Technical University of Denmark

42186 Model-based Machine Learning

Final Report

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

Spring 2020

For completeness, the first three pages of this report includes many repeated sections compared to Milestone 1. If in a hurry, skip to Results section at the end of page 3 to see the important content of this summary.

Introduction

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.[1] 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. The aim of this project is to develop a probabilistic model and attempt to predict the electricity load of households in the US.

Research question

Based on the motivation to forecast electricity consumption, the proposed model will only 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 wind speed. The research question of this project is:

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

Data

The dataset "Great Energy Predictor III"[2] includes one year of hourly measurements from 1448 buildings of different types. This project focuses on households.[3] When working with the dataset, it was discovered that the measurements were very odd for some buildings and the measurements weren't complete for all 145 houses, therefore we focused only on buildings with complete measurements (54 houses) to focus on the models and not on imputing missing measurements or atmospheric variables. The raw variables are listed in Appendix B - Milestone 1. For more information regarding the dataset go to Appendix B and/or EnergyPredictor.ipynb.

Models

As the problem has both temporal (weather and meter readings) and non-temporal attributes (e.g. square feet and residential area) we intend to use a bayesian linear regression approach and a LDS approach as well. We used a normal linear regression model as our baseline. Notice that the generative stories can be found in Appendix A.

Bayesian Linear Model

The probabilistic regression model investigated is a Bayesian Linear Regression model, this was used for both forecasting a single building and all the buildings with complete measurements (54 buildings). By labeling meter reading values as $\mathbf{y}$ and their corresponding building attributes as $\mathbf{X}$, and assuming Gaussian priors β and α. Given these assumptions, we create our graphical model (Figure 1) and the following joint distribution:

Factorized Joint Distribution

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

[1] Arriaga, Ignacio J. Regulation of the power sector. London New York: Springer, 2013. Print.

[2] ASHRAE - Great Energy Predictor III. How much energy will a building consume? at https://www.kaggle.com/c/ashrae-energy-prediction

[3] There are 147 building_id in the building_metadata file but two of them are not found in the metered data (772, 1397).

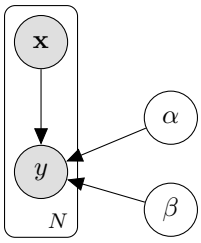


Figure 1: Bayesian Linear Model

Hierarchical Model

The hierarchical model, is based on the concept of hierarchical regression. The dataset consists of multiple observations from various sites. In comparison to a vanilla regression model, where the bias for each attribute is the same for all sites, in this model we allow different sites to have different biases. This model only works when forecasting all the buildings, not for a single building.

Hierarchical Model: Factorized Joint Distribution

$$p(\alpha, \beta, \mathbf{y_n} \mid \mathbf{X_n}) = p(\alpha_\mu)p(\alpha_\sigma)p(\beta)p(\sigma)(\prod_{site=1}^{S} p(\alpha \mid \alpha_\mu, \alpha_\sigma)) \prod_{n=1}^{N} p(\mathbf{y} \mid \alpha_{site} + \beta^T \cdot X_n, \sigma) \tag{2}$$

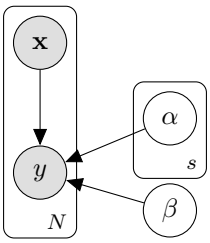


Figure 2: Hierarchical Model. Hyper priors α_μ and α_σ not in the figure.

Temporal Models - LDS-AR and LDS-ARx

The temporal models (linear dynamical systems) investigated are a variety of linear dynamic auto regressive models. Including lag 1-168, and with/without the weather data as an input. Notice that for LDS-AR24 and LDS-AR168 with and without weather inputs, only beta parameters 1, 24, and 168 are used, not the whole lags between $t-1$ and $t-24$ or $t-168$, respectively. More information can be found in the notebook, these models are used for single building forecast, not for all buildings at once.

