Ph125.9X Capstone Grown-ups

Jonathan Behar

6/16/2019

1. Overview

The salary housing market and often result in a disparity in the income pay gap. Between men and woman, in this paper, I will try to perform exploratory data analysis and develops a machine learning algorithm to predict whether an individual will have an income earning greater than 50,000 annually utilizing data derived from a data-set on Kaggle.

Several predictive algorithms were trained, including a GLM, KNN, Random Forest, and rPart models using a series of continuous and categorical predictors or covariates to explain variances in earning potential. The accuracy of the models was utilized to evaluate performance and determine the 'best-fit' prediction.

2. Data-set and Package

Loading the Data-set

This data-set used from Kaggle was originally extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The following code is used to download the data-set.

```
url <-
"https://github.com/jdb443/ph125.9x_Capstone_Project/blob/master/grownhuman.csv?raw=true"
salary<- read.csv(url)
```

Loaded Packages

The packages used to perform the analysis include:

- tidyverse
- caret
- corrplot
- gridExtra
- ggplot2
- dplyr
- data.table
- ROCR

- rpart
- randomForest

3. Data Exploration

General properties of the data-set

The income data-set has about 32000 plus rows and 15 columns.

```
dim(salary)
```

The structure of this data-set indicates that missing values in the file has been represented by a question mark (?). These were replaced in the table with NA's and tallied, revealing the presence of several missing data points.

```
sapply(salary,function(x) sum(is.na(x)))
```

Dependent Variables

The categorical outcome to be predicted is income. It has two categories greater than 50,000 and less than or equal to 50,000, with 24.080% of the entries representing individuals earning more than 50,000 a year and 75.919% representing individuals earning less than 50,000 a year. Based on over 32000 data points.

Table 1. Category Totals

income	Count
<=50K	24720
>50K	7841

Independent Variables/A Priori Predictors

The headers provided in each column in the data-set are listed below.

1. Age: age
2. Employment status: workclass
3. Demographic weighting in the data-set: fnlwgt
4. Educational achievement: education
5. Numerical representing/ranking of educational achievement: education.num
6. Marital status: marital.status
7. Occupation: occupation
8. Familiar relationships: relationship
9. Race: race
10. Sex: sex

11. Capital gain: capital.gain
12. Capital loss: capital.loss
13. Hours of work per week: hours.per.week
14. Country of origin: native.country

Several of the predictors also had various levels or unique attributes. Table one shows that the demographic weighting (fnlwgt)has over 26,000 levels, and as a result, is also excluded from the analysis. It is removed because such a large number of levels would suggest that it will not group individuals into identifiable demographic categories that can significantly contribute to the prediction of an individual's income level.

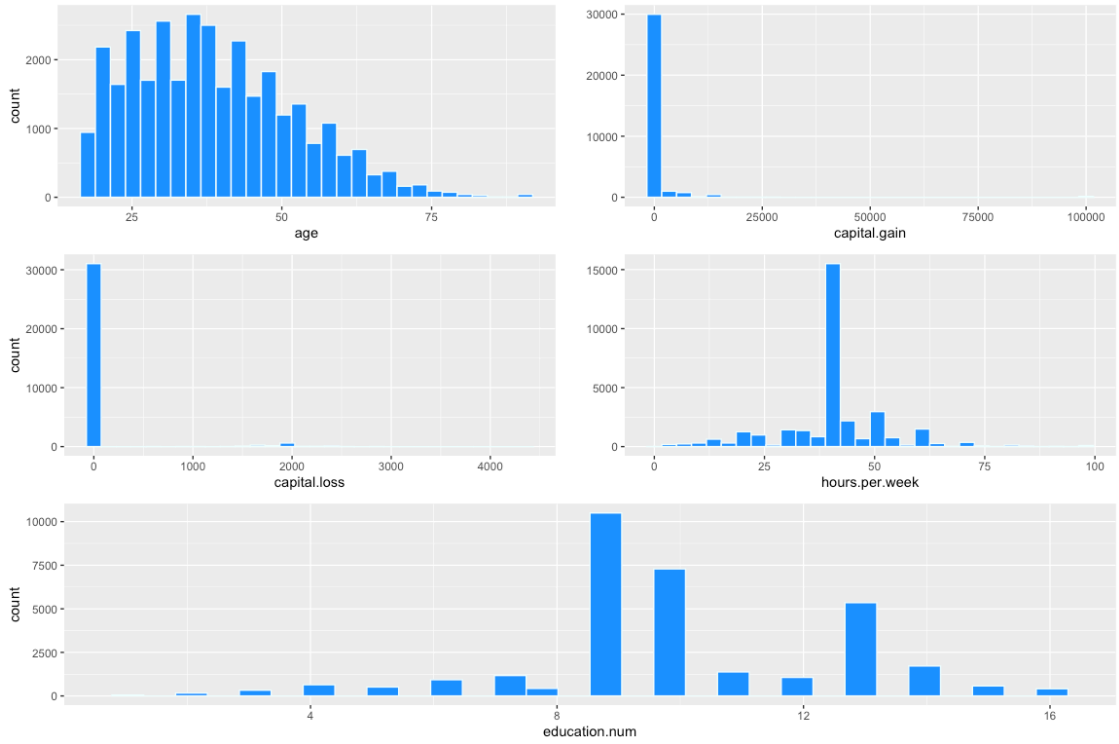
Each predictor gives information regarding the characteristics of an individual. However, several predictors have similarities. For example, "education" and "education.num" or education number provides a similar representation for educational achievement, and thus one of the two can probably be omitted or split into a separate table. The relationship is also omitted from the ongoing analysis as the two predictors "marital.status" and "relationship" share similar vital attributes, as a person's marital status describes a person's relationship with another.

