

RMBI3110 Assignment 3 --- Classification

Li Chung Yat

Student ID: 20422979

Lo Hau Yi

Student ID:20421470

As the dataset contains '?' in attributes native country, occupation and workclass which is not meaningful for any data analyses, we decided to perform Data Preparation first.

Data Preparation

Data Cleaning:

Using `df.isna().sum()` function, we observe that there are not NaN value in raw data. However, there are '?' value found in workclass, occupation and native country independent variables. They are properly missing value. This can be caused by privacy issues or careless.

If we take a deeper look on those missing value, there is a pattern can be found. There are a total of 2799 responds which has value '?' under variable workclass. Specifically took a look on these 2799 observations, they clearly show the fact that if workclass has value '?', occupation will also has value '?'. However, it is not necessary for workclass to be '?' when occupation is '?'. Among 2809 observations which contain value '?' in occupation, 2799 of them do also have value '?' in workclass, the remaining 10 has value 'Never-worked'. All these traits boiled down to the conclusion: 2799 interviewees refused to disclose their workclass and occupation due to privacy issues and the rest '?' exists in occupation is because interviewees indeed did not have any working experience. Dropping rows contain '?' in occupation is the most efficient way to cleanse these missing value. Credit to the large dataset provided, we still have lots of records which provide sufficient information required for generating a precise classification rule after dropping these '?' in occupation and workcalss. Native country also contains a few '?' in raw data. These '?' are properly missing value obtained from the survey. Instead of taking the approach we done to occupation and workclass, grouping these '?' to category 'United-States' reserved these valuable information and there are not much effect caused as 91.4% of interviewees were born in America.

<code>isna().sum():</code>		number of records in native-country:	
<code>age</code>	0	United-States	42103
<code>workclass</code>	0	Mexico	903
<code>fnlwgt</code>	0	Philippines	283
<code>education</code>	0	Germany	193
<code>educational-num</code>	0	Puerto-Rico	175
<code>marital-status</code>	0	Canada	163
<code>occupation</code>	0	El-Salvador	147
<code>relationship</code>	0	India	147
<code>race</code>	0	Cuba	133
<code>gender</code>	0	England	119
<code>capital-gain</code>	0	China	113
<code>capital-loss</code>	0	Jamaica	103
<code>hours-per-week</code>	0	South	101
<code>native-country</code>	0	Italy	100
<code>income</code>	0	Dominican-Republic	97
		Japan	89
		Guatemala	86
		Vietnam	83
		Columbia	82

Data Transformation:

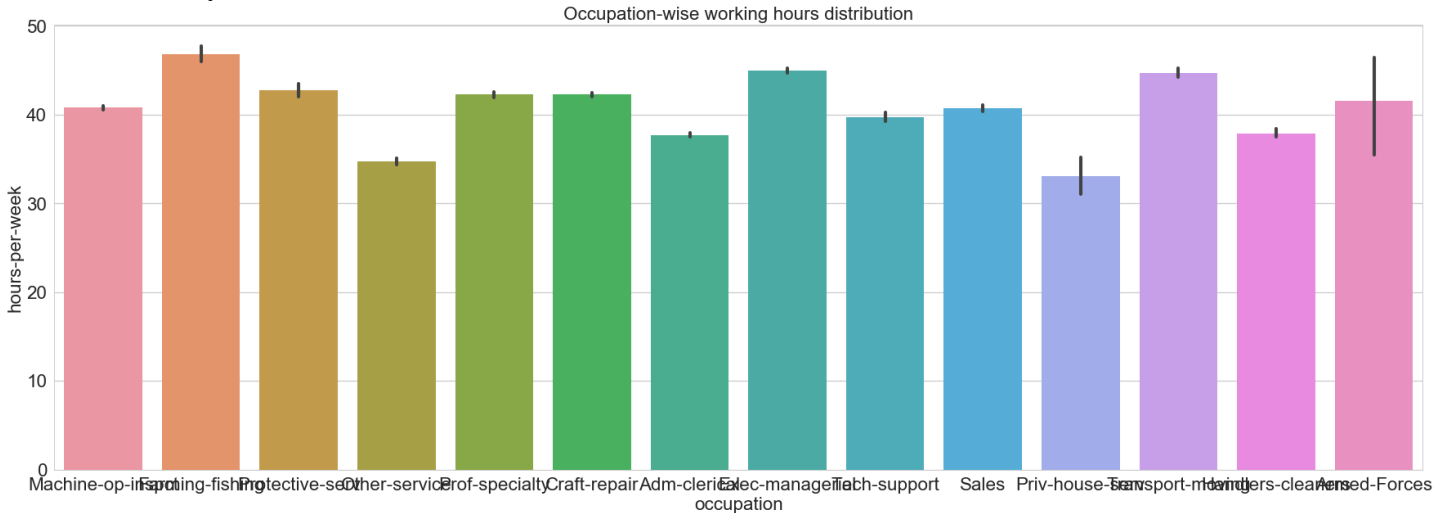
By adopting the Min-Max transformation, all numerical data are standardized. The reason behind is because standardized data will generate less biased fitted model. Besides, all categorical data are transform to dummy variables using the function `get_dummies()`. A categorical variable with n categories split into n-1 dummy variables as 1 category has to be the base case to avoid unprecise fitting in logistic regression.

