Evaluating a University-Wide Mathematics Assessment for Use in Introductory Statistics Course Performance

Will Boyles[1] & Alan Dabney[1]

[1] Department of Statistics, Texas A&M University, College Station, Texas, USA

Correspondence: Will Boyles, Department of Statistics, Texas A&M University, College Station, TX., 77843, USA. E-mail: will@stat.tamu.edu

Received: April 5, 2023 Accepted: June 22, 2023 Online Published: June 23, 2023

doi:10.5430/ijhe.v12n3p116 URL: https://doi.org/10.5430/ijhe.v12n3p116

Abstract

Students' prior math skills are a strong indicator of potential for success in introductory statistics courses regardless of the emphasis on mathematical ability necessary for the course. Thus, many undergraduate instructors utilize some form of math skills assessment as a review for students or for students to identify weak areas in their mathematics preparation. Such evaluations may provide instructors and students with an early indication of the student's chances for success in a course and if additional assistance is required. This study illustrated the efficacy and importance of a standardized math assessment developed and administered as a university-wide mathematics placement test in student success prediction from several modeling approaches.

Keywords: statistics education, course success, prior math knowledge

1. Introduction

Previous research has suggested a relationship between prior mathematics skills of undergraduate students and their success in introductory statistics courses regardless of the style, content, or delivery method of the course. This research has largely focused on assessments of math skills that include previous college math course grades, high school math courses and grades, short review assignments, and standardized test math scores to varying degrees of efficacy. Some universities have developed and implemented mathematics placement tests for use in identifying the correct first mathematics course for students as they enter the institution, but there does not appear to be a large scale study of utilizing this type of assessment for helping students identify which introductory statistics courses will be the best for their success or if they will need to seek to improve their mathematical background in preparation for the course they plan to take. This research study will look at retrospective data from a large, public research university to investigate the impact of using the results from a university-wide mathematics placement assessment in predicting the success of students enrolled in several different introductory statistics courses across two years.

Introductory statistics courses often serve as requirements for not only statistics majors but other majors as well. Because of this reliance upon and requirement of statistics within other majors, especially STEM-focused majors, there is a greater importance afforded to ensuring that students have the best opportunity to succeed and learn in their introductory statistics courses. Within STEM majors, studies have shown that early failures and adversity in STEM coursework can lead to extended degree completion delays and potential departure from STEM majors or higher education altogether (Cromley, Perez, & Kaplan 2016).

Mathematics skills prior to undergraduate enrollment has been shown to be another important factor in predicting the success and retention of students in STEM courses of study (Belser, Shillingford, Daire, Prescod, & Dagley, 2018, Spencer 1996). Previous studies have investigated multiple different methods of assessing prior mathematics knowledge before introductory courses. These methods have included math skills quizzes or review assignments which have long been proposed and validated (O'Neal, Chissom, & Ittenbach, 1986; Johnson & Kuennen, 2006; Lunsford, Poplin, & Pederson, 2018). The studies have also investigated and compared math skills quizzes to standardized exam math scores, such as the SAT or ACT, but no studies have been conducted on the comparison to institutional pre-enrollment assessments for introductory statistics courses.

Prior mathematics knowledge is not the only proven factor of student success in these types of courses. Studies have often considered gender in building predictive models for student success, but as shown in the 2019 report on Women, Minorities and Persons with Disabilities in Science and Engineering from the National Center for Science and

Engineering Statistics, the relationship between gender and success varies from one STEM field to the next (National Center for Science and Engineering Statistics [NCSES], 2019). As such, gender remains important to examine with regard to success in STEM coursework but still demands further research. Similarly, ethnicity has been shown to have a significant impact on student's STEM success but remains a factor that necessitates further research (Belser et. al. 2018).

This research endeavored to examine whether the institution-wide assessment may prove useful in building a model that can accurately predict students that are at high risk for being unsuccessful in their introductory statistics course. To that end, this study employed a retrospective collection of data from students enrolled in introductory statistics courses at a large research university over the course of multiple semesters to investigate the feasibility of using an established pre-enrollment mathematics skills assessment. If feasibility and validity is shown for this type of assessment, then researchers and advisors will be provided with another potential metric for prior math knowledge in determining which students will be successful and which might need further resources or assistance.

