Use of Decision-Tree Approach for Identification of Dominant Predictive Factors in Standardized Visualization Testing

Jorge Rodriguez, PhD, MBA
Department of Engineering Design, Manufacturing, and Management Systems Western Michigan University

Luis G. Rodriguez-Velazquez,
PhD Department of Engineering
University of Wisconsin - Waukesha

Abstract

Predictive analytics is a subject that has become useful in forecasting behaviors and performances, which could be of benefit when attempting to predict the students' performance on a visualization standardized test. Predictive analytics uses a variety of algorithmic approaches, being decision-tree one of them, and an approach that has been recognized for its applicability and the fact that its outcomes can be represented graphically. Decision-tree is considered an approach that generates a model based on the probabilities extracted from the data being analyzed.

Some initial modeling using a small dataset has been reported, and results were obtained based on performance (i.e., minimum overall score on standardized visualization test – PSVT:R) and demographics (i.e., four characteristics were analyzed – status, gender, ethnicity, CAD experience). The objectives pursued for this report are twofold:

i) increasing the size of the dataset being utilized in the model building and validation phases, and compare the new results for performance predictions to the ones previously reported, ,and

ii) establishing predictive parameters based on grouping and trends of the performance data, in order to attempt to define common predicting factors.

The ultimate goal of these studies is to have objective information that can help in the definition of specific academic interventions in course content or in content delivery.

Introduction

The topic of predictive data analytics has received substantial attention in the recent past due in part to its potential to provide a competitive advantage in a globalized economy, which has resulted in the almost imperative need for focused or customized

services, thus deriving in this global trend of collecting and analyzing all kinds of data. Most of the attention and applications of this concept relate to consumer sciences, but the applicability of predictive data analytics has extended to processes and trends analysis, which has more direct relation to engineering and manufacturing. Data analytics is considered a generic term used to refer to a set of quantitative and qualitative approaches that are applied to provide the basis for some decision making (Big, 2017). Specific objectives that are being pursued when using data analytics are increase in productivity, additional business profit, or expected performance or behavior (Data, 2017).

Predictive data analytics is primarily utilized to establish an expected performance, specifically in academics besides the administrative tasks like enrollment and satisfaction of students, it was extensively used in technical applications, but not in pedagogical studies where the objective is to establish an expected academic performance or behavior, such as spatial visualization skills. There is a variety of tests that have been applied to measure spatial visualization skills of students (Strong 2002, Yue 2008), and there are numerous studies that have collected and analyzed information regarding demographics, spatial visualization skills, and academic performance (Prieto 2009, Sorby 1999). Of interest are studies where spatial visualization skills have been linked to abilities to do engineering and technology work, and subsequent studies that have provided a relationship between those skills of students and their performance in engineering courses, particularly for engineering graphics and design courses (Sorby, 2005). Similarly, there are reports that indicate the value in improving visualization skills when looking at the performance in learning in technology and engineering courses Koshevnikov, 2006), indicating improvement of such skills as the complexity of the problem increases (Titus 2009) which is the basis for looking at performance in a standardized test such as Purdue Spatial Visualization Test with Rotations (PSVT:R) (Guay, 1977).

This study reports on the application of a predictive data analytics approach to spatial visualization scores with the objective of establishing dominant predictive questions that define expected high performance. The data utilized in this study is from the PSVT:R. The goal is to have information that helps in directing interventions to be implemented for development of spatial visualization skills.

Methodology

Scores for each of the questions in the PSVT:R test where utilized as dataset. The test was administered to students taking introductory engineering graphics courses, and the results were collected for a previously reported study (Rodriguez, 2016a). This study

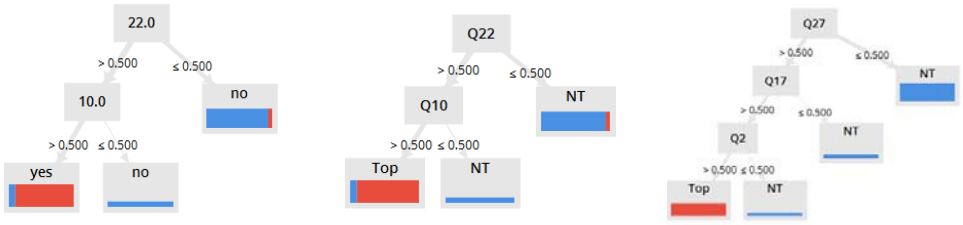
focuses on identifying the answers for each one of the 30 questions in the PSVT:R test and the final total score (maximum of 30) as potential dominant factors, no demographic data is utilized even though it was collected. Similarly, a new parameter is introduced, 'top performer,' which is used as defined as the prediction criterion, being top performer indicates that the total score in the test is equal or above a given value.

