Predicting Income from 1996 US Census Data

Methods

Data on 48,842 individuals from the 1996 US Census were consolidated from an SQLite database and stored as a .csv file. R statistical software was used for analysis. Data were split into training, testing, and validation (60%, 20%, 20%) using simple random sampling without replacement. Descriptive statistics, missing values, and distribution of values were recorded. Data were transformed and levels of categorical variables were collapsed when appropriate. Two continuous variables, *Capital Gains* and *Losses*, were binned using splits determined by a decision tree. A decision tree was used as a baseline model, with the binary variable "Over-50k" used as the target. Variables relationships with the target were investigated, as well as interactions between variables. A logistic regression model was evaluated, and stepwise and backwards selection techniques were used.

Results and Recommendations

Several factors were found to be significant predictors for whether an individual made more than $50,000. The most significant variables included occupation, education level, age, and post-social insurance gains and losses. These six variables explain 57.5% of the variability in the data set. The strongest relationship overall was that individuals with capital losses between $1,564-$1,816 were **1.96x less likely** to make above $50,000. The strongest positive relationship was that individuals with a doctorate degree were **1.4x more likely** to make above $50,000.

The recommended model is a logistic regression model using the log-transformed age variable, binned capital gains and losses, education level, occupation, and relationship status. This model is the simplest model that retains a majority of the predictive power. Additional variables and interactions could be included, but the gain in predictive power is minimal. A ROC curve and AUC values from the validation data are provided comparing the different models tested (Figure 1). The recommended model will provide a probability for an individual making more than $50,000, and depending on the cost associated with false positives or false negatives, a cut-off for probability can be recommended.

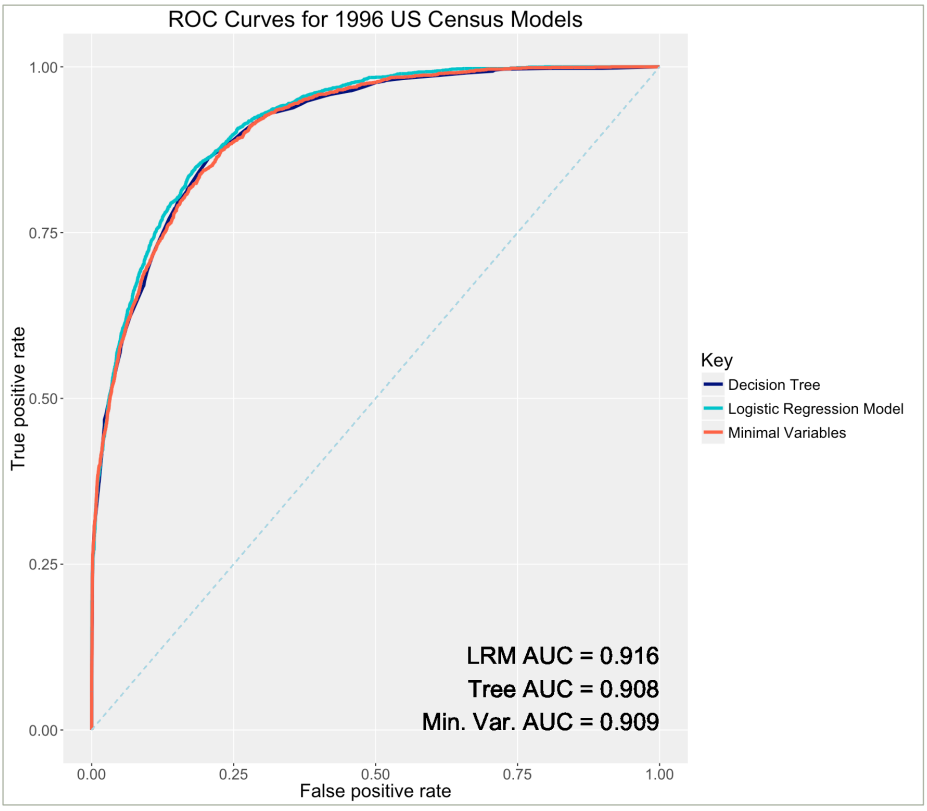

Figure 1. ROC Curves

Cory Everington