The dependent variable income also transform into binary variable for logistic regression as well. Using 1 to imply income greater than 50K and 0 to symbolize the case $\leq 50K$.

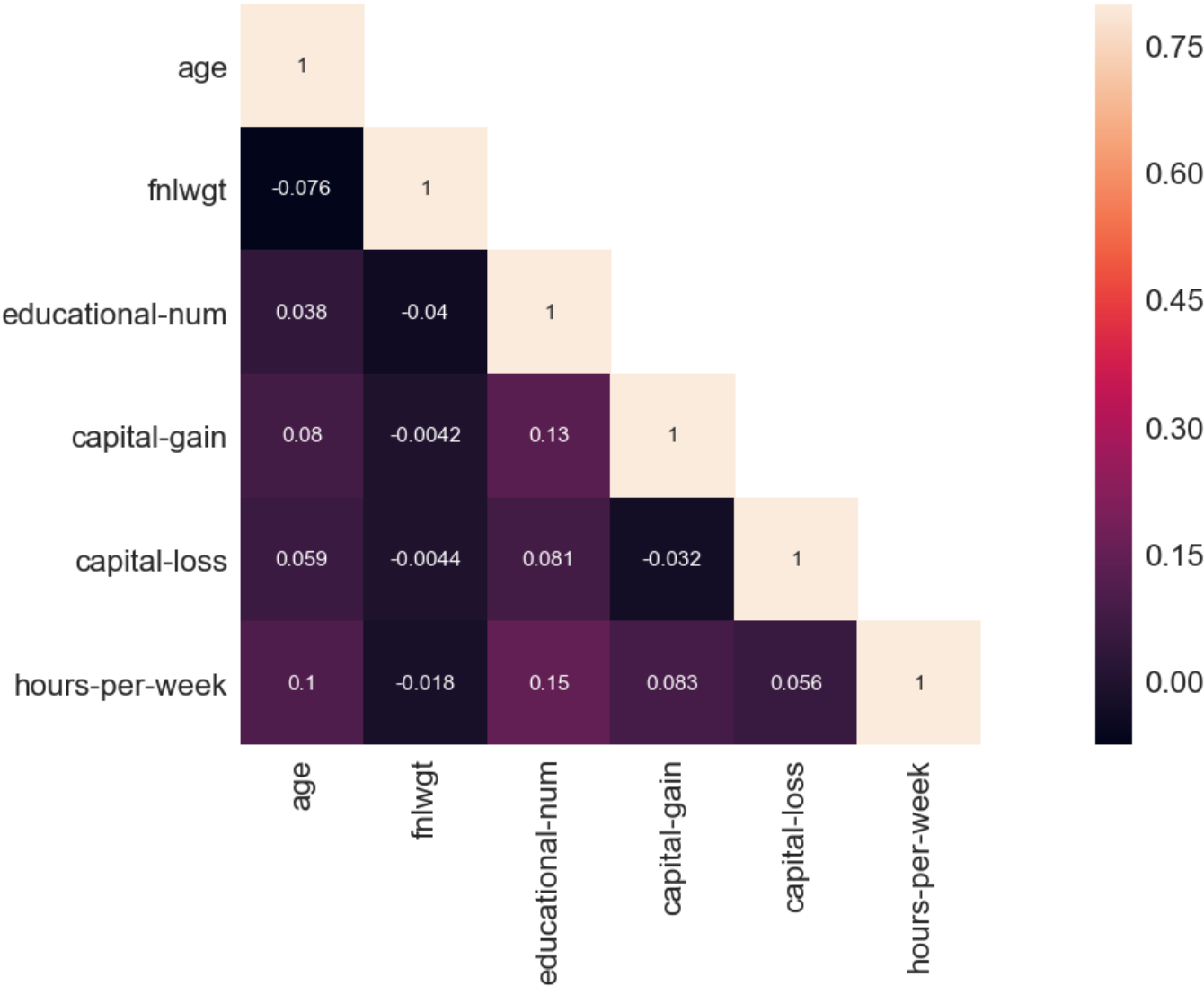
Standardization, however, will be performed after exploratory data analysis in order to study raw data in their original scale so as to give meaningful interpretations.

