title: “Income Prediction”

author: “Lin Li”

date: “July 21, 2020”

Introduction

In this report, the **Adult** was used to create a **Income Prediction Algorithm** that can be used to predict whether a person makes over $50K a year.

The **Adult dataset** consists of 32561 observations with 15 variables.

The data was pulled directly from the kaggle website (https://www.kaggle.com/uciml/adult-census-income). The data can’t be automatically downloaded unless a registration with the website is finished. Thus, the dataset will be uploaded with this report in https://github.com/l98033110/Havardx-Capstone.

The raw dataset was explored, cleaned up, wrangled to become a more useable subset and then split into trainset dataset and the testset dataset.

Accuracy is the target parameter to improve.

Two models were trained using trainset and evaluated on testset. More effective model is **Random Forest**.

Using this method, an **Accuracy** of **0.83** was obtained. In the last part of report, codes for tuning of rain forest model are provided but not run due to very slow response from my laptop. Some of parameters in codes can be adjusted accordingly based on results. I believe that a higher accuracy can definitely be achieved after further tuning.

Data Analysis

Read Data

The raw datasets were pulled directly from the kaggle website and saved to a file called income_data.

Adult Dataset

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90         ?  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66         ? 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
```

```
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3                 ?     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             18  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States  <=50K
```

Data Preprocessing

All duplicated obeservations are removed. Column fnlwgt stands for final weight which is not useful in prediction and is removed. capital gain and loss columns are removed and combined into one column (capital_net) using the formula gain - loss. All observations with ? input are removed and income_num column is added with inputs according to values in income column (1 for “>50K”, 0 for “<50K”). All character columns are converted to factor.

```
## # A tibble: 2 x 2
##   income     n
##   <chr>  <int>
## 1 <=50K  24698
## 2 >50K    7839
```

Above table shows number of observations for each income level. There are only two income levels.

Processed Dataset

```
##   age workclass    education education.num marital.status
occupation
## 1  82   Private      HS-grad             9        Widowed   Exec-
managerial
## 2  54   Private      7th-8th             4       Divorced Machine-op-
inspct
## 3  41   Private Some-college            10      Separated    Prof-
specialty
## 4  34   Private      HS-grad             9       Divorced     Other-
service
## 5  38   Private         10th             6      Separated      Adm-
clerical
## 6  74 State-gov    Doctorate            16  Never-married    Prof-
specialty
##     relationship  race    sex hours.per.week native.country income
capital_net
## 1  Not-in-family White Female             18  United-States  <=50K       -
4356
## 2      Unmarried White Female             40  United-States  <=50K       -
```

