CEE 498DS Project 11: Building Energy Predictions - Project Report

This manuscript (permalink) was automatically generated from cathyxinchangli/cee498ds-project11@870d6e5 on December 5, 2020.

Authors

- **Xinchang 'Cathy' Li**
  · cathyxinchangli
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Benjamin Smakic**
  · mkbenja
  Department of Aeronautical & Vehicle Engineering, Royal Institute of Technology, KTH; Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Zhiyi Yang**
  · zhiyiy2
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Mingyu Sun**
  · TBD
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

Abstract

This is the abstract. Testing to see if it will show up.

By utilizing modern electric meters, it possible to collect and store enormous amount of data about household energy consumption. This data can be used to predict energy consumption and help energy providers manage energy (electricity) output and plan for energy peaks/lows.

Introduction

By utilizing modern electric meters, it possible to collect and store enormous amount of data about building energy consumption. This data can be used to predict energy consumption and help energy providers manage energy (electricity) output and plan for energy peaks/lows. It can also help the indiviudal resident keep track of their energy usage and perhaps even recommend energy saving actions.

As the human population increases, it is vital to lower the energy footprint of each individual and save the limited resources earth has to offer to future generations. Household energy consumption is a big part of our total energy consumption, and new technology offers new ways of decreasing it.

This project focuses on predicting building energy consumption based on the Kaggle competition "ASHRAE - Great Energy Predictor III". The main goal of the project is first and foremost to create an AI-model that can predict the energy usage of a building as accurately as possible, based on different input data such as air temperature, building size etc. In addition to that, different AI-models are explored to investigate which one works the best for this particular competition.

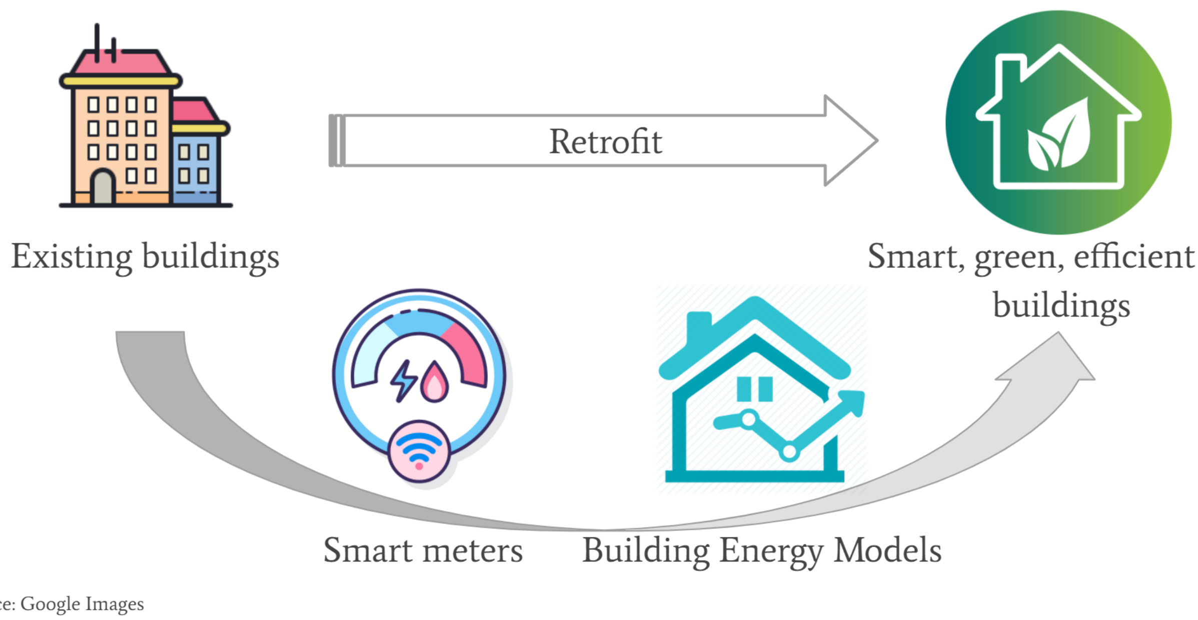

Figure 1: Building retrofit flowchart.

