5002 Assignment 2 Feature Engineering

Student ID:20560903
ZHANG, Peiyi
pzhangan@connect.ust.hk

1. Abstract of my feature engineering

I will briefly introduce my feature engineering and the result of my modeling here. The detailed process and reasons will be shown in the following parts.

For Existing Features				
ID	Feature Name	type	Modification	Reason
1	age	int	Standardization	Distributed mainly between 20-40
2	workclass	string	(1) fill NA with 0 (2) replace "never-worked" with "without-pay" (3) dummy	These two values have similar meaning and have a low count
3	fnlwgt	int	(1) log1p transformation (2) standardization	As the variance of fnlwgt is very large, we shrink it into a small interval
4	education	string	(1) replace "1st-4th", "preschool" with "primary" (2) dummy	These two values belongs to low education level and both have a low count.
5	education-num	int	standardization	May not need to be changed
6	Marital-status	string	(1) replace "Married-AF-spouse" with "Married-civ-spouse" (2) dummy	As they have similar meaning and the former has a lower count.
7	occupation	string	(1) fill NA with 0 (2) dummy	
8	relationship	String	Dummy	
9	race	String	Dummy	
10	sex	string	Dummy	
11	Capital-gain	int	standardization	We keep the zeros as when interpolation on them, the accuracy decrease.
12	Capital-loss	int	standardization	
13	Hours-per-week	int	standardization	
14	Native-country	string	dummy	

(2.2) Add new features We assume education, working-hour and the native-country contribute to your fortune. But the features above may ambiguous or discrete. Therefore, we add new features into the model:				
ID	Feature Name	type	Rules	Reason
15	High-edu	Binary	If education-num>12: Return 1 Else: Return 0	As 12 is the division of going to the university and the 75% quantile.
16	Work-hard	binary	If hour-per-week>45: Return 1 Else: Return 0	45 is the 75% quantile of the working hour per week
17	development	string	**<u>'USA'</u>**: native-country in ['United-States', ' 0'] **<u>'western'</u>** : native-country in [' England', ' Germany', ' Canada', ' Italy', ' France', ' Greece', ' Philippines'] **<u>developing</u>:** native-country in [' Mexico', 'Cuba','Puerto-Rico', ' Honduras', 'Jamaica', 'Columbia', ' Laos', ' Portugal', 'Haiti', ' Dominican-Republic', ' El-Salvador','Guatemala','Peru','Trinadad & Tobago', 'Outlying-US(Guam-USVI-etc)', ' Nicaragua', ' Vietnam', ' Holand-Netherlands'] **<u>'eastern'</u>:** native-country in [' India', ' Iran', ' Cambodia', ' Taiwan', ' Japan', ' Yugoslavia', ' China', ' Hong'] **<u>'polandteam'</u>:** native-country in [' South', ' Poland', ' Ireland', ' Hungary', ' Scotland', ' Thailand', ' Ecuador']	Further add the countries information into the data for modeling

Since then, we have total 17 features for modeling. I build a XGboost model and the accuracy on the validation set is 88.21%

In the following parts, I will illustrate my feature engineering in 2 aspects: overview of the dataset, change on existing features and adding new features.

2. Overview of the dataset

Firstly, I observe the basic information of the dataset:

	age	fnlwgt	education-	capital-	capital-	hours-per-

			num	gain	loss	week
count	34189	34189	34189	34189	34189	34189
mean	38.646 14	189792. 1	10.0771	1073.524	87.64544	40.45284
std	13.679 42	105407	2.565457	7451.486	403.3667	12.48263
min	17	12285	1	0	0	1
25%	28	117847	9	0	0	40
50%	37	178449	10	0	0	40
75%	48	237624	12	0	0	45
max	90	1490400	16	99999	4356	99

