Anomaly detection in energy consumption of the buildings

Pavel Kuzmin
e-mail: pavloon@bk.ru

I. Problem description

Three buildings are given. Each building has a number of different measurements gathered with some time step and some data about the building. Information about weather (for building or some point not far from the building) is also presented. First of all the data was reformatted as a table like: Timestamp, measurement 1,measurement 2, temperature, is holiday,etc. In terms of mathematics the problem can be presented in the following way.

Energy consumption observed by the meter for fixed time step is presented. If cumulative energy consumption is presented (like for site “038” for meter id “38_9686”) it should be numerically differentiated. As an example site “334_61” is considered. The energy consumption for the site is shown at *fig.1*.

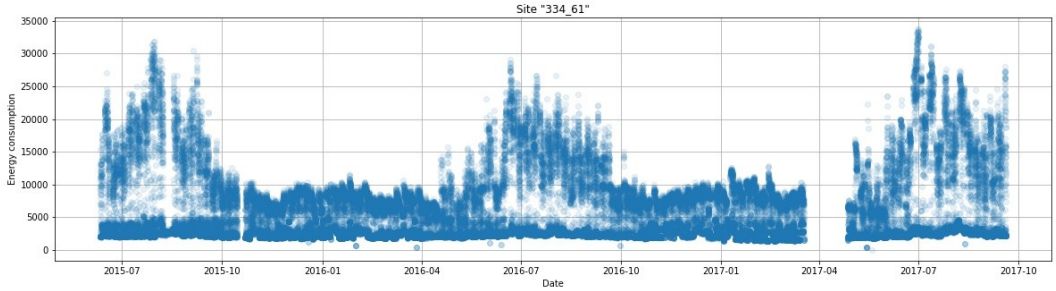


Fig 1.Energy consumption for site “334_61”

Energy consumption $\widetilde{W}$ can be presented as a sum of function of time t and several parameters $x_1, x_{2,\dots}, x_n$ and random noise ξ : $\widetilde{W} = f(t, x_1, x_{2,\dots}, x_n) + \xi$ (1)

If we can build this function and measure the noise, then any additive can be considered as an abnormal additive η : $\widetilde{W} = f(t, x_1, x_{2,\dots}, x_n) + \xi + \eta$ (2)

The following method is designed to find the abnormal energy consumption in this sense.

II. Solution

Key parameters

Several parameters were used as main parameters to find the estimation $\widehat{W}$ of normal energy consumption $f(t, x_1, x_{2,\dots}, x_n)$. These parameters are:

- temperature, t
- time of day(in hours), h
- label that average temperature for the day $\bar{t}$ is higher then some threshold . Looks like the threshold controls turning on some systems of the building,

- label of working day, w_{day}
- other parameters were used individually for small improvement in quality (year, month, day of week for site “038” and year, month, day_of_week, values of meters with ids 875, 896, 925 for site “234_203”).

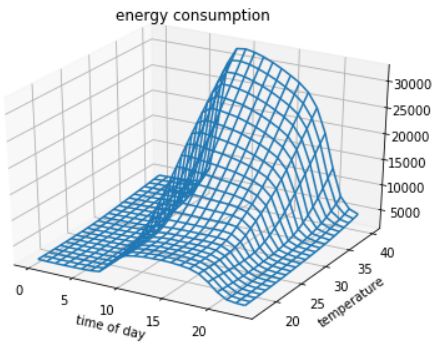
$\widehat{W}$ for each site was found independently for different values of parameters $\bar{t}$ and w_{day} :

1. $\bar{t} > 17^o, w_{day} = 1$
2. $\bar{t} < 17^o, w_{day} = 1$
3. $\bar{t} > 17^o, w_{day} = 0$
4. $\bar{t} < 17^o, w_{day} = 0$