Table 2. Unique Levels

	Unique Levels
age	73
Workclass	9
fnlwgt	21648
education	16
educationnum	16
maritalstatus	7
occupation	15
relationship	6
race	5
sex	2
capitalgain	119
capitalloss	92
hoursperweek	94
nativecountry	42

Further examination of the remaining continuous predictors by histogram plots displayed in Object 1 shows that Capital gain and capital loss consists of mainly zeros.

Object 1. Frequency of Predictors



Zeros account for 91.7% of the entries within the capital gain predictor and 95.3% zeroes in the capital loss predictor, which is a total of 29849 and 31042 entries successively from each of the respective predictors. The significant number, so zeros in both predictors are neither skewed to being related to a person earning less than or equal to 50,000 or greater than 50,000 and thus excluded from the ongoing analysis.

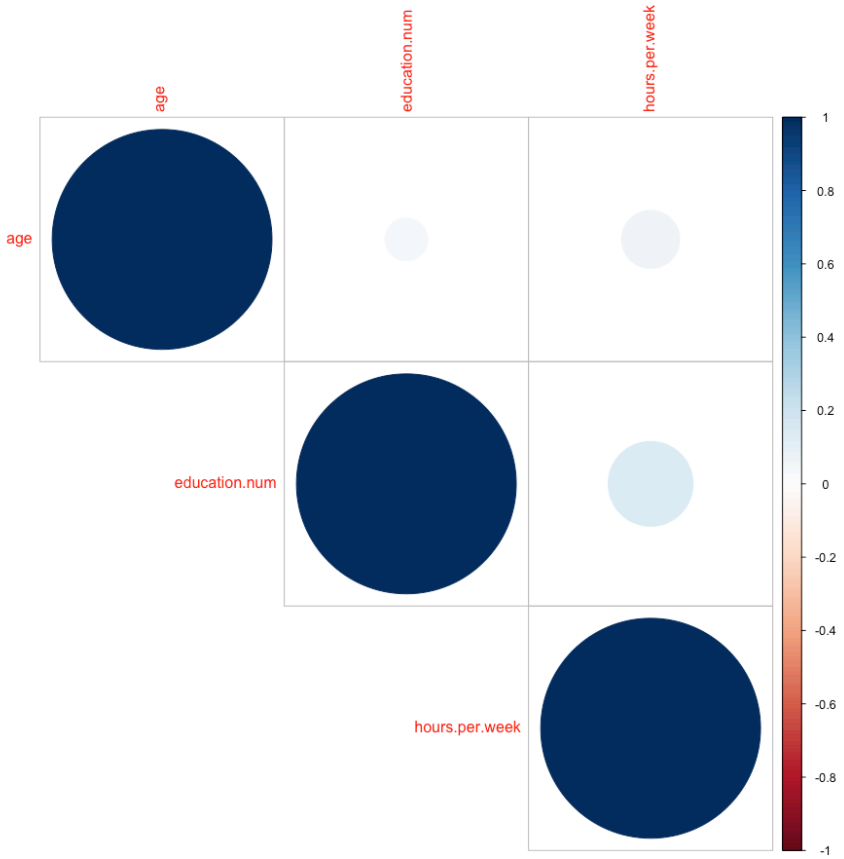
```
sum(salary$capital.gain==0)/length(salary$capital.gain)
sum(salary$capital.gain==0)
sum(salary$capital.loss==0)/length(salary$capital.loss)
sum(salary$capital.loss==0)
```

The correlation of the remaining determinants age, education.num, and hour.per.week, which are displayed in Object 2 show that they are not highly correlated.