Literature review

Zhang, Grolinger, & Capretz (2109): Forecasting Residential Energy Consumption: Single Household Perspective

In the paper "Forecasting Residential Energy Consumption: Single Household Perspective" (Zhang, Grolinger & Capretz 2019), the authors attempt to predict energy consumption in residential households, with focus on single households.

According to the authors it is more difficult to predict single household energy (electricity) consumption, compared with e.g. workplace energy consumption. The reason is that single households often differ in energy consumption patterns while workplace patters tend to be more similar. Also, if big workplace buildings or multi-family residential buildings are analysed, any anomalies tend to cancel each other out (with a big enough dataset).

Data set

The data set used originates from an electricity provider in London, Ontario, Canada. It consists of hourly smart-meter readings of electricity usage (kWh) of 15 households between 2014 and 2016.

Firstly, the data set used might not be sufficiently broad. Tracking only 15 households will most likely not capture a variety of electricity consumption patterns. However, it was deemed enough in this case. Furthermore, the data set comes from one city with a certain climate, which means that different environmental prerequisites are not considered. Perhaps using other cities from a different part of the world would lay a foundation for a more advanced ML-algorithm (Machine Learning algorithm). As it is now, the ML-algorithm might be inaccurate for other parts of the world. In addition to this, Zhang, Grolinger & Capretz state that residents of London, Ontario, Canada tend to heat their homes gas heating systems, which affects the electricity consumption drastically.

Secondly, the data set used is pre-processed in different ways. Any missing readings of electricity consumption is replaced with the average value of the previous reading and the next reading, missing weather condition is replaced with the weather condition of the previous hour etc. This is perfectly good way of replacing missing data. However, the authors do not give an explanation as to why this method was chosen, if there are any consequences and if there are other methods of making the data set complete.

Exploratory Data Analysis

The EDA performed by the authors is illustrated in the form of electricity consumption graphs and heat-maps. The patterns show that most households live regular lives (the authors do not define what "regular lives" mean, though it can be understood by the context). However, there are some exceptions where irregularities occur (e.g. empty homes during the summer when consumption otherwise is the highest), which significantly reduce the precision. This could be improved by adding vacancy detection which could be implemented in the ML-algorithm (which raise privacy concerns). Zhang, Grolinger & Capretz realize that the top three most important variables (i.e. the variables that correlate the most with the output energy consumption) are "temperature", "hour of the day" and "peak index". Peak index is a variable that captures important energy usage peaks, such as peak hours, days, seasons etc.

Prediction model and results

The authors used Support Vector Regression (SVR) to predict the energy consumption. SVR is a supervised machine learning algorithm, which means that it is compares an input with an output and is trained by comparing predicted results with true results. It was chosen due to time and computational hardware constraints. No other evaluations or comparison of other machine learning algorithms were made by the authors, so it is difficult to understand why SVM is faster and require less processing power. Zhang, Grolinger & Capretz present the results for home #1 in figure 1. They managed to predict electricity consumption well. The most inaccurate parts are peaks that arise due to random variations. Furthermore, the authors present a table with results for all 15 homes. According to them, time-based splitting is used to check parameter stability over time, which makes the algorithm more accurate. In this case however random sampling performs better in cases where some residential customers have irregular and uncertain patterns. These uncertain patterns make time-based splitting more inaccurate over time. Therefore, both methods are employed. Lastly, mean absolute percentage of error (MAPE) was utilized to measure the performance of the algorithm, which is a widely used performance metric.

Figure 1: Observed electricity consumption compared to predicted electricity consumption for house #1 (of 15) (Zhang et al.). Insert figure Figure 2: Performance results of the predication model for all homes. Insert figure #### Conclusion The biggest strength of the paper is the execution of the chosen methods to achieve desired results. It is a relatively successful attempt at predicting single households, which tend to be more unpredictable compared to multi-family or corporate residential buildings. The biggest weakness is the justification for the chosen methods. The authors do an excellent job of utilizing the chosen methods, but there is little thought put in to why these methods where chosen or why they did certain things. By including a more extensive evaluation and justification for method choices, the target audience and other researchers in the same field can understand better and continue the research. However, it makes the paper longer and more complex, which can be negative for the readers and the target audience.

