Predict whether income exceeds $50K/yr based on census data

Haichen Dong

5/16/2019

Project Overview

In this project, I will apply machine learning techniques to predict whether income exceeds $50K/yr based on census data. In the initial stage, we would like explore the data by using some statistic methods and graphics. After understanding the data, data cleaning and feature selection will be applied in data file for future data analysis. In this process, data records with missing field will be removed, and some features do not fit in our analysis models also be removed. Data will be randomly separated to two parts, 80 percent data for training and 20 percent for testing. Finally we can implementing some machine leaning models to predict the income, checking whether income exceeds $50K/yr. Base on the accuracies of all the models, we can pick best model for this predicting.

Project Data

1. Data Collection

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). Data can be downloaded from: https://www.kaggle.com/uciml/adult-census-income/downloads/adult-census-income.zip/3

2. Data Attributes:

income: >50K, <=50K

age: continuous

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

fnlwgt: continuous

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool

education-num: continuous

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

sex: Female, Male

capital-gain: continuous

capital-loss: continuous

hours-per-week: continuous

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

Prepare Data:

1. Feature Selection and Data Clean:

Some features in this dataset are not fit my analysis model.

- "capital.gain", too many zeros (see Figure 1).
- "capital.loss", too many zeros (see Figure 2).
- "native.country", most values are U.S. (see Figure 3).
- "education-num", same feature with "education".
- "fnlwgt", too many unique values and no significant different between >50K and <=50K (see Figure 4).

All the data with missing value(s) has been removed to clean data and improve the result.

