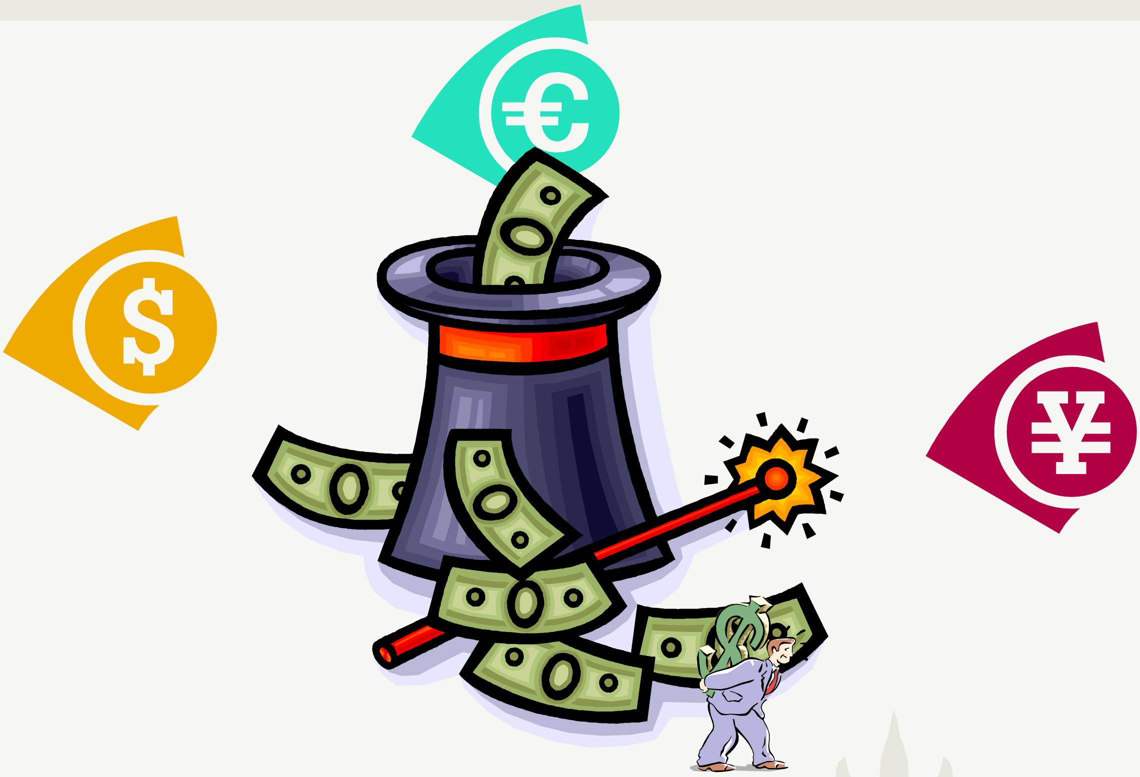
資料探勘與知識發現~期末報告



Using Adult DB to Predict Yearly Salary

Greater or Less than 50K in 1994

Adviser ： Prof. San-Yih Hwang

Student ： Hung, I-Chun (b924020024)
Lee, Nan-Kuang (b924020031)
Tseng, Shih Hui (b924020036)

資料探勘與
知識發現

Using Adult DB to Predict Yearly Salary Greater or Less than 50K in 1994

Adviser：　Prof. San-Yih Hwang
Student：　Hung, I-Chun (b924020024)
Lee, Nan-Kuang (b924020031)
Tseng, Shih Hui (b924020036)

Content

A.Clarify the motivation and the background

- **Motivation**

Because we are the graduate student, most of us ate going to work. For understanding how to prepare to earn higher salary, we found this data, and try to find the model to learn.

- **Background**

Data Extraction was done by Barry Becker from the 1994 Census database (*http://www.census.gov/*). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))

B.Step 1：Translate business problem into DM problem

- **Identify the people whose salaries are greater or less than 50K.**

In the step one, we do the translation to make original problem turn into a concrete goal. So, we use "classification" data mining techniques to do the direct data mining. This process of building a classifier starts with examples of records that have already correctly classified. And in the result, we expect our model can reach higher correct prediction rate (lower error rate).

C.Step 2：Select appropriate data

- **Indication of what attributes were being predicted Salary greater or less than 50K.**

Data Extraction was done by Barry Becker from the 1994 Census database (*http://www.census.gov/*). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). At the same time, we set "greater or less than 50K" about salary as the class label.

Attributes	
Age	continuous
Work class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt	continuous

Attributes	
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
Class Label	
Class	>50K, <=50K

D.STEP 3：Get to know the data

◆ **This data set is reasonable and able to use.**

For example, We draw a histogram (Figure 1) contains number of people and age as follows. It shows up many people are 18-50 years old and this range represents a normal working age. According to its distribution, we think it’s reasonable. The followings are distributions or relations figures based on each attribute and people amount.