```
3900
## 3      Own-child White Female             40  United-States  <=50K        -
3900
## 4      Unmarried White Female             45  United-States  <=50K        -
3770
## 5      Unmarried White   Male             40  United-States  <=50K        -
3770
## 6 Other-relative White Female             20  United-States   >50K        -
3683
##   income_num
## 1          0
## 2          0
## 3          0
## 4          0
## 5          0
## 6          1
```

Explore/Visualize/Clean Categorical Data

Education

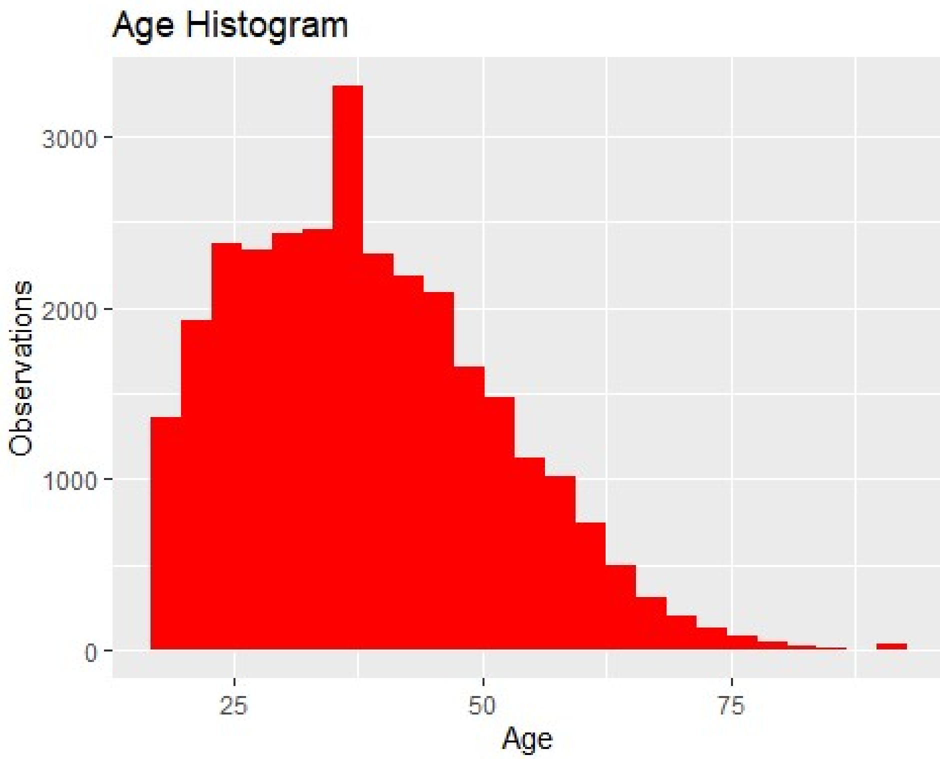
In below tables, n is number of observations and means of each education level are calculated using new added column income_num (1 for ">50K", 0 for "<50K"). Higher education level may result in a higher possibility of well-paid employment.And education and education.num are the duplicate information for prediction. Thus education column is removed.

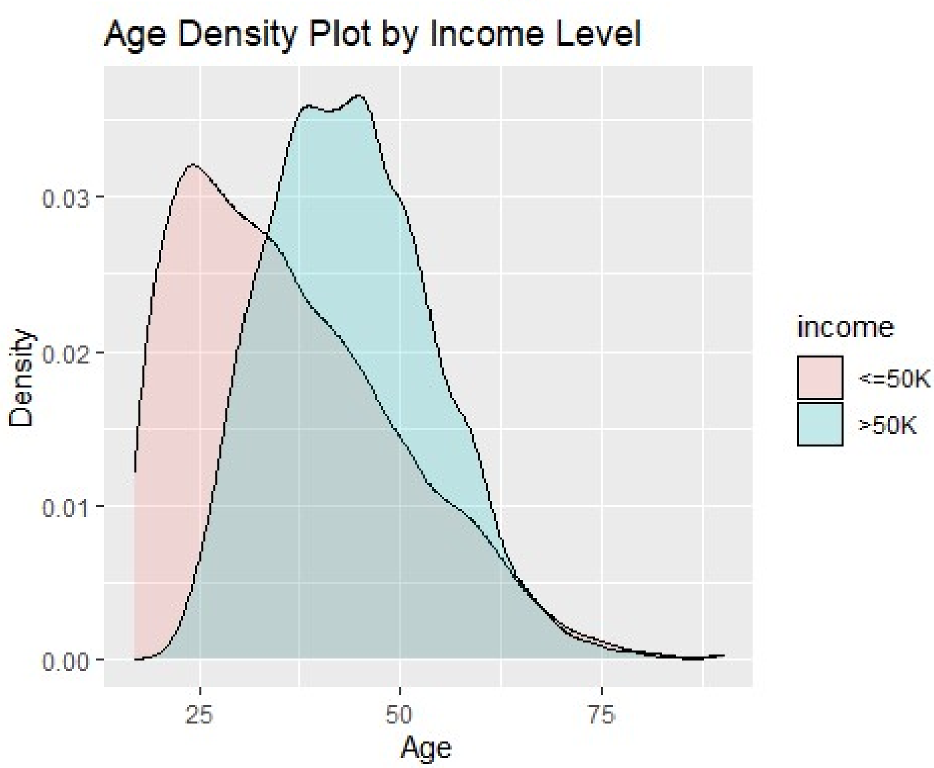
```
## # A tibble: 16 x 3
##    education      mean     n
##    <fct>         <dbl> <int>
##  1 Prof-school  0.749    542
##  2 Doctorate    0.747    375
##  3 Masters      0.565   1626
##  4 Bachelors    0.422   5042
##  5 Assoc-voc    0.263   1307
##  6 Assoc-acdm   0.254   1008
##  7 Some-college 0.200   6669
##  8 HS-grad      0.164   9834
##  9 12th         0.0769   377
## 10 10th         0.0720   820
## 11 7th-8th      0.0629   556
## 12 11th         0.0563  1048
## 13 9th          0.0549   455
## 14 5th-6th      0.0418   287
## 15 1st-4th      0.0403   149
## 16 Preschool    0           44