Figure 1

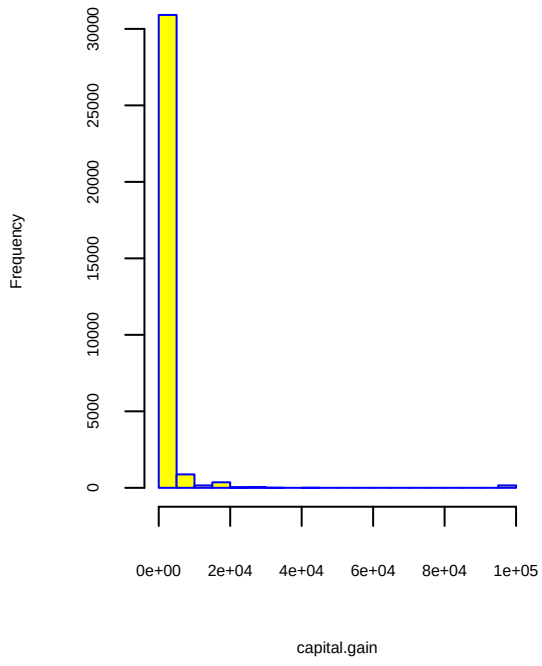


Figure 2

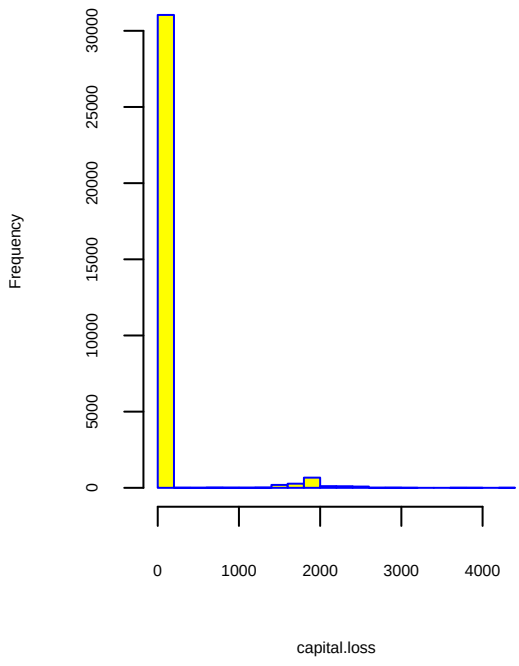
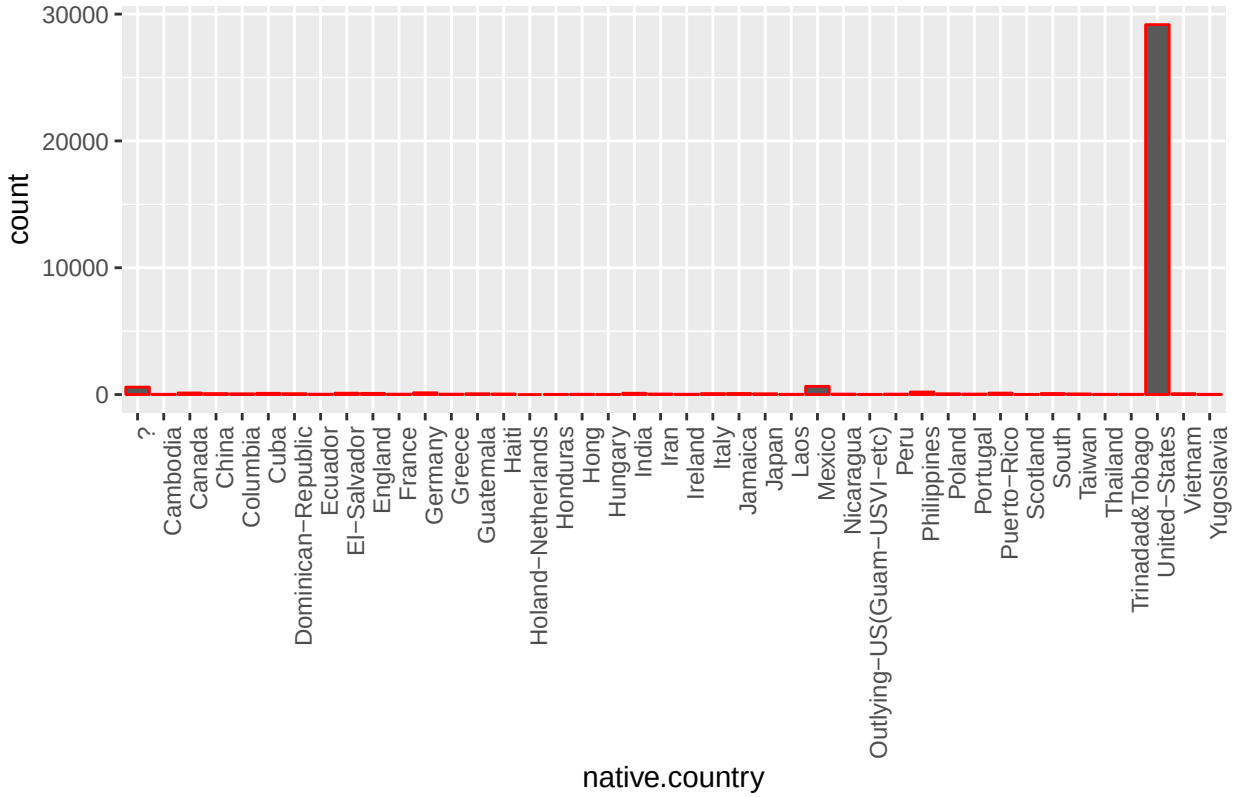
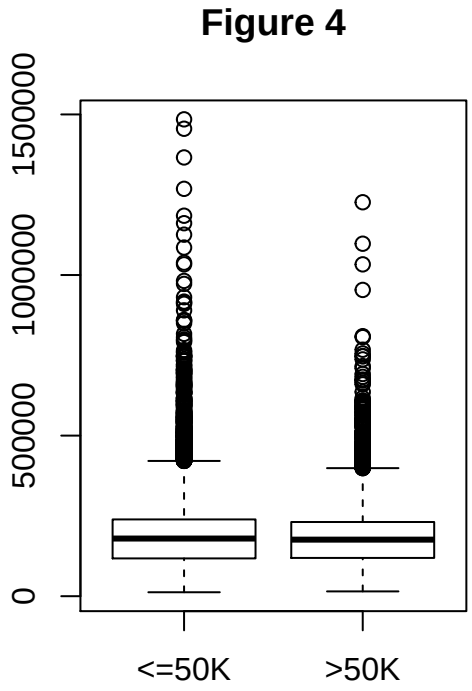


