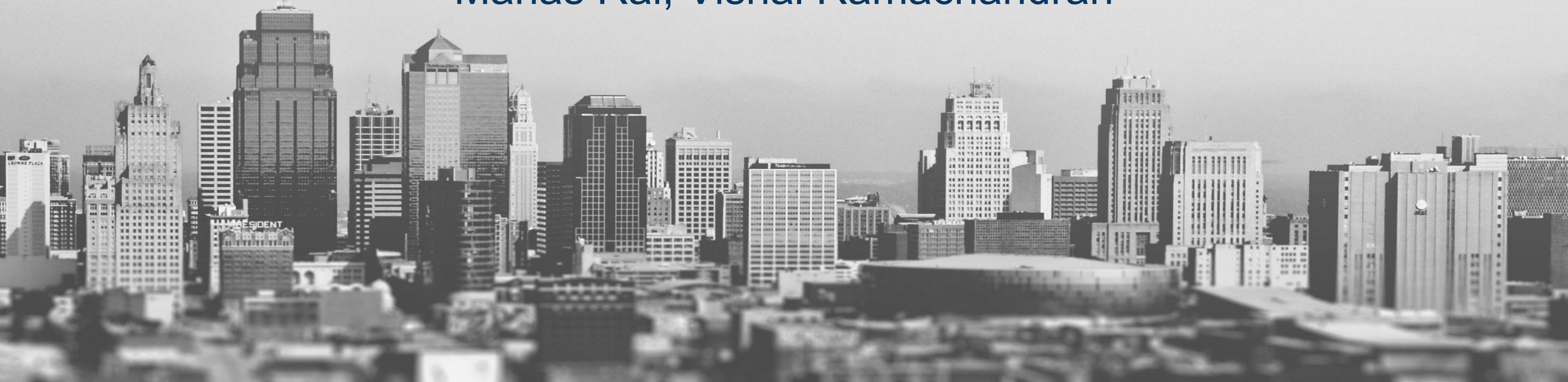
Great Energy Predictor

Team: High on Data

Members: Anisha Alluru, Apoorva Jasti, Aurangzaib Ahmed Siddiqui,
Manas Rai, Vishal Ramachandran



Agenda

- Project Overview
- Data Description and Handling
- Exploratory Data Analysis
- Modeling and Prediction
- Recommendations and Conclusion

Project Description

Background

Today, significant investments are being made to improve building efficiencies but owners makes payments based on the **difference between their real energy consumption and what they would have used without any retrofits**. The latter values come from a model.

Objective

We have chosen a kaggle case competition focusing on **building accurate models of metered building energy usage** provided by ASHRAE, an institute that serves to advance the arts and sciences of heating, ventilation, air conditioning, refrigeration and their allied fields.

https://www.kaggle.com/c/ashrae-energy-prediction/data

Technology Stack

Data Description

Data collected over 3 years across 16 sites for around 1500 buildings.

- Building metadata
  - Floor count, area, age, primary use, site, etc.
- Weather data
  - Site level hourly data of temperature, wind speed, precipitation, cloud coverage
- Meter Readings
  - Building level hourly data of energy consumption readings collected from 4 different meters
  - 20 million training data points and 40 million test data points

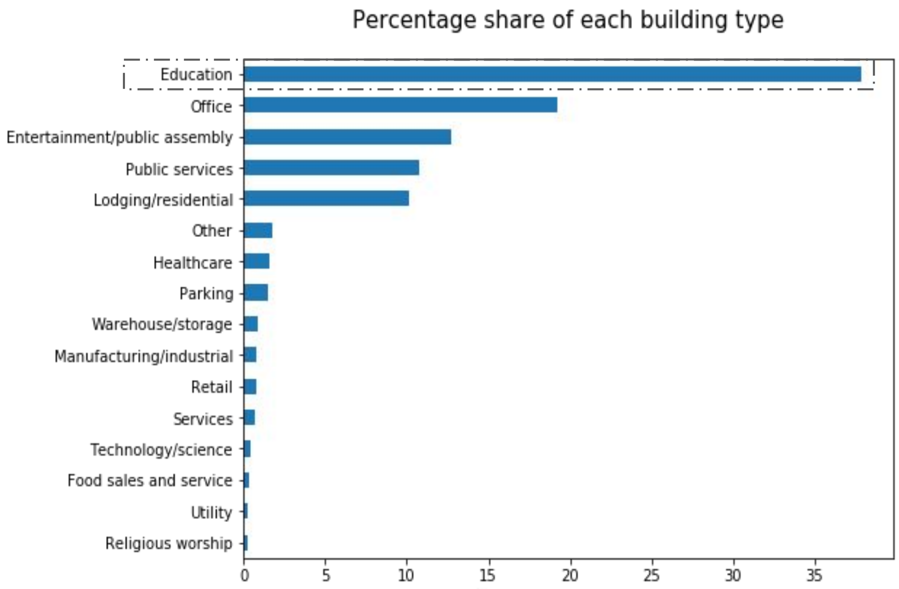
60 million data points - Databricks & Colab were like

