THE RESEARCH AGENDA

Machine Learning Methods for Course Enrollment Prediction

By Lucy Shao, Martin Ieong, Richard. A. Levine, Jeanne Stronach, and Juanjuan Fan

Accurately forecasting course enrollment rates in higher education is of great concern in order to minimize unnecessary administrative costs as well as burden to both students and faculty. This research aimed to first recreate course enrollment predictions based on a conditional probability analysis using student data from San Diego State University (SDSU) and then to improve upon those predictions by applying classification and regression trees (CART) and random forest. The authors incorporated student demographic and academic information into algorithms to ascertain their influence on improving course enrollment prediction accuracy. They used these strategies to predict enrollment in General Chemistry at SDSU, a course with historically varied and large enrollment numbers in the multiple hundreds per semester. The authors then determined which factors were the most influential to the General Chemistry enrollment number using a variable importance metric derived from tree-based algorithms.

In higher education, accurately forecasting the enrollment rates for individual courses helps the university allocate appropriate resources, including the number of seats for the course and course sections, assignment of classroom or lab space, assignment or hiring of instructors, assignment of teaching assistants, and management of wait-lists, in addition to study of course conflicts across a university. Unnecessary administrative costs can be incurred when the demand for courses is not accurately predicted. If a course is over-enrolled, the department must rush to allocate additional space and instructors for the additional students. If the course is under-enrolled, a class or class sections may be cancelled, creating challenges for student class schedules and instructor work schedules. Both scenarios lead to potentially immense stress on students, instructors, and administrators, typically at the start of a term.

At San Diego State University (SDSU), the course prediction model currently in place is heuristic, relying mostly on recent semester enrollments while making

adjustments as students enroll in real time. Such practice serves its purpose in most circumstances, but historically has failed to give accurate predictions in regard to courses with high enrollment variance. One of these classes is Chemistry 200: General Chemistry, taken by students in their first or second years. For a number of STEM majors, the course is a core prerequisite for further courses in the student program of study. There are several factors that may possibly contribute to this course having a high variation in its enrollment rate. Most noteworthy is the course success rate and number of students repeating the course each semester to satisfy a major pre-requisite.

The authors first looked at the current commonly used methodologies for predicting course enrollment. The most prominent methods assume a model of static linear growth at the university, *i.e.* the student population is projected to grow a certain percentage annually and therefore the course enrollments are increased by a proportional amount. Another popular method involves a flowchart or conditional probability analysis wherein students are placed into category buckets based on perceived characteristics that would impact their likelihood to enroll in a particular course (*e.g.* freshman vs. transfer student). This latter method then identifies the proportion of students in each bucket that ultimately enroll in the course in question to create a prediction for future terms.

Although the literature is limited in research on course prediction, a few papers are worthy of mention. Kraft and Jarvis (2005) posited a conditional probability system of course prediction that achieved a very high level of predictive accuracy, reported at approximately a 2 percent error rate in most circumstances at Clemson University. This prediction system buckets students into broad demographics such as student status upon matriculation and when the student successfully meets course prerequisites. The model then finds the number, and proportion, of students in each bucket at each successive step. The final enrollment prediction number is found by running the entire student body through conditional enrollment proportions. The paper claims that making predictions on a per person basis, of whether or not individual students would enroll in a particular course, is not viable because of high variance. Other limitations of this study are that a limited number of factors are considered in a conditional probability analysis, and the interactions between variables are rarely considered. One of the deficiencies of the Kraft and Jarvis (2005) conditional probability analysis method is that it places students into buckets after the actual enrollment was finalized. This could lead to much better predictive performance when predicting the next year's enrollment rate because of year-by-year variability. In a machine learning sense, there should be training and testing data split for better validation of the flowchart model in the conditional probability analysis; the two percent error rate does not represent the true performance of the model. In general, machine learning in data science refers to the use of computer algorithms to make accurate predictions of the outcome.

At the University of Virginia, a cohort-based method is used (University of Virginia 1999). This method, called the Grade Progression Method, suggests that as a cohort of students moves from year to year, a certain portion of the students will drop from the curriculum of interest, but most of the students will enroll in the course of interest in the same year of study. Using historical data, it is possible to determine what proportion of students will choose to move on to other curriculum paths and then, using the cohort size of each year, create a prediction of how many students in the original cohort will remain and eventually enroll in the course of interest. Five years of historical data are used to stabilize the prediction number. This approach is consistent only when cohorts have a predictable plan of study from year to year, which most often occurs at the graduate level. At San Diego State University, such consistent paths are not applicable for most majors, particularly at the undergraduate level for large enrollment courses with multiple hundreds or more students per semester.

The University of California employs a multiple participation rate calculation that separates students on the basis of ethnicity to find their participation rate, *i.e.*,

the proportion of students in that ethnicity that ultimately enroll. A prediction coefficient is created and multiplied by the entire student ethnicity population in question (University of California 1999). Four years of historical data is used in order to stabilize the prediction coefficient. The University of California has found that this method is successful for long-range predictions. Unlike either of the two methods cited above, this approach uses demographic information in the form of ethnicity in order to categorize students into different groups that have demonstrated some difference in enrollment patterns. However, it has yet to be seen if additional demographic information and prior academic performance data could be added to this model to create a more accurate predictor.

As the effectiveness of prediction methods may vary from university to university, this paper begins by reproducing conditional probability analysis using data from SDSU. The authors then propose two tree-based models of course prediction that consider historical data in the form of student demographic information, historical academic performance (*i.e.*, before matriculating to San Diego State University), and current academic information such as student standing or declared major. The first model was built using a classification and regression tree (CART) algorithm (Breiman, *et al.* 1984) to create a single decision tree. By introducing additional data in the form of historical student information that is known by the university prior to students' registration period, this decision tree aims to improve upon the predictive accuracy of pre-existing models. The second model is a random forest algorithm (Breiman 2001) that built a forest of decision trees and often achieves better prediction via model averaging. The authors aimed to report the most reliable prediction results using machine learning techniques while considering comprehensive factors that could contribute to the variation in the course enrollment number. The authors note that Soltys, *et al.* (2021) proposed a machine learning framework for predicting enrollment of applicants to a university. As course enrollment requires a different set of features and specific considerations with respect to pre-requisite courses and student performance at the university, this paper compliments Soltys, *et al.* (2021) well toward successful applications of machine learning to solve enrollment management problems.

Data

The data used for this study comprised all students who enrolled at San Diego State University from 2010 through 2019. Across the years in question, there were approximately 83,000 students who enrolled at SDSU who were available as data points for the models. The data set contained 45 covariates that described in different capacities a student's academic history and demographics. This information included gender, disability status, SAT scores, first-generation status, minority status, and high school GPA, among others (*see* Appendix A, Table 4, on page 24).

