Tutku Kılıçaslan
54201
ENGR 421
Homework #7

CLASSIFICATION PROBLEM: CREDIT CARD DEBT DELAYS

Objective

The main objective is to classify each customer payment behavior according to their various types of data. We are given three different data sets and three time intervals. Our aim is to predict if the customers in the data set will pay their debt in the given time intervals. At the end, we label the customers as zeros and ones according to their payment behavior. Our measurement will be AUROC.

Observations

As I observed, we have lots of null values and some type of categorical values like R4. We need to handle them.

Brainstorm

- Most proper algorithms for our problem are **LDA, Neural Network and Decision Tree and PCA.**

- We can get higher accuracy by using **different algorithms**.

Combination of different algorithms gives us higher accuracy. For our data PCA and LDA can be suitable since our data is massive.

- We can get higher accuracy also by using **different parameters**.

Some algorithms are sensitive to parameters. Therefore a combination of algorithms with different parameters can give us better results.

- We can get higher accuracy by using **different input representations.**

We can give different input to the algorithm and we can get higher accuracy by using these different results from different algorithms. For our data, we can use the PCA method to extract features and obtain higher accuracy.

My Solution Method

Since our data has lots of different features and our target values are only ‘0’ and ‘1’, I thought that we need to extract features and then apply our training algorithm.

Before extracting features, I needed to do one-hot encoding to organize some features. After that I need to handle null values. This was a challenging problem because the methods depend on which algorithm we use. While assigning zero for a neural network sounds reasonable, it might be misleading for decision trees. Since I use LDA, I decided to assign column mean.

After that I applied a PCA algorithm to process my data apriori. This enable us decreasing number of feature to 143 which is very significant in terms of computational.
After preprocessing part, I applied LDA fitting algorithm. I found LDA algorithm on the internet so it applies bootstrapping by default.

I applied k-fold cross validation at first. However, I struggled by obtaining good AUROC level.Therefore, I left cross validation as default.

As I implemented the LDA model, I obtained accuracy as 0.8651019 after 25 replications for the first training data set. After this, I calculated the AUROC value for my data. I received the value
0.7464968. This value is relatively successful for our problem. Our classifier is able to distinguish the customer who pays the debt and the customer who delays the debt more than 1 days.

Here I assumed that my algorithm will work well for other training sets as well. Therefore, I applied the same procedure for the second and third data set. In the ideal case, we could have applied different algorithms for these three training data sets individually. This would be another approach.

The LDA algorithm gave me the accuracy 0.9389133. I also obtained 0.546388 AUROC for the second data set. This result shows me that my approach is not proper for the second data set. Because this says to us that my classifier barely distinguishes the customer who delays the debt more than 31 days and less than 31 days.

With the same approach, I obtained an accuracy level of 0.8588933 which is very good and 0.6248622 as AUROC.

Therefore, I assumed that the payment behavior of the customers in test sets can be predicted by this model.

Improvement

In the solution key, we were given an xgboosting model. The solution was quick and dirty. Since our measurement is AUROC I tried to improve it by implementing LDA.

I improved the AUROC values from **0.9830176** to **0.7464968** for the first training set. At first it may seem non-improvement. However, where AUROC has almost value 1, it means that our algorithm memorizes the data and does not learn it. For a good classifier we need the value more than 0.5 and less than 1. 0.7-0.9 are very good classifiers.

Conclusion

During the semester, we have learnt to implement different algorithms. However, one algorithm does not always give the best solution. We can improve our solution by taking the advantage of diversity. In my solution, I improved the diversity of my input representation by PCA and then applied LDA. My performance measurement was **AUROC**. At the end of the implementation, I obtained **improvement** in the data sets. It is important that this solution

may not be the best solution but it solves our problem and improves the error with a good amount.

References

- https://www.r-bloggers.com/computing-and-visualizing-lda-in-r/
- https://www.rdocumentation.org/packages/caret/versions/6.0-78/topics/preProcess
- https://campus.datacamp.com/courses/helsinki-open-data-science/clustering-and-classification?ex=7