Technical University of Denmark

42186 Model-based Machine Learning

Milestone 1

Edvard Foss, s191652
Jorge Montalvo Arvizu, s192184
Evangelos Stavropoulos, s193183

Spring 2020

Introduction

Energy load forecasting has always been a critical activity in the power sector. Given the physical constraints of the commodity, generation has to equal the demand at exactly the same time and adjust its production every time the load changes.[1] This motivates the grid operator to predict the necessary energy output of electricity generators to balance the grid and serve the customers every second of each day. Recently, load forecasting has been receiving major importance given the surge on utility-scale and distributed renewable technologies, i.e. electricity generated at the same spot where the load is consumed, further constraining the grid as transmission resources are allocated differently given the continuous fluctuation of renewable output.[2] Therefore, it is of upmost importance to anticipate the load consumed at every node of the grid and comply with the recent challenges of the power grid of the future. This project aims to develop a probabilistic model and attempt to predict the electricity load of 145 different buildings in the US from the ASHRAE's Great Energy Predictor III dataset.[3]

Research question

Given the motivation to forecast the electricity consumption, the proposed model will focus on households and aims to grasp the human behaviour behind the consumption of each household and its interaction with physical variables such as temperature and pressure. Therefore, the research question of this assignment is:

> *How accurately can we forecast a period of 24 hours for each household given measured past observations and external atmospheric variables?*

1 Data

The dataset "Great Energy Predictor III" includes one year of hourly measurements from 1448 buildings of different types. This assignment will focus on households, therefore only 145 houses will be used.[4] The raw variables are:

- **building_id** (categorical) - Foreign key for the building metadata.
- **meter** (numerical) - The meter id code. {0: electricity, 1: chilledwater, 2: steam, 3: hotwater}.
- **timestamp** (date) - When the measurement was taken
- **meter_reading** (numerical) - The target variable - electricity consumption in kWh
- **site_id** (categorical) - Foreign key for the weather station
- **primary_use** (categorical) - Indicator of the primary category of activities
- **square_feet** (numerical) - Gross floor area of the building
- **year_built** (categorical) - Year when the building was opened
- **floor_count** (numerical) - Number of floors of the building
- **air_temperature** (numerical) - Air temperature in degrees Celsius
- **cloud_coverage** (numerical) - Portion of the sky covered in clouds, in oktas
- **dew_temperature** (numerical) - Dew temperature in degrees Celsius
- **precip_depth_1_hr** (numerical) - Millimeters of precipitation depth per hour
- **sea_level_pressure** (numerical) - Pressure in millibar/hectopascals

[1] As of now, electricity cannot be stored economically (yet) and the bulk of electricity generation has to adjust instantly.
[2] Arriaga, Ignacio J. Regulation of the power sector. London New York: Springer, 2013. Print.
[3] ASHRAE - Great Energy Predictor III. How much energy will a building consume? at https://www.kaggle.com/c/ashrae-energy-prediction
[4] There are 147 building_id in the building_metadata file but two of them are not found in the metered data (772, 1397).

- **wind_direction** (numerical) - Compass direction (0-360°)
- **wind_speed** (numerical) - Wind speed in meters per second

After an initial inspection of the filtered dataset, it was decided to remove the following variables:

- **meter** - the dataset is filtered for electricity {0}
- **primary_use** - the dataset is filtered for 'Lodging/residential'
- **floor_count** - missing values for residential at 87.6% and the variable **square_feet** already represents the space
- **cloud_coverage** - missing for residential at 47.6% and it isn't representative to the model's story
- **wind_direction** - there's no data about the geometry and position of the building, so it doesn't fit into the model's story
- **precip_depth_1_hr** - missing for residential at 24.5% without information to impute or ignore