LDS-AR1: General Factorized Joint Distribution & PGM

Example for AR-1:

$$p(y_t \mid y_{t-1}, \beta, \sigma) = p(\beta) \prod_{t=1}^{T} p(y_t \mid \beta y_{t-1}, \sigma) \tag{3}$$

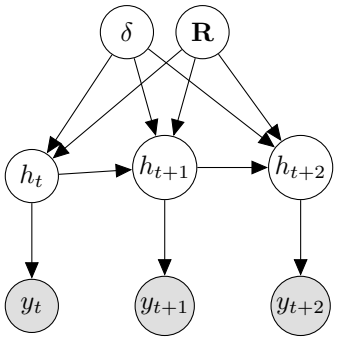


Figure 3: AR-1 Model

LDS-AR1 plus weather inputs

The temporal LDS-AR model was extended to include the weather observations by adding inputs as observations $\mathbf{W}$ to the model, and the parameters η as latent variables(fig. 4).

LDS-AR1 w/ Weather Dependencies: General Factorized Joint Distribution & PGM

Example for AR-1:

$$p(y_t \mid y_{t-1}, \beta, \sigma) = p(\beta) \prod_{t=1}^{T} p(y_t \mid \beta y_{t-1} + eta \cdot x_{t-1}, \sigma) \tag{4}$$

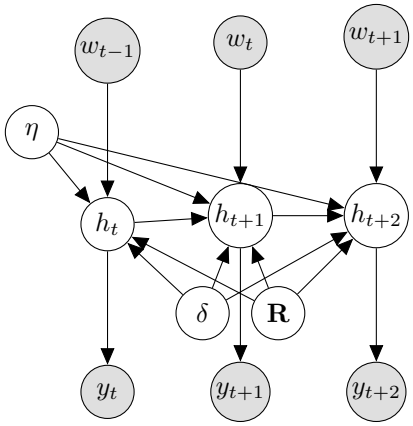


Figure 4: LDS-AR-1 model with weather dependencies

Results

Forecasting all buildings at once

From the results in Figure 5 and Table 1, it can be seen that the linear and hierarchical regression are the best performers when forecasting July 1, 2016 for all the buildings. The bayesian regression is a significant underperformer in this setting, since it forecasts negative values - this could be fixed by restricting the target variable to only be positive variables (maybe a truncated normal or half cauchy distribution) instead of using the normal distribution. Notice the first part of the plot shows zero values for many buildings, this was a hard test to our models to forecast, therefore we selected this date.

	CORR	MAE	RAE	RMSE	R2
Linear Regression	0.961	11.309	0.215	19.997	0.913
Bayesian LR	0.963	54.472	1.036	57.261	0.288
Hierarchical	0.959	11.907	0.227	20.016	0.913

Table 1: Error calculations for all regression models

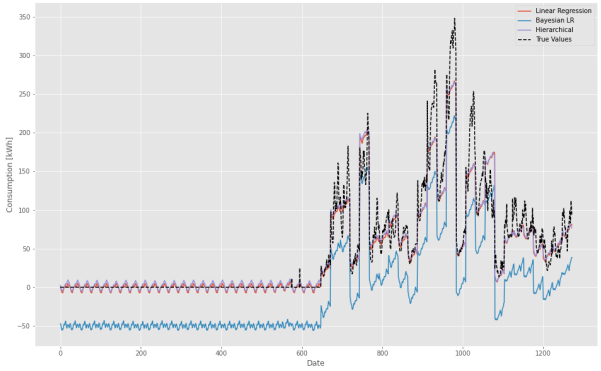


Figure 5: Comparison of the meter readings for the Linear, Bayesian and Hierarchical regression vs. true values. The x axis is a continuous axis of 24 hours for the 54 forecasted buildings, i.e. the first 24 ticks correspond to some building, the next 24 ticks to the next one, and so on.

Forecasting a single building

All models were executed to forecast 24 hours in February 1, 2016, and July 1, 2016. This was intended to test the performance of all models when training with 1 month of information vs. 6 months of information. The errors (MAE, RAE, RMSE and R2) values for all the models are shown in Table 2 and 3.

