Random Forests for Evaluating Pedagogy and Informing Personalized Learning

Kelly Spoon
Computational Science Research Center
San Diego State University
kellyspoon@gmail.com

John C. Whitmer
Blackboard
john.whitmer@blackboard.com

James P. Frazee
Instructional Technology Services
San Diego State University
jfrazee@mail.sdsu.edu

Andrew J. Bohonak
Department of Biology
San Diego State University
abohonak@mail.sdsu.edu

Joshua Beemer
Computational Science Research Center
San Diego State University
joshbeemer@hotmail.com

Juanjuan Fan
Department of Mathematics and Statistics
San Diego State University
jjfan@mail.sdsu.edu

Jeanne Stronach
Analytic Studies & Institutional Research
San Diego State University
jstronac@mail.sdsu.edu

Richard A. Levine*
Department of Mathematics and Statistics and
Analytic Studies & Institutional Research
San Diego State University
rlevine@mail.sdsu.edu
*Corresponding Author

Random forests are presented as an analytics foundation for educational data mining tasks. The focus is on course- and program-level analytics including evaluating pedagogical approaches and interventions and identifying and characterizing at-risk students. As part of this development, the concept of individualized treatment effects (ITE) is introduced as a method to provide personalized feedback to students. The ITE quantifies the effectiveness of intervention and/or instructional regimes for a particular student based on institutional student information and performance data. The proposed random forest framework and methods are illustrated in the context of a study of the efficacy of a supplemental, weekly, one-unit problem-solving session in a large enrollment, bottleneck introductory statistics course. The analytics tools are used to identify factors for student success, characterize the benefits of a supplemental instruction section, and suggest intervention initiatives for at-risk groups in the course. In particular, we develop an objective criterion to determine which students should be encouraged, at the beginning of the semester, to join a supplemental instruction section.

1. INTRODUCTION

With the ever increasing availability of data on students, sources ranging from student information system databases and warehouses, learning management systems (LMS), and course assessments, we are in a position to perform in-depth statistical analyses of pedagogical innovations and intervention strategies. These studies include evaluations of online and hybrid instructional modalities (e.g., see Means et al., 2010), instructional technologies (e.g., see Rossman and Chance, 2014), or supplemental instruction components (e.g., see Goomas, 2014). We may also identify and characterize so-called at-risk students and provide appropriate interventions towards significantly increasing the odds of success in a course or program. Such assessments are invaluable to instructors, advisors, and administrators as strategic plans, resource allocation strategies, and curricular maps are developed with an eye on student success and retention. For the general educator, a generic learning analytics infrastructure is required to mine the educational data and automate analyses for evaluation of intervention strategies within student success studies. In this paper, we consider random forests as an analytics methodological foundation underlying such an infrastructure. We propose a series of statistical summaries and inferential tools naturally arising from the random forest framework for educational data mining tasks.

A random forest is a collection of classification and regression trees (CART) that uses a recursive partitioning algorithm to divide observations (students in our case) into homogeneous groups relative to an outcome of interest (Breiman, 2001). As an ensemble of individual decision trees, random forests have proven successful in prediction and classification and also present a means for identifying important factors in the prediction process (Chapter 8 of James et al., 2013). We propose to use random forests for the following analytics tasks in student success studies.

- Rank the importance of inputs (student information such as demographics and academic preparation, LMS data, and course assessments) with respect to success in a course or program.
- Use importance ranking as an initial step in regression model building for evaluating pedagogies and interventions.
- Rank the importance of inputs predicting success under pedagogical innovations and intervention strategies.
- Identify at-risk students (before the start of a course or prior to an institutional course drop deadline/census count). In particular,
  - quantify the impact of pedagogical innovations and intervention strategies,
  - characterize students benefitting from pedagogical innovations and intervention strategies,
  - identify at-risk subgroups, and
  - identify successful pedagogical and intervention strategies for subgroups.

We bring the concept of individualized treatment effects (ITE) from the personalized medicine literature (Dorresteijn, 2011) to study the impact of a pedagogical innovation or intervention strategy and characterize students benefitting from such "treatments". As the name suggests, ITEs quantify the difference in an outcome of interest (e.g., final exam score) between treatment and control. The advantage is that, through an ensemble learning approach, the technique provides estimation of that difference *for all* subjects (students), whether they experienced only the treatment or only the control modality. The ITE thus allows us to predict student outcomes under intervention regimes and, through this quantification, characterize students benefitting from the intervention.

Extensive literature reviews of the educational data mining research field have been conducted by Romero and Ventura (2010), Baker and Yacef (2009), and Peña-Ayala (2014). The majority of the contributions to the literature on educational data mining attempt to identify factors associated with student drop out or failure with regards to a particular course, program or school. Dekker et al. (2009) uses multiple data mining techniques, including decision trees and random forests, to predict which Electrical Engineering students will drop out after the first semester of studies before enrolling in the program. Zhang et al. (2010) outlines a university 'knowledge base system' that utilizes data mining methods, specifically naive bayes, support vector machines and decision trees, to predict student dropout. Delen (2010) compares seven different data mining techniques, including decision trees and random forests, to predict which first-year students would register for the following semester. Kotsiantis et al. (2004) compares six machine learning algorithms for predicting whether a student would pass or fail the final exam for a particular class. A common theme across these contributions to the literature is the desire to "accurately predict the at-risk students and hence optimize the allocation of limited resources to retain them" (Delen, 2010) or "identify at-risk students and allow for more timely pedagogical interventions" (Macfayden and Dawson, 2010).

While each of these studies mentions directing the students identified as at-risk to tutoring (Zhang et al., 2010), generic interventions (Delen, 2010; Macfayden and Dawson, 2010), or additional help or attention (Kotsiantis et al., 2004; Dekker et al., 2009), no analyses are performed to whether the intervention will help these students succeed or persist. Furthermore, misclassification of at-risk students, or merely limited definition of such a classification, may result in the intervention being applied to students who will not benefit or not being applied to students who will benefit from it. Superby et al. (2006) offers a possible correction for this phenomenon by instead categorizing students as "low-risk," "medium-risk" and "high-risk". In their study on student dropout using multiple data mining methods, including random forests and decision trees, the medium-risk group was defined as "students, who may succeed thanks to the measures taken by the university". The outcome risk groups in Superby et al. (2006) were created by using grades obtained in the first month of class, simply categorizing students scoring less than 45% as high-risk and students scoring higher than 70% as low-risk. However, no discussion is made about the claim that the medium-risk students will benefit from measures taken by the university or what those measures might be. We propose the ITE approach as a method that directly identifies students who may benefit the most from a particular intervention, thus allowing for an effective allocation of resources.

Many studies compare random forests (for example, Superby et al., 2006; Dekker et al., 2013; Delen, 2010; Kuyoroshade et al., 2013; and Sharabiani et al., 2014) with different data mining tools in terms of prediction accuracy, however little discussion has focused on random forests' useful attributes other than prediction. Kim et al. (2014) uses random forests to compare student success in two types of classes, utilizing variable importance rankings from the random forests to investigate differences in the attributes associated with success in different learning environments. ITEs offer a way to further this work, utilizing the random forests for two (treatment) groups to make predictions and advise students appropriately.

At the institutional scale, Van Barneveld et al. (2012) and Norris and Baer (2013) present a

conceptual framework for what they call analytics in higher education. The papers cite a number of university initiatives to incorporate student success dashboards within learning management systems and reporting tools for University administrators. We consider these applications as excellent first steps for flagging at-risk students (for example, Purdue University Signals, Arnold and Pistilli, 2012, and UMBC Check My Activity, Fritz, 2011), evaluating student behavior relative to historical/comparative norms (for example Arizona State University's eAdvisor System, Phillips, 2013) and descriptive reporting for institutional research departments (for example University of Michigan Tableau driven M-Reports, www.bi.mich.edu). Our proposed random forest analytics approach takes the next significant step in automating efficacy study analyses and informing faculty and institutional stakeholders on pedagogical approaches and intervention strategies that may improve success within a course or program.

To illustrate our proposed random forest learning analytics framework, we present the results of a study of a supplemental instruction component in a large enrollment, bottleneck introductory statistics course at San Diego State University in Fall 2013. In that semester, the supplemental instruction component was a separate 1-unit active problem-solving section in which students could voluntarily enroll. There are three goals to the study. First, we wish to identify factors important for student success to aid the instructors in course redesigns including the instructional approach, assessments, and refinements of this new supplemental instruction component. Second, we evaluate an admittedly ad-hoc criterion, based on a beginning of semester math assessment quiz, for advising at-risk students into the supplemental instruction sections. We also propose a random forest based criterion to identify students with the greatest potential for benefitting from the supplemental instruction section using the complete set of student data available. Third, we apply ITEs to characterize students benefitting from the supplemental instruction section as a means of expanding the program and perhaps identifying alternative interventions to improve the course success rate. The purpose of this illustration is primarily to introduce the random forest for course/program-level analytics and encourage the reader to consider their own specific applications within this framework. Secondarily, the study itself adds an empirical study to the literature on supplemental instruction and learning communities in large enrollment, quantitative methods service courses.

The paper unfolds as follows. In Section 2. we detail the data for our illustrative application of studying the success of a supplemental instruction component in a large enrollment introductory statistics course. In Section 3. we briefly review random forests and proposed analysis tools for predicting student success, including variable importance ranking and model building, then present results from our study data. In Section 4., we present ways to use random forests and variable importance rankings to identify characteristics that may indicate a student will perform better in a particular intervention. And in Section 5., we present the methodology for calculating individualized treatment effects from random forests, simulation results to assess these estimated treatment effects, and results from our study data. Section 6. presents a concluding summary about the random forest analytics methodology proposed and discusses extensions of the approach and institutional strategies arising from our success study.