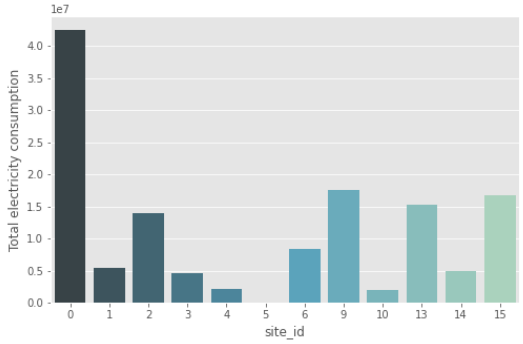
Table 1 and **Table 2** show the descriptive statistics of the raw variables included in the model without any transformation. The standard deviation of the target variable **meter_reading** is bigger than the mean, hence the distribution is very heavy tailed; this indicates a possible necessary transformation to 'help' the model by normalizing the values. Also, there is considerable missing data for variable **year_built**, at 43.4% of the whole dataset. Further statistics can be found in the extensive summaries on the notebook.

measure	building_id	meter_reading	site_id	square_feet	year_built
count	1229082.0	1229082.0	1229082.0	1229082.0	695561.0
mean	704.9	108.6	6.9	86018.8	nan
std	510.5	141.5	5.7	95665.8	nan
min	6.0	0.0	0.0	2000.0	1900.0
25%	134.0	21.0	1.0	37100.0	1956.0
50%	773.0	58.8	6.0	57334.0	1975.0
75%	1186.0	145.0	13.0	102774.0	2002.0
max	1447.0	12571.0	15.0	745671.0	2013.0

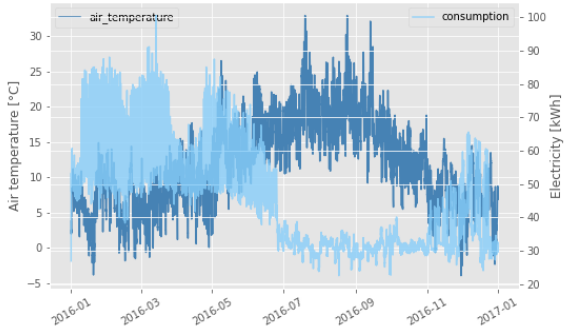
Table 1: Descriptive statistics (I/II)

measure	air_temperature	dew_temperature	sea_level_pressure	wind_speed
count	1219932.0	1219772.0	1183642.0	1216778.0
min	-28.9	-31.7	973.5	0.0
25%	10.0	1.7	1012.5	2.1
50%	18.0	10.6	1016.5	3.1
75%	24.4	17.8	1021.0	4.6
max	47.2	26.1	1046.0	18.5

Table 2: Descriptive statistics (II/II)



(a) Total energy consumption per site_id



(b) Electricity consumption and air temperature over the year for building ID 135

Figure 1: Total yearly energy consumption per site_id and yearly plot for building ID 135

Figure 1a shows the total yearly electricity consumption per **site_id**, the plot shows that the buildings at '0' consume in total more energy than most of the other sites. Particularly, site 5 is extremely low compared to the others, to the point that it isn't event plotted. On the other hand, **Figure 1b** shows the yearly behaviour of electricity consumption vs. air temperature over 2016. The relationship between these variables seems negative, as can be seen from the correlation **Figure 2** as well. This can be interpreted as when the temperature decreases, the need for electrical heating increases for that specific building. Notice that this was a chosen building at an specific site; other buildings at other sites may show a different behaviour (maybe even the opposite).

