Class projet
Forecasting the energy performance of buildings

INSA Toulouse,
UF ‘Machine Learning’

September 26, 2020

The problem considered is to forecast the energy performance of buildings with statistical and machine learning techniques, based on 780 simulated data.

Project organization and deliverables.

The project has to be done by groups of 4. Deadline: December 2, 2020.
As deliverables, two Jupyter notebooks are expected, one in R, the other in Python. More over, one of them must be a technical report: it must include an interpretation of the results, an introduction, a conclusion, etc. It should be provided in pdf format.

Dataset.

The given dataset has been adapted from simulated data, available on the UCI website `http://archive.ics.uci.edu/ml/datasets/Energy+efficiency`. Indeed, for the sake of realism, we have added noise for the continuous variables. Furthermore, for simplicity, we have created a single output variable by adding the two seasonal ‘load’ variables. This output variable, called ‘Energy’, quantifies the building performance. In addition, we have created a qualitative variable ‘Energy efficiency’ with levels ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, obtained by slicing the ‘Energy’ variable with the thresholds: 30, 35, 45, 55, 65, 75.
The input variables are: relative compactness, surface area, wall area, roof area, overall height, orientation (north, east, south, west), glazing area, glazing area distribution. This latter variable has six levels: uniform (25% each side), 55% north (and 15% for the others), 55% east (and 15% for the others), 55% south (and 15% for the others), 55% west (and 15% for the others) and no glazing.

Problem goal.

We consider here the classification problem: to predict the energy efficiency.

... / ...

Questions.

Data analysis.

The aim of the section is to control and understand the data, which is a useful preliminary step. The questions below are the basics that you should do. Feel free to complete them with your own ideas.

1. Start with some unidimensional descriptive statistics of the dataset. Can you see anomalies?
2. Continue with a multidimensional descriptive analysis. In particular, using visualization techniques (e.g. scatterplot, conditional plot), which variable(s) seem to be the most influential on the output? Can you see interactions?
3. Consider the quantitative variables, except the Energy one. Do a principal component analysis. Can you see clusters? Do they correspond to the energy classes?
4. Still about the quantitative variables, use a clustering technique. Conclusion?

Models.

Now we consider the prediction problem with a machine learning point of view, i.e. by focusing on the model performance. What best performance can we expect? Below some guiding questions.

1. First of all, split the data into a training set and a test set. Why is this step necessary when we focus on performance?
2. Here, we consider the classification problem directly. Compare the performance of a linear model (logistic regression) with/without penalization, an optimal tree, random forest, boosting and SVM. Justify your choice (e.g. kernel for SVM), and tune carefully the parameters. Interpret the results and quantify the improvement brought by non-linear models.
3. Now, we first consider the regression problem and then classify using the given thresholds. Same question as before.
4. What approach is the best to predict energy classes: direct classification or regression+thresholding?
5. Interpretation and come-back to data analysis. Are your results consistent with the preliminary data analysis, e.g. about non-linearities, influence of variables (or variable importance)?