2. RANDOM FOREST ANALYTICS ILLUSTRATION: STUDY DATA

In Fall 2013, four sections of Stat 119, an introductory business statistics course, were offered at SDSU. This course is a requirement for majors within the College of Business Administration and acts as a statistics service course for other majors such nursing and criminal justice. Two of the sections were offered in a standard face-to-face format, where all lectures took place on campus, as either two 75-minute lectures (Traditional TTh) or three 50-minute lectures (Traditional MWF) per week. The remaining two sections were offered in a "hybrid" format, where one 75-minute lecture was delivered on campus and the other 75-minute lecture was delivered synchronously online, but also archived for later viewing using the Blackboard Collaborate classroom (Hybrid T and Hybrid Th). Hybrid T, Hybrid Th, and Traditional TTh were all taught by the same instructor, a lecturer with over 20 years of experience. Traditional MWF was taught by a graduate teaching assistant (GTA) with 3 years of experience. Both instructors have Masters degrees in Statistics, have extensive experience teaching elementary statistics courses, and employ similar teaching methodologies.

2.1. STAT 119A

Students enrolled in Stat 119 could voluntarily enroll in a 1-unit supplemental instruction section, Stat 119A. Eight of these sections of 35 seats were offered in Fall 2013 with an average enrollment of 25 students. These sections were taught by graduate teaching assistants and met twice per week for 50 minutes each. The material covered in the large lectures was reviewed in the first class meeting of the week and students worked in groups to solve example problems in the second class meeting.

The focus of the analyses herein was to determine the impact of these optional supplemental instruction sections on student success in Stat 119, focusing on which students would have the greatest potential benefit from enrolling in Stat 119A. Since students can add Stat 119A only during the census period (from start of classes to the drop deadline two weeks into the semester), all analyses in this paper focus on the data available prior to census. Such predictions are most prudent in allowing advisors and instructors to make recommendations to students at the beginning of the semester.

Students were encouraged in class and via email by their instructors several times over the first two weeks to join a section of Stat 119A. An initial mathematical preparation assessment, Quiz 0, was given online via MyStatLab in the first week of class. Students scoring less than 70% on this quiz were sent additional emails by their instructor encouraging them to enroll in a Stat 119A supplemental instruction section. As further incentive, students successfully completing Stat 119A were awarded two percentage points towards their course grade. Of the 1059 students enrolled in Stat 119, 196 students also enrolled in Stat 119A.

Table 1 presents summary statistics for students who did and who did not enroll in 119A. A higher percentage of female students and upperclassmen enrolled in Stat 119A. Additionally, the students enrolled in Stat 119A had weaker math skills as measured by their SAT math scores and Quiz 0 score.

As we are primarily interested in analyzing the effect of enrollment in the Stat 119A sections,

Table 1: Demographics of students, stratified by Stat 119A enrollment.

	Enrolled in Stat 119A	Not Enrolled in 119A
Demographics		
% Female**	64.8	45.3
% Freshman**	63.3	76.8
SAT math**	522.2 (83.2)	560.9 (79.1)
HS GPA	3.4 (0.5)	3.5 (0.5)
Previous Math Level	Pre-calculus	Pre-calculus
Quiz 0 Score*	0.76 (0.16)	0.79 (0.14)
Outcome Measures		
Final Exam	200.93 (72.98)	196.98 (74.40)
% Successful	76.0	72.2

* statistically significant difference at 0.05 level, ** at 0.01 level

we excluded from all subsequent analyses in this paper the 27 students who were initially enrolled in 119A but did not receive credit for completing the course.

2.2. Outcome Measures

For this paper, we focused on two different outcome measures: score on the common final exam and an indicator of successful completion of the course.

2.2.1. Final Exam Score

The final exam was scored out of 300 points and consisted of 40 multiple choice problems. Figure 1 shows the overall distribution of final exam scores for all 1059 students who initially enrolled in Stat 119. Of those 1059 students, there were 83 who did not take the final exam and received a 0 score. Eleven of these students were already excluded for not completing Stat 119A and the remaining 72 of these students were removed from the analyses with final exam score as an outcome, leaving us with 960 students in all analyses with final exam as the outcome variable.

The average final exam score for students enrolled in Stat 119A was 218.27 while the average score for students not enrolled in Stat 119A was 214.10; the difference between the means was not statistically significant.

As an outcome measure, the final exam score has two potential limitations. First, the TraditionalMWF class had the second lowest mean score, despite being one of the stronger classes - this can be attributed to the fact that the previous exams in this class were both free response and multiple choice, so the students may not have had the same level of preparation for a final exam that consisted of only multiple choice problems. Additionally, the 73 students who did not sit for the final exam and were not enrolled in Stat 119A are of particular interest in that we would like to know if being enrolled in Stat 119A might have kept them engaged in the course.

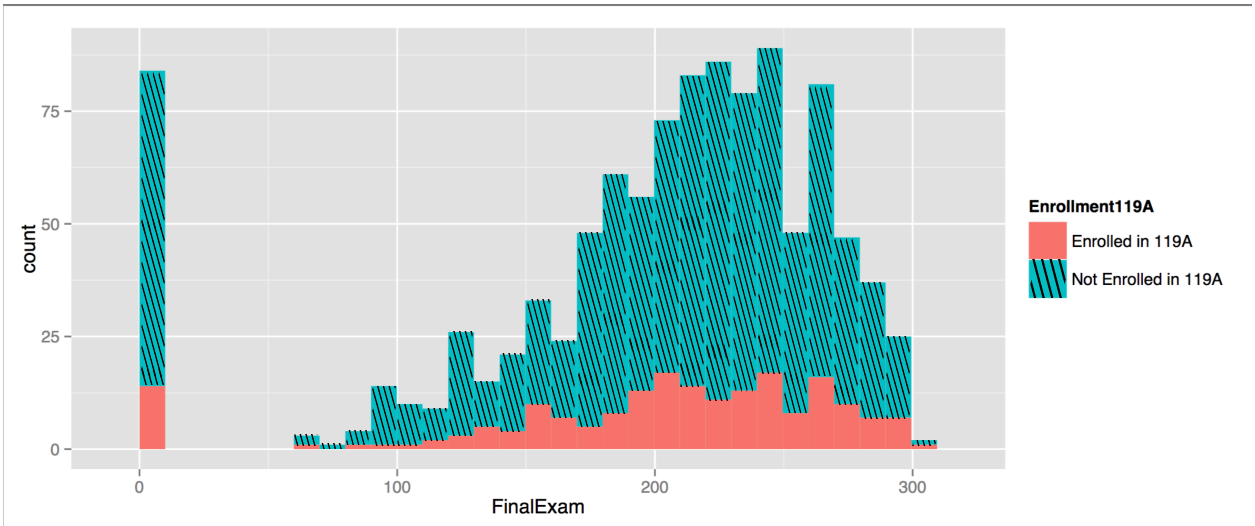

Figure 1: Final exam scores for all 1059 students enrolled in Stat 119, including 0 scores. The rather board and smooth distribution and lack of peak of 100% suggest that the exam has relatively good discriminatory power.

2.2.2. Successful Completion of Course

The second outcome measure, successful completion of the course, was chosen so that individual differences between instructors would be lessened and those students who chose not to sit for the final exam could still be included in the analyses. Successful completion was defined as receiving a C or better in the course. This cutoff aligns with the requirements for most majors serviced by the course. Furthermore, SDSU allows students earning a grade below C to repeat the course and replace the grade. There were 1032 students included in all analyses with successful completion as the outcome measure.

Looking at these 1032 students, there was a statistically significant difference ($p = 0.0032$) between the success rates of students enrolled in 119A (83.4%) and those not enrolled in 119A (72.2%).

2.3. Covariates

The 45 covariates analyzed came from a number of sources: the course gradebook, learning management systems (specifically Blackboard Learn 9.1 and Pearson's MyLabs), and institutional data.

2.3.1. Gradebook Data

During the first two weeks of the semester, students took a 'Quiz 0' to assess mathematical preparation and algebra skills. All four sections also had a participation assignment during week 2: a graded worksheet in TraditionalMWF and clicker questions in the other three sections. Table 2 shows the descriptive measures for these covariates.

Table 2: Descriptive statistics for the covariates available from the instructor gradebooks.

Variable	Mean (sd)	Median	Minimum Value	Maximum Value
Quiz 0 Score	0.76 (0.22)	0.82	0	1
Quiz 0 Time (minutes)	28.32 (10.91)	28.88	0	45.00
Homework 1 Score	0.96 (0.14)	1	0	1
Homework 1 Time (minutes)	82.25 (57.25)	67.95	0	712.18
Homework 1 Date (julian day)	246.91 (6.57)	246	238	343
Homework 1 Late (days after due date)	0.47 (4.08)	0	0	94
Week 2 Participation	0.79 (0.40)	1	0	1

2.3.2. LMS Data

The sections of Blackboard that had statistics tracking enabled were not used within the first two weeks of the semester. Unfortunately, the version of Blackboard used during this semester had no other measures of student engagement.

Time-on-task data for the homework and quizzes was available through the Pearson's MyLabs product. The Quiz 0 and first homework assignment were due within the first two weeks of the semester, so all data for these two assignments was included in the analyses: score, time-on-task, date submitted and number of days late the assignment was submitted.

