Predicting Adult Census Income Level in R

Juhe Nie

2020/2/14

Introduction

In this project, we will use the US adult census income data to create predictive model to predict if the income of any individual is greater than or less than USD 50000. The datasets used for this analysis is donated to the public site http://archive.ics.uci.edu/ml/machine-learning-databases/adult. We will use three datasets from this website: "adult.data", "adult.test" and "adult.names". The "adult.data" set is used to build training model, the "adult.test" is used to do final test. We use the "adult.names" to extract variable names and add them as column names for training and test sets.

We first download and read these three datasets and change the missing data from " ?" to "NA".

Our datasets include 15 variables: six are integers and nine are factors. Their basic description is in the table below. The task of this project is to predict the variable "income" using other objects.

Variable name	Description	Type
age	age of the individual	continuous
workclass	class of work	categorical (8 levels)
fnlwgt	final weight	continuous
education	the highest education level	categorical (16 levels)
education.num	number of education years	continuous
marital.status	marital status of the individual	categorical (7 levels)
occupation	occupation of the individual	categorical (14 levels)
relationship	present relationship	categorical (6 levels)
race	race of the individual	categorical (5 levels)
sex	sex of the individual	categorical (2 levels)
capital.gain	capital gain made by the individual	continuous
capital.loss	capital loss made by the individual	continuous
hours.per.week	average number of working hours for each week	continuous
native.country	native country of the individual	categorical (41 levels)
income	income of the individual	categorical (2 levels)

We find that 7.37% of data in train set and 7.49% of in test set is "NA", to make our data cleaner, we remove all "NA"s from train set and test set.

Now there are 30162 observations in train set and 15060 in test set.

Methods

Visualization

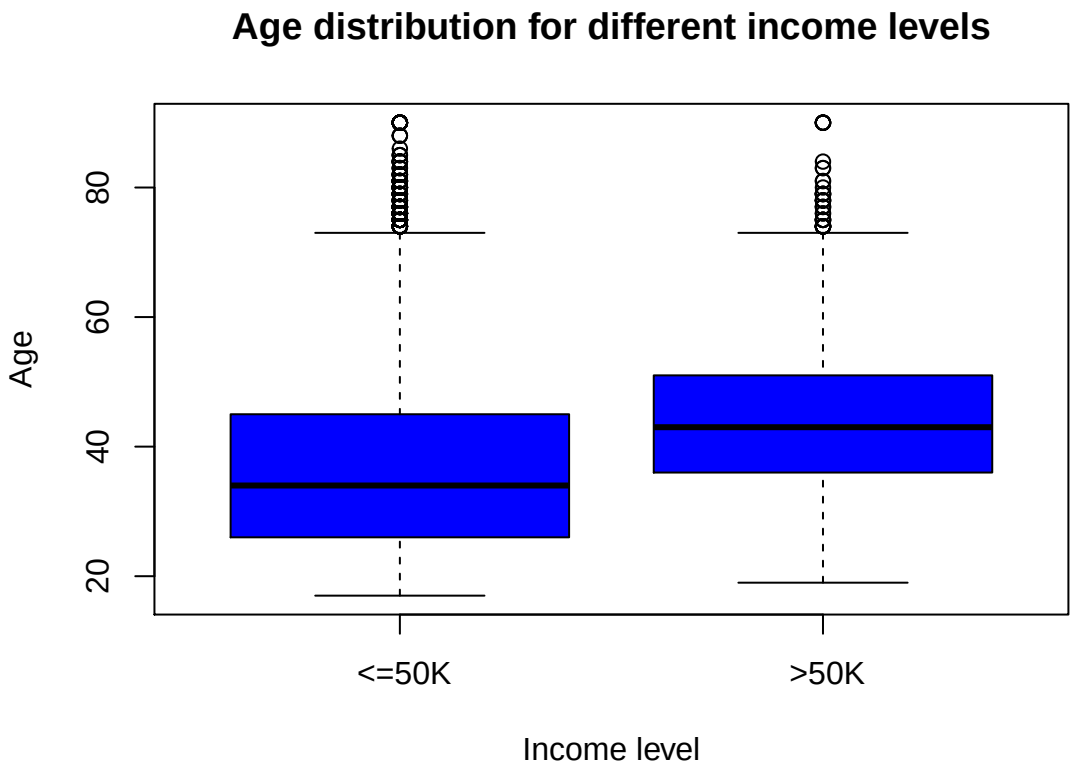
We want to do more data exploration of the train set and make the data visualizable. There are 30162 entries in train set, among which 22654 individuals have income lower than 50K and 7508 individuals have income higher than 50K.

```
## # A tibble: 2 x 2
##   income       n
##   <fct>    <int>
## 1 " <=50K" 22654
## 2 " >50K"   7508
```