The software used in this study is RapidMiner, a commercially available data analytics software that has the option to analyze and visualize datasets applying different approaches, thus comparing results. Because the objective of this study s to identify dominant factors (i.e., questions) that predict high level of performance, the Decision-Tree approach has been applied. This approach has been identified has a good general purpose technique, with acceptable reliability in predictions, and it is a technique that provides graphical output that is very helpful in following the predictive model developed (Best, 2017). A decision-tree is a tree like collection of nodes that defines a decision on specific parameters to a class or an estimate of a numerical target value (i.e., final test score). Each node represents a splitting rule for one specific Attribute (i.e., answer to each test question). This approach reduces the error in an optimal way for the selected criterion (top performer) (RM, 2017).

Results

The dataset for this study consisted of 156 test records. As indicated before, this dataset was collected at two different institutions, and they have no statistically significant difference by being from two campuses (Rodriguez, 2016b). A total of 152 records were used for the machine learning stage where the prediction model is being built, the rest of the records were used for validation of the prediction model generated by the decision-tree approach.

The first objective is to increase the number of records in the dataset being used for model building. The dataset used here is almost 6 times the size of the dataset used in the previous pilot study (Rodriguez, 2018), and the previously reported result indicating that Q22 is the dominant factor when specifying a top score of 25 or higher (Figure 1). Interesting situation is that such dominant factor is not the same one when the top score is modified, which takes us to the second objective, and important issue is that only question answers are being used for the predictive model.



(a) T25, small n (b) T25, large n (c) T26, large n

Figure 1. Decision tree for two different datasets. Small n=27, large n = 152.

For the generation of predictive models to identify dominant factors, ten decision-tree models were generated for top scores ranging from 21 and above to 30. The dominant factor(s) for each case are reported in Table 1. As it can be observed, the primary dominant factor is not a single one for the tested range, it varies from Q23 at the lower end (scores 21 to 23), to Q30 for the high end (scores 29 and 30), with different primary dominant factor for the scores in between.

It is of interest to relate these results to previously reported results that extract dominant factors in the standardized test, in particular a study by Ernst 2017 where a factor analysis is performed, indicating the need to consider at least three principal components (factors) to have an acceptable level of inclusion of data variance. Performing a Factor Analysis (FA) on the dataset utilized in this study it renders similar result of requiring at least three factors to have an acceptable level of the data variance explained, as seen in the Scree graph in Figure 2. There is no match in terms of the specific test questions considered as principal components by the FA, and the ones identified as dominant predictive parameters, which is expected given the nature of the two studies, but indicating the need of performing a clustering approach to have better agreement (Farias 2017).

Table 1. Summary of Dominant Factors for Top Performers

Test Score+	21	22	23	24	25	26	27	28	29	30
% as Top Performer	73	73	73	50	46	35	23	12	8	4
1st Dominant Question	23	23	23	12	22	27	29	29	30	30
2nd Dominant Question	19	19	19	19	10	17	27			
3rd Dominant Question						2				

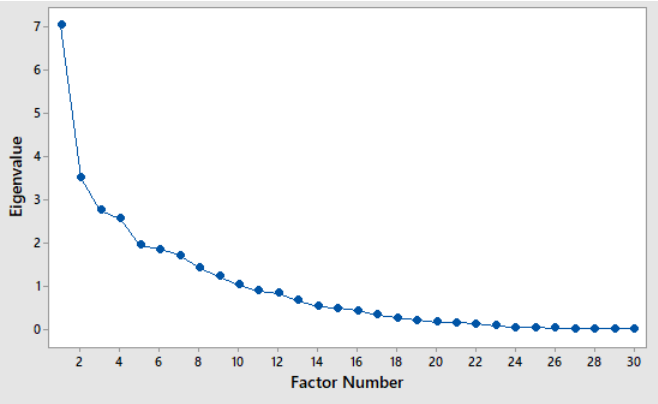


Figure 2. Scree plot for Factor Analysis of the dataset used in this study.