2.3.3. Institutional Data

Institutional data was collected from the SDSU Student Information Management System (SIMS), the Office of Financial Aid (OFAS), EOP, and the Housing Office. As a summary, Table 3 displays the inputs available and Table 4 displays the descriptive statistics for the quantitative variables.

2.4. Missing Data

Two variables had missing data for 202 students: SAT scores and the previous math course data. Table 5 shows the number of missing observations for the two variable types. The data for these 202 students was imputed using the `mice` package in R (R Core Team, 2013; Van Buuren et al., 2011).

3. Random Forest

In this paper, we explore multiple ways to utilize random forests in a learning analytics setting, emphasizing approaches to identify at-risk students and to determine how to characterize students who would benefit the most from a particular intervention.

Table 3: Summary of institutional data included in analyses as covariates

Course Information
Section Number
Instructor
Class Format (Traditional or Hybrid)
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
Institutional Programs
Honors Program
On-campus Housing in Dorms
Learning Community - specialized dorms
Compact for Success - scholarship program
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 - high school, transfer institution, SDSU
Semester of highest math class taken
Calculus level - applied calculus, calculus 1, calculus 2, calculus 3
Number of statistics classes taken - 0, 1 or 2
Indicator for AP Stats taken
Indicator for AP Calculus taken
Demographic Information
Gender
Age
Ethnicity
Disabled
First generation college student
Low income indicator
Pell grant

Table 4: Descriptive statistics for the quantitative covariates available from institutional data.

Variable	Mean (sd)	Median	Minimum Value	Maximum Value
SAT Verbal	510.62 (100.08)	520	290	800
SAT Quantitative	555.40 (80.49)	560	260	800
High School GPA	3.49 (0.45)	3.56	1.75	4.32
Age	19.40 (2.23)	18.72	17.05	38.35
Math Level	4.79 (1.49)	4	1 (Algebra 1)	8 (Above Calculus 2)

Table 5: Contingency table showing the 202 students with incomplete data by missing variable.

	Missing SAT	Not Missing
Missing Previous Math	28	14
Not Missing	160	830

3.1. Methods: Random Forest for Identifying At-Risk Students

All analyses were performed using random forests in R (R Core Team, 2015; Liaw and Wiener, 2002). A random forest is a collection of classification and regression trees (CART). Classification and regression trees (CART) use binary split rules on the given set of covariates/inputs to divide students into groups that are increasingly similar for the outcome of interest. For an example, Figure 2 shows the best regression tree, from the R tree package `rpart`, predicting final exam score for students in an SDSU elementary business statistics course. The boxes represent the nodes of the tree, in which the average final exam score, number of students (n) in that node, and percentage of students in that node relative to the class as a whole are shown. The top-most node is called the root node. It is characterized by a decision rule, shown underneath the node in the Figure 2 graphic, to split the students into two groups: in this case students with a Quiz 0 (q0) score less than 78% are sent down the tree to the left, students with a Quiz 0 (q0) score greater than or equal to 78% are sent down the tree to the right.

Nodes characterized by an analogous decision rule are called internal nodes. For example, the 365 students sent down the tree to the left of the root node (representing 38% of the class) have an average final exam score of 200 (out of 300). This next internal node splits students according to a high school GPA (hsgpa) less than (sent down the tree to the left) or greater than or equal to (sent down the tree to the right) 3.8. Students progress through the tree, split according to binary decision rules at each internal node. At last, students collect at the bottom of the tree in terminal nodes. The tree in Figure 2 has 12 terminal nodes.

In a random forest, the data (students in the example of Figure 2) are randomly sampled with replacement to create a sample the same size as the data set. A CART-type growing algorithm is then run to construct a tree on this bootstrap sample. In our application in the R package `randomForest`, the decision rules are determined by minimizing the within-node impurity (e.g., mean squared error in the regression tree context or misclassification rate in the classification tree

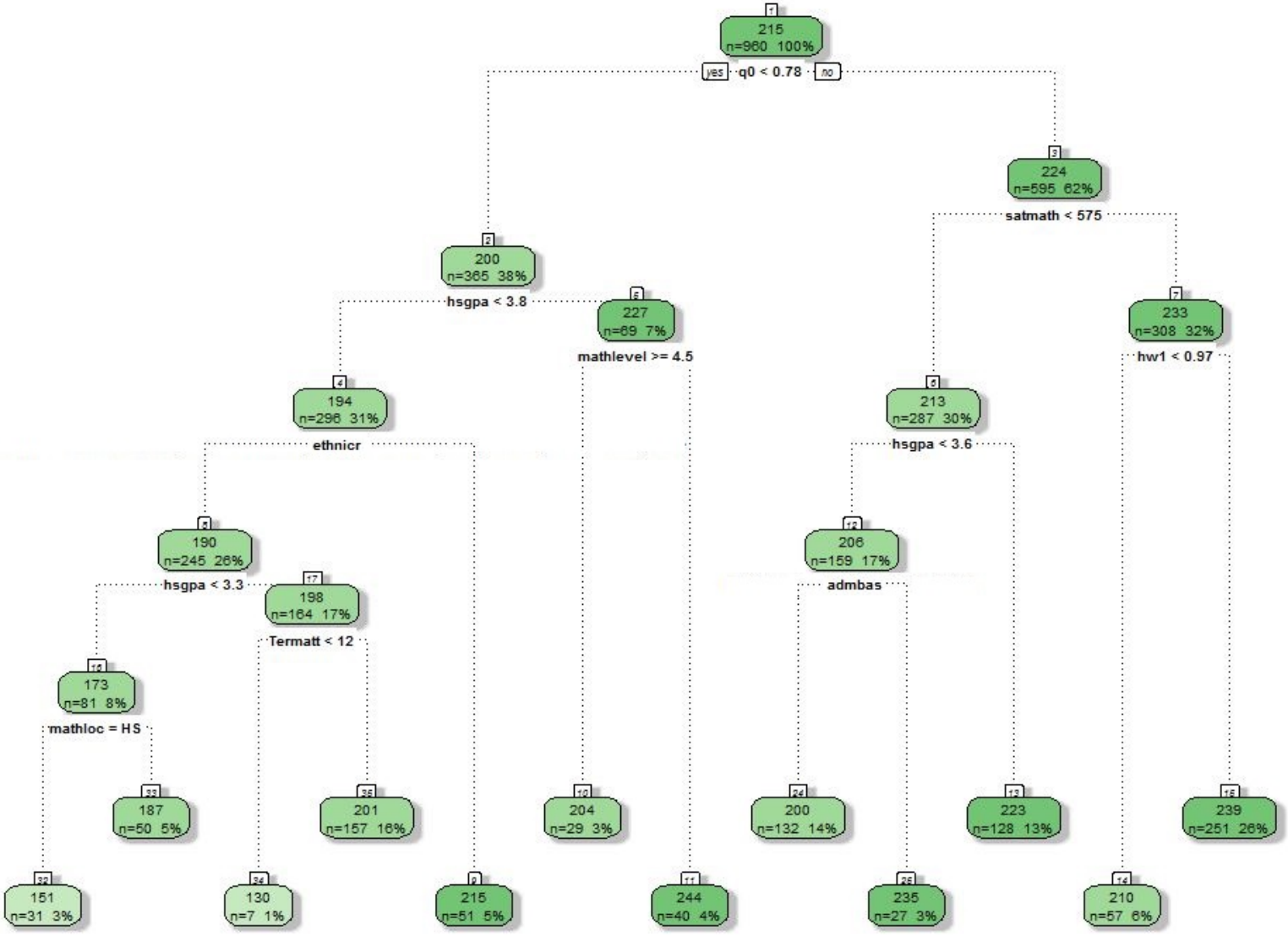

Figure 2: An example of a regression tree created using final exam scores (out of 300) as the outcome measure. A final exam score of 210 (70%) can be thought of as the threshold for passing the exam. Each node (box) presents the number of students, percentage of students, and average final exam scores for the students in that node. Underneath the root and each internal node is the split rule characterizing that node. The ethnicity decision rule (ethnicr) sends categories Asian, Filipino, and International to the left. The admission basis decision rule (admbas) sends U.S. resident first-time-freshman to the left.

context) over all possible split variables (inputs/covariates) and split thresholds for each selected splitting variable. Unlike CART applications for identifying an optimal tree however, at each node a random forest randomly samples a subset of covariates over which to choose the binary decision rule for splitting. Furthermore, a random forest does not implement a pruning algorithm, each tree in the forest grown until the terminal nodes reach a minimum number of observations or are completely homogeneous relative to the covariates and/or outcome of interest. Each tree in a random forest is then different and potentially sub-optimal, but aggregating outcomes over a collection of trees may improve prediction accuracy, provides an estimate of variability, and allows for a ranking of variable importance in predicting the outcome of interest (James et al., 2013).

3.2. Results: Random Forest for Identifying At-Risk Students

In this section, we use random forest to identify students at-risk of failing Stat 119. A student is sent down each tree in the forest, noting the terminal node in which the student is placed. In a classification tree context, the predicted probability of success in a course is presented as the proportion of students in that terminal node that earned a passing grade. In a regression tree context, the predicted final exam score for the student is presented as the average final exam scores of students in that terminal node. By averaging these measures across the trees in the forest, we may estimate the predicted likelihood of success, course grade, or final exam grade, flag at-risk students as having a predicted success probability or final exam score below a given threshold, and direct these students towards interventions.