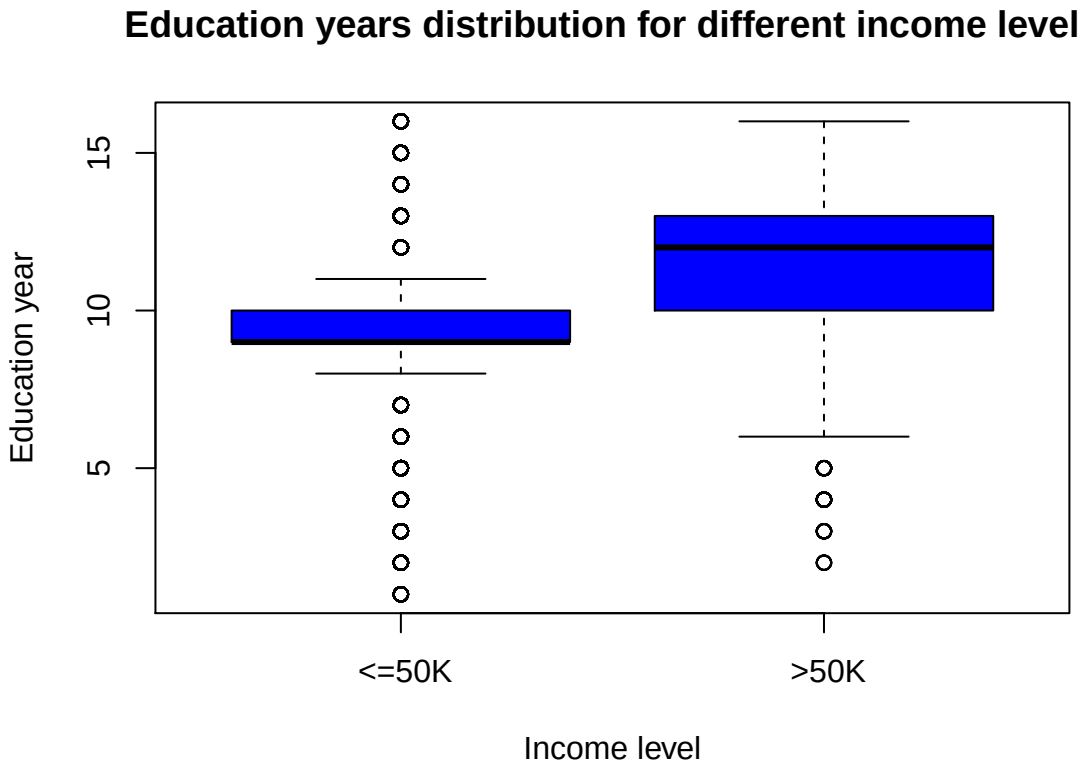
Continuous variable

We then go into the continuous variables: age, education year, capital loss, capital gain, fnlwgt and working hours per week. We use boxplot to illustrate the distribution of each variable for different income levels. Because fnlwgt has no relationship to income prediction, we will not analyze this variable and also remove it from our train set.

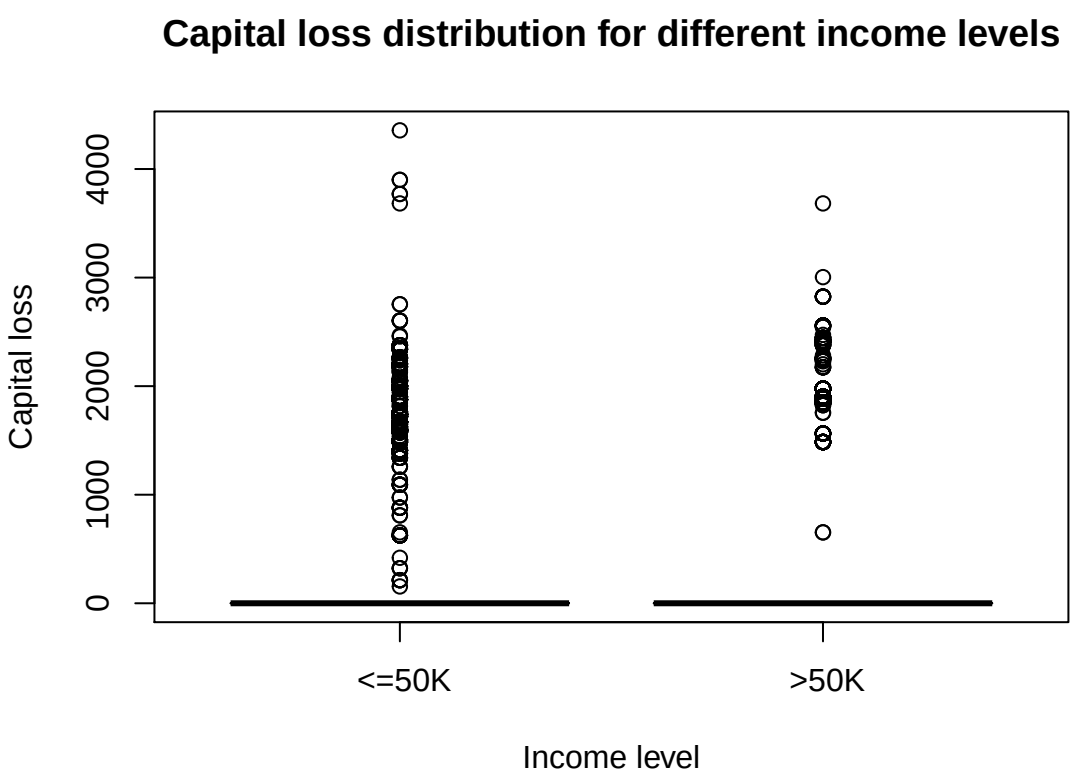
The variable "age" has a wide range and variability. Its distribution and mean are quite different for income level lower than 50K and higher than 50K, so we think "age" will be a good predictor of "income".

