The Implementation of Learning Analytics
in Assessing Course Redesigns for College Level Statistics Courses

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Statistics

by

Joshua R. Beemer

Spring 2015

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Joshua R. Beemer:

The Implementation of Learning Analytics

in Assessing Course Redesigns for College Level Statistics Courses

Richard Levine, Chair
Department of Mathematics and Statistics

Barbara Ann Bailey

Barbara Bailey
Department of Mathematics and Statistics

Andrew Bohonak
Department of Biology

5/7/2015

Approval Date

DEDICATION

Dedicated to my friends, family, and all those who have helped me through life.

ABSTRACT OF THE THESIS

The Implementation of Learning Analytics
in Assessing Course Redesigns for College Level Statistics Courses
by
Joshua R. Beemer
Master of Science in Statistics
San Diego State University, 2015

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 in two learning analytics problems: 1) comparing an online and standard face-to-face offerings of an upper division applied statistics course that is a curriculum bottleneck at San Diego State University; 2) evaluating a new supplementary instruction component to a large enrollment introductory statistics course recognized as presenting an undesirably high repeatable grade rate. The ITE in particular allows us also to characterize students that benefit from pedagogical innovations (e.g., online or traditional course offerings) and intervention strategies (e.g., supplemental instruction). We discuss the general implications of this analytics framework for assessing pedagogical innovations and interventions strategies, identifying and characterizing at-risk students, and optimizing the individualized student learning environment.

TABLE OF CONTENTS

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

ACKNOWLEDGMENTS

I would like to thank Dr. Levine, Dr. Fan, Dr. Bailey, Dr. Bohonak, Jeanne Stronach, and the Analtyics Studies and Institutional Research department for providing the necessary tools and knowledge needed to bring this thesis to fruition.

CHAPTER 1
INTRODUCTION

In this thesis, we consider ensemble learners for assessing pedagogical innovations and intervention strategies and, within these methods, introduce the concept of individualized treatment effects (ITEs) to characterize students succeeding in a course. We use our techniques to evaluate the impact of learning modalities on student success in two bottleneck statistics courses.

The first chapter presents a success study of an online applied statistics course, Stat 350A, that is normally offered in a standard lecture setting. The ensemble learning method and ITEs are used to assess student performance in the online and standard courses, identify factors important in predicting student success, characterize students benefitting from the online mode of instruction, assess student performance in the follow-up course Stat 350B.

The second chapter presents a success study of a voluntary supplemental instruction component recently introduced to a large enrollment introductory statistics course Stat 119. The recitation course provides student with extra time and exposure to course practice problems and materials in a smaller, active learning environment. The ensemble learning method and ITEs are used again to assess student performance after enrolling in the supplemental instruction course and characterizing students benefitting from that course.

Overall, the goal of these analyses is to identify students that are ‘at-risk’, predicted to make a repeatable grade (C- or below), and then ultimately provide such students optimal options to improve their chances of achieving the course learning outcomes and making a C or better in the courses. As part of the methods development in each study, we explore the efficacy of ensemble learning approaches. In particular, we illustrate advantages of the ensemble learner over individual learners. We also identify aspects where the ensemble learner may be improved, particularly for educational data mining problems.

CHAPTER 2
ASSESSING INSTRUCTIONAL MODALITIES: INDIVIDUALIZED TREATMENT EFFECTS FOR PERSONALIZED LEARNING

2.1 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 30% less direct state support per student in 2013-2014 than it did in 2007-2008 (California State University, 2013). 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 for example 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 sub-groups.

In this paper, we propose an ensemble learning approach as an analytics engine for student success studies. In particular, we introduce the concept of individualized treatment effects for characterizing at-risk students to the educational data mining literature. We apply this approach to analysis of 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. This success study serves as an illustration of our proposed analytics approach, encouraging readers to consider their own specific application within this framework. 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).

The machine learning tools we propose for our analytics engine are random forests and lasso (James et al., 2013, Chapters 8 and 6 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 towards zero. Depending on choice of the penalty tuning parameter, the lasso thus serves as a regression estimation and model selection routine. 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, via an ensemble learning approach. These predictions allow for a form of personalized learning by providing instructors and advisors a measure by which they may orient a student towards 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.

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) presents 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) presents a review of online statistics instruction from 1999 to 2009. The paper 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) finds, 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) considers 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) finds 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) finds 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 paper then, individualized student management criteria have not been developed and applied and the focus has been on the introductory statistics course.

To illustrate our proposed learning analytics framework, we present results from a study of 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. There are four goals to the study: 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.

2.2 A Learning Analytics Framework: Ensemble learning and individualized treatment effects

The machine learning tools forming the engine for the proposed analytics framework are random forests and lasso (James et al., 2013, Chapters 8 and 6 respectively). The random forest is a collection of decision trees, constructed off of a bootstrap sample of the original data set (Breiman, 2001). CART (classification and regression trees) is at its base, providing a recursive partitioning algorithm to identify binary decision rules to divide observations (students in our case) into groups that are increasingly similar with respect to the response of interest (e.g., final exam score or course grade). In a random forest application, the binary decision rule at the root node and each internal node of the tree is chosen as the best split over a randomly selected subset of inputs. Trees in a random forest are typically not pruned. Each tree in the forest may thus be "sub-optimal" and/or over fit the data. However, ensemble learning over many trees in a forest may substantially improve predictive performance and also allows for a variable importance ranking of the inputs. In this paper we use the `randomForest` package in *R* (2014), which chooses the binary splits by minimizing within-node impurity (Liaw and Wiener, 2002). The lasso fits a regression model to the full set of inputs but shrinks coefficient estimates towards zero via a penalized least squares criterion (Tibshirani, 1996). The lasso thus performs variable selection and estimation in a single routine based on a regression model.

We choose to use both random forests and lasso as we have found in our context that the lasso has better prediction accuracy. Though we use the lasso to identify important predictors, random forests are specifically oriented to provide variable importance rankings. Additionally, random forests provide a more flexible learning framework, handling non-linear relationships, in particular variable interactions (James et al., 2013, Chapter 8). We thus complement random forest predictions with the lasso, particularly in quantifying individualized treatment effects.

We use the variable importance rankings from a random forest as an initial screening of inputs towards constructing a predictive model of student success. The inputs/independent variables in our setting are collected from institutional databases that describe student background and performance at SDSU. The dependent variable of student success is measured by either course final exam score or course grade. The variable importance rankings are also used to identify important inputs in predicting student success as part of assessing the pedagogical innovation.

Individualized treatment effects (ITE) are used to quantify individual differences in the outcome with and without treatment, particularly for studies in medicine (see for example 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). We may transfer the ITE idea to compute for each student, say, the predicted change in final exam score from taking the online offering as opposed to the standard face-to-face offering of the applied statistics course.

We compute the ITE for each student as follows. We construct two random forests for the outcome of interest (say final exam score in our application), one for the students in the treatment group (online class) and another for the students in the control group (standard face-to-face sections). We send the students from the treatment group down the trees in the control group random forest to predict the response (performance on the final exam) if they had in fact been in the control group (enrolled in a standard face-to-face section). The difference between the student's actual outcome under the treatment (actual final exam score in the online class) and the predicted outcome from the control group random forest (predicted score from the standard face-to-face section 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 (predicted final exam score from the online class random forest) and the actual outcome under the control (actual final exam score in the standard face-to-face section).

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.

