Predicting energy consumption of a building based on historic usage rates and weather data

Tamara Šumarac
Department of Physics
Harvard University
Cambridge, MA 02138
tsumarac@g.harvard.edu

Abstract

One of the biggest problems facing the world today is global warming. With electricity and heat production being one of the biggest contributors to the greenhouse gases, there is a demand in constructing a more energy-efficient buildings and retrofitting the old ones. An estimate of how much energy would improved building consume relative to the old one can incentivize investors to pursue these renovations. For this purpose it is useful to have an accurate prediction of how much energy would a building consume in future given how much that same building was consuming in the past. In this project I will use the data [1] of historic usage rates and observed weather over 1 year period across hundreds of buildings in order to make energy consumption predictions. In this report I investigated several feature encodings in combination with several regression learning algorithms including linear regression, decision tree regression and random forest regression in order to obtain the best prediction of the data. I compared these results with the results one would obtain using automated regression model generated by autosklearn, in order to understand what kind of predictions a non-expert in the field would be able to make.

1 Introduction

Emission of greenhouse gasses into the atmosphere is known as the main cause of global warming. Through human activity, specifically burning fossil fuels, deforestation and farming, large amounts of these gasses are released into the atmosphere every year, causing our planet to warm up at a rapid speed. Energy generated by fossil fuels is mostly used for electricity and heating, and for the purposes of optimizing this energy consumption there is a demand in constructing a more energy-efficient buildings and retrofitting already built ones. To motivate investors to pursue renovations needed to optimize energy consumption of existing buildings, an accurate estimate of how much energy would improved building consume relative to the old one is important. One can make these estimates using programs based on physical principles that take as input climate data and many building features data such as geometry, materials used for constructing, positioning of the building and its orientation. These programs have been greatly utilised, however they can become very costly to handle when one wants to define large number of retrofits [2]. An alternate approach to estimating energy consumption would be to make these estimates using historical usage data. In this method, building features, building historical usage data and climate data are used together to predict building's energy consumption by applying a learning process. In this project I am using the latter approach to estimate buildings energy consumption. In subsections to follow I will describe the data I will use for the project along with the objective function and learning algorithm that I plan to apply to it.

Preprint. Under review.

1.1 Data

In this project I would like to use the data [1] of historic electricity usage rates and observed weather over one year period across hundreds of buildings in order to predict how much energy would these buildings consume in the future.

The input features for the learning can be split into three categories: consumption features, building features and weather features, where the output of this model, target data, would be the energy consumption in units of kWh. Consumption feature includes timestamp feature that will indicate at which point in time was the electricity reading taken. This feature will be encoded in several ways as described in Methods section, and I will compare how much these encodings are important. Building features include square feet, year built and floor count, and weather features include air temperature, cloud coverage, dew temperature, precipitation depth, sea level pressure, wind direction and wind speed.

1.2 Evaluation criteria

Since we would like to predict energy consumption of a building, which is a number, I will use root mean squared error (RMSE) as a natural evaluation metric:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i - y_i)^2} \tag{1}$$

where g_i is our estimated value and y_i is the target value.

1.3 Learning algorithm

There is a vast literature of different approaches used for building energy consumption predictions. In review [3] authors have discussed the performance of Artificial Neural Networks, Support Vector Machine and hybrid methods for load forecasting. In [2], in addition to previously mentioned techniques, authors have also discussed application of Gaussian distribution regression models and hierarchical clustering for this problem. In [4] authors have investigated the effectiveness of Long Short-Term Memory (LSTM) technique and its variations in prediction of electricity consumption data.

For the milestone report I compared predictions I get using various regression models including linear regression, decision tree regression and random forest regression model. As a next step in the project, I would like to investigate LSTM with this data.

2 Methods

2.1 Data processing and data selection

For this project I did two types of data processing and selection: one for the features data and one for the target data. Details on the work done is provided in appendix A.

2.2 Feature encoding

For this report I did two different types of encoding and I compared how these different encodings affect the goodness of the fit. I also compared if we were to do no encoding to the features, what would the results look like in that case.

First encoding is what I called simple encoding in my code [5]. For this encoding I replaced the timestamp feature with hour, month and year features. After this, I normalized all the input features, including these and this was the input for my learning algorithm.

Second encoding, that I called polynomial encoding, was inspired by Moon et al [6]. Given that energy consumption has clear daily variations, if we encode day as a numeric feature, then even though 0000hrs follows right after 2359hrs, they will numerically look like they are far apart. This encoding solves this problem by substituting day feature with day_sin and day_cos feature, with frequency that corresponds to 24h. I used this type of encoding for day, week and year feature with frequency of 24h, 7 days and 365 days respectively in order to reflect the cyclical nature of electricity consumption. I also added additional feature called is_weekday to handle lower amplitudes of energy consumption over weekend relative to weekdays. Wind_direction feature in units of degrees has similar issue as described above, where 0 and 359 degrees are numerically far away but physically they are close. Thus I substituted wind_speed and wind_direction features by combining them together into wind_x and wind_y features. After these encodings, I normalized all the input features before sending them to the learning algorithm.

