[20201181] [Jihwan Oh]

IE30301-Final Report: Classification of individual's annual income

1. Introduction

We have 33272 individual's annual income data with other 14 variables on 1994 US. There are total 15 features.

Feature information

1. **workclass**: employment status of an individual.

   ▶ {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}

2. **education**: education achieved by an individual.

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

3. **marital-status**: marital status of an individual.

   ▶ {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}

4. **occupation**: occupation of an individual.

   ▶ {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}

5. **relationship**: represents what this individual is relative to others.

   ▶ {Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}

6. **race**: race of individual

   ▶ {White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}

7. **sex**: biological sex of the individual

   ▶ {Female, Male}

8. **country**: origin country of individual

   ▶ {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}

9. **age**: individual age (numeric)
10. **education-num**: Education level (numeric)
11. **fnlwgt**: Weights of individuals assigned by the Census Bureau based on a series of observations (numeric)
12. **capital-gain**: Individual capital gain (numeric)
13. **capital-loss**: Individual capital Loss (numeric)
14. **hours-per-week**: Working hours per week (numeric)
15. **income**: An individual's annual income. **(Target variable)**

▶ {>50K,<=50K}

'Income' is the target variable. So, we should predict the individual's annual income with other variables. Before we analysis this dataset, we can set some hypothesis following each reasons through feature variables and intuition.

Hypothesis1: If individual have higher education-num, then income will be higher.

- Education-num refers to individual's education level. Higher education level means that he/she is smart and is more likely to work as a manager than a worker in society. In general, the income manager is higher than worker.

Hypothesis2: If age is close to 50, the income will be higher.

- Usually, the leader or manager of company is close to 50 years old. Their gain will highest in all age.

Hypothesis3: If working hour per week is higher, the income will be higher.

- There are minimum hourly wage system in our society. Therefore, working time is proportional to gain.

Hypothesis4: If race is close to white, then the income will be higher.

- In 1994 US, there are racism. Therefore, white people can get more gain.

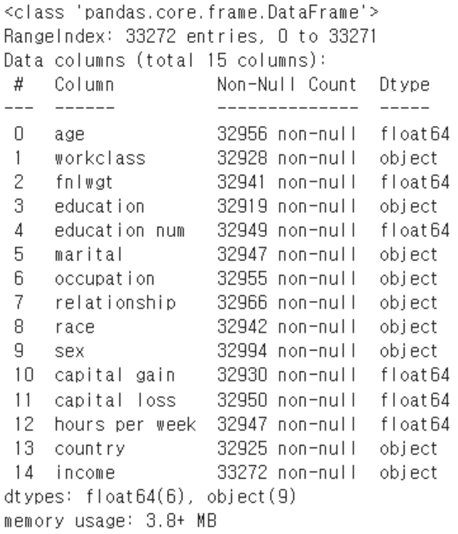
Hypothesis5: Male will have higher income than female.

- In 1994 US, there was patriarchy. Therefore, income of male will higher than female.

We should analysis this data based on these hypothesis. Then, I will check the hypothesis, then predict individual's annual income based on other variables and confirmed hypothesis.

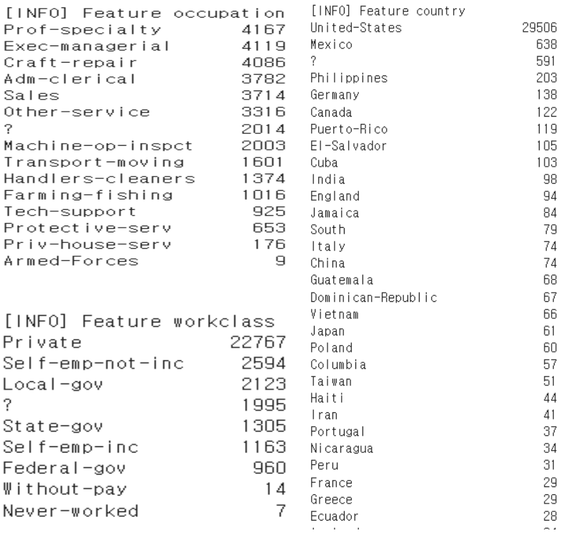
2. Exploratory data analysis

1. Check missing values, undefined values and duplicate data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33272 entries, 0 to 33271
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32956 non-null  float64
 1   workclass       32928 non-null  object
 2   fnlwgt          32941 non-null  float64
 3   education       32919 non-null  object
 4   education num   32949 non-null  float64
 5   marital         32947 non-null  object
 6   occupation      32955 non-null  object
 7   relationship    32966 non-null  object
 8   race            32942 non-null  object
 9   sex             32994 non-null  object
 10  capital gain    32930 non-null  float64
 11  capital loss    32950 non-null  float64
 12  hours per week  32947 non-null  float64
 13  country         32925 non-null  object
 14  income          33272 non-null  object
dtypes: float64(6), object(9)
memory usage: 3.8+ MB
```

Figure 1. Data information



