Class projet
Forecasting the energy performance of buildings

INSA Toulouse, ModIA
UFs 'Analyse de données' & 'Eléments de modélisation statistique'
Olivier Roustant & Cathy Maugis-Rabusseau

September 14, 2020

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

Project organisation and deliverables.

The project has to be done by groups of 3. Each group must include students both from INSA and ENSEEIHT. There are 15 hours scheduled during the class.
Three deliverables are expected: technical report + R code + Python code. We encourage you to write a R Markdown and a Jupyter notebook, and provide the pdf from one of these files. The report must include an interpretation of the results, an introduction, a conclusion, etc.

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 must do. Feel free to complete them with your expertise.

1. Start with some descriptive statistics of the dataset.
2. Using visualization techniques, what variables seem to be the most influential on the output? Can you detect interactions?
3. Before using the proposed classes, can you see clusters in the data? Use a clustering technique to justify your answer.

Linear models.

In modeling, always start by simple models. This section is devoted to the huge class of linear models. Will you find a 'suitable' one? Below some questions to guide you.

1. Study a suitable model to linearly explain the variable 'load' as a function of the quantitative variables of the dataset.
2. Study a suitable model to linearly explain the variable 'load' as a function of the other variables of the dataset.
3. Let 'Energy efficiency bis' be the binary variable defined by 1 if 'Energy efficiency' belongs to $\{"A", "B"\}$ and 0 otherwise. Study a suitable a suitable generalized linear model to explain the variable 'Energy efficiency bis' as a function of the other variables.
4. Study a suitable generalized linear model to explain the variable 'Energy efficiency' as a function of the other variables.

Non-linear models.

In this section, we consider the prediction problem with a machine learning point of view, i.e. by focusing on the model performance. As a first approach, we only consider random forest models in addition to linear models. 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 the logistic regression, an optimal tree, and a (tuned) random forest. 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 thresholded regression?
5. Interpretation and come-back to data analysis. Are you results consistent with the preliminary data analysis, e.g. about non-linearities, influence of variables (or variable importance)?