Axa Assignment

Hugo SCHINDLER

February 13, 2021

I spent two evenings on your homework after school. It is divided in two parts: Julia[1] pre-processing and Python training and optimization. I chose Julia at the beginning because I like this language. But given the small size of the data-set, deep learning methods were not possible. So I chose more traditional classifiers. Scikit earn[2] is interfaced with Julia, but Bayesian python optimization[3] is very good. This is why the second part of my work is based on python.

My results are available in the coma separated values file result_schindler_hugo.csv. I can, of course, give you more details if you wish. Excuse me, I have a lot of work with Georgia Tech.

1 Pre-processing

I made an HTML export of the Julia Pluto[4] notebook. Here are the steps:

- Data cleaning: missing values imputation. Some dates were missing, I replace them with today.
- Transform dates strings into date objects to be able to compute easily duration what are more understandable by our models.
- One hot encoding training and scoring data-sets. No new class were found in the scoring data-set.
- Remove useless attributes like the index.

I exported this pre-processing to CSV files.

2 Training

I made an HTML export of Python notebook.

- After importing data, I set up a min max scaler to not break the one-hot encoding.
- I split the data into training and testing partitions 80/20%.
- The training dataset is imbalanced. To solve this, I chose a SMOTE sampler[5] on the training dataset.
- I picked three classic classifiers. After trying to optimize them with the bayesian optimization. The best model I found was AdaBoostClassifier[6]. The results are summed up inside Table 1

[1] https://docs.julialang.org/en/v1/
[2] https://scikit-learn.org/stable/
[3] https://github.com/fmfn/BayesianOptimization
[4] https://github.com/fonsp/Pluto.jl
[5] https://imbalanced-learn.org/stable/generated/imblearn.over_sampling.SMOTE.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

	precision	recall	f1-score	support
non résilié	0.98	0.92	0.95	166
resilié	0.69	0.91	**0.78**	34
accuracy			0.92	200
macro avg	0.83	0.91	0.87	200
weighted avg	0.93	0.92	0.92	200

Table 1: Best AdaBoost Classifier test results