This study considered four different modeling approaches to determine whether models considering this assessment tool are useful in predicting whether a student will be unsuccessful in introductory-level statistics courses. The model approaches considered in this study were logistic regression, random forests, support vector machines, and neural networks. Each model was trained on the same data, then validated for tuning on another subset of data, and finally, compared on a subset of data put aside for testing. The models were compared on a balanced weighted accuracy metric to determine the most effective predictive model. This study found that the random forest model had the highest accuracy in identifying unsuccessful students, and the variable importance metrics illustrate that, while not the most important variable, the assessment tool under consideration was useful in the best model.

2. Methods

2.1 Course Descriptions

This study drew students from four introductory statistics courses offered at the main Texas A&M University campus, where each course is comprised of multiple sections each semester. Each course serves a unique purpose to fulfill the varying needs of the many students enrolled at the institution. The numberings for these courses are STAT 201, 211, 302, and 303. Despite the course numberings, all courses included in this study are a first course in statistics with no statistics course prerequisites. The undergraduate catalog outlines that only one of 201, 302, or 303 may be taken to fulfill the requirements of a degree.

STAT 201 is designed to serve as a general education fulfillment for students across the university with a minimal emphasis on mathematics. From the course catalog description, it covers such concepts as "Data collection, tabulation and presentation; elementary description of the tools of statistical inference; probability, sampling and hypothesis testing; applications of statistical techniques to practical problems." STAT 201 is the one statistics course included on the list of "mathematics" courses that are part of Texas A&M's Core Curriculum.

STAT 211 is intended to serve as a first course of statistics for statistics, mathematics, engineering, and other math-intensive STEM majors. It is calculus-based and covers "Introduction to probability and probability distributions; sampling and descriptive measures; inference and hypothesis testing; linear regression, analysis of variance."

STAT 302 and 303 are similarly designed courses with different major audiences in mind. Both minimize mathematical emphasis and cover "Introduction to concepts of random sampling and statistical inference, estimation and testing hypotheses of means and variances, analysis of variance, regression analysis, chi-square tests." The main difference is that STAT 302 is intended for biological science undergraduates while STAT 303 is designed for social science undergraduates.

2.2 Learning Environment

The general framework for the four introductory courses offered during fall and spring semesters is three in-person lecture hours per week across 14 weeks. The lecture hours are either provided through three 50-minute lectures or two 75-minute lectures each week. This is slightly different during summer semesters where the university has a ten-week semester or two consecutive five-week course semesters. All the course sections involved in this study during the summer semester fell into one of the two five-week sessions. For these five-week sessions, courses meet five times per week for 95 minutes.

In these courses, instructors generally employ a cumulative final exam and two or three in-semester exams which are all largely multiple-choice questions. In addition, instructors utilize homework assignments and sometimes quizzes. This is the general course make-up across sections and courses, though, occasionally, instructors might have small alterations to this framework. Instructors are also afforded the opportunity to determine the grade breakdown for each of these assignment categories with no standardized formula for this aspect of courses, although the Department of Statistics and the university as a whole have informal target grade distributions. The final grades, however, do follow a standardized breakdown on a ten-point scale where the possible results are A, B, C, D, and F for students that complete the courses.

This describes the general format for the learning environment established in the courses investigated during this study. There is a notable exception to this framework included in this study during the 2020 spring and summer semesters. Due to the global COVID-19 pandemic, the university was forced to finish out the second half of the spring semester and the following summer semester in a fully online format. The university also allowed all students to change their grading format for the spring semester, switching from the standard format to the pass/fail grading framework.

2.3 Study Design

Data were retrospectively collected and provided by the university's Office of the Registrar for 11 483 unique students from Fall 2018 to Summer 2020. Students that appeared multiple times in the initial dataset either repeated the course or took multiple different courses. The earliest appearances of these students is retained in the study. Prior to data collection, the study was approved by the university IRB. The provided data included statistics course, statistics course grade, statistics course professor, age (when enrolled in the course), gender, race/ethnicity, student classification (e.g. first-year, second-year, third-year, fourth-year), GPA, high school GPA, ACT and SAT scores, major, first-generation status, athletics status, participation in learning communities, and scores on a Math Placement Exam (MPE) that was developed and validated by the Department of Mathematics and has been used by the university to evaluate incoming students' baseline mathematical skill since 2012. The assessment covers basic algebra and precalculus skills. When collected, the data were de-identified by the registrar and sent in an encrypted file through the university's secure file sharing service, and researchers were not provided with the link between study identification numbers and student identifiers to protect privacy and confidentiality.