```
[INFO] Feature occupation
Prof-specialty         4167
Exec-managerial        4119
Craft-repair           4086
Adm-clerical           3782
Sales                  3714
Other-service          3316
?                      2014
Machine-op-inspct      2003
Transport-moving       1601
Handlers-cleaners      1374
Farming-fishing        1016
Tech-support            925
Protective-serv         653
Priv-house-serv         176
Armed-Forces              9


[INFO] Feature workclass
Private               22767
Self-emp-not-inc       2594
Local-gov              2123
?                      1995
State-gov              1305
Self-emp-inc           1163
Federal-gov             960
Without-pay              14
Never-worked              7
```

```
[INFO] Feature country
United-States                 29506
Mexico                          638
?                               591
Philippines                     203
Germany                         138
Canada                          122
Puerto-Rico                     119
El-Salvador                     105
Cuba                            103
India                            98
England                          94
Jamaica                          84
South                            79
Italy                            74
China                            74
Guatemala                        68
Dominican-Republic               67
Vietnam                          66
Japan                            61
Poland                           60
Columbia                         57
Taiwan                           51
Haiti                            44
Iran                             41
Portugal                         37
Nicaragua                        34
Peru                             31
France                           29
Greece                           29
Ecuador                          28
```

Figure 2. Undefined data

By Figure 1, there are some missing values in each feature variables, but not many. There are no missing values in target variable(income). Variables type are setting correctly by categorical to object type, numerical to float type. On the other hand, by Figure 2, there are undefined data '?' in some categorical variables. We should convert to NaN. In addition, there are 601 duplicated data. Therefore, we should remove them.

2. Check the imbalance ratio of target variable.

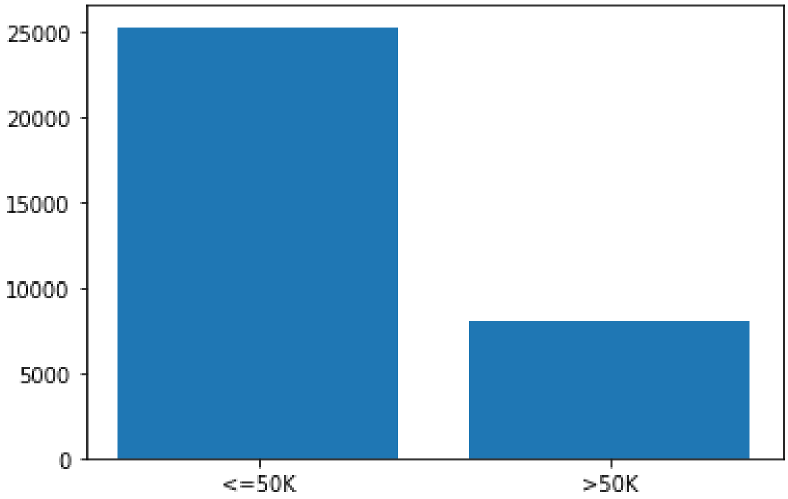


Figure 3. Ratio of target variable

By Figure 2, the number of people who income lower 50K is 3.16 more times than income higher 50K. It is quite different , but still there are enough higher 50K data for training and testing. Therefore, it is not imbalanced.

3. Check the outlier of data.

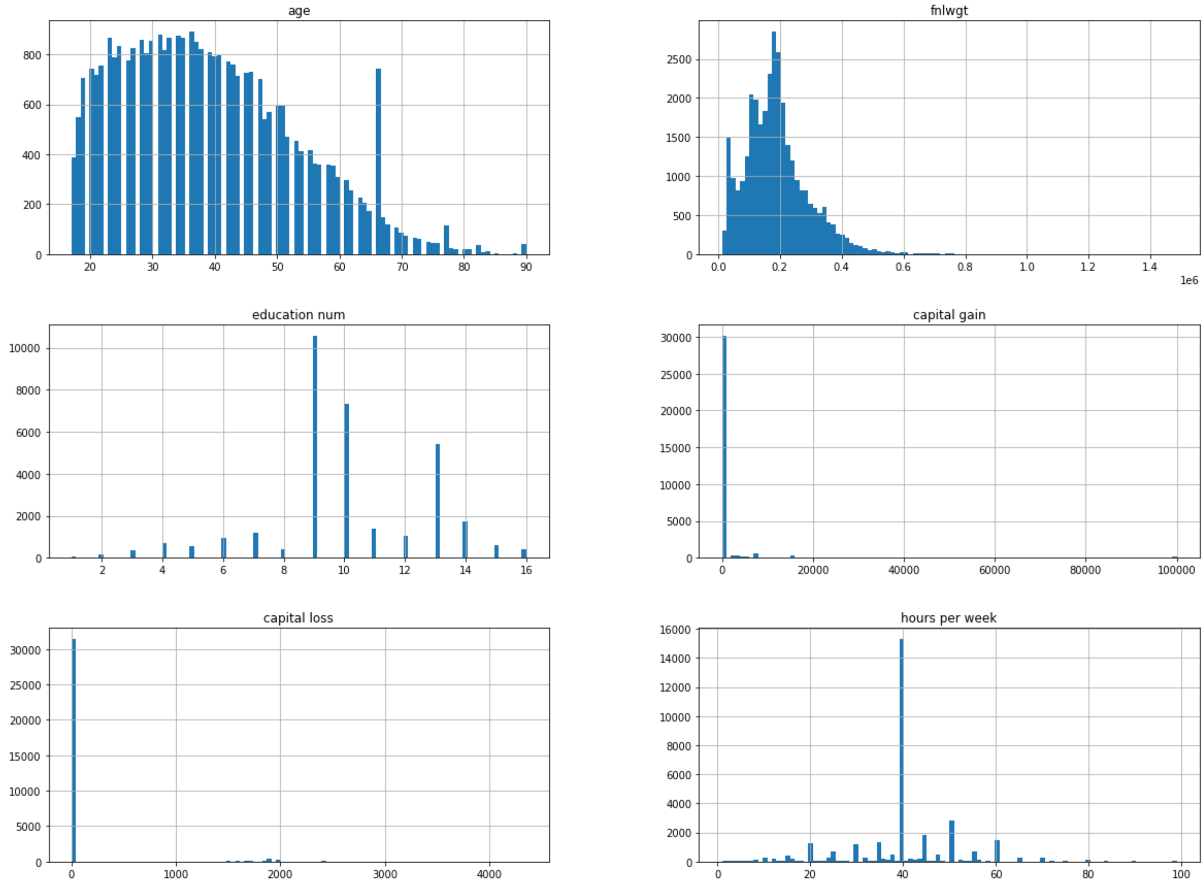


Figure 4. Histogram of numeric variables

In this histogram, it seems to there are outliers in capital gain and capital loss variable. In addition, above 90% of data has ‘0’ value in capital gain and capital loss variable. Therefore, the two variables are very biased. For that reason, we do not consider two variables. In data processing phase, we will remove these two variables.