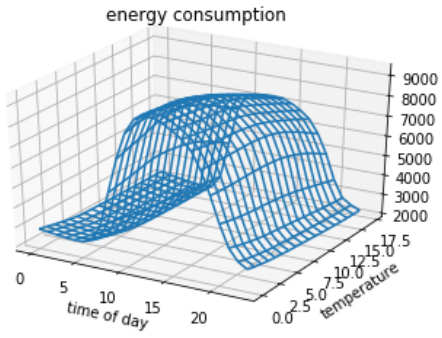
Model 1. Neural network

$\widehat{W}$ as a function of parameters for the site was found as feed forward neural network for each area of parameters independently (4 networks for each building). The neural networks had the following parameters: 5 hidden layers with ReLU activations and 1024 neurons each. The dropout layer with probability of dropout 0.5 was inserted after each hidden layer. Mean absolute error was used as a loss function, because it is known that there is a number of anomalies in the dataset. The network was optimized with adam optimizer for 5-10 epochs. After training $\widehat{W}$ was calculated with this network for the same points. High dropout was used to prevent overfitting.

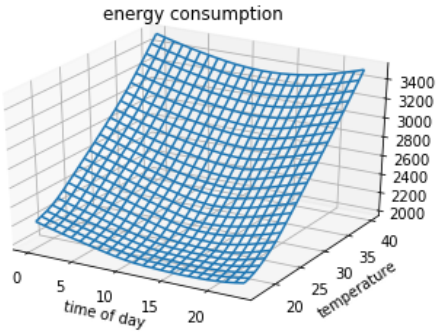
$\widehat{W}$ for the building “334_61” for four combinations of parameterss $\bar{t}$ and w_{day} is shown at *fig.2 a-d*.



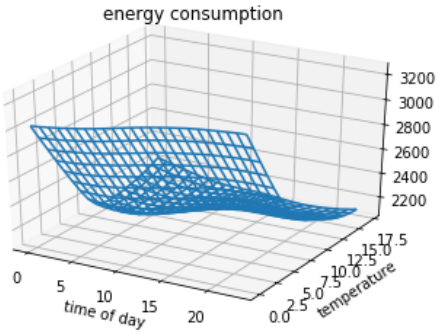
a) average day temperature >17 degrees, working day



b) average day temperature < 17 degrees, working day



c) average day temperature >17 degrees, weekend



d) average day temperature < 17 degrees, weekend

Fig 2. Surfaces of found normal energy consumption for site "334_61" for different values of average day temperature and for working days and weekends.

After getting $\widehat{W}$, the difference with observed consumption $\widetilde{W}$ can be calculated:

$$\Delta = \widehat{W} - \widetilde{W} \quad (3)$$

this difference is normalized to become metric in some sense:

$$\kappa = \left| \frac{\Delta - \bar{\Delta}}{\sigma_\Delta} \right| \quad (4)$$

where $\bar{\Delta}$ is the estimation of mean for Δ (just arithmetic average) and $\sigma_\Delta = \frac{1}{n-1} \Sigma (\Delta - \hat{\Delta})^2$ is the estimation of std for Δ .

If the number of abnormal points in the dataset is known and fixed, thresholds p for κ can be estimated to get the desired number of abnormal points. All points with $\kappa > p$ are marked as anomalies. Because the number of abnormal points is unknown in the competition and there may be abnormal points in another sense, different thresholds were used to get high score on leader board. A good starting point for finding the threshold is 3 (99.7% quantile of normal distribution). The resulting thresholds were 2.5 – 4.5 for different buildings and areas of parameters. The graph like one at *fig.3.* was used to select the value of threshold to get desired rate of abnormal points for each of 4 models for each building. The scatter plot of observed energy consumption and estimated normal energy consumption is shown at *fig. 4.*

