Final Project Report

Beshad Talayminaei & James Collins

May 17, 2015

Summary

The goal of this project was to facilitate interaction with and understanding of US census data. To this end we used Public Use Microdata Sample (PUMS) data sets provided by the US Census Bureau for the state of Maryland for the years 2012 and 2013. We developed a Shiny application to serve as a user interface for interacting with this data. We used a k-nearest neighbor approach to develop a predictive model for income. The end result of this project is a functional, and user friendly Shiny application that enables the user to subset and view US census data in various ways. However, the error rates for the predictive model were quite high and perhaps in the future more time could be spent finding ways to improve this model.

Introduction

Census data can provide a lot of interesting information and can offer many insights into various aspects of the society. When looking at the raw Public Use Microdata Samples using excel or in CSV format it is difficult to interact with or understand the data because there are over 95,000 observations and as over 200 variables. Therefore, one of our major tasks was to create a user interface to provide users with the ability to subset the dataset in various ways to better make sense of the data. We focused on developing clean and easy to understand graphs and plots. We decided to use the Shiny web application framework for R, as it provides fairly easy ways to add interactivity. In order to add some more concrete and useful information to the application we created a model for predicting income based on some of the variables available in the PUMS dataset for the state of Maryland. After trying various linear and non-linear models, we found that a k-nearest neighbors model predicts income with the highest accuracy although this model is still very prone to error.

Materials & Methods

Public Use Microdata Sample datasets are untabulated records about individual people or housing units. In this project, we focused soley on data collected for individual people. The combined datasets for 2012 and 2013 dataset include over 95,000 observations along with over 200 variables. For the purposes of this project, obviously not all 200 some variables were required. We selected only a few of these variable which we thought would have the greatest influence on income. The variables we selected include age, sex, and race of the individual. In addition we thought that level of education as well as and individual's field could have a significant impact. Lastly, level of income of individuals was crucial for this investigation.

As college students who have a vested interest in our future prospects we were particularly interested in examining data for individuals who have attained at least a bachelor's degree. Observations for which any of the variables mentioned above were missing, were also excluded from the data. Through this process, the team reduced the number of observations to just over 24,000. The figure below shows a small sample of the final data used in this project. The figure below shows a sample of the data used for the purposes of this project.