4. Check the correlation between variables.

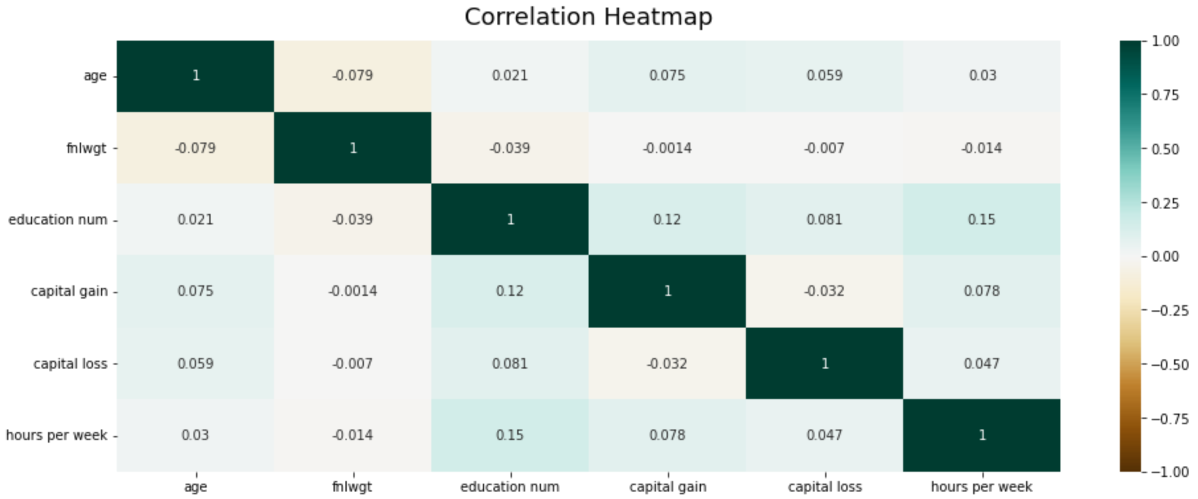


Figure 5. Correlation between numeric variables

There are no high correlated variables. Therefore, we don't worry about the multicollinearity. The variables are almost independent.

5. Split data into train and test data set.

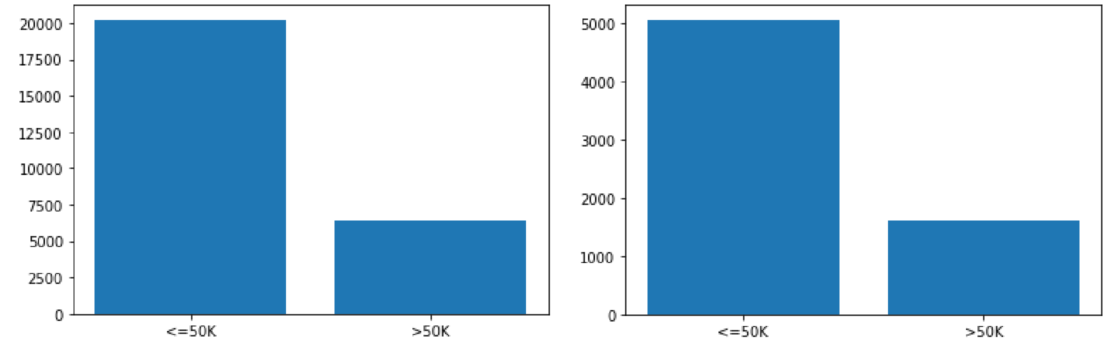


Figure 5. Ratio of train set (left) and test set (right)

Data is split into 80% train data and 20% test data.

6. Test the hypothesis through EDA.

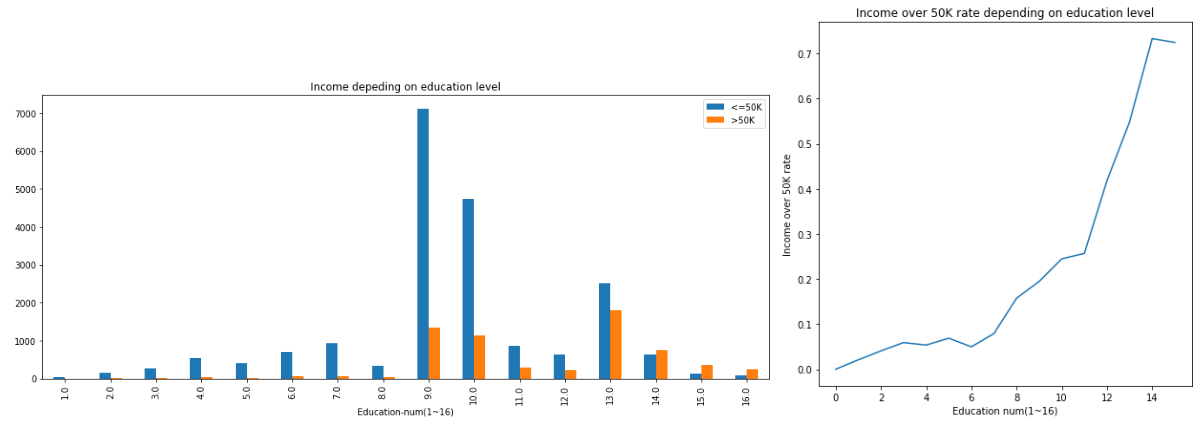


Figure 6. Education level VS Income

By Figure 6, Education level is proportional to individual's annual income clearly. Therefore, '**Hypothesis1**: If individual have higher education-num, then income will be higher.' seems to correct. Additionally, I think that education level is highly correlate with work class and occupation. Because, the education level usually determine their job and work circumstance. Therefore, we can suppose that education, work class and occupation variables represented by education num variable. But work class has large bias and education is exactly same as education num.

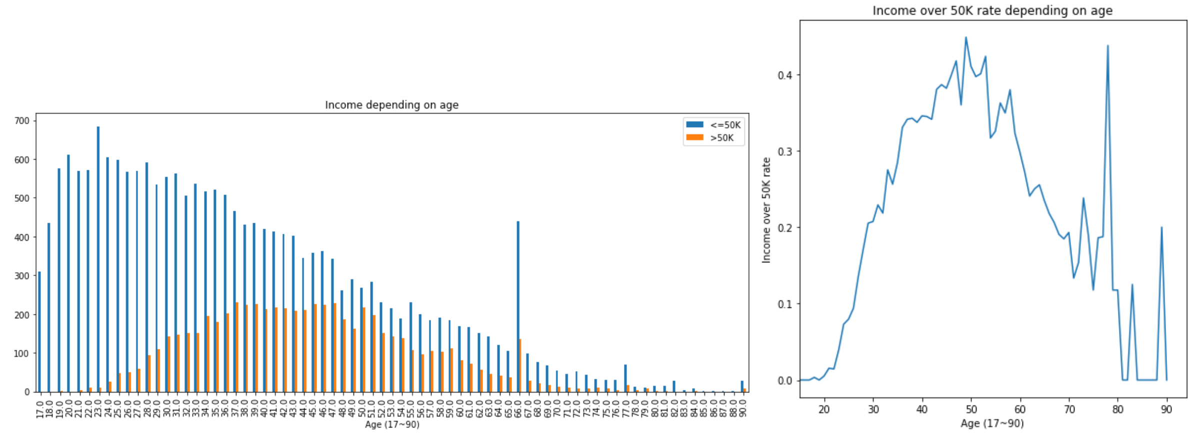


Figure 7. Age VS Income

By figure 7, they gain high income who is age closer to 50. But in age 75-80, they get also high income. According our data, the number of data in age of 75-80 is very small. Therefore, it may not be an accurate analysis in age of 75-80. An accurate analysis can be found with more data. But what's certain is that around the age of 50, they get the age of 50. '**Hypothesis2:** If age is close to 50, the income will be higher.' seems to correct.