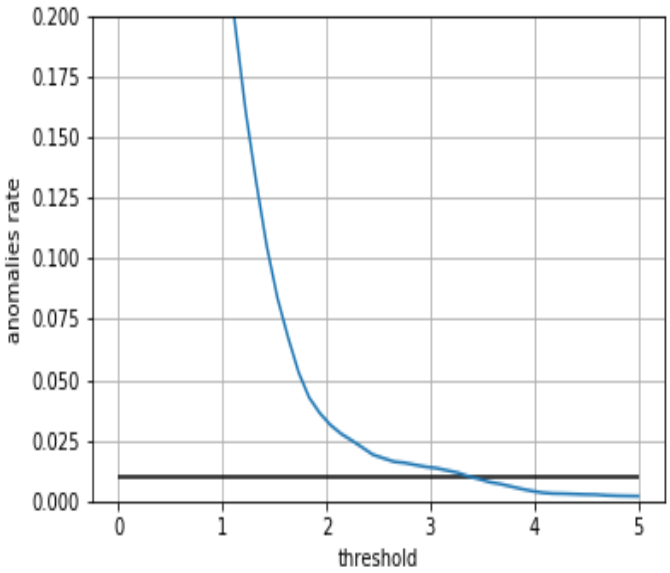


Fig 3. Anomalies rate versus threshold. Black line shows the rate of 1% anomalies in data. The intersection with the blue line shows the corresponding threshold.

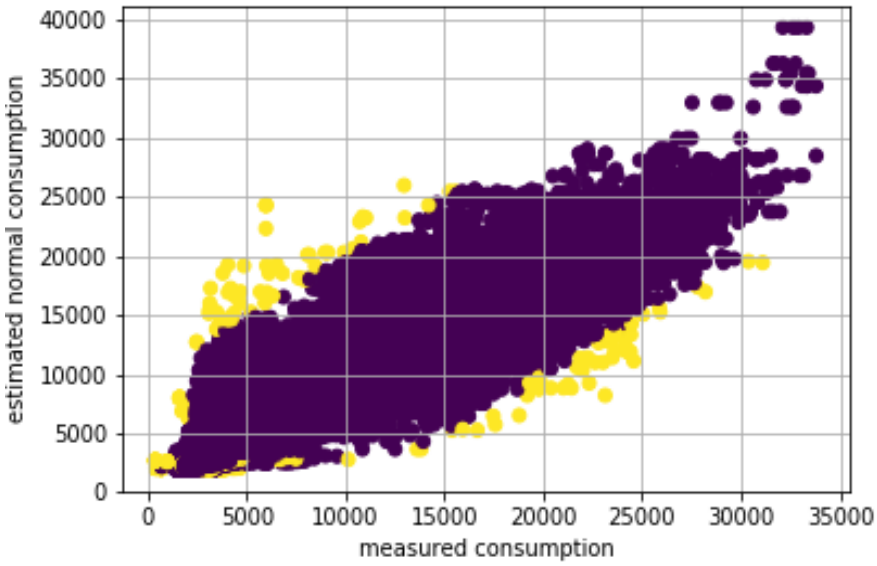


Fig 4. Scatter of observed energy consumption and estimated by model 1 normal value of energy consumption. Anomalies are shown with yellow points

Model 2. K nearest neighbors

KNN with Manhattan distance and 1000 nearest neighbors was used to gather another estimations of normal energy consumption. More features were used to get estimations with KNN. Features “mean temperature for day”, “mean temperature for night and daylight hours”, “day of year” were added. All other operations (calculating metric, getting threshold) were just the same as in the first model. The scatter plot of observed and estimated normal energy consumption is shown on *fig 5*. Abnormal points are marked with yellow.

