**Predicting High (>$50k) or Low (<$50k) Income from US Census Data**

**A Comparison of Classifying Algorithms**

**Project Proposal**

**10/3/2016**

**Prepared For**

**Dr. Raed Seetan**

**CPSC605-88 – Data Mining and Analysis**

**Prepared By**

**Trace Tidwell**

**wct1001@sru.edu**

**Abstract**

With the ever increasing quantity of data being collected, machine learning techniques are being utilized to solve complex problems more and more frequently. Both the public and private sectors are amassing massive amounts of data, and they know that there is great value to be had if put to good use. Some of the more well-known companies that use machine learning, including classification algorithms, are Amazon, Facebook, and Google. Classification algorithms, or classifiers, are used to determine to which class observations belong, or classify them. In its simplest form, classification is an example of pattern recognition. By using training data containing observations whose class membership is already known, the classifier can then determine to which class a new set of observations, whose class is unknown, belongs. Specific examples of companies using classification algorithms are banks using customer data to determine which loan applicants are risky and which are safe and marketers using customer profiles to determine who will or will not purchase a product. There are many different types of classifiers, and, depending on the dataset, certain classifiers may perform better for a given task.

**Introduction**

Income inequality has long been a hot-button issue in American politics. Two of the most popular methods for combating income inequality are redistribution of wealth from higher earning households to lower earning households, and helping lower income households develop the tools necessary to increase their incomes, such as job training and education grants. This analysis will focus on the latter method. Most people seem to intuitively understand that certain factors can increase income. For example, it comes as no surprise that the number of years of education is very strongly correlated with higher income, as is working in certain fields, such as medicine, law, or STEM. Currently, the primary method used to determine such factors is a review of census data. From there, statistics are assembled stating things like, “Among earners in the upper quintile, 77% worked full-time and 61% of households had two or more income earners, compared to the national percentages of 54% and 39%, respectively,” or, “The median income for households with at least a Bachelor’s degree is $94,934 versus $58,044 for the general population.” However, these are simply descriptive statistics. Our goal is to build a model whereby users can input their own information (or future information) and predict whether or not they can expect to have a household income above the median.

Using census income data, we can examine which variables contribute to incomes that are greater than the median and which variables contribute to incomes that are less than the median. The results can then be used to make recommendations about which factors play the strongest role in predicting higher (and lower) incomes. In this particular study, a comparison of five classifiers – decision trees, support-vector machines, naïve Bayes, K-nearest neighbors, and random forests – will be tested to determine which one can most accurately predict if citizens of the United States have household incomes above or below the median value. Multiple iterations will be performed on different subsets of the data for each dataset and each algorithm. The predicted results of the algorithms will then be compared to the actual results, and the algorithm with the highest average score will be deemed the best. We can also determine the weights of the respective variables to see which ones contribute most heavily. To test the classifiers, three different data sets will be used: Census Income Data Set, Census-Income (KDD) Data Set, and US Census Data (1990) Data Set. The data sets contain information from the 1990, 1994, and 1995 censuses, and all three can be found at the UC Irvine Machine Learning Repository.

Deliverables for the project include a written analysis, the R script written to perform the comparison, the datasets used for the comparison, and a bibliography. In addition to analyzing the results of the comparison, recommendations will be made about which factors to pursue and avoid in order to increase one’s chance at having an income above the median. A secondary goal, though perhaps not possible within the time constraints of the project, would be to build a web app where users can input certain variables such as marital status, education level, field of employment, etc. and have the app return a yes/no prediction of whether household income is greater than the median income.

Sources (to be cited properly in Final Project)

<http://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-05.html>

<http://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html>

<https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>

<https://cran.r-project.org/web/packages/rpart/rpart.pdf>

<ftp://cran.r-project.org/pub/R/web/packages/e1071/vignettes/svmdoc.pdf>

<https://cran.r-project.org/web/packages/e1071/e1071.pdf>

<https://cran.r-project.org/web/packages/kknn/kknn.pdf>

<http://archive.ics.uci.edu/ml/datasets/Census+Income>

<http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>

<http://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29>