2.4 Statistical Methods

The final grade measurements for students include the typical letter grades of A, B, C, D, E, and F for completed courses. Since this study is focused on determining whether students are successful or not, grades of A, B, and C were treated as successful for the purpose of this study. For students that did not complete the course, grades of I, Q, or W will be considered as not successful for the purpose of this study. This study does contain the semester in which COVID-19 forced the university to switch to online course delivery mid-semester; some students were allowed to switch their grading scheme to S/U grading for that semester only. As such, grades of S are successful, and grades of U are considered not successful. The binary outcome of success is the response variable in this study for all modeling approaches.

Several variables were altered for ease of interpretation, standardization across students, or generalization beyond the specific university for this study. With the wide variety of grade frameworks present in high schools, it was necessary to standardize high school GPAs based upon the GPA scale provided in the data by the registrar. Due to some missingness in this aspect, a new binary variable was generated to indicate whether a student had a standardized high school GPA to investigate whether there was a significant difference between students with or without this variable. With two different standardized college entrance exams in the ACT and SAT, there was not one test that all students had taken. To address this issue, all students that had an ACT score and no SAT score had their score converted to an SAT score according to the concordance table as generated by ACT. If the student had an SAT score, then that score was utilized. This allowed for the creation of a new variable that contained both the original SAT scores and the converted ACT scores. The course instructors were divided into two categories to generate a new variable indicating whether the instructor was a faculty member or graduate student. To avoid potential numerical issues in building models, a variable representing college was generated. The levels of the new variable correspond to the college that offered the major in which the student was enrolled.

With the data processed and separated, classification models were fit to predict the success of the students as represented by the variable generated from final grade information. Four different models were fitted and investigated. A good overall reference for the classification models used is the book An Introduction to Statistical Learning by James, Witten, Hastie, and Tibshirani (2021). First, a common classification model for binary data is logistic regression, as outlined by Hosmer, Lemeshow, and Sturdivant (2013). With its ability to estimate class membership probabilities, logistic regression allows a thorough examination of the relationships between data characteristics and the response through the parameters or coefficients indirectly. This type of model is useful in this study for determining if the relationship between a covariate and the response merits concern and further examination. Equation 1 below is a simple example of a binary logistic regression model where π represents the probability of student being successful in the course. The log odds is a linear function of MPE where an increase of 1 in the student's MPE is expected to increase the log odds by β_1. The multiple logistic regression model in this study includes predictors of college, type of instructor, SAT score, MPE score, GPA, athletics status, age, classification, course, gender, ethnicity, first generation status, and semester. The model also included interaction terms between MPE score and GPA, college, SAT score, and course, in addition to an interaction between gender and ethnicity.

$$ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 * MPE \quad (1)$$

Another common model used in similar classification scenarios is a random forest, introduced by Breiman (2001) as a continuation of work by Ho (1995) and Amit and Geman (1997). In the framework of classification, the random forest provides prediction probabilities based on bagging by providing the proportion of trees that assign an observation to a specific category. While random forests do provide these prediction probabilities, they are considered a “black box” algorithm and provide much less interpretive information on the actual relationship between dependent and response variables. The most common information on these relationships from random forests is provided as variable importance factors, where the variables with the largest factors are considered more important by the random forest. The random forest model in this study includes predictors of college, type of instructor, SAT score, MPE score, GPA, athletics status, age, classification, course, gender, ethnicity, first generation status, and semester with no interaction terms considered.

The third model used is the support-vector machine initially proposed by Boser, Guyon, and Vapnik (1992) as a statistical learning classification method. The support-vector machine model is useful in many applications for separating data into binary “labels” based on characteristics of the data as provided to the model by using the maximum margin separating hyperplane, or the hyperplane that provides the greatest distance between the separated labels and the separating hyperplane. The set-up of this model does not easily correspond to prediction probabilities in some statistical software, and parameter interpretation of an implemented support-vector machine is difficult and not always useful. The support-vector machine model in this study includes predictors of college, type of instructor, SAT score, MPE score, GPA, athletics status, age, classification, course, gender, ethnicity, first generation status, and semester with no interaction terms considered.

Finally, a neural network classification model was considered (Ripley, 1993; Weiss and Kulikowski, 1991; Zhai and Sands, 2022). Specifically, a multilayer perceptron was considered, which supplements the classical feedforward algorithm with backpropagation, enabling updated input weights for use in the network layers. Unfortunately, the implementation of the neural network in the analysis resulted in numerical issues: regardless of our model tuning efforts, our neural network models predicted success for all students. This prevented meaningful comparison with the other classifiers, and neural networks were removed from the remainder of the analysis.