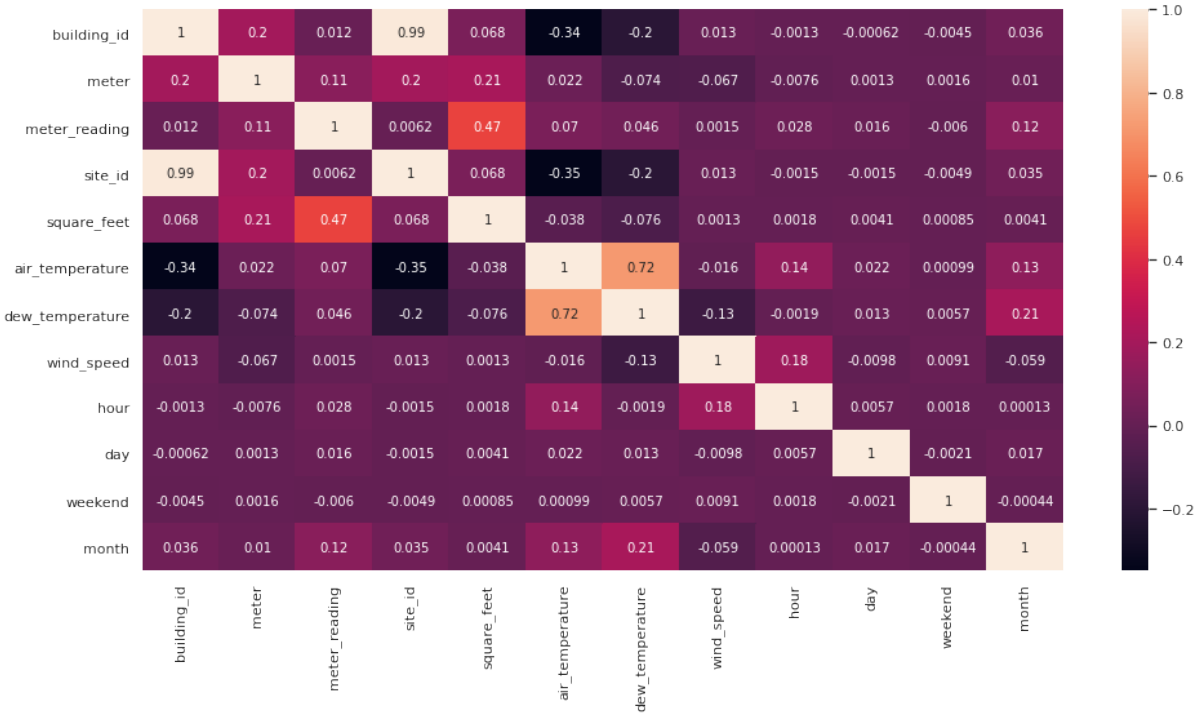


Figure 2: Ranked correlation matrix of selected variables

Probabilistic Graphical Model (PGM)

Considering that problem has both temporal, like weather and energy consumption and non temporal characteristics, like square feet and residential area, we intend to use a mixed model as our proposed solution. Specifically we will use a state-space model, together with a Bayesian Linear Regression model, to model both each dependency separately.

1.1 Bayesian Linear Model

The first model we construct is Bayesian Linear Regression model. Given our dataset consisting of meter reading values y and their corresponding building attributes $\mathbf{X}$, assuming Gaussian priors β and α we can write:

$$\beta \sim \mathcal{N}(\mu, \sigma) \tag{1}$$

$$\alpha \sim \mathcal{N}(\mu, \sigma) \tag{2}$$

$$y \sim \mathcal{N}(\alpha + \beta \cdot \mathbf{X}, \sigma) \tag{3}$$

Given these assumption we create our graphical model (figure 3), and the corresponding generative process:

1. Draw coefficients $\beta \sim \mathcal{N}(\beta \mid \mathbf{0}, \lambda\mathbf{I})$
2. For each feature vector $\mathbf{X\_n}$
   (a) Draw target $y_n \sim \mathcal{N}(y_n \mid \beta^T \cdot \mathbf{X_n}, \sigma^2)$

1.1.1 Factorized Joint Distribution

$$p(\alpha, \beta, \mathbf{y_n} \mid \mathbf{X_n}, \mu, \sigma) = p(\beta \mid \mu, \sigma)p(\alpha \mid \mu, \sigma) \prod_{n=1}^{N} p(\mathbf{y_n} \mid \alpha, \beta, \mathbf{X_n}, \sigma) \tag{4}$$