Figure 3


30000
20000
10000
0
count
?
Cambodia
Canada
China
Columbia
Cuba
Dominican−Republic
Ecuador
El−Salvador
England
France
Germany
Greece
Guatemala
Haiti
Holand−Netherlands
Honduras
Hong
Hungary
India
Iran
Ireland
Italy
Jamaica
Japan
Laos
Mexico
Nicaragua
Outlying−US(Guam−USVI−etc)
Peru
Philippines
Poland
Portugal
Puerto−Rico
Scotland
South
Taiwan
Thailand
Trinadad&Tobago
United−States
Vietnam
Yugoslavia
native.country



2. Data Summary:

There are 30,718 records and 9 features in this dataset. The "income" is the field we will predict. There are about 25% data with >50K and 75% data with <=50K. There are two continuous featrues: age and hours.per.week; and seven categorical features.

```
## [1] 30718    10
```

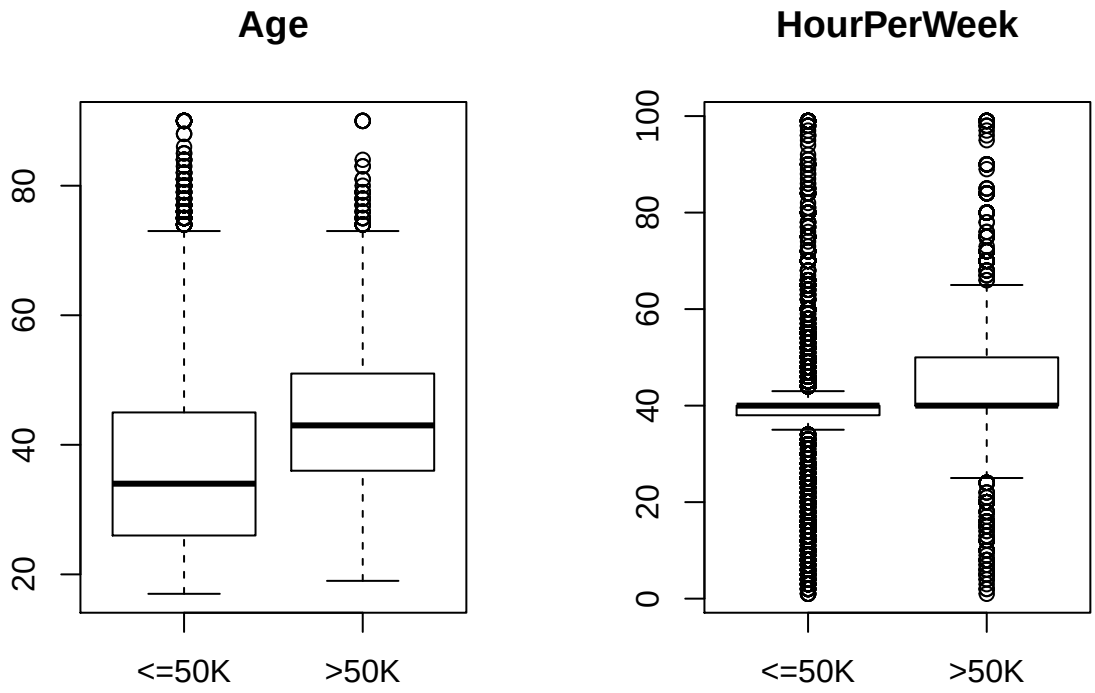
```
##
##     <=50K      >50K
## 0.7509603 0.2490397
```