The authors split the data into training and testing sets to give an accurate model performance evaluation. The authors created models through gaining information from training data, and then predicted on testing data. For example, to predict the number of enrollments in fall 2019, the training data contained the information from previous falls, and the testing data used the information about fall 2019. So, it is important to keep training data and testing data of the same structure in order to produce accurate predictions. In the analysis, the authors created individual data sets for each fall semester in order to obtain timely information, such as the number of years between the time a student graduated from high school and the target fall semester and the number of days between a student's completion of the prerequisite for the course and when the target fall semester began. For each fall semester, the authors curated the data including only the students who entered the university within five years from that fall semester. This constraint removed earlier cohorts that may have had different patterns from the current cohorts. Since the authors were creating models based on the previous three years of data, there were three training sets. They took an ensemble of the three models created across these three training sets to obtain the final model.

Conditional Probability Analysis

Methods

The authors first recreated the Kraft and Jarvis (2005) course enrollment prediction system. As mentioned earlier, this prediction system buckets students into broad academic backgrounds such as student status upon matriculation and whether the student successfully meets course prerequisites. The approach then creates conditional probabilities using the proportion of the students who take the target course in each bucket from previous years. Since every university has different practices and student bodies, it was important to make sure that the results derived from this flowchart-based analysis at Clemson University had a similar level of predictive accuracy at SDSU. This analysis built upon the Kraft and Jarvis (2005) method by first replicating their model using data from San Diego State University.

There were a few differences between the original analysis and the analysis that is the focus of this paper. In the original study, students were separated into three different buckets based upon their majors: one of either a General Engineering major, a major that needs the course in question as a prerequisite for a future major course, or other major. The General Engineering bucket was needed because of a limitation in their data source that made the data concerning General Engineering students inaccurate. The SDSU data did not have such a limitation. As such, this conditional probability model contained only two buckets for whether or not a student needed the course in question as a prerequisite for a future major course. In sum, the method grouped students by transfer status, prerequisite met term, whether the course is a major prerequisite, and whether the students had taken General Chemistry before. Furthermore, the authors' analysis had a separate layer between major groups and final course enrollment that analyzed whether or not a student had failed the General Chemistry prerequisite. At Clemson University, the course in question did not have any prerequisites whereas General Chemistry at SDSU had a prerequisite that bars certain students from enrolling in General Chemistry otherwise. The detailed structure of the conditional probability analysis is shown in Appendix A, Figure 5 (on page 27).

Table 1 (on page 15) shows the percentage of enrollment for each student population bucket that fell under a particular category, as was previously discussed. For example, the first cell of the first row of Table 1 is the conditional probability of enrollment among first-time freshmen who needed General Chemistry as their major prerequisite; in other words, 32.79 percent of the students who were first-time freshmen and needed General Chemistry as their major prerequisite took General Chemistry from fall 2015 to 2017. The enrollment prediction using this probability 0.3279 on 2018 fall data was 303 students. The actual fall 2018 enrollment number for first-time freshmen who needed General Chemistry as their major prerequisite was 412.

The conditional probabilities were obtained from each fall semester over the past three years, and the conditional probabilities displayed are an average of the three sets of conditional probabilities. The authors also made sure that each bucket of students was mutually exclusive. For example, in the flow chart, the group of "In pre-req two terms ago" did not contain any students from the group of "In pre-req last term."

Results

Table 1 shows the average enrollment rates of each student population subgroup from 2015 to 2017 as well as the predicted enrollment numbers for fall 2018. Through this bucketing method, it became clear that the conditional probability analysis did not have the same level of accuracy across each of the buckets. Some of the subgroups, in particular students in the last group who had not passed the prerequisite or taken the class in the last two terms, were over-predicted by this method ("Other Students" row in Table 1). On the flip side, some of the subgroups were under-predicted, for example the first subgroup of first-time students requiring this course for their major ("First-time" row in Table 1).

The largest group of students who enrolled in General Chemistry fell under the category of first-time

TABLE 1 ➤ Conditional Probability Analysis: Conditional Enrollment Rates for Predicting Fall 2018

Condition(s)			Predicted Enrollment Rate (%)[1]	Fall 2018 Prediction	Fall 2018 Actual (n)
New Students					
First-Time and Met Prerequisite	Major Prerequisite		32.79	303.33	412
	Everyone Else		22.33	22.33	25
Transfer and Met Prerequisite	Major Prerequisite		31.13	5.29	5
	Everyone Else		32.92	10.20	10
Continuing Students					
Passed Prerequisite Last Term	Major Prerequisite		35.00	40.97	71
	Everyone Else		48.48	40.73	62
Passed Prerequisite Two Terms Ago	Major Prerequisite		12.75	7.90	6
	Everyone Else		10.43	2.60	4
In Course Last Term	Major Prerequisite	Passed	4.22	17.94	5
		Failed	20.07	5.02	3
	Everyone Else	Passed	2.89	3.55	2
		Failed	18.16	0.55	0
In Course Two Terms Ago	Major Prerequisite	Passed	1.20	3.50	3
		Failed	4.00	0.92	1
	Everyone Else	Passed	0	0	1
		Failed	0	0	1
Other Students	**All**		5.00	162.00	37
Overall				**627.07**	**650**

[1] Average of previous three Fall term enrollment rates (actual) used for calculating current Fall enrollment prediction.

students who needed the course as a major course prerequisite. Table 2 (on page 16) predicts that 313 students would enroll in the 2019 fall semester, whereas 396 students actually enrolled. This analysis showed that the overall strategy of bucketing students into different groups depending on individual characteristics could be an effective strategy; however, although the overall enrollment prediction is 78.8 percent accurate, the accuracy for individual groups deviated significantly. But the authors also predicted the 2018 fall enrollment rate using 2015–2017 data, and the accuracy was better. The actual enrollment number was 650 for fall 2018, but the authors predicted only 627 based on the flowchart method (*see* Table 1), an error rate of 3.6 percent. These illustrations show that the year-to-year variation in General Chemistry enrollment could create predication accuracy challenges for the conditional probability analysis.

The primary drawback to the conditional probability analysis method is that it uses only limited academic information. The inclusion of demographic information and even academic performance information prior to college may allow for adjustment to the model. The flowchart groupings are also rather subjective based upon the analyst's perceived notions of what background differences would affect likelihood to enroll. In particular, by introducing a tree-based model, the authors aimed to overcome these drawbacks and create a statistically defensible strategy to identify which groups of students have differing likelihoods of enrolling in

TABLE 2 ➤ Conditional Probability Analysis: Conditional Enrollment Rates for Predicting Fall 2019

Condition(s)			Predicted Enrollment Rate (%)[1]	Fall 2019 Prediction	Fall 2019 Actual (n)
New Students					
First-Time	Major Prerequisite		35.47	62.8	157
	Everyone Else		12.98	1.7	11
Transfer	Major Prerequisite		27.98	2.5	6
	Everyone Else		35.62	2.9	4
Continuing Students					
Passed Prerequisite Last Term	Major Prerequisite		44.92	53.0	75
	Everyone Else		55.38	48.7	60
Passed Prerequisite Two Terms Ago	Major Prerequisite		11.06	6.3	9
	Everyone Else		11.60	2.6	2
In Course Last Term	Major Prerequisite	Passed	2.70	9.7	17
		Failed	17.11	7.9	7
	Everyone Else	Passed	1.94	2.4	2
		Failed	11.71	1.4	1
In Course Two Terms Ago	Major Prerequisite	Passed	1.17	5.5	2
		Failed	3.53	1.2	0
	Everyone Else	Passed	0.34	0.4	0
		Failed	5.55	0.2	0
Other Students	**All**		3.35	102.8	43
Overall				**313.0**	**396**

[1] Average of previous three Fall term enrollment rates (actual) used for calculating current Fall enrollment prediction.

