American Journal of Pharmaceutical Education **2018; 82 (7) Article 6980.**

COMMENTARY

The Application of Classification Trees to Pharmacy School Admissions

Samuel C. Karpen, PhD,[a] Steve C. Ellis, MS[b]

[a] College of Veterinary Medicine, University of Georgia, Athens, Georgia
[b] Bill Gatton College of Pharmacy, East Tennessee State University, Johnson City, Tennessee

Submitted January 26, 2018; accepted May 8, 2018; published September 2018.

In recent years, the American Association of Colleges of Pharmacy (AACP) has encouraged the application of big data analytic techniques to pharmaceutical education. Indeed, the 2013-2014 Academic Affairs Committee Report included a "Learning Analytics in Pharmacy Education" section that reviewed the potential benefits of adopting big data techniques.[1] Likewise, the 2014-2015 Argus Commission Report discussed uses for big data analytics in the classroom, practice, and admissions.[2] While both of these reports were thorough, neither discussed specific analytic techniques. Consequently, this commentary will introduce classification trees, with a particular emphasis on their use in admission. With electronic applications, pharmacy schools and colleges now have access to detailed applicant records containing thousands of observations. With declining applications nationwide, admissions analytics may be more important than ever.[3]

Keywords: predictive analytics, admissions, decision tree

INTRODUCTION

Classification trees attempt to separate data into maximally homogenous groups in terms of an outcome of interest. For example, the restaurateur depicted in Figure 1 wants to know when customers will wait to be seated at his restaurant, as opposed to leaving to go elsewhere.[4] Accordingly, he/she recorded information about 12 customers. The restaurateur found that the number of patrons in the restaurant best differentiated customers in terms of waiting. When there were no patrons in the restaurant, no one waited; when there were some patrons, everyone waited; and when the restaurant was full, two customers out of six waited. Restaurant type, however, was not as informative as crowd size, because as many people waited as did not wait at each type of restaurant: at the French and Italian restaurants, one waited and the other did not, while at the Thai and burger restaurants, two waited, and two did not. In other words, no clear distinction between waiting and not waiting could be drawn. Figure 1 also introduces another attribute: whether or not the customers were hungry. None of the satiated customers in the "full" group waited, but half of the hungry ones did. The next step would be to determine which variable(s) split the hungry customers into homogenous groups of waiters and non-waiters.

Corresponding Author: Samuel C. Karpen, College of Veterinary Medicine, University of Georgia, 501 D.W. Brooks Dr., Athens, GA 30602. Tel: 423-439-6883. Fax: 706-542-5460. E-mail: sckarpen@uga.edu

Classification trees' clearest advantage is their interpretability. By portraying the analysis as a series of binary classifications, they provide a straightforward graphic for non-statisticians, unlike logistic regression, which can be difficult to interpret when many variables are included. Classification trees, however, can capture complex relationships without the added interpretational difficulty. Furthermore, when the relationship between one's predictors and the dependent variable is markedly non-linear, classification trees tend to yield much more accurate predictions than logistic regression.[5-7]

Despite their advantages, classification trees are prone to overfitting.[5-7] That is, classification trees are apt to model both the relationship between the variables of interest and that data's idiosyncrasies. For example, if a researcher is trying to model the relationship between students' demographic information and first-year PharmD grades, and the data contains four students from New York with high P1 GPAs, a categorization tree may indicate that living in New York is an important predictor of P1 GPA when it is just a quirk of that particular data. As a consequence, the tree algorithm will not make accurate predictions when it is applied to new data with new idiosyncrasies.

Because of this limitation, it is important to build one's classification tree on an initial dataset (the training set) and then apply it to a new dataset (the test set) to determine how well it generalizes before deploying it. Traditionally, the analyst randomly splits the data 80-20, using 80% of the data for the training set and 20% for the test set.[7] Analysts may also choose to use k-fold cross-validation

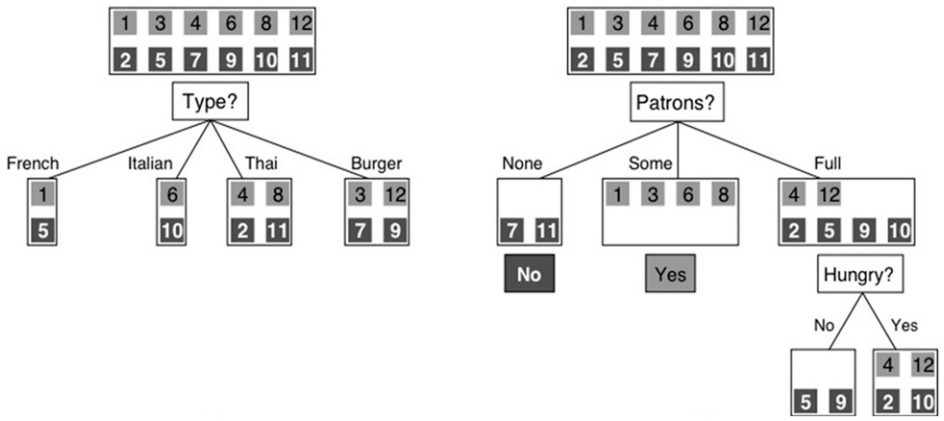

Figure 1. Number of Patrons in a Restaurant Does a Better Job of Splitting Potential Customers into Groups of Waiters and Non-waiters Than the Type of Food Served.