As an illustration of the ability of a random forest to predict a continuous variable, the random forest for the 960 students in the final exam outcome data set had a resulting out-of-bag MSE of 2025.95. While these predictions were better than those obtained using training and test data sets with linear regression, advising students to enroll in Stat 119A based on low predicted final exam scores (which account for only 30% of the overall grade in the course) is not as straight-forward as using the successful completion outcome.

The successful completion outcome can be easily interpreted for advising students in to the Stat 119A supplemental instruction course. For example, a student predicted to not successfully complete the course in more than 50% of the trees in the random forest could be flagged as at-risk and advised to sign up for the supplemental instruction course. The random forest for the 1032 students in the successful completion data set had an out-of-bag error rate of 22.10%.

To assess the effectiveness of the random forest for identifying and advising at-risk students, the random forest predictions were compared to the current advising indicator, Quiz 0. The instructors in this study currently use an arbitrarily chosen cutoff of 70% on Quiz 0 to advise students to enroll in Stat 119A. We wish first to determine an optimal cutoff on Quiz 0 as an at-risk indicator and then compare performance with an at-risk indicator from the random forest predictions. Figure 3 presents the ROC curve from this comparison.

In an ideal setting, we would have enough sections of Stat 119A to allow every student to enroll in the supplemental instruction course, allowing us to maximize sensitivity only. However, maximizing sensitivity results in an optimal cut-point of 100% on Quiz 0 and would not be feasible with our limited resources. Given the desired balance of enrolling students who could benefit from the supplemental instruction while staying within the number of sections budgeted for Stat 119A,

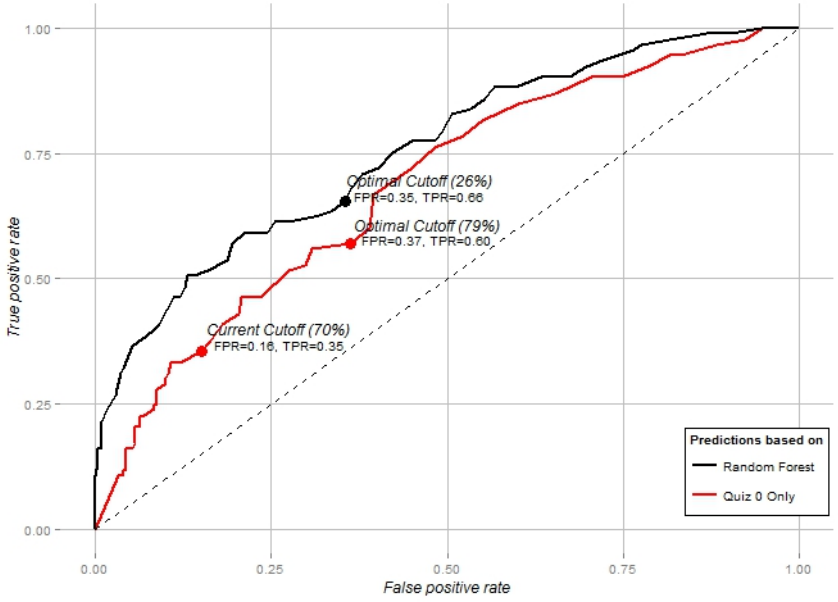

Figure 3: ROC curves based on random forest predictions and Quiz 0 scores as cutoffs for determining students "at risk" of not successfully completing the course. The graphic identifies three cut-points: optimal cutoff using the random forest predictions (26% cut-point on predicted probability of success), an optimal cutoff using Quiz 0 (79% score), and the cutoff currently used by the Stat 119 instructors (70% score) to encourage students to enroll in Stat119A for additional help.

it was decided to use a method that simultaneously maximizes sensitivity and specificity (Gallop et al., 2003) for a resulting optimal cut-point of 79%. The 79% cut-point identifies 444 (43.0%) students as at-risk and increases the correctly identified at-risk students by 50% as evidenced in Figure 3. The 79% cutoff aligns nicely with the first split rule cutoff of 78% from our regression tree for the final exam score outcome in Figure 2.

The ROC curve from the random forest prediction (Figure 3) is obtained by splitting the data set into thirds, using two thirds as a training data set to create the forest and using the remaining third to test the predictions from the random forest. Note that the random forest prediction presents a better ROC curve than the ROC curve based on predictions using Quiz 0 only.

A drawback of this method is that while we may better identify students who are at risk of failing using random forests or an optimal cutoff on Quiz 0, this method does not assess whether or not these students would actually receive a benefit if enrolled in Stat 119A. We shall explore this aspect in the next subsections.

4. Using Random Forest to Assess Pedagogical Interventions

In this section we will discuss three ways to use random forests to evaluate a pedagogical strategy or intervention. These approaches include the following.

- Using variable importance rankings for the entire data set to determine if the intervention is among the most important variables for success in the course
- Identify variables to use in stepwise regression methods to determine if the intervention is statistically significant.
- Identifying variables most important to predicting success in subsets of the population.

4.1. METHODS: VARIABLE IMPORTANCE RANKINGS

One useful feature of random forest is the ability to identify variables which may be most important in predicting a particular outcome, even in the presence of multicollinearity (James et al., 2013). For regression trees, the `randomForest` package (Liaw and Wiener, 2002) reports two measures of variable importance averaged over all trees in the forest: percent increase in mean squared error (MSE) and increase in node purity. Suppose we are looking at the importance of a variable $V1$. The percent increase in MSE is computed by permuting the variable $V1$, over the data set, and computing the percent increase in MSE between the permuted data set and the original data set, averaging across all trees in the random forest. For the increase in node purity, the difference between the residual sum of squares before and after a split using $V1$ is measured and summed over all splits and all trees in the random forest. For classification trees, the randomForest package reports two measures of variable importance: mean decrease in accuracy and mean decrease in Gini index. The mean decrease in accuracy is similar to the percent increase in MSE measure for regression trees. The decrease in accuracy is computed by permuting the variable $V1$, over the data set, and computing the classification accuracy between the permuted data set and the original data set, averaging across all trees in the random forest. For the mean decrease in Gini index, each time variable $V1$ is used in a node of a tree, the Gini index of the resulting child nodes is calculated and the difference between those child nodes and the original split node is summed over all nodes that use $V1$. These measures are computed for every covariate in the data set. A sorted list of these measures presents a rank order of variable importance.

The variable importance ranking may be used by instructors and administrators to identify factors important in predicting student success as part of curriculum development, course assessments, and intervention assessments. Of particular use in this paper, we also check where in the ranking list an intervention strategy indicator appears. This inspection gauges the importance of an intervention, especially in the presence of predictors correlated with that intervention. From an analysis standpoint, the variable importance ranking suggests a first pass on variable selection in data sets with a large number of predictors/inputs. For example, in constructing a regression model of student success, a subset of predictors with the highest variable importance ranking may be used as the initial model in the model building phase.

4.2. RESULTS: VARIABLE IMPORTANCE RANKINGS

In this section, we used the methods discussed in Section 4.1. to determine what factors had the greatest impact on student success in Stat 119: Elementary Business Statistics. A random forest with 10,000 regression (final exam score) or classification (successful completion of course) trees was used to create a list of the most important variables for predicting student success. Table 6 displays the variable importance rankings for both success measures. The performance variables

Table 6: Variable importance rankings for the top 25 of 45 covariates for the final exam and successful course completion outcome measures.

Rank	Final Exam	Successful Completion
1	High school GPA	High school GPA
2	SAT comp	Homework 1
3	SAT math	Quiz 0
4	Quiz 0	SAT math
5	Homework 1	SAT comp
6	Admission basis	Quiz 0 time
7	SAT verbal	Homework 1 time
8	Homework 1 date	Homework 1 date
9	HS grad year	Homework 1 late
10	Math location	Age
11	Math period	SAT verbal
12	Dorm	Math level
13	Quiz 0 time	Math period
14	Math level	Week 2 participation
15	Enrollment status	Dorm
16	Student level	**Enrolled in 119A**
17	Age	HS grad year
18	First semester	Math location
19	Homework 1 late	Term units attempted
20	Online units	Admission basis
21	Compact for success	Ethnicity
22	EOP	Online units
23	Full time	Major
24	**Enrolled in 119A**	Enrollment status
25	Ethnicity	Calculus level

of high school GPA and SAT scores ranked amongst the top variables for both outcomes. Though further down the list, a few other variables caught our interest. The 'previous math class' variables, which includes when the last math class was taken (Math period), where it was taken (Math location), and what the level of the course was (Math level), appeared in the top fifteen. The initial course gradebook data appeared as important predictors, Quiz 0 and Homework 1 in the top five variables. Thus, beyond the academic preparation variables of high school GPA and SAT scores, establishing success early in the course on these initial assessments are important indicators of overall course success. Residential life appears amongst the first "non-academic" variables in the rankings. As we will see later in this analysis, commuters show significantly weaker performance in the course.

The intervention variable of Stat 119A enrollment appeared on both lists as top 25 in importance (Table 6), a higher ranking for successful completion than for the final exam score as an

outcome. This ranking may strike one as low, however we would not expect Stat 119A enrollment to outright trump the academic preparation variables: we expect students with stronger academic backgrounds, as measured by high school GPA, SAT scores, Quiz 0 scores, and math level to perform better in Stat 119. To verify this and to further explore if enrollment in Stat 119A had a significant impact on success in the course, the top variables identified by the random forest were then included in a stepwise regression model selection process.