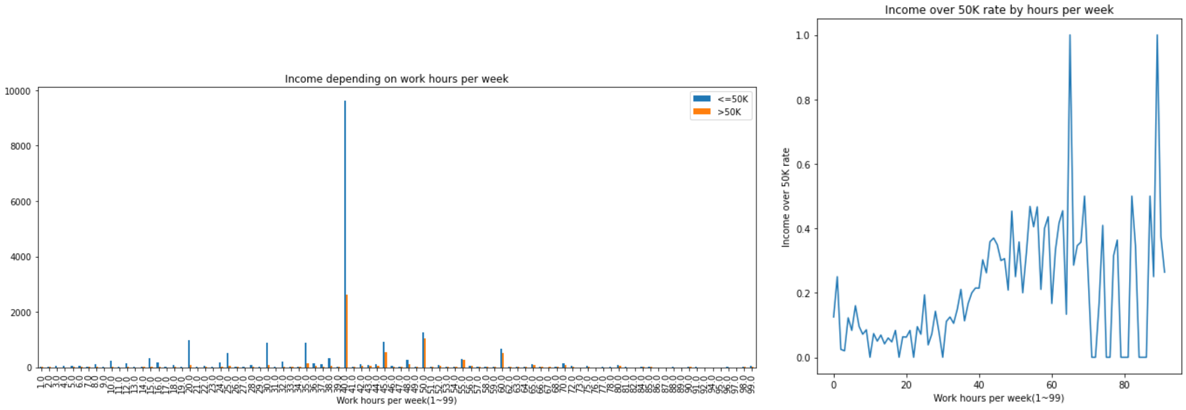


Figure 8. Work hour per week VS Income

By figure 8, the data is quite bias to 40 hours. Therefore, there is a possibility that the analysis results are inaccurate. But it seems reliable except for sections with very small numbers of data as over 60 hours. Consider this circumstance, the income rate in high working time (40-60 hours) is higher than the low working time (10-30 hours). But it does not continue to increase. Therefore, we should revise our original Hypothesis3. **Revised_Hypothesis3:** If working hour per week is higher, the income will be higher during 10 to 60 working hours.

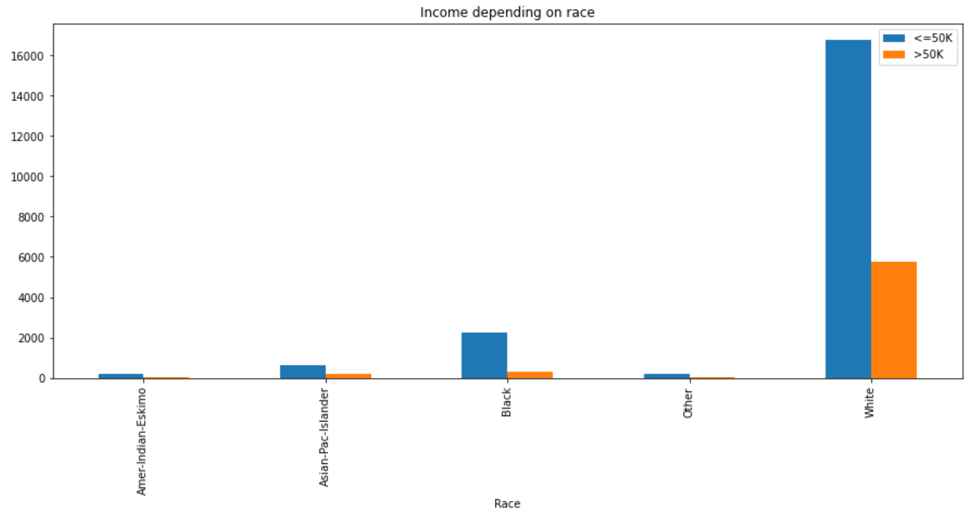


Figure 9. Race VS Income

By figure 9, the income over 50K rate is 25.53%, 12.43%, 26.26%, 10.76% and 9.05% each White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo and Other. Therefore, we should revise our original Hypothesis4. **Revised_Hypothesis4 : Asian-Pac-Islander and White get higher income, Amer-Indian-Eskimo and Other get lower income.**

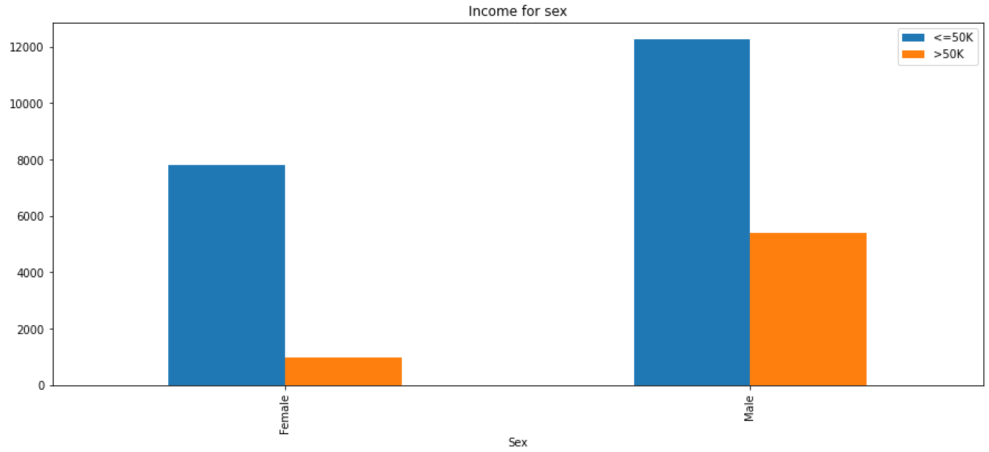
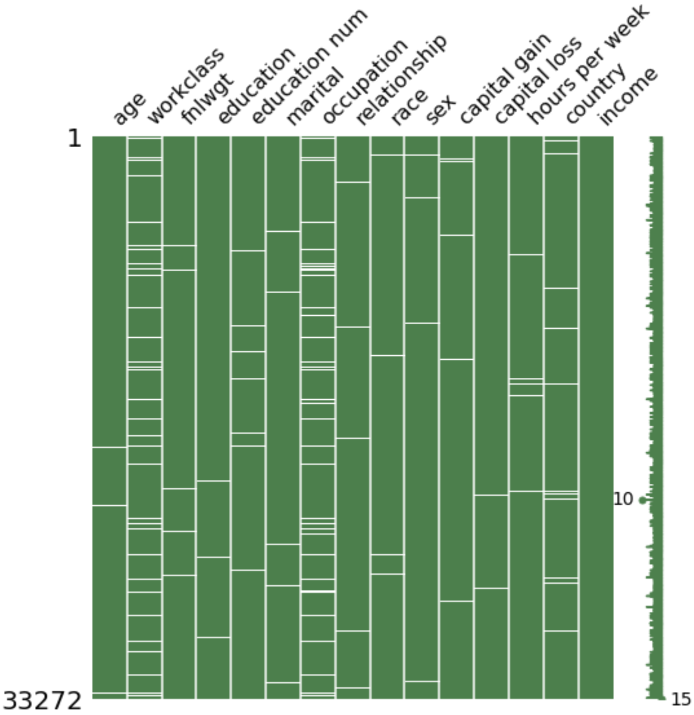


Figure 10. Sex VS Income

By figure 10, The income over 50K rate is 11%, 30.5% each Female and Male. Therefore, '**Hypothesis5:** Male will have higher income than female.' seems to correct.

3. Preprocessing