to increase generalizability. To further improve a tree's generalizability, researchers can use random forests. To generate a random forest, the algorithm randomly selects a subset of variables from all variables and generates a classification tree based on each subset. For example, pharmacy programs may be interested in predicting whether an applicant will progress normally. When using common variables such as applicant age, pre-pharmacy science GPA, biology PCAT, and applicant's highest degree, the algorithm may select PCAT and pre-pharmacy science GPA, and generate a tree based on only those variables. In addition to randomly selecting variables, the algorithm also randomly selects cases, such that the randomly selected variables are only used to build a tree on a subset of the cases. The predictions of the randomly generated trees are then combined to produce a more generalizable estimate. Randomly generating and combining trees improves generalizability because each tree only produces estimates based on part of the data; hence there is a great deal of variance in the trees. Since the trees are unlikely to be dependent on one another (because they are generated from a different set of variables) some will overestimate the outcome and others will underestimate the outcome; thus, their aggregate estimate should be an accurate representation of the outcome in the population. Consequently, this tree should generalize well. Analysts, however, should still use training and test datasets or k-fold cross-validation with random forests, and regression for that matter.

Classification trees also suffer from non-statistical limitations in that they require advanced statistical training and programming ability. Most assessment staff do not have this skillset. Additionally, classification trees require very large datasets for optimal performance – ideally 3,000+ cases. Most pharmacy colleges do not have that many student records. If a college is large, well-established, and has the necessary expertise, however, classification trees can be valuable tools.

Example

The classification tree in Figure 2 displays the relationship between enrollment status and several variables available in pharmacy school applications. The tree can be interpreted as follows: The zeroes and ones at the top of each box show the outcome for the majority of applicants in the group – every group in which at least 50% of the applicants enrolled is labeled with a one. In the first box, which represents the data before any splitting, there is a zero in the top position because most applicants who were extended an offer did not enroll. The proportion indicates the percent of people in each group who enrolled. In the first box, the 0.48 means that 48% of students who were extended an offer eventually enrolled. Finally, the percentage indicates the percent of all applicants who are in that group. This number is 100% in the first box because it represents all of the data before it is split into groups.

The box in the lower left corner of the tree represents the information in which a school would be most interested (ie, the type of applicant most likely to enroll). As shown, 36% of applicants with at least one bachelor's degree holding parent (ParentEd >=.5 is "yes"), who attended a four-year college (FourYear >=.05 is "yes") and earned a pre-pharmacy GPA above 2.6 (PrePharmGPA>=2.6 is "yes") enrolled. Collectively, this group represented 50% of all applicants. It should be noted that if a variable is coded as 1 and 0, like parent education (ParentEd), the software that generated the tree denotes splits by less than or greater than 0.5. Less than 0.5 is "the group coded as 0" whereas greater than 0.5 is "the group coded as 1." In the case of the classification tree in Figure 2 FourYear >=0.5 is "no" means that the applicant did not attend a four-year institution because "four year institution" was coded as 1.

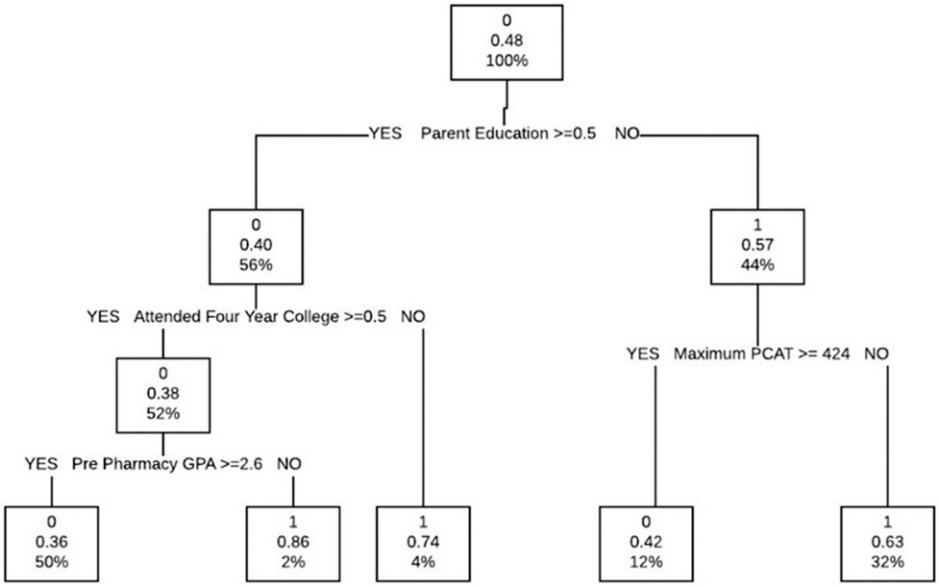

Figure 2. The tree displays an overall predication (0=less than 50% of the group enrolled, 1=more than 50% of the group enrolled), the proportion in each group who enrolled, and the percent of the total sample constituted by a given group. For example, 57% of applicants whose parents did not have a four-year degree enrolled (ParentEd >=no). This group represented 44% of all applicants.