2.3 STUDY DATA

2.3.1 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 four-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 as 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*, Sixth Edition by R. L. Ott and M. Longnecker 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 75-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 amongst 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).

2.3.2 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 minutes) 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. The students in the online offering were thus assessed twice per week. The standard face-to-face offering entailed two 75-minute lectures per week in the classroom. Assessments included homework assignments 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 Summer 2013 offering enrolled 57 students, while the four traditional sections enrolled a total of 157 students. Table 2.1 presents inputs from the student information database for this study. Most variables are self-explanatory, though the table caption details two inputs (pre-major and admission status) 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 upwards 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.

In this paragraph, we present summaries in the order of traditional sections and then online sections. The average age was 24.4 (sd 6.0) and 25.1(sd 6.8) years respectively, each modality having 30% of enrollees female. Fifty-five percent of the enrollees in the traditional sections entered SDSU as first-time freshman, 47% in the online class. The median number of term units attempted was 12 units (4 courses) and 6 units (2 courses) respectively. Both offerings had students presenting an average SDSU GPA of 3.0 (sd 0.6). The average grade in

Table 2.1. Study inputs from the SDSU student information database. 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 status categorizes students as entering SDSU as first-time freshmen or transfer students and as California residents, out-of-state, or international students.

Course indicators	Data on Units
Online/Traditional	Units attempted that semester
Semester enrolled	Units earned that semester
Last Stat course taken	Total units attempted
Grade in last Stat course	Total units earned
	Total # online course units previously

University-level data	Demographics
SDSU GPA	Age
Total GPA	Gender
Student level	Ethnicity
Major	Low income
Pre-major indicator	Pell grant
Admission status	EOP
	1st generation college student
	County resided
	Previous institution

previous statistics course was 3.2 (sd 0.8) and 3.0 (sd 0.9) respectively. The average number of online units was 2.3 units (median 0 units) and 3.6 units (median 3 units).

Figure 2.1 presents 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 non-science students primarily using the course as a general math elective. Stat 250 is the SDSU elementary statistics 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, 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. Finally, Figure 2.1 presents the final grade distribution for the traditional and online sections. The online students seem to perform better, a larger number of As and Bs, offset by fewer Cs and Ds. In the following section, we dive deeper into an exploration of this difference.

2.4 Success Study: Online vs. standard face-to-face offerings

2.4.1 Student performance

In this section, we test for a significant difference in student performance between the online and standard face-of-face offerings of Stat 350A. We first use the random forest to choose a subset of the most important variables for model selection. We then apply a regression stepwise model selection procedure with AIC as the selection criterion to choose the final model. Not surprisingly, 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 binomial regression on the responses found that students in the online class have a higher probability of success on the final exam than students in the traditional sections ($p = 0.0004$; estimated regression coefficient 0.44 with standard error 0.13). Analogously, a regression on a log-transformed course GPA found that students in the online class displayed a 13% higher course grade than students in the traditional class ($p = 0.02$).

The lasso did not lend itself to this analysis as the procedure does not provide coefficient standard error estimates nor explicit variable rank order of importance. Recent work by Lockhart et al. (2014) towards constructing significance tests for the lasso may nonetheless allow such applications in the near future.

2.4.2 Variable importance ranking

In this section, we use the random forest to identify the most important variables in predicting success on the final exam amongst the students in the online class and separately amongst the students in the standard face-to-face class. 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.

Analogous to the findings in Section 2.4.1, Table 2.2 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 lasso does not provide a variable importance ranking per se, but it does provide us a form of model selection by shrinking regression coefficients to zero. A lasso fit on each subset identified the same set of predictors in Table 2.2 as having non-zero coefficients, with two exceptions. First, the number of online units previously taken appeared with a non-zero coefficient in analyses of both the online and traditional class subsets. Second, in the analysis of the online class subset, the indicator of a first generation college student appeared with a non-zero coefficient, not the equal opportunity program (EOP) indicator.

The random forest does not provide us with an effect size on these variables; as we allude to in Section 2.5, this is an avenue of our current research interests. Nonetheless, the lasso coefficient estimates suggest that first generation college students and transfer students are less successful in the course. Based on these initial findings, both from the random forest and lasso analyses, it may behoove the instructor to focus on pedagogical strategies to aid EOP, first time college, and transfer students in progressing successfully through the online offering.

2.4.3 Individualized treatment effects

Table 2.2. 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
GPA	GPA
Grade in last Stat course	Major
EOP	Total # units attempted & earned
Admission basis	Grade in last Stat course
Major	Age
Total # units attempted	

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.

Before presenting the ITE, we present an evaluation of the predictive performance of the lasso and random forest in this application. An out-of-bag prediction error analysis found that in predicting course GPA, the lasso and random forest present root mean squared error rates of 0.69 and 0.72 respectively. The lasso thus out-performs the random forest, though each predicted course GPA within about 0.7 of a grade point. In predicting the final exam score, the lasso and random forest present root mean squared error rates of 0.18 and 0.16 respectively. Random forest out-performs the lasso, though each predicted final exam score below 20 percentage points. The high error rate of effectively a letter grade in both course GPA and final exam score is a consequence of the relatively small sample size (57 online section students; 157 traditional section students).

Table 2.3 presents the average individualized treatment effects for students enrolled in the online class and for students enrolled in the traditional sections. The students in the online class are predicted to score markedly worse in the traditional sections, in terms of both class grade and final exam score (difference reported is observed outcome in the online class minus the predicted outcome if in a traditional section). The students in the traditional offerings are predicted to score slightly better in the online class (difference reported is the predicted outcome if in the online class minus the observed outcome in a traditional section). The differences though are not statistically significant as indicated by the relatively large standard deviations.

Table 2.3. Average individualized treatment effects for course grade (on a 4-point GPA scale) and final exam score (percentage) for students enrolled in the online class and in the traditional classes. Standard deviation is reported in parentheses.

	Online offering	**Traditional offering**
Random forest		
Course grade	0.67 (0.55)	0.01 (0.76)
Final exam	0.14 (0.13)	0.01 (0.18)
lasso		
Course grade	0.37 (0.55)	0.12 (0.71)
Final exam	0.07 (0.12)	0.02 (0.18)

In Table 2.4, 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. We also construct a comparison group with the same number of students (20%), but with estimated individualized treatment effect of approximately zero. Students with lower GPAs and at a more advanced year (seniors) benefitted more from the online offering. In addition, a larger percentage of students having taken Stat 119 or an elementary statistics course at a community college displayed larger ITE. A smaller percentage of students coming out of an AP Statistics course fell into the top 20% ITE group. Though not shown, the two groups were comparable (44%) in terms of percentage of students having taken Stat 250. In terms only of course grade performance, ITE suggests that Pell grant, upper division, and statistics major students benefit from the online offering.

2.4.4 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 spring semesters. Stat 350B is a required course in

Table 2.4. Variables significantly distinguishing students who benefited from the online offering (ITE Top 20%) in terms of final exam score and course grade.

	p-value	**ITE Top 20%**	**Comp gp**
Final exam			
Total units attempted	0.09	117	106
Total GPA	0.02	3.00	3.18
Campus GPA	0.04	2.92	3.12
Age	0.10	25.5	23.8
Online units	0.002	1.5	3.1
Previous stat course	0.06		
Stat 119		20%	9%
AP Stat		0%	5%
Community College		39%	24%
Average ITE	< 0.0001	0.17	-0.07
Course grade			
Total units attempted	0.003	123	104
Total units earned	0.03	117	102
Total GPA	0.002	2.89	3.17
Campus GPA	0.003	2.75	3.09
Pell	0.08	27%	12%
Upper division student	0.07	77%	56%
Major	0.0006		
Statistics		53%	26%
Computer Science		12%	40%
Average ITE	< 0.0001	0.85	-0.03

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 2.4.1 to test the difference between the online and traditional sections in terms of persistence and performance 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 2.1 that the online class enrolled a larger percentage of computer science majors than the traditional sections. As seen in Figure 2.2, 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 2.2 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 status (online vs. traditional) and grade; we did not control for any other inputs. We focus on class 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 status 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 grades between students from the online class and students in the traditional sections.

