Udacity ML Nano Degree Capstone Project
Ashrae Energy Prediction

Kally Wenying Wu, October 31st 2019

1. Domain Background

Buildings consume about 40% of the total energy use in the United States. In recent years, significant investments have been made to improve building energy consumption to lower operational cost and reduce environmental footprint. Predicting energy used by heating, ventilating, and air-conditioning systems is important for HVAC diagnostics, system control, system identification, as well as energy management and optimization.

Under the current pay-for-performance financing plan, building owners make payments based on the difference between their real energy consumption and what they would have used without any retrofits. The latter values come from an estimation model. However, building energy estimation models are challenging to build and current methods of estimation are fragmented and many of them do not scale well.

2. Problem Statement

The goal of this project is to use machine learning algorithms to building energy estimation models based on historic usage rates and historic weather data to predict building energy usage across four energy types: chilled water, electric, hot water and steam meters. Machine learning algorithms product accurate energy consumption forecasts, and they can be used to implement energy saving policies. My goal is to work on more accurate prediction to improve the efficacy of building energy conservation measures and lower the cost of pay-for-performance financing.

3. Datasets and Inputs

I will be using datasets from a Kaggle competition – Ashrae, Great Energy Predictor III. The Kaggle datasets come from over 1000 buildings over a three-year timeframe. The dataset includes three years of hourly meter readings from over one thousand buildings at several different sites around the world.
The datasets include the following files:

train.csv
Historic meter reading data by timestamp for the building

- building_id - Foreign key for the building metadata.
- meter - The meter id code. Read as {0: electricity, 1: chilled water, 2: steam, 3: hot water}. Not every building has all meter types.
- timestamp - When the measurement was taken
- meter_reading - The target variable. Energy consumption in kWh (or equivalent). Note that this is real data with measurement error, which we expect will 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

building_meta.csv
Building metadata including the building use, square ft area, year build

weather_[train/test].csv

Weather data from a meteorological station as close as possible to the site. Dataset includes precipitation, cloud coverage, air temperature and etc.

- 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 contains the meter, building id and timestamp that we will be predicting for.

- 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 contains all the future data that we need to predict on.

4. Solution Statement

First, I will train a baseline model using linear regression.
Next, I will experiment with more computationally expensive models such as support vector machine (SWM), random forest, and k-nearest neighbors to make predictions.
Last, I will work with different boosting algorithms, and deep neural networks (LSTM).
I will use the evaluation metrics (RMSLE) to compare the performances of these solutions against the baseline model.

5. Benchmark Model

For the baseline, I will train a simple linear regression model after preprocessing training data.

6. Evaluation Metrics

The evaluation metric for this competition is Root Mean Squared Logarithmic Error (RMSLE). We use RMSLE here instead of RMSE because both predicted and true values are huge numbers and RMSLE penalizes the underestimation of the actual values more severely than it does for overestimation.

The RMSLE is calculated as:

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:
n is the total number of observations in the data set,
pi is your prediction of target, and
ai is the actual target of i.
log(x) is the natural logarithm of x.

7. Project Design

a. Data Preprocessing

Joining train.csv with weather_train.csv files based on site_id
Converting timestamps to datetime objects

b. Exploratory Data Analysis

Conducting exploratory data analysis on the joined dataset to discover existing patterns and pick meaningful variables for feature engineering

c. Feature Engineering

Applying results from exploratory data analysis and create features for building machine learning algorithms

d. Training Models and Evaluating Models

Training models using different algorithms, and evaluating results using RMSLE.

References

1. Kreider, J.F., and Haberl, J.S. Sat . "Predicting hourly building energy use: The great energy predictor shootout -- Overview and discussion of results". United States.
2. ASHRAE – Great Energy Predictor III https://www.kaggle.com/c/ashrae-energy-prediction
3. Building energy consumption prediction, a comparison of five machine learning algorithms, http://cs109-energy.github.io/