	CORR	MAE	RAE	RMSE	R2
Linear Regression	0.768	14.395	0.700	19.495	0.382
Bayesian LR	0.767	33.592	1.634	37.257	0
AR1	-0.384	26.374	1.283	29.761	0
AR2	-0.469	30.021	1.460	37.112	0
AR24	0.551	21.480	1.045	25.191	0
AR168	0.488	21.392	1.041	26.134	0
AR1x	0.039	28.007	1.362	31.647	0
AR2x	-0.174	61.612	2.997	75.397	0
AR24x	0.681	17.241	0.839	22.044	0.210
AR168x	0.814	10.713	0.521	14.576	0.654

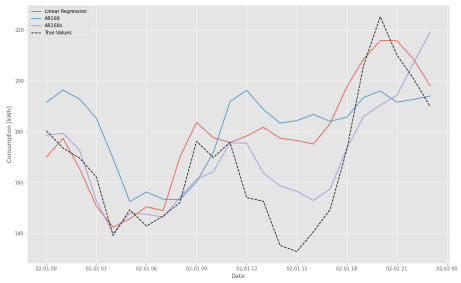
Table 2: Error calculations for all models with 1 month training data

From Table 2 it can be determined that the auto regressive model with lag 168 and including weather input (AR168x) is the superior model in regards to our error calculations when training on 1 month of data. The forecasts from the best performing models, in comparison with the true values, for 02/01/2020, with one month of training data, are shown in Figure 6a.

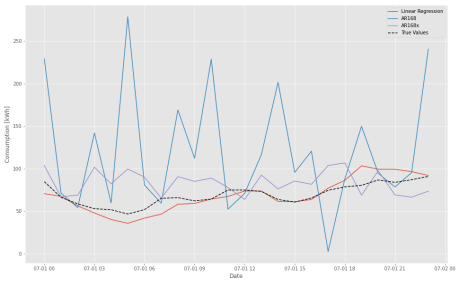
	CORR	MAE	RAE	RMSE	R2
Linear Regression	0.916	7.063	0.671	9.439	0.402
Bayesian LR	0.914	7.039	0.669	9.284	0.422
AR1	-0.050	48.700	4.628	60.300	0.000
AR2	0.173	62.648	5.953	83.470	0.000
AR24	0.176	88.016	8.364	97.638	0.000
AR168	-0.022	60.490	5.748	86.939	0.000
AR1x	-0.070	17.388	1.652	20.643	0.000
AR2x	0.330	15.227	1.447	18.773	0.000
AR24x	-0.014	14.618	1.389	19.242	0.000
AR168x	-0.155	20.441	1.942	24.220	0.000

Table 3: Error calculations for all models with 6 months training data

However, for 6 months of training data one can see from Table 3 that the error increases for the AR168x and the error from the regression models is significantly improved when forecasting day 07/01/2020. The reason for this should be further investigated, but it may be either a difficult day to forecast when depending on temporal information. The comparison of the different forecasted meter readings is shown in Figure 6b.



(a) Comparison of meter readings AR168x vs. true for 02/01/2020 and the best models after 1 month of training data.



(b) Comparison of meter readings AR168x vs. true for 07/01/2020 and the best models after 6 months of training data.

Figure 6: Single building forecast comparison with data availability 1 month vs. 6 months.

Conclusion

When forecasting all the buildings at once, the hierarchical model performs almost the same as the linear regression model; while the bayesian linear regression model forecasts negative values. This could be further investigated and attempt to fix. However, the time-dependency of energy consumption is "hard-coded" with the time-wise dummy variables. Given that the results of the LDS-AR168x model were very promising, it might be a better option to focus on improving the LDS-ARx models instead. When forecasting a single building, the LDS-AR168x outperforms the reference models and shows great potential at forecasting 24h based when trained with 1 month of data. However, the LDS-ARx models with 6 months of training data increase its error estimate compared to with 1 month of training. This might be because Notice that all LDS-AR models without weather inputs perform very badly compared to the linear models and LDS-AR models with inputs.