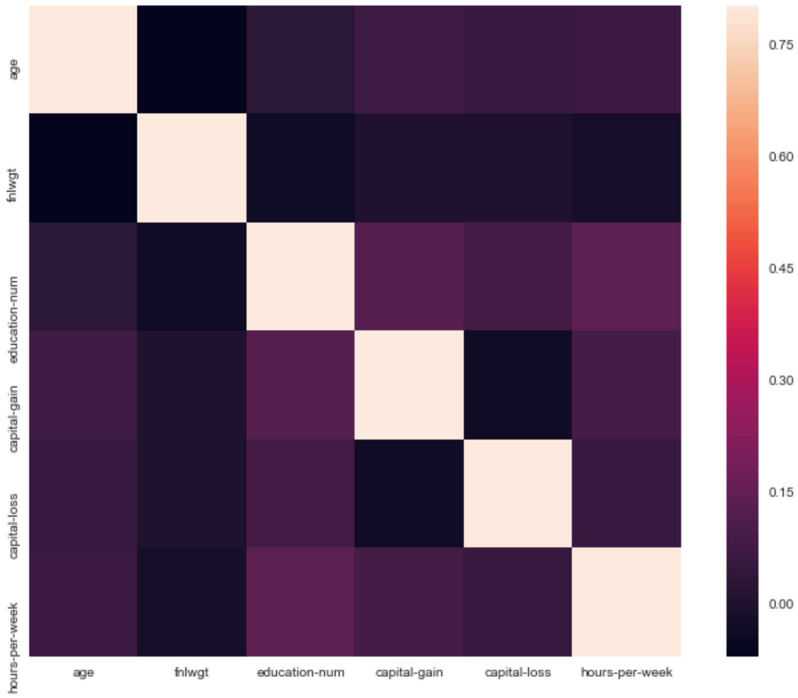
(1) Descriptive properties

There are 14 features in total, including 6 int and 8 string. The max, min, mean and quantile of the numerical features are shown below:

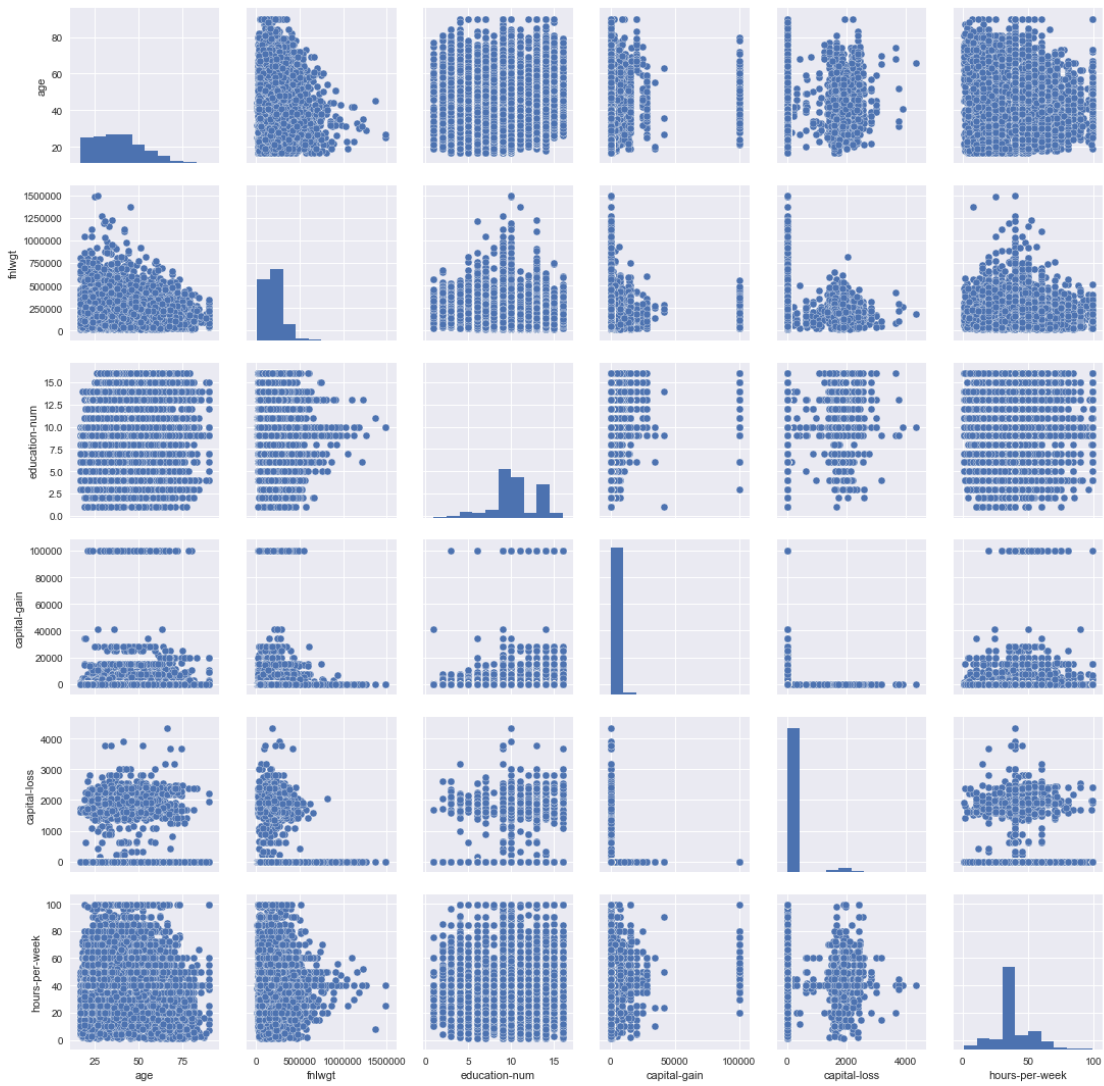
We can see that: more than 75% of capital-gain and capital-loss is 0, existing default value or missing value; the fnlwgt feature has a large standard variance. I will process them later.