By our hypothesis, we don't need {work class, fnlwgt, education, marital status, relationship, capital gain, capital loss, native county}. We should remove these column first. Then we remove missing values and duplicated values and undefined values('?'). Finally, we encode categorical feature and scale numeric features. In this process, we should do this task separately in train set and test set. Finally, we use {age, race, sex, hours per week, income}. When we remove missing value data, the cleansed data distribution is almost same as original data. Because, there are not many missing values like figure 11. For example, in Figure 12, it is change of age distribution after remove missing values (upper: before remove, lower: after remove). It seems missing variables are quite lot in age variable which we use for analysis. But, the distribution is almost not changed. Therefore, we can remove missing values.



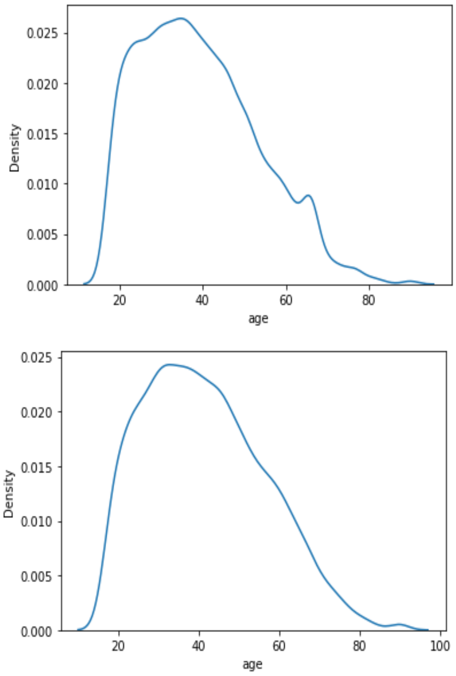


Figure 11. Missing value distribution

Figure 12. Change of Age distribution

4. Model train & test

1. Models

In this data set, data size is quite big. So, we have to limit the max depth and set appropriate number of split and leaves in Decision tree and Random forest. We set optimal hyper-parameters each. In KNN model, the optimum value of k is known to be similar to the square root of the number of samples. Therefore, we select it's near value. In addition, this data is not close to normal distribution. Therefore, I select the distribution-free model which like Decision tree, Random forest and KNN, Support vector machine. These models are not need to distribution assumption. And this data, I