Comparing the accuracy of decision trees and logistic regression in personnel selection

Proposal

Kyle C. Marks

Alexander T. Jackson, Ph.D.

Ying Jin, Ph.D.

Don Zhang, Ph.D.

Edgar Kausel, Ph.D.

Middle Tennessee State University

DECISION TREES IN PERSONNEL SELECTION

Abstract

The current study aims to compare the accuracy of decision trees to logistic regression in a

personnel selection context. To make this comparison, two studies are proposed. The first study

uses simulated applicant data. For each applicant, there will be a cognitive ability score,

conscientiousness rating, and a structured interview score. Job performance will be simulated as

a function of the simulated scores. Additionally, different selection ratios will be applied to the

simulated data to mimic how organizations select applicants and to determine whether the

selection ratio has an impact on the accuracy of each analytic approach. A second purpose of

this study is to examine whether the decision strategies used by decision makers in a real

selection context (graduate school admission decisions) reflect the strategies that the decision

makers *should* be using. For each graduate school applicant, the predictors of undergraduate

grade point average (GPA) and graduate record examination (GRE) scores will be used. To

measure performance, final graduate GPA will be used. The performance data will be used to

determine the accuracy of the decision tool in a real selection context. Additionally, there will be

information on the decision of whether the applicant was admitted to the program. Whether an

applicant was admitted will be matched to the prediction made by the regression analysis and the

decision tree analysis to determine if what was done matches what the regression and decision

tree models. This is meant to determine which analysis more closely matches how individuals

actually make decisions.

Summary
Background
Actuarial methods such as logistic regression have been shown to be more accurate in predicting applicant's future performance when compared clinical judgments (Dawes, 1971, 1979). Despite this, organizations continue to use suboptimal methods when making decisions (Highhouse, 2008). Decision trees have gained continued research interest as a fast and easy to interpret actuarial method (Raab & Grigerenzer, 2015). Decision trees are an attractive alternative due to the ease of interpretation. With methods like logistic regression, prior statistical knowledge is needed in order to interpret the model. Decision trees require no such prerequisites to interpret. This has led to numerous studies in fields ranging from finance to medicine that have demonstrated decision trees to be as accurate as other actuarial methods (Sinha & May, 2004).

> *Hypothesis 1:* Decision tree analysis will be as accurate as logistic regression in predicting performance outcomes in the simulated environment. Accuracy of decision trees will not be significantly different from that of regression.
> *Hypothesis 2:* Decision tree analysis will be as accurate as logistic regression in predicting acceptance into a graduate school program. Accuracy of decision trees will not be significantly different from that of regression.

There are three commonly used selection ratios: fixed quota, threshold selection, and mixed quota/threshold decision (De Corte, 1999, 2002). These different ratios will be used to simulate how different organizations hire employees.

> *Hypothesis 3:* More stringent selection ratios will reduce both logistic regression and decision tree models ability to accurately identify the top performing applicants.
> *Hypothesis 4:* Decision trees will be more robust to changes in selection ratio than logistic regression.

Methods

Study One
A population will be simulated using conscientiousness, GMA, and structured interview scores. Coefficients for the regression equation can be seen below (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000). The population will then be split into a training sample and a test sample consisting of 250 applicants, the average number of applicant to many white collar jobs (Joyce, 2016). The training sample is used to create both a logistic regression and decision tree model. The test sample is used to apply the models and obtain classification accuracy. This will be done under three different selection ratios, fixed quota (performance in the top ten percent), fixed threshold (performance above .371) and mixed model (performance in the top ten percent and above .371). To analyze this data a factorial ANOVA will be used to compare methods.

$$Perfor\widehat{\widehat{mance}} = \beta_0 + .43 * GMA + .24 * Consientiousness + .58 * structuredinteview + Error$$

Study Two
This study will use archival data from MTSU's I/O psychology graduate program to compare the decision trees to logistic regression in a real sample. Predictors will be GPA, GRE verbal, and GRE quantitative with the outcome variable being whether or not they were selected to the program. The data will be split into thirds with one third being training data and the other test data. An independent samples t-test will be used to compare methods.

DECISION TREES IN PERSONNEL SELECTION

References

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, K. H., & Gilliland, S. W. (2000). The incremental vaidity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, *53*(2), 325–351. https://doi.org/10.1111/j.1744-6570.2000.tb00204.x

Dawes, R. M. (1971). A Case Study of Graduate Admissions. *American Psychologist*, *26*, 180–188. https://doi.org/10.1037/h0030868

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

De Corte, W. (1999). A note on the success ratio and the utility of fixed hiring rate personnel selection decisions. *Journal of Applied Psychology*, *84*(6), 952–958. https://doi.org/10.1037/0021-9010.84.6.952

De Corte, W. (2002). Sampling variability of the success ratio in predictor-based selection. *The British Journal of Mathematical and Statistical Psychology*, *55*(Pt 1), 93–107. https://doi.org/10.1348/000711002159716

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial & Organizational Psychology*, *1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Joyce, S. P. (2016). The internet revolution : Digital disruption in recruiting & job search. *Career Planning & Adult Development Journal*, *32*(2), 14–21.

Liu, Y., Lin, D., Xiao, T., Ma, Y., Hu, Z., Zheng, H., … Gao, Y. (2011). An immunohistochemical analysis-based decision tree model for estimating the risk of lymphatic metastasis in pN0 squamous cell carcinomas of the lung. *Histopathology*, *59*(5), 882–891. https://doi.org/10.1111/j.1365-2559.2011.04013.x

Raab, M., & Gigerenzer, G. (2015). The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology*, *6*(OCT), 1–6. https://doi.org/10.3389/fpsyg.2015.01672

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, *21*(3), 249–280. https://doi.org/10.1080/07421222.2004.11045815