For differentiating between the models above and determining the best model for prediction, a model selection process was employed. As a classification problem, this scenario will rely largely on some form of classification error rate. Classification models can be compared on a simple classification error rate by determining the simple proportion of incorrect classifications out of the entirety of the data used in the classification. This provides an overall assessment of misclassification but has drawbacks in the specific information provided. The same issue is found in the accuracy rate which determines the proportion of correct classifications out of the entirety of the classification data. Both metrics lack the ability to differentiate between the types of error which can be detrimental to instances where one misclassification type has more severe consequences than another. To address this issue, sensitivity and specificity were employed as ways of looking at the different types of classification accuracies instead of errors, sometimes labeled as false positives and false negatives. In this case, the positive is represented by those with what was labeled as success in their statistics course, and negative is represented by those labeled as unsuccessful in their statistics course. Sensitivity, sometimes referred to as a true positive rate, describes the proportion of those with success in the course that were accurately predicted as being successful in the course. Specificity, on the other hand, represents the proportion of those labeled unsuccessful who were accurately predicted to be unsuccessful. With these metrics, there is now information on both types of accuracies, and thus indirectly both types of errors, but there are now two metrics to consider while determining the best model. In optimization problems, optimizing over two different metrics adds further complexity. This can be addressed by combining sensitivity and specificity into a single metric like balanced accuracy. Balanced accuracy is a weighted average of sensitivity and specificity. With the weighted average, emphasis can be placed on one type of accuracy compared to the other. This allows classifications to be analyzed by examining overall accuracy with the flexibility to emphasize a more important misclassification, depending on context. In this research, both the basic balanced accuracy and a balanced accuracy with greater weight placed on specificity were compared across the models. Both metrics will be looked at and the models selected using each will be compared to understand the differences and similarities. The balanced accuracy with higher weight on specificity was chosen because this research, while looking into the overall relationship between math skills and course performance, is primarily interested in predicting performance to identify those students who have a higher probability of being unsuccessful. To this end, specificity is weighted to 75% of the accuracy and leaves sensitivity to account for the remaining 25% of the accuracy. Higher emphasis on specificity corresponds to higher importance placed on correctly identifying students that would be unsuccessful, as there may be supplemental resources such as tutoring or recitations that could be assigned to help students identified as having lower chances of success.

While looking at the metrics selected, further model selection is required to “tune” model parameters and decide on a cutoff to use for predictive probabilities. The probability cutoff is a value established as a point where probability values above a threshold are considered to fall into one binary category while the probabilities that are below the threshold fall in the other category. The default value of cutoff for binary classification problems is 0.5. This value follows logically but is not always the optimal choice for a scenario based upon accuracy. In this research, an optimal cutoff is established through comparison of the balanced accuracy metric concerning high specificity across a selection of potential cutoffs. As shown in Figure 1, potential cutoffs between 0.5 and 0.99 were implemented for both models on the validation data subsets. The cutoff with the highest balanced weighted accuracy was selected and used for testing prediction accuracy. Due to only recovering prediction probabilities for random forests and logistic regression, the cutoff is only applied to these two models. The other issue in model selection is tuning parameters required by two of the three model options, random forests and support-vector machines. Random forest models have two commonly adjusted tuning parameters. First is the number of variables randomly considered at each decision split for every tree. The default is often the square root of the number of total variables available in the data. The other tuning parameter is the minimum terminal node size, where smaller values indicate a larger potential tree. The default value for this model parameter is 1 for classification. Tuning the random forest parameters occurred prior to identifying the optimal cutoff, and the tuned parameters were utilized in the cutoff selection process. For support-vector machines, there are again two model tuning parameters. These are gamma, which represents a parameter used in the kernel function, and the cost of a violation of constraints imposed in the support-vector machine model.