Edwards, New, & Parker (2012): Predicting future hourly residential electrical consumption: A machine learning case study

The article explores seven machine learning (ML) techniques on their performances in predicting next hour electricity consumption of buildings, with a focus on residential buildings. Sensor-based energy modeling uses high-frequency sensor data and ML algorithms to statistically derive building energy forecasting models that can help improve building efficiency. Previous studies have applied such models to commercial buildings where high-frequency sensor data are available, but not to residential buildings due to a lack of sensors in homes and consequently a lack of frequent sensor data. The authors address this gap by employing a new high-frequency residential dataset to test proven and emerging ML techniques on predicting next hour residential energy consumptions. The ML techniques used include: Linear Regression, Feed Forward Neural Networks (FFNN), Support Vector Regression (SVR), Least Squares Support Vector Machines (LS-SVM), Hierarchical Mixture of Experts (HME) with Linear Regression Experts, HME with FFNN Experts, and Fuzzy C-Means with FFNN. They use cross validation for parameter tuning, and select the best model based on three performance metrics, namely Coefficient of Variance (CV), Mean Bias Error (MBE) and Mean Absolute Percentage of Error (MAPE).

The authors first validate the ML techniques on the ASHRAE Great Energy Predictor Shootout dataset, which contains hourly sensor data for a commercial building of year 1989. They find that FFNN performs best on this dataset, with CV at ~11%, competitive with the top-3 competition winners. Then they apply these techniques to the new dataset, the Campbell Creek dataset, which contains full-year (2010) measurements at 15-minute interval from sensors installed on three Tennessee homes. The results show unsatisfactory results from FFNN (CV = ~32%), close to the baseline Linear Regression (CV = ~34%). LS-SVM is selected as the overall best technique for modeling the Campbell houses, with an average CV of ~26%. The overall larger model errors in the residential dataset are attributed to variant occupancy behaviors in homes that lead to more complex energy use patterns, as compared to those of commercial buildings, which tend to vary only between workdays and weekends/holidays. The also find that statistically different training and testing data may cause LS-SVM to fail when generalizing to the testing data, which is the case for the ASHRAE dataset. The performance of LS-SVM is improved after randomizing the training and testing data.

This paper provides a comprehensive analysis on the application of multiple ML techniques to building energy data, which could serve as a guide to our project. The discussions on the differences between commercial and residential buildings are particularly helpful for our feature engineering and model selection, as we are expected to model 16 building types, commercial and residential included. The authors are also able to identify and prove the cause of failure for LS-SVM on the ASHRAE dataset, which may help us in model tuning and selection. Despite these merits, the paper does not address the differences in performance for FFNN, HME-FFNN and FCM-FFNN, the three closely related methods, shedding doubts on whether it is necessary to include them all. In addition, it would be very informative if the paper included the computational expenses required for training each model, which could be an important metric in model selection especially for real-life applications.

Methods

Exploratory Data Analysis

Five data files are provided in the competition dataset. The file names and the contents of each file are detailed below (adapted from the [competition site]()).

train.csv * `building_id` - Foreign key for the building metadata. * `meter` - The meter id code. Read as {0: electricity, 1: chilledwater, 2: steam, 3: hotwater}. Not every building has all meter types. * `timestamp` - When the measurement was taken * `meter_reading` - The target variable. Energy consumption in kWh (or equivalent, except site 0 electric meter readings which are in kBTU). Note that this is real data with measurement error, which may impose a baseline level of modeling error.