2.3 Learning algorithm

For this report I compared the results of linear regression, decision tree regression, random forest regression and autosklearn regression in combination with different feature encodings as described in Feature encoding section in order to find the best prediction model for the data. For all the models I dedicated month of June to be the test data, and the rest of the months of the year to serve as training data for the models.

3 Results

Using root mean squared error as a metric, the Table 1 provides the breakdown of results that different learning algorithms in combination with different feature encodings give.

Model	Polynomial	Simple	No encoding
Linear regression	0.0514	0.111	15.9
Decision tree	0.0348	0.0261	0.097
Random forest	0.0260	0.020	0.095
Autosklearn	0.0215	0.05171	0.096

Table 1: RMSE of different regression learning models in combination with different feature encodings. All the data is obtained in 2-4min timescale except for data for autosklearn that was run for 1h for each table input.

Table above shows that the best results are obtained for random forest model with simple feature encoding. Visual comparison of predictions of this model w.r.t. target data is presented in Figure 1.

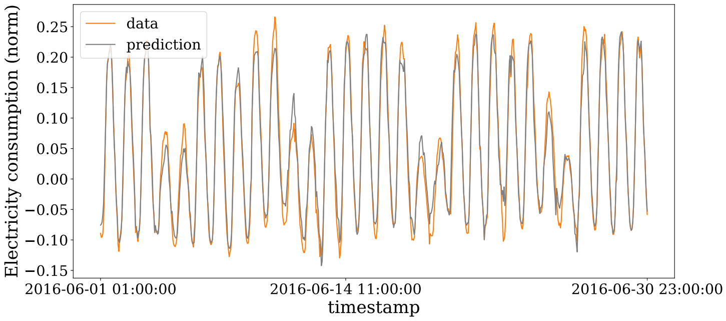

Figure 1: Random forest regression model with simple feature encoding fit for the month of June 2016 of averaged electricity consumption of 118 educational buildings. The values of electricity consumption presented here are normalized w.r.t. training data mean and standard deviation values.

An interesting question that comes to mind is whether it would be possible to find good predictor for the data without having any expert knowledge about what feature encoding to do or what regression model to use. An automated toolkit that can choose a good algorithm and feature processing such as autosklearn can be used to answer this question. As can be seen in the table, if we did no encoding before sending the data to autosklearn, we would not be able to make very good predictions, with metric obtained in this case being 5 times larger than the best value we were able to obtain with random forest regression. However, if we were to do a polynomial feature encoding, autosklearn would be able to provide results that are comparable with the best results we were able to obtain. This suggests that autosklearn could do well with automation of finding the right learning algorithm to fit the data, however, automation of feature processing is something that seems not to be optimized, at least for this data.

With regression models, we didn't have any memory feature of what the consumption value was in the past, and I believe that this type of information could improve the fit. The next step for the project would be to incorporate this type of memory feature through LSTM and to see if this architecture can improve the energy consumption fit.

References

[1] URL: `https://www.kaggle.com/c/ashrae-energy-prediction`.

[2] Saleh Seyedzadeh et al. "Machine learning for estimation of building energy consumption and performance: a review". In: *Visualization in Engineering* 6 (2018).

[3] Mohammad Azhar Mat Daut et al. "Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review". In: *Renewable and Sustainable Energy Reviews* 70 (2017), pp. 1108–1118.

[4] Daniel L. Marino, K. Amarasinghe, and M. Manic. "Building energy load forecasting using Deep Neural Networks". In: *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society* (2016), pp. 7046–7051.

[5] URL: `https://github.com/TamaraSumarac/6.862_MachineLearningProject`.

[6] Jihoon Moon et al. "A comparative analysis of artificial neural network architectures for building energy consumption forecasting". In: *International Journal of Distributed Sensor Networks* 15.9 (2019).

A Data processing details

As described in the main text, I did two types of data processing and selection: one for the features data and one for the target data.
A significant proportion of features data was missing in the downloaded data files. The breakdown of the missing data for the input features is the following: 53.4% of the year built, 75.5% of the floor count, 9.5% of the cloud coverage, 0.1% of the dew temperature, 36.0% of the precipitation depth, 7.6% of the sea level pressure, 4.5% of the wind speed and 0.2% of the wind direction. In order to have the largest number of data for the training, I filled these missing elements for all the input features. For the weather data features, I filled the missing data by taking the average of the closest non missing values of that feature, whereas for the building data I filled the missing data for the feature with the mean of that feature for that building's primary use group.
For the target data, the breakdown of the missing data is the following: 36% of the buildings had consumption data that was missing for more than 7 days and 83% of the buildings had missing data for more than 1 day. Given that consumption data has large daily and weekly variations, and given that this data should be what we predict, I chose, as a first step, not to include the data of the buildings that have missing values for more than a day. Also, as a first step, I ended up working only with buildings whose primary use is educational, which ended up being 118 buildings that I used for learning in total. As a next step in this project, I would extend this to different types of buildings using one-hot encoding for the primary use feature.