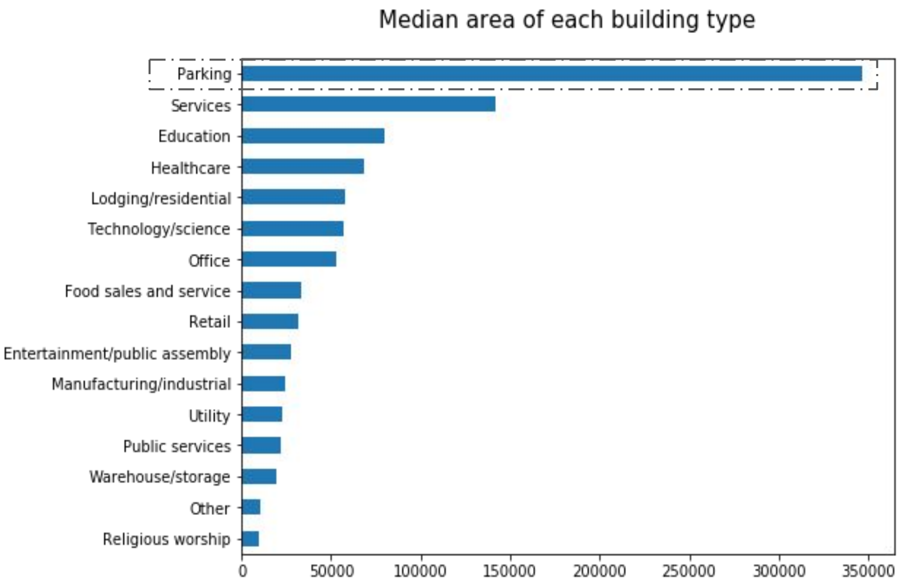
Data Size Reduction

- 60 million data points - Around 8 GBs of data before feature engineering

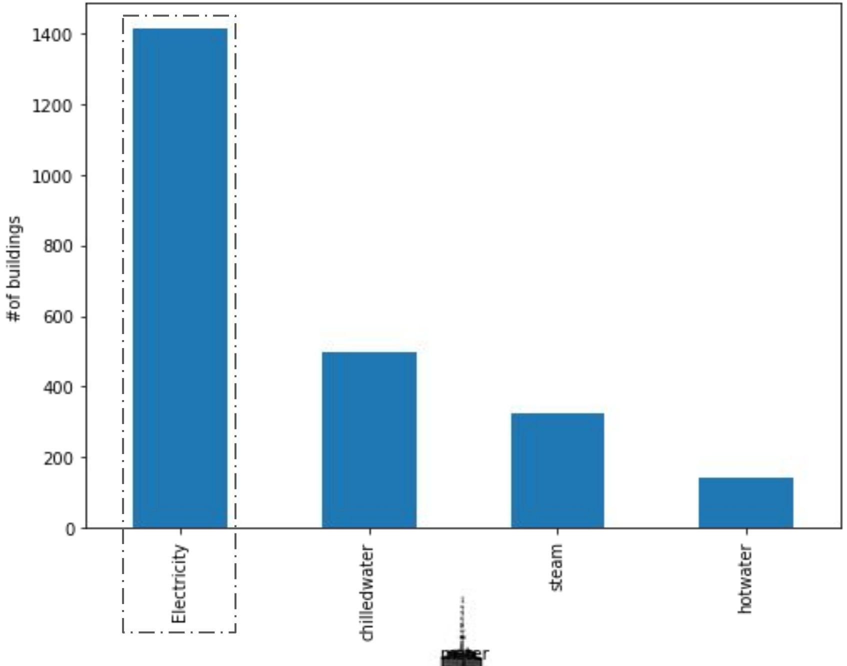
- Reduced data size by almost 50%
  - int64 -> int32 and int16
  - float64 > float32 and float16
  - Removed columns with >70% nulls
  - Label Encoding categorical features to int8