General Chemistry. The Kraft and Jarvis (2005) conditional probability analysis provided a good starting point from which, as explained in the next section, the authors improved through a tree-based approach for more optimal splits between student groups.

Classification Tree Analysis

Methods

A decision tree is a flowchart-like structure in which: each internal node represents a "test" on an attribute (*e.g.*, whether a coin flip comes up heads or tails); each branch represents the outcome of the test; and each leaf node represents a class label (decision taken after computing all attributes) (Wikipedia: The Free Encyclopedia 2020). An example of a decision tree is shown in Figure 1 (on page 19). The root, or top-most, node splits students as to whether the date their prerequisite was fulfilled is at least 37 days after matriculation (send student to the left) or less than 37 days after matriculation (send student to the right). Internal nodes below the root node in the decision tree are similarly characterized by decision rules that send students down the tree to the left or the right. All branches end in so-called terminal nodes where students are collected. In Figure 1, the terminal nodes do not have a split rule, but instead present the number of students who end up in the terminal node (bottom number) and the probability a student in that node enrolls in the General Chemistry course (top number). For example, the first terminal node in Figure 1 (at the left most bottom end of the tree) shows that 90 percent of the students met their prerequisite more than 37 days after the term started, including those students who never met the prerequisite.

The terminal node shows that there was a probability of 0.0000031 for those students to enroll in the General Chemistry course. Note that students who fulfilled the pre-requisite at least 37 days after the semester began were sent from the root node to this first terminal node. The students who fulfilled their prerequisite 75 days before the term started were sent to the second terminal (from the left) in Figure 1 (on page 19).

Note that the conditional probability analysis also produces a decision tree. But the flowchart therein was created by the authors' intuition without validating if the splits were data informed. In particular, the authors needed machine learning algorithms to help identify the best tree splits. The authors incorporated supervised learning methods using an algorithm to predict the target variable (here enrollment status).

Classification and regression tree (CART) operates by recursively splitting the sample space in such a way that each split increases the purity of the resulting two spaces (James, *et al*. 2013). Purity can be measured in a number of different ways; for the purposes of this analysis, the Gini index is used:

EQUATION 1.

$$\text{Gini Index} = 1 - \sum_{k=1}^{K} p_k^2$$

Where:
K = number of classes and
p_k = probability of an object being classified to a specific class

For our analysis $K=2$, with the two classes being *enrolled* and *unenrolled*. The Gini index ranges from zero to one with smaller values signifying purer nodes. For example, a node containing observations with only one category/class would have a Gini index value of zero. The best split is defined as the one that reaches the maximum reduction in Gini index between the parent node and the average of the two child nodes, weighted by the proportion of observations sent to the two child nodes.

With this method of recursive splitting, the purity of the resulting leaf nodes would always increase to some extent, and this may lead to overfitting in the resulting model. To this end, CART employs a pruning process based on a cost-complexity measure, which consists of the level of misclassification at each terminal node and a penalty term that increases as the tree gets larger. Using the cost-complexity measure, the tree is then pruned such that the resulting tree is the best taking into account both the overall fit and tree complexity. For the authors' analysis, pruning was not the best approach because of the imbalance in the data. Only approximately 500 students per semester were enrolled in General Chemistry out of a total of approximately 400,000 students in the data set. In such a scenario, the splitting algorithm would select the majority class in most scenarios, as simply defaulting to the majority class without a single split would yield an approximately 98 percent accuracy. This would produce a predictively accurate but ultimately impractical and unusable tree. Instead of pruning, this analysis used a minimum number of observations for each leaf node as a stopping rule.

The authors coded the algorithm in R and used the R package rpart (Therneau, Atkinson, and Ripley 2019) for this classification tree analysis. They used the rpart package default value for the minimum number of observations in a node, which is 20. This stopping rule stops splitting if a node does not contain at least 20 observations, preventing the model from overfitting on nodes with few observations.

In a CART model, each observation is run down the tree and through each respective split until it finally reaches a terminal node, where it is classified into one of two possible classes, in this case either enrolling in General Chemistry or not. The authors' final prediction took the proportion of positive cases (in the training data from prior years) in each terminal node and multiplied those proportions by the number of students (in the new data from the year to be predicted) that ended up in each respective terminal node to arrive at the predicted enrollment in General Chemistry. Again, like the flowchart method, the authors created training data by year, but instead of using the variables of prerequisite met date, they used the number of days from the matriculation period. The authors also divided the model

training process into sets based on the number of prior cohorts. For example, in the analysis, to predict 2018 enrollment, the authors trained the model separately on 2015, 2016, and 2017 data. The final enrollment prediction was then an ensemble average of these three trees fit for each prior year.

The proposed tree-based algorithm utilized individual observations to predict the chance that a student would enroll in a course, aggregated those predictions, and used that aggregation as a prediction for overall course enrollment. Instead of using a traditional prediction methodology where new observations are individually run down the tree and given a binary indicator of whether or not they fall in a class, SDSU's model summarized the tree terminal nodes as proportions, counted how many observations from the prediction year fell in to each terminal node, and then multiplied this count for each terminal node by its respective proportion in order to arrive at a final prediction number. In contrast to the flowchart method, which only takes into consideration limited student academic information at the time of enrollment, tree-based methods are capable of considering a much wider variety of data. In this study, the tree-based methods used demographic information, academic performance prior to entering university, as well as other covariates (*see* Appendix A, Table 4, on page 24) in order to create a more holistic system of prediction for course enrollment.

Results

The decision trees trained from fall 2017 and fall 2018 can be seen in Figures 1 (on page 19) and 2 (on page 20). Each node shows the number of students in the node (bottom number) and the probability they would enroll in the General Chemistry course (top number). Other than the terminal nodes, below each node is the splitting rule characterized by the variable and threshold for sending students down the tree to the left or right. The decision tree trained from different years' data was slightly different, but the main factors that the trees considered were similar: student academic standing, if and when the student had previously taken General Chemistry, and how many days before the target semester the student met the prerequisites for General Chemistry being the major determinants of enrollment probability. SAT comprehensive score, high school graduation year, and entry term also appeared as split rules in the trees. The detailed structures of the trees trained from fall 2015 and fall 2016 are shown in Appendix A, Figures 6 and 7 (on pages 27 and 28). The decision tree method found that recent high school graduates are more likely than others to enroll in General Chemistry. The variable abbreviations shown in the decision tree outputs can be referenced in Appendix A, Table 4 (on page 24). The decision tree method predicted 335 students would enroll for fall 2019, the actual value being 396 students enrolled. The decision tree error rate for fall 2019 was 15.5 percent, which was a slight improvement compared to the error rate of 21.2 percent for the conditional probability analysis. Similarly, the decision tree error rate was smaller when predicting fall 2018. The decision tree method predicted that 656 students would have taken General Chemistry, compared to the actual enrollment of 650 students, resulting in an error rate of 0.8 percent.

