CMSC 12300 Final Project

Andy Liu and Nelson Auner

June 2, 2013

1 Introduction

1.1 Logistic regression of census data

Our project investigates predicting whether a household self reports that its income is over or under \$50,000, based on census data. We aim to find the best classification method within the subset of logistic regression classifiers. Logistic regression transforms the predictors to

$$\pi(X) = \frac{e^{(X\beta)}}{1 + e^{(X\beta)}} \tag{1}$$

this projects the space of $X\beta$ to the interval $[0, 1]$, allowing us to view the result as the probability that our outcome is 1. Although more complicated than a basic regression mode $Y = X\beta$, logistic regression allows us to predict a binary outcome, while resulting in more interpretable coefficients β than black box methods such as neural networks or random forest.

1.2 Goals and summary of results

Our main goals, put succintly, were to find and compare the best prediction method by testing various combination of predictors. For each prediction method, we wanted to quantify the effectiveness of that algorithm. Given the large proportion of observations with missing values, we also wanted to examine the effects of filling in missing observations with values.

2 Description of the Data set

Our analysis is on the Census-Income (KDD) Data Set, a classic machine learning dataset as well as applicable machine The census data we used is made freely available from the UC-Irvine Machine Learning Repository The most pertinant characteristic of the data set is that the ”over 50K” variable only takes on a value of true for 6.2

3 Initial analysis

Our inital analysis was done in R, and much work was required to find and adapt the necessary statistical methods for use in python. We used the Akaike Information Criteria (AIC), defined as $AIC = 2k - 2ln(L)$, where k is the number of the parameters in the model, and L is the maximized value of the likelihood function - in our case, the logistical regression equation defined above. By maximizing AIC, we produced a subset of predictors that AIC evaluated as the "best" predictor sets. We noticed that among single-variable predictive methods, the education variable was one of the most effective, with a false positive error rate of 5