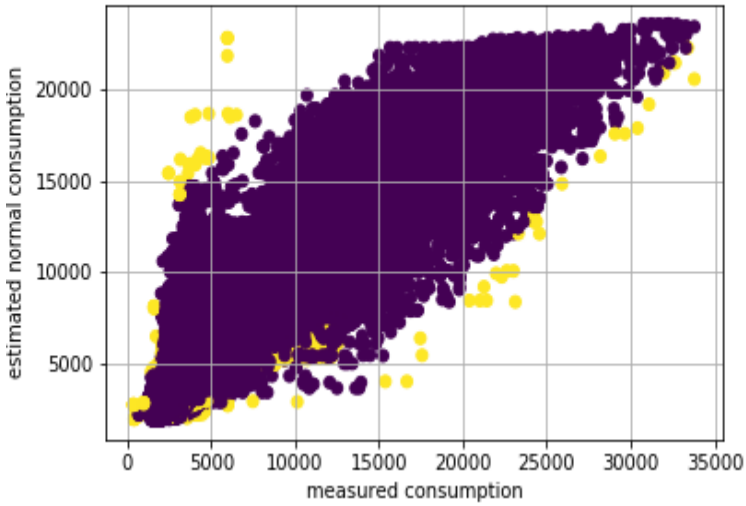


Fig 5. Scatter of observed energy consumption and estimated by model 2 normal value of energy consumption. Anomalies are shown with yellow points

Final labeling

The points marked with both model 1 and model 2 were finally labeled as anomalies. The energy consumption over time for the site “334_61” with marked anomalies by model 1, anomalies by model 2, and finally labeled anomalies are shown on fig.6.

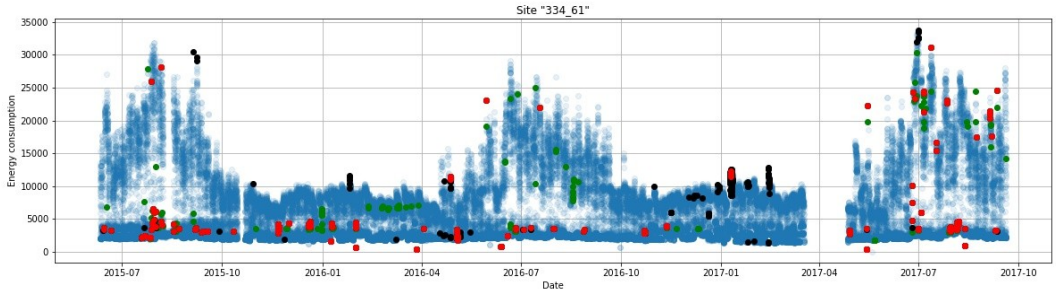


Fig 6. Energy consumption of site "334_61". Green points show outliers labeled with the help of neural net, black points - knn, red points - final labeling.

III. Types of anomalies found with the method

1. Extremely high/low energy consumption over short period of time. Can warn about the systems of the building which need more attention in next scheduled inspection of the systems. See Fig 7 a and b. Several rare spikes on the right part of the graphs are seen.