The decision tree-based method was more accurate than the original flowchart-based method. The final predictions and associated error rates can be seen in Table 3 (on page 21). The most likely reason behind the improvement was the introduction of demographic information and academic performance prior to enrolling at SDSU. In this way, the decision tree was able to use all available information to build a better predictive model of the conditional probabilities.

Random Forest Analysis

Methods

Random forest is a popular machine learning algorithm that uses decision trees in an ensemble fashion to improve predictive accuracy. As a further step in the analysis, a random forest algorithm was implemented to investigate its predictive capabilities in this situation.

In much the same way that CART increases the level of model complexity compared to a conditional probability analysis, random forest builds upon the complex-

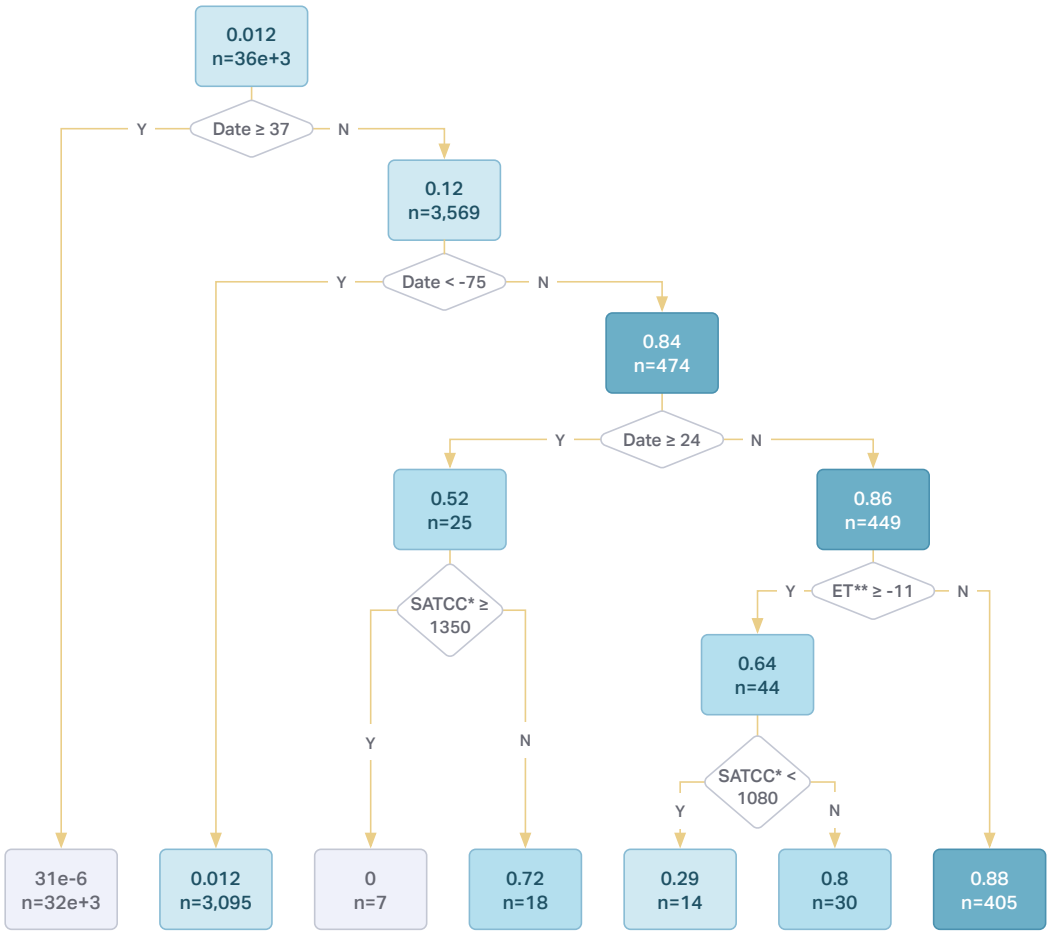

FIGURE 1 ➤ Classification Tree Analysis: Decision Tree Trained From 2017 Data

* SATCC = SAT_CompConv; ** ET = EntryTerm

ity found in CART by using an aggregate of multiple decision trees instead of a single decision tree. The algorithm takes an ensemble average of the results from the multiple trees in the forest of trees in order to form a final prediction. In contrast to the tree building method utilized by CART, which attempts to utilize all possible information at every step, random forest randomizes the model building process in a number of ways in order to produce a set of varied individual trees.

First, instead of using the entire sample to grow a tree, random forest uses only a random subset of the original data in a process called bagging: given an original sample size *n*, a training set is produced by randomly sampling with replacement from the original sample *n* times. Second, at each node split, random forest randomly samples from the feature space. Recall that CART, on the other hand, considers every possible predictor to determine the split rule for a given node. A recommended default number of predictors (termed "mtry"), which the authors used in their analysis, is the square root of the number of predictors (Breiman 2001). Third, random forest does not include a pruning step. The idea is to grow a forest of over-fit trees. But by randomly choosing the sample for each tree and the predictors over which to split at each node, the authors produced a diverse set of trees over which an ensemble prediction may perform, potentially much better than a single decision tree.

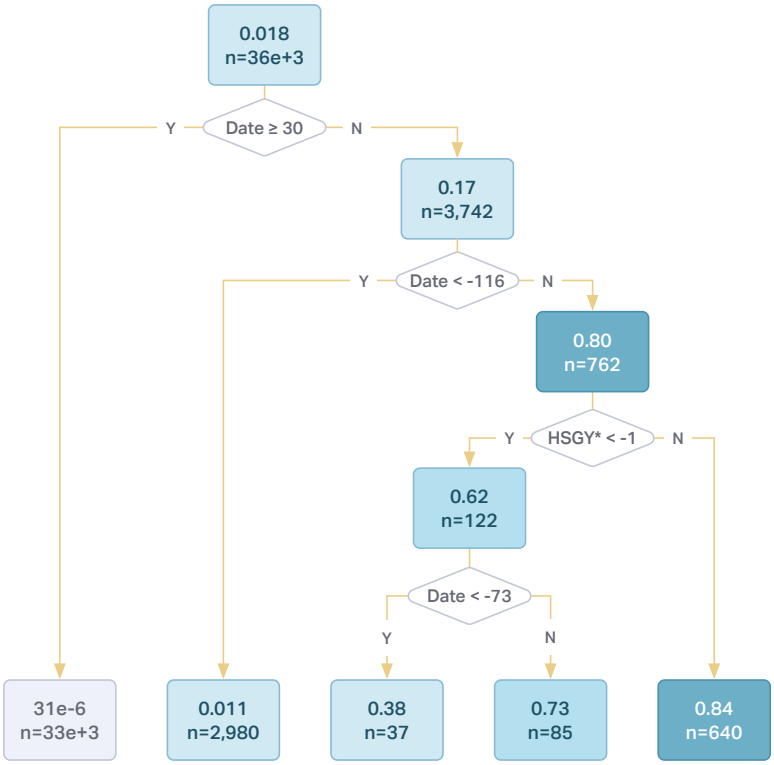

FIGURE 2 ➤ Decision Tree Trained from 2018 Data

* HSGY = HS_GradYr

The stopping criteria of each tree is that the samples in the node have the same responses, have the same features, or meet the minimal node size or maximum tree depth (Breiman 2001). For this analysis, concerning the large sample size of the data, a forest of 1,000 trees were grown. The Gini index was used as the split rule and a maximum number of splits of ten was used for each tree.