Future Work

Regarding the forecast of all buildings at once, an improvement to the hierarchical model might be to use also hyperpriors for each beta parameter of each house regarding weather variables. Increasing the pooling of the model and adjusting for each house. Regarding the forecast of a single building, the LDS models could further be improved by testing different combinations of lags and mixing it with a hierarchical model to include the hyperpriors per site and even per building type. Even though this project focused on only households, the dataset also includes other types of buildings which fits very well with a hierarchical modelling approach. Due to the limitations of time and the problem's complexity it was decided to leave the mixed model to future work, i.e. combining the temporal model with a hierarchical. Some problems arised when mixing numpyro and normal pyro, therefore we decided to focus on testing different simpler models instead of making a single very complex model.

Appendix

Appendix A: Generative processes and extra figures

Bayesian Linear Regression:

$$\beta \sim \mathcal{N}(\mu, \sigma) \tag{5}$$

$$\alpha \sim \mathcal{N}(\mu, \sigma) \tag{6}$$

$$y \sim \mathcal{N}(\alpha + \beta^{\mathbf{T}} \cdot \mathbf{X}, \sigma) \tag{7}$$

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

Hierarchical Model:

Given **site** number of sites we have:

$$\beta \sim \mathcal{N}(\mu, \sigma) \tag{8}$$

$$\alpha \sim \mathcal{N}(\mu \mid \text{site}, \sigma) \tag{9}$$

$$y \sim \mathcal{N}(\alpha + \beta \cdot \mathbf{X} \mid \text{site}, \sigma) \tag{10}$$

1. Draw coefficients $\beta \sim \mathcal{N}(\beta \mid \mathbf{0}, \lambda \mathbf{I})$
2. Draw coefficients $\alpha_\mu \sim \mathcal{N}(\alpha_\mu \mid \mathbf{0}, \lambda \mathbf{I})$
3. Draw coefficients $\alpha_\sigma \sim \mathit{HalfCauchy}(\alpha_\sigma \mid \mathbf{5})$
4. For each **site**
   (a) Draw bias $\alpha_{\text{site}} \sim \mathcal{N}(\alpha_\mu, \alpha_\sigma)$
5. For each feature vector **X_n**
   (a) Draw target $y_n \sim \mathcal{N}(y_n \mid \beta^T \cdot \mathbf{X_n}, \sigma^2)$

LDS-AR1 - extendable to AR2, AR24, and AR168

1. Draw coefficients $\delta \sim \mathcal{N}(\beta \mid \mathbf{0}, \lambda \mathbf{I})$
2. Draw coefficients $\mathbf{R} \sim \mathit{HalfCauchy}(\alpha\sigma \mid \mathbf{5})$
3. For each timestep $t \in 2, ..., T$
   (a) Draw $h_t \sim \mathcal{N}(h_t \mid \delta_1 y_{t-1}, R^2)$
4. For each timestep $t \in 1, ..., T$
   (a) Draw target $y_t \sim \mathcal{N}(y_t \mid \beta^T \cdot \mathbf{h_t}, \sigma^2)$

LDS-AR1 with inputs - extendable to AR2, AR24, and AR168

1. Draw coefficients $\delta \sim \mathcal{N}(\beta \mid \mathbf{0}, \lambda \mathbf{I})$
2. Draw coefficients $\mathbf{R} \sim \mathit{HalfCauchy}(\alpha\sigma \mid \mathbf{5})$
3. For each timestep $t \in 2, ..., T$
   (a) Draw $h_t \sim \mathcal{N}(h_t \mid \delta_1 y_{t-1} + \eta \mathbf{x}_{t-1}, R^2)$
4. For each timestep $t \in 1, ..., T$
   (a) Draw target $y_t \sim \mathcal{N}(y_t \mid \beta^T \cdot \mathbf{h_t}, \sigma^2)$

