BOCCONI UNIVERSITY

Prediction of Energy Consumptions

20596 – MACHINE LEARNING. Final Data Challenge [Graded]
Prof. Daniele Durante

May 5, 2019

PROBLEM DESCRIPTION

The dataset consists of information from 14810 **customers of Eni Gas e Luce**. For 10000 of these customers you known their energy consumption [in kilowatt hour] on January 12, 2019 [denoted as `y`], and you posses additional input variables describing several features of each customer and his/her daily consumption behavior in 2018. For the other 4810 customers, you have information on the inputs, but you do not have access to their energy consumption on January 12, 2019.

Your goal is to predict `y` for the held–out 4810 customers.

There are **378 input variables**, which are described below.

- **Variables for the customer's daily energy consumptions in 2018. In particular, for each day** `d` $= 1,2,\ldots,$ **and month** `m` $=$ `jan,feb,...,dec` **the following variable is available**
  - `Xdm2018`: energy consumption [in kilowatt hour] on day `d` of month `m`

  A total of 365 variables are available on the energy consumption — one for each day in 2018.
- **Characteristic variables of the customer and of the company contract**.
  - `TYPE_CUST`: type of power consumer [`RESIDENZIALI`-residential, `MICROBUSINESS`-microbusiness, `MULTISITO`-multisite, `CONDOMINIO`-condominium, `IMPRESE`-company]
  - `GENDER`: gender of the power consumer [`M`-male, `F`-female]
  - `AGE_CUST`: age of the power consumer
  - `BILLING`: billing frequency [`BIMESTRALE`-bimonthly, `MENSILE`-monthly]
  - `PAYMENT`: type of payment [`BOLLETTINO POSTALE`-postal payment slip, `DOMICILIAZIONE`- domiciliation payment, `BONIFICO`-wire transfer]
  - `TYPE_OFFER`: type of power offer [7 possible types]. Labels `Prezzo fisso`, `Prezzo indicizzato` and `Prezzo indicizzato arera` mean fixed price, indexed price and arera index price, respectively

- FAMILY_OFFER: family of the product offer [25 possible categories]. Labels Altro, Cambiocasa, Prezzo Certo, Raddoppio sicuro, scelta sicura, Sconto su regolata, Sottocontrollo, SuMisura, SuMisura con Profili, and tutela simile mean Other, ChangeHouse, Safe Price, Doubling guaranteed, safe choice, Settlement discount, Under control, Tailored, Tailored with Profiles and similar protection, respectively
- CAP, CITY, PROVINCE and REGION: postal code, city, province and region where the Point of Delivery [POD] is located
- ALTIMETRIC_ZONE: categorical variable representing the altitude zone of the corresponding municipality [Pianura-flat land, Collina Litoranea-coastal hill, Collina Interna-internal hill, Montagna Interna-internal mountain, Montagna Litoranea-coastal mountain]
- CLIMATE_ZONE: Categorical variable representing the climatic zone of the corresponding municipality

A total of 13 variables are available on the characteristics of each customer and of the company contract.

NOTE: The dataset has missing values denoted with NA, and some of the categorical inputs have rare categories which may appear in the training set but not in the test set [or viceversa]. You are free to deal with these issues as you prefer [e.g. *imputing missing data in a reasonable way, collapsing rare categories with the most occurring ones, excluding variables with many missing values or rare categories from the analysis, etc* ...].