Table 7 shows that the educational background measures of SAT math score, high school GPA, admission basis and location of last math class were all significant predictors of final exam performance. For the gradebook data, a 10% increase in homework 1 score or quiz 0 score is associated with almost a 30 point increase, equivalent to a letter grade difference, in final exam score. The linear regression model had an out-of-bag MSE of 2050.50 slightly underperforming the random forest out-of-bag MSE of 2025.95.

Table 7: Linear regression inferences for model predicting final exam score (out of 300).

	Estimate	Std. Error	p-value
Intercept	99.511	53.907	0.065
High school GPA	23.193	4.481	<0.001
SAT math	0.082	0.022	<0.001
Quiz 0	26.882	8.510	0.002
Homework 1	28.854	11.499	0.012
Admission basis			0.000
FTF from CA			
FTF Foreign	37.843	8.586	
FTF Not from CA	-9.436	4.772	
Transfer	-1.651	11.235	
Homework 1 date	-0.091	0.050	0.069
Math location			0.009
High School			
SDSU	-0.087	5.231	
Transfer	21.798	8.233	
Dorm	10.589	4.175	0.011
Quiz 0 time	-0.353	0.160	0.028
Math level	1.600	1.051	0.128
Student level			0.148
Freshman			
Sophomore	-7.310	5.564	
Junior	6.795	7.860	
Senior	14.954	9.057	
Compact for Success	-11.423	6.273	0.069
Full time	15.307	8.562	0.074
Enrolled in 119A	13.873	3.963	<0.001

For the successful completion outcome, a stepwise logistic regression routine based on AIC was performed on the most important variables identified by the random forest. The final model had an AUC of 80.27% and an out-of-bag error rate of 20.37%. It had 9 variables in common with the final exam outcome measure including the indicator for enrollment in 119A, SAT math score, high school GPA, Quiz 0, homework 1, level and location of last math class, admission basis, and whether or not the student was living in the dorms on campus; see Table 8. Interestingly, participation in week 2 of the course, as measured by clicker or worksheet problems, is a significant predictor of completion, but not significantly related to final exam score.

Table 8: Logistic regression inferences for successful completion outcome measure.

	Estimate	Std. Error	p-value
Intercept	-237.9	132.9	0.074
High school GPA	1.158	0.258	<0.001
Homework 1	5.052	0.907	<0.001
Quiz 0	0.850	0.373	0.022
SAT math	0.004	0.001	0.001
Homework 1 date	-0.006	0.003	0.071
Homework 1 late	0.072	0.034	0.032
Math level	0.168	0.060	0.005
Math period	0.012	0.007	0.080
Participation week 2	0.754	0.190	<0.001
Dorm	0.847	0.216	<0.001
Math location			
Reference: High School			
SDSU	-0.028	0.248	0.909
Transfer	1.101	0.569	0.053
Admission basis			
Reference: FTF from CA			
FTF not from CA	-0.758	0.258	0.003
FTF Foreign	1.029	0.457	0.024
Transfer	0.466	0.656	0.478
Enrolled in 119A	1.211	0.268	<0.001

Based on this initial analysis, enrollment in Stat 119A is a statistically significant indicator for success in Stat 119 controlling for other factors. From Tables 7 and 8, students enrolled in Stat 119A would be expected to have a 13.9 point (4.6%) increase in final exam score and a factor of $\exp(1.211) = 3.36$ increase in odds of successfully completing Stat 119. From the logistic function, holding all other variables fixed, the estimated probability of successful course completion increases from 23% to 77%, approximately, when enrolling in Stat 119A.

4.3. METHODS: VARIABLE IMPORTANCE RANKINGS FOR INTERVENTION SUBSETS

In intervention efficacy studies, the random forest may be used to study the impact of the intervention strategy relative to student success through the variable importance ranking and so-called individualized treatment effects (ITE) as presented in Section 5. By an efficacy study, we imagine students are split into an intervention (treatment) group and a control group. We may grow a random forest, and construct variable importance rankings, for each group, namely the intervention group data set and the control group data set. The variable importance rankings may be compared between the intervention and control groups, specifically those variables which appear high in one data subset and not in the other. Such comparisons may suggest to instructors and administrators ways to improve on the intervention strategy or to determine which factors the intervention may potentially be mitigating with respect to a particular success measure.

4.4. VARIABLE IMPORTANCE RANKINGS FOR INTERVENTION SUBSETS

To determine the factors that may serve as the strongest predictors of success in Stat 119 depending on enrollment in Stat 119A, separate random forests were constructed for the students enrolled in Stat 119A and for those not enrolled in Stat 119A. Out-of-bag MSE for predicting final exam score on these two subsets was 2336 and 1999 respectively; out-of-bag error rate for predicting success completion of the course on these two subsets was 17% and 22% respectively. The factors most strongly associated with success in the course for both groups can be found by looking at the variable importance rankings for both random forests.

Similar to the findings for the final exam outcome in Section 4.2., Table 9 identifies high school GPA, SAT math score, homework 1 score, and homework 1 date as being important predictors of success in both modes of instruction. Surprisingly, Quiz 0 score was not one of the top 20 important variables for students in the 119A sections while it was for students not enrolled. In all previous and subsequent analyses, Quiz 0 is a significant predictor of success in Stat 119. One possible explanation for the exclusion of Quiz 0 in the variable importance rankings for the students enrolled in Stat 119A is that the additional help and practice mitigates the math deficiencies that may be associated with a lower Quiz 0 score. Other variables that may be closely associated with math ability - math level and SAT math score - are still among the top variables for students enrolled in Stat 119A. However, SAT math dropped from the second most important variable to the ninth, which may be further evidence of how Stat 119A can help overcome testing and math deficits.

Table 10 presents the variable importance rankings for random forests constructed on the binary successful completion outcome. Students not enrolled in Stat 119A had all measures for homework 1 (score, date, late, time) within the top 15 variables, where those enrolled in Stat 119A had only the homework 1 score and late indicator. This discrepancy may be a consequence of Stat 119A enrollee characteristics. The amount of time a student spends working on homework 1 and how early before the deadline a student starts homework 1 are indications of work ethic and the amount of out-of-class practice the student may continue to exercise throughout the semester. These two attributes are more likely to appear in the students enrolled in Stat 119A.

While high school graduation year appears among the top variables for those students enrolled in Stat 119A, there are similar measures of age and period last math class was taken for the students not enrolled in Stat 119A. Ethnicity was on the overall variable importance but was dropped from

Table 9: Variable importance rankings for the final exam outcome measure across intervention groups, variables identified as important in both groups are presented in bold font.

Rank	Not Enrolled in 119A	%IncMSE	Enrolled in 119A	%IncMSE
1	**High school GPA**	49.275	**High School GPA**	29.995
2	Quiz 0	42.733	**Math level**	19.336
3	**SAT comp**	42.511	**High school grad year**	11.817
4	**SAT math**	41.759	First semester	9.468
5	**Homework 1**	39.482	**Compact for success**	9.296
6	**Admission Basis**	28.861	**Enrollment status**	9.199
7	**Math period**	23.791	**Homework 1 date**	9.199
8	Math location	22.979	**SAT math**	8.864
9	**Dorm**	22.963	**SAT comp**	7.357
10	**Homework 1 date**	22.312	Gender	7.206
11	**High school grad year**	22.034	Took AP Stats	6.854
12	Quiz 0 time	21.717	**Homework 1**	6.711
13	**Math level**	21.384	**SAT verbal**	6.479
14	**SAT verbal**	20.497	**Age**	6.290
15	Student level	20.341	**Dorm**	6.031
16	**Compact for success**	20.029	**Admission basis**	3.906
17	**Enrollment status**	17.575	**Math period**	2.153
18	**Age**	16.299	EOP	1.699
19	Homework 1 late	15.755	First generation	1.646

Table 10: Variable importance rankings for the successful completion outcome measure across intervention groups, variables identified as important in both groups are presented in bold font.

Rank	Not Enrolled in 119A	Enrolled in 119A
1	**High school GPA**	**SAT math**
2	**Homework 1**	**High school GPA**
3	**Quiz 0**	**Homework 1**
4	**SAT math**	Ethnicity
5	**SAT comp**	Age
6	**Quiz 0 time**	**Quiz 0**
7	Homework 1 time	**SAT comp**
8	Math level	First semester
9	**Homework 1 late**	Math location
10	Homework 1 date	High school grad year
11	Participation week 2	Enrollment status
12	Age	**SAT verbal**
13	**SAT verbal**	**Homework 1 late**
14	Math period	Online units
15	Admission basis	**Quiz 0 time**

the model during the stepwise logistic regression process in Section 4.2.. Ethnicity appears near the top of the variable importance ranking for students enrolled in Stat 119A. Table 11 delves deeper into the ethnicity comparison, showing that all groups except Southeast Asian and Filipino perform as well if not better when enrolling in Stat 119A.

For both outcome measures, final exam score and successful completion, the demographic variables present as less important than academic performance, particularly for students not enrolled in Stat 119A. Gender appears only as an important predictor for final exam score of students enrolled in Stat 119A, ethnicity appears only as an important predictor of successful completion for students enrolled in Stat 119A, and age appears as a more important variable for students enrolled in 119A.