When applied to the test set, the example tree predicted 70.6% of the cases accurately. To further improve the tree's accuracy, researchers could include additional attributes (eg, whether a student has a competing offer) or use a random forest. Accordingly, the random forest on the training data predicted 77% of the cases in the test dataset correctly. If the leadership (deans, associate/assistant deans, etc.) is comfortable with 77%, then the tree will be ready for use in admission decisions.

CONCLUSION

Categorization trees can be used to facilitate a variety of decisions across the pharmacy education spectrum. One potential use is developing a predictive model to aid in admissions decision as described by Muratov and colleagues.[8] In this project, the authors developed a pharmacy school performance predictive model to identify applicants for interviews based on their likelihood of academic success. While the Muratov and colleagues' model relies primarily on cognitive factors, such admissions models also can include non-cognitive factors that may expand the utility of the model for an institution. For example, to aid in identifying applicants who fit the school's mission or, as illustrated in this paper, to identify applicants most likely to enroll, which in turn can inform the recruiting process. A potentially innovative use of categorization trees is in support of a school's assessment of student APPE readiness as addressed in Standard 24.3 of ACPE Standards 2016. Determining APPE readiness is challenging because readiness includes skills commonly referred to as soft skills; students' proficiency in these areas (or lack thereof) often does not manifest until they have begun APPEs. Examining data from relevant courses and experiences during the didactic years may allow for earlier identification of students who may need additional preparation prior to entering the APPE portion of the curriculum. Indeed, classification trees, and techniques like them, may be useful in any situation where prediction is required.

Classification trees are also very useful for exploratory analyses. For example, if an institution wishes to identify the demographic characteristics of struggling students, a classification tree could be a valuable tool. To this end, the authors of this commentary built a classification tree to visualize the relationship between application information: PCAT scores, pre-pharmacy math/science GPA, whether the student attended a four year institution, whether the student was from our region, age, and gender, and whether the student failed at least one class during their first year. While the results were not entirely surprising in that students with low pre-pharmacy math/science GPAs and low biology PCAT scores were more likely to fail a course than those with higher scores, the tree also indicated that students outside of the traditional age range (less than 22 or greater than 32) were prone to struggle. While this solitary tree should not be used for prediction, it uncovered a potential issue that may have otherwise gone unnoticed.

It is increasingly important that Doctor of Pharmacy programs develop methods for analyzing programmatic data in ways that facilitate decision making. Big data and associated analytics have already shown promise at universities

that are willing to invest in them. After using big data analytics to identify at-risk students (and assign them to appropriate interventions), Georgia State University saw a 6% increase in graduation rate over three years, a half-semester decrease in the amount of time required to graduate, and better performance in STEM courses by first-generation students.[9] While it was the school's intervention that ultimately increased student performance and retention, the statistical models showed them were to target their efforts. Similarly, pharmacy programs can use tools like classification tree analysis to gain insight into their applicant pools and student performance in ways that go beyond traditional analyses.

ACKNOWLEDGMENTS

The authors thank David Roane for his comments related to this work.

REFERENCES

1. Cain J, Conway JM, DiVall MV, et al. Report of the 2013-2014 Academic Affairs Committee. *Am J Pharm Educ*. 2014;78(10): Article S23.
2. Baldwin JN, Bootman JL, Carter RA, et al. Pharmacy practice, education, and research in the era of big data: 2014-2015 Argus Commission report. *Am J Pharm Educ*. 2015;79(10):Article S26.
3. AACP Council of Deans administrative board meeting. Oral presentation at AACP interim meeting. February 25, 2017, Rio Grande, Puerto Rico.
4. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. Essex, UK: Pearson Education; 2014.
5. Breiman L. Random Forests. *Machine Learning*. 2001;45(1): 5-32.
6. Hayes, T Usami S, Jacobucci R, McArdle. Using Classification and Regression Trees (CART) and random forests to analyze attrition: results from two simulations. *Psychol Aging*. 2015;30(4): 911-929.
7. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*. 2009;14(4):323-348.
8. Muratov E, Lewis M, Fourches D, Tropsha A, Cox WC. Computer-assisted decision support for student admissions based on their predicted academic performance. *Am J Pharm Educ*. 2017; 81(3):Article 46.
9. Kamenetz A. How one university used big data to boost graduation rates. *NPRED*. October 30, 2016. http://www.npr.org/sections/ed/2016/10/30/499200614/how-one-university-used-big-data-to-boost-graduation-rates.