2.5 Discussion

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

To illustrate 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 and in course grade, with the same instructor as the standard face-to-face sections. The most important predictors of success 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.

The success study was planned in Fall 2012, after the offerings of each of the standard face-to-face sections. We thus do not have measures of statistics concept competencies (CAOS, delMas et al., 2007), student attitudes towards statistics (SATS, Schau, 2000), nor student anxiety towards statistics (STARS, Cruise et al., 1985) as part of the study (see also the analysis of DeVaney, 2010 in comparing online and traditional offerings of introductory statistics courses). 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.

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, in educational data settings we find ITE predictions are highly variable (see for example, the standard deviations in Table 2.3). Our current research aims to unify the approach within a machine learning context towards 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 students with lower GPA, more University units (i.e., more seniors), and who are more economically challenged benefitted from the online offering. Mature students and statistics majors also 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 proposed analytics infrastructure, an ensemble learning approach leading to predictive models, variable importance rankings and individualized treatment effect estimates, 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.

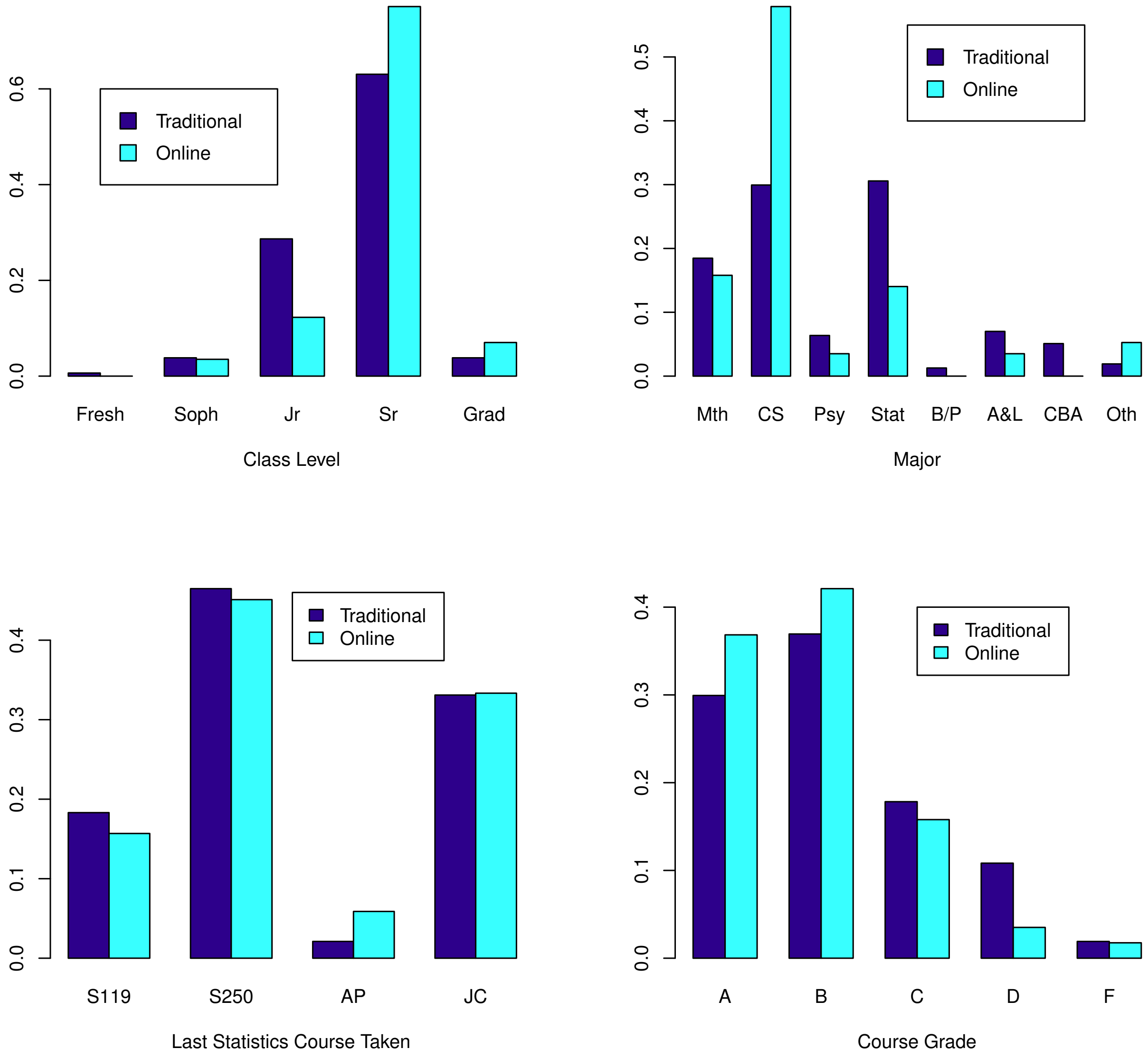


Figure 2.1. From left to right, relative frequencies of class level, majors, previous statistics course taken, and grade in course, stratified by enrollment in online or traditional course. The major categories on the top right bar plot are math, computer science (CS), psychology (Psy), statistics (Stat), biology/physics (B/P), arts & letters (A&L), business (CBA), and other. On the lower left bar plot, JC denotes junior or community college.