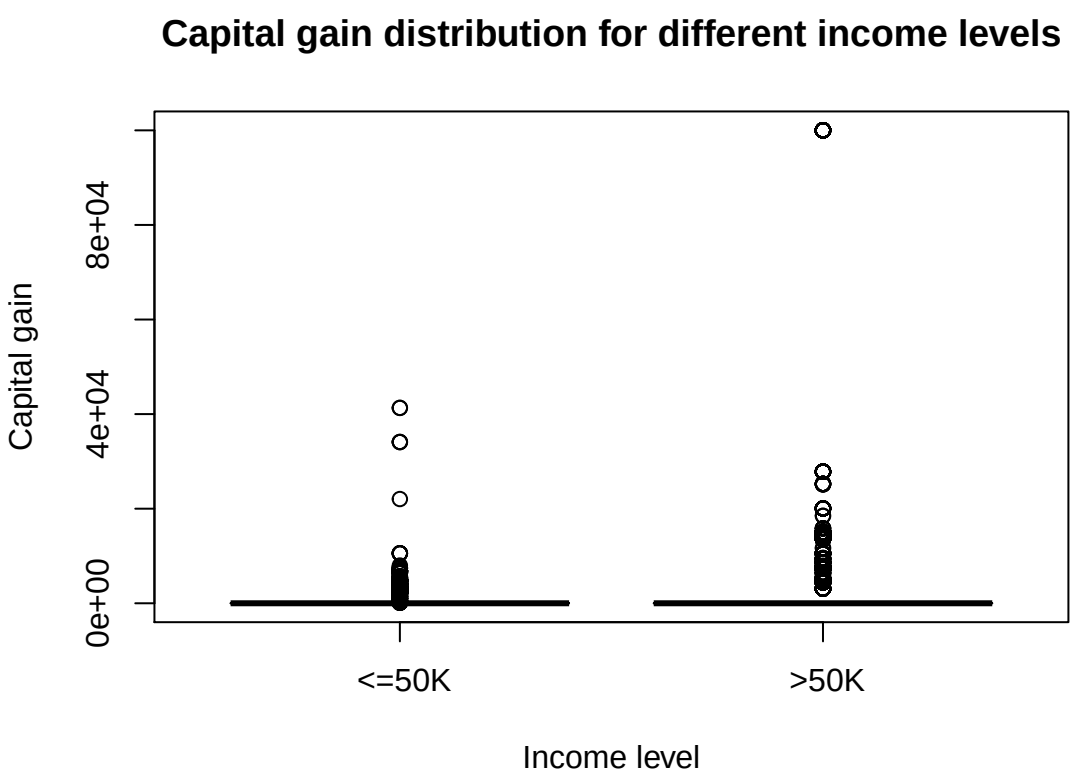


The distribution of education years is quite different between different income levels, and it also has a good variability, so "education.num" is a good predictor.

