TÉCNICO+
FORMAÇÃO AVANÇADA

1st DASH - 2021

Forecasting Project

Deadline – May 30nd 2022 @ (23:59)

PROJECT GOAL

Critical application of data science techniques to forecast time series in several domains.

Students are asked to explore one dataset and, in accordance with their findings, adequately select and learn models from the available data, as well as assess and relate those models.

Additionally, students should be able to criticize the results achieved, hypothesize causes for the limited performance of the learnt models, and identify opportunities to improve the mining process.

DATA

The datasets to explore in the project have to be multivariate and a continuous target variable, in order to allow the train of classification forecasting models. Students are encourage to use datasets in their area of expertise, either chosen among the options available or any other source.

METHODOLOGY

The project should be developed according to one of the standard data science processes, for example CRISP-DM. Among those steps, only *data profiling*, *data preparation*, *modelling* and *evaluation* will be considered.

Students are encouraged to use **python** (using *scikit-learn*), but can use **R** or any other language. Support for and code examples in python are available.

Data Profiling

In terms *of data profiling,* each variable shall be described against time, given particular attention to the granularity perspective.

Remember that data profiling is used as a mean to best understand the data and mostly for identifying the required transformations to apply to the original data, in the following step. These transformations aim to improve the performance of mining techniques, to be applied during the modeling phase.

In particular, students should perform a statistical analysis of the dataset in advance and summarize relevant implications in the report.

Data Preparation

At this stage, time series should be transformed in accordance to the properties of the original dataset, identified during the previous step. In this manner, students are asked to apply the usual preparation tasks, explaining their expected results.

Whenever students choose to not apply one of the studied tasks over the data, that option has to be justified based on the data characteristics. If there is no suspicion that one of the task approaches is more appropriate than another, both should be applied and the results obtained evaluated.

Are of particular importance the imputation *of* ***missing values*** and ***type transformation*****,** since the *sci-kit learn* methods' implementation do not deal neither with missing values nor with symbolic variables. Symbolic variables have to be transformed to numerical ones or discarded. Note that in the temporal context, missing value imputation may be approached through more informed techniques, such as the 'most probable value' since use can use time as the sorting variable.

The aggregation regarding the best granularity identified, and the differentiation of the different time series are expected to contribute significantly to results improvement

In all cases, the application of each one of the preparation techniques should be assessed. This evaluation should be made by comparing the modeling results before and after the application of each technique, verifying the impact of each one on the final results.

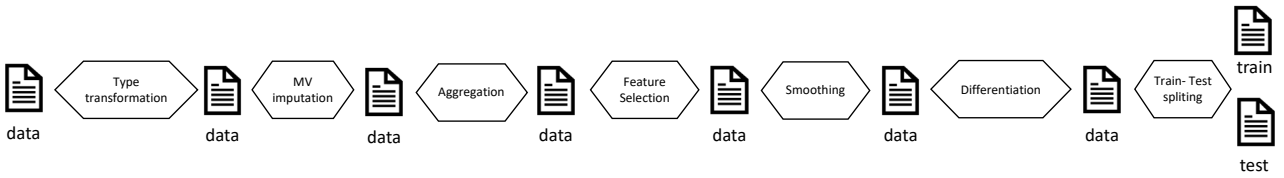
Suggested methodology

Choosing the best set of transformations to apply to the data, before moving on to the modeling phase, is not trivial. This choice implies the experimentation of each preparation task over the data, followed by the recognition of the improvements on the results obtained. However, the different preparation tasks, and their alternatives, do not have the same impact on all modelling techniques, and therefore their choice has to be careful.

With infinite time and resources, we could try out all the combinations among preparation tasks and modeling techniques, and find the best models from them.

However, there would be a combinatory explosion, which would disturb our understanding of the phenomena and the impact of each of the preparation tasks per se. It is therefore suggested, students follow a simplification of the process, evaluating the impact of each preparation task separately.

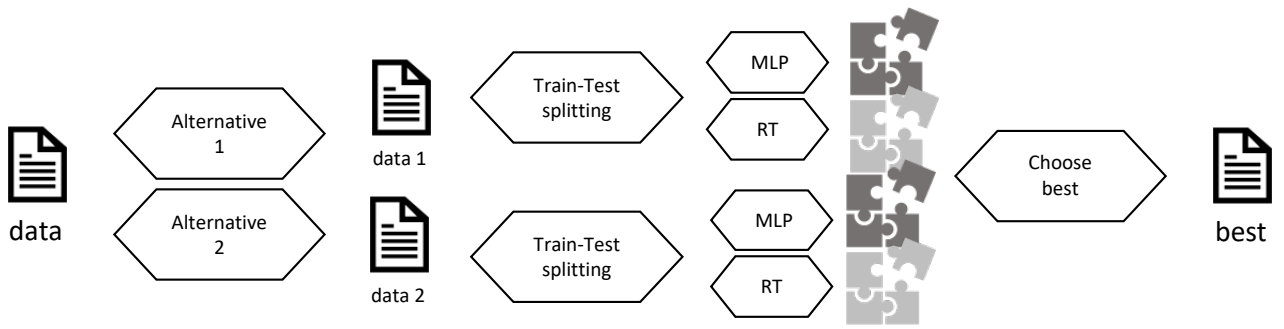
The following figure illustrates the proposed process.