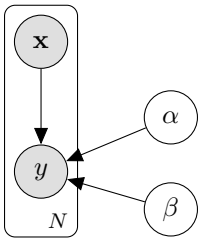


Figure 3: Bayesian Linear Model

1.2 Temporal Model

To model our temporal dependencies, we chose a State-Space Model, where the observations y_t are assumed to be generated by the hidden state h_t, where the hidden states are assumed to be first order Markovian variables (see fig. 4). Specifically:

$$\delta \sim \mathcal{N}(0, 1) \tag{5}$$

$$\mathbf{R} \sim LogNormal(0, 1) \tag{6}$$

$$h_t \sim \mathcal{N}(\delta \cdot h_{t-1}, \mathbf{R}) \tag{7}$$

$$\beta \sim \mathcal{N}(0, 1) \tag{8}$$

$$y_t \sim \mathcal{N}(\beta \cdot h_t, \sigma) \tag{9}$$

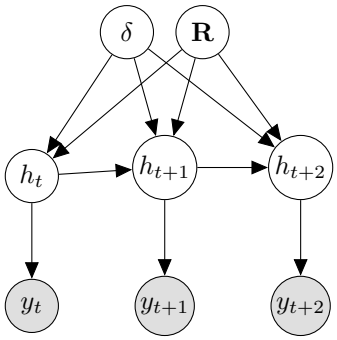


Figure 4: State-Space Model

1.2.1 Weather Dependencies

It is straight forward to extend the model to also the weather observation, by adding those inputs as observations $\mathbf{W}$ and the parameters η as latent variables(fig. 5). So by adding this dynamic eq. (7) becomes:

$$h_t \sim \mathcal{N}(\delta \cdot h_{t-1} + \eta \cdot \mathbf{w_t}, \mathbf{R}) \tag{10}$$

where $\mathbf{w_t}$ is a vector of weather features and η is sampled from a Normal distribution as $\eta \sim \mathcal{N}(m, \sigma)$.

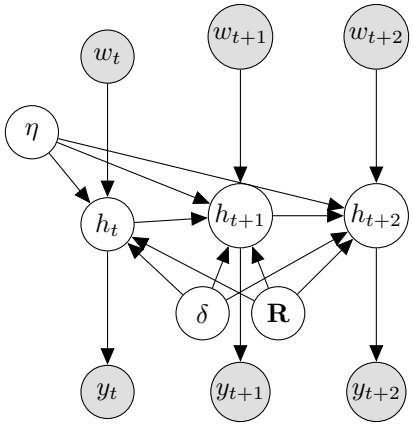


Figure 5: State-Space Model with weather Inputs

j

1.3 Mixed Model

For the mixed model we are going to use the bayesian linear model to estimate the parameters β in eq. (9). That leads us to:

$$\beta_{hh} \sim \mathcal{N}(\mu, \sigma) \tag{11}$$

$$\beta_{hi} \sim \mathcal{N}(\alpha_i + \beta_{hh} \cdot \mathbf{X}) \tag{12}$$

$$h_t \sim \mathcal{N}(\delta \cdot h_{t-1} + \eta \cdot W_t, \sigma) \tag{13}$$

$$y_{i,t} \sim \mathcal{N}(\beta_{hi} \cdot h_t, \sigma_i) \tag{14}$$

Another possible choice would be to model the bias instead of the weights. It is possible to extend the model to a hierachical structure to take into account different bias parameters for different residential areas.

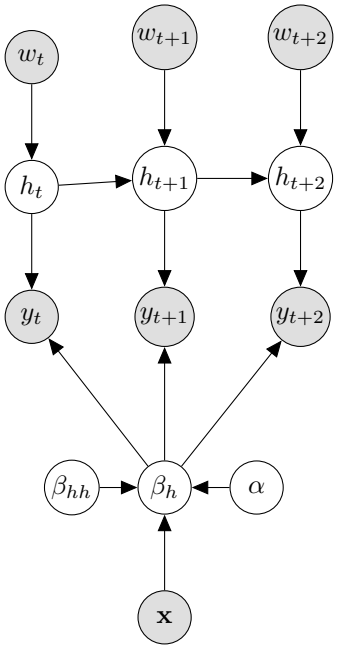


Figure 6: Mixed Model