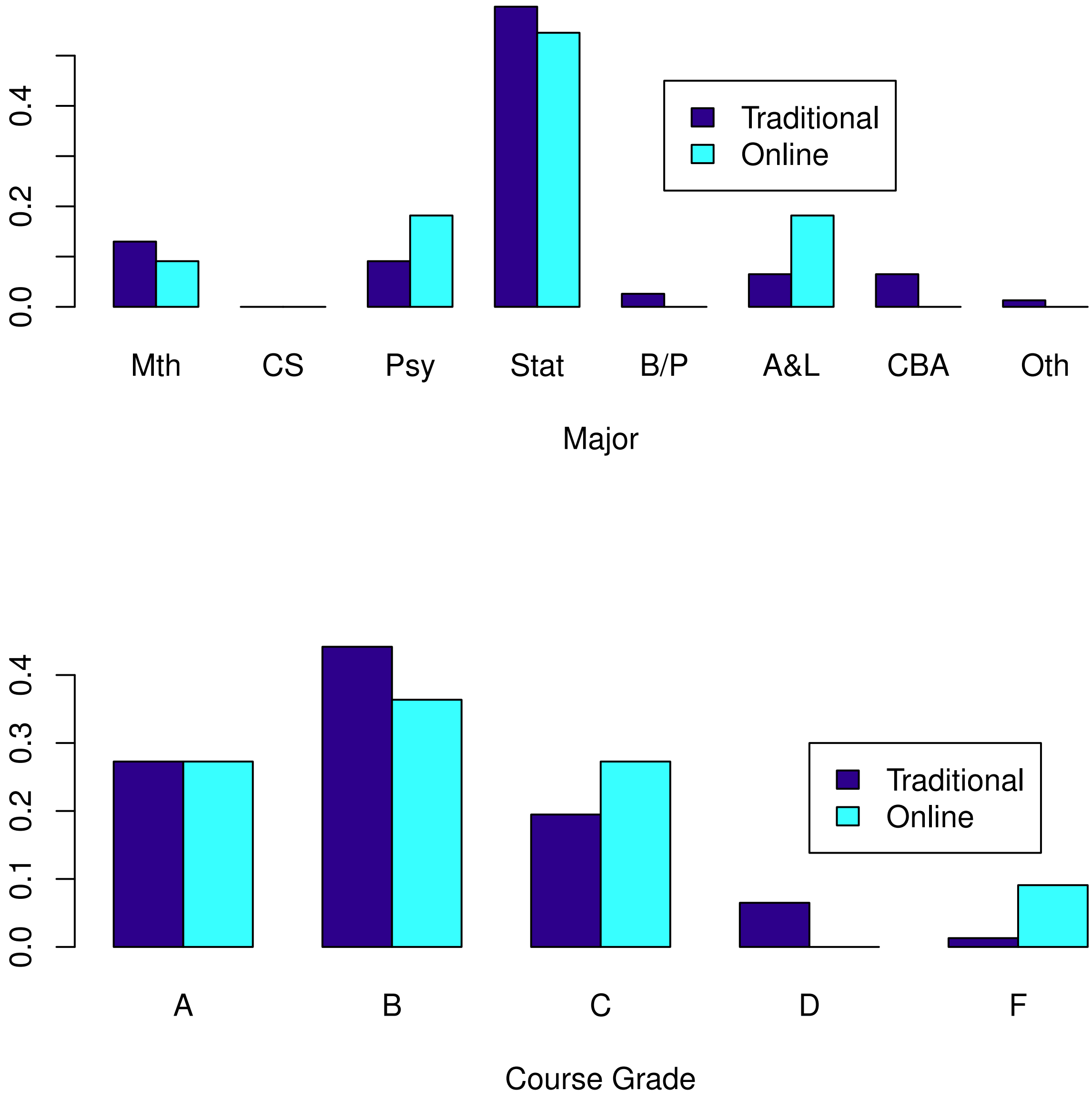


Figure 2.2. Percentage of majors (top) and grades (bottom) for students from the Stat 350A online class (11 students) and traditional sections (77 students) completing the follow-up course Stat 350B.

CHAPTER 3

INTEGRATING ENSEMBLE LEARNING TECHNIQUES FOR ASSESSING A STATISTICS RECITATION COURSE

3.1 INTRODUCTION AND LITERATURE REVIEW

In this chapter, we further explore an expanded ensemble learning approach to estimating individualized treatment effects (ITE). In our previous work (Beemer et al., 2015 and Spoon et al., 2015), we find that ITE estimates are quite variable, leading to potentially imprecise evaluations of the "treatment" (in our situations, e.g., pedagogical innovations or intervention strategies). As discovered in the previous chapter, combining ITE estimates from random forests and the lasso improves the precision of the ITE estimates. Ensemble methods over a wider class of learners has been shown in the literature to greatly improve prediction accuracy. For example, Moon et al. (2007) proves that in the case of classification, an ensemble average over a suite of classifiers will improve accuracy over a single classifier. The case is not so clear cut in a regression context, though for example Moreira et al. (2007) presents a number of approaches to create ensembles to improve prediction accuracy.

A few student success studies take advantage of ensemble learning. Kotsiantis et al. (2010) connects 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) assesses student performance through an application of the data mining tool `Rminer`. The paper compares several ensemble techniques (Random Forest, SVM, etc.) using three responses: binary outcome, multi-class factor, and continuous outcome.

Daum (2014) defines ensemble methods as learning techniques that are combined to improve overall performance, resulting in 'simpler' learning techniques that improve performance. Opitz et al. (1999) explains further that there must be disagreements between classifiers in order to improve classification through ensemble methods. Canuto et al. (2007) expands on the necessity of diversity between classifiers. The work shows that the best set of classifiers would have uncorrelated errors, a combination of which would then diminish the overall prediction error.

Sewell (2011) notes that there is no taxonomy of ensemble methods, and several articles explore different ways of combining learning techniques. These include boosting, bagging, and the mixture of experts methods (Marsland, 2009). Furthermore Alpaydin (2010) investigates seven methods to combine multiple learners, which also includes bagging and boosting, as well as voting, error-correcting output codes, mixtures of experts, stacked generalization, and cascading. Sewell (2011) summarizes chronologically papers that have utilized ensemble learning and is a valuable article for further sources on ensemble learning techniques.

In this paper, we will consider a simple, yet commonly applied version of model averaging presented in Alpaydin (2010). In particular, we identify weights to determine the contribution of a learner to the ensemble. These weights are based on the inverse of the mean square error for each learner. Additionally, an equal weighted average of learner's predicted probabilities was found and an optimal cut point was utilized to produce ensemble predictions. We will illustrate the technique by evaluating a supplementary instruction component in a large enrollment introductory statistics course Stat 119: Elementary Business Statistics.

3.2 Ensemble Learning Methods

In this section each ensemble learning technique and their process for fitting models and classifying data is briefly described. More in depth detail about the techniques and how to utilize them using *R* (R Core Team, 2014) can be found in James et al. (2013). Linear and logistic regression were performed alongside these techniques but are not described below.

3.2.1 Classification/Regression Tree Based Techniques

Classification and regression trees (CART) are introduced in the previous Chapter. In this subsection, we briefly introduce the random forest, namely a collection of CART-grown trees, as well as bagging and boosting. We note that bagging and boosting can be used with alternative learning techniques, not limited to CART applications.

3.2.1.1 BAGGING

Bagging uses classification or regression trees as a foundation for modeling data. Each tree is grown using a randomly chosen subset of the training data, which is obtained via a bootstrap method known as bootstrap aggregation. Prediction is performed by averaging predicted values over all the regression trees or tallying a majority vote to choose a class prediction over all classification trees. Consequently, this process reduces the misclassification rate or variance and improves the prediction error.

3.2.1.2 RANDOM FORESTS

As with bagging, random forests utilize bootstrap aggregation to select subsets of training data to grow individual trees. However rather than scanning over the full set of predictors to determine the decision rule at each node in the tree, a (small) specified number of inputs are chosen randomly. This subset is then used to choose the variable on which the split will be made. This process is continued through the tree growing routine, until the terminal nodes are homogeneous or a minimum number of observations is reached in a terminal node. Unlike CART, random forests do not implement a pruning step.

3.2.1.3 BOOSTING

Unlike bagging and random forests, Boosting does not use bootstrap methods to create subsets of training data. Instead Boosting grows trees on a modified version of the original data. Individual trees are grown sequentially, with each subsequent tree being grown utilizing knowledge from the previous trees. In particular, the trees are fit to the residuals of the model fit instead of the outcome Y. These trees are then added to the model to update the residuals and the process is repeated.

3.2.2 lasso Regression

Least absolute shrinkage and selection operator (lasso) regression is a shrinkage method similar to ridge regression. The residual sum of squares is minimized by using a shrinkage parameter, λ, multiplied by the sum of the absolute value of the coefficients, β_j from $j = 1, \ldots, p$ (Tibshirani, 1996):

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \, .$$

Here, y_i denotes the dependent variable/response/output and x_i denotes the independent variable/input for observations $i = 1, \ldots, n$.