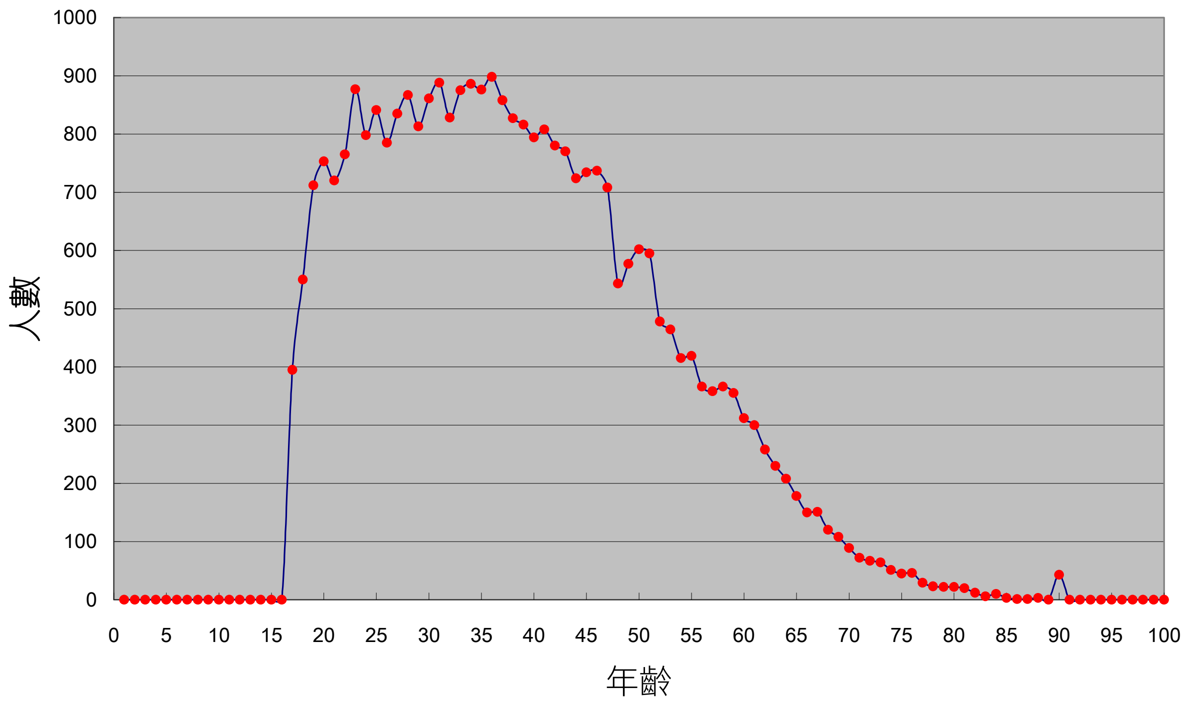


Figure 1: It shows the age distribution. We found most people fall on the 16-60 interval. In the reality, this situation is reasonable.

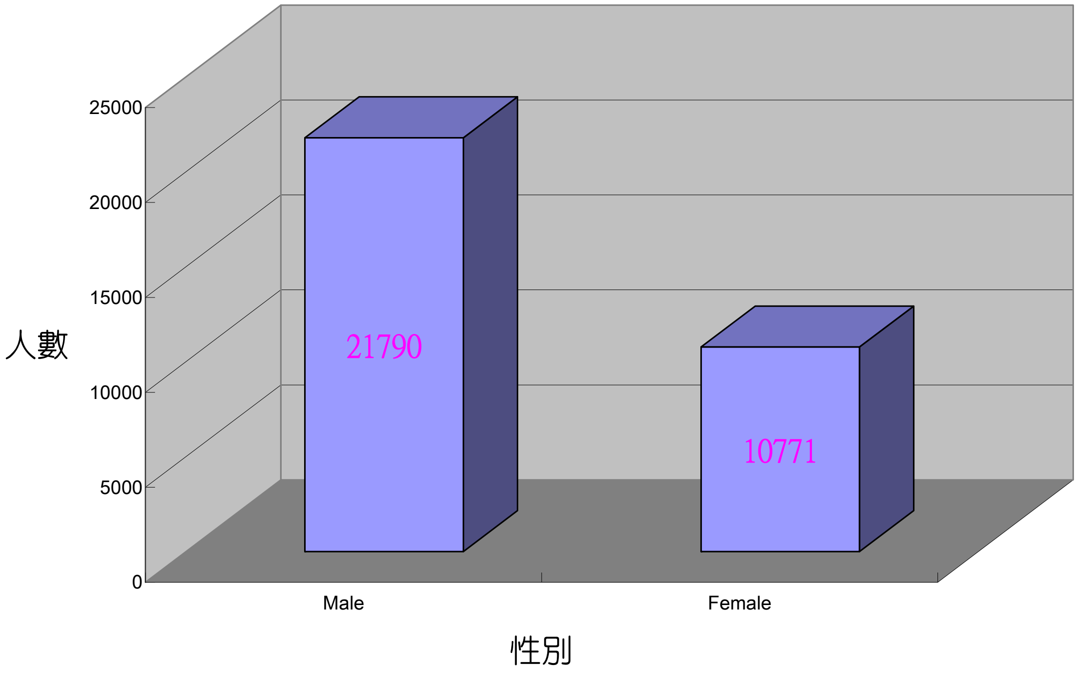


Figure 2: It shows the Male and Female ratio is about 2:1. It's reasonable. In the reality, male of working percentage is higher than female.

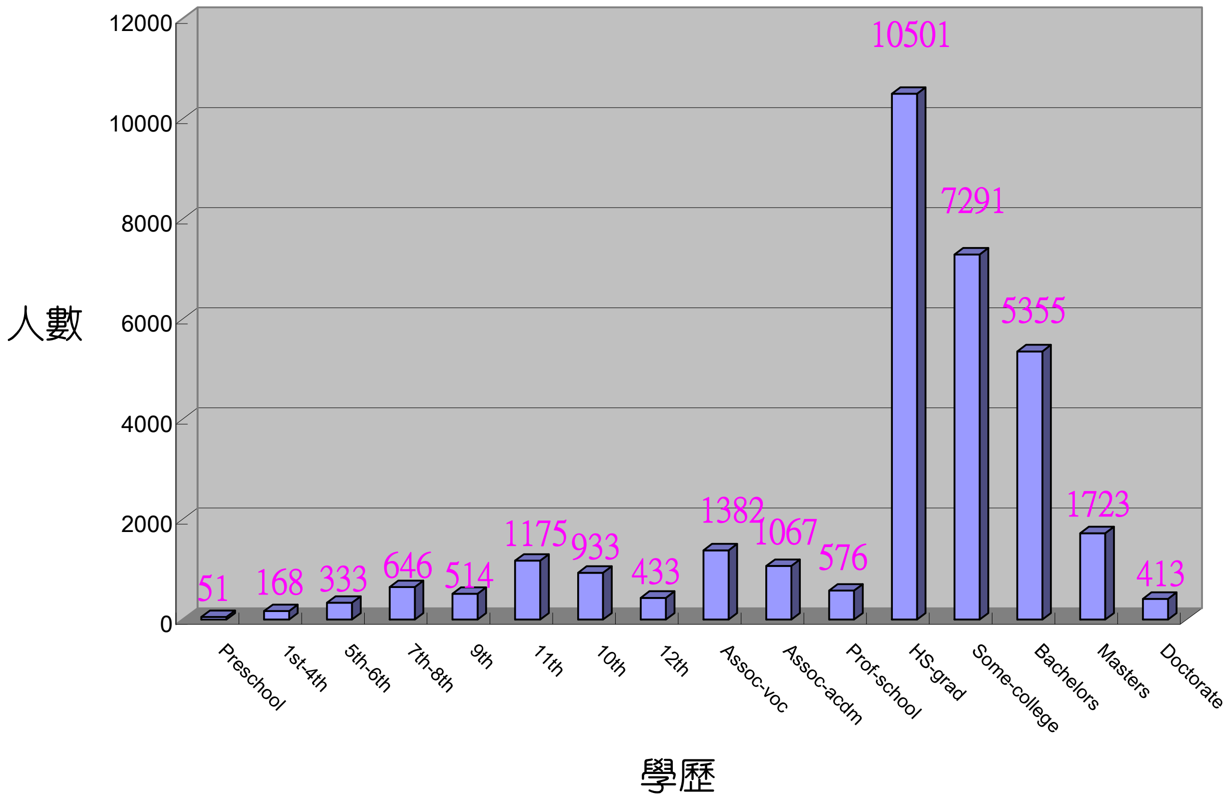


Figure 3: In the data we got, most of people have HS-grad or better than it.