5. Individualized Treatment Effects

As in Section 4.3., we may predict the student success outcome from the random forest, but in this case we have two random forests at our disposal: for the intervention group and for the control group. We may use these forests to compute so-called *individualized treatment effects* (ITE). ITE is a measure found in the biomedical statistics literature used to quantify differences in outcomes between treatment and control groups. (Dorresteijn et al., 2011)

Let us label our two random forests as the intervention group forest and the control group forest. To compute the ITE, we first send students from the intervention group down the trees in the control group forest, thus providing a prediction performance for the intervention group students if they had not received the intervention. We next send students from the control group down the

Table 11: Percentage of students successfully completing Stat 119 by 119A enrollment and ethnicity. The Pacific Islander category had only 3 students, none of which signed up for 119A, so they were not included in the table.

	Overall		No 119A		119A	
Ethnicity	% Successful	n	% Successful	n	% Successful	n
Asian	87.5%	66	86.6%	60	100%	6
Southeast Asian	83.9%	31	84.6%	26	80%	5
White	82.5%	389	81.1%	333	91.1	56
International	79.8%	48	75.6%	41	100%	7
Filipino	78.0%	41	83.9%	31	60.0%	10
Multiple Ethnicities	77.8%	72	76.9%	52	80%	20
Other Hispanic	77.8%	72	77.8%	63	77.8%	9
Mexican American	74.2%	190	74.5%	141	73.5%	49
Other	70.0%	30	62.5%	24	100%	6
African American	61.8%	34	55.0%	20	71.4%	14

trees in the intervention group forest to predict how control group students would have performed if they had received the intervention. We may then compare performance for each student under intervention and control even though they appeared in only one. In particular, the individualized treatment effect is then calculated as the difference between how a student performed if in the intervention group (the actual scores for those in the intervention group and the predicted scores based on the intervention group tree for those in the control group) and how a student performed without the intervention (the predicted scores based on the control group tree for those in the intervention group and the actual scores for those in the control group).

The individualized treatment effects may be used in two ways. First, each semester we may predict ITE for a new class to not only flag at-risk students, but identify intervention strategies that may improve success in the course. In this paper we focus on the study of a single, supplemental instruction intervention strategy. In the discussion of Section 6., we expand on this idea for studying a suite of strategies to create a personalized set of interventions for potentially each student.

Second, ITE may be used to characterize students who will benefit the most from a given intervention. In the case of regression trees, we suggest taking the top quartile of students based on their individualized treatment effects and investigate how students in the top quartile differ on the predictor variables available from a comparison group with no treatment effect. For classification trees, we suggest comparing the students who have a treatment effect from the intervention (either those students not successful in the control group but were predicted to be successful by the intervention group trees or those students that were successful in the intervention group but were predicted to not be successful by the control group trees) from those with no treatment effect. Instructors and administrators may use these characterizations of the successful student to refine and/or create intervention strategies, identify population subgroups for further study, and more effectively devote resources towards intervention strategies and student success.

5.1. SIMULATION STUDY

The impact of being enrolled in the optional recitation section on final exam score is not a measured variable. In order to assess the strength of the ITEs to properly measure these individualized treatment effects, we ran a simulation study where the actual ITEs could be obtained.

One hundred simulated samples and corresponding random forests of 1000 interaction trees and linear models were used to assess the ability to predict these individualized treatment effects. The simulated data set consisted of 5000 observations with 12 variables from a discrete uniform distribution, $X_{ij} \sim$ discrete uniform [0, 0.1, 0.2, ..., 1.0], $i \in [1, 5000]$, $j \in [1, 12]$. To assess the ability of the ITEs to control for the selection bias inherent in our study, since students self-select into the optional recitation assignments, we assigned each observation in the simulation a propensity score based on four of the variables (X_3, X_4, X_5, and X_6). This propensity score was determined using a voting scheme where

$$\text{votes for } p_i = \begin{cases} 0.1 \text{ if } X_{ij} \in [0.7, 1.0] \\ 0.5 \text{ if } X_{ij} \in [0.4, 0.6] \\ 0.9 \text{ if } X_{ij} \in [0, 0.3] \end{cases} \quad i \in 1:5000, j \in 3:6$$

To determine if a particular observation would be in the treatment or control group, treatment assignment was generated from a Bernoulli trial using the propensity score for the individual,

$$\text{trt}_i \sim \text{Bernoulli}(p_i).$$

We created an outcome influenced by main effects from a subset of variables (X_2, X_4, X_6, and X_8) and the treatment group, as well as interaction effects from a subset of variables (X_1, X_2, X_3, and X_4).

$$Z_{ij}^{-} = \begin{cases} 1 \text{ if } X_{ij} \leq 0.5 \\ 0 \text{ if } X_{ij} > 0.5 \end{cases}$$

$$Z_{ij}^{+} = \begin{cases} 1 \text{ if } X_{ij} > 0.5 \\ 0 \text{ if } X_{ij} \leq 0.5 \end{cases}$$

$$\begin{aligned} y_i = 2 + 2 \cdot trt_i + 2 \cdot Z_{i2}^{-} + 2 \cdot Z_{i4}^{-} + 2 \cdot Z_{i6}^{-} + 2 \cdot Z_{i8}^{-} + 2 \cdot Z_{i1}^{+} \cdot trt_i + \\ 2 \cdot Z_{i2}^{+} \cdot trt_i + 2 \cdot Z_{i3}^{+} \cdot trt_i + 2 \cdot Z_{i4}^{+} \cdot trt_i + \epsilon_i, \epsilon_i \sim \text{Normal}(0, 1) \end{aligned}$$

The direct effect of the covariates to both the outcome and treatment assignment was chosen to be in an opposite direction as the interaction effect with the treatment. This was chosen as it matched what we hypothesized to be true about our application data - that students who would get the most out of the optional recitation class may be those who are least likely to sign up for that intervention and perform well in the absence of the intervention.

After the predictive models were made using both a random forest and linear model, a testing data set was created and run through the ITE procedure to obtain estimated ITEs for the observations. The actual ITEs of an observation in this simulated test data set could be calculated using:

$$\text{actual ITE}_i = 2 + 2 \cdot Z_{i1}^{+} + 2 \cdot Z_{i2}^{+} + 2 \cdot Z_{i3}^{+} + 2 \cdot Z_{i4}^{+}$$

where the estimated ITEs were calculated using $\hat{y}_{Ci}$ from the control group model and $\hat{y}_{Ti}$ from the treatment model as follows:

$$\text{estimated ITE}_i = \begin{cases} y_i - \hat{y}_{Ci} \text{ if } trt_i = 1 \\ \hat{y}_{Ti} - y_i \text{ if } trt_i = 0 \end{cases}$$

Figure 4 shows the resulting mean squared error for individualized treatment effects calculated using predictions from both a random forest and a linear model built from each of the 100 simulated data sets on a testing data set of the same size, along with the variance of the actual treatment effects. These boxplots show that the random forest performs well in estimating the treatment effects associated with a particular treatment with selection bias present.

5.2. Study Data: Individualized Treatment Effects

We now wish to quantify performance difference upon completing Stat 119A and characterize students succeeding under this supplemental instruction section. As in Section 3.2., two random forests were constructed - one for the students who enrolled in and completed Stat 119A and one for the remaining students who did not enroll in Stat 119A. The students who were not enrolled in Stat 119A are then sent down the trees in the intervention group random forest (enrolled and completed Stat 119A) to predict how these students would have performed if they had enrolled in Stat 119A. Analogously, students enrolled in Stat 119A are sent down the trees in the control group random forest (not enrolled in Stat 119A) to predict how they would have performed if they had not enrolled in Stat 119A. The individualized treatment effect is then calculated as the difference between how a student performed if enrolled in Stat 119A (the actual scores for the students enrolled in 119A and the predictions based on intervention group random forest for those not enrolled in Stat 119A) and how a student performed if not enrolled in Stat 119A (the predictions based on the control group random forest for those enrolled in Stat 119A and the actual scores for those not enrolled in Stat 119A).

The goal for these individualized treatment effects is to determine which students are at risk of failing the course if not enrolled in Stat 119A, allowing us to appropriately advise students into the Stat 119A sections. We emphasize that the analysis produces an ITE for each student, in our case being the change in score on the final exam and change in odds of successfully completing the course, if one enrolled in Stat 119A (vs. not enrolled). In practice, we may then flag individual students with large ITEs and intervene accordingly. For purposes of illustration in this section however, we present average metrics over student groups rather than details of ITEs for each of the 960 students in the final exam analyses and 1032 students in the successful completion analyses.

The average predicted treatment effect is 9.97 points out of 300 on the final exam, which is similar to the coefficient from the regression model for Stat 119A as a predictor given in Table 7 of 13.9 points.

