MSBD5002 KNOWLEDGE DISCOVERY AND DATA MINING, FALL 2018
ASSIGNMENT 2

Zhe WANG
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
zwangec@connect.ust.hk
STU NO.:20550960

1 FEATURE ENGINEERING

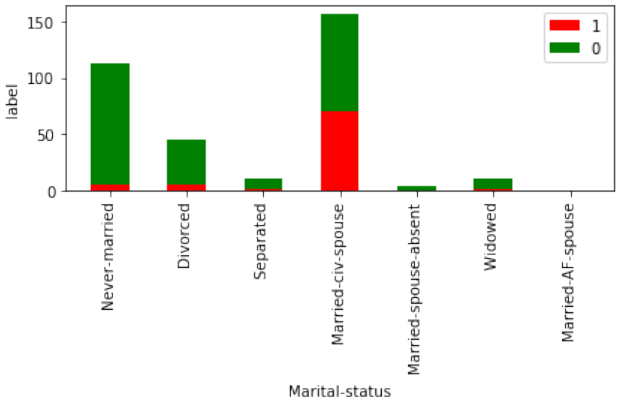
Firstly, the feature values in different columns have different kinds of types. Most machine learning models don't support string type feature values. So we have to transfer the original string type into number. For the missing values, the data set use '?' to stand for missing values. Since we will factorize the data, the '?' will be transferred into same value in same column, which means we use a same value to stand for the missing value in one column, so we don't need to process the missing value, just factorize it.
The factorize function can transfer different feature values into different integers, the result is showing as below:

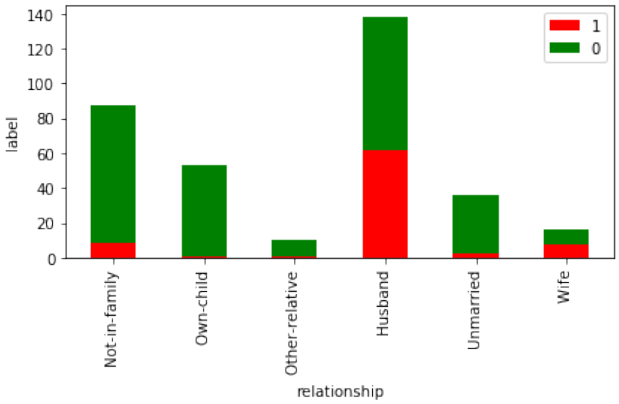
age	workclass	fnlwgt	education	education-num	Marital-status
25	1	258299	1	12	1
29	1	208250	2	11	2
50	1	143460	3	6	3
25	1	228773	4	4	1
21	2	41104	2	11	1
25	1	89155	3	6	1
58	1	157750	2	11	4

The 'workclass' and the 'Marital-status' features' values have been transferred into integers. I tried to factorize all the features, but the accuracy will decrease. That's because the information included in the integer feature values was partly lost after factorizing.

Secondly, I analyzed the label file, the label has 8168 '1' label and 26021 '0' label,the negative label is more than 3 times larger than the positive label. I also analyzed the relationship between the feature values and the labels then choose two feature with significant relationship:

From the picture we can see the feature values ' Married-civ-spouse' from 'Marital-status' and the ' Husband' from 'relationship' can successfully predict almost half of the label rightly, which is far more accurately than the original 0.25, so I combine the two feature values into another feature(if the two labels satisfied the label value, the new feature values is 1, else the value is 0) then used the new data-set to train, the K-Fold result is a little higher than before. So I used this new feature to create the testing result.





2 Model Selection

I tried many ensemble learning model, including XGBOOST, ADABOOST, LightGBM, LR based Bagging, Random Forest and Gradient Tree Boosting.the K-Fold result is showing as below:

XGBOOST	ADABOOST	LightGBM	Bagging	Random Forest	GTB
0.876	0.872	0.859	0.826	0.859	873

The result of XGBOOST and Gradient Tree Boosting is very close, so I choose the XGBOOST model with a little higher score. Then I used grid search to find the proper parameters values of the XGBOOST, **GridSearchCV** function can test different parameters by combining them together greedily.

```
parameters_xgb = {'n_estimators': list(range(500,2501,100)), \
'learning_rate': [0.01, 0.02, 0.05, 0.1, 0.2, 0.3], \
'max_depth':list(range(3,15,2)), 'min_child_weight': \
list(range(1,10,2))}
```

After the grid search, I got the best parameter showing as below:

```
xgb:{'n_estimators': 1000, 'learning_rate': 0.1, 'max_depth': 3,\
 'min_child_weight': 9}
```

But after testing, I got a conclusion that expect the 'n_estimators' parameter. Using the default parameter values seemed to get a better K-Fold score. So I only choose the grid search result {'n_estimators': 1000} for my final model.

3 WHY NOT VOTING

I tried the voting method to build all the ensemble model together, but in the K-Fold cross validation, the voting result will decrease the accuracy of the best model. That's obviously, the voting is used to avoid over-fitting of single model. Considering about the highest accuracy 87.5%, it's far from over-fitting. So I believe the single XGBOOST model can have higher accuracy than the voting model. That's why I didn't choose the voting model.

4 CONCLUSION

I used two feature values to create a new feature to increase the accuracy of my model prediction. Then I factorized the all features that are not integer to fit the model requirement. I chose the XGBOOST model with parameter {'n_estimators': 1000} to make the final prediction. I append the train set and the test set together to factorize, then split them and used the train set with the train label to train the model and used the test set to figure out the final result.