building_meta.csv * `site_id` - Foreign key for the weather files. * `building_id` - Foreign key for training.csv * `primary_use` - Indicator of the primary category of activities for the building based on EnergyStar property type definitions * `square_feet` - Gross floor area of the building * `year_built` - Year building was opened * `floor_count` - Number of floors of the building

weather_[train/test].csv
Weather data from a meteorological station as close as possible to the site. * `site_id` * `air_temperature` - Degrees Celsius * `cloud_coverage` - Portion of the sky covered in clouds, in oktas * `dew_temperature` - Degrees Celsius * `precip_depth_1_hr` - Millimeters * `sea_level_pressure` - Millibar/hectopascals * `wind_direction` - Compass direction (0-360) * `wind_speed` - Meters per second

test.csv
The submission files use row numbers for ID codes in order to save space on the file uploads. test.csv has no feature data; it exists to help get predictions into the correct order.

- `row_id` - Row id for your submission file
- `building_id` - Building id code
- `meter` - The meter id code
- `timestamp` - Timestamps for the test data period

sample_submission.csv
A valid sample submission containing `row_id` to match your predictions.

All floats in the solution file were truncated to four decimal places. There are gaps in some of the meter readings for both the train and test sets. Gaps in the test set are not revealed or scored.

Missing values

Here are the number (percentage) of missing values in each dataframes: 1. **Missing data in bldg_meta**:
* primary_use: 0 (0.0%) * year_built: 774 (53.4%) * square_feet: 0 (0.0%) * floor_count: 1094 (75.5%)
2. **Missing data in weather_train**: * air_temperature: 55 (0.0%) * cloud_coverage: 69173 (49.5%) * dew_temperature: 113 (0.1%) * precip_depth_1_hr: 50289 (36.0%) * sea_level_pressure: 10618 (7.6%) * wind_direction: 6268 (4.5%) * wind_speed: 304 (0.2%)
3. **Missing data in weather_test**: * air_temperature: 104 (0.0%) * cloud_coverage: 140448 (50.7%) * dew_temperature: 327 (0.1%) * precip_depth_1_hr: 95588 (34.5%) * sea_level_pressure: 21265 (7.7%) * wind_direction: 12370 (4.5%) * wind_speed: 460 (0.2%)

As we can see, some features such as `year_built`, `floor_count` and `air_temperature`, are missing over half of all entries. Special attention should be given to these variables when filling in the missing values to avoid losing raw training data while minimizing artificial influence.

Building Metadata

There are in total 16 sites, labeled 0~15 (`site_id`), containing 1449 buildings. Each building is identified with a unique `building_id` independent from the `site_id`, from 0 to 1448. The number of buildings in each site differ greatly, as shown in 2.