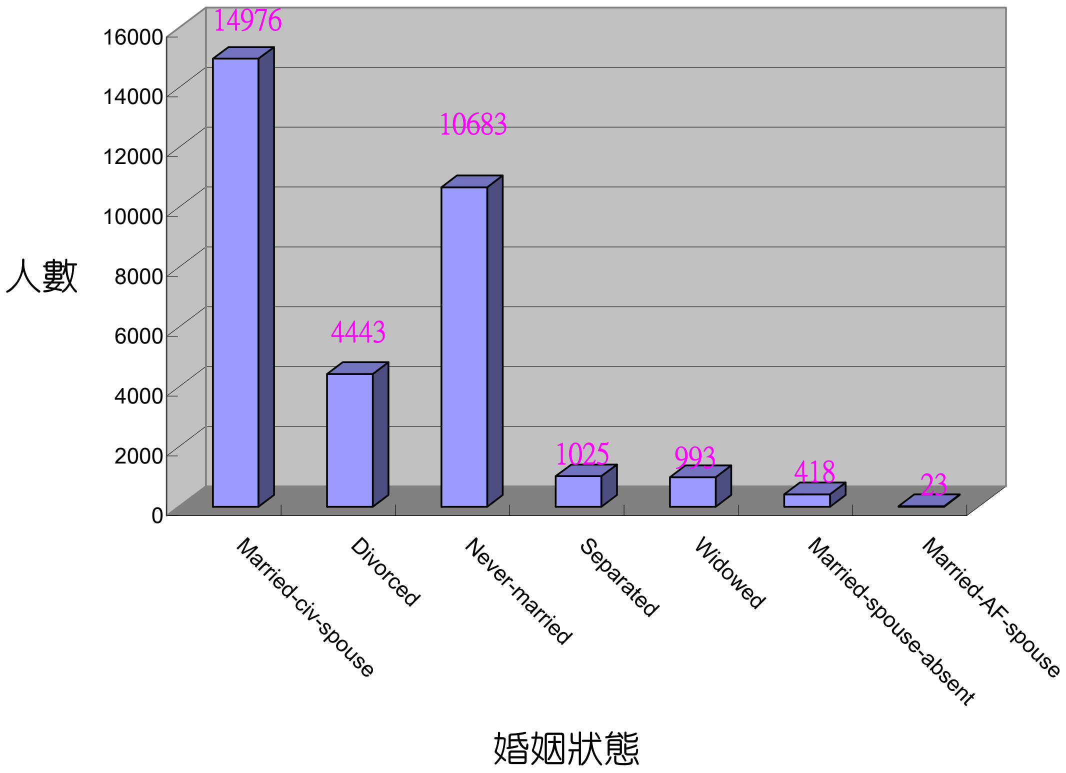


Figure 4: The most part is the people who married. The second part is the people who are single.

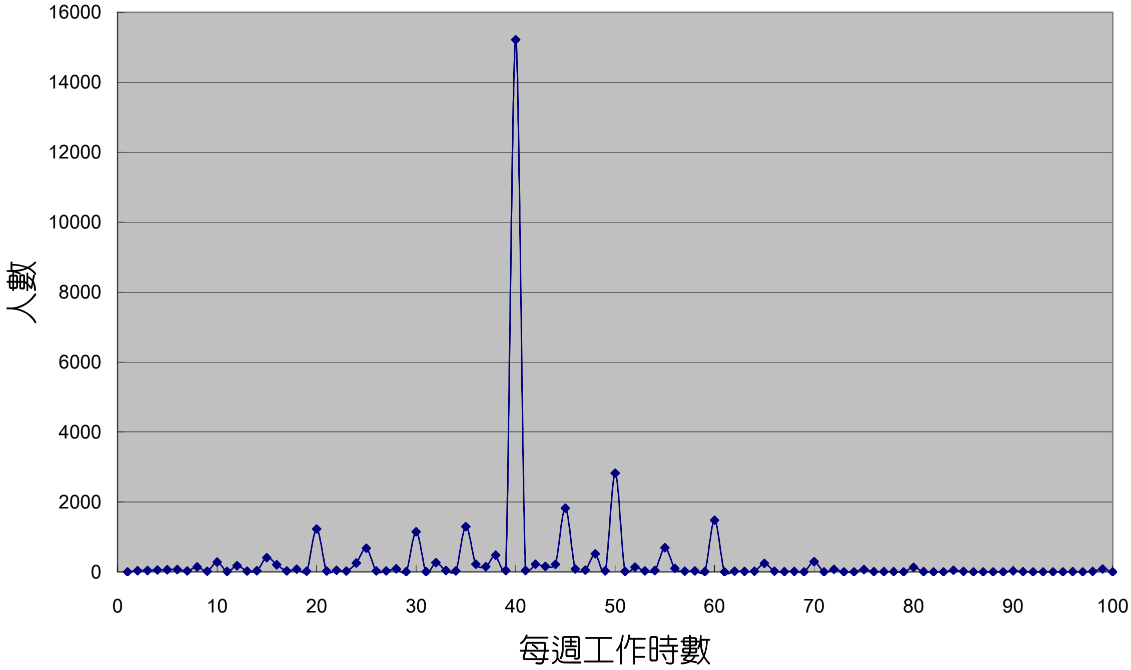


Figure 5: Most part of this data shows a common situation 40hrs per week. It accords with the reality.