Tables 12 and 13 present the ITE results for the final exam score outcome. For this analysis, we separate the top 25% of the students in terms of the estimated individualized treatment effect. We

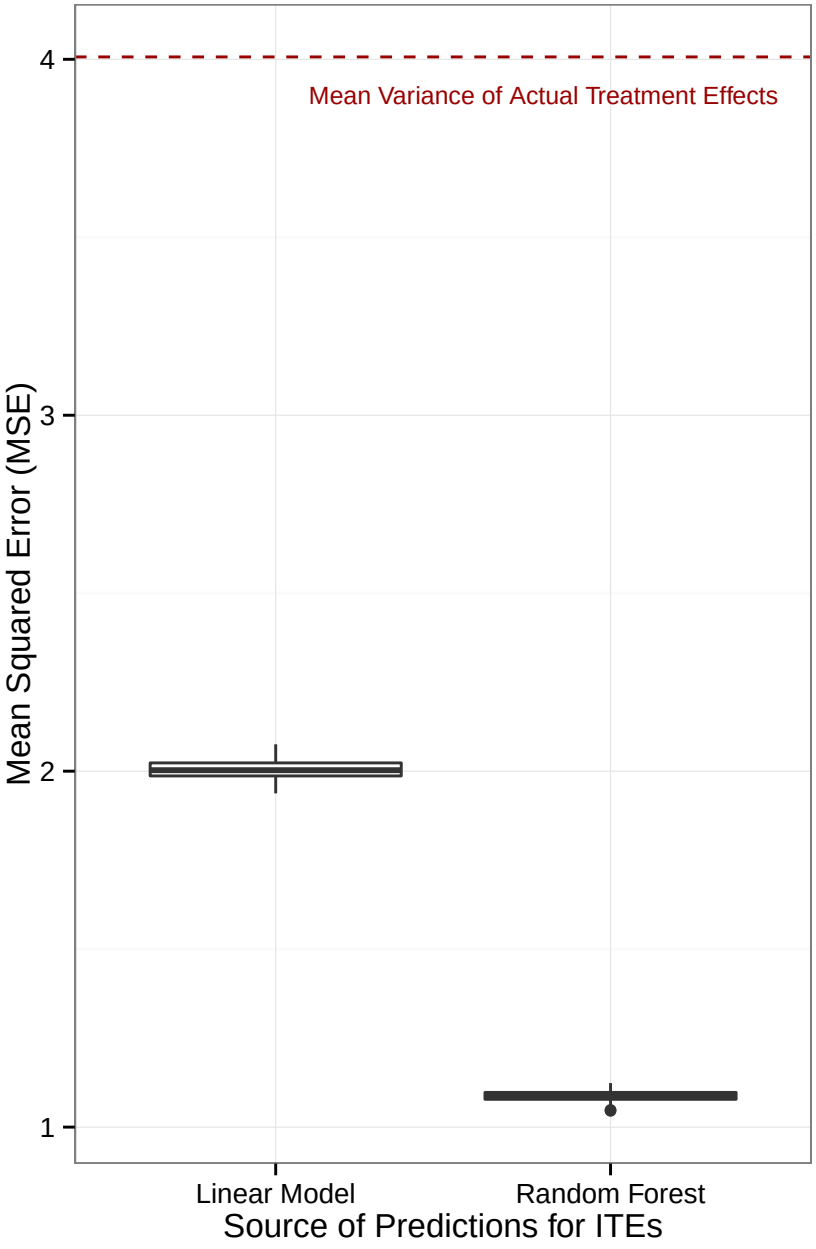

Figure 4: Boxplots showing the mean squared error (MSE) for the predicted individualized treatment effects calculated using a linear model and a random forest. The mean variance of the actual individualized treatment effects from the simulated data sets is shown for reference.

Table 12: Summary of continuous covariates for individualized treatment effects on the final exam score. Only covariates with significant differences between the different treatment effect groups are shown. The top 25% group is comprised of students with individualized treatment effects in the top 25% of ITE. The comparison group has the same number of students (25%), but with an estimated ITE of approximately zero.

	Top 25%	Comp Group	p-value
	Mean (sd)	Mean (sd)	
SAT comp	1025.7 (150.0)	1076.4 (154.2)	<0.0001
SAT math	535.6 (76.1)	558.4 (78.6)	0.0003
SAT verbal	490.0 (99.0)	518.0 (96.6)	0.0004
High school GPA	3.41 (0.50)	3.51 (0.42)	0.003
Age	19.83 (2.90)	19.15 (1.77)	<0.0001
Units attempted	13.9 (2.2)	14.4 (1.8)	0.001
Online units	1.86 (3.95)	1.20 (2.80)	0.01
Homework 1	93.1 (2.0)	96.4 (1.2)	0.002
Homework 1 late	1.09 (7.07)	0.29 (2.50)	0.01
Quiz 0	71.1 (2.4)	76.4 (2.2)	0.002
Final exam	174.79 (59.65)	217.37 (20.90)	
Treatment Effect	**72.10 (28.27)**	**6.22 (16.73)**	

also construct a comparison group with the same number of students (25%), but with estimated individualized treatment effect of approximately zero (i.e., no estimated treatment effect). Not surprisingly, in Table 12, the average SAT score, high school GPA, homework 1 score, and quiz 0 score are all lower for the students who have the largest treatment effect. Table 13 suggests that instructors and advisors should recommend Stat 119A enrollment to transfer students, commuter students, EOP students, Compact Scholar Program students, and part-time students; a larger percentage of each of these students appear in the group with the largest ITE as compared to the other group.

For the successful completion outcome measure, the data was divided into students who had a positive effect from Stat 119A (either did not pass without taking Stat 119A but were predicted to pass taking Stat 119A or those who passed taking Stat 119A but were predicted to fail without taking Stat 119A) and those who had no effect from Stat 119A. The summaries for both groups are included in Tables 14 and 15. Similar to the final exam score outcome, students predicted to have successfully completed the course when enrolled in Stat 119A had lower SAT scores, high school GPA, math level, homework 1 score and quiz 0 score. Transfer and non-resident students, commuter students, and upper-level students present as larger portions in the group gaining from Stat 119A as compared to the other group.

In Tables 13 and 15, enrollment in Stat 119A was included to see if the students were self-selecting properly into the recitation sections. Interestingly, the students who were predicted to gain the most on the final exam were more likely to be enrolled in Stat 119A whereas the students who were predicted to pass the class only if enrolled in Stat 119A were less likely to be enrolled

Table 13: Summary of binary covariates for individualized treatment effects on the final exam score. Only covariates with significant differences between the different treatment effect groups are shown. The top 25% group is comprised of students with individualized treatment effects in the top 25% of ITE. The comparison group has the same number of students (25%), but with an estimated ITE of approximately zero.

	Top 25%	Comp Group	p-value
Variables	Percentage	Percentage	
Enrolled in 119A	24.2%	15.2%	0.004
Student level			0.02
Freshman	70.4%	77.7%	
Sophomore	12.5%	12.9%	
Junior	10.8%	6.5%	
Senior	6.2%	2.9%	
Enrollment status			0.004
New - FTF	60.4%	72.1%	
New - Transfer	7.1%	4.0%	
Continuing	32.5%	24.0%	
Admitted to major	5.8%	10.8%	0.03
Admission Basis			0.02
FTF from CA	72.9%	78.3%	
FTF not from CA	11.2%	12.9%	
FTF Foreign	4.2%	3.1%	
Transfer	11.7%	5.6%	
Math Location			0.006
High School	67.5%	78.3%	
SDSU	19.6%	13.8%	
Transfer	12.9%	7.9%	
EOP	17.9%	11.9%	0.03
First semester	67.5%	76.0%	0.02
Dorm	44.2%	58.3%	0.0004
Compact for success	10.0%	5.8%	0.05
Full time student	93.3%	97.7%	0.007

Table 14: Summary of continuous covariates for individualized treatment effects on the successful completion outcome. Only covariates with significant differences between the different treatment effect groups are shown.

	Effect from 119A (n=253)	No Effect (n=779)	p-value
Variables	Mean (sd)	Mean (sd)	
SAT comp	1022.98 (158.05)	1075.24 (156.28)	<0.0001
SAT math	532.09 (83.14)	560.43 (79.21)	<0.0001
SAT verbal	490.89 (102.54)	514.82 (97.53)	0.0016
High School GPA	3.27 (0.48)	3.53 (0.44)	<0.0001
Math Level	4.31 (1.47)	4.84 (1.50)	<0.0001
Age	19.94 (2.92)	19.33 (2.05)	0.0029
Units attempted	13.91 (2.30)	14.30 (1.99)	0.0191
Online units	2.20 (4.12)	1.42 (3.27)	0.0080
Participation week 2	0.59 (0.49)	0.83 (0.38)	<0.0001
Homework 1	87.3 (2.6)	96.7 (1.2)	<0.0001
Homework 1 time	73.91 (51.03)	83.08 (58.12)	0.0196
Homework 1 late	1.57 (5.61)	0.42 (4.21)	0.0040
Quiz 0	63.8 (3.0)	77.7 (2.1)	<0.0001
Quiz 0 time	25.89 (13.89)	28.54 (10.31)	<0.0001

in Stat 119A.

6. DISCUSSION

We propose random forest as a tool for identifying factors associated with success in an educational setting, identifying factors associated with success relative to an intervention strategy or pedagogical innovations, identify covariates for regression modeling of success against an interventional strategy or pedagogical innovation indicator, and for predicting student success and identifying at-risk students who may benefit from an intervention or pedagogical innovation. While other machine learning methods may be used to make predictions on student success in a course or to compute individualized treatment effects, we choose random forest for the added benefit of variable importance rankings. Additionally, random forest has been shown to be a consistent high-performer in machine learning applications (for example see Caruana et al., 2006, Caruana et al., 2008, and Fernandez et al., 2014). We thus focus on random forest in this paper, though note that one of our current lines of research is the study of ensemble learning methods for combining classifiers/predictions.

We further propose computation of individualized treatment effects to determine the characteristics of students who would benefit the most from a particular intervention or pedagogical innovation. Random forest predictions for the individualized treatment effects could be used to advise students towards an intervention that leads to higher estimated probability of success. While we assumed that our intervention could only improve student performance in the course, this method

Table 15: Summary of binary covariates for individualized treatment effects on the successful completion outcome. Only covariates with significant differences between the different treatment effect groups are shown.