Appendix B: Data Section from Milestone 1

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

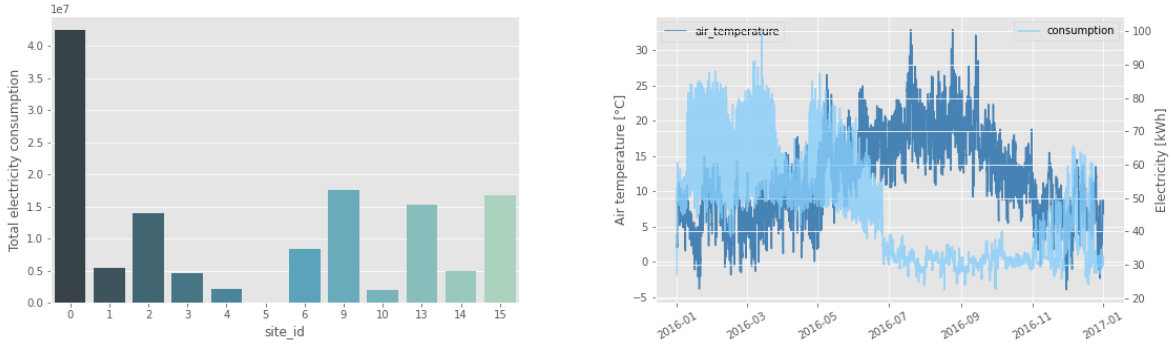
Table 1 and **Table 2** show the descriptive statistics of the raw variables included in the model without any transformation. The standard deviation of the target variable **meter_reading** is bigger than the mean, hence the distribution is very heavy tailed; this indicates a possible necessary transformation 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).

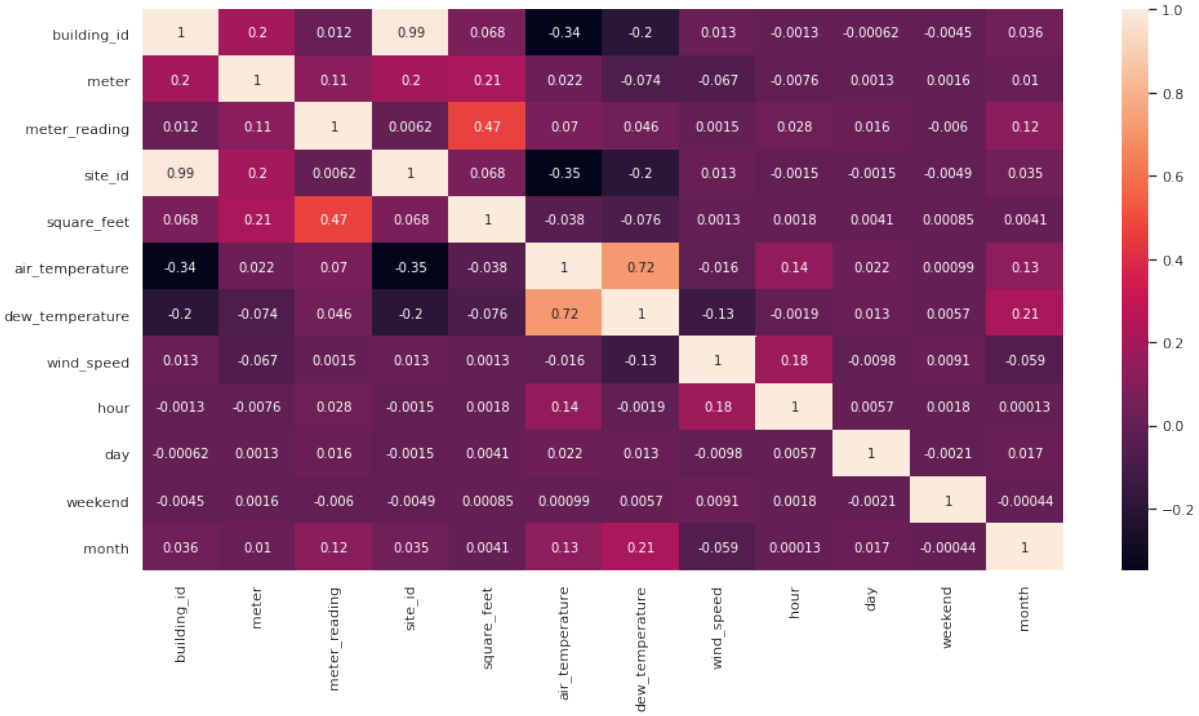


Figure 2: Ranked correlation matrix of selected variables