LAB OF INFORMATION SYSTEMS AND ANALYTICS

# Energy Utilization Prediction

## FROM ENVIRONMENT PHYSICAL PROPERTIES DATA


Università Ca'Foscari Venezia

H-FARM®

Brocco Mattia & Piccolo Giulio

# Introduction

In the age of smart homes, the ability to predict energy consumption not only may produce a saving, but may even result in more money generation by giving excess energy back to Grid (e.g. in the case of solar panels usage); it can furthermore be crucial to detect abnormal energy use patterns, to be part of an energy management system for load control and to model predictive control applications where the loads are needed. The aim of this study is to predict Appliance energy usage based on data collected from different sensors through regression analysis.

The subject of the project comes from dual sources, the former relies in the interest for the matter, i.e. the possibility to compute and predict the energy consumption (in Watthour) of an house that per se is low-energy, once data on physical parameters from both within and outside the house itself were gathered; the latter source arose at the time when the previous project on the "Prediction of house prices from Airbnb and road accidents data, New York 2017" was abandoned – after a massive work of data cleaning on three datasets and data collection on NY neighborhoods and their coordinates – due to high inconsistency between the results of the models and the relative performance.

Four statistical regression models were trained with train/test split and measured in terms of out-of-sample accuracy: (a) multiple linear regression (b) polynomial regression (c) decision tree regressor and (d) random forest regressor. The best model (DTR), in accordance with the benchmark measure of accuracy, was able to exceed the threshold of ninety percent.

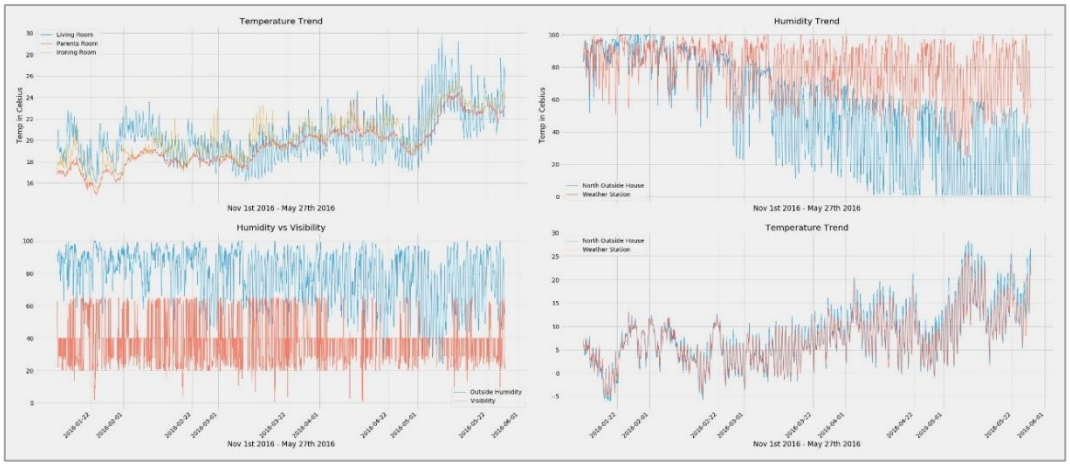
# Exploratory analysis

At the first glimpse the data gathered from the UCI Repository website seemed to reflect an astonishing quality of data collection, therefore data cleansing activities occupied the smaller share of the whole amount of time spent on the project. The dataset is built in such a way that each observation measures the power at a 10-minute span rate. The variables involved in the prediction are summarized below.

| Variable | Description |
|---|---|
| Appliance (Wh) | The target variable of the concerned study |
| Temperature (°C) from sources within the house | A total of nine variables from the same amount of rooms within the house |
| Humidity (%) from sources within the house | A total of nine variables from the same amount of rooms within the house |
| Other atmospheric properties recorded from the Chievres weather station | Temperature (°C), Pressure (mmHg), Humidity (%), Windspeed (m/s), Visibility (km), Dew Point (°C) |
| Two identical random variables | |
| Date (yyyy-mm-dd hhr:mm:ss) | |

The purpose of the project did not include investigation of the behavioral consumption during a timespan thus, the feature "date" has been neglected. The analysis of the dataset has risen to the surface that the column "light" was made of zeros for more than 60% of its length and, since it was another measure of power, it was not considered during the study.