Table 3. Correlation of Predictors

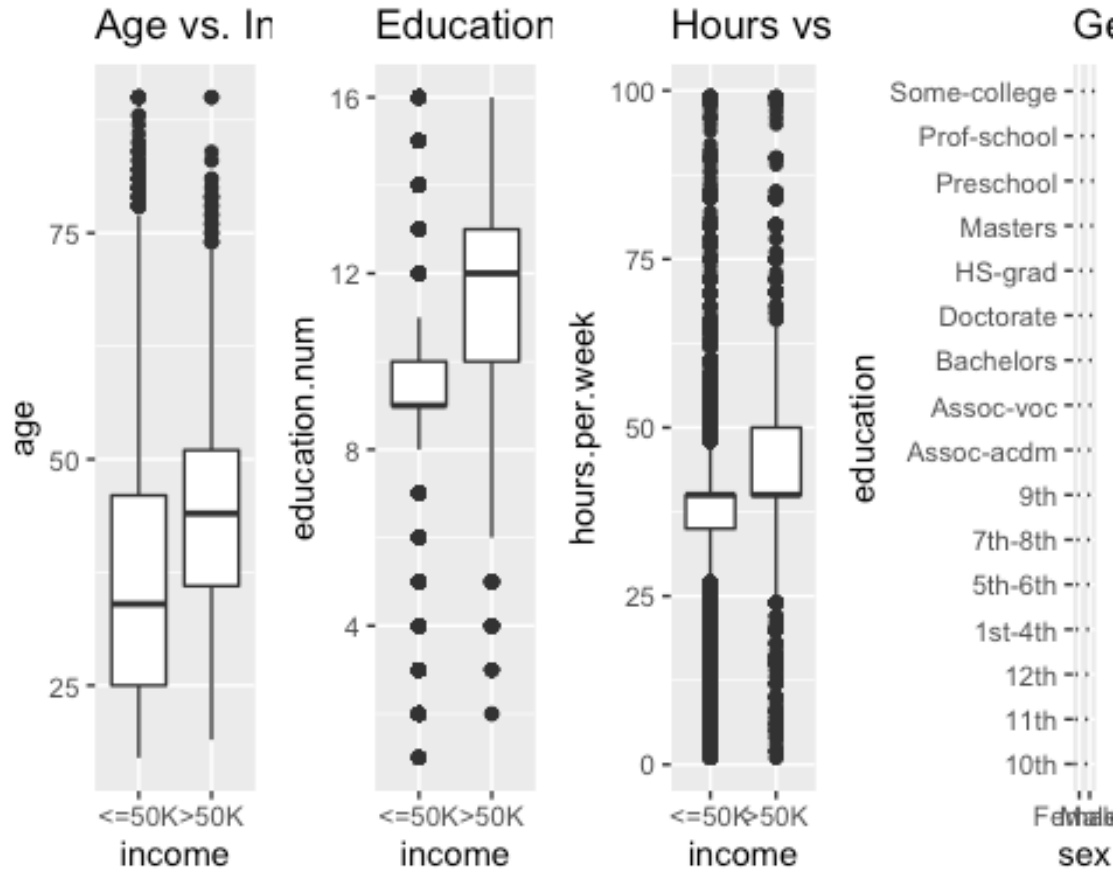
	age	education.num	hours.per.week
age	1.000	0.037	0.069
education.num	0.037	1.000	0.148
hours.per.week	0.069	0.148	1.000