This criterion penalizes the coefficients in the model, which may then present estimated coefficients that are close to or equal to zero. Consequently, the lasso procedure is a form of variable selection in that non-important predictors may be penalized to zero.

3.2.3 Naive Bayes Classifier

Naive Bayes classifier is based on Bayes' Theorem, however the classifier assumes naively that the independent variables are conditionally independent of each other given the dependent variable (Mitchell, 2010). This assumption is generally incorrect but succeeds in making the estimation simpler. Since the class-conditional marginal densities are assumed to be independent, estimation is performed using a simple one-dimensional Gaussian kernel density (Hastie et al., 2013).

3.2.4 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a classification method that makes use of continuous independent variables to classify a categorical dependent variable, whether binary or multiple classes. LDA models each dependent variable class density as multivariate normal. It is then assumed that the covariances are the same among all the classes of the dependent variable, allowing for linearity of the decision boundaries. LDA then fits the linear boundaries that best divide the data into classes, and these decision boundaries are then used to classify the data. Quadratic discriminant analysis provides an extension to LDA when covariances are not presumed the same across classes (Hastie et al., 2013).

3.2.5 Support Vector Machines

Support vector machines are based off of maximal margin classifiers, which utilize linear hyperplanes as decision boundaries. In order to form linear hyperplanes, maximal margin classifiers require the unrealistic assumption that the classes be separable. Support vector machines deal with the constraint by producing a linear boundary in a transformed feature space. The transformation to the new feature space, where the classes are now separable is done, through kernels. There are four kernel types that are typically considered: linear, polynomial, sigmoid, and radial basis functions.

3.2.6 K-Nearest Neighbors

K-Nearest Neighbor analysis is a simple learning technique in which data is classified based on similar observations within a certain distance from itself. For each observation in the test data, a training observation that has the shortest distance away is found, called the k-nearest neighbor. The neighbors are then used to classify that test data observation. The distance between the observations can be calculated using different measures, for example Euclidean, Manhattan, and Minkowski distances. The R function `kknn` uses a Minkowski distance for distances between x_i and x_i' for n pairs of coordinates,

$$\left(\sum_{i=1}^{n} |x_i - x_i'|^p \right)^{1/p} ,$$

to find the k-nearest neighbors. The parameter p has a default value of two in which case the Minkowski distance equals the Euclidean distance. The R function additionally uses kernel functions to weigh the k-nearest neighbors based on their distance from the test observation, and then takes a majority vote for classification.

3.2.7 Ensemble Learning Method

Each technique described in this section has its own advantages and disadvantages, particularly in terms of prediction and classification. Combining the techniques can create an overall better fit, allowing learners that over-fit the data counterbalance those that under-fit the data for example. As mentioned in the introduction of this chapter, we predict by assigning weights to each learner and then either compute a weighted average across the individual learner predictions (continuous response) or a weighted majority vote (binary response) to potentially provide a more accurate set of predictions than that from a single learner.

3.3 Impact of a Recitation Course on Student Success in Introductory Statistics

As part of a course redesign project in Fall 2013, a focus was place on improving DFW rates in Stat 119: Elementary Business Statistics, an introductory statistics course. DFW is defined by a student receiving a D (a repeatable grade), failing, or withdrawing from the course. In the past six years the DFW rate for Stat 119 has hit undesirably high levels of 23% to 32%. To combat the higher DFW rates in the introductory statistics course, a voluntary recitation course Stat 119A was introduced. The recitation course consists of two one-hour classes a week taught by a graduate teaching assistant (GTA). In the recitation course, the GTA reviews and answers students' questions about material covered in the main course. The GTA also leads active problem solving sessions on additional exercises beyond the lecture and homework assignments, giving students more exposure to statistical concepts. Stat 119 in Fall 2013 had 1059 students. Approximately 18%, 196, of Stat 119 students in Fall 2013 enrolled in the recitation course. Table 3.1 summarizes the covariates used in the analysis for the Stat 119 recitation course.

Table 3.1. Summary of institutional data included in analyses as covariates

Course Information
Instructor
Class Format (Traditional or Hybrid)
Participation Week 2
Homework 1 Grade
Homework 1 Time Taken
Homework 1 Late Grade
Quiz 0 Grade
Quiz 0 Time Taken
University-level data
College Description - proxy for major
Admission basis - First-time freshman, transfer
Major status
Student level
Enrollment status - full-time or part-time
Number of online units completed
Term Units Attempted
Institutional Programs
On-campus Housing in Dorms
Learning Community - specialized dorms
Compact for Success - scholarship program
Educational Opportunity Programs
Admissions Information
SAT scores - verbal and quantitative
High School GPA
Previous Math Experience
Highest math class completed - algebra, pre-calculus, calculus...
Location of highest math class taken
Calculus level - Applied calculus, calculus 1, calculus 2, calculus 3
Number of statistics classes taken
Indicator for AP Stats taken
Indicator for AP Calculus taken
Demographic Information
Gender
Age
Ethnicity
First generation college student
Low income indicator
Pell grant

We consider three outcomes in the analysis: final exam score (on a scale of 0 to 300), final grade in the course (on a 4-point GPA scale), and repeatable grade indicator (binary response of whether student received a 'C' or better). These three measures of student success in the course were used to investigate if there is a positive impact when a student participates in the Stat 119A recitation course. Table 3.1 summarizes the covariates used in the analysis for the Stat 119 recitation course. Table 3.1 summarizes the covariates used to analyze the impact Stat 119 recitation course on these three outcomes.

3.3.1 Individualized Treatment Effects

Individualized treatment effects (ITE) are traditionally used in medical studies due to possible variation between patients in reaction to treatments (see for example Dorresteijn et al., 2011). This is comparable to our study where there is conceivable variation in student's reception to the "treatment", namely the recitation course. Individualized treatment effects allow us to identify the variation in final exam score or final course grade from participation participating in the recitation course.

Students are split into control and recitation groups, that is those who had not taken the recitation course and those who had, respectively. Each learning technique was trained using both subsets. Predictions for the recitation group are found by feeding the recitation group subset into the model fits that were trained on the control group, and vice versa for finding predictions for the control group. Namely, for students who enrolled in the recitation course, we are able to predict performance if they had not enrolled, and for students who did not take the recitation course, we are able to predict performance if they had enrolled. These predictions can now be used to find the individualized treatment effects for all students, the difference in performance for each student under treatment (enroll in recitation course) and control (do not enroll in recitation course). For the students in the recitation course their predictions are subtracted from their observed outcomes, and for those not in the recitation course, their observed outcomes are subtracted from their predictions.

3.3.2 ITE results

In our study, the average individualized treatment effect for final exam score was 9.3 with a standard error of 1.48. The average ITE for final course grade was 0.45 with a standard error of 0.03. Both final exam score and final course grade had average ITE significantly greater than zero ($p < 0.0001$). Though the improvement in final exam score from taking the recitation course is not large relative to the 300 point scale of the final exam, taking the recitation course leads to an increase of almost half a grade point, on average, in final course grade.