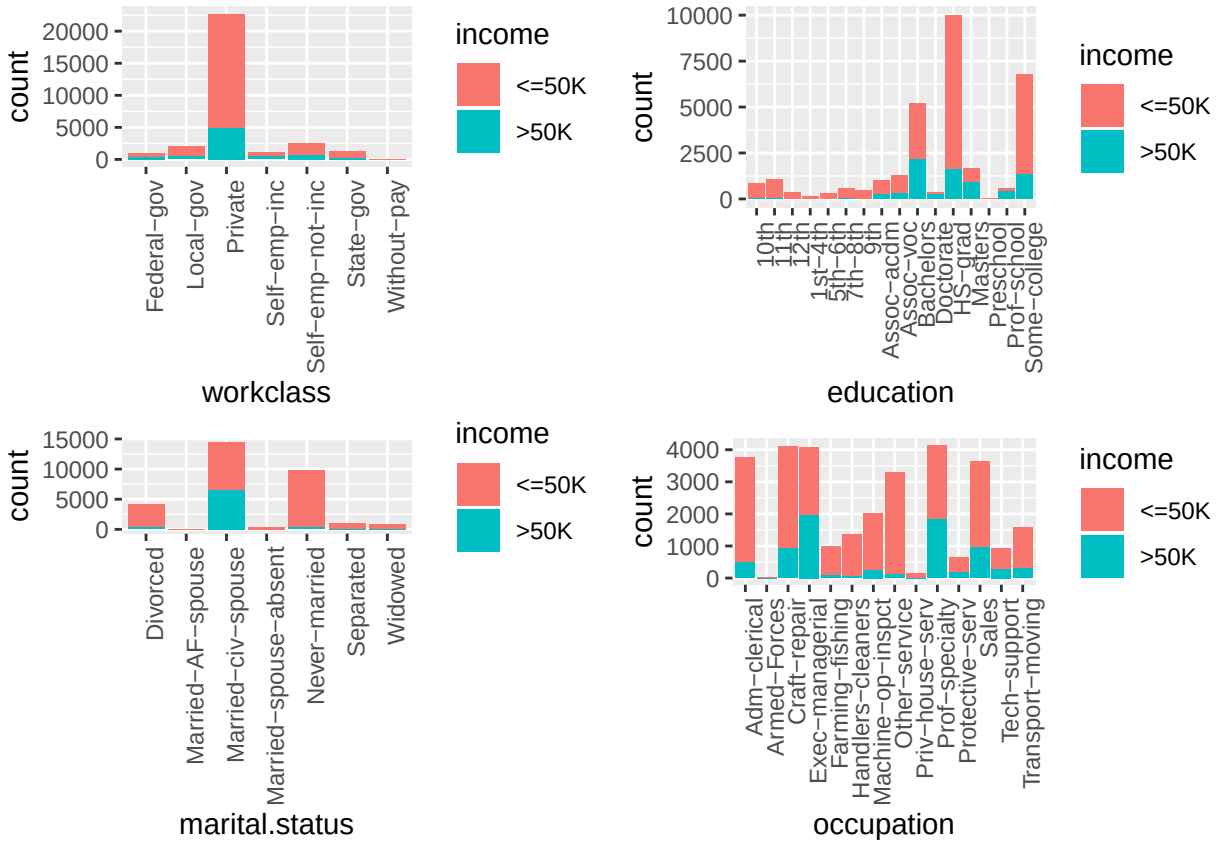
```
##       age                    workclass            education
##  Min.   :17.00   Private         :22696   HS-grad     :9968
##  1st Qu.:28.00   Self-emp-not-inc: 2541   Some-college:6775
##  Median :37.00   Local-gov       : 2093   Bachelors   :5182
##  Mean   :38.44   State-gov       : 1298   Masters     :1675
##  3rd Qu.:47.00   Self-emp-inc    : 1116   Assoc-voc   :1321
##  Max.   :90.00   Federal-gov     :  960   11th        :1056
##                  (Other)         :   14   (Other)     :4741
##                marital.status            occupation
##  Divorced              : 4258   Prof-specialty :4140
##  Married-AF-spouse     :   21   Craft-repair   :4099
##  Married-civ-spouse    :14339   Exec-managerial:4066
##  Married-spouse-absent:  389   Adm-clerical   :3770
##  Never-married         : 9912   Sales          :3650
##  Separated             :  959   Other-service  :3295
```

```
##  Widowed              :  840   (Other)        :7698
##          relationship                    race           sex
##  Husband       :12704   Amer-Indian-Eskimo:  286   Female: 9930
##  Not-in-family : 7865   Asian-Pac-Islander:  974   Male  :20788
##  Other-relative:  918   Black             : 2909
##  Own-child     : 4525   Other             :  248
##  Unmarried     : 3271   White             :26301
##  Wife          : 1435
##
##  hours.per.week    income
##  Min.   : 1.00   <=50K:23068
##  1st Qu.:40.00   >50K : 7650
##  Median :40.00
##  Mean   :40.95
##  3rd Qu.:45.00
##  Max.   :99.00
##
```