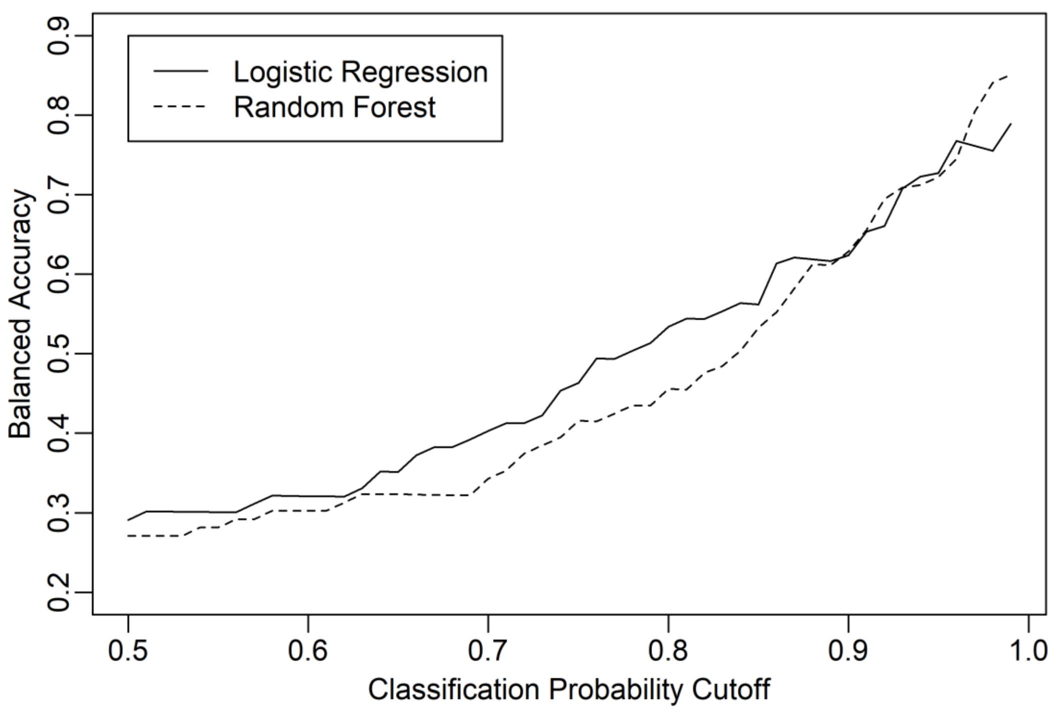

Figure 1. Classification Probability Cutoff Validation

Description: This figure illustrates the change in balanced accuracy during the validation of the choice of the cutoff for the classification probability. The solid line shows the logistic model, and the random forest model is shown in the dashed line.

In the final analyses of this study, the predictive power of the three different selected models was examined via comparison of their high specificity balanced accuracy metric. In order to ensure that the models do not overfit the data, the data was split into “training,” “evaluation,” and “testing” subsets. The data had initially been split into three different data sets via stratified subsampling to ensure that all categories were represented in each data set, and the resulting data subsets did not have structural differences across student demographics. With this method, approximately 60% of the data was used for model training, 20% for the validation processes, and 20% for testing the selected models.

3. Results

Descriptive statistics for the data collected show that there were more male students enrolled in these courses during the study, and the majority had an ethnicity of white. The distribution of ages of the students contains outliers above 40 years old and even students above 60 years of age. The majority of students, however, fell between the ages of 19 and 25 with a median age of 20. Most of the students were classified at the time of course enrollment as third- or fourth-year students by a wide margin with slightly more students falling into the third-year classification. Which course a student enrolled in differed between male and female students regarding 211 and 302 where 211 had more male students and 302 had more female students. 201 and 303 seemed to have roughly equal enrollment between male and female students. Most students were enrolled in 211 and 302 in this study, and 201 and 303 had approximately half as many students. The majority of course sections offered in the time period for the study were taught by faculty members, so a majority of the students were taught by faculty members instead of graduate students.

In a first look at the three different models as established in the training process, the random forest seems to have the best accuracy in predicting whether a student will be successful. SVM has the second-best training accuracy, and the logistic regression model has a slightly worse training accuracy. Validation and testing of the models will indicate whether these are due to overfitting or are sustainable in future analyses.

In the validation process, the tuning parameters selected for random forest and SVM were not drastically different from the default values frequently utilized in analysis. In the selection of prediction probability cutoffs for the random forest and logistic regression models, the higher values led to higher balanced accuracies. This is expected given the greater emphasis placed on specificity in the accuracy metric chosen for this study. One concern that arose in the validation process was the steep decline in accuracy for the SVM model which is an indicator of potential overfitting in the training set.

