Emre Uludağ
50209
Hw07

In this 7th homework I am working on a dirty dataset to predict the results of 3 different target groups. These groups of people are:

- People who will delay credit card payment more than 1 day (target1)
- People who will delay credit card payment more than 31 day (target2)
- People who will delay credit card payment more than 61 day (target3)

I applied the following steps as I was working on getting the results:

-First I read the dataset from the csv files:

hw07_training_data.csv
hw07_test_data.csv
hw07_training_label.csv

respectively and drop the id columns.

-First I eliminated some meaningless columns from the data

-Then I classified this data into numerics and non-numerics.

-I applied one hot encoding to the categorical data

-I applied standard scaling to the numeric data

-I combined these two dataframes into one

-I applied dimensionality reduction in the form of PCA to the data set.

-I fixed the number of columns to 100 with PCA-Dimensionality Reduction.

-For the three target groups I created training and test set.

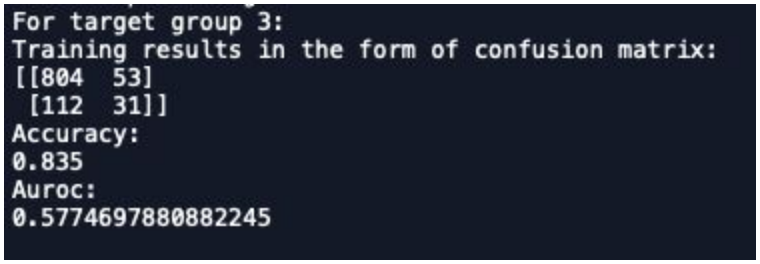
-I separated the data set with a ratio of 80% 20%

-I trained my three training data sets with the training set labels with these splits.

-This allowed me to choose the best algorithm for each target group.

```
For target group 1:
Training results in the form of confusion matrix:
[[1836   62]
 [ 204   98]]
Accuracy:
0.8790909090909091
Auroc:
0.6459186735427324
```

```
For target group 2:
Training results in the form of confusion matrix:
[[1684   15]
 [  92    9]]
Accuracy:
0.9405555555555556
Auroc:
0.5401400940564922
```



Results were like the following:

-Other results for AUROC with different algorithms(top 3 strongest algorithms) are noted on the code.

-With the chosen algorithms, I then trained the whole data set (the data set that I preprocessed with the same methodology) for predicting labels for target 1-2-3 test data. I applied the predict() function of sklearn, saved the results in :
y_predict_real_1,
y_predict_real_2,
y_predict_real_3,
respectively.

-Finally I wrote the results in these variables to csv files:
hw07_target1_test_predictions.csv
hw07_target2_test_predictions.csv
hw07_target3_test_predictions.csv
respectively.