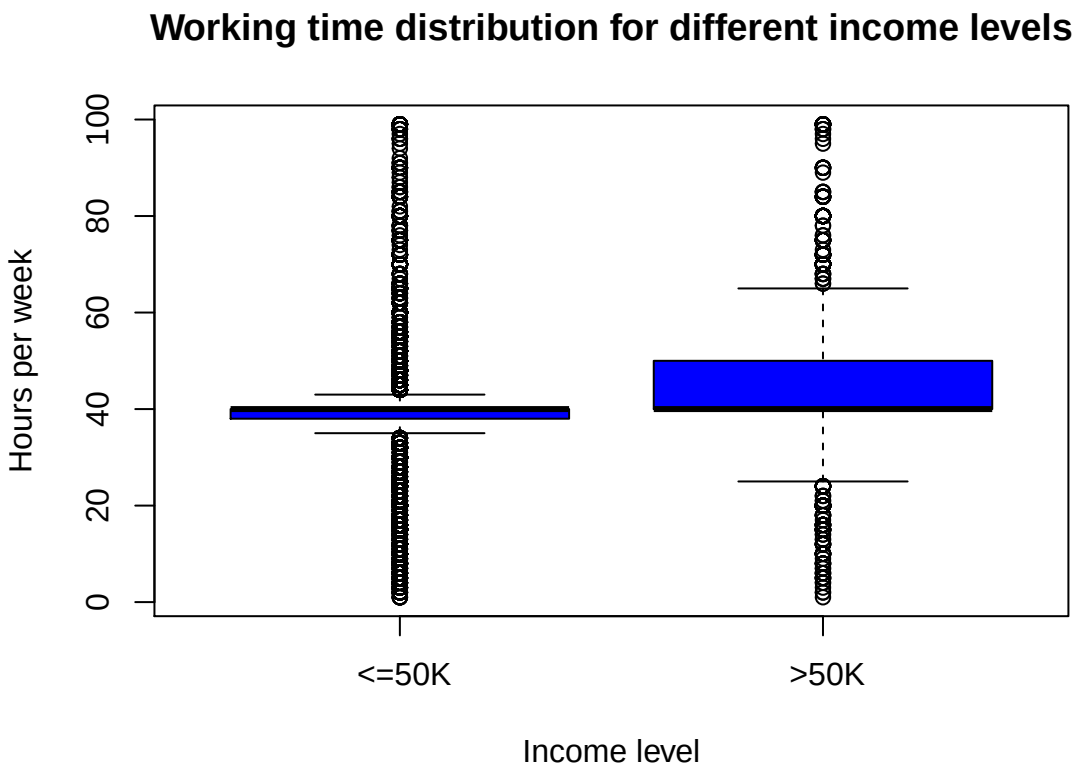


The capital gain and capital loss don't show many differences in different income levels from their boxplots. We find that 91.59% of data in capital gain and 95.2% of data in capital loss have the value 0. We think capital.gain and capital.loss are not good predictors and they will not be used in our prediction model since they don't show much variance.


Capital loss distribution for different income levels
Capital loss
0
1000
2000
3000
4000
<=50K
>50K
Income level



Working hours per week shows a good variability, implying it is a good predictor.



We then want to see whether there are some correlations between age, education years and working hours. The table below shows that these variables are independent.

```
##                      age education.num hours.per.week
## age            1.00000000    0.04352609      0.1015988
## education.num  0.04352609    1.00000000      0.1525221
## hours.per.week 0.10159876    0.15252207      1.0000000
```