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##     WAGP FOD1P SCHL RAC1P AGEP PUMA OCCP SEX HINS1 MAR
## 1 398000  3606   24     2   56 1003   10   1     1   1
## 2  60000  5200   24     2   59 1003 2000   2     1   1
## 3  68000  6200   22     2   56 1107 2145   2     1   1
## 4  49000  2307   22     1   25  502 2330   2     1   5
## 5  65000  2414   21     1   23  901 1530   1     1   5
## 6  30000  6201   22     1   45  301  800   2     2   3
```

A main focus of analysis in this project was creating an accurate model for predicting income based on other variables. To this end we used various analysis methods such as linear regression, random forest, and k-nearest neighbors. We began the analysis using a simple linear regression model to assess the signifcance of each of the parameters used to predict income, and to determine whether a linear model can accurately make this prediction. The first set of analyses showed that many of the variables have great significance in predicting income. These variables were used in various combinations and with different interactions and many had a p value very close to 0. However, the adjusted R-squared value for all of the interaction terms was also significantly low. The highest adjusted R-squared value attained was around 0.2, meaning that the model could only represent about 20 percent of the data.

Looking at Q-Q plots of the data, we saw that many subsets of this data are fairly normally distributed but with large tails.. To compensate for this we decided to give the user the option of looking at the log of income instead of raw income for the Q-Q plot in the Shiny app. We once again tried a linear regression model for predicting log of income and as before, the results showed very low p values but also a very a low adjusted R-squared value. However, knowing that the parameters were indeed significant in predicting income, the we decided to use the same parameters in different models to try and get more accurate results.

Next we tried using a random forest model with 10 fold cross validation to predict both raw income and log of income. However the mean error accross these 10 folds was dissapointing. The random forest model was also ineffective in improving prediction accuracy. We then decided to look into the k-nearest neighbors model in order to, once again, more accurately predict income. To use knn we broke up the data into a training set and a test set. The training set included all observations from the year 2012 and the test set included observations from the year 2013. The goal was to see how accuratly knn can predict income in 2013 using records from 2012.

We assesed the performance of the model by comparing what the incomes th model predicted for 2013 tow what the actual values were and padded the actual values with different error ranges. Seeing that the knn function in R does not treat the income variables as continuous, we decided that setting a range in which a prediction is deemed acceptable was crucial. Since income is a variable that can see a significant amout of fluctuation, we decided it would make more sense to check if predictions fell within an acceptable error range around the actual value. For example, if the model was able to consistently predict that an individual would make $80,000 with an acceptable error range of +- $10,000, this model might be very informative. Deciding on this range was a difficult task since we wanted the user to be able to subset data based on income. Therefore, we decided that it would be best if the choice of acceptable error range was left to the user of the Shiny application via a simple slider.

Results

The first goal of this project was to allow users to easially interact with, subset and better understand the data in the Public Use Microdata dataset for the state of Maryland in the years 2012 and 2013. We developed a very easy to use interface using Shiny that allows users to subset the data in various ways and view that data on a few graphs so as to acquire a better understanding of the dataset.

The sidebar pane of the shiny app allows users to subset the data based on age, income, degree type, sex, race, and major category. The first pane in the Shiny app displays the subset of data in a histogram that plots frequency of income. The user can select what granularity they want by using the bin slider at the bottom of that pane. The second pane displays a Q-Q plot using the qqnorm function. The user has the option of viewing the Q-Q plot for raw income or for log of income. The third pane displays the distribution of income geographically. The PUMS datasets have Public Use Microdata Area (PUMA) codes that indicate what geographic area in the individual lives in. More information about PUMAs can be found here: https://www.census.gov/geo/reference/puma.html We were able to find a shapefile for these areas from the U.S. census bureau's website. Using these codes and the shapefile we were able display a map of these areas and shade them according to various user-defined options. The user can shade theses areas by the number of observations per area in the selected subset or by the mean income of observations per area in the selected subset or by the sum of incomes per area in the selected subset. The last pane goes into greater detail about how to use the application and how to interpret the different visualizations.

The second goal of this project was to create an accurate model for predicting income based on a few sigfinicant parameters available in the dataset. Age, sex, and race proved to be very significant parameters in predicting income with p values so small that R displayed them as 2e-16. Furthermore, level of education along with the field of degree also had p values very close to 0. As described above we attempted to fit a model using various modeling techniques such as linear regression, random forest, and k-nearest neighbors. Of these models, the k-nearest neighbors performed best with the lowest error rate given the same acceptable error ranges.

Some insights into this dataset can be gained from the knn model however the predictive power of the model is a bit dissapointing. Given a large error range for prediction, the model could predict income for some subsets with accuracy as high as 100 percent. However, this was only really the case for very small and specific subsets of data typically in the higher income ranges. Obviously, given the fact that the income figures in the dataset range from $0 to $250,000, with the majority falling in the $0 to $100,000 range, the model is not predicting much of anything. For ages 18-65 with incomes between $1000 and $250,000 and an accepable error of +-$15,000 the model is able to predict with about 24% accuracy. Although this accuracy is higher than the accuracy of other models previously attempted, it was still disappointing to see that the accuracy is not very good and would be virtually unuseable for any practical applications.

Before making the final decision that, given our dataset and predictors, these results were inconclusive, the we made one last attempt to transform the data used in the k-nearest neighbor model in order to account for the highly tailed distribution. A logarithmic function of base e was applied to the training wage vector as well as the test wage vector. Obviously, the format by which ranges were determined had to change as well to account for the logarithmic function. This change resulted in a marginal increase in predictive accuracy. givent the same parameters as above we were able to achieve about 26% accuracy which is still not very informative or useful.

Given the low accuracy of the various predictive models used to for income in this dataset, we decided that the results of our final predictive model are inconclusive. This is not a complete surprise given the variability of income on individual basis. One factor that could have added more unpredictability is that the we decided to only look at individuals with a bachelor's degree or higher. Doing so limited the data to under half of what was originally made available by the US Census Bureau. Perhaps including observations without college degrees may have made it easier to predict income and might improve the model. Maybe adding in other factors like occupation or type of job (government, private sector) would also help improve the model. Then again perhaps income is just not as predictable using the given predictors as one might assume.

Conclusion

We were able to create a user friendly web application through the use of the Shiny framework. This application allows the user to interact with the census data in various ways and better understand it through different data visualization techniques. We also attempted to create an accurate model for predicting income based on an individuals age, sex, race, level of education, and field of degree. However, the accuracy of the various predictive models, such as linear, random forest, and k-nearest neighbors were not convincing enough for us to recommend using this model for any serious applications.