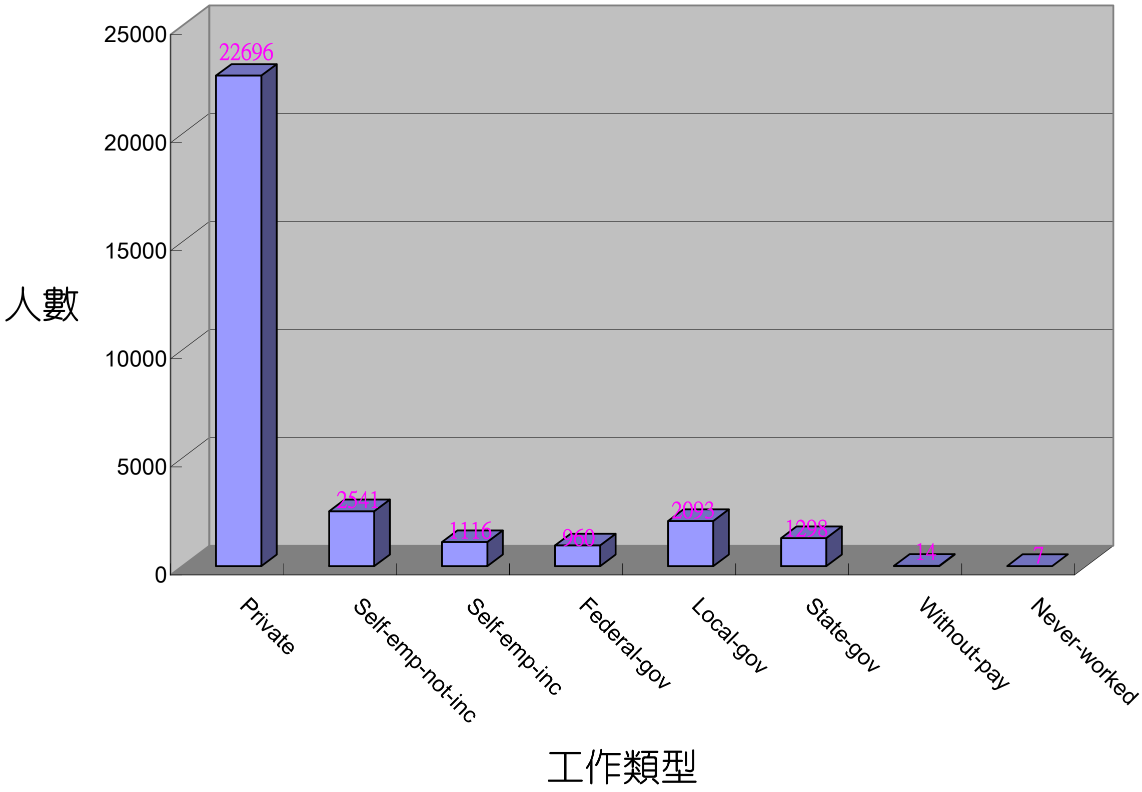


Figure 6: The type of "Private" work has high percentage.

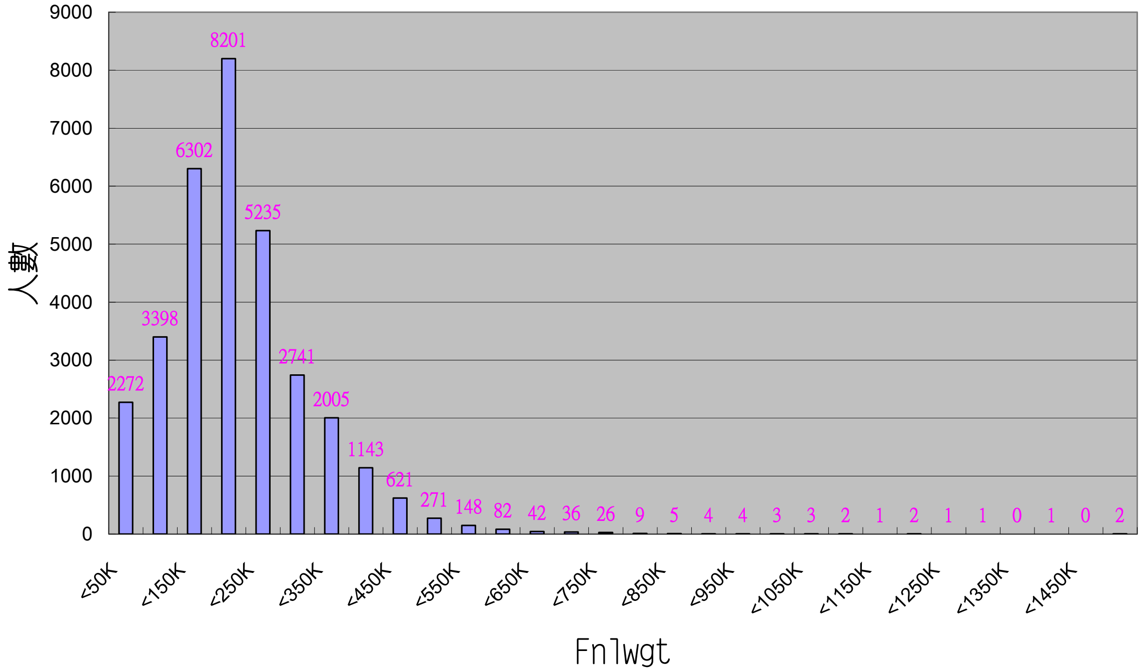


Figure 7: Fnlwgt and people amount. Almost all people are fall on 0~450K interval.