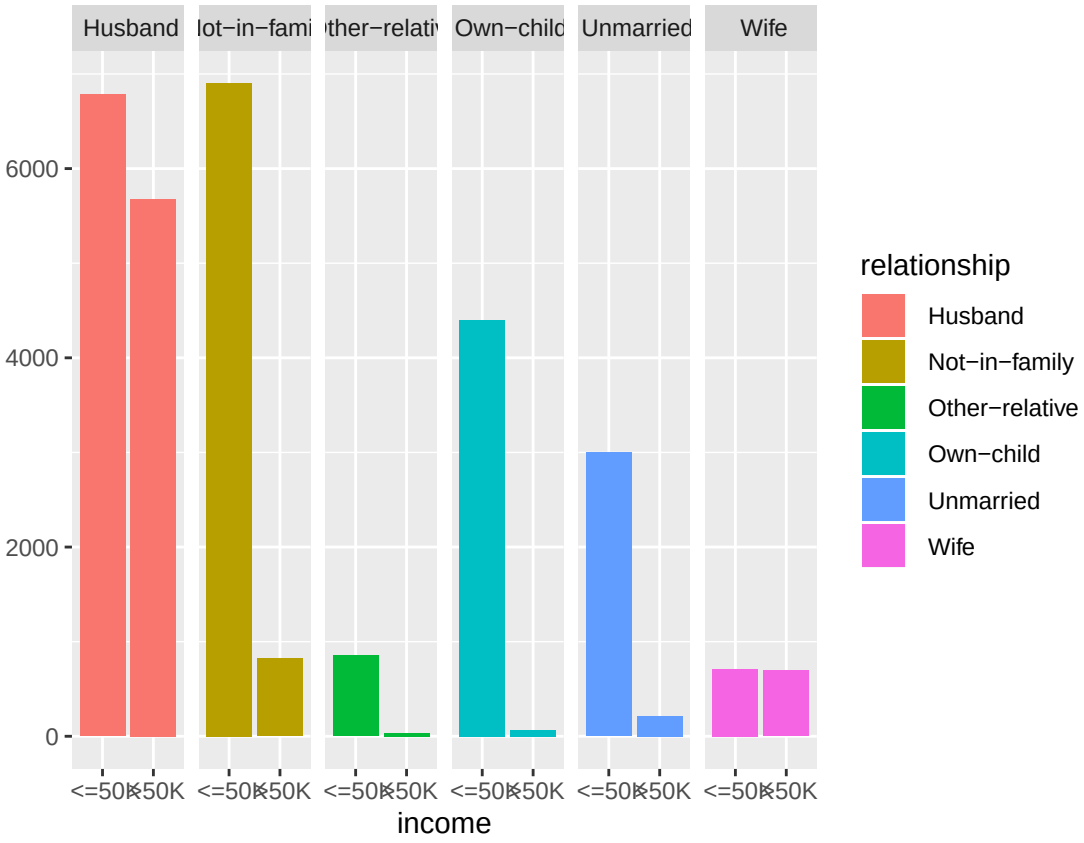
Categorical variable

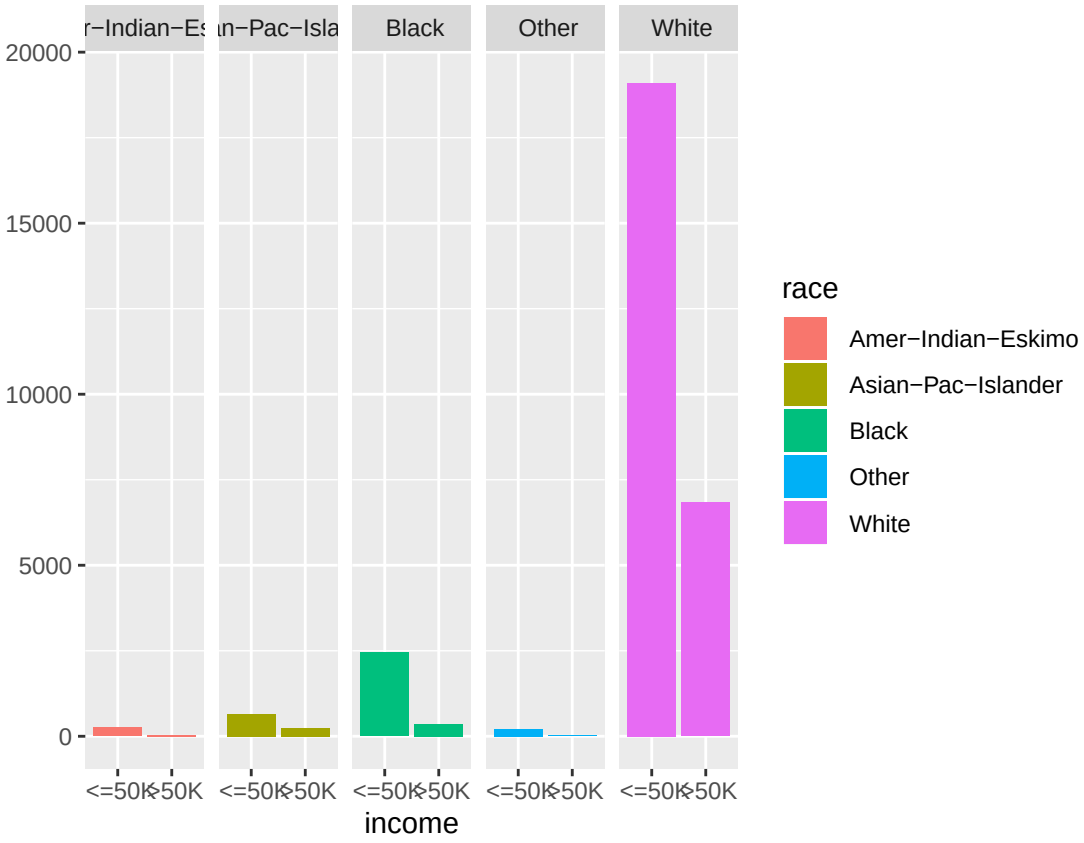
We then go into the categorical variables: “sex”, “relationship”, “race”, “marital.status”, “workclass”, “occupation”, “education” and “native.country”. Since we have found that “education.num” (number of education years) is a very good predictor and the information from “education” is quite similar to “education.num”, to avoid overweighting on education area, we will remove “education” variable from datasets.

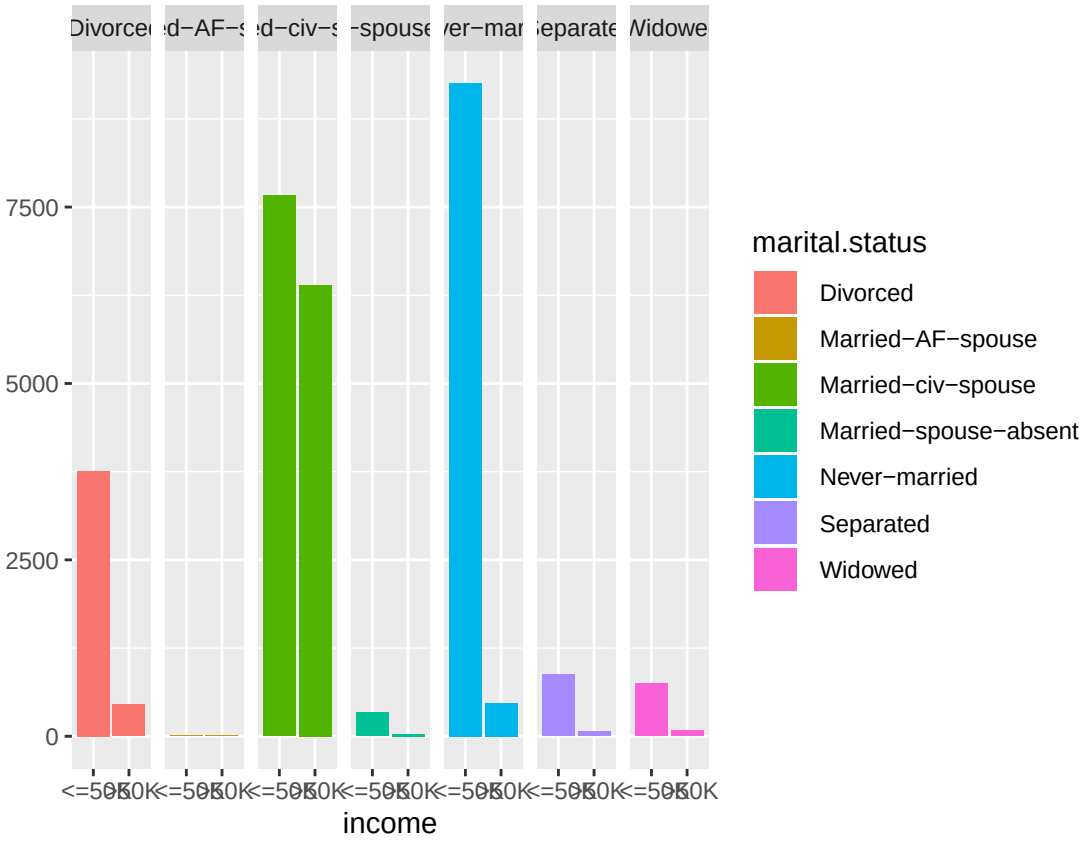
Regarding sex, the table below shows that 88.6% female have income lower than 50K and 68.6% male have income lower than 50K. Sex shows a good variance, implying it is a good predictor.