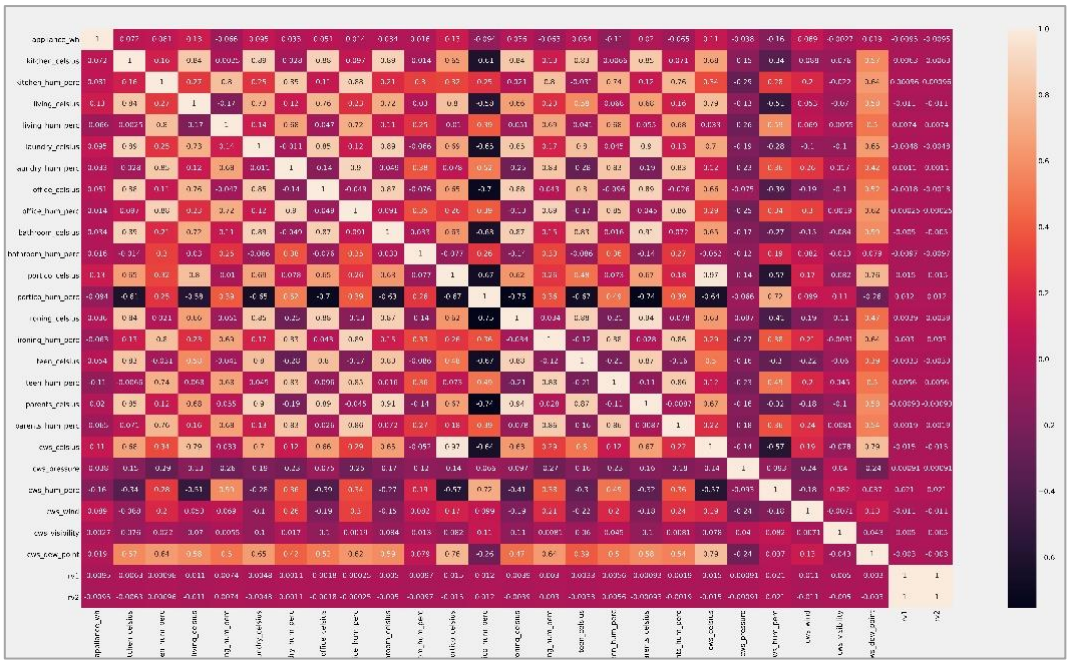
At a second place a short analysis on how variables could be related was carried out, and the result are presented through these four visual comparisons. The first plot emphasizes the variability of the temperature in the living room, while the last the similarity between the two measures examined. For what concerns humidity and visibility, the second and third graphs show how the measures involved seem to differ by a constant.

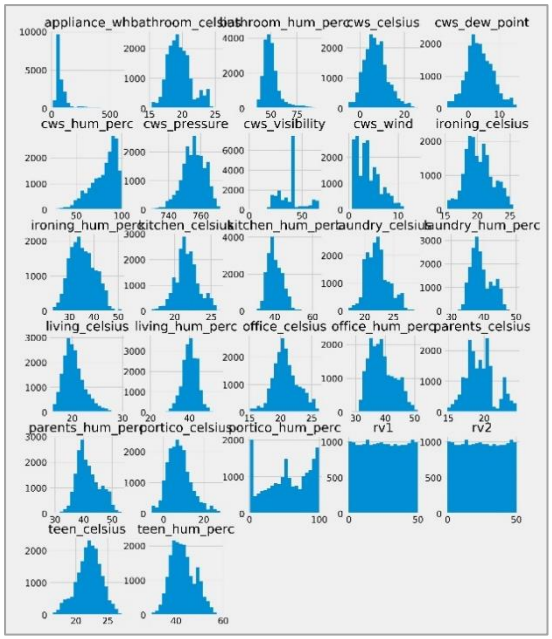



## Correlation matrix and distributions

A deeper understing of the correlation between variables is given by the coorelation matrix. The extent of correlation across the variables is fairly high. All Temperature's variable show weak positive correlation with the target variable, while all measures of temperature are quite storngly correlated each other (e.g. parental room temperature with laundty, bathroom, ironing room, teenager's room). The strongest correlation occurs between the values coming from the sensor placed outside the house and the one from the nearest weather station, that is quite a reasonable statement. For what concernes the measures of humidity sensors correlation idexes are slightly lower.

Random variables have, apparently, no role to play: matching their distribution and their correlation we can safely state that they are duplicated with shuffled entries.