Exploratory data analysis

Macro data analysis:



This is a plot of working hours-per-week against occupation. From this plot, we can find some interesting observations. All occupations follow a standard working hour and seldom do they deviate from this standard, except Armed-Forces which is understandable as special force and police, as an example, definitely will have different working hours. Farming and Fishing, the agriculture industry, has the longest working hours whereas private house services has lowest working hours. However, all occupations have an average 40 hours generally speaking.



From the correlation matrix, numerical attributes seems to have week correlation with each other.

Micro data analysis:

We decided to split the cleansed data into two groups based on the dependent variable income, $\leq 50K$ and $> 50K$. This slice is for studying the major differences between the two target groups so as to give us a general picture on how their characteristics differ from each other. Pie chart will be one of the best visualizing tools to show composition of variable and significant characteristics in that perspective.

Before visualizing data, a numerical statistic summary is carried out to analyse numerical data in dataset,

```
mean working hours-per-week for class income<=50K 39.38354858282049
mean working hours-per-week for class income>50K 45.69024689196288
mean age for class income<=50K 36.756320245008816
mean age for class income>50K 44.011819296095254
mean educational level for class income<=50K 9.639478778423046
mean educational level for class income>50K 11.612064437051304
```

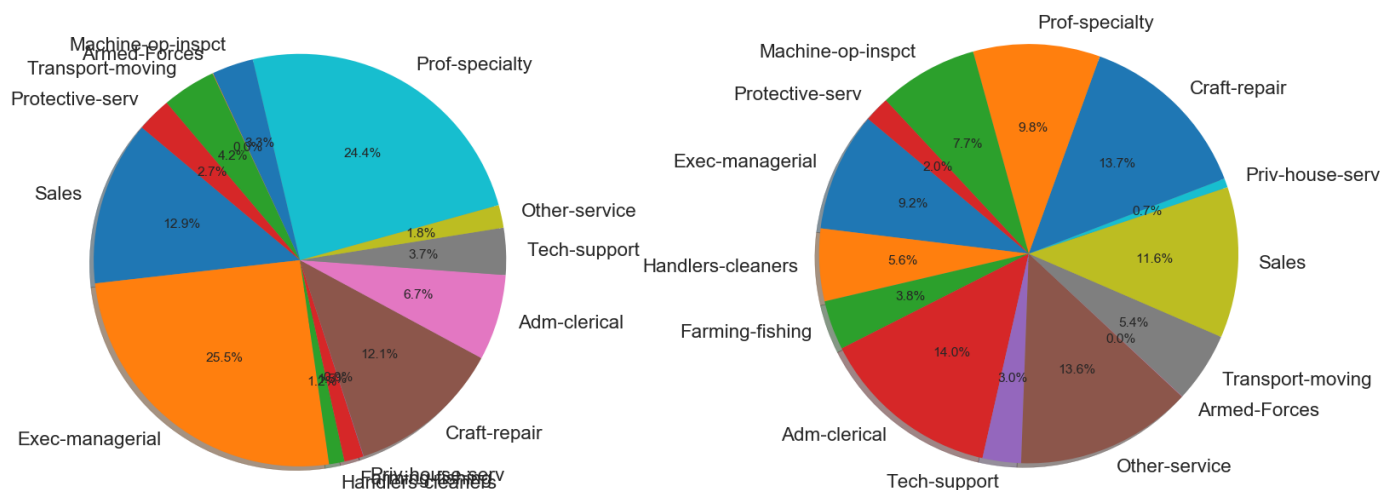
Higher income class in general has higher working hours (45.7 hours per week), this could be the reason why they have higher income.

Average age of interviewees whose income level is higher than 50K is 44. Compare to the $\leq 50K$ class, with average age 37, age seems to have relationship with people's income. However, age might not be the best interpretation for the difference of income level because it is not necessary to have higher salary if you are older. Working experience seems to be the better explanation for this observation as people with longer working experience have higher grade in a company. Higher the grade, higher the salary, this makes more sense to understand the relationship between age and income level.