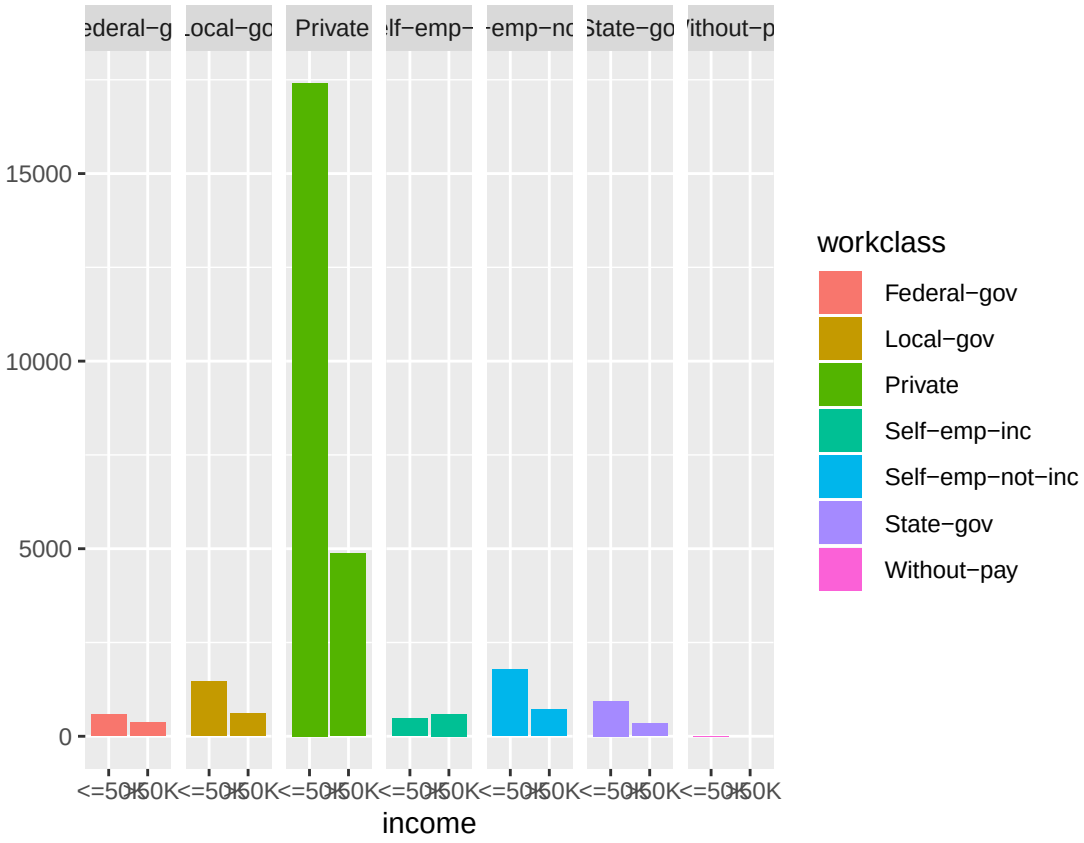
```
## # A tibble: 2 x 3
##   sex       lower higher
##   <fct>     <dbl>  <dbl>
## 1 " Female" 0.886  0.114
## 2 " Male"   0.686  0.314
```