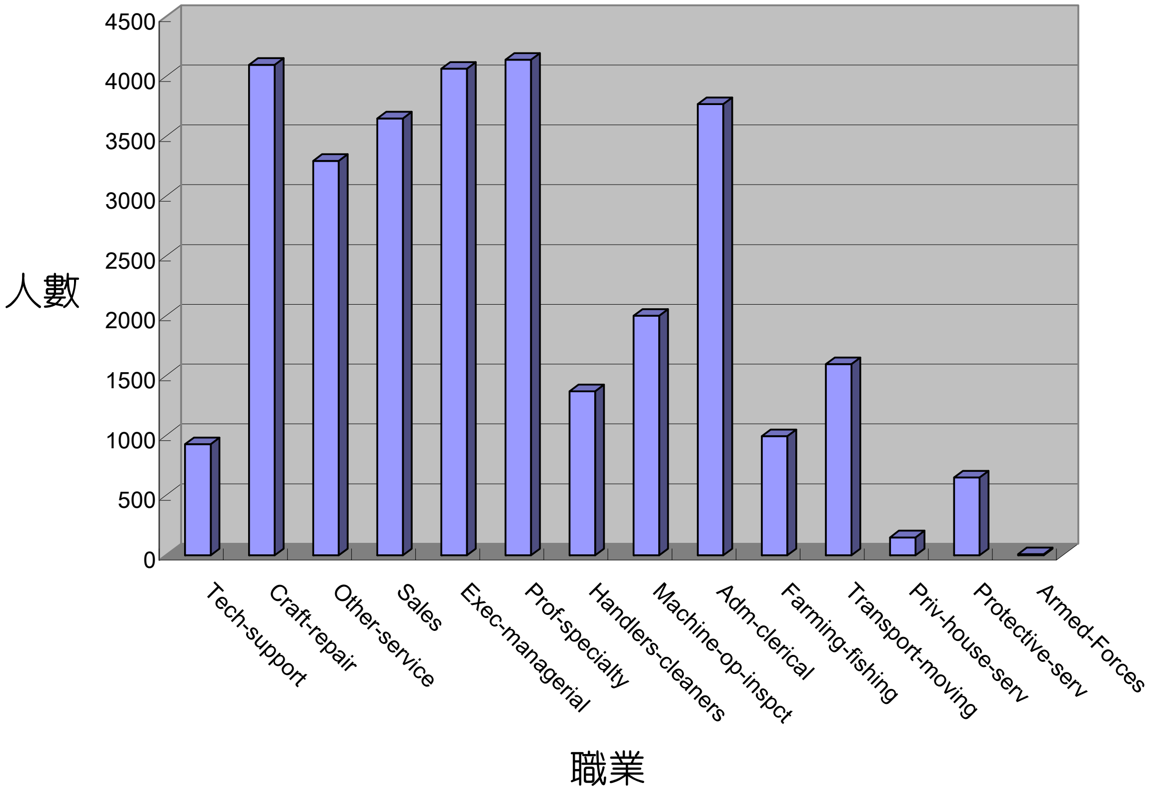


Figure 8: Occupation and people amount. The data we got contains 14 occupations.

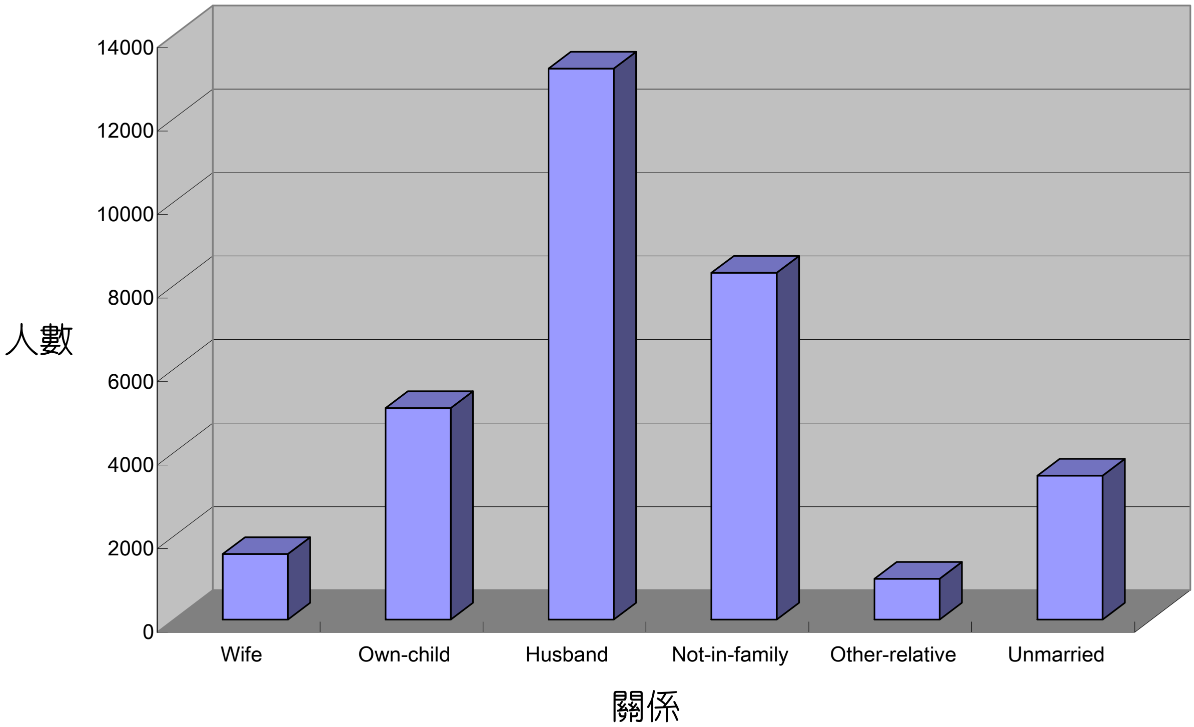


Figure 9: Family character and people amount. As the figure2 shows, male’s working percentage is higher. So, in this figure “Husband” label has the most amounts.

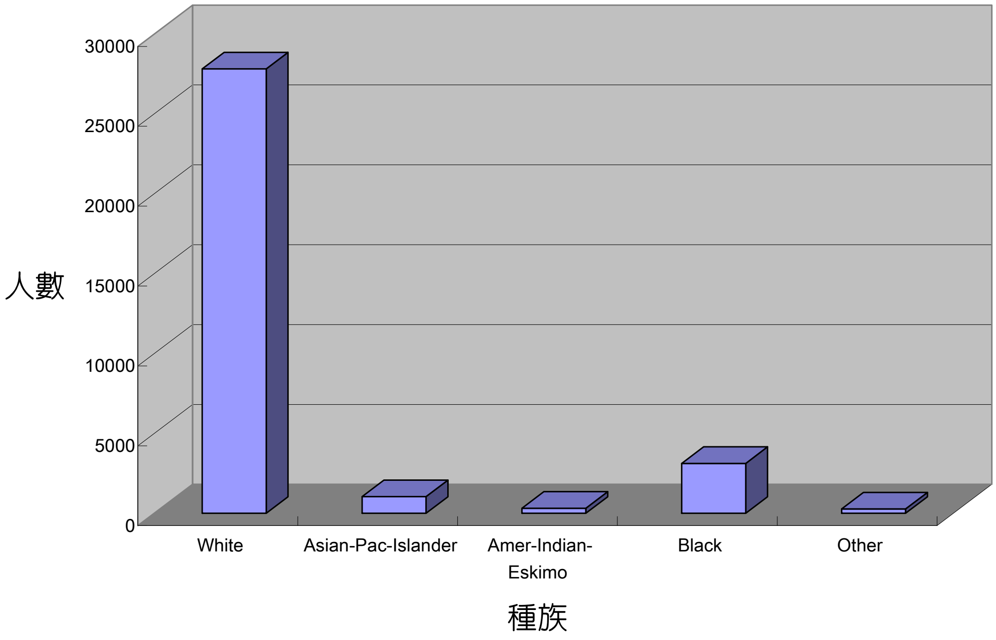


Figure 10: In this data we got, the white people are a large part.