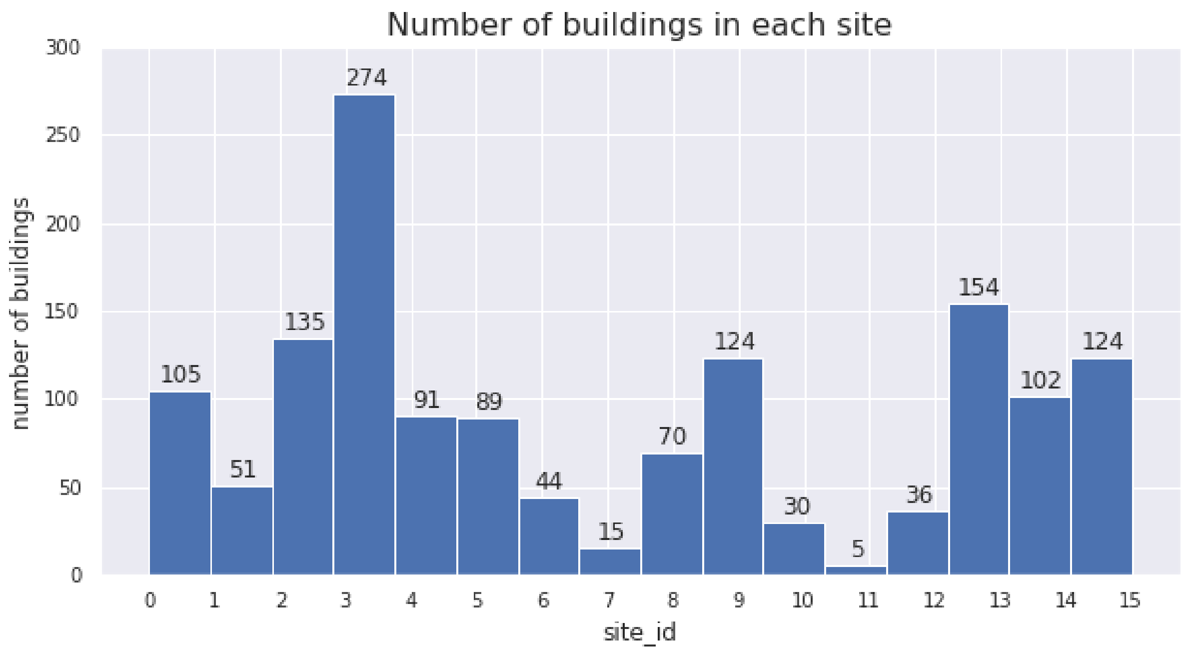

Figure 2: Number of buildings in each site.

There are 16 primary use types, with a mix of residential and commercial buildings, mostly built after the 1950s. Both the building square footage and floor counts are approximately logarithmically distributed, meaning most buildings are relatively small, single- to multi-story buildings (3). There's a fairly strong correlation (correlation coefficient = 0.53) between building square footage and floor counts, as we could expect, and a small but positive correlation between the constructed year and the building size (0.11).

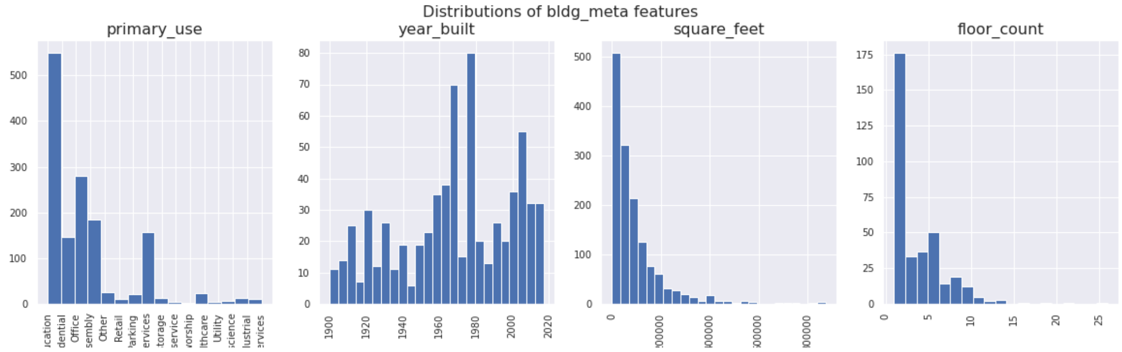

Figure 3: Distribution of features in building_metadata.

Target Variable: Meter Readings

ASSIGNED TO: Mingyu

Weather Data

`weather_train` has 2016 hourly weather data, and `weather_test` has 2017~18 hourly weather data. The time series plots for all variables of both the training and test periods are shown below (4).