This analysis used the randomForestSRC (Ishwaran and Kogalur 2020) R package because it could use proportions as predictions from each decision tree. In contrast, the more popular R package randomForest (Breiman and Cutler 2018) combines binary predictions when producing the ensemble, which is less accurate (Malley, *et al*. 2012).

Similar to the conditional probability analysis and CART methods of the previous sections, the authors trained the final model as an ensemble of three models from the prior three years of data. Specifically, 2015, 2016, 2017 cohorts were used to predict 2018 enrollment numbers, and 2016, 2017, 2018 cohorts were used to predict 2019 enrollment numbers.

Results

The actual General Chemistry enrollment for fall 2019 was 396 students, and the random forest predicted enrollment to be 394 students, giving an error rate of 0.8 percent, a substantial improvement over both the flowchart method and decision tree method. The actual General Chemistry enrollment for fall 2018 was 650, and the random forest predicted enrollment to be 658, giving an error rate of 1.3 percent, similar to the decision tree method. A comparison of the performances is listed in Table 3 (on page 21).

A nice feature of the random forest approach is that it can also identify which variables are most import-

TABLE 3 ➤ Actual vs. Predicted Enrollment, Fall 2018 and Fall 2019

Term	Actual Enrollment	Predicted Enrollment, by Method					
		Conditional Probability Analysis		Classification and Regression Tree		Random Forest	
		n	Error (%)	n	Error (%)	n	Error (%)
Fall 2018	650	627	3.5	655	0.8	658	1.3
Fall 2019	396	312	21.2	335	15.5	393	0.8

ant for predicting enrollment (Breiman 2001). These variable importance values are found for each variable by measuring the difference in prediction accuracy between the original data and a randomly scrambled version of the variable. The top fifteen most important variables are shown in Figure 3. The most important variables were prerequisite met date, entry term, freshman status, whether the class had been taken before, and high school graduation year. Different from the flowchart method and decision tree method, random forest used more variables for prediction, as cross referenced in Figure 1 (on page 19) and Table 1 (on page 15). Although some of the variables were not as important as, say the top five in the variable importance plot of Figure 3 (on page 22), the random forest model took a lot of other variables into consideration due to its much more complicated structure.

Conclusions and Discussion

This paper aimed to improve upon existing course enrollment prediction models by applying two tree-based algorithms. The authors found that the proposed decision tree approach was able to improve upon the current state-of-the-art conditional probability analysis slightly, and the proposed random forest model was able to further improve upon both methods. In Figure 4, the enrollments over the past few years are visualized. From these statistics, it is clear that the enrollment numbers generally were not static, with fall 2019 having a much smaller enrollment number than previous years. Because of this variability, the three methods predicted this change in enrollment to varying degrees. Although all three predictions were respectably reasonable compared to an initial prediction based on the one-, two-, or three-year average of previous enrollment, both tree-based methods were more accurate than the conditional probability analysis.

The comparison of the three methods' performance is shown in Table 3 (on page 21). The decision tree was able to give further insights into the relevance of each variable in predicting whether or not a student would enroll in General Chemistry. The CART algorithm searched through each variable and found the variable that was able to create the most homogeneous child nodes. The conditional probability analysis split students by transfer status, prerequisite met term, whether the course is a major prerequisite, and whether the students had taken General Chemistry before. Whereas, the decision tree method did not consider major prerequisite as an important split, it considered SAT comprehensive score, high school graduation year, entry term, and ethnicity as important variables. The decision tree, thus, improved the conditional probability analysis by measuring which variables were able to identify more homogenous subgroups through the algorithm. Analogous to the decision tree, random forest identified prerequisite met date, entry term, freshman status, whether General Chemistry had been taken before, and high school graduation year as the most important variables. Random forest also incorporated SAT comprehensive score, though to a lesser degree. Due to the random forest's much more complicated model structure, it captured year-to-year variation and resulted in an improved prediction.

These results advanced upon the insights gained from the conditional probability analysis. The flowchart