## # A tibble: 16 x 3
##    education.num   mean     n
##            <int>  <dbl> <int>
```

```
##  1               15 0.749    542
##  2               16 0.747    375
##  3               14 0.565   1626
##  4               13 0.422   5042
##  5               11 0.263   1307
##  6               12 0.254   1008
##  7               10 0.200   6669
##  8                9 0.164   9834
##  9                8 0.0769   377
## 10                6 0.0720   820
## 11                4 0.0629   556
## 12                7 0.0563  1048
## 13                5 0.0549   455
## 14                3 0.0418   287
## 15                2 0.0403   149
## 16                1 0          44
```

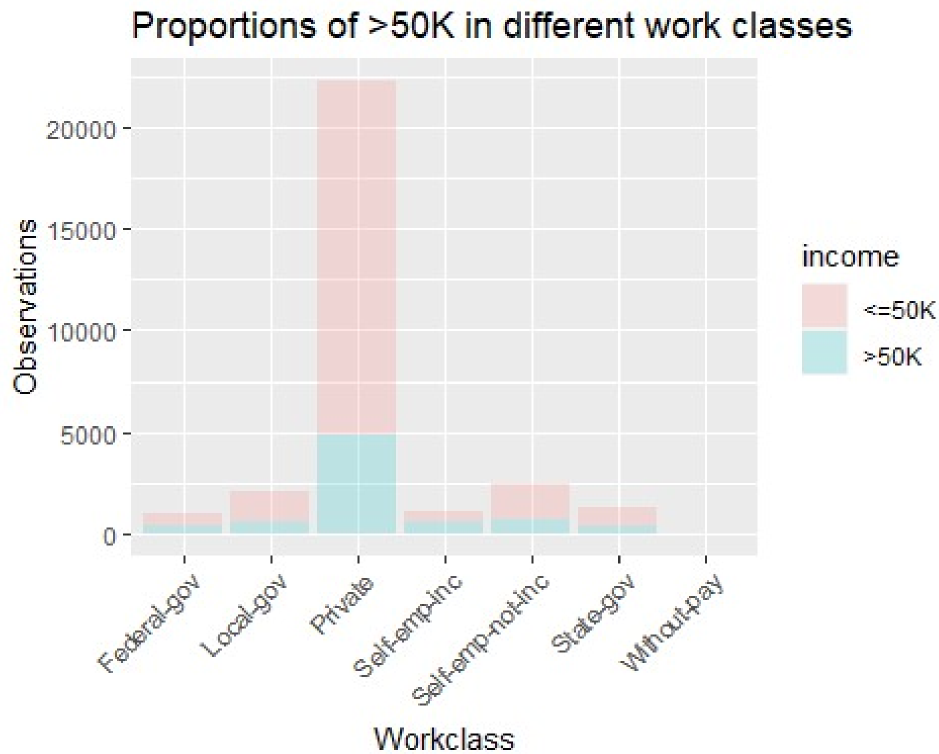
Age

Below figures indicate that older people have higher possibility of higher income.



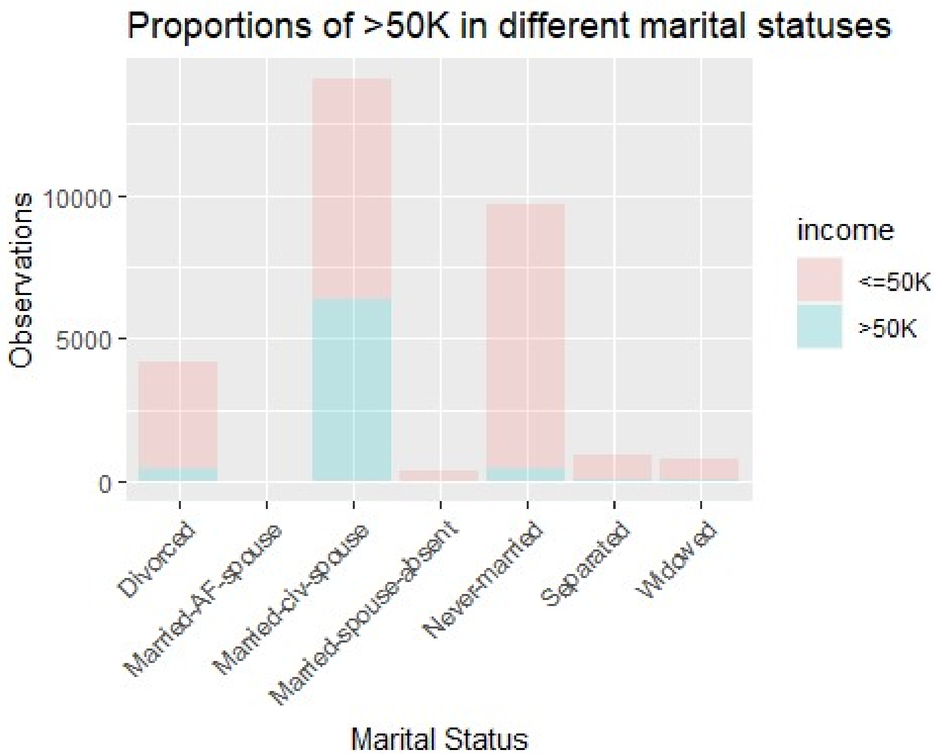


Work Class



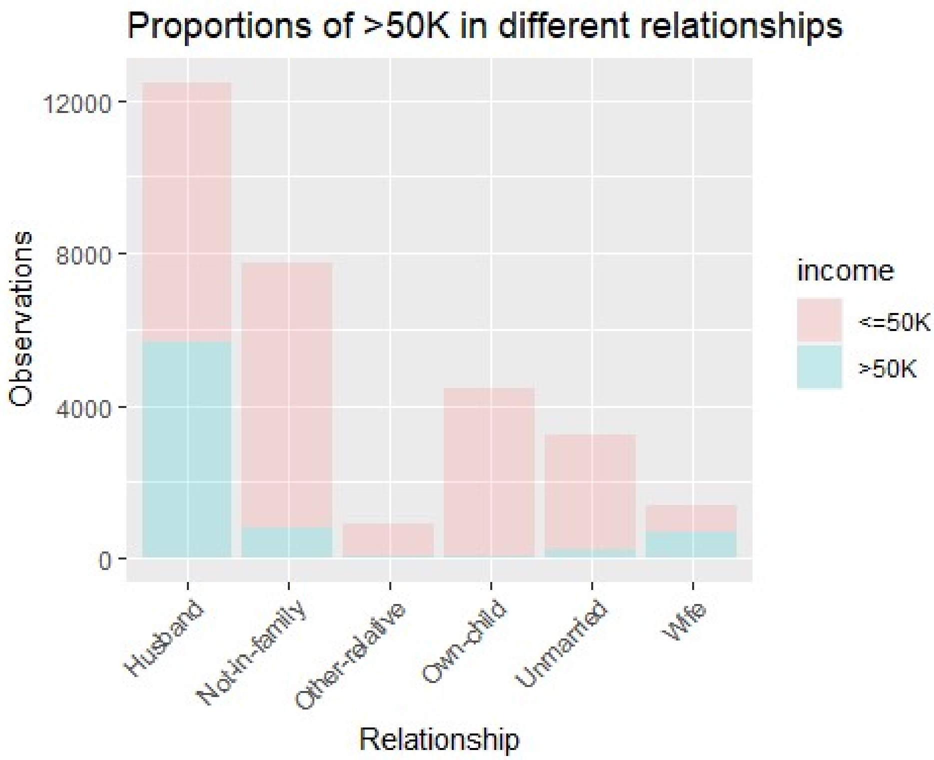
The private sector has the most people who earn more than 50K per year and has the largest number of population. However, in terms of the proportion, the self-employed people are the winner.

Marital Status



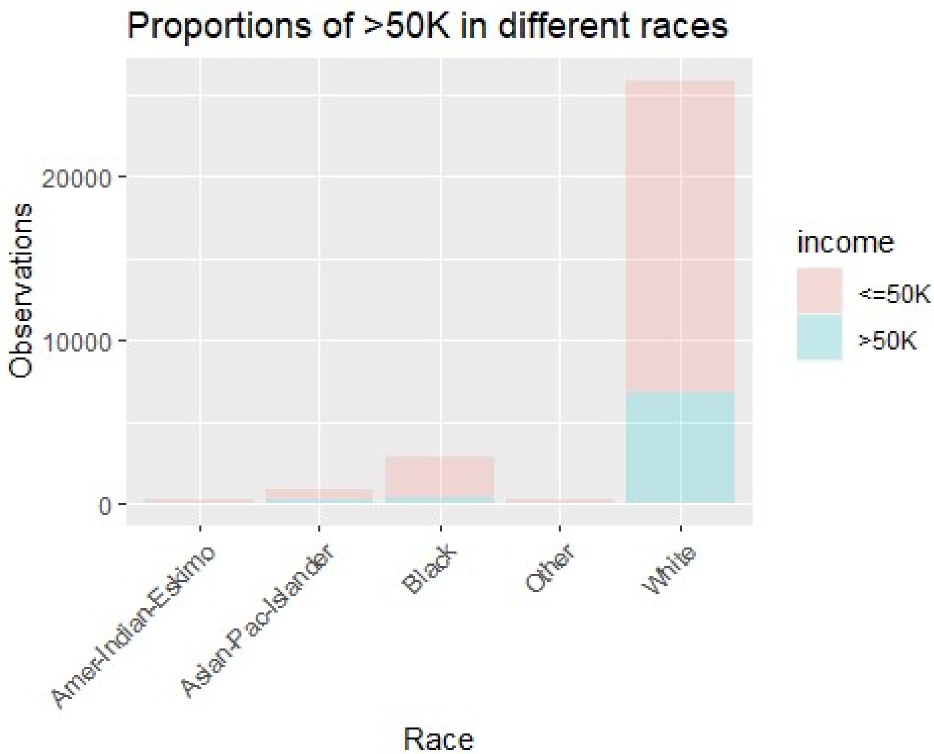
The figure indicates that married status has the most people who earn more than 50K per year and has the largest number of population.

Relationship