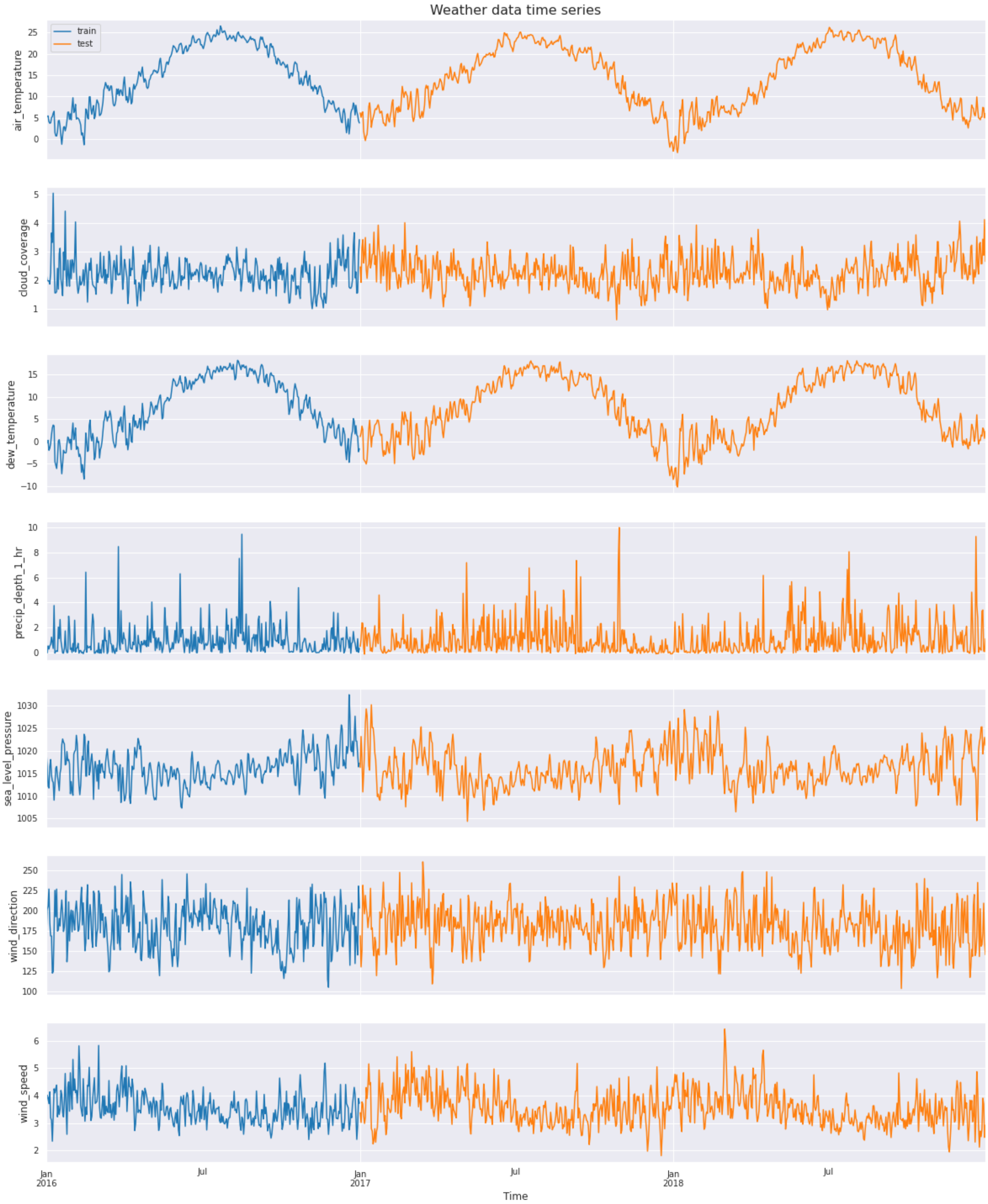

Figure 4: Time series plots of weather variables.

Correlations

Putting `building_metadata`, `train` and `weather_train` together, we can generate the correlation between each features and the target variable. The heat map below (5) shows that the correlation between variables range from -0.32 to 0.98, but no individual features have significant correlation with the target variable `meter_reading`. The top 5 most features most correlated with `meter_reading` are building square footage (0.13), number of floors (0.13), year of construction (0.11), meter type (0.077), and sites (0.047). This suggests that building metadata are potentially important predictors for our machine learning models, and the missing values need to be treated with care.

In addition, some features show rather strong correlations with each other, such as: `square_feet` and `floor_count` (0.58); `air_temperature` and `dew_temperature` (0.75); `wind_direction` and `wind_speed` (0.43). This may provide insights to imputation of the missing values.