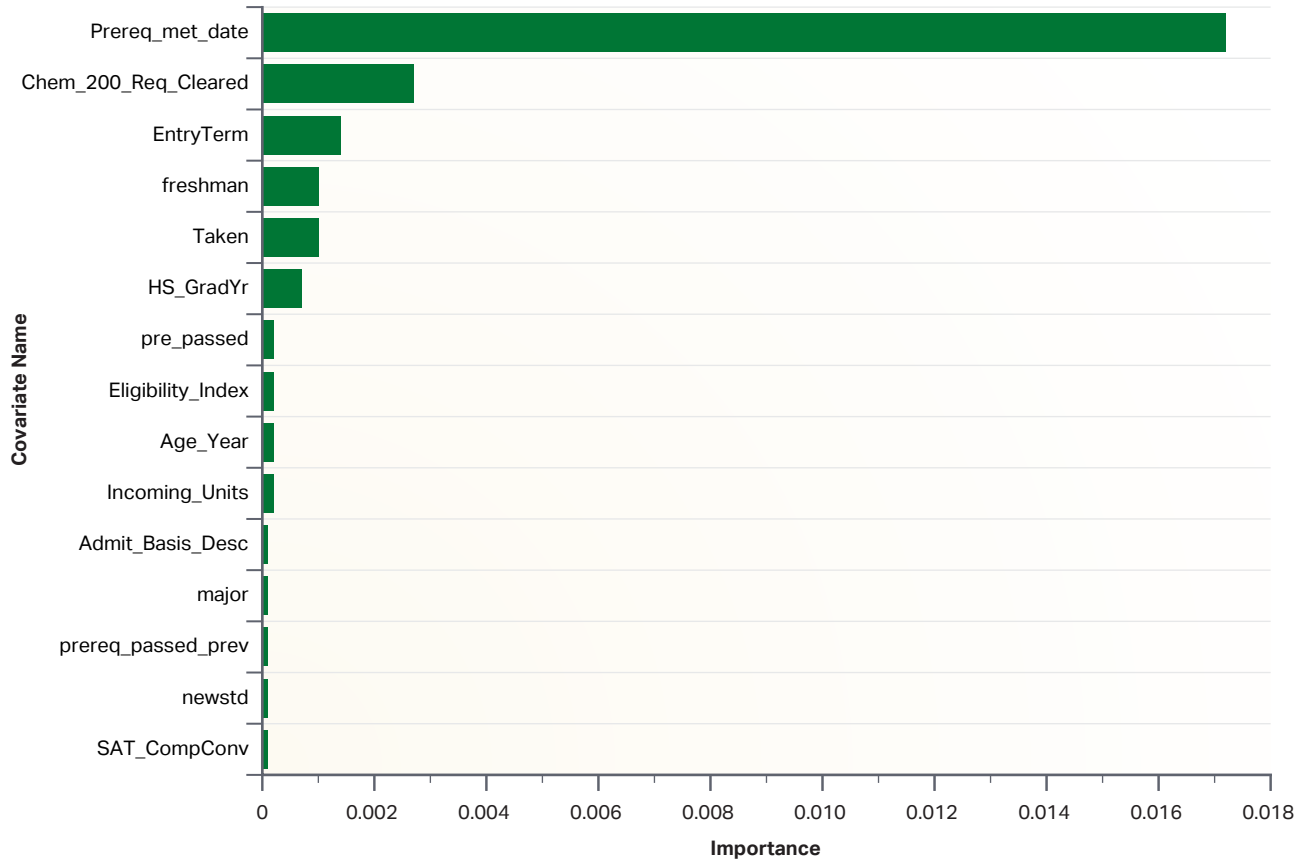

FIGURE 3 ➤ Variable Importance for Random Forest Application

posited several categories that could have influenced whether or not a student would enroll in a particular course. The decision tree model showed that some of the categories presented in the conditional probability model either were not necessary or, on the other hand, were more important than previously thought. In particular, the model showed that, in the presence of student academic background, there was no need to make a separate distinction for students whose major requires General Chemistry as a prerequisite. It also showed that prior academic information, especially high school graduation year and SAT comprehensive score, could influence how soon a student takes General Chemistry upon entering the university.

Accurately predicting course enrollment for future terms is a critical task for any university. Machine learning methods are powerful tools to aid in university planning and minimize the unnecessary administrative costs associated with allocating additional or fewer seats for students than was originally planned. The proposed tree-based approaches pinpoint areas of interest for future studies. The tree-based algorithms showed that prior GPA, age, and high school graduation time are useful in course prediction for General Chemistry in addition to the previously theorized metrics of student performance in university.

Enrollment patterns of other courses may differ from that of General Chemistry. The authors' methods

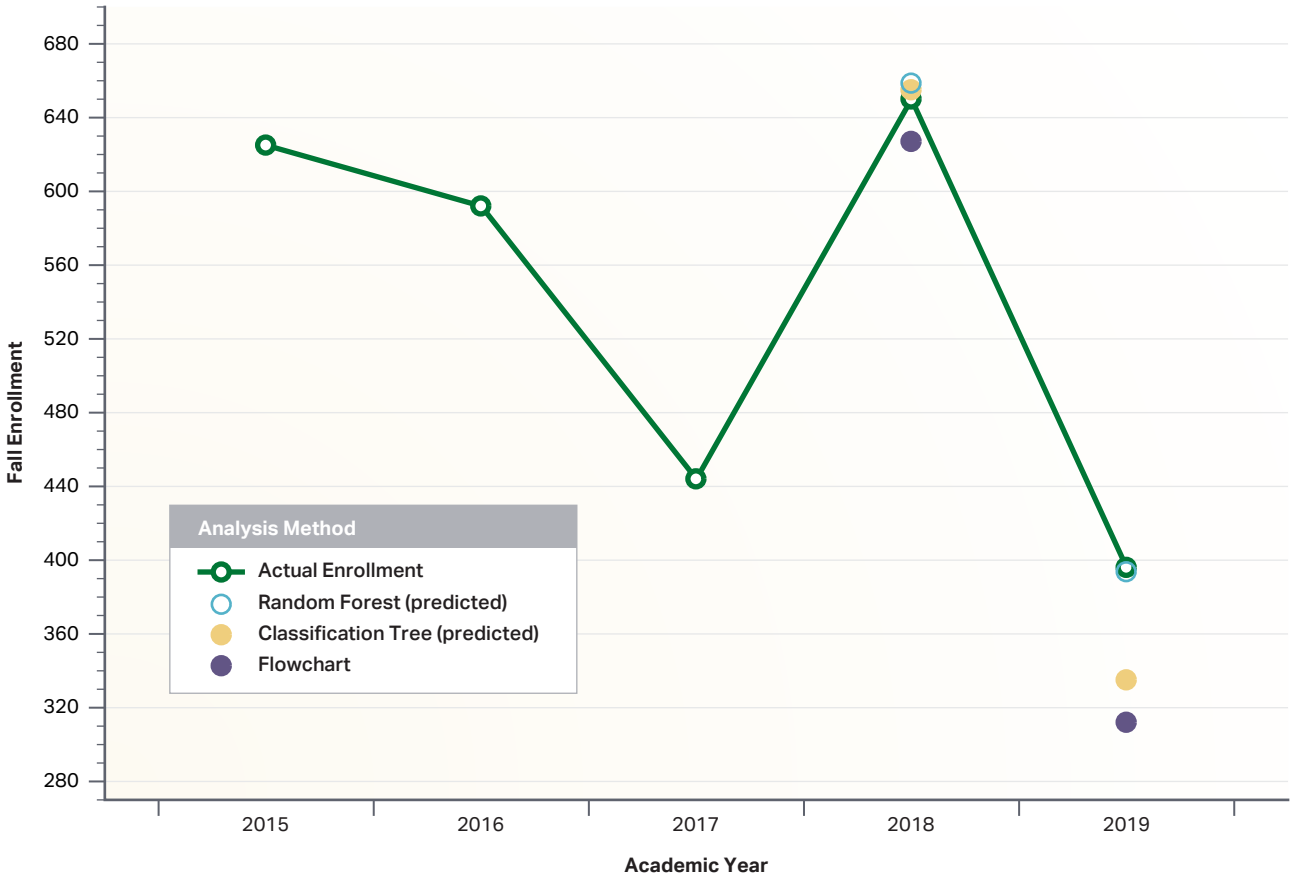

FIGURE 4 ➤ Enrollment Numbers Visualized

may analogously be applied to each individual course to predict future enrollment. Further research is currently under way to explore whether or not the patterns of enrollment seen here hold true for other large enrollment courses with hundreds of students per semester, especially courses where there are no prerequisites, as these courses may be harder to predict.

As presented in Appendix A, Table 4 (on page 24), many variables were used in these predictive models including potentially sensitive data such as ethnicity, disability, and parent education. The authors note that the data was completely de-identified before analysis. In addition, variables were used only to aid in prediction accuracy and not for enhancing nor limiting student access to any class.