The ITEs were split into three subgroups: the top 25%, the middle 20% (centered around 0), and the bottom 25%. These subgroups were then analyzed to identify average characteristics of students that benefited the most, were not effected, and were hindered by the recitation course respectively. The individualized treatment effects showed that there was a positive impact from the recitation course on final exam grade and final grade in the course. The subgroups of individualized treatment effects illustrate that students in the top group with the largest treatment effects for both final exam score, Table 1, and final course grade, Table 4, had lower SAT math, SAT verbal, high school GPA, homework 1 grade, and current math level than students in the middle and lower groups. Students in the top group scored, on quiz 0, in the D+ or C- range, 69% to 71%, compared to the middle and low groups, scoring in the C+ to B+, 77% to 80%. Quiz 0 is a beginning of semester assessment of mathematics preparedness for the introductory statistics course. Students who scored lower than 70% on the quiz were strongly advised to enroll in the Stat 119A supplemental recitation course. These findings thus could be considered further as a good indicator for preemptive warning for the student's possible success or lack there of in the course.

Students with the largest treatment effects for final exam score, Table 2, had a significantly higher proportion of students in EOP and first generation students, while having a significantly lower proportion of students who live on-campus (dormitories). Additionally Table 5 shows analogous results for final course grade, with a higher proportion of first generation and EOP students, and a lower proportion of students living on-campus in the top ITE group. The distribution of where a student took their last math course was significantly different for the top and middle ITE groups with respect to final exam score, Table 3, and significantly different for both the middle and bottom groups with respect to the final course grade, Table 6.

Table 3.2. Final exam grade individualized treatment effects mean (sd) for three ITE subgroups and p-values for tests comparing these three ITE subgroups.

	ITE Top 25%	ITE Mid 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Term Units Att.	14.01 (2.16)	14.48 (1.74)	14.4 (2.17)	0.01	0.05	0.66
SAT Math	536.75 (77.63)	560.15 (78.25)	568.58 (85.26)	0.002	<0.0001	0.28
SAT Verb	491.79 (96.49)	516.8 (97.3)	518.5 (104.88)	0.01	0.004	0.86
HS GPA	3.42 (0.48)	3.52 (0.4)	3.53 (0.47)	0.02	0.02	0.82
Math Level	4.76 (1.45)	4.79 (1.5)	4.91 (1.63)	0.80	0.29	0.45
Age	19.66 (2.55)	19.16 (1.96)	19.4 (2.1)	0.02	0.23	0.22
Online Units	1.9 (3.98)	1.04 (2.59)	1.63 (3.48)	0.01	0.43	0.04
Participation Week 2	0.78 (0.41)	0.78 (0.41)	0.84 (0.37)	0.93	0.09	0.13
Homework 1 Grade	0.93 (0.2)	0.97 (0.09)	0.97 (0.11)	0.01	0.005	0.83
Homework 1 Time Taken	84.68 (56.19)	77.4 (46.55)	81.2 (54.95)	0.14	0.49	0.44
Homework 1 Late Grade	1.09 (7.07)	0.08 (0.51)	0.21 (1.92)	0.03	0.07	0.30
Quiz 0 Grade	0.71 (0.24)	0.77 (0.23)	0.8 (0.19)	0.02	<0.0001	0.11
Quiz 0 Time Taken	29.22 (12.34)	27.88 (10.37)	28.1 (10.12)	0.22	0.28	0.83

Table 3.3. Input breakdown for three final exam grade ITE subgroups and p-values for tests comparing these proportions across the three ITE subgroups.

	ITE Top 25%	ITE Middle 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Sex (M)	50%	54%	52%	0.51	0.65	0.89
EOP (1)	18%	10%	12%	0.04	0.10	0.67
Dorm (1)	45%	62%	57%	<0.0001	0.01	0.32
Class Type (Trad.)	22%	24%	22%	0.85	0.98	0.77
Low Income EFC (1)	33%	32%	25%	0.93	0.04	0.09
Pell Indicator (1)	30%	29%	24%	0.88	0.18	0.32
Faculty (Prof. A)	89%	88%	91%	0.86	0.54	0.36
Major Stat (Major)	6%	10%	12%	0.17	0.02	0.46
First Gen NCES (1)	22%	13%	14%	0.01	0.02	0.90
First Gen Some Coll (1)	39%	28%	28%	0.03	0.02	0.97
Learning Community (1)	16%	20%	22%	0.36	0.11	0.63
Compact (1)	10%	6%	5%	0.14	0.06	0.92
Student Level (Lower)	70%	81%	76%	0.01	0.18	0.20
Enroll. Status (First-Time)	66%	78%	71%	0.01	0.24	0.12

Table 3.4. Continued: Input breakdown for three final exam grade ITE subgroups and p-values for tests comparing these proportions across the three ITE subgroups.

	ITE Top 25%	ITE Middle 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Math Location				<0.0001	0.13	0.14
HS	68%	84%	76%			
SDSU	21%	12%	15%			
TRANS	11%	5%	8%			
Stat AP (1)	13%	13%	16%	0.97	0.52	0.57
Calc AP				0.36	0.62	0.63
0	78%	73%	77%			
1	20%	22%	19%			
2	3%	5%	5%			
College				0.86	0.17	0.10
Business	17%	15%	15%			
Sciences	57%	59%	51%			
Liberal Arts	27%	25%	35%			
Admin. Basis				0.52	0.01	0.02
First Time Fresh.	4%	3%	6%			
Lower Div.	73%	77%	82%			
Upper Div.	22%	20%	12%			

Table 3.5. Final course grade individualized treatment effects mean (sd) for three ITE subgroups and p-values for tests comparing these three ITE subgroups.

	ITE Top 25%	ITE Middle 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Term Units Att.	14.04 (2.04)	14.53 (1.97)	14.51 (1.94)	0.011	0.012	0.90
SAT Math	537.5 (83.19)	564.95 (74.96)	571.4 (83.34)	<0.0001	<0.0001	0.41
SAT Verb	491.08 (108.53)	520.62 (96.62)	524.91 (97.84)	0.003	<0.0001	0.66
HS GPA	3.36 (0.5)	3.53 (0.39)	3.59 (0.45)	<0.0001	<0.0001	0.15
Math Level	4.76 (1.46)	4.83 (1.52)	4.93 (1.59)	0.62	0.22	0.50
Age	19.81 (3.07)	19.13 (1.73)	19.2 (1.68)	0.004	0.008	0.69
Online Units	1.53 (3.63)	1.28 (3.16)	1.49 (3.12)	0.43	0.89	0.50
Participation Week 2	0.73 (0.44)	0.81 (0.39)	0.89 (0.31)	0.05	<0.0001	0.03
Homework 1 Grade	0.93 (0.19)	0.97 (0.1)	0.97 (0.1)	0.009	0.004	0.71
Homework 1 Time Taken	79.81 (51.22)	84.68 (62.01)	84.71 (56.19)	0.38	0.33	0.98
Homework 1 Late Grade	1.06 (7.07)	0.18 (1.98)	0.1 (0.86)	0.07	0.04	0.61
Quiz 0 Grade	0.69 (0.26)	0.8 (0.19)	0.8 (0.19)	<0.0001	<0.0001	0.94
Quiz 0 Time Taken	28.4 (13.02)	27.96 (9.69)	28.36 (9.41)	0.68	0.97	0.67

Table 3.6. Input breakdown for three final course grade ITE subgroups and p-values for tests comparing these proportions across the three ITE subgroups.

