Memorandum

To: Executive Stakeholders meeting

From: Justin Kreft

Re: Predicting if a customer makes over $50k based on census data

Date: April 6, 2016

Overview

The purpose of the following analysis was to produce a model that predicts if a new customer makes over $50k based on information features typically provided in census data. My analysis examined 48,842 records from the US Census, assessed four different machine learning models, and produced a final model was able to correctly predict if a new observation makes over $50k with an accuracy of 85% based on 8 key variables (see Table 1). Of these variables, education number, capital gain, and marital status were by far the most influential in the analysis.

Observations from Key Features

In the attached visualization, you can see three charts demonstrating the relationships between these three variables and the someone making over $50k. Unsurprisingly, capital gains greater than five thousand account for large amounts of the population that make over 50k. Generally, capital gains are a strong indicator for prediction. Also, as an individual’s education level increases, the divergence between over $50k and under $50k populations increase. The attached chart also compares distributions of education levels across marital statuses for the over and under $50k subgroups.

The bottom chart examines the performance of the predictive model against capital gains and education level. Most prediction errors were made in the where capital gains were zero and education level was high (e.g. greater than an associates degree).

Methodology

The data from the census records was extracted and cleaned for analysis, and four different models (RandomForest, LogisticRegression, K Nearest Neighbor, and DecisionTree) were examined over three rounds of evaluation. The first round determined baseline performance of the models with all 13 original feature variables (see Table 1). The second round eliminated relationship and education\_level as features due to conflicts or duplication with other more informative features; it also eliminated country, race, and workclass as features because these resulted in categories that did not produce groups that were either mostly over $50k or mostly under $50k.[[1]](#footnote-1) In contrast, educational level creates relatively pure groupings (e.g. persons with only an 8th grade education rarely make over $50k).

The final model (K Nearest Neighbor) was selected because it had the best overall performance. K Nearest Neighbor predicts if a person makes over $50k or under $50k based on a other people that share similar characteristics data (i.e. its neighbors). The final model (KNN@10) makes its prediction based on the majority vote of a person’s 10 closest neighbors.

Analysis of the dataset was able to produce a model (KNN@10) that correctly predicts if a new instance is over\_50k by using 8 of the original 13 variables. When assessed against a test set, the model correctly predicted 0.852 of instances, indicating that the model could be generalized to assess if new customers make over $50k.[[2]](#footnote-2)

1. The removal of these features reduced the noise in the dataset and improved performance for three out of four models at p<.05. [↑](#footnote-ref-1)
2. Data was split into training (.75) and test sets (.25) from the cleaned data. Validation/model evaluation was accomplished through 10-fold cross-validation to maximize the amount of data available for training. [↑](#footnote-ref-2)