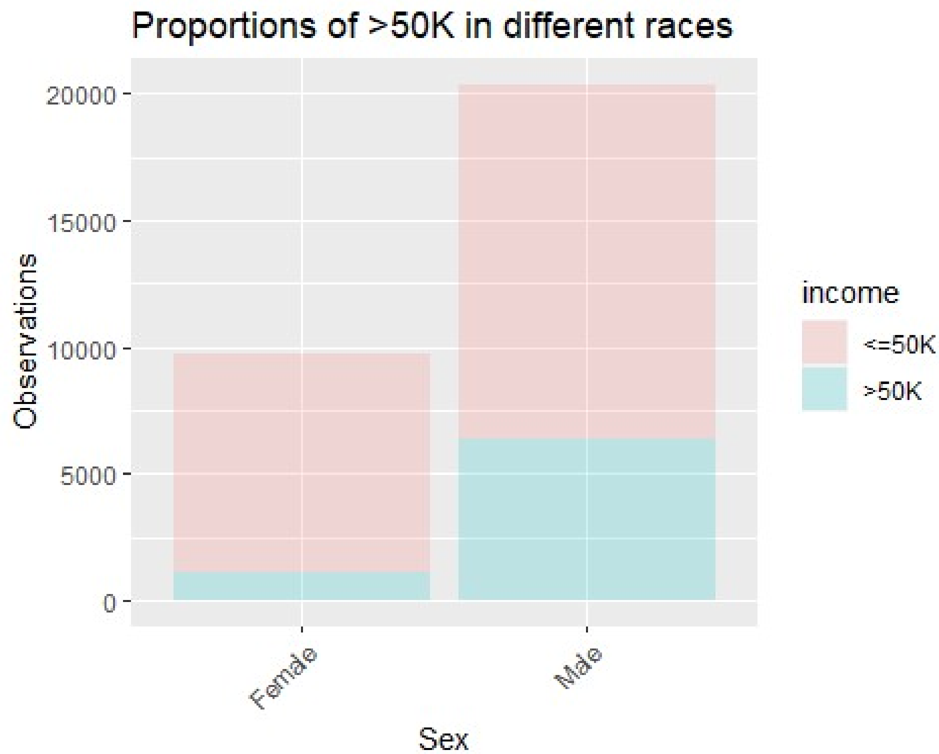
Husband and wife only contribute marital status and gender information which are indicated by sex and marital status columns. The figure indicates that husband and wife have higher proportions of observations who earn more than 50K per year which is already reflected in Marital Status. According to graphs, marital status and relationship tell us the same thing. So relationship column is removed from the model.

Race



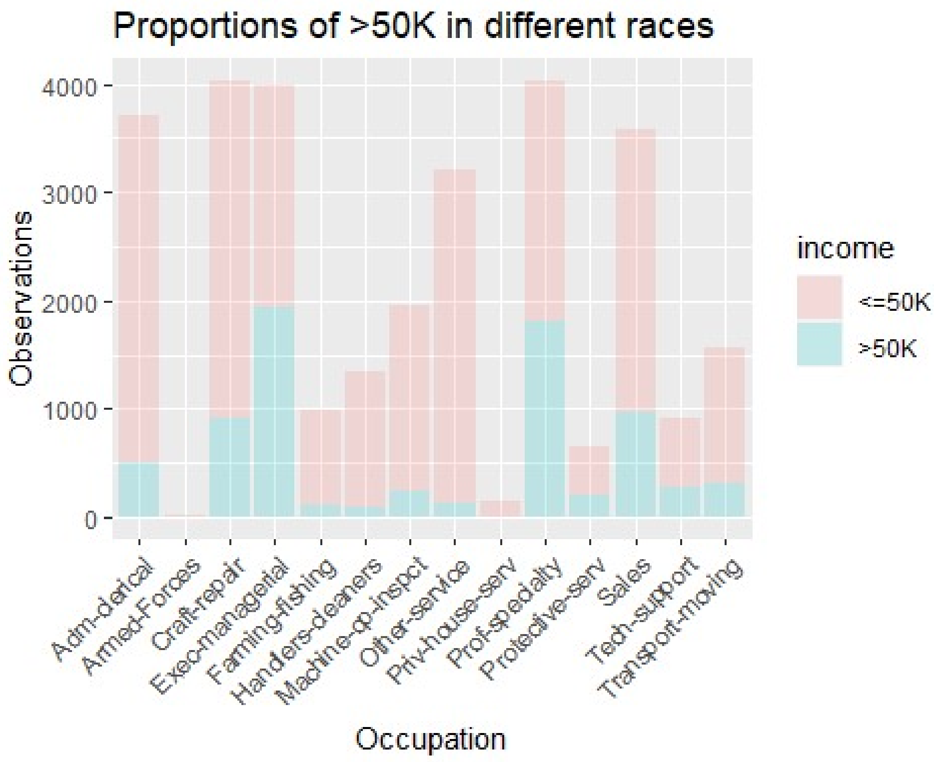
White has the highest proportion of observations who earn more than 50K per year.

Sex



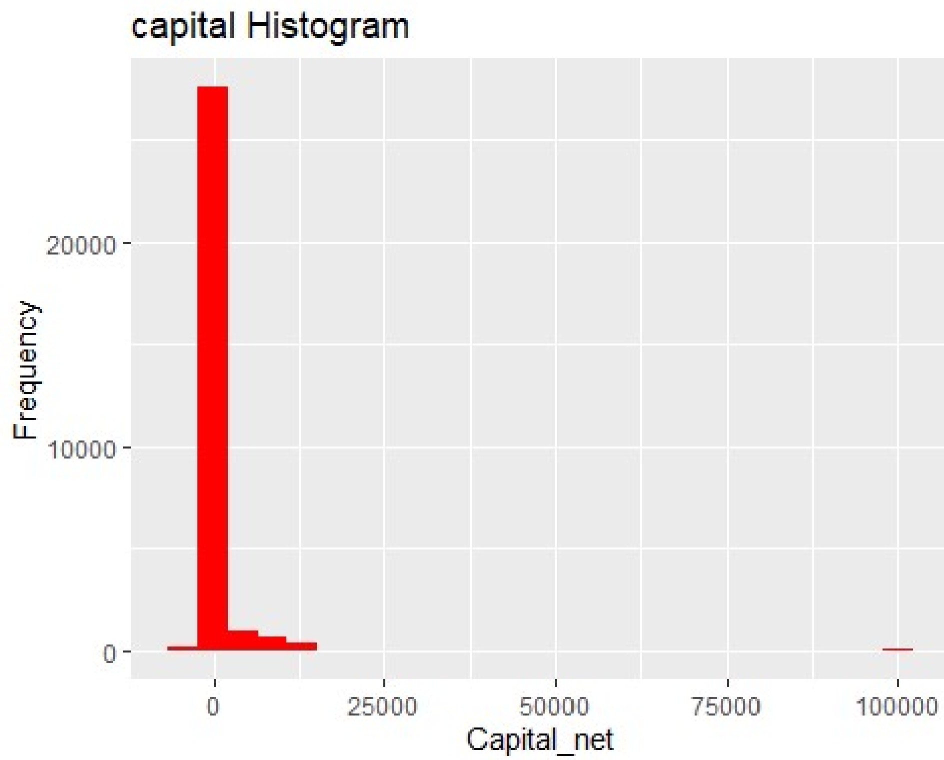
