat 350A and the elementary statistics courses for business and science, Stat 119 and Stat 250, respectively, include these instruments as part of the data collection. We also are creating common final exam questions that target specific learning outcomes throughout our lower division statistics curriculum as part of assessment efforts. These questions will focus on statistical concepts, data analysis skills, and statistical communication.

We do not have access to AP scores nor is our data fine enough to identify all college-level statistics courses taken by a student prior to the Stat 350A course of interest. In particular, a student may take an AP Statistics course and then an introductory statistics course at SDSU such as Stat 119. Our dataset would only indicate that the student took Stat 119 as the most recent statistics course taken. As a future endeavor, we are looking into itemizing this variable. It is of interest to investigate differences in students who took an introductory statistics course at a community college or university, between those that also took an AP Statistics course and those that did not.

In the direction of personalized learning, the individualized treatment effect introduced provides a method for characterizing students who benefit from the online course offering, as compared to the standard face-to-face offering. While a natural means of quantifying the impact of a treatment on outcome, we find that ITE predictions in educational data settings are highly variable. Our current research aims to unify the approach within a machine learning context toward improving precision of ITE estimates. Nonetheless, this method allows us to identify clusters of students who may benefit most from, in our study, either an online offering or traditional offering for purposes of advising majors or identifying at-risk students. Interestingly, in our study we found that older students, including transfer students, and students attempting fewer term units showed improved performance in the online modality. These findings may be a function of the greater engagement attained in the online offering. The course was fast-paced with online quizzes every other "class." The instructor presented synchronous online lectures, with online chat and discussion options, and interacted with the students in online office hour discussion rooms and face-to-face office hours. Although students in the standard face-to-face sections were graded on weekly homework assignments, the quizzes in the online class seemed to keep the students more engaged and on top of the material. With that said, an additional caveat worthy of mention is that the summer session audience typically includes more motivated students. In this study, the online class was older (though not a statistically significant difference) and contained a greater percentage of seniors and computer science majors aiming to put this required course behind them in their drive to graduate.

The findings from our particular study may be used to advise students as they choose between an online or standard face-to-face offering of a course. We note as well that the proposed analytics infrastructure is not limited only to efficacy studies of online course offerings. We may apply the method to predict differences under any pedagogical innovation or intervention strategy. These "treatments" may encompass variations in online deliveries (e.g., synchronous or asynchronous; videos produced by instructor, publisher, or third party such as Khan Academy; MOOCs), instructional technology (learning management system, discussion boards, online office hours, etc.), format (online only, hybrid/blended, supplemental instruction, active learning environments, tutoring, etc.), and assessment (online quizzes, online homework, etc.). In fact, under a reasonable sample of training data, instructors and advisors may design individualized learning modules, through a suite of text and lecture material, online and traditional formats, problem sets, and interventions to optimize ITE-based predicted performance for each student in a course.

Funding

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

Alpaydin, E. (2009), *Introduction to Machine Learning* (2nd ed.), Cambridge, MA: MIT Press.

California State University (2013), *2014–2015 Support Budget*, November 2013. Available at: *http://www.calstate.edu/budget/fybudget/2014-2015/executive-summary/documents/2014-15-Support-Budget.pdf*

Cruise, J. R., Cash, R. W., and Bolton, L. D. (1985), "Development and Validation of an Instrument to Measure Statistical Anxiety," in

American Statistical Association Proceedings of the Section on Statistical Education, pp. 92–98.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing Students Conceptual Understanding After a First Course in Statistics," *Statistics Education Research Journal*, 6, 28–58.

de Jong, N., Verstegen, D. M. L., Tan, F. E. S., and O'Connor, S. J. (2013), "A Comparison of Classroom and Online Asynchronous Problem-Based Learning for Students Undertaking Statistics Training as Part of a Public Health Masters Degree," *Advanced in Health Sciences Education*, 18, 245–264.

DeVaney, T. A. (2010), "Anxiety and Attitude of Graduate Students in on-Campus vs. Online Statistics Courses," *Journal of Statistics Education*, 18. Available at: *https://doi.org/10.1080/10691898.2010.11889472*.

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., and Cook, N. R. (2011), "Estimating Treatment Effects for Individual Patients Based on the Results of Randomised Clinical Trials," *BMJ*, 343, d588. Available at: *https://doi.org/10.1136/bmj.d588*.

Gundlach, E., Richards, K. A. R., Nelson, D., and 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. Available at: *https://doi.org/10.1080/10691898.2015.11889723*.

Gibbs, A. L. (2014), "Experiences Teaching an Introductory Statistics MOOC," in *Proceedings of the 9th International Conference on Teaching Statistics* (ICOTS 9), Flagstaff, AZ, USA, July 2014.

Hapfelmeier, A., and Ulm, K. (2013), "A New Variable Selection Approach Using Random Forests," *Computational Statistics and Data Analysis*, 60, 50–69.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, New York: Springer.

Long, P., and Siemens, G. (2011), "Penetrating the Fog Analytics in Learning and Education," *EDUCAUSE Review*, 46, 31–40.

Lu, F., and Lemonde, M. (2013), "A Comparison of Online Versus Face-to-Face Teaching Deliver in Statistics Instruction for Undergraduate Health Science Students," *Advances in Health Sciences Education*, 18, 963–973.

Means, B., Toyama, Y., Murphy, R., Bakia, M., and Jones, K. (2010), *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies*, Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Mills, J. D., and Raju, D. (2011), "Teaching Statistics Online: A Decade's Review of the Literature About What Works," *Journal of Statistics Education*, 19. Available at: *https://doi.org/10.1080/10691898.2011.11889613*.

Ott, R. L., and Longnecker, M. T. (2008), *An Introduction to Statistical Methods and Data Analysis* (6th ed.), New York: Cengage Learning.

Polikar, R. (2006), "Ensemble Based Systems in Decision Making," *IEEE Circuits and Systems Magazine*, 6, 21–45.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, Available at: *http://www.R-project.org/*.

Schau, C. (2000), "Survey of Attitudes Toward Statistics," in *Commissioned Reviews on 250 Psychological Tests*, eds. J. Maltby, C. A. Lewis, and A. Hill, Lampeter, Wales: Edwin Mellen Press, pp. 898–901.

Scherrer, C. R. (2011), "Comparison of an Introductory Level Undergraduate Statistics Course Taught With Traditional, Hybrid, and Online Delivery Methods," *INFORMS Transactions on Education*, 11, 106–110.

Simmons, G. R. (2014), "Business Statistics: A Comparison of Student Performance in Three Learning Modes," *Journal of Education for Business*, 89, 186–195.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620.

Tishkovskaya, S., and Lancaster, G. A. (2012), "Statistical Education in the 21st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform," *Journal of Statistics Education*, 20. Available at: *https://doi.org/10.1080/10691898.2012.11889641*.