After training and validation, the finalized models were compared across the balanced accuracy metric for prediction in the testing data set aside for this purpose. As shown in Table 1, SVM performs the worst with an accuracy metric of 0.35 which is worse than its validation accuracy. The logistic regression model’s accuracy increased slightly from its validation accuracy to 0.77. The random forest model remained the model with the highest accuracy and increased to 0.79.

In the variable selection process, MPE scores were selected through least absolute shrinkage and selection operator (LASSO) in the logistic regression framework which indicated that MPE scores are a useful predictor. To verify significance, likelihood ratio tests were utilized for each selected variable. In the likelihood ratio tests shown in Table 2, MPE scores were found to be significant but were not the most significant predictor. For the random forest model, MPE scores were in the top three most important variables. These results illustrate that while the MPE might not be the most significant or important predictor, it is still useful in predicting students' success in the random forest model.

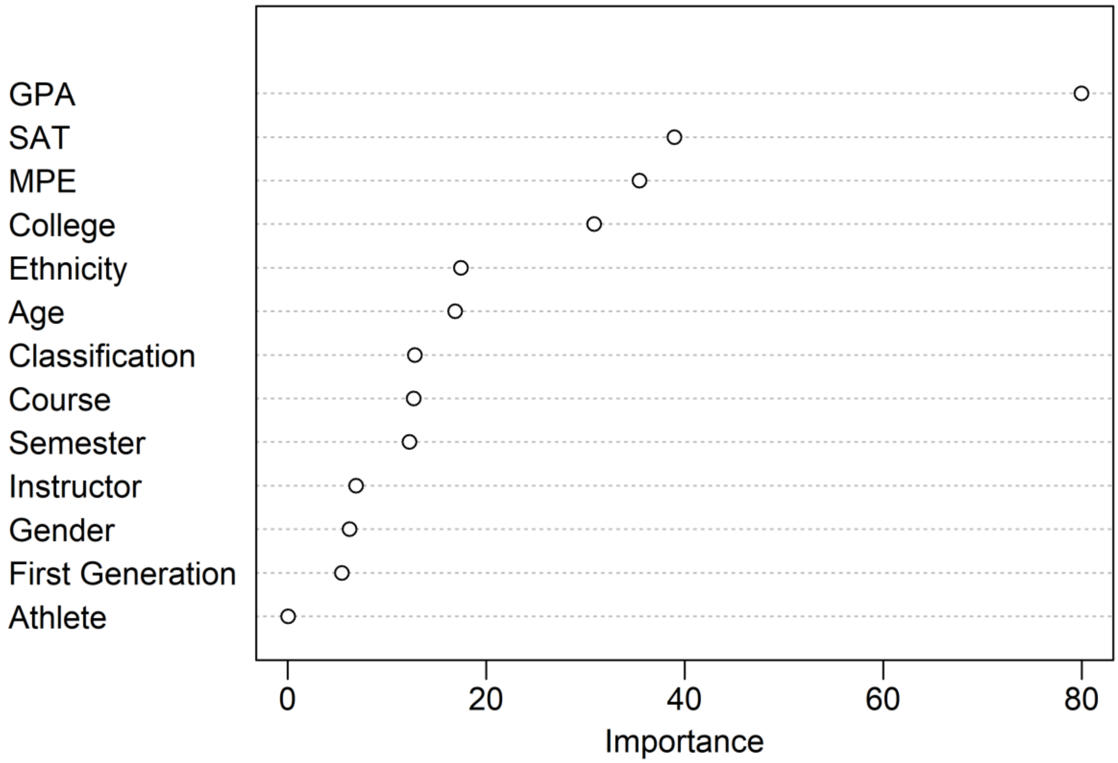

Figure 2. Random Forest Importance

Description: This figure shows the importance of variables from the random forest model in descending order where the importance is determined by the mean decrease in node impurities across all trees as measured by the Gini Index.

As expected and shown in previous studies, undergraduate GPA was a significant and important predictor of success in introductory statistics courses. In the random forest model, GPA was found to be the variable with the highest importance, and the logistic regression model found that GPA had the lowest p-value. Logistic regression model results indicated a positive relationship between GPA and the odds of student success.

Other variables were found to either be significant predictors or have high importance, but there was disagreement between the models about which predictors were significant or important. The student's classification, their college, semester, and the type of course instructor were all found to be significant predictors within the likelihood ratio tests for logistic regression. Model results showed that second-year students performed the worst, and fourth-year students performed the best. Most colleges performed similarly with only students in interdisciplinary studies performing worse. Students performed worse in spring semesters, and summer semesters saw slightly higher results than fall semesters. Graduate student instructors had higher rates of success than faculty according to the logistic regression model. In the random forest model, the student's college was still important, but SAT score was now an important variable for prediction. Instructor type and student classification had lower importance in the random forest model.