Conclusions

Predictive data analytics approaches provide a valid insight when looking for dominant factors that will help define possible pedagogical interventions, as is the case in this study. For the dataset utilized there is a good agreement in terms of the type of factors (i.e., question number) that define top performance in a standardized skill visualization test. Given the design of the PSVT:R, more involved manipulations are indicative of higher performance by the student. Comment is that only the first dominant factor in the predictive model is considered, and in such case there are two models that do not follow the expected trend (i.e., for scores of 24 and 26), which indicates that further investigation is required.

Regarding the parameters for this study, one issue is that even when a larger dataset has been utilized, it might need to have a substantially larger set for better generation of predictive models. A second issue is the possible use of a different predictive analytics approach, this is a field that is constantly being improved and there might be a 'better trap' out there already. Both of these issues are currently being considered.

References

Best Data Analysis Tools (2017) https://www.softwareadvice.com/bi/data-analysis-comparison.

Big Data Analytics (2017) What it is and why it matters, https://www.sas.com/en_us/insights/ analytics/big-data-analytics.html.

Data Analytics (2017) Make Better Business Decisions, https://www.qlik.com/us/lp/ppc/qlik-sense- desktop/business-analytics?sourceID1=google&Campaign_Type=Non-Brand&KW=data %20%26 %20analytics&k_clickid=24c8bad7-5950-446c-9aa6-

68027642b4fd&gclid=EAIaIQobChMIIraI
we6V2QIVT7nACh39LAGpEAAYAiAAEgJR4_D_BwE.

Ernst, J. V., Williams, T. O., Clark, A. C. and Kelly, D. P. (2017) Factors of Spatial Visualization: An Analysis of the PSVT:R, Engineering Design Graphics Journal, vol. 81, no.1, winter 2017.

Farias, H. (2017) Machine Learning vs Predictive Analytics, Concepta https://conceptainc.com/?hstc=753710.b1677a14cd71d7df609ab70b58cb063c.1540727194782.1540727194782.154073 2836484.2&_hssc=753710.1.1540732836484&_hsfp=3519403555.

Guay, R. (1977) Purdue Spatial Visualization Test – Visualization of Rotations. West Lafayette, IN Purdue Research Foundation.

Kozhevnikov, M. and Thornton, R. (2006) Real-Time Data Display, Spatial Visualization Ability, and Learning Force and Motion Concepts, Journal of Science Education and Technology, vol. 15, no. 1, Springer Science & Busines Media.

Prieto, G. and Velasco, A. D. (2010) Does spatial visualization ability improve after studying technical drawing? Quality and Quantity, Research Note, Springer.

RapidMiner (2017) User's Manual, https://docs.rapidminer.com/downloads/RapidMiner-v6-user- manual.pdf.

Rodriguez, J. and Rodriguez-Velazquez, L. G. (2016a) Comparison of Spatial Visualization Skills in Two Approaches to Entry-Level Graphic Courses, Proceedings of ASEE Annual Conference.

Rodriguez, J. and Rodriguez, L. G. (2016b) Comparison of Spatial Visualization Skills in Courses with Either Graphics or Solid Modeling Content, Proceedings of ASEE-EDGD Mid-year Conference.

Rodriguez, J. and Rodriguez-Velazquez, L. G. (2018) Application of Data Analytics Approach to Spatial Visualization Test Results, Proceedings of ASEE Annual Conference.

Sorby, S. A. (1999) Developing 3-D Spatial Visualization Skills, Engineering Design Graphics Journal, 63(1), 21-32.

Sorby, S. A. (2005) Assessment of a New and Improved Course for the Development of 3-D Spatial Skills. Engineering Design Graphics Journal, 69(3), 6-13, 2005.

Strong, S. and Smith, R. (2002) Spatial Visualization: Fundamentals and Trends in Engineering Graphics, Journal of Industrial technology, vol. 18, no. 1.

Titus, S. and Horsman, E. (2009) Characterizing and Improving Spatial Visualization Skills, Journal of Geoscience Education, vol. 57, no. 4.

Yue, J. (2008) Spatial Visualization by Realistic 3D Views, Engineering Design Graphics Journal, vol. 72, no. 1, winter 2008.