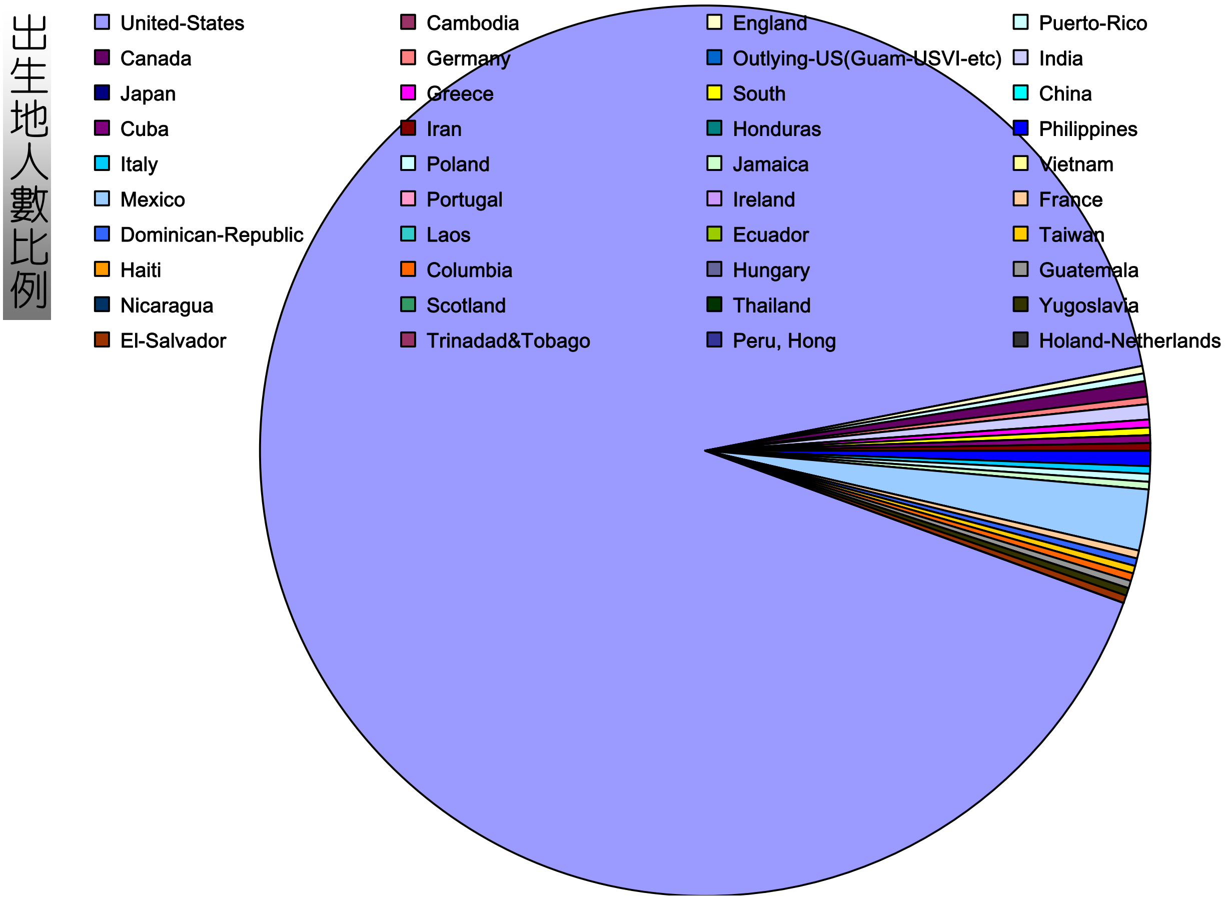


Figure 11: As the Figure10 shows, the most people's native country is US.

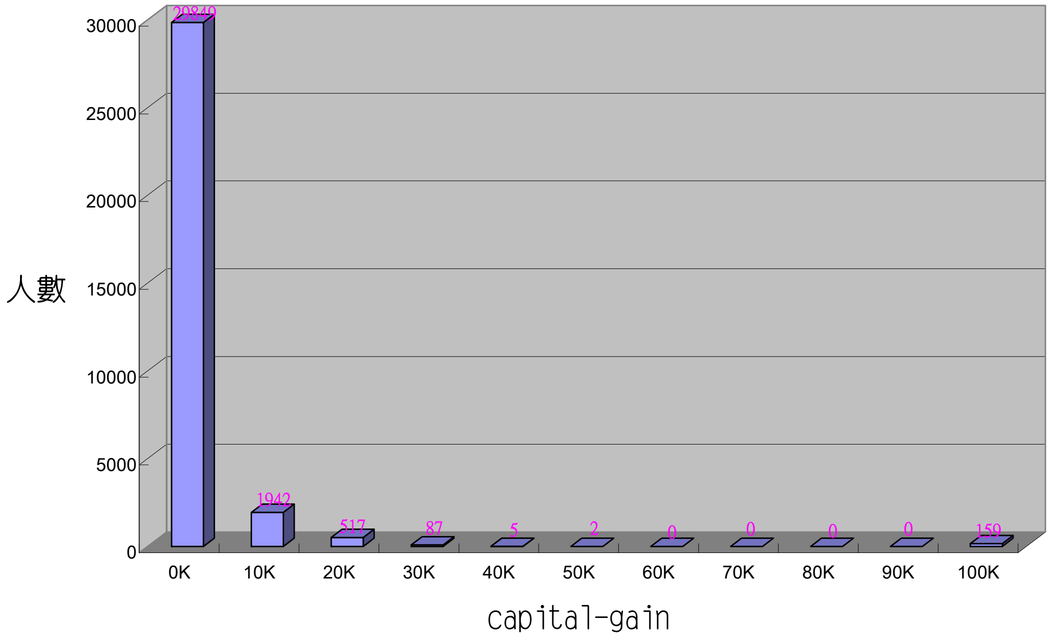


Figure 12: Most people's capital-gain is 0. It represent most people don't have extra profit by selling assets.