Object 2. Illustration of Correlation of Predictors



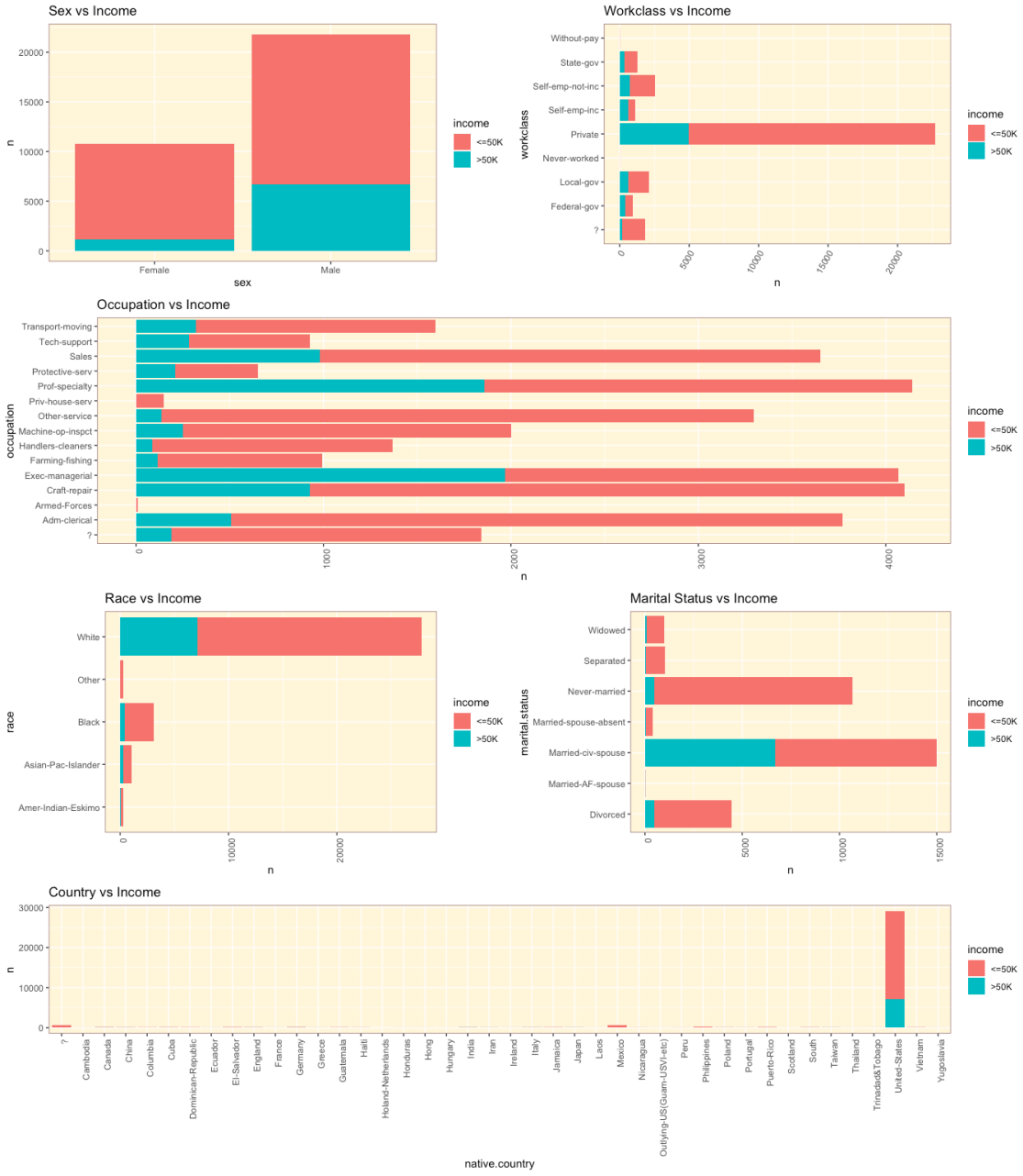
The three non-correlated continuous predictors show variations with income level. The box plots suggest that older individuals, those with higher educational achievements and working more than 40 hours per week have higher numbers earning more than 50,000 annually (greater than 50,000) and can significantly influence variation.

Object 3. Continuous Predictors and Income



Exploration of the categorical predictors indicates that one predictor (native.country) does not have significant variation and can be excluded from the ongoing analysis. The histograms also suggest that white males in the private job class with a professional specialty or in executive, managerial positions who are married-civ spouse have more significant numbers earning more than 50,000 annually (greater than 50,000) and can significantly influence variation.

Object 4. Categorical Predictors and Income



4. Preprocessing and Model Fit Results

Removing predictor columns

Considering the findings after the analysis of the raw data-set was updated, and several predictors were removed. These included fnlwgt, education, relationship, the capital.gain, the capital.loss, native.country, and rows, including missing data.

Data Partitioning

The categorical dependent variable is transformed to a binary factor with 0 representing an earning of less than or equal to 50,000 and 1 representing an earning of greater than 50,000 and then the data-set was then split into a training set denoted as train and a validation set denoted as test, which is used to test the algorithm for predicting income. The test set represents 10% of the data-set.

```
set.seed(1)
test_index <- createDataPartition(y = salary$income, times = 1, p = 0.10,
list = FALSE)
train <- salary %>% slice(-test_index)
test <- salary %>% slice(test_index)
```