(2) Correlation of different numerical variables:

The lighter the color, the higher the correlation: there exists no Multicollinearity.



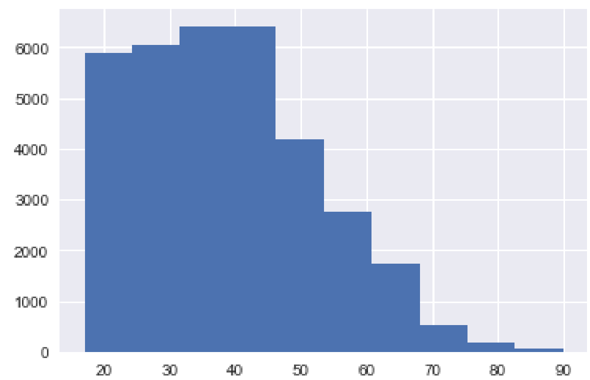
We can go further on the relationship of the variables, shown in the scatter plot:



According to this scatter plot and the property of the features, we make some change on the features.

3. Change on existing features

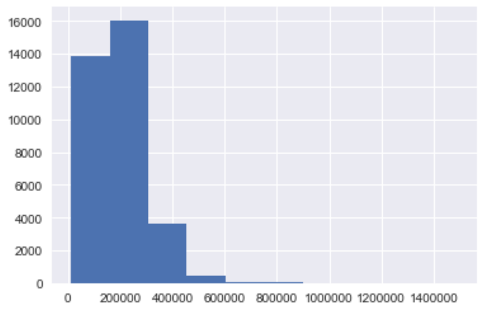
Age: it is mainly distributed between 20-40 and we just standardize it incase of losing distribution information.

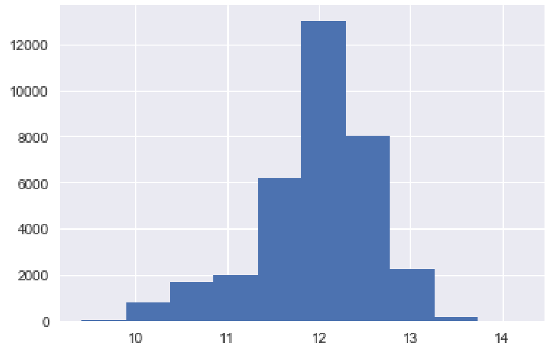


Workclass: we fill the NA with 0, making it as a new value.

```
Private             23702
Self-emp-not-inc     2713
Local-gov            2218
0                    1950
State-gov            1393
Self-emp-inc         1192
Federal-gov           995
Without-pay            26
```