Educational level for higher income class is also larger than that of lower class which is a common sense. Educational level with 9.6 is equivalent to level of some-college, maybe refers to top-up degree or college level diploma whereas 12 is equivalent to associate degree or close to bachelor's degree.

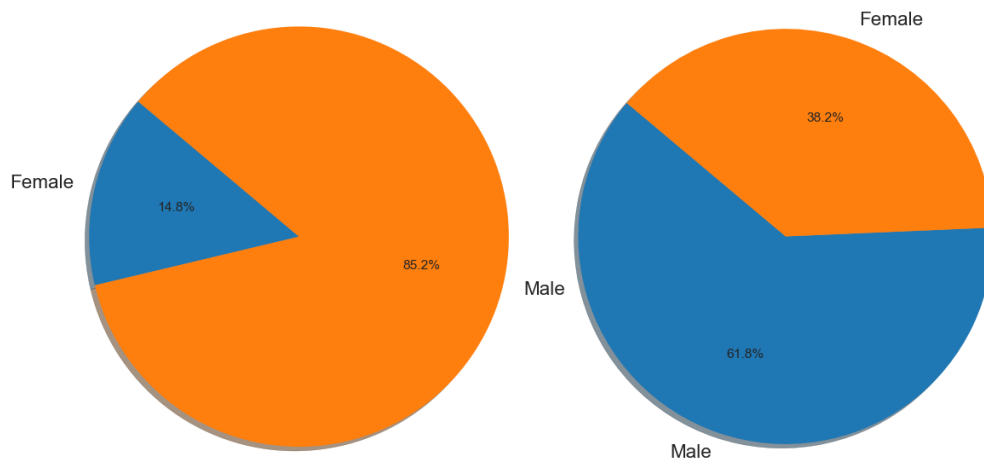
Pie charts on the left are class $> 50K$ and right are class $\leq 50K$,

Occupation:



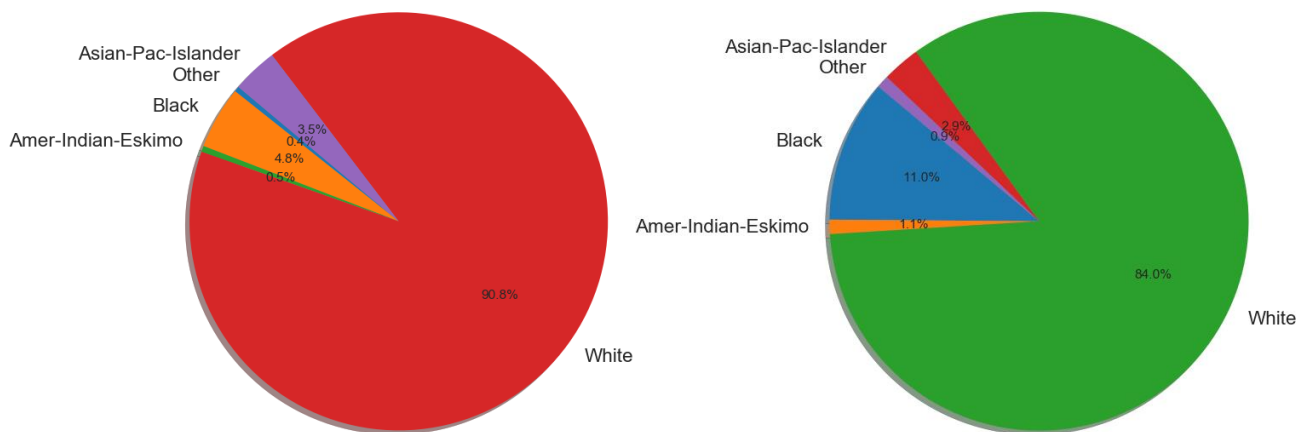
50% of interviewees in class with higher income are working in occupations in professional specialty and managerial roles. Specialists and managers in fact are much more well-paid in modern society. There are no obvious pattern can be observed from class $\leq 50K$, they are almost evenly distributed among all occupations.