reduced to 5 dimension data. It's not many. In DT,RF, KNN models, there are curse of dimensionality. If variable dimension is too high, the model can be overfitting and has poor performance. But this data, I reduced to 5 dimension data. Therefore these models can be good classification model. Finally, It was considered whether to use Linear SVM or Non-linear SVM. In this 4 dimension data, their classification boundaries are often not clear. Therefore, I choose RBF Non-linear SVM model. RBF means that Radial Basis kernel Function. It is close to linear SVM, but boundary is curved.

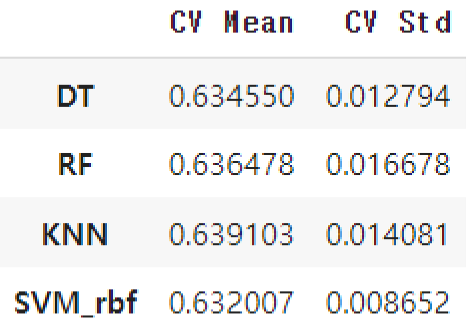
2. Hyper-parameter tuning

When we train the model, there are hyper-parameter which is user defined parameter. The performance of the model depends on it. Therefore, we should find the hyper-parameters close to optimal, each model. I use GridSearchCV form sklearn package.

5. Result

1. Cross Validation

We use 4 model for data analysis which are Decision tree, Random forest, k-nearest neighborhood(KNN), Support vector machine(SVM). We should use 10-fold cross validation method. Because., KNN and Random forest are easy to overfit. We prevent overfitting issue to use cross validation.



	CV Mean	CV Std
DT	0.634550	0.012794
RF	0.636478	0.016678
KNN	0.639103	0.014081
SVM_rbf	0.632007	0.008652

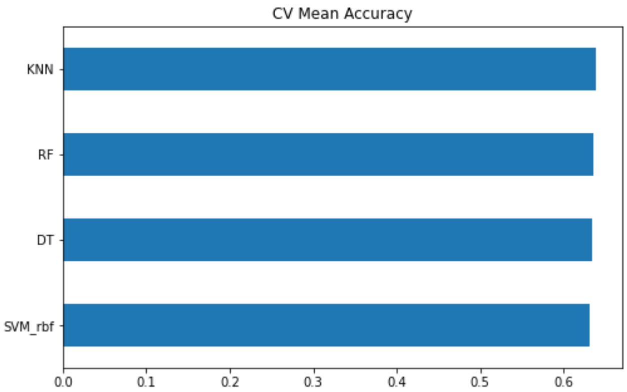


Figure 13. CV mean and standard deviation of accuracy

In Figure 13, there are mean accuracy of 10-fold cross validation. By this figure, we can see that KNN have the highest accuracy as 0.639103 and SVM have the lowest accuracy.