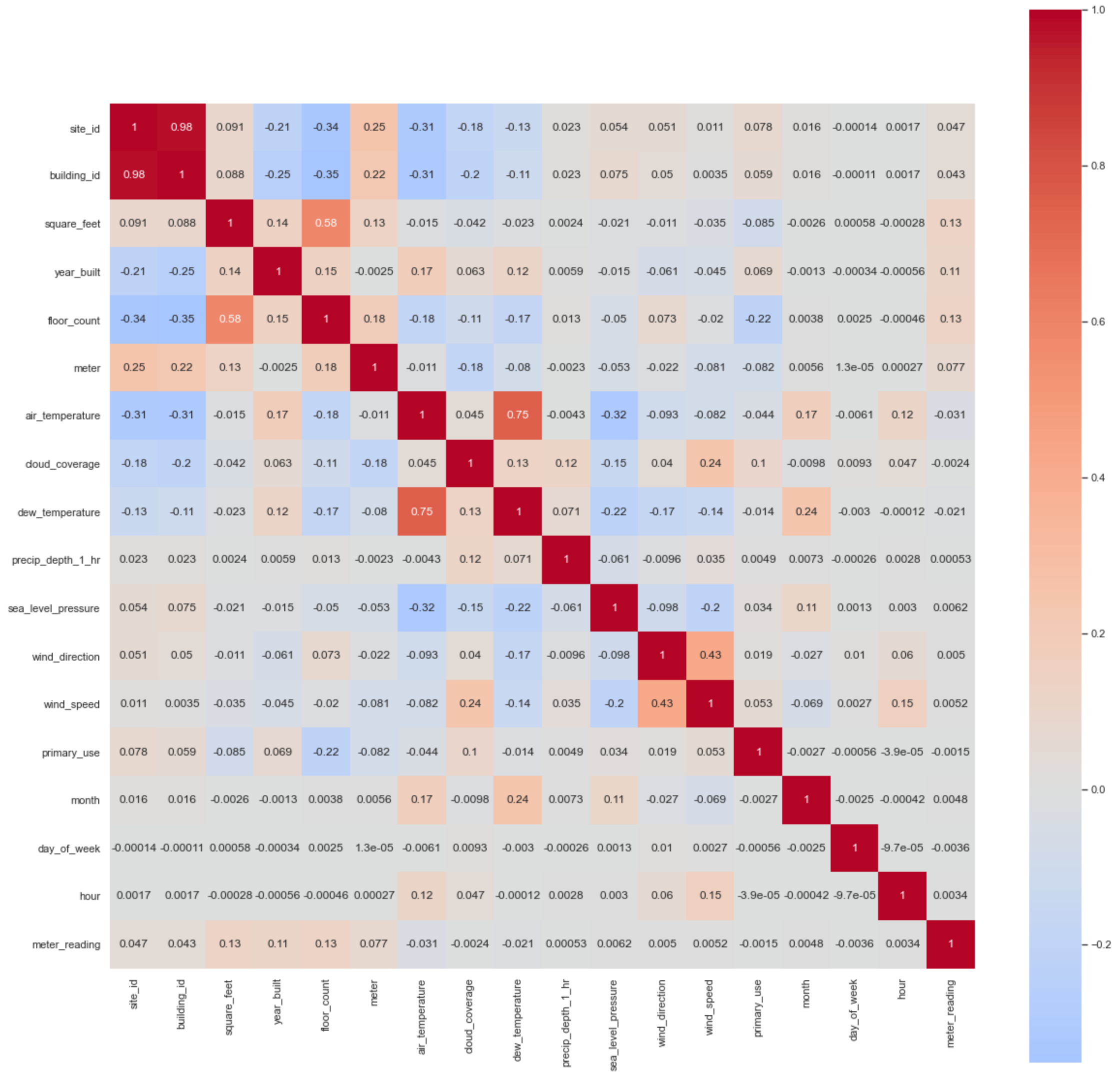

Figure 5: Correlations heatmap.

Machine Learning Models

Baseline: Linear Regression

Neural Network: RNN-LSTM

Tree-based Model: LightGBM

References