According to the graph, male employees have higher proportion of observations who earn more than 50K per year.

Occupation

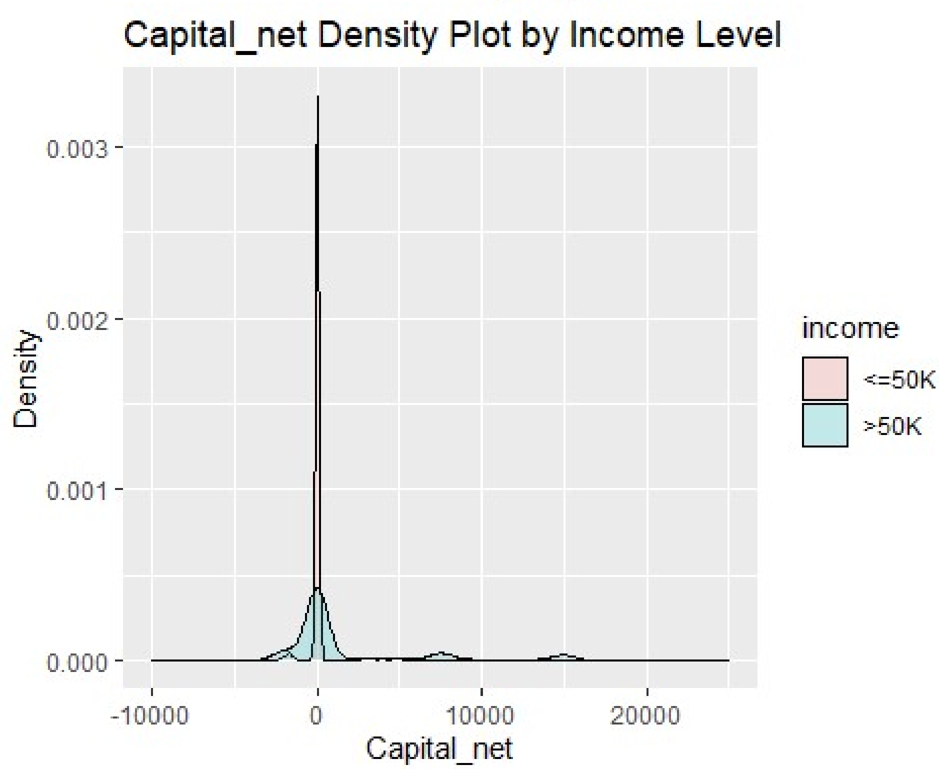


According to the graph, Exec-managerial and Prof-specialty have highest proportion of observations who earn more than 50K per year. Blue collar such as Handers-cleaners and sales make less salaries.

Capital_net



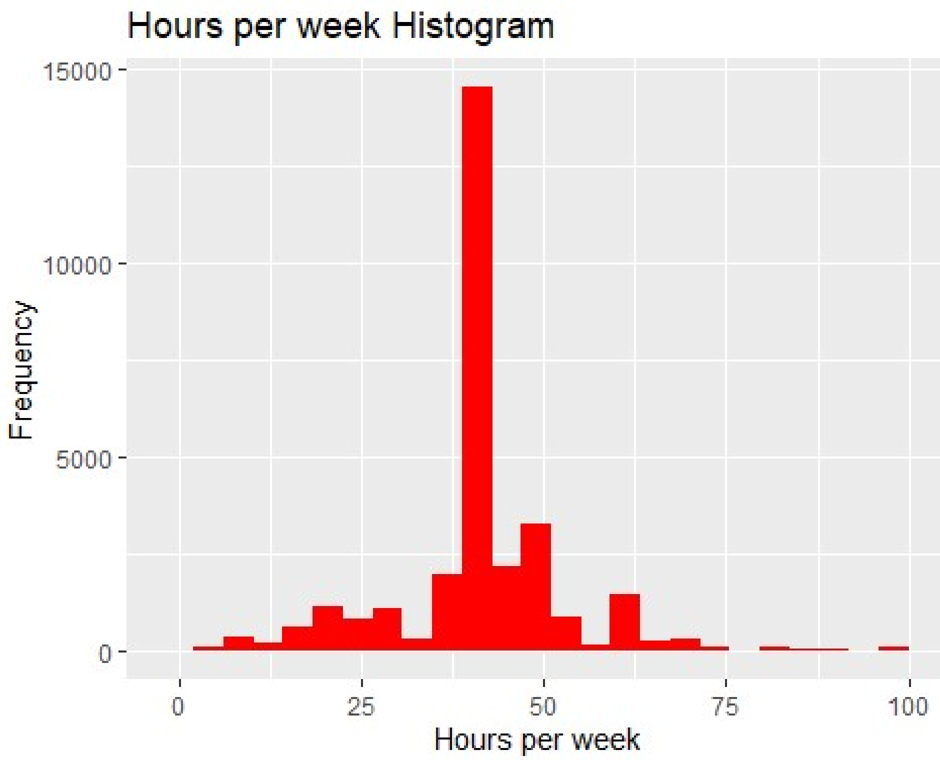
capital Histogram
Frequency
20000
10000
0
0
25000
50000
75000
100000
Capital_net

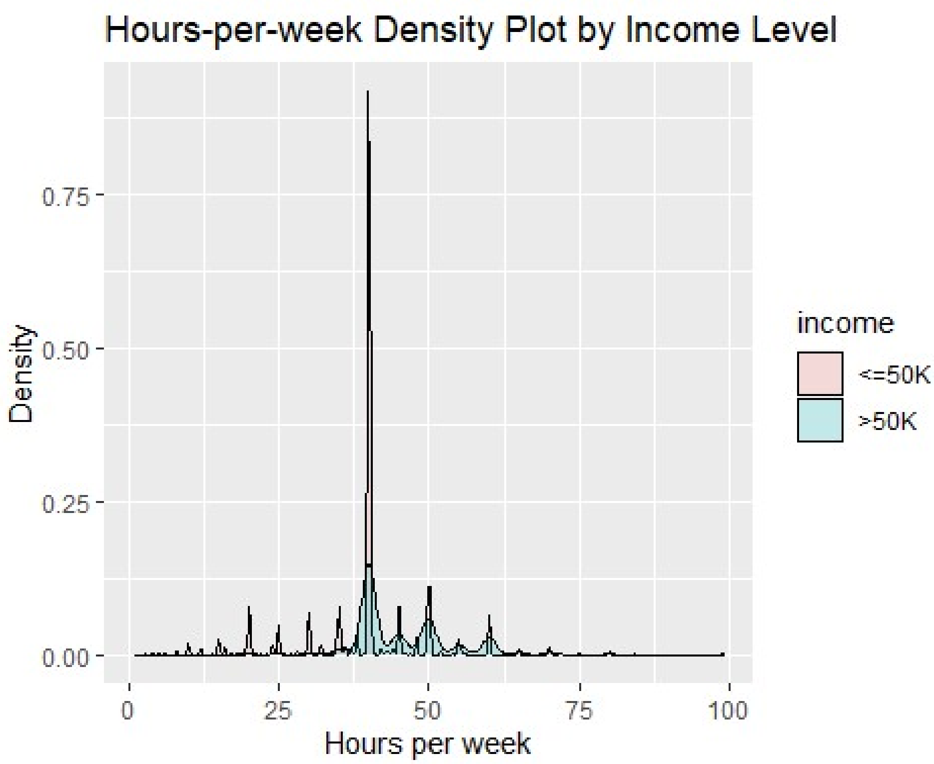


Capital_net Density Plot by Income Level
Density
0.003
0.002
0.001
0.000
-10000
0
10000
20000
Capital_net
income
<=50K
>50K