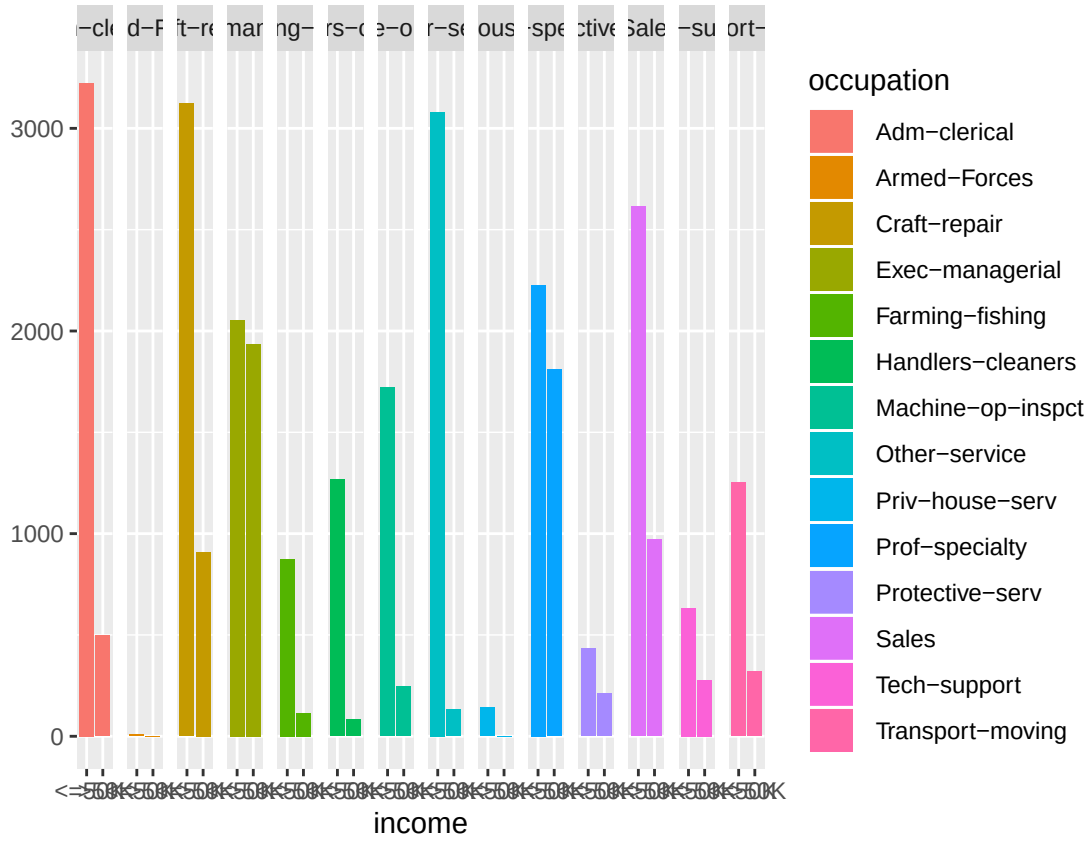
The five figures below illustrate that relationship, race, marital status, workclass and occupation are all good predictors for income level due to the variant behaviours of factor levels for each variable.


Husband
lot−in−fami
ther−relati
Own−child
Unmarried
Wife
6000
4000
2000
0
<=50K
>50K
income
relationship
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife


r−Indian−Es
n−Pac−Isla
Black
Other
White
20000
15000
10000
5000
0
<=50K
>50K
income
race
Amer−Indian−Eskimo
Asian−Pac−Islander
Black
Other
White


Divorce
d−AF−s
d−civ−s
−spouse
ver−mar
eparate
Widowe
7500
5000
2500
0
<=50K
>50K
income
marital.status
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed


ederal−g
ocal−go
Private
lf−emp−
emp−no
State−go
ithout−p
15000
10000
5000
0
<=50K
>50K
income
workclass
Federal−gov
Local−gov
Private
Self−emp−inc
Self−emp−not−inc
State−gov
Without−pay



Regarding native country, we notice that there are many different factor levels (41 countries)from different continents, but more than 90% of individuals are from United States. Due to the data from the other 40 countries are insufficient, we don't think these data can show the relationship between native countries and income levels well, so we will not use native country as a predictor in our prediction model.

Prediction model

Now we will start building prediction model. In this case, we will try three different mothods: logistics regression, decision tree and support vector machine (SVM). As we mentioned before, we first remove "fnlwgt", "capital.gain", "capital.loss", "education" and "native.country" from train set.