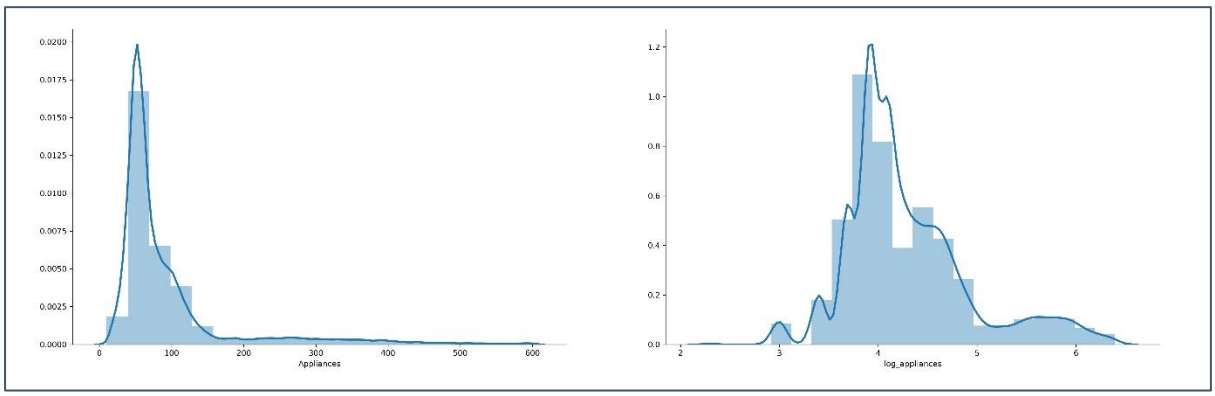

Further considerations were drawn over the distribution of the variables. The corresponding set of histograms reveals the extent to which many features, including the target variable, are characterized by distribution skewness.

In particular, the focus moved on the target variable, due to the negative effects that skewness may have on the results of the regressions. Therefore, for all the models applied in the study the target variable is meant to be considered as

$$target = ln\ (target)$$

On the other hand, for the concerns on the feature variables, only for the implementation of the feature selection through the Lasso Regression the matrix of features is preprocessed in order to make it resemble a normal distribution.

The element-wise standardization of the target variable according to its natural logarithm (on the right in the graph below) reflects in a significant reduction of the skewness, that eventually brings an increase in the accuracy.




## Prediction models

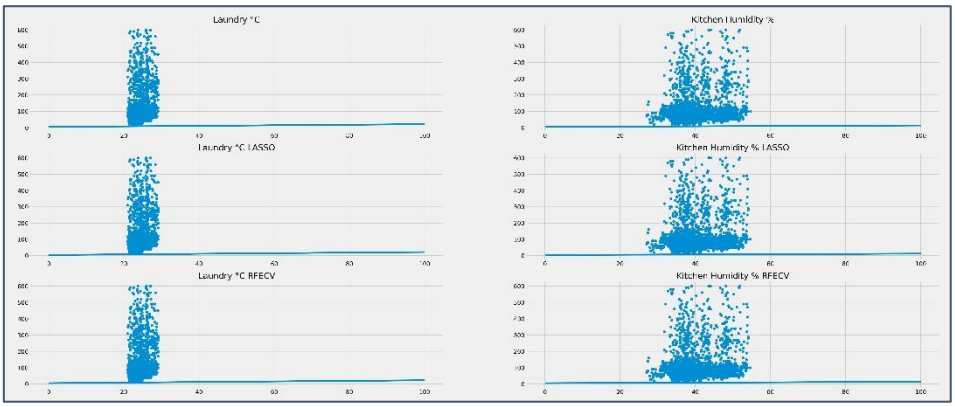
Throughout the study on the energy consumption several models have been scrutinized and evaluated in accordance with more than one metric; indeed, for the both the linear and polynomial models the $R^2$ score, Mean Squared Error and variance are provided.

| Model | $R^2$ score | MSE | Variance |
|---|---|---|---|
| Linear | 0.15096 | 0.24602 | 0.08766 |
| Linear (Lasso) | 0.10000 | 0.27402 | 0.13243 |
| Linear (RFECV) | 0.09841 | 0.2745 | 0.10624 |
| Polynomial (degree 2) | -0.98545 | 0.57532 | 0.60886 |
| SVR (degree 2) | 0.15006 | 0.24628 | 0.04959 |

Additionally, an SGDRegressor was implemented, but without significant results in terms of both explained variance and error in the prediction of the model.

The outcome obtained highlights how all linear models implemented perform almost at the same, unsatisfactory, level. The abovementioned results change – in particular, improve – only more than slightly by increasing the degree for both the polynomial regressor and the support vector regressor models. The "Variance" refers to the variance of the array of predicted values.