	ITE Top 25%	ITE Middle 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Sex (M)	55%	54%	49%	0.93	0.20	0.31
EOP (1)	17%	9%	14%	0.03	0.36	0.23
Dorm (1)	48%	63%	56%	0.002	0.15	0.13
Class Type (Trad.)	23%	24%	25%	0.94	0.64	0.81
Low Income EFC (1)	37%	29%	26%	0.11	0.02	0.62
Pell Indicator (1)	33%	27%	25%	0.21	0.07	0.71
Faculty (Prof. A)	89%	89%	88%	0.98	0.88	0.92
Major Stat (Major)	5%	8%	16%	0.44	0.00	0.01
First Gen Nces (1)	22%	13%	14%	0.04	0.06	0.86
First Gen Some Coll (1)	39%	28%	29%	0.02	0.03	0.89
Learning Comm. (1)	18%	23%	25%	0.22	0.07	0.71
Compact (1)	10%	4%	6%	0.02	0.09	0.64
Student Level (Lower)	72%	81%	78%	0.04	0.26	0.40
Enroll Status (First-Time)	72%	78%	71%	0.18	.99	0.18

Table 3.7. Continued: Input breakdown for three final course grade ITE subgroups and p-values for tests comparing these proportions across the three ITE subgroups.

	ITE Top 25%	ITE Middle 20%	ITE Low 25%	Top v. Mid	Top v. Low	Mid v. Low
Math Location				0.004	0.01	0.87
HS	69%	82%	81%			
SDSU	18%	11%	13%			
TRANS	13%	6%	6%			
Stat AP (1)	10%	14%	19%	0.21	0.01	0.25
Calc AP				0.99	0.38	0.50
0	76%	76%	75%			
1	20%	21%	19%			
2	3%	4%	6%			
College				0.11	<0.0001	0.01
Business	20%	14%	12%			
Sciences	59%	58%	46%			
Liberal Arts	20%	27%	42%			
Admin Basis				0.002	<0.0001	0.24
First Time Fresh.	4%	7%	4%			
Lower Div.	69%	79%	85%			
Upper Div.	27%	13%	12%			

3.3.3 Example Students

The ITEs may be used, at the beginning of a course, to flag students that may benefit from the recitation course. For purposes of illustration, we selected a handful of students to identify possible demographic and educational markers that indicate such an “at-risk” student.

Table 3.8. Summary characteristics of example students by response.

	SAT Math	SAT Verbal	HS GPA	Dorm	Low Income	Quiz 0
Final Exam Score	500.7	474	3.32	6.7%	68.8%	77.5%
Final Course Grade	555	436.3	3.13	50%	56.3%	74%

A common theme for students who benefited from the supplemental course were those who did not live in on-campus housing (residential life), had a lower high school GPA (between 2.8 and 3.4), and had lower SAT math and verbal scores. On the other hand, students who had a higher high school GPA (between 3.6 and 4.0), higher SAT math and verbal scores, and had taken Calculus or Calculus AP, but did not participate in the supplemental course would have also benefited from it. The example students had a high percentage classified as low income and averaged lower on quiz 0 than other students.

This latter group may be a function of student confidence. Students who did well in high school and took higher-level math classes were over-confident in their preparation for the course, particularly in terms of math skills. The students thus chose not to enroll in the recitation course and consequently performed below their potential. The former group consists of students who knew they needed the extra help and consequently benefited from participating in the supplemental course.

3.4 Ensemble Learning Performance Evaluation

For analyses with the final exam score and final course grade as responses, combining the different learning techniques by simply averaging over their predictions did not produce a more accurate prediction than the best stand-alone learner. We thus explored taking a weighted average of learners to penalize weaker learners. The weight was created from the inverse of the mean squared prediction error over analysis of a randomly selected test data set not used for training. For this analysis the data was split into a train set, 50%, a test set, 25%, and validation set, 25%.

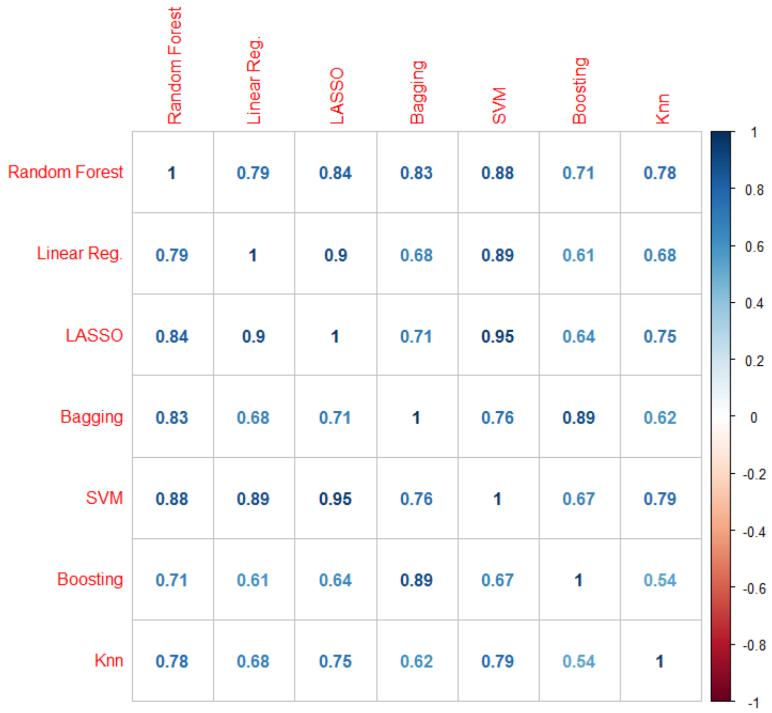
Over repeated random splits of the data into training and testing sets, the weighted average learner out-performed all individual learners for final exam score and final course grade 18% and 20% of the time, respectively. Furthermore, the weighted average learner ranked in the top two 90% of the random splits, and always placed amongst the top three learners. Table 8 illustrates a scenario where the ensemble method outperforms the individual learners for predicting both final exam score and final course grade.

Table 3.9. Ensemble learning performance by MSE as a measure of model accuracy.

Method	Final Exam Score MSE	Method	Final Course Grade MSE
Ensemble	2053.9	Ensemble	0.788
Random Forest	2072.9	Random Forest	0.797
SVM	2091.1	LASSO	0.809
Boosting	2103.6	Linear	0.828
LASSO	2117.9	SVM	0.846
Linear Reg.	2167.2	Bagging	0.860
K-Nearest Neighbor	2175.5	Boosting	0.861
Bagging	2175.6	K-Nearest Neighbor	0.873

It should be noted that a major cause for concern in creating an ensemble learner, as referenced in the introduction from Canuto et al. (2007) and Opitz et al (1999), is the high correlation between different learners (see Figure 1). Bagging and boosting are a a good example of highly correlated learners (correlation coefficient 0.89; not surprising due to the similarity in these learning techniques). This can cause the ensemble method to perform poorly compared to the individual learners.