3. Check Feature Data :

From the following graphics, we can see the feature data are different in >50K and <=50K.




count
20000
15000
10000
5000
0
Federal−gov
Local−gov
Private
Self−emp−inc
Self−emp−not−inc
State−gov
Without−pay
workclass
income
<=50K
>50K
count
10000
7500
5000
2500
0
10th
11th
12th
1st−4th
5th−6th
7th−8th
9th
Assoc−acdm
Assoc−voc
Bachelors
Doctorate
HS−grad
Masters
Preschool
Prof−school
Some−college
education
income
<=50K
>50K
count
15000
10000
5000
0
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed
marital.status
income
<=50K
>50K
count
4000
3000
2000
1000
0
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
occupation
income
<=50K
>50K

4. Prepare Training/Testing Data

Separate our data to training data and testing data. createDataPartition function has been used to randomly separate data, with 80 percent for training and 20 percent for testing. After separation, >50K and <=50K rates in both dataset are very same as the original dataset 25% vs 75%. There are 24574 records in traing dataset and 6144 records in testing dataset.

```
## [1] 24574    10
```

```
## [1] 6144   10
```

```
##
##         0         1
## 0.7509563 0.2490437
```

```
##
##         0         1
## 0.7509766 0.2490234
```

Build Models

1. Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. We also can group some data items for more simple and conclusive.

```
##    Accuracy Sensitivity Specificity
##   0.7903646   0.8047248   0.7470588
```

2. Logistic Regression

Logistic Regression is the go-to method for binary classification problems.It gets better accuracy then Naive Bayes in our dataset.

```
##    Accuracy Sensitivity Specificity
##   0.8149414   0.9523190   0.4006536
```

3. Stepwise Logistic Regression

We use both direction in stepwise logistic regression to improve the accucracy from 0.81 to 0.82.

```
##    Accuracy Sensitivity Specificity
##   0.8240560   0.9161248   0.5464052
```

4. Nearest Neighbor

KNN has easily been the simplest to pick up. Despite it's simplicity, it has proven to be incredibly effective at certain tasks. We have auto picked the K for best result. We found 7 is the optimal number for our dataset.

```
##    Accuracy Sensitivity Specificity
##   0.7991536   0.8784135   0.5601307
```

5. Random Forest

The last model that I would like to use is Random Forests. we have optimized our tree with minNode and predFixed parameters.

```
##   predFixed minNode
## 1         2       1
```

```
##    Accuracy Sensitivity Specificity
##   0.8221029   0.9289120   0.5000000
```

6. Ensembles

After trying all those models, I would like improve the final result by combining the results of two different algorithms, Random Forest and Stepwise Logistic Regression. The result is little better than any of them.

```
##    Accuracy Sensitivity Specificity
##   0.8240560   0.9055050   0.5784314
```

Conclusion

In this project, I use five different models to predict data. Nine features have been used in all the models. From the simple model, Stepwise Logistic Regression wins the best model for predicting person's income exceeds $50K/y in this census data; Naïve Bayes has good balance in sensitivity and specificity; and Ensembles can improve the final results. For further work, I would like try group some data items to make data simple, like the age or education. Also try reducing some features.

```
## # A tibble: 6 x 2
##   method                       accuracy
##   <chr>                           <dbl>
## 1 Naive Bayes                     0.790
## 2 Logistic Regression             0.815
## 3 Stepwise Logistic Regression    0.824
## 4 Nearest Neighbor                0.799
## 5 Random Forest                   0.822
## 6 Ensembles                       0.824
```

Project File Link at GitHub : https://github.com/Haichen-Dong/Adult_CensusData_Project