Appendix A: Supplemental Reference Material

TABLE 4 ➤ Variables Used for Course Enrollment Predictions

Variable Name	Variable Description	Value(s)/Value Type	Summary Statistics[1]		
			Result	Low	High
University Data					
EntryTerm	Entry Term	Fall, Spring	–	–	–
STDLVL	Student Status upon Admission	Freshman, Transfer, Readmit, Non-Degree Seeking	–	–	–
SMAJOR	Student Major	206 Different Majors	–	–	–
Chem_200_Req_Cleared*[2]	Prerequisite Met Status	Yes, No	0.1080	–	–
Attempt_1_Period[2]	Semester of First Enrollment	Fall, Spring, Summer	–	–	–
Attempt_1_Grade	Final Grade from First Enrollment[2]	A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, W	–	–	–
Attempt_2_Period	Semester of Second Enrollment[2]	Fall, Spring, Summer	–	–	–
Attempt_2_Grade	Final Grade from Second Enrollment[2]	A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, W	–	–	–
Attempt_3_Period	Semester of Third Enrollment[2]	Fall, Spring, Summer	–	–	–
Attempt_3_Grade	Final Grade from Third Enrollment[2]	A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, W	–	–	–
Date	Date Prerequisites Were Fulfilled[2]	MM-DD-YYYY	–	–	–
Study_Abroad_Ever_Desc*	Studied Abroad	Yes, No	0.0386	–	–
Honors_Desc*	Honors Student	Yes, No	0.0309	–	–
Compact_Desc*	Compact Scholar	Yes, No	0.0638	–	–
FAST_Desc*	FAST Program Participant	Yes, No	0.0230	–	–
Athlete_Desc*	Student Athlete	Yes, No	0.0177	–	–
Summer_Bridge_Desc*	EOP Summer Bridge Participant	Yes, No	0.0115	–	–
Demographic Data					
Age_Year	Age in Year	Integer	18	16	78
URM*	Under Represented Minority	Yes, No	0.3330	–	–

[1] For categorical variables, the proportion of positive cases
[2] Reference course: General Chemistry (CHEM 200)
* Indicator Variable with values of 1 (yes) and 0 (no)

TABLE 4 ➤ Variables Used for Course Enrollment Predictions

Variable Name	Variable Description	Value(s)/Value Type	Summary Statistics[1]		
			Result	Low	High
Students_of_Color*	Student of Color	Yes, No	0.0426	–	–
Hispanic*	Hispanic	Yes, No	0.2920	–	–
Male*	Gender	Yes, No	0.4240	–	–
FirstGen_NCES_Desc*	First Generation College Student Status	Yes, No	0.1880	–	–
		No High School	0.0682	–	–
		Some High School	0.0595	–	–
		High School Graduate	0.1590	–	–
Parent1_Edu_Desc	Parent Education (1)	Some College	0.1510	–	–
Parent2_Edu_Desc	Parent Education (2)	2-Year College Graduate	0.0630	–	–
		4-Year College Graduate	0.2550	–	–
		Postgraduate	0.1290	–	–
		Unknown	–	–	–
Military_Desc*	Military Status	Yes, No	0.0833	–	–
Disability_Desc*	Disability Status	Yes, No	–	–	–
Admissions Data					
In_Service_Area_Desc*	Local Area Admittance	Yes, No	0.4580	–	–
HS_GradYr	High School Graduate Year	Integer	–	–	–
CSU_Eligible_Desc*	CSU Eligibility Status	Yes, No	0.9380	–	–
SAT_Comp	SAT Composite Score	Integer	1140	420	1590
SAT_Math	SAT Math Score	Integer	570	200	800
SAT_Ver	SAT Verbal Score	Integer	570	200	800
SAT_CompConv	SAT Composite Score (converted from ACT score)	Integer	–	–	–

[1] For categorical variables, the proportion of positive cases
[2] Reference course: General Chemistry (CHEM 200)
* Indicator Variable with values of 1 (yes) and 0 (no)

TABLE 4 ➤ Variables Used for Course Enrollment Predictions

Variable Name	Variable Description	Value(s)/Value Type	Summary Statistics[1]		
			Result	Low	High
Incoming_GPA	Incoming GPA	Decimal	3.59	2.00	4.50
Incoming_Units	Incoming Units Taken	Integer	21	0	309
HS_MathProficient*	High School Math Proficiency (as determined by performance in CSU ELM exam)	Yes, No	0.8960	–	–
Fall_MathProficient*	Fall Matriculation Math Proficiency (as determined by performance in CSU ELM exam)	Yes, No	0.9010	–	–
HS_EnglishProficient*	High School English Proficiency (as determined by performance in CSU EPT exam)	Yes, No	0.8990	–	–
Fall_EnglishProficient*	Fall Matriculation English Proficiency (as determined by performance in CSU ELM exam)	Yes, No	0.9100	–	–
Major*	Student's major requires course[2] as prerequisite	Yes, No	–	–	–
Prereq1*	Fulfilled Prerequisites 1 Term Ago	Yes, No	–	–	–
Prereq2*	Fulfilled Prerequisites 2 Terms Ago	Yes, No	–	–	–
Class1*	Course[2] Taken Last Term	Yes, No	–	–	–
Class2*	Course[2] Taken Two Terms Ago	Yes, No	–	–	–
Freshman*	Freshman	Yes, No	0.5730	–	–
Transfer*	Transfer Student	Yes, No	0.3580	–	–
Other*	Undergraduate Readmit or Non-Degree Seeking	Yes, No	0.0651	–	–
Taken	Course[2] Taken	Yes, No	0.0979	–	–
Eligibility_Index	CSU Eligibility Index		4128	420	5070

[1] For categorical variables, the proportion of positive cases
[2] Reference course: General Chemistry (CHEM 200)
* Indicator Variable with values of 1 (yes) and 0 (no)