The application of the preparation tasks should follow the order illustrated on the figure, carefully chosen to avoid duplication of efforts. Note that `sklearn` doesn't allow for training models over data with *missing values* or *symbolic variables*, and so, the solution of those problems has to be the first to be applied.

In the majority of situations, each preparation task has diverse parametrizations, but each one may have different impact on the modeling techniques' results.

The proposal here is to process each preparation task and then assess the impact of each alternative. Such impact has to be measured over the models trained over the datasets resulting from those alternative transformations.



The figure above illustrates the application of a generic transformation with two alternative techniques. And it works as follows:

First, we apply the different two alternatives to a single dataset, generating two new datasets. Then, we apply a train-test split and train new models from each alternative dataset. We suggest the use of both Regression Trees and MLP to train these models, resulting in four different models. Our choice is supported by the difference on the nature of the approaches , which limit the chances of choosing an approach best suited for a particular modeling technique.

After training the different models, we chose the preparation technique that presents the best improvement when compared with the previous dataset. In this manner, after the training we may face 4 possibilities:

- none of the alternative of preparation task applied improve the results: so, we should keep the previous dataset and proceed for the next step;
- one of the alternatives lead to the training of better models using both modeling techniques: so, we chose the dataset resulting from this transformation to proceed for the next step;
- the alternative supporting the improvement is different for each modeling technique: so, it is necessary to evaluate in which of the models the improvement is higher, and choose the approach responsible for that increase;
- the improvements are residual: so, it is our choice to continue with the previous dataset, or to follow with the preparation technique that theoretically should present higher improvements.

Remember that you should only consider applying the technique, i.e. using one of the resulting files, if in fact there is an improvement in the performance of the models when compared to the performance in the original dataset.

There are two exceptions to this rule:

- the imputation of *missing values*, since sklearn does not allow the application of training algorithms in data with missing *values;*
- and the separation in train and test, in order to warrant an unbiased and independent evaluation of models' performance.

Another important aspect, is that each technique only applies to solve a specific situation. It makes no sense to impute *missing values,* in datasets without *missing values*, for example.

A word about ***feature selection***. Given the different impact on the different algorithms studied, it shall be measured for each of them individually, and therefore shall be studied as one of the key factors for models performance.

Modeling

The project just considers the forecasting task. Remember that the goal is not just to describe the best models learnt, but to understand the impact of the available options on the produced models performance.

The forecasting task has to explore the application of *Regression Trees, Multi-layer perceptrons, Random Forests, GradientBoosting, ARIMA and LSTMs*. For this purpose, the use of a target variable is mandatory. Evaluation of the obtained models should be done as usual, through confidence measures and evaluation charts. A thorough comparison of the adequacy of the models should be presented taking into consideration the adequacy of their behavior against the properties of each dataset and their observed performance.

For this purpose the analysis of each classification technique should be done at three different levels:

- the analysis of the impact of the different parameters on models performance;

- the description of the best model found for each forecasting technique;
- the comparison of different best models, explaining the different achievements with the different techniques.

REPORT

The report file should be named **`report_X.pdf`** (replacing X by the team number) and submitted through Moodle. It should follow the template, without changing the margins and fonts used, and should have at most **10 pages**. Each additional page with analysis won't be considered, but an appendix for data profiling charts is allowed.

The report may be written in Portuguese or English. It should describe all the experiments made over the data, from their profile to the discovered models. Beside the placed choices, preparation performed, applied parameterizations and found models for each dataset, their interpretation and critical analysis are mandatory.

Delivery

The project has to be delivered through Moodle system. Only one report per team has to be submitted.

The submission deadline is the one stated on the first page at 23:59.

EXCELLENCE

A project that applies the suggested data mining techniques over the given dataset and provides a clear and *sound analysis of the collected results is not necessarily an excelling project*.

Excelling projects have three major characteristics.

First, they show an acute understanding of the data characteristics and their impact on the discovery, formulating hypothesis to explain differences in performance.

Second, robust assessments go beyond simple performance indicators, studying different and adequate parameters, and deriving trends from the experiments.

Third, poor results are not acceptable, and there is always something that we can learn from the data.

Plagiarism

Plagiarism is an act of fraud. We will apply state-of-the-art software to detect plagiarism. Students involved in projects with evidence of plagiarism will be reported to the IST pedagogical council in accordance with IST regulations.

EVALUATION CRITERIA

The project will be evaluated as a *whole*. Nevertheless, we provide below a decomposition of the total project score for the purpose of guidance and prioritization:

1. **Data profiling** (**10%**)
2. **Data preparation** (**10%**)
3. **Classification**
    a. Regression Trees (**5%**)
    b. Random Forests (**10%**)
    c. Gradient Boosting (10%)
    d. Regression models (**10%**)
    e. Multilayer perceptrons (10%)
    f. LSTMs (**10%**)
4. **Evaluation and critical analysis (25%)**

Good Work ! ! !