Fit Model

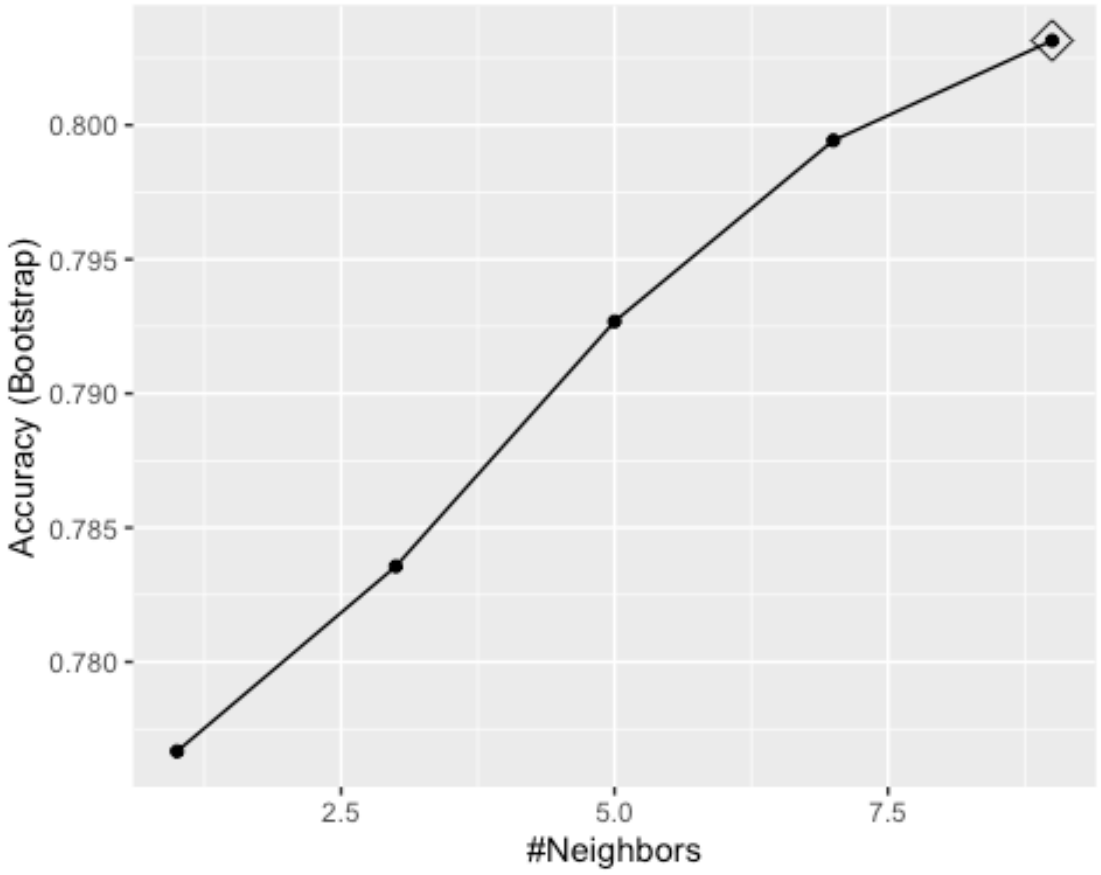
- GLM: Logistic Regression

```
fitglm <-glm(income ~.,family=binomial(link='logit'),data=train)
phat<- predict(fitglm, test, type = "response")
yhat<- ifelse(phat>0.5, 1,0) %>% factor()
confusionMatrix(yhat, test$income)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8369665
```

- KNN: k-Nearest Neighbor

```
fitknn <- train(income ~ ., method = "knn", tuneGrid = data.frame(k = seq(1,
10, 2)), data = train)
ggplot(fitknn, highlight = TRUE)
```




```
fitknn$bestTune

##   k
## 5 9

fitknn <-knn3(income ~.,data=train, k=9)
yhatknn<-predict(fitknn, test, type="class")
confusionMatrix(yhatknn, test$income)$overall["Accuracy"]

##  Accuracy
## 0.8234572
```

Recursive Partitioning and Regression Trees: RPART

```
fitrpart<-train(income~., method="rpart", data=train)
yhat<-predict(fitrpart, test, type="raw")
confusionMatrix(yhat, test$income)$overall["Accuracy"]

##  Accuracy
## 0.8249923
```

- Random Forest

```
fitrf<- randomForest(income~.,data= train)
yhat<-predict(fitrf, test, type="class")
confusionMatrix(yhat, test$income)$overall["Accuracy"]

##  Accuracy
## 0.8406509
```

5.Conclusion

After using the Adult Income Census data from Kaggle. Many inferences can be drawn about the data. After the use of several models, the highest accuracy of 0.8406509 was established by a Random Forest model with predictors age, workclass, education, occupation, race, sex, hours of work per week and marital status.

Link to Github Movielens Project