There were variables that were still selected by the models but were not found to be as significant or important. This represents variables that might be useful in prediction but are not the driving force behind any decisions made by the models. For the logistic regression model these include ethnicity, gender, course number, athletics status, age, and first-generation status. The variables listed that are not included in Table 2 did not have interaction terms or multiple categorical levels and were examined via regression p-value. None had significant results. Of the variables, athletics status had the lowest importance in the random forest model. With the data containing semesters that occurred before and during the pandemic response, the study allowed an investigation of whether the pandemic had an impact on the rate of success in the courses. A preliminary look at the rates shows that there do not appear to be any major changes in the rate of success.

Table 1. Balanced Accuracy Results

	Logistic Regression	Random Forest	SVM
Training	0.8354	0.9977	0.9449
Validation	0.7599	0.7730	0.3847
Testing	0.7703	0.7921	0.3492

Description: This table contains the best balanced accuracy measures for each of the three model types from the training, validation, and testing processes.

Table 2. Likelihood Ratio Test

Variable	LRT p-value
MPE	0.0404
Instructor Type	0.0029
College GPA	<0.0001
Gender	0.5076
Classification	<0.0001
Semester	0.0212
Course	0.0696
College	0.0109
SAT	0.2553
Ethnicity	0.3142

Description: This table contains the likelihood ratio test results for the variables in the logistic regression model in the training process.

4. Discussion

Based on the results of the analyses, there are a few approaches to addressing or responding to the information learned from this study. In determining the best approach to this research, individuals and institutions will need to carefully consider and examine their own needs, circumstances, and concerns before adapting this research to their own processes. If an institution does not currently employ a pre-enrollment math skill assessment this could be used to justify implementation, or the general framework of this study could be used to assess the validity of an assessment developed in the future. If there is an assessment already established, the institution or individual could utilize the assessment for success prediction as outlined in this study or further research as mentioned below. When implementing the pre-enrollment assessment as a predictor of success, it will be necessary to think through who will be using the information and how they might be using the information.

As stated in the Results section, the random forest model provided the highest potential balanced accuracy in the model testing process. For the purposes of this study, the best balanced accuracy is the goal and thus leads to the selection of the random forest model for prediction of students that might need extra resources to be successful. The results for the random forest model as illustrated in Figure 2 show that while not the most important variable, the MPE score is still a somewhat important variable and would likely merit inclusion in multivariable analysis for other studies including the same variables. A potential drawback arises from the lack of clarity as to what would be the driving factors behind whether a student is identified as needing the resources or not. If there is the potential for concerns as to why a student might be offered, encouraged, or even required to use extra resources, the black-box nature of the random forest algorithm would lead to difficulties in the explanation of any decision made based on this research.

If the black-box nature of a random forest model is too much of a concern or drawback for use by an individual or institution, the logistic regression model was shown to be a useful predictive model with a modest drop-off in balanced accuracy during the model testing process. This decrease in accuracy is the trade-off for a model with a decision-making process that would be easier to explain and understand.

The support vector machine model, while performing the best in the training accuracy, did substantially worse in predictive accuracy during both the validation and testing process. This leads to the conclusion that the model severely overfit the initial data and was not useful for future prediction. The support vector machine is a black-box model much like the random forest model, and thus would not provide any benefit over either other model.

4.1 Strengths

Of the studies involved in examining the importance of math skills prior to an introductory statistics course, few collected data over a period longer than 1 or 2 semesters. One of the strengths of this study is the ability to collect large amounts of data across several years' worth of courses in multiple different semester frameworks. The number of subjects in this study is larger than most other studies focused on questions around this topic. Previous studies also rarely examined institutional pre-enrollment assessments as the source for knowledge about prior math skills, often relying on start-of-semester assignments or standardized tests such as the math score from the SAT or ACT.

The large study size at a research university also allows this research to cover students from a wide breadth of disciplines in multiple styles of introductory courses which is not something seen in many other studies. This study also involves many different instructors both from faculty and graduate students, whereas prior work has largely been limited to a few faculty members as the instructors of the studied courses. With these many different instructors, the approaches to these courses will also naturally differ slightly from section to section.