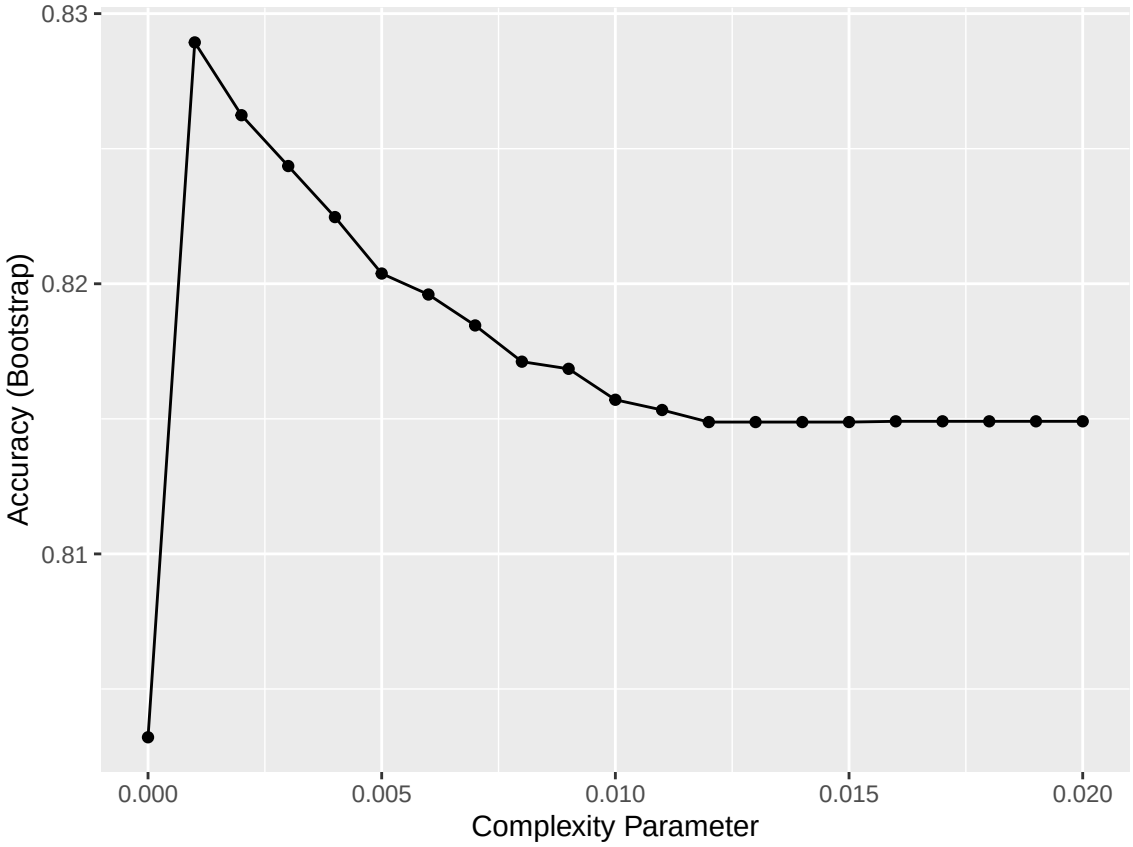
We then divide our train set into two parts: 90% in learning set and 10% in validation set. The learning set is used to build model and validation set is used to verify the models. There are 3017 entries in validation set, which is good enough to make validation. At the same time, we want to as many data used in training as possibile, so dividing train set into 90% and 10% is reasonable.

Model 1: Logistics Regression

In model 1, we apply the logistics regression method. We found that there is a warning message when we use workclass as a predictor, so in this model we will only use the other 8 predictors to build model. We use confusion matrix, in which " <=50K" is regarded as positive class, to calculate overall accuracy, sensitivity and specificity. The overall accuracy is 0.8283, the sensitivity is 0.8612 and the specificity is 0.6951. This logistics regression model performs better when predicting true positive than predicting true negative.

Model 2: Decision Trees

In model 2, we use the dicision tree as predictive model. We train the learning data with all the predictors with complexity parameter cp from 0 to 0.02. As the plot below shows, the tree has the best accuracy when cp is 0.001.



Here we describe the tree briefly: the tree starts division with whether marital.status is "Married-civ-spouse", and then divided according to education years, then divided based on different attributes in different branches.

We use confusion matrix again to test the result in validation set: accuracy is 0.8373, sensitivity is 0.8690, specificity is 0.7124. The sensitivity is higher than specifity.

Model 3: SVM

In model 3, we use support vector machine model with all predictors. The accuracy of SVM is 0.8369, sensitivity is 0.8600 and specificity is 0.7342.

Results

Here we summarize the results in validation set in the table below:

Model	Accuracy	Sensitivity	Specificity
logistics regression	0.8283	0.8612	0.6951
decision tree	0.8373	0.8690	0.7124

Model	Accuracy	Sensitivity	Specificity
SVM	0.8369	0.8600	0.7342

For these three models, they all have better sensitivity than specificity. This makes sense because we have more lower than 50K observations than higher than 50K ovservations in our dataset. Decision tree has the best overall accuracy result and sensitivity, while SVM performs the best in predicting true false (specificity).

We then use these three models to predict the test set. The result is presented below:

Model	Accuracy	Sensitivity	Specificity
logistics regression	0.828	0.8639	0.6842
decision tree	0.8321	0.8658	0.6961
SVM	0.8321	0.8593	0.7117

In our final test, still all these three models have better sensitivity than specificity. The decision tree model and SVM model have the best overall accuracy. Decision tree performs best on sensitivity and SVM performs best on specificity.

Conclusion

In this project, we use US adult census income data to predict individual's income levels with multiple variables. We first download and clean the train set and test set. We then explore train_set data and make visualization in order to see which variables are good predictors for income level and which are not. Next, divide train_set data into learning set and validation set. We use three models to train learning data respectively and use validation set to verify. The three models are logistics regression model, decision tree model and SVM model. Finally, we use these three model to predict the income level for test_set. The result shows that all these three models have better sensitivity than specificity. The decision tree model and SVM model have a better overall accuracy than logistics regression, decision tree has the best sensitivity and SVM has the best specificity.