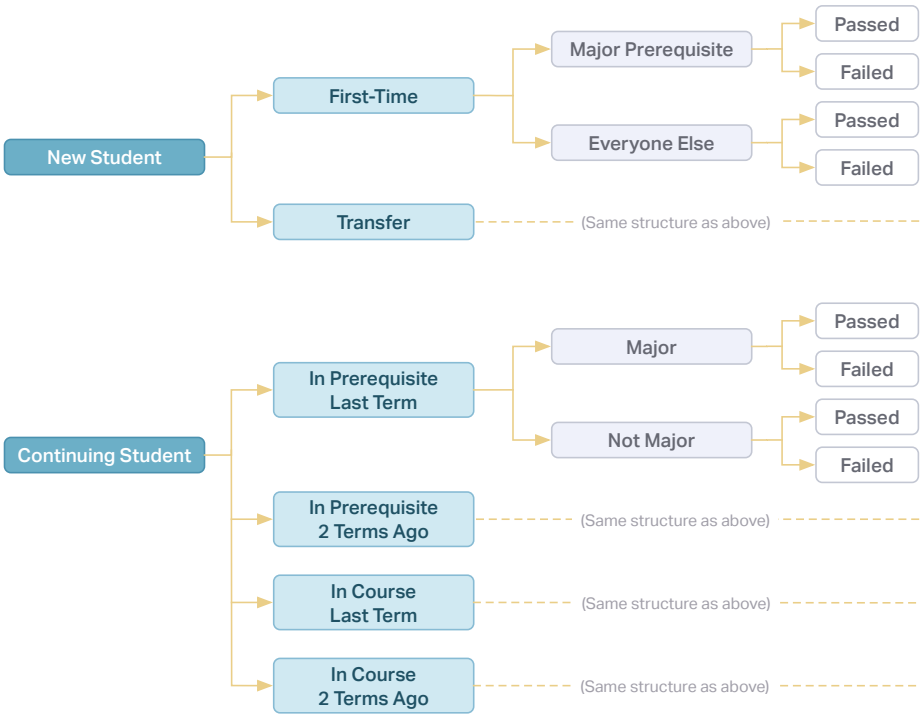

FIGURE 5 ➤ Flowchart Structure for the Conditional Probability Analysis

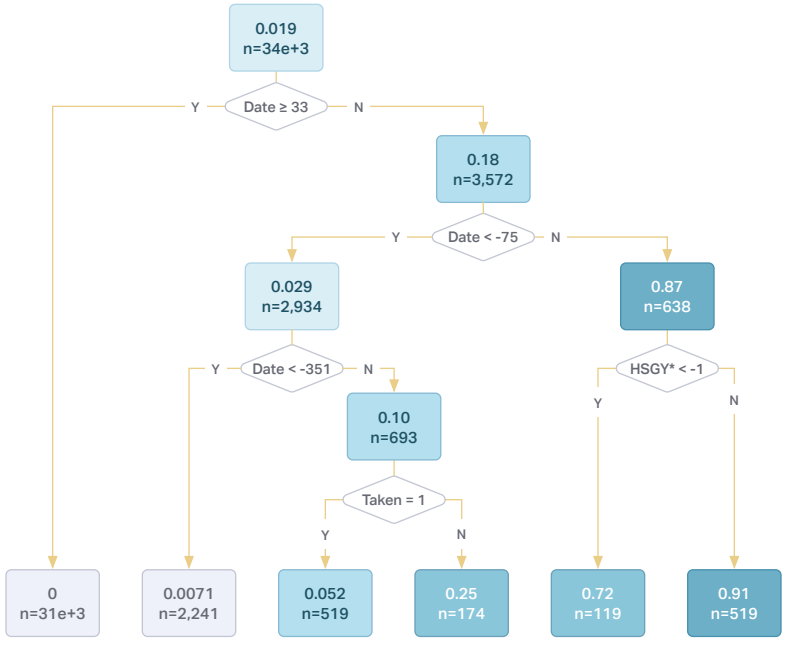

FIGURE 6 ➤ Classification Tree Analysis: Tree Trained Using 2015 Data

* HSGY = HS_GradYr

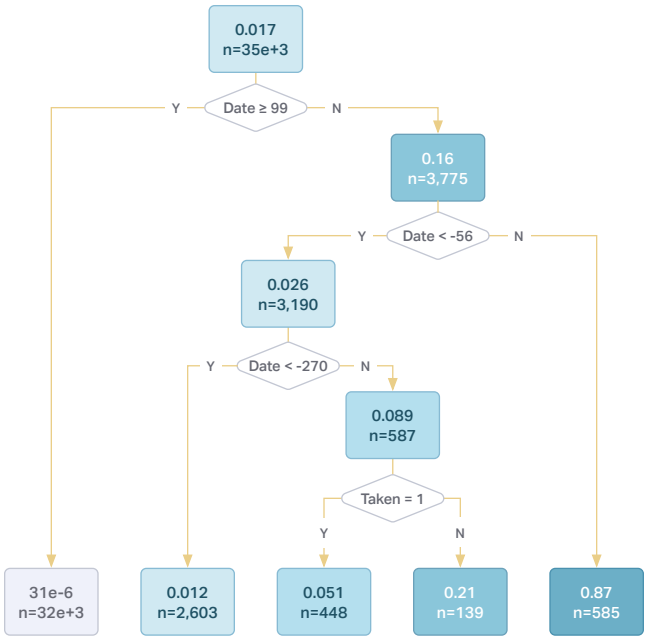

FIGURE 7 ➤ Classification Tree Analysis: Tree Trained Using 2016 Data

ACKNOWLEDGMENT

This research was supported in part by the National Science Foundation grant 1633130.

References

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Boca Raton, FL: Taylor & Francis Group.

Kraft, C. R., and J. P. Jarvis. 2005. *An Adaptive Model for Predicting Course Enrollment, Master's* (thesis). Clemson University, Clemson, SC.

Breiman, L. 2001. Random forests. *Machine Learning*. 45: 5–32.

University of California. 1999. *Educating the Next Generation of Californians in a Research University Context: University of California Graduate and Undergraduate Enrollment Planning Through 2010*. CA: Author. Available at: <ucop.edu/institutional-research-academic-planning/_files/enrollplan99.pdf>.

University of Virginia. 1999. *Methodology for School Enrollment Projections*. Charlottesville, VA: Demographics Research Group, Weldon Cooper Center for Public Service.

Malley, J., J. Kruppa, A. Dasgupta, K. Malley, and A. Ziegler. 2012. Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*. 51: 74–81.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. Classification. In *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

Wikipedia: The Free Encyclopedia. 2020. Decision tree. April 18. San Francisco: Wikimedia Foundation.

Breiman, L., and A. Cutler. 2018. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. Available at: <cran.r-project.org/package=randomForest>.

Ishwaran, H., and U. B. Kogalur. 2020. *RandomForestSRC Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. Available at: <cran.r-project.org/package=randomForestSRC>.

Therneau, T., B. Atkinson, and B. Ripley. 2019. *rpart: Recursive Partitioning and Regression Trees*. Available at: <cran.r-project.org/web/packages/rpart>.

Soltys, M., H. Dang, H., G. R. Reilly, and K. Soltys. 2021. Enrollment predictions with machine learning. *Strategic Enrollment Management Quarterly*. 9(2): 11–18.

About the Authors

Lucy Shao

Lucy Shao is a Ph.D. student in Biostatistics at the University of California, San Diego. Previously, she worked at Analytic Studies & Institutional Research at San Diego State University on student learning and institutional operation. Currently, her research focuses on causal inference and high-dimensional statistics. Lucy received an M.S. in Statistics from UC San Diego and a B.S. in Applied Mathematics from UCLA.

Martin Ieong

Martin Ieong is a Data Scientist at Revolution Medicines, where he analyzes genomic and bioinformatics data in the field of cancer research. He previously worked in the software development industry, providing technical assistance and statistical knowledge to create data science solutions. Martin received an M.S. in Statistics from San Diego State University, and a B.A. from UC San Diego in Linguistics.

Richard A. Levine

Richard A. Levine is Professor of Statistics at San Diego State University and Faculty Advisor overseeing the Statistical Modeling Group in SDSU Analytic Studies and Institutional Research. He is former Chair of the SDSU Department of Mathematics and Statistics and past Editor of the *Journal of Computational and Graphical Statistics*. He is Associate Editor for Statistics of the Notices of the American Mathematical Society and is a fellow of the American Statistical Association. Professor Levine received his Ph.D. in Statistics from Cornell University.

Jeanne Stronach

Jeanne Stronach is an experienced and effective higher education leader with 20 years of experience in institutional research. Specializing in successful team-building with internal and external partners, innovative resource management and effective strategic planning. Achievements include building self-service visualizations to support data-informed decision-making as well as spearheading a cross-divisional Data Champions program to build data community and boost data literacy. Jeanne earned her B.A. at UC San Diego and her M.A. at DePaul University.

Juanjuan Fan

Juanjuan Fan, Ph.D., is a Professor of Statistics and Data Science in the Department of Mathematics and Statistics, and serves as a Faculty Advisor at the Analytic Studies & Institutional Research (ASIR), at San Diego State University. Her research interests include survival analysis, decision trees and random forests, and observational study data. Working with her students and collaborators, she has published many papers assessing student success studies and solving various problems in educational data mining. Professor Fan received her Ph.D. in Biostatistics from the University of Washington.