4.2 Limitations

Due to the nature of the study and unforeseen circumstances, there are limitations to the analysis and conclusions that can be drawn from this study. The first and most prominent limitation is due to the nature of the data collection in this study. As the data is collected retrospectively, the information is limited to what is collected and stored by the registrar for the university. As such there are factors shown in the extant literature to be significant predictors of course success that were unable to be included in the study. This includes variables such as employment status, statistical anxiety, study hours per week, etc.

Because this study involved semesters that overlapped with the global COVID-19 pandemic, there were potential concerns over whether these semesters would have a significant impact on the likelihood of student success. In the course of the research, these concerns were investigated as outlined in the Results section. The examinations of the concerns showed there was not a significant impact due to any changes occurring in the two altered semesters that were included in the data collection.

4.3 Future Considerations

Future plans for this area of research involve enrolling students in a study at the start of a semester and sending out a survey to gather the information from students directly. This survey would include information not available through the retrospective data collection process and shown in previous literature to be a significant predictor of success. This type of data collection would be more feasible with the framework of a concurrent study. Future proposed research would also consider examining comparisons between university-wide assessment and standardized math pre-assessments from statistical education literature.

To further verify the validity of these analyses, expanding this research to other similar and different institutions would provide a better understanding of the results and their portability from one institution to another. It would also afford the opportunity to examine further different styles of introductory courses. Finally, it might be of interest to move beyond introductory statistics courses to introductory courses in other disciplines that might rely upon math skills or concepts.

5. Conclusion

The main lesson gleaned from this research is that, if available, an institution-wide pre-enrollment math skill assessment, such as the Math Placement Exam, has the potential to serve as a predictor for student success in introductory statistics Courses. Discovering this allows for another metric that can be used by faculty, advisors, and administrators for identifying students in need of resources and examining the efficacy of current courses and resources. For schools that do have the types of pre-enrollment assessments this provides an example study for determining the validity of their own assessments and what kinds of models might be expected to provide accurate predictions of student success.

Acknowledgements

The authors would like to thank and acknowledge Scott Crawford, Darren Homrighausen, and Vickie Williamson for their support, thoughts, and feedback in the research process for this project. We would also like to express appreciation to the Office of the Registrar at Texas A&M University for collecting and processing all the data used in this research.

References

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation, 9*(7), 1545-1588. https://doi.org/10.1162/neco.1997.9.7.1545

Belser, C. T., Shillingford, M., Daire, A. P., Prescod, D. J., & Dagley, M. A. (2018). Factors influencing undergraduate student retention in STEM majors: Career development, math ability, and demographics. *Professional Counselor*, *8*(3), 262-276. https://doi.org/10.15241/ctb.8.3.262

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). https://doi.org/10.1145/130385.130401

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 4-11. https://doi.org/10.1177/2372732215622648

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. https://doi.org/10.1002/9781118548387

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning, Second Edi. *New York: Springers*. https://doi.org/10.1007/978-1-0716-1418-1

Johnson, M., & Kuennen, E. (2006). Basic math skills and performance in an introductory statistics course. *Journal of statistics education, 14*(2). https://doi.org/10.1080/10691898.2006.11910581

Lunsford, M. L., Poplin, P. L., & Pederson, J. G. (2018). From research to practice: Using assessment and early intervention to improve student success in introductory statistics. *Journal of Statistics Education, 26*(2), 125-134. https://doi.org/10.1080/10691898.2018.1483785

National Center for Science and Engineering Statistics (2019). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019 | NSF - National Science Foundation*. (2019). Nsf.gov. https://ncses.nsf.gov/pubs/nsf19304/

O'neal, M. R., Chissom, B.S., Ittenbach, R.F. (1986). Predicting Success and Measuring Math Skills Improvement in an Introductory Statistics Course. Retrieved from https://eric.ed.gov/?id=ED289907

Ripley, B. D. (1993). Statistical aspects of neural networks. *Natworks and Chaos-Statistical and Probabilistic Aspects*, 40-123.

Spencer, H. E. (1996). Mathematical SAT test scores and college chemistry grades. *Journal of Chemical Education, 73*(12), 1150. https://doi.org/10.1021/ed073p1150

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc..

Zhai, H., & Sands, T. (2022). Comparison of Deep Learning and Deterministic Algorithms for Control Modeling. *Sensors, 22*(17), 6362. https://doi.org/10.3390/s22176362

Copyrights