Gender:



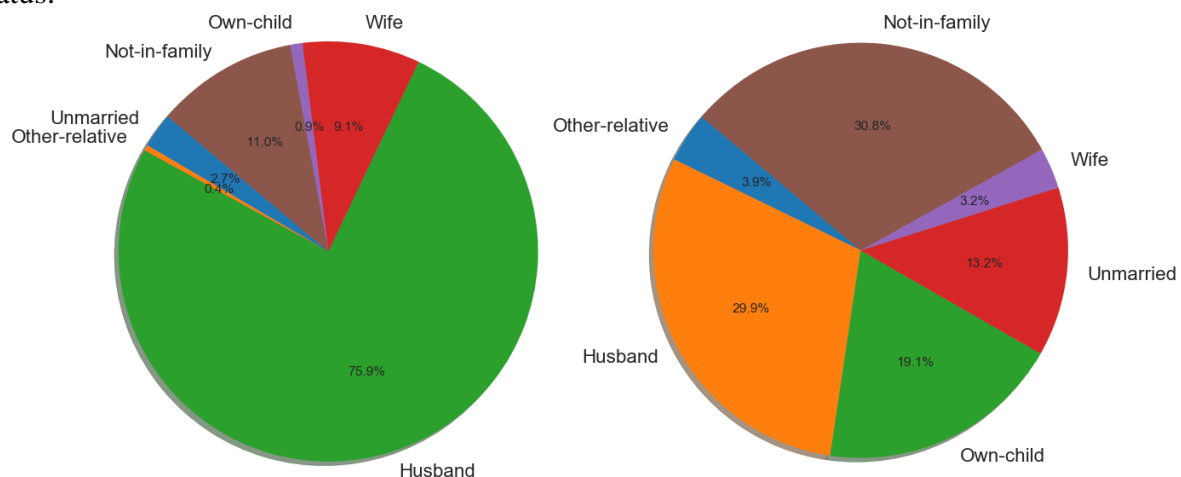
Although the pie charts show quite extreme distribution for gender attribute, the proportion of male and female is not even for this case, male:female ratio is around 2:1. If we scale female interviewees' responds to fit even gender distribution, the lower income class does acquire 1:1 scale. However, in class >50K, male outnumbers female data significantly. Perhaps, gender equality still has room for improvement in workspace.

Race:



As we can see in the above pie charts, most of the respondents in the survey are white people. Combined with prior study on native country, we are confident to claim that the interviews are conducted in United States. Despite the fact that black people are the second largest population in America, they only occupied 4.8% population in high income class. Meanwhile, 3.5% of high incomes are Asian, most of them should be immigrants who have high educational level, like doctoral or master degree holders in professional area.

Family status:

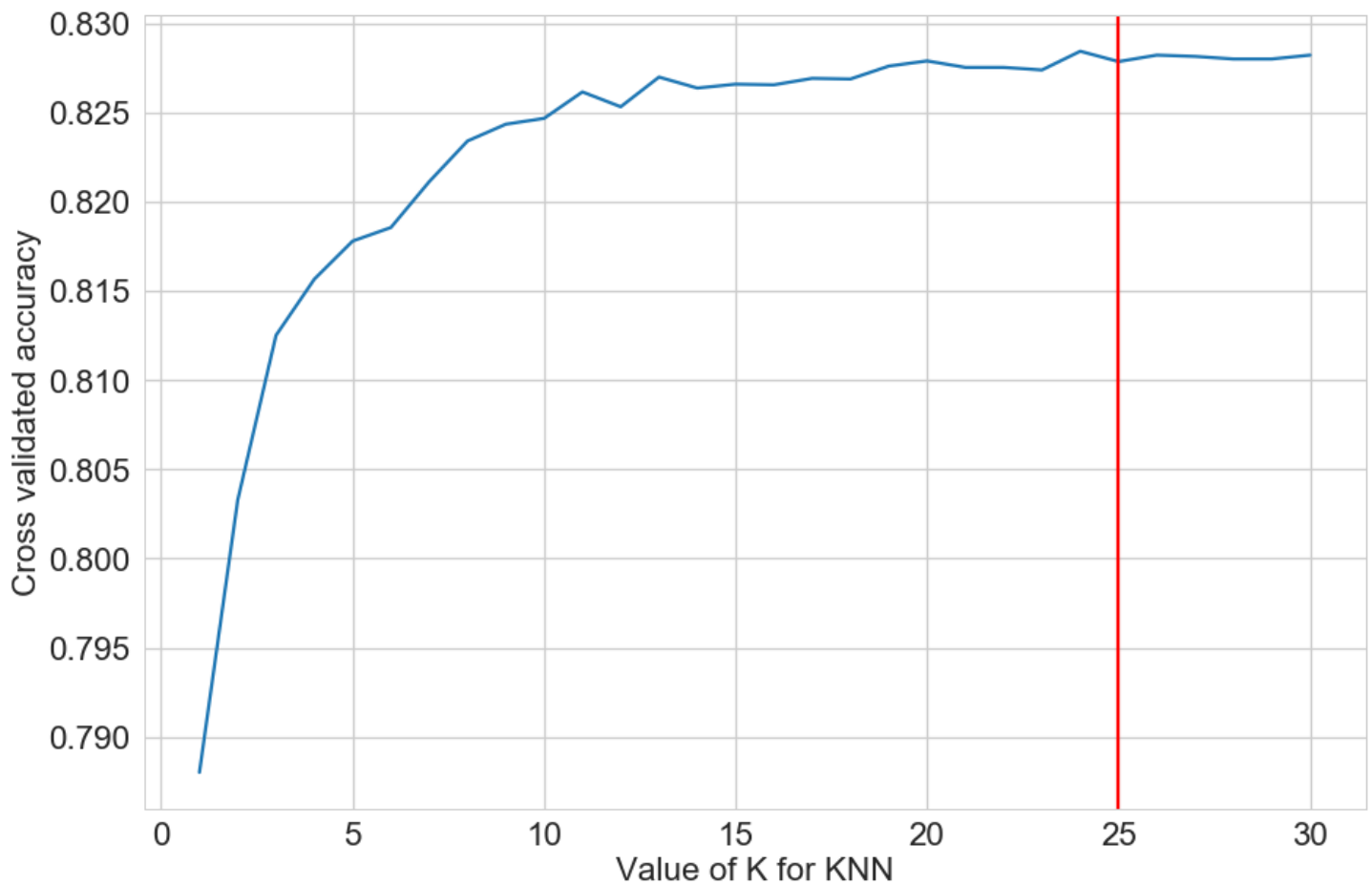


The study on Family status will be more informative if we interoperate the data by married and unmarried. Obviously, we can see proportion of unmarried in lower income class is much higher. Perhaps people will lower income are still relatively young for marriage or they are not financially affordable to set up a family.

Generating training dataset and testing dataset

Instead of using `df.sample(frac=x)`, `train_test_split` of `sklearn` is used. Training dataset is formed by 60% random samples from both classes and testing dataset is the remaining 40%.

Perform K-nearest neighbour, use cross validation to select K



Using 10-fold cross validation, K=25 gives the best accuracy on test dataset

By fitting training data and testing data in KNN learning model with K = 25,

Train_data Accuracy: 0.84

Test_data Accuracy: 0.83

The fitted model has high accuracy and strong prediction power, it performs quite well on both training dataset and testing dataset.

Although the cross-validation gives the best result when K = 24, however, it is not a good choice of K because when K equals to an even number, sometimes the model will encounter a situation that the decision option evenly spread. For instance, 12-12 when K = 24, the rule cannot do a proper prediction but only randomly assign the undecidable record. Therefore, K = 25 is being selected in this case.

Perform logistic regression on the training set and interpret the result.

First fit:

Please check the summary statistic of first fit data in submitted code

VIF:

VIF Factor		features
0	1.7	age
1	1.1	fnlwgt
2	8372.4	educational-num
3	1.1	capital-gain
4	1.0	capital-loss
5	1.2	hours-per-week
6	3.1	workclass_Local-gov
7	6.8	workclass_Private
8	2.2	workclass_Self-emp-inc
9	3.6	workclass_Self-emp-not-inc
10	2.3	workclass_State-gov

The graph on the left is a snapshot of VIF among 96 variables.

Using a function previously adopted in multiple linear regression, the variables that have $VIF > 10$ will be dropped. It is because $VIF > 10$ implies these variables have strong multicollinearity with other variables.

Columns 'education-num', 'native-country_United_States', 'race_White', 'workspace_Private', 'hours-per-week', 'marital_status_Married_civ_spouse' are dropped.

Based on the knowledge we had from exploratory data analysis, most of the above features are the features that majority of respondents have. Thus, they will exist strong multicollinearity,

After dropping VIF, we fit the data in logistic regression again.

The re-fitted summary can be found in ipynb codes submitted.

