Data mine and forget it?: A cautionary tale

Yuri Tada, Ph.D.[1], Judith Orasanu, Ph.D.[2], & Norbert Kraft, M.D.[1]

[1]San Jose State University/NASA Ames Research Center, [2]NASA Ames Research Center

ABSTRACT

With the development of new technologies, data mining has become increasingly popular. However, caution should be exercised in choosing the variables to include in data mining. A series of regression trees was created to demonstrate the change in the selection by the program of significant predictors based on the nature of variables.

INTRODUCTION

Data mining is advertised as a method of extracting hidden predictive information from large databases. In recent years, with the development of new technologies that allow users to navigate through their data in real time, the use of data mining has become increasingly popular. However, companies and analysts tend to put automatically collected data through a data mining program without careful consideration of the variables. The original data may not only be incomplete, noisy, and inconsistent, but also contain highly correlated measures and uninteresting/irrelevant variables. Many data mining techniques are based on statistical theories (e.g., CART) and hence, careful thought is warranted on just what input variables should be included to arrive at meaningful solutions, even for an exploratory procedure like data mining.

RESEARCH QUESTIONS

- **Can data mining uncover truly meaningful relationships when variables are indiscriminately entered?**
  - **What's the effect of including highly correlated performance variables?**
  - **The effect of including disaggregated ancillary performance variables?**

METHOD

A series of regression trees was constructed for two main outcome variables to examine the interpretability of resulting trees based on the type of input variables. Data came from a study involving team performance by 24 ad hoc teams who worked on six computer-based distributed team search missions over three days on six 75-minute computer-simulated search scenarios set in Mars (Orasanu, et al., 2008).

Participants worked in separate booths, communicating via email and audio headsets. They had to plan a search strategy, share information, and manage limited resources while dealing with time-pressured tasks and dynamically changing conditions. The Distributed Dynamic Decision-making (DDD) software and search scenarios were developed by Aptima, Inc.

VARIABLES

***Team Performance obtained from* 24 teams (*N* = 120, female *n* = 48, male *n* = 72)**

Outcome Variables: (1) Total Team Points earned, and (2) Overall Mission Success

Ancillary Performance Variables:

Team Points in Moderately Difficult scenarios, Team Points in Difficult scenarios, SM (Seismic Monitor) scores of various importance at Days 1, 2, & 3, ET (Emergency Task) scores at Days 1, 2, & 3, & Collaboration score

***Individual Differences*: Big Five Inventory (BFI; John et al., 1991), aggregated across team members**

***Group Dynamics:* Group Environment Scale (GES; Moos & Humphrey, 1974)**

***Cognitive Ability*: WinSCAT (Kane, Short, Sipes & Flynn, 2005), aggregated across team members**

RESULTS

Regression Tree Set #1: All potential predictors included

- **The first model used all distinct variables other than the main outcome variables as predictors, including secondary performance measures.**

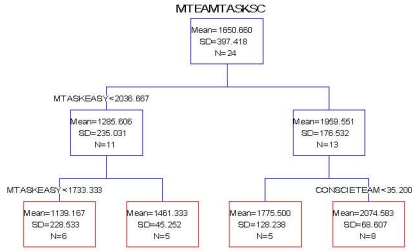

***Figure 1*. Regression tree of total team points earned with all variables included as possible predictors**

- **Team points in moderately difficult scenarios overwhelmingly predicted the outcome**
- **The proportional reduction in error [PRE] at the first split was .75 for total team points**

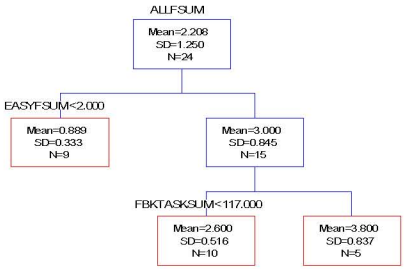

***Figure 2*. Regression tree of overall mission success with all variables included as possible predictors**

- **Mission success in moderately difficult scenarios overwhelmingly predicted the outcome**
- **PRE at the first split was .67 for overall mission success**

RESULTS

Regression Tree Set #2: Exclusion and aggregation of predictors

- **The next tree excluded those closely related performance measures as input variables, but retained the ancillary performance measures for individual days**
- **Number of successfully processed 500-pt tasks in Day 1 was found to be predictive of both outcome variables, although there was no obvious reason for the particular day's measure to be predictive of performance**

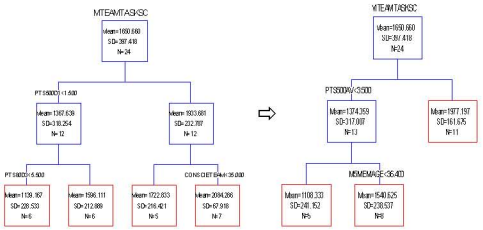

***Fig. 3*. Regression trees of total team points earned modified by aggregation of predictors**

- **Number of successfully processed 500-pt tasks across days were averaged in the left tree to stabilize the variability**
  - **The same variable, now averaged, was selected as a predictor, suggesting that it was not specific to that particular day**
  - **PRE at the first split increased from .529 to .596**

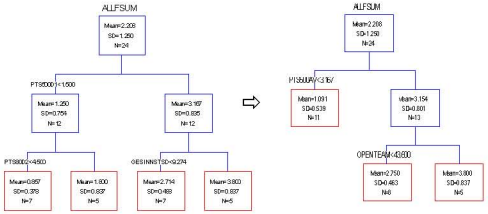

***Figure 4*. Regression trees of overall mission success modified by aggregation of predictors**

- **Number of successfully processed 500-pt tasks across days were averaged in the left tree to stabilize the variability**
  - **The same variable, now averaged, was selected as a predictor on overall mission success as well**
  - **PRE at the first split increased from .613 to .705**

Regression Tree Set #3: No performance predictors

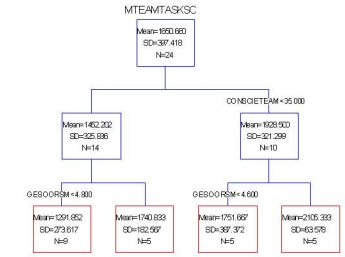

***Figure 5*. Regression tree of total team points earned without performance predictors**

- **PRE for the tree was .629 = ~64% of total team points accounted for by individual/team dynamic measures**
- **Teams performed better if they were not so high on average consciousness but very orderly and organized as a team**

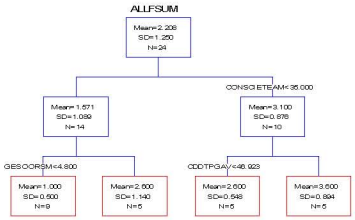

***Figure 6*. Regression tree of overall mission success without performance predictors**

- **PRE for the tree was .677 = ~68% of overall mission success accounted for by individual, team dynamic & cognitive measures**
- **Teams performed better again if they were not so high on average consciousness but had performed exceptionally in code substitution as a team**

CONCLUSIONS

- **Including predictors indiscriminately can obscure the meaningful relationship**
  - **Predictors that are highly correlated with the outcome variable may not be useful**
  - **Aggregation can stabilize unsystematic variability, if multiple measures of the same variable exist**
- **Think carefully before choosing predictors to include in data mining**
  - **Data mining is not a magic bullet → Garbage In, Garbage Out still holds, as in any data analysis**