	Effect from 119A (n=253)	No Effect (n=779)	p-value
Variables	Percentage	Percentage	
Enrolled in 119A	3.8%	20.1%	<0.0001
Student Level			0.0324
Freshman	69.8%	75.9%	
Sophomore	11.9%	11.7%	
Junior	10.2%	8.3%	
Senior	8.1%	4.1%	
Enrollment Status			0.0003
New - FTF	55.3%	69.3%	
New - Transfer	5.5%	4.4%	
Continuing	39.1%	26.3%	
College Description			0.0285
Business	47.2%	55.0%	
Arts & Letters	22.1%	13.7%	
Health & Human Services	6.8%	9.3%	
Professional & Fine Arts	9.4%	6.0%	
Undergraduate Studies	6.8%	6.5%	
Sciences	5.1%	7.0%	
Engineering	2.1%	2.1%	
Education	0.4%	0.4%	
Admission Basis			<0.0001
FTF from CA	65.1%	80.4%	
FTF not from CA	16.6%	9.4%	
FTF Foreign	6.0%	3.8%	
Transfer	12.3%	6.4%	
Math Location			0.0025
High School	65.1%	76.0%	
SDSU	23.4%	14.8%	
Transfer	11.5%	9.2%	
Ethnicity			<0.0001
White	37.9%	40.2%	
Mexican American	20.4%	20.3%	
Asian & Pacific Islander	8.5%	10.7%	
African American	5.1%	3.1%	
Other	28.1%	25.7%	
First Semester	60.9%	73.7 %	0.0001
Dorm	39.1%	55.7%	<0.0001
Learning Community	9.8%	21.0%	0.0002
Taken AP Calc	13.2%	26.0%	0.0001

could be especially useful in other advising situations such as choosing amongst different teaching modalities (e.g., online, hybrid, flipped, standard offerings) or course selection. This approach could easily be extended for multiple interventions that are mutually exclusive and occur at the same time point. Future analyses will investigate how to handle synchronous interventions and interventions that are possible at different time points in the course. In all, the random forest machinery provides for a level of personalized learning, instructors and advisors able to create a "cocktail" of pedagogical strategies and interventions that best suits each student based on their institutional data workup.

We demonstrated our random forest methodology through a study of the impact of an optional supplemental instruction component in a large enrollment introductory statistics course. This introductory statistics course success study showed that the supplemental instruction component significantly increased the odds of success for students enrolling in that section. Furthermore, we identified a 79% score as an optimal cutoff on a beginning of semester math assessment quiz for advising students into the supplemental instruction section. This level is higher than the arbitrarily chosen 70% quiz score currently used in the course. Flagging students by the quiz score is a reasonable strategy for identifying potential at-risk students, though we also present a superior measure which advises students with lower than a 26% course success probability estimate from the random forest predictors into the supplementary instruction section.

The variable importance and individualized treatment effect analyses of the success study lead to one small pedagogical reform and two institutional student success initiatives relative to this bottleneck statistics course at SDSU. On the instructional end, the importance of performing well on the very first homework of the semester, measured by score, time on task, and on-time submission, motivated the instructors to put greater emphasis on engagement during the first two weeks of the course. On the institutional end, Tables 6, 7, 9, and 13 indicate significant improvement amongst Compact Scholars Program students in the supplemental instruction section. The SDSU Compact Scholars Program is aimed towards at-risk students through a partnership with a local public high school district. In Fall 2015 we will implement a learning community, requiring all Compact Scholars to enroll in an active problem-solving section along the lines of Stat 119A, though catered specifically for their academic needs. Furthermore, the study indicated that students with weaker academic background (SAT scores, high school GPA, math level, quiz 0 math assessment) as well as transfers benefitted from the supplemental instruction section. The Department is considering a "math bootcamp" prior to the semester as an intervention, for these students to review requisite algebra and pre-calculus material and help remove mathematics as a hurdle for success in introductory statistics.

Though the ITE approach provides a natural means through random forest predictions of quantifying the impact of a "treatment" on a desired outcome, educational data sets with a highly diverse set of students and potentially weak signal lends to highly variable ITE estimates. For example, ITE standard errors in Table 12 are on the high side, in fact students estimated with a negative ITE. Of course, this is not necessarily suggestive that the supplemental instruction component has a negative impact on these students, but merely suggesting that no significant benefit is observed given the standard errors. Nonetheless, we are exploring ensemble learning approaches, teaming other machine learning tools with random forests, to improve the prediction accuracy of the ITE

estimates.

The proposed method follows the approach in the personalized medicine literature of constructing two regression models to draw inferences on individualized treatment effects. A primary contribution in this paper is the use of random forests to this end. However, we are often interested in treatment effects for subpopulations. For example, in our application, what is the improvement in final exam score for Compact Scholar Program students enrolling in Stat 119A? Amongst Compact Scholar Program students, by how much are we increasing the odds of turning a D or F student into a B or C student? Subpopulation ITE estimates under the current scheme may have undesirably large standard errors. Our current research aims to unify the ITE approach completely within the random forest construct.

REFERENCES

ADAMIC, L. A., LUKOSE, R. M., PUNIYANI, A. R., AND HUBERMAN, B. A. 2001. Search in power-law networks. *Physical Review E 64,* 4.

ARNOLD, K.E. AND PISTILLI, M.D. 2012. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge LAK'12* , 267-270.

BAKER, R.S. AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1), 3-17.

BREIMAN, L. 2001. Random Forests. *Machine Learning* 45, 5-32.

DEKKER, G.W., PECHENIZKIY, M. AND VLEESHOUWERS, J.M. 2009. Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining.*

DELEN, D. 2010. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498-506.

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.

FILELLA, X., ALCOVER, J., MOLINA, R., GIMENEZ, N., RODRIGUEZ, A., JO, J., CARRETERO, P. AND BALLESTA, A.M. 1995. Clinical Usefulness of Free PSA Fraction As an Indicator of Prostate Cancer. *International Journal of Cancer*, 63, 780784.

FRITZ, J. 2011. Classroom Walls That Talk: Using Online Course Activity Data of Successful Students to Raise Self-Awareness of Underperforming Peers. *The Internet and Higher Education* 14, 89-97.

GALLOP, R.J., CRITS-CHRISTOPH, P., MUENZ L.R. AND TU, X.M. 2003. Determination and Interpretation of the Optimal Operating Point for ROC Curves Derived through Generalized Linear Models. *Understanding Statistics*, 2(4), 219242.

GOOMAS, D.T. 2014. The Impact of Supplemental Instruction: Results from an Urban Community College. *Community College Journal of Research and Practice* 38, 1180-1184.

JAMES, G., WITTEN, D., HASTIE, T. AND TIBSHIRANI, R. 2013. *An Introduction to Statistical Learning*. Springer, New York.

KIM, J.H., PARK, Y., SONG, J. AND JO, I.H. 2014. Predicting Students' Learning Performance by Using Online Behavior Patterns in Blended Learning Environments: Comparison of Two Cases on Linear and Non-linear Model. In *Proceedings of the 7th International Conference on Educational Data Mining*, 407-408.

KOTSIANTIS, S., PIERRAKEAS, C. AND PINTELAS, P. 2004. Predicting Students' Performance In Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence*, 18(5), 411-426.

KUYORO'SHADE, O., OLUDELE, A., OKOLIE SAMUEL, O. AND NICOLAE, G. 2013. Framework of Recommendation System for Tertiary. *Framework*, 2(04).

LIAW, A. AND WIENER, M. 2002. Classification and Regression by randomForest. *R News* 2(3), 18–22.

MACFADYEN, L.P. AND DAWSON, S. 2010. Mining LMS data to develop an early warning system for educators: A proof of concept. *Computers & Education*, 54(2), 588-599.

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. U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, Washington, D.C.

NORRIS, D.M. AND BAER, L.L. 2013. Building Organizational Capacity for Analytics. EDUCAUSE.

PENA-AYALA, A. 2014. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4), 1432-1462.

PHILLIPS, E.D. 2013. Improving Advising Using Technology and Stat Analytics. *Change*, 48-55.

R CORE TEAM 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

RIDDLE, D.L. AND STRATFORD, P.W. 1999. Interpreting Validity Indexes for Diagnostic Tests: An Illustration Using the Berg Balance Test. *Physical Therapy*, 79, 939950.

ROMERO, C. AND VENTURA, S. 2010. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40.6, 601-618.

ROSSMAN, A.J. AND CHANCE, B.L. 2014. Using Simulation-Based Inference for Learning Introductory Statistics. *Wiley Interdisciplinary Reviews: Computational Statistics* 6, 211-221.

SHARABIANI, A., KARIM, F., SHARABIANI, A., ATANASOV, M. AND DARABI, H. 2014. An enhanced bayesian network model for prediction of students' academic performance in engineering programs. *In Global Engineering Education Conference (EDUCON), 2014 IEEE* (pp. 832-837). IEEE.

SUPERBY, J.F., VANDAMME, J.P. AND MESKENS, N. 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining* (pp. 37-44).

VAN BARNEVELD, A., ARNOLD, K.E. AND CAMPBELL, J.P. 2012. Analytics in Higher Education: Establishing a Common Language. *EDUCAUSE Learning Initiative Paper*, 1-11.

VAN BUUREN, STEF. AND GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.

Wickman, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

ZHANG, Y., OUSSENA, S., CLARK, T., AND HYENSOOK, K. 2010. Using data mining to improve student retention in HE: a case study. In *Proceedings of ICEIS 12th International Conference on Enterprise Information Systems, Portugal.*