Classification Assignment

Name: Adham Mohamed
ID:20192945

Promoter: Toon Calders
Mentor: Sam Pinxteren

Universiteit
Antwerpen

Data Mining Course

April 18, 2020

1 Code Description

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.
In our project using classification algorithms to know which one has the best accuracy.
firstly i analyzed the data set for know how can i manipulate with the features, i found that 5 columns have NaN values so i filled it with the mod to decrease the loss of data set, secondly,there are string columns so i used (Label Encoder())library to change it to numerical, now my data sets is clean and i can build my classifier,thirdly, i have tried some classification algorithms as shown in1

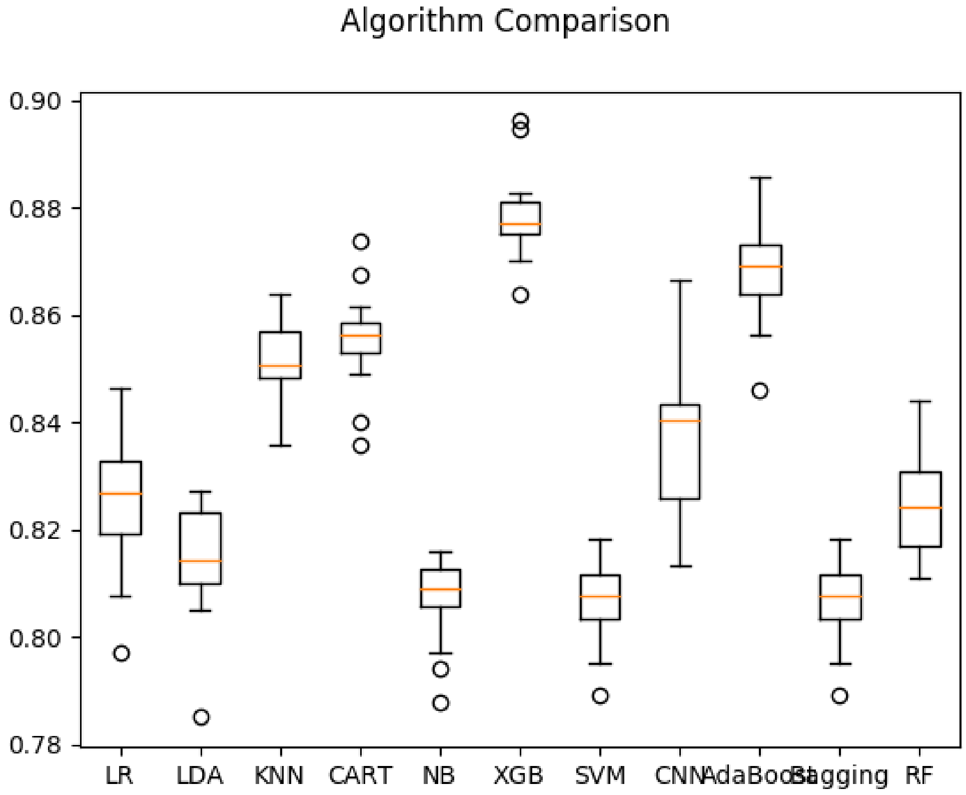
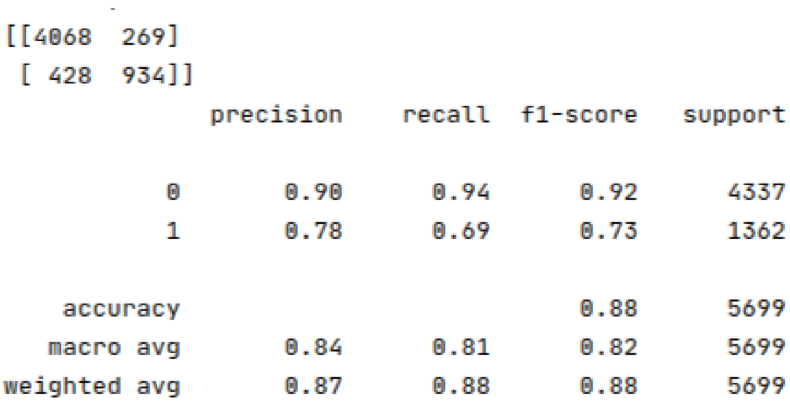


Figure 1: Algorithms Comparison
.

finally 'XGBClassifier' is the best classifier for this problem and has accuracy (88%) and the classification report has the precision, recall and F1 scores after XGBoost hyper parameter tuning by doing a grid search.as shown below in 2.



```
[[4068  269]
 [ 428  934]]
              precision    recall  f1-score   support

           0       0.90      0.94      0.92      4337
           1       0.78      0.69      0.73      1362

    accuracy                           0.88      5699
   macro avg       0.84      0.81      0.82      5699
weighted avg       0.87      0.88      0.88      5699
```

Figure 2: Classification Report and confusion matrix

.

the main program it is been attached with name (classifier.py),there are also (test.py) with some visualizations for test and train data sets distribution .

2 The Solutions

(Q1):my expected profit for high income people, 10% of them are likely to accept the offer,on average the total high income people is 3096 person, so the profit 10% of them with accuracy 100% is :
profit for 100% accuracy =(number of high-income *accuracy * 0.1 *980)-(the mailing costs)= (235805.18169012427)Euro.

(Q2): the IDs of the people in (potential-clients.csv) to send a promotional package to, i will attach the file(ids.txt).

(Q3): my estimation of the total profit depends on what the profit and the costs, so the profit 10% of them with actual accuracy (88%) is :
the income for (88%) accuracy = (number of high income * accuracy * 0.1 * 980)=(266765.1816901243)Euro.
there are almost 12% from that people will be low-income people so they will cost 310 Euro per person.
and the cost for that people is :
(number of the high income*(1-accuracy)*0.05*310)=5795.5477939089205 Euro.
the Total profit =(Income -cost-mailing costs)=230009.63389621535 Euro.
*you will find all the accounting in the (profit.py) file