Capital_net of most observations sit around zero regardless of income levels. The graphs show that the capital_net is not very useful for classification. So it is taken out of the model dataset.

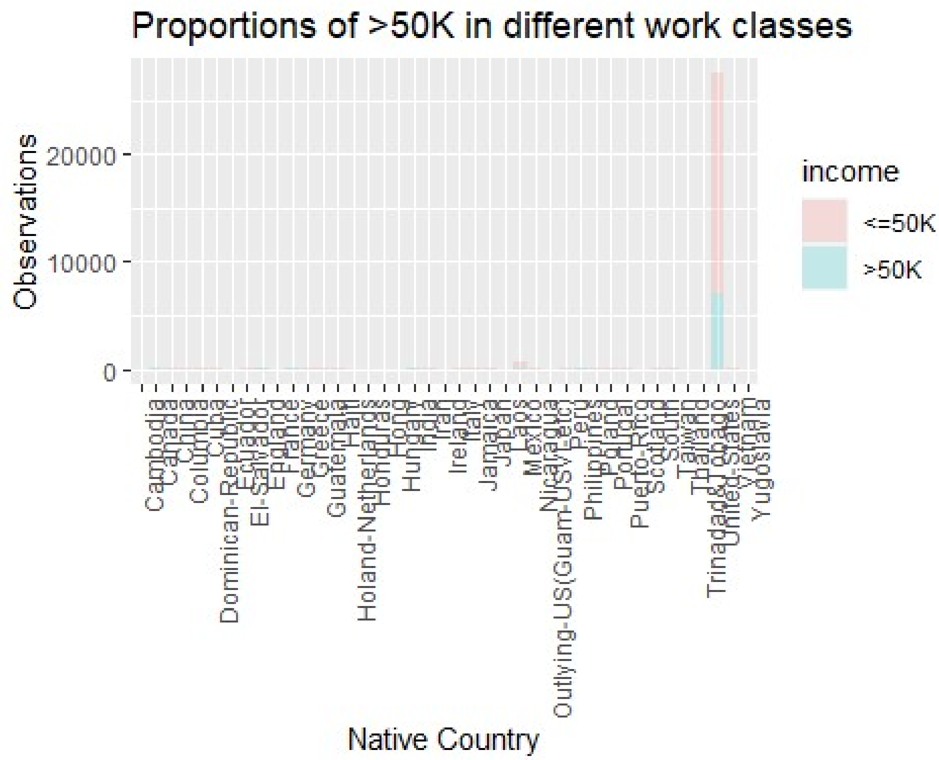
Hours per Week

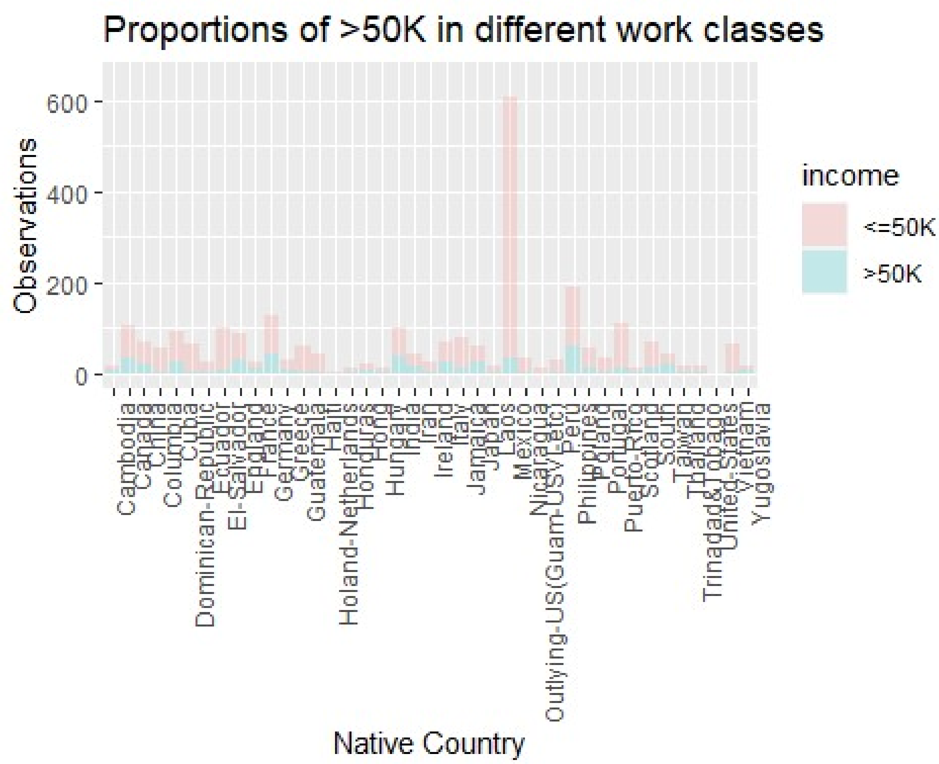




Greater hours per week may result in a higher possibility of high income.

Native Country





Most people are from US according to the 1st graph. People from Mexico are the second largest population in the dataset. The 2nd graph is zoomed in graph of the 1st one indicating that most of people from Mexico are making salaries less than 50K per year.

Final Dataset Used for Models

