Scalable Machine Learning and Deep Learning : Lab 1

Sina Erndwein
Samuel Leonardo Gracio

December 2, 2018

1 Short explanation of the lab

The lab was divided into two parts : the first part is a "FILL IN" part where a big part of the code was already given and the second part was a more free part including an end-to-end project for a large-scale Machine learning algorithm. In the first part, we worked with a "price housing" data set in Scala and managed different tasks. The main purpose was to find the best model to predict the different prices of houses depending on several attributes. The used models were a Linear regression model, a decision tree regression model, a random forest regression model and a gradient-booster forest regression model.

For the second part, both Python to plot graphs and Scala to implement the different models were used. Our program doesn't have any parameters, you just need to execute it.

2 Results

2.1 House Prizing

The results before tuning can be seen in table 1.

Table 1: Comparison of Models

Model	RMSE test data
Linear Regression	68877.26
Decision Tree	67719.21
Random Forest	65880.55
Gradient-boosted tree	56288.71
Hyperparameter tuning	53671.89

So, finally, the best result is given by the Gradient-boosted tree regression model. Hyperparameter-tuning shows that the results can be improved.

2.2 Creditcard Dataset

The first step for the more free end-to-end algorithm was to explore the data. This was done by plots which can be seen in figures 1, 2 and 3. All in all the data-set includes 30.000 entries.

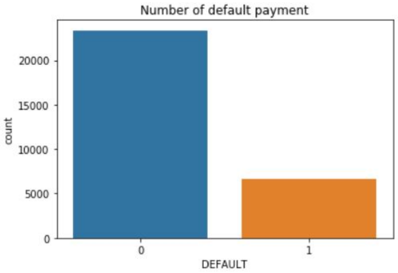
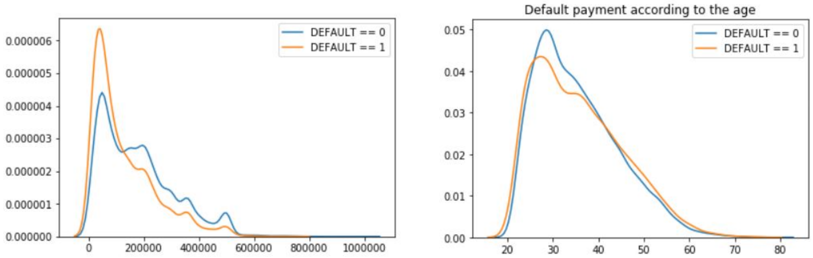


Figure 1: Distribution of Default Values



(a) Relationship between Default status and credit on credit card

(b) Relationship between Default status and Age

Figure 2: Relationships between different features and Default status

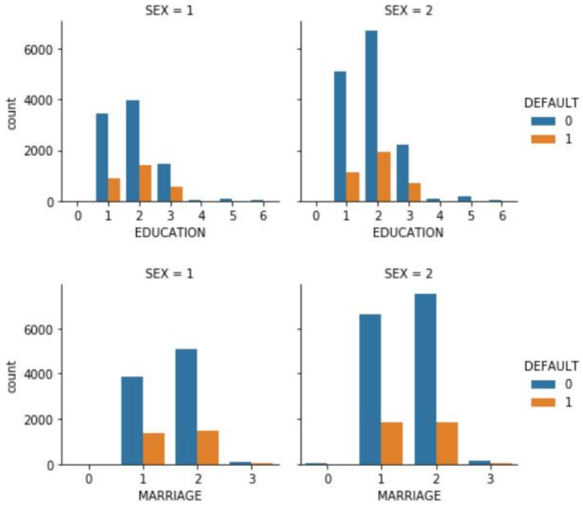


Figure 3: Relationship between Sex, Education and Default status

3 Conclusions

In table 2, the results for the second data set can be seen which is evaluated by the ROC score.

Table 2: Comparison of Models

Model	ROC test data
Logistic Regression	0.5
Random Forest	0.6524
Decision Tree	0.6609

In order to choose the best model, we need to check which model has the best score but also which is the simplest and fastest model. It seems that the **decision tree classifier** and the **random forest classifier** have almost the same results.

Fbetween these two models, we could compute the variance and the bias of each model and try to choose the best one.

What more would you do with this data? Anything to help you devise a better solution?

In order to improve our solution, we could filter the data according to the different graphs we have made. Some features are not important and might changes the result.

What we could also do is to search for better parameters and fine-tune them more or try other models.