fnlwgt: we reduce the variance from 105407 to 0.6283

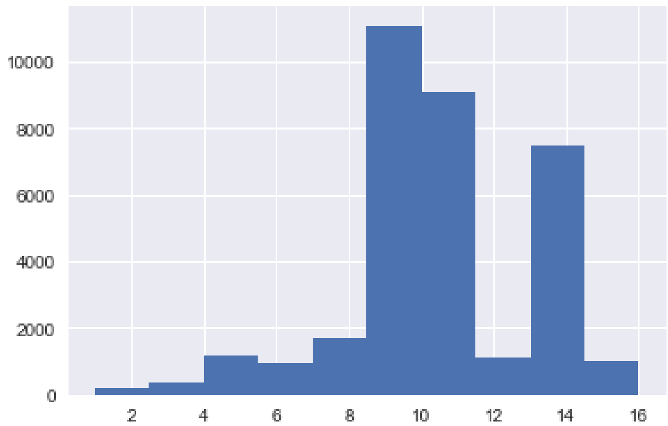




Education: we merge (1st-4th and preschool) into one value as "primary": as they all belong to the low education level and each of them has a low value. After the merging, the feature may become more outstanding

Education-num: we reserve the raw education-num data and add a new feature 'high-education' by splitting the education-num at 12.

Why 12? As it's the division of going to the college and also the 75% quantile of the data, which is the same as the label'1''s percentage.



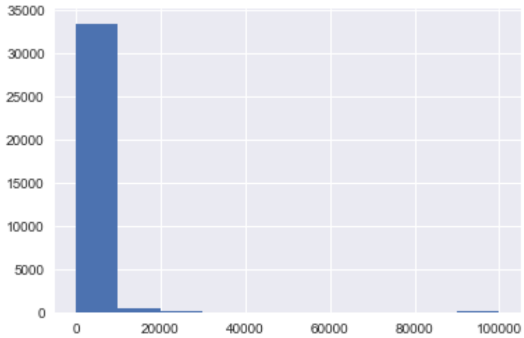
High-education:

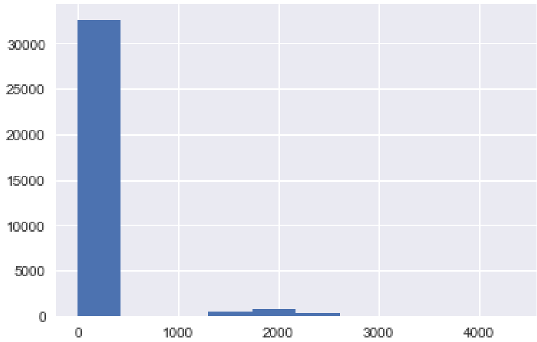
$$\text{high} - \text{education} = \begin{Bmatrix} 1, if\ education - num > 12 \\ 0, if\ education - num \leq 12 \end{Bmatrix}$$

Marital-status: we replace "Married-AF-spouse" with "Married-civ-spouse", as the count of "Married-AF-spouse" is very low, which has the similar meaning with "Married-civ-spouse".

Occupation: we replace the NA with 0 as the new value.

Capital-gain & capital-loss: as Interpolation decrease the accuracy, we do not change the values.





Hours-per-week: no change except standardization.

Work-hard: we extract feature from hours-per-week, as shown below:

$$\mathrm{work} - \mathrm{hard} = \begin{Bmatrix} 1, if\ hour - per - week > 45 \\ 0, if\ hour - per - week \leq 45 \end{Bmatrix}$$

development: we extract the geography information and divide the native-country into 5 values, according to the below chart:

Development_value	Native-country
'USA'	' United-States', ' 0'
'western'	' England', ' Germany', ' Canada', ' Italy', ' France', ' Greece', ' Philippines'
'developing'	' Mexico', ' Cuba',' Puerto-Rico', 'Honduras', 'Jamaica', 'Columbia', 'Laos', 'Portugal', 'Haiti', ' Dominican-Republic', 'El-Salvador', 'Guatemala', 'Peru', 'Trinadad&Tobago', 'Outlying-US(Guam-USVI-etc)', 'Nicaragua', 'Vietnam', ' Holand-Netherlands'
'eastern'	' India', ' Iran', ' Cambodia', ' Taiwan', ' Japan', ' Yugoslavia', ' China', ' Hong'
'polandteam'	' South', ' Poland', ' Ireland', ' Hungary', ' Scotland', ' Thailand', ' Ecuador'