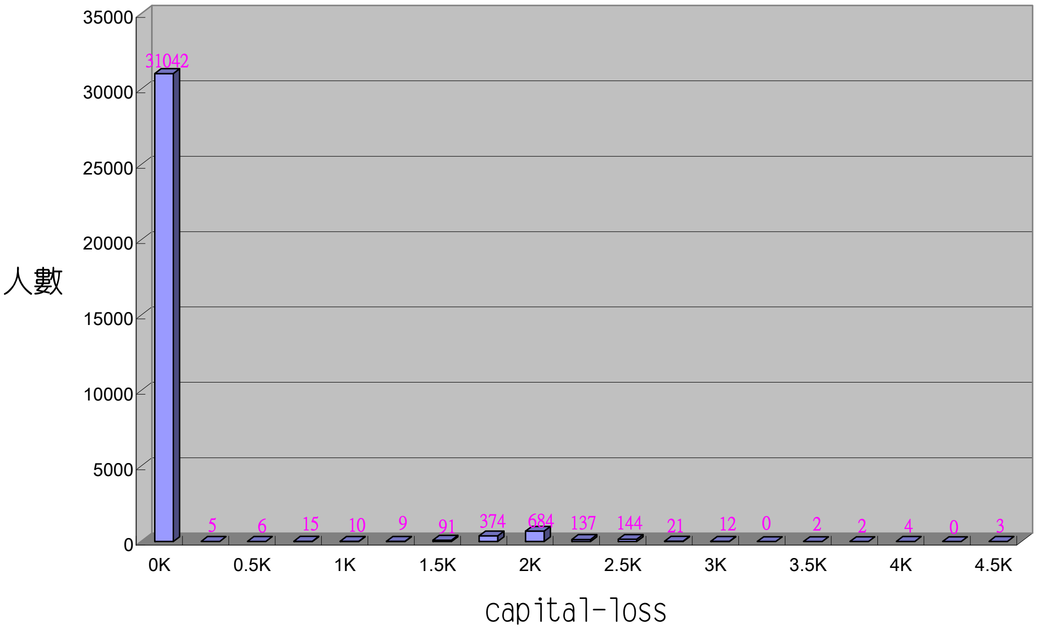


Figure 13: Most people's capital-loss is 0. It represent most people don't have extra loss by selling assets.

E.Step 4：Create a Model Set

1. The data contains many personal characteristics, and most studies think there are possible relationships between personal salary and these characteristics.
2. The quantity of data is large, so we suppose the data is balanced.（*The quality is 32500）
3. Since this data set is from 1994, we use the data from the same year to be the test data. We will try to find the other year data to be the future research.

F.Step 5：Fix Problems with the Data

◆ **Categorical variables with too many values**

- The education variable is transferred to education-num variable. Since the 12-year compulsory education in American, we suppose that people with unfinished compulsory education are most earning lower salary（No conspicuous difference.） We grouped the classes from Preschool to 12th.

- And since the same meaning of education and education-num, we decide delete the variable education-num to avoid double use.

12th
11th
10th
9th → **Unfinished**
7th-8th
5th-6th
1st-4th
Preschool

◆ **Numeric values with skewed distribution and outliers**

We try to find any outliers in these attributes. We use the mean and standard deviation to calculate every numeric attributes to find outliers. Then we find the age exist the outliers.

- **Age** (mean: 38.4, standard deviation:13.37)

  38.4+3*13.37 = 78.51

  38.4-3*13.37 = -1.71

  We remove the age more than **80** years old.

◆ **Missing value**

We replace the missing value with the average value of every attributes.

- Categorical Variables：the attribute that appear most times
- Numeric Variables： the average value

◆ Since the quantity of the data is large, the problems won’t cause many trouble.

G.Step 6：Transform Data to Bring Information to the Surface

◆ **Capture Trends**

Since the data is just the collection of one year, and our research time series is year. We didn’t do this step.

◆ **Create Rations and Other combination of Variables**

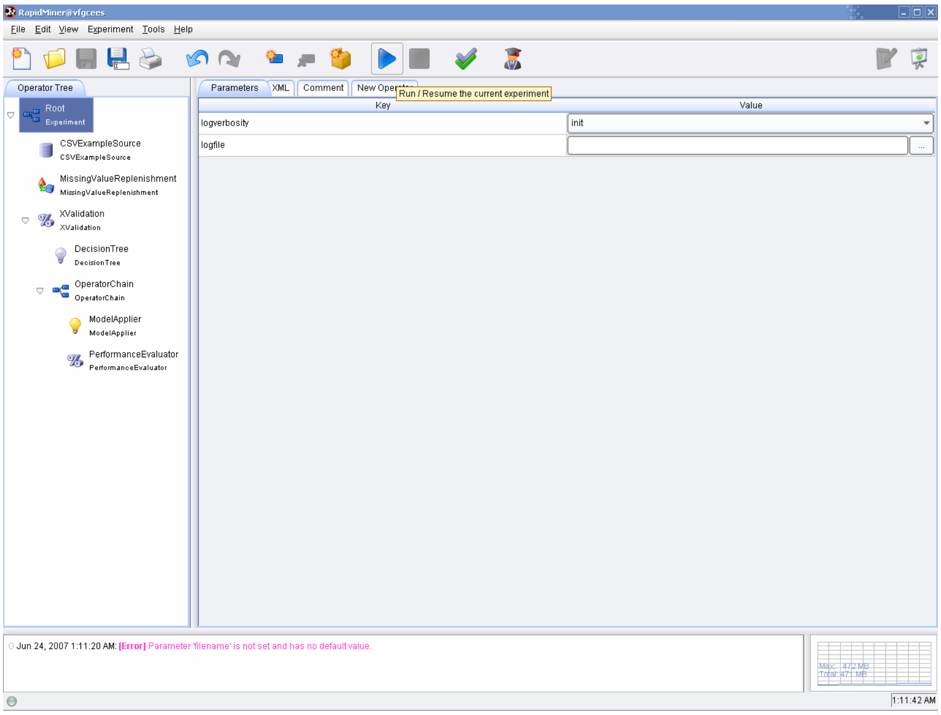
We didn’t find relative rations about this mining research, so we didn’t do the step too.