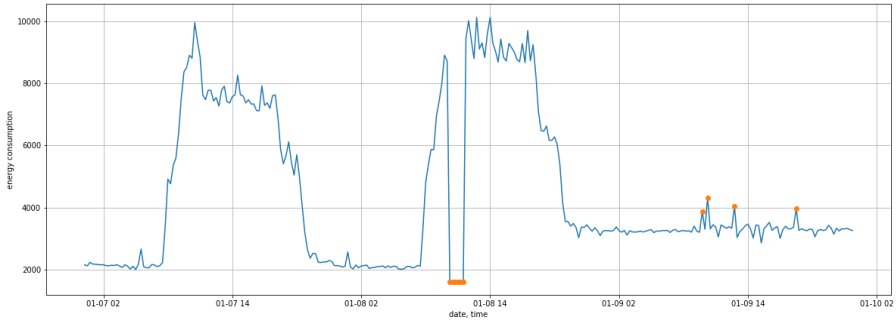


Fig 7a. Extremely low consumption and several smaller outliers

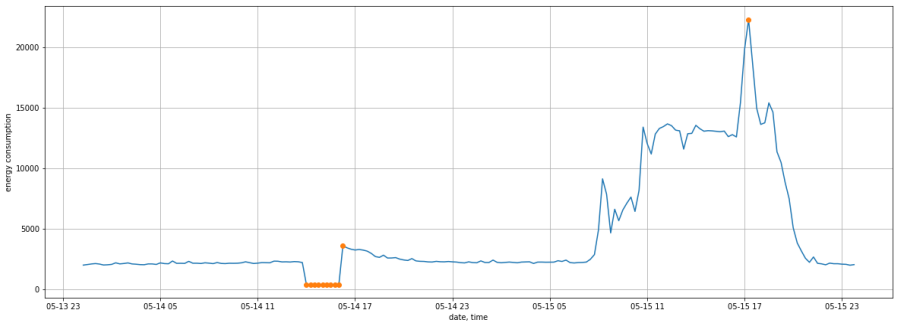


Fig 7b. Low consumption and single abnormal point

2. High/low energy capacity for the time of day. Can be helpful to find malfunction of equipment of the building. Like central conditioning system for example. See fig 8.

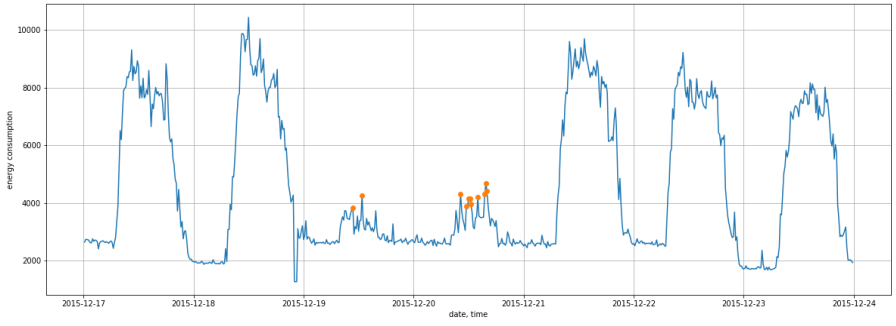


Fig. 8. High energy consumption for the day.

3. Extremely High/low energy capacity for the day of week. See extremely low consumption at fig 7 a,b and fig 9. Such patterns can warn about some critical malfunction of energy system of the building.

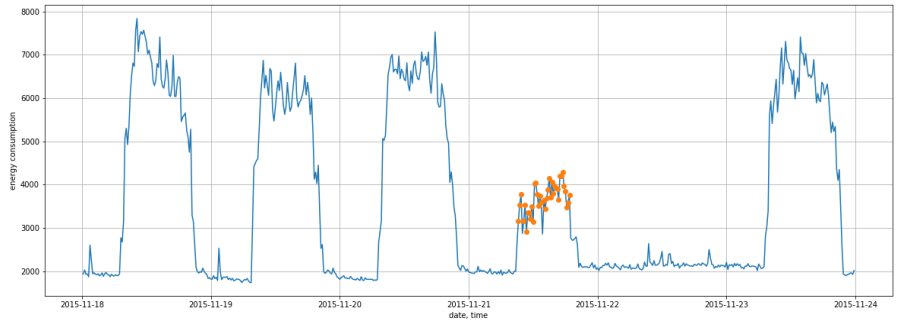


Fig. 9. Huge number of outliers for the day.

IV. Possible improvements

- More precise information about holidays can improve the models quality extremely.
- The information about the equipment operation modes can also help to improve the model. For example the temperature of turning the central conditioning system or heating systems on and off.
- Exploring the possible usage and the goals can help in feature engineering. For example understanding which type of abnormal behavior is really abnormal for our needs.
- One more model can be added to catch low and high frequency processes as long term and shot-term non-stationarity.
- Looking on the problem as time-series task and search not anomalies for all the dataset, but search outliers based on previous seen data only – marking only previously unseen data.

V. Scenarios of using in production

There are several possible scenarios to use this model in production.

- Automatic labeling of new data alike to the one in train set. After that overview by the assessors is needed.
- Can be used for remote real time control of correct operation of equipment in the building. Especially can be helpful if some time-series model is built and used along with present model.

Feel free to e-mail me with any questions and discussions.