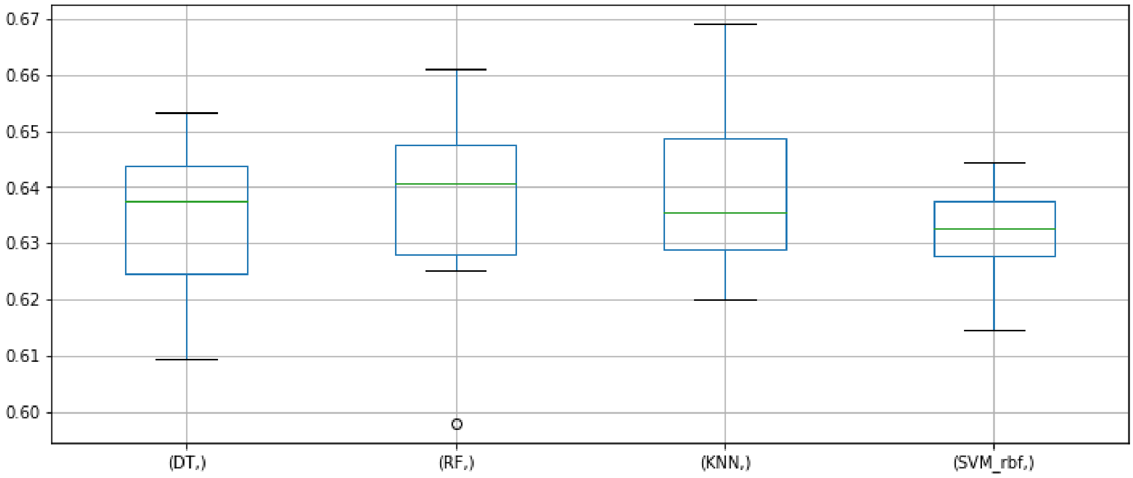
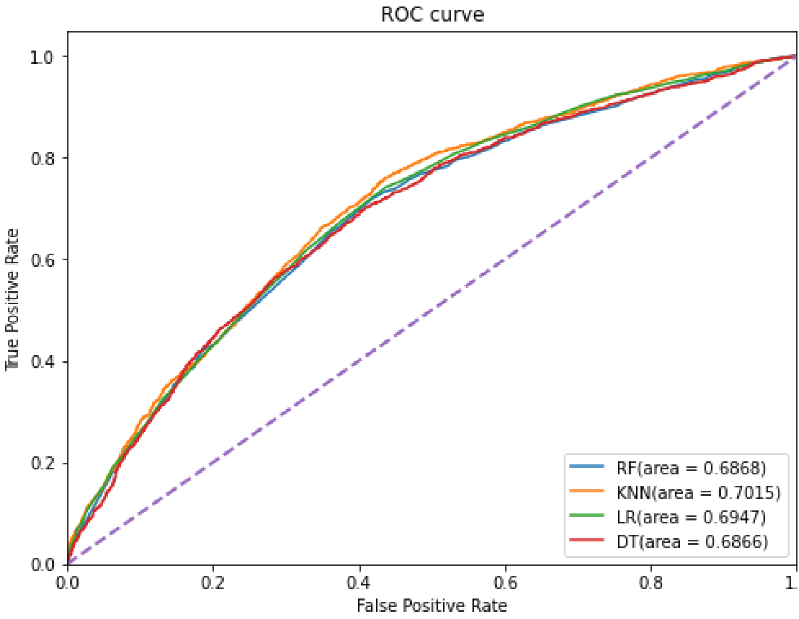


Figure 14. Error bar of 10-fold CV

In Figure 14, there are mean accuracy and standard deviation of 10-fold cross validation with error bars.

2. ROC Curve



	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree	0.6785	0.5487	0.3845	0.4522	0.6868
Random forest	0.6814	0.5574	0.3725	0.4466	0.7015
KNN	0.6785	0.5541	0.3500	0.4290	0.6947
SVM	0.6705	0.5755	0.1718	0.2646	0.6866

This is overall result of evaluation.

The Random forest have highest accuracy and AUC as 0.6814 and 0.7015. Therefore, in our model, Random forest is Best model and SVM is worst model.

6. Discussion & Conclusion

There are 4 feature variable {age, race, sex, hours per week} and target variable {income}. According to my evaluation result, the model is quite poor. Because the AUC and Accuracy is not good as close to 0.7. So, we can interpret that the 4 feature variable {age, race, sex, hours per week} is not enough to explain the target variable. Therefore, my hypothesis is valid but not enough. For this reason, I think why this happened? I think that the reason is that I remove many feature variables. So, I miss the important variable which in removed. Therefore, for improve my hypothesis, I should more data analysis in correlation of feature variables into target variables. In other side, I should improve the model like find more optimal hyper-parameters. Additionally, I should study more knowledge in variables through better EDA.