Machine Learning Evaluation

Fabrice Rossi

! General rules

Read the full document before starting to code.
The use of AI tools during this test is authorised for coding purposes. You may consult course documents, the wikipedia, etc. The test is strictly personal and you cannot seek help from anybody.

1 Setting

In this evaluation, you are given two data sets in the following files:

- `dataset.csv`: the core data set in which the target variable is known;
- `evaluation.csv`: the evaluation data set for which you are expected to produce predictions.

1.1 Data description

For technical reasons, the data sets are presented in a very abstracted form with non informative variable names (and modalities for the nominal variables). They contain:

- an identifier variable `idx` used to identify uniquely each object described in the data sets;
- 25 explanatory variables;
- and in the core data set only a target variable named `target`.

The core data set contains 4819 observations while the evaluation data set contains different observations.

1.2 Expected results

Your goal is to build the best possible predictive model on the core data set. This model has to predict `target` using the explanatory variables. Notice that this is a regression problem and you have to chose an adapted quality metric (as well as report adapted metrics).

You have to report in a pdf document:

- the expected performances of your model on future data, expressed with adapted metrics;
- the performances of your model on the core data set, using adapted metrics and adapted visual illustrations.

The report should contain additional content such as remarks on the data, a description of the way you built your model, a variable importance analysis, etc. Remember that you can only use AI tools for the code, not for any other content.

You have also to produce predictions for the evaluation data set:

- the predictions must be stored in a CSV file named `predictions.csv`;
- the CSV must contain exactly two columns:
  - the first column must be named `idx` and must contain the identifiers of the objects in `evaluation.csv`;
  - the second column must be named `target` and must contain the estimated value for `target` given by your best model.

> ❗ Important
>
> This file will be processed by an automated script. If you do not respect the format and the instructions, the script will fail and you will fail the corresponding part of the evaluation.

1.3 Testing service

During the lab, you will be allowed to upload your `predictions.csv` file once on moodle to get an idea of your performances. This is not an automated service so do not expect fast answers.

1.4 Final submission

At the end of the lab, you have to upload on moodle an archive file with the following content:

- the small report in pdf;
- the CSV file with the predictions;
- your full code;
- your data sets.

The code must be executable *as is* from the directory in which it will be extracted. In particular you **must** use local file names only.

> ❗ Important
>
> Notice that if you code attempts to install a package or to modify things outside of the extraction directory, you will fail the evaluation.
> For R users, `pacman` does install packages, so **do not use it** in this code.

2 Some advice

In no particular order:

- use a resampling strategy
- try first a simple and fast model such as a decision tree and switch only to something more complex (e.g. random forests) when your code is fully validated
- `idx` is **not** an explanatory variable
- while you may produce a small report from jupyter-notebook, it will be ugly and your grade will suffer from that
- some variables are not numerical. You must pre-process them for instance by using a so-called one-hot-encoding. Use first a simple strategy (like one-hot-encoding) and move to something more complex only if needed