The top 12 variables that impact the prediction result the most are:

Positively impact,

Capital-gain, capital-loss, education_Prof-school, education_Doctoral and education_Master

Negatively impact,

native_country_Scotland, workclass_without-pay, native_country_holand-Netherlands, native-country-Honduras, relationship_Own-child, native_country_Laos and occupation_Priv-house-serv

However, if we do the hypothesis testing on fitted model against saturated model, the p value is larger than 0.05 which implies that the fitted model performance is not satisfactory.

On training dataset:

Confusion matrix:

	P_1	P_0
A_1	19313	1453
A_0	2764	4089


Classification report:

	precision	recall	f1-score	support
0	0.87	0.93	0.90	20766
1	0.74	0.60	0.66	6853

avg / total 0.84 0.85 0.84 27619

$$\text{Accuracy: } \frac{19313 + 4089}{27619} = 0.847$$

On testing dataset:

Confusion matrix: 

	P_1	P_0
A_1	12877	968
A_0	1843	2726

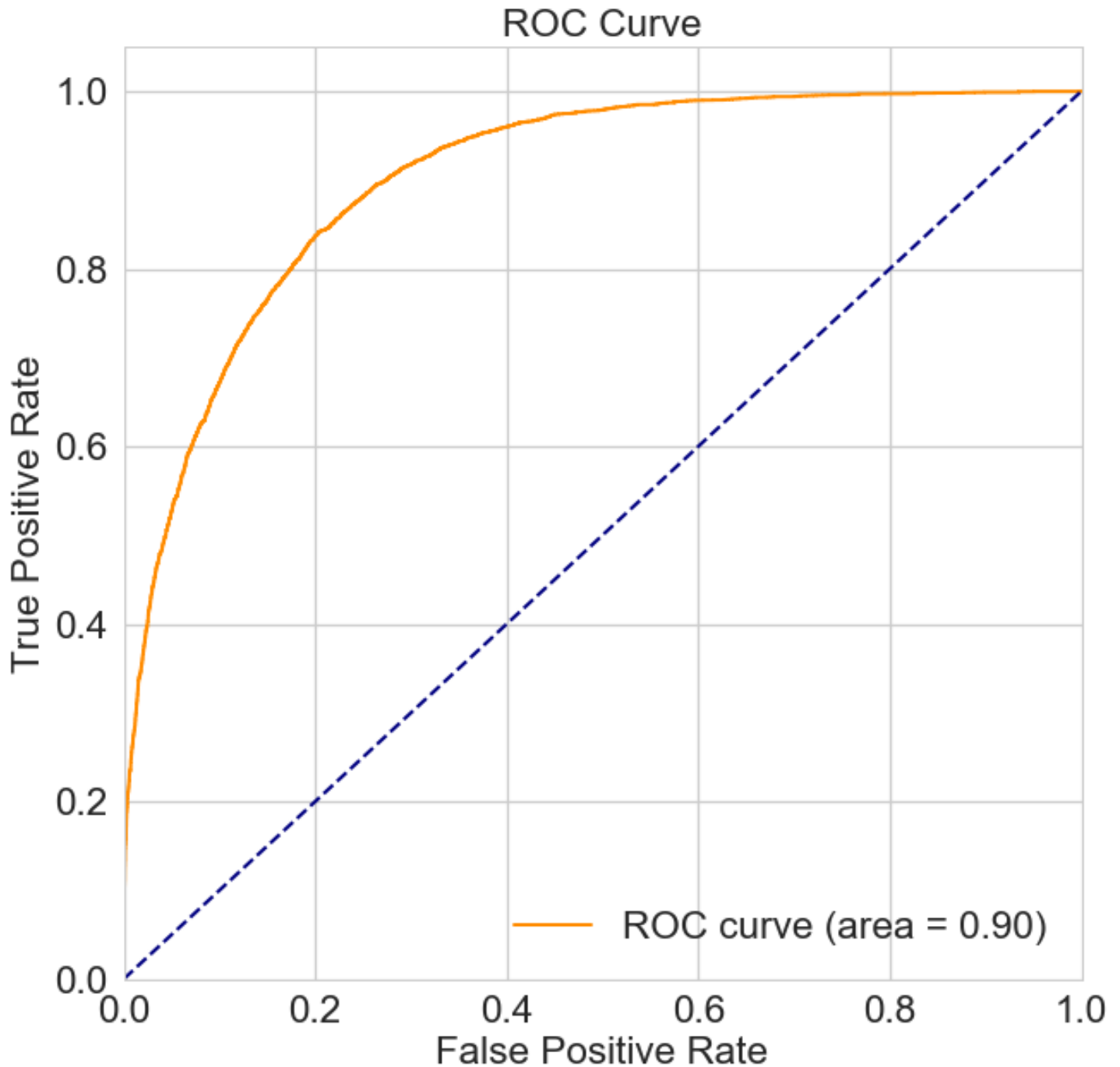
Classification report:

	precision	recall	f1-score	support
0	0.87	0.93	0.90	13845
1	0.74	0.60	0.66	4569

avg / total 0.84 0.85 0.84 18414

$$\text{Accuracy: } \frac{12877 + 2726}{18414} = 0.847$$

ROC curve:



The area under ROC curve is 0.9 which implies that the fitted model in logistic regression has an outstanding performance. Type I error (α) and Type II error (β) are low. As the x-axis of ROC plot is the β , smaller the β , the fitted model is less likely to wrongly reject correct predictions. Whereas y-axis is the sensitivity of the model, higher the sensitivity, more trustworthy the fitted model is. Therefore, the model has a good performance overall.

Please write a short summary (less than 1 page) about this case. Please Indicate the business implication.

Classification results overview:

The ultimate goal for classification is to make accurate prediction on unseen data. Take into consideration of both KNN-classification and logistic regression's result, they both have good accuracy on training dataset meaning the fitted model is well-fitted. More specifically speaking, the nearest 25 neighbours are good predictors for each other in the KNN case and rule generated by fitted features in logistic regression model are useful. Despite the fact that hypothesis testing with saturated model and fitted model in logistic regression suggested that the fitted model is not good enough compare to saturated model. This can be solved by stepwise features selection using AIC or BIC. Both methods also perform extremely well in testing dataset, this reflects that the logistic regression rule generated or the nearest neighbours method have strong prediction power with unseen data, they are precise and objective to determine unseen data as well as not being over-fitted by training data.

KNN result analysis:

We performed 10-fold cross validation on the selection of K ranging from 1 to 31. The running time is long, not to mention we only selected from the range 1 to 31, a large dataset like this, with over 40000 records, should definitely test on a larger range K so as to achieve the best accuracy in cross validation and choose corresponding K in order to generate best result in KNN. However, the process will take so long and not preferable taking time cost into consideration.

Logistic Regression analysis:

The training time is much lower comparing to that of KNN, it is more cost-effective. The positively correlated significant features are Capital-gain, capital-loss, education_Prof-school, education_Doctoral and education_Master and the negatively correlated ones are native_country_Scotland, workclass_without-pay, native_country_holand-Netherlands, native-country-Honduras, native_country_Laos and occupation_Priv-house-serv.

Obviously, logistic regression clearly show that education level are crucial factor for a people to have high income. People graduated from professional school or holding doctoral or master degree tend to get higher pay. It is because high pay jobs, like professionals and managers, require high educational level. For example, in order to be a data scientist, usually, require a master degree in machine learning or related topics. Capital gain and capital loss seems to be strange in which both of them are positively related to income level of a person. However, it actually make more sense than the case which they related to income level oppositely. It is because only people with decent income are capable to make any investment, capital gain or loss only reflect the fact that whether the respondent is a good investor or not.

Most of the negatively related variables are from the attribute native-country. Although the regression result show that people from Scotland, Netherlands, Honduras and Laos tend to have low income level, the dataset contain only few of the respondents who are coming from these country. With a small sample size, it is not persuasive to use this fitted model as an estimator for unseen data to conclude that people who native from these places are low income. The test dataset has high accuracy because test data are also originate from the same survey, the result will be more persuasive if interviewees' originated from are evenly spread in the dataset. Other than native-country, workclass_without-pay also has strong negative relationship to predict a person income level but the name of the variable explained the reason for that. People working in private house services also tends to have low income level, we believed this predictor probably refers to domestic maids or house cleaners, similar to those in Hong Kong, they indeed are under-paid.

Business Implication:

Banks are the group that would be interested in people income. Bank always want to figure out the valuable group of customers who will have a lump sum of saving and eager to invest in financial products. As a bank manager, if there are reliable source we can obtain these raw data and the 2 classification methods developed above, I can develop specific marketing plan or services to the corresponding targeted potential customer group. For example, for the high income class, the banks should send salesmen to promote financial products or encourage them to save up in our bank or provide fix income saving plan to low income class.

Besides, if I were a luxury product salesman, based on the logistic regression result, I will learn more about investment related topics so that I can have a common interest to discuss on with my targeted customers.