```
##   age workclass education.num marital.status        occupation  race
sex
## 1  82   Private             9        Widowed   Exec-managerial White
Female
## 2  54   Private             4       Divorced Machine-op-inspct White
Female
## 3  41   Private            10      Separated    Prof-specialty White
Female
## 4  34   Private             9       Divorced     Other-service White
Female
## 5  38   Private             6      Separated      Adm-clerical White
Male
## 6  74 State-gov            16  Never-married    Prof-specialty White
Female
##   hours.per.week native.country income
## 1             18  United-States  <=50K
## 2             40  United-States  <=50K
## 3             40  United-States  <=50K
## 4             45  United-States  <=50K
## 5             40  United-States  <=50K
## 6             20  United-States   >50K
```

The final dataset is split into the trainset training dataset and the testset testing dataset using function createDataPartition. Testset is 10% of final dataset.

Models and Results

Decision Tree

The accuracy of decision tree is below:

```
## Accuracy
## 0.825539
```

Random Forest

The accuracy of random forest is below:

```
##
## Call:
##  randomForest(formula = income ~ ., data = trainset)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
```

```
##
##         OOB estimate of  error rate: 17.13%
## Confusion matrix:
##       <=50K >50K class.error
## <=50K 18433 1936  0.09504639
## >50K   2711 4044  0.40133235

##  Accuracy
## 0.8334992
```

Random forest is a better model for the dataset considering quite a few features used for prediction.
Below codes are not run in my computer due to very slow response. A better accuracy can be achieved by tuning parameters.

grid <- data.frame(mtry = c(1, 5, 10, 25, 50, 100))
control <- trainControl(method="cv", number = 5)
train_rf <- train(income ~ ., data = income_data,
method = "rf",
ntree = 150,
trControl = control,
tuneGrid = grid,
nSamp = 5000)
ggplot(train_rf)
train_rf$bestTune

Conclusions

Random forests are a very strong machine learning approach for categorical prediction with many features. The initial random forest model can achieve an accuracy of 0.83. Further parameter tuning work can be done in provided codes to improve accuracy.But computation time of random forest is very long. Thus some future work need to be done to optimize codes for tuning parameters of random forest and thus reduce computation time significantly.