Exploring the Use of Classification Trees in Categorizing Survey Respondents on Perceived Survey Burden

Arcenis Rojas, Scott Fricker, Lucilla Tan

U.S. Bureau of Labor Statistics

Background

A classification tree model uses predictor variables to predict unit membership in one of the dependent variable's classes using one or more statistical methods such as discriminant analysis or cluster analysis and can be non-parametric. Some common algorithms are CART, AID, THAID, CTREE, QUEST, CRUISE, and GUIDE.

Goal: to develop a classification tree model to predict respondent's perception of survey burden from burden-related questions asked at the end of the survey.

The Consumer Expenditure Survey (CE)

- Data on household characteristics, expenditures, income, and taxes. See http://www.bls.gov/cex/
- Two independent surveys: Interview and Diary
  - Interview is a multi-wave recall survey that covers mainly large expenditures; *this is the data source for this study*.
  - Diary respondents generally report smaller, day-to-day expenditures over two 1-week periods.
- Paradata on contact attempts and periodic post-survey assessment questions asked of interviewers and respondents
  - *This study utilizes post-survey burden-related questions asked after 5th wave from the Interview survey only.*

Select Classification Tree Options

- Classification or regression tree: is the dependent variable categorical or continuous?
- Interaction detection: Should variable interactions be considered for splits?
- Univariate or linear splits: Should the algorithm prioritize univariate or linear splits? Should only 1 type of split be used?
- Equal or estimated priors: Should the probability of each class be equal to that of the other classes?
- Misclassification costs: Are some misclassifications more egregious than others?

Study sample

- 655 single-person consumer units who had completed their 5th wave interview between April and September 2013.
- Dependent variable: Direct burden question; has 4 response options ranging from "Not burdensome" to "Very burdensome". (Figure 1)
- Predictors: 14 burden-related questions, ranging between 3 to 5 response options. (Figure 2)

Figure 1: Direct burden item response distribution **(dependent variable)**

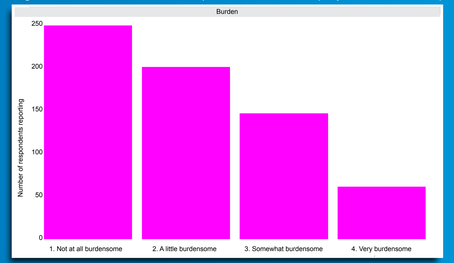


Figure 2: 14 burden-related items response distributions **(predictors)**

Methodology

1. **Recoding of burden-related items.**
   Predictor variables were recoded so that the direction of burden went from low to high to correspond with the dependent variable

2. **Dealing with Noisy Data**
   Initial classification trees using GUIDE's default options yielded models with high misclassification rates. We categorized the sources of the noise as methodological or data-driven.

Methodological problem:

CE is not designed as an attitudinal survey → measurement of respondent experiences is inherently difficult

Data-driven problems:

- When predicting the main burden variable, which is coded as four possible response options, equal misclassification costs could cause misclassification, e.g., the algorithm applied the same cost of misclassifying a "Not burdensome" response as a "Very burdensome"
- The frequency distribution of the main burden variable is uneven, which requires the use of estimated priors.
  - Response of "Not at all burdensome" is five times more likely than "Very burdensome"

3. **Software used:** GUIDE version 21.2

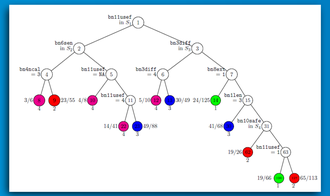
GUIDE References

1. Loh, W.-Y. (2002), Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, vol. 12, 361-386.
2. To download GUIDE and for further references please visit: http://www.stat.wisc.edu/~loh/guide.html

Intermediate Models

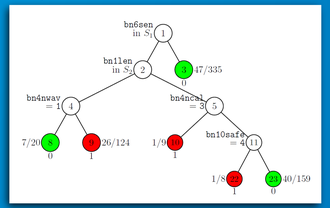
Model 3

Equal misclassification costs and equal priors did not correspond with underlying data properties, thus, the model yielded high misclassification rates (52.6%)



Model 14

Using estimated priors and collapsing the main burden variable into a binary variable (0 = "Not at all burdensome" or "A little burdensome", 1 = "Somewhat burdensome" or "Very burdensome") assigned probabilities of classification that corresponded more closely to the underlying data and yielded a lower misclassification rate of 18.6%



Final Model and Findings

In the final model we collapsed the main burden variable into a binary variable where "Very Burdensome" was in its own class. We did this to further reduce the probability of misclassifiying any other type of response as "Very Burdensome." Running the model with 10 cross-validations yielded a misclassification rate of 5.5%

Findings

The final model suggested that among the 14 burden-related items, 5 were highly predictive of the "very burdened" category of the direct burden question: Useful, Sensitivity, Number of Calls, Difficulty, and Effort. Specifically, Wave 5 respondents who perceived their survey experience to be "very burdensome" were those who rated:

- *"time and effort spent on the survey to be not well spent"*, and *"the survey questions were very sensitive"* , OR
- *"time and effort spent on the survey to be at least a little well spent"*, but *"prior to the interviewer there were too many contact attempts"* , and *"answering the survey questions was not very easy"*, and *"effort exerted to answering the survey was not moderate (i.e., a little OR a lot)"* .

Arcenis Rojas, Economist. Branch of Information & Analysis , Consumer Expenditure Survey, U.S. Bureau of Labor Statistics
Email: Rojas.Arcenis@bls.gov | Telephone: 202-691-6884

U.S. Bureau of Labor Statistics BLS