The analysis on the ‘C’ or better binary response resulted in the classification ensemble having a higher accuracy than any other learning technique. The classification ensemble was found by averaging the predicted probabilities from each learner and then



	Random Forest	Linear Reg.	LASSO	Bagging	SVM	Boosting	Knn
Random Forest	1	0.79	0.84	0.83	0.88	0.71	0.78
Linear Reg.	0.79	1	0.9	0.68	0.89	0.61	0.68
LASSO	0.84	0.9	1	0.71	0.95	0.64	0.75
Bagging	0.83	0.68	0.71	1	0.76	0.89	0.62
SVM	0.88	0.89	0.95	0.76	1	0.67	0.79
Boosting	0.71	0.61	0.64	0.89	0.67	1	0.54
Knn	0.78	0.68	0.75	0.62	0.79	0.54	1

Figure 3.1. Correlation matrix plot for individual learners.

setting a threshold of 0.77, found using the R function `optimal.cutpoints`, for predicting a 0/1 response. This threshold left the ensemble with a lower accuracy, 75.87%, than lasso regression, which was the highest with 82.27%. The ensemble had a specificity and sensitivity of 76.1% and 75.3%, and lasso had a specificity and sensitivity of 89.6% and 69.6%. To optimize accuracy a cutoff point of 0.50 was used instead of 0.77. Using this new cutoff the ensemble presented with the highest accuracy at 80.52%, the lasso regression presenting the second highest accuracy at 79.94%. Though the difference between the ensemble and lasso regression is small, the ensemble does 1.4% to 3.4% better in accuracy than the other learners (see Table 3.10).

The ensemble method additionally has the second highest area under the curve (AUC), behind only the lasso regression. This can be explained by how the lasso regression model performance is optimized compared to the other learning techniques. While random forest, bagging, naive Bayes, etc. are optimized based on accuracy, lasso optimizes its penalty

Table 3.10. Classification accuracy in predicting a grade of ‘C’ or better and AUC for each learner and for the ensemble method.

Method	Accuracy	AUC
Ensemble	80.52%	0.805
LASSO	79.94%	0.8258
SVM	79.07%	0.7797
LDA	79.07%	0.7937
Random Forest	78.78%	0.8013
Boosting	78.49%	0.7143
K-Nearest Neighbor	78.49%	0.726
Bagging	77.91%	0.7184
Niave Bayes	77.03%	0.7798

parameter, λ, which in turn optimizes specificity and sensitivity over accuracy. The ROC curves, Figure 2, illustrate that the ensemble outperformed all other learners.

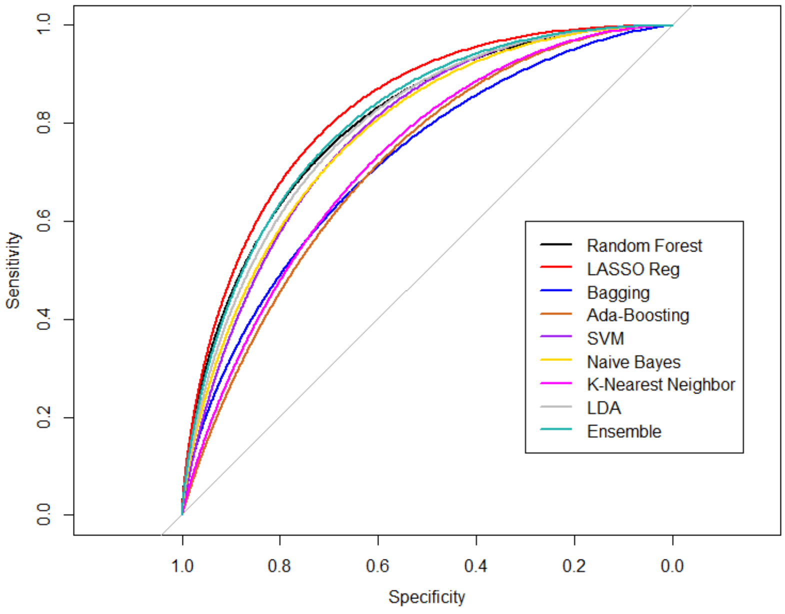


Figure 3.2. ROC curves for each learner and for the ensemble method.

3.5 Discussion

We show that the supplementary instruction component to Stat 119 does help students with weaker academic backgrounds succeed in the course. We are also able to identify characteristics that may be considered when advising students into the voluntary Stat 119A course.

Combining different learning techniques to create an ensemble learner resulted in a higher accuracy for the repeatable grade response, C or better. As for the area under the ROC curve, the lasso was the only technique to beat the ensemble because of how the model is optimized compared to the other learners. Ensemble learning also outperformed the individual learners with the lowest mean squared error, presenting in the top three learners over repeated sampling of training and test data sets. Analyzing student success by means of predicting outcomes using ensemble methods proved to be difficult, but promising. With this information and other supporting papers we would still recommend an ensemble learning approach.

Future explorations in optimizing the ensemble learner method will include stacking and cross-validation approaches to combine individual learners. These approaches are significantly more computationally intensive, but have shown success in the literature. Furthermore, the correlation between individual learners in Figure 1 suggests exploring an ensemble of a subset of learners, rather than the complete set of individual learners. Once the ensemble learning method has been optimized, an *R* package will be constructed to make future learning modalities analysis smoother and readily available for educators.

BIBLIOGRAPHY

[1] Alpaydin, E. (2010), Introduction to Machine Learning, *Adaptive Computation and Machine Learning*, second edition, The MIT Press, Cambridge, MA.

[2] Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5-32.

[3] California State University. (2013). *2014-2015 Support Budget*. November 2013. Retrieved from: http://www.calstate.edu/budget/fybudget/2014-2015/executive-summary/documents/2014-15-Support-Budget.pdf

[4] Canuto, A. M. P., Abreu, M. C. C., de Melo Oliveira, L., Xavier, Jr, J. C. and de M. Santos, A. (2007), Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles, *Pattern Recognition Letters* 28(4), 472486.

[5] Cortez, P., Silva, A. (2008). Using Data Mining To Predict Secondary School Student Performance. Technical Report, University of Minho, Portugal.

[6] 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.

[7] Daum II, H. (2014). Ensemble Methods. *A Course in Machine Learning* (pp. 150-156).

[8] de Jong, N., Verstegen, D. M. L., Tan, F. E. S., 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.

[9] 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(2), 28-58.

[10] DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-Campus vs. online statistics courses. *Journal of Statistics Education* 18 (1).

[11] 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.

[12] 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 (1).

[13] 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.

[14] Hastie, T., Tibshirani, R., Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.).

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

[16] 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*, 529-535.

[17] Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 1822.

[18] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics* 42, 413-468.

[19] Long, P. and Siemens, G. (2011). Penetrating the Fog Analytics in Learning and Education. *EDUCAUSE Review* 46(5).

[20] 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.

[21] Marsland, S. (2009), *Machine Learning: An Algorithmic Perspective*, Chapman & Hall/CRC Machine Learning & Pattern Recognition, CRC Press, Boca Raton.

[22] Means, B., Toyama, Y., Murphy, R., Bakia, M., Jones, K. (2010). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, Washington, D.C.

[23] 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 (2).

[24] Mitchell, T. (2015). Updated Chapter 3. *Machine Learning*. New York: McGraw-Hill.

[25] 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.

[26] 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.

[27] Opitz, D., Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, 11, 169-198.

[28] Ott, R. L. and Longnecker, M. T. (2008). *An Introduction to Statistical Methods and Data Analysis*, Sixth Edition. Cengage Learning, New York.

[29] Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 21-45.

[30] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[31] Schau, C. (2000). Survey of attitudes toward statistics. In J. Maltby, C. A. Lewis, and A. Hill (Eds.), Commissioned Reviews on 250 Psychological Tests (pp. 898-901). Lampeter, Wales: Edwin Mellen Press.

[32] 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.

[33] Sewell, M. (2008). Ensemble learning. RN, 11(02).

[34] Simmons, G. R. (2014). Business Statistics: A comparison of student performance in three learning modes. *Journal of Education for Business* 89, 186-195.

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B 58, 267-288.

[36] 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 (2).