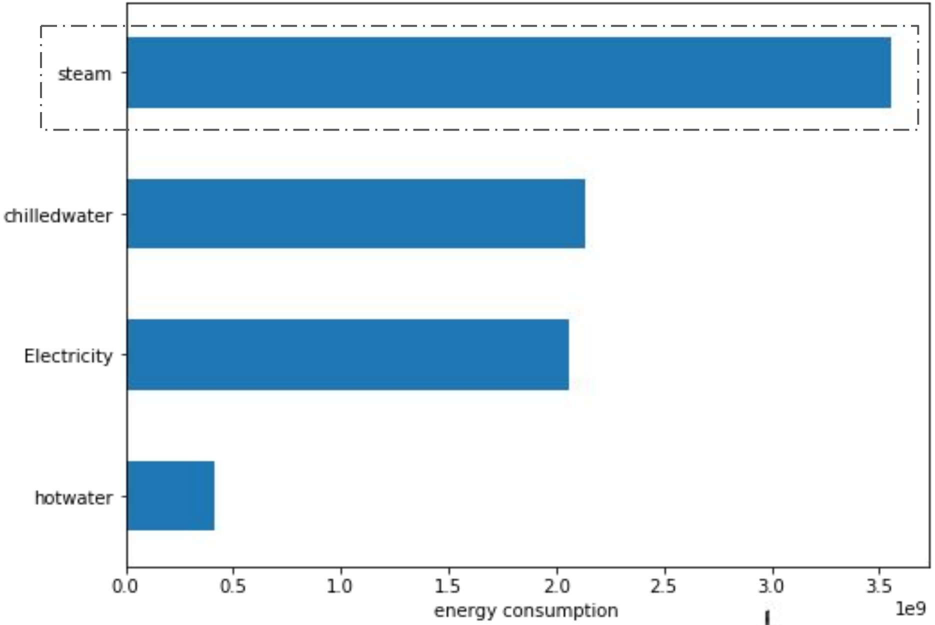
Educational buildings have maximum presence while Parking takes the most space



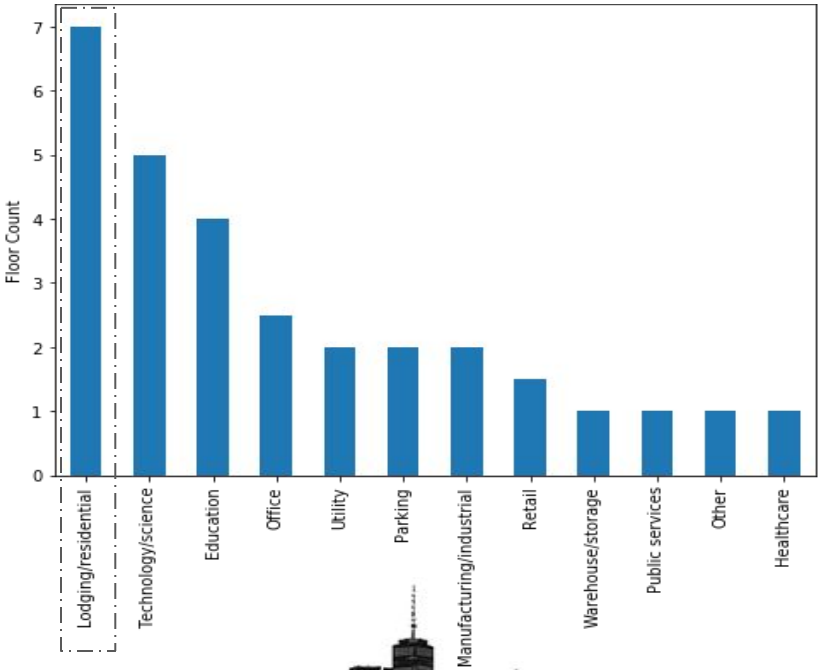