◆ **Convert Counts to Proportions**

The data don’t contain counts, so we ignore this step.

H.Step 7：Build Models

We used RapidMiner to help us to mine the data, and the version is 4.0 beta. Figure 14 is what the software look like.



Root
Experiment
CSVExampleSource
CSVExampleSource
MissingValueReplenishment
MissingValueReplenishment
XValidation
XValidation
DecisionTree
DecisionTree
OperatorChain
OperatorChain
ModelApplier
ModelApplier
PerformanceEvaluator
PerformanceEvaluator

Key	Value
logverbosity	init
logfile	

```
Jun 24, 2007 1:11:20 AM: [Error] Parameter 'filename' is not set and has no default value.
```

1:11:42 AM

Figure 14

We used the CSV file which we edited as the source. After the software filtrated out missing values, we had the decision tree like Figure 15.

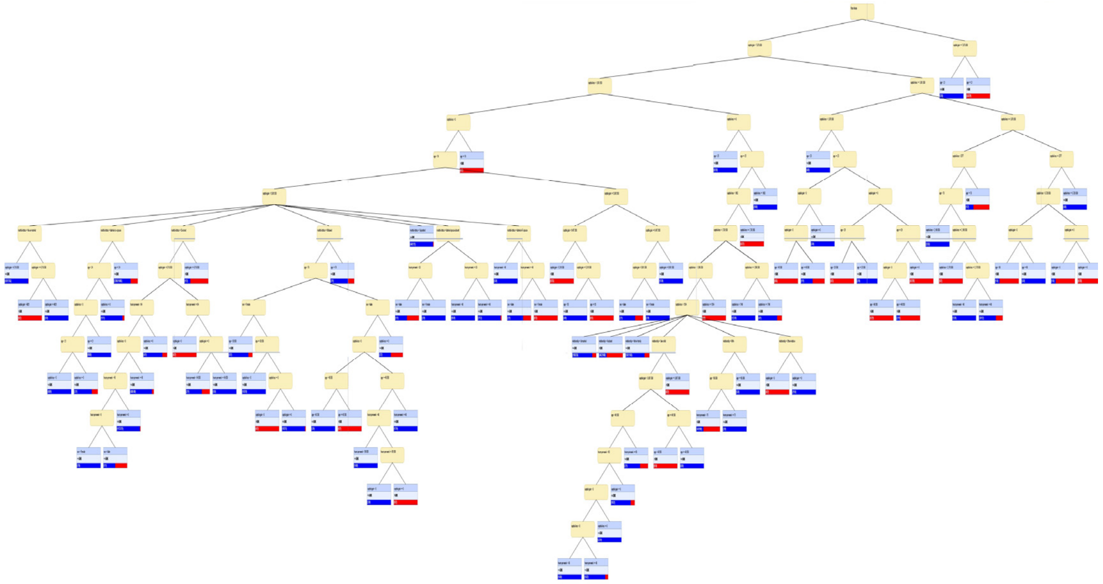


Figure 15

As picture, this decision tree is very unbalance and left-leaning. Figure 16 is part of the decision tree which near root part.

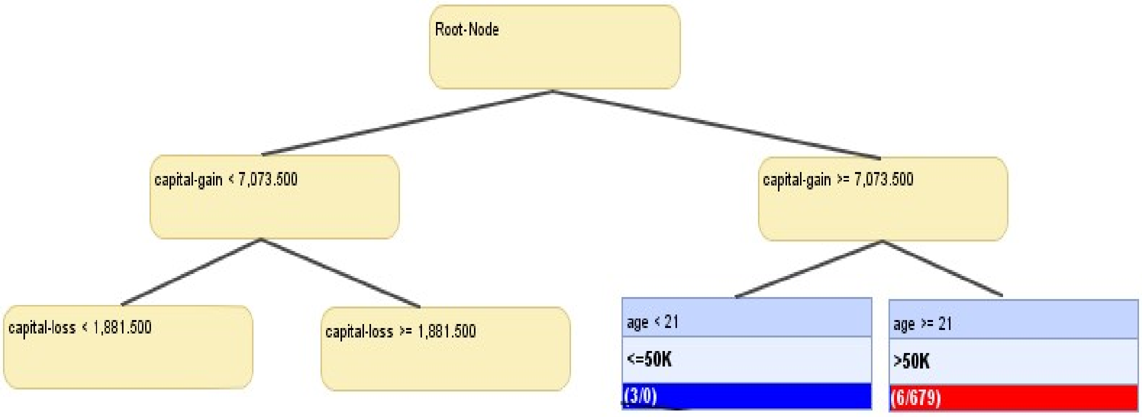


Figure 16

We can see that the most conspicuous indicator is **capital-gain**, while **capital-gain** > 7073.5 and **age** > 21, the record will be labeled by "> 50K". And if **age** < 21, the record will be labeled by "<= 50K". On the other side, to label whether "> 50K" or "<= 50K" is not that easy though.

※ (3/0) means there are 3 records were labeled by <= 50K and it is true. And (6/579) means there are 579 records were labeled by > 50K and it is true but 6 are false.

I.Step 8：Access Models

We used part of our source data as the test data. And the result is follow.

	True <= 50K	**True > 50K**	**class precision**
Pred. <= 50K	16991	861	95.18%
Pred. > 50K	7624	6964	47.74%
class recall	69.03%	89.00%	
SUM	24615	7825	32440

class precision = 16991 / (16991+861) = 95.18%
= 6964 / (7624+6964) = 47.74%

class recall = 16991 / (16991+7624) = 69.03%
= 6964 / (861+6964) = 89.00%

Accuracy = (16991+6964) / (16991+6964 + 861+7624) = 73.84%

Lift = 191.97%

According the table, the accuracy of our decision tree is 73.84%. And it is good at prediction the record which is true <= 50K. Also, the Lift value is 191.97%, it means the more the better!

J.Conclusion

1. Comparing the mining result and our pre-assumption.
   - ◆ Before mining, we assume the most conspicuous indicator is **education**, since the direct thinking.
   - ◆ But the mining result is **capital-gain**, so we think that people make more money who can use more money to invest.
2. In the mining process, we have learned that it's better to understand the meaning of all attributes, or the process will block very often.
3. Fix data problem is a very important step. After fixing data problem, the mining result will be more accuracy.