The addition of two methods for eliminating features - Lasso regression and the Recursive feature elimination that exploits cross validation – does not increase the coefficient of determination of the Linear Regression models. This means that the linear models are not capable enough to explain the variance in the features that is predictable from the target variable. Nevertheless, the output shows that the MSE of the models relies in the range $0.2 < MSE < 0.6$ where data on which they are tested has a mean of 4.23, that is an encouraging value in terms of the error the model is affected by.




The measure of accuracy to which each model refers to, and that is used to establish a comparable benchmark across the different regression models, is computed as:

$$Accuracy = 100 - MAPE = 100 - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

The investigated models present high accuracy according to this term, but it must be intended as a mean of comparison between different models. The choice fell on the mean absolute percentage errors due the contingency faced, where, by the definition of power, prediction cannot be negative, and from the exploitation of different measures combined (among which the MSE).

For what concerns the Decision tree Regressor, the main improvements came from the use of GridSearchCV as a mean of tuning parameters. Eventually, known both the variance and the MSE of the Tree, it was possible to identify the best ensemble method to improve the Tree itself. Since the implementation of tuning made the variance fall to zero, the only estimate that can be lowered was the bias (that is included in the MSE); therefore, the most suitable method consisted in Boosting the Tree. The outcome shows that the extent to which AdaBoosting lowered bias is so low that Bagging performed almost the same improvement.

The scale of comparison between different models increased drastically for Random Forests, due to the fact that the parameter tuning, and the exploitation of other Libraries of Python widened the range of aspects to reflect on.

A Random Forest was initiated, and then tuned twice through a randomized search with 2-fold cross validation and subsequently twice through grid search with 2-fold cross validation (that is consistently faster to be computed than the former one). Finally, an attempt on GBM Regressor was deployed and tuned through a Bayesian optimization with 5-fold cross validation that improved the abovementioned regressor.

| Model | MSE | Variance | Accuracy |
|---|---|---|---|
| Decision Tree | 0.82111 | 0.43254 | 82.75 % |
| Decision Tree (Grid Search) | 0.29281 | 0.00000 | 91.48 % |
| Decision Tree (Boosting) | 0.29255 | 0. 00000 | 91.46 % |
| Decision Tree (Bagging) | 0.29263 | 0. 00000 | 91.46 % |
| Random Forest | 0.48287 | 0.12586 | 84.60 % |
| Random Forest (Randomized Search) | 0.39272 | 0.09173 | 86.32 % |
| Random Forest (after Grid Search) | 0.38645 | 0.08877 | 86.34 % |
| LGBM Regressor | 0.33910 | 0.08278 | 87.96 % |
| LGBM Regressor (Bayesian Optimization) | 0.27598 | 0.02193 | 89.94 % |

# Problems

The main constraint related to the concerned project was time: indeed, a first intensive activity on data cleansing on the project mentioned in the introduction (to which there are references on other attached file) took away a lot of time and it was eventually discarded due to high ambiguity and inconsistency of the results obtained.

Taking inspiration from the Latin brocard "*utile per inutile non vitiatur*", which means that the useful is not affected by the useless, when the decision to head in a different direction by changing the topic of the study was taken, we delighted in applying one clustering technique (kNN) on geo-spatial data, up to obtaining the one hundred most significant clusters, starting from the coordinates of about seventy thousand Airbnb houses in NY City.

Since 90's the researchers have been running the never-ending marathon of "collecting enough data", however, today the problem could be the opposite one. The work on clustering exposed us to the problem according to which too many data, when not correctly processed, could swamp even best processors. Rendering on a WebMap millions of data points is not something trivial and we must carefully process, at our best, the data.

During the study on Energy Consumption, issues on understanding how to evaluate a regressor emerged, together with difficulties in gauging the parameters of the many regressors that were tested. In this sense, ambiguity on the property ".score" that regressors in scikit-learn own still fluctuates, and on the coefficients of determination accordingly.

Then, the parameter tuning on Random Forests occupied a significant share on the total amount of hours dedicated to this project due to high execution times (even by running chunks on Google Datalab). Finally, a disclaimer on results due to the high contextuality of the algorithms, that produce different results by changing the training sample.

Last, but not least, was the resolution of the problem related to the identification of a version-control system to cope with the parallel development of two projects simultaneously. Learning and using GIT and GitHub have been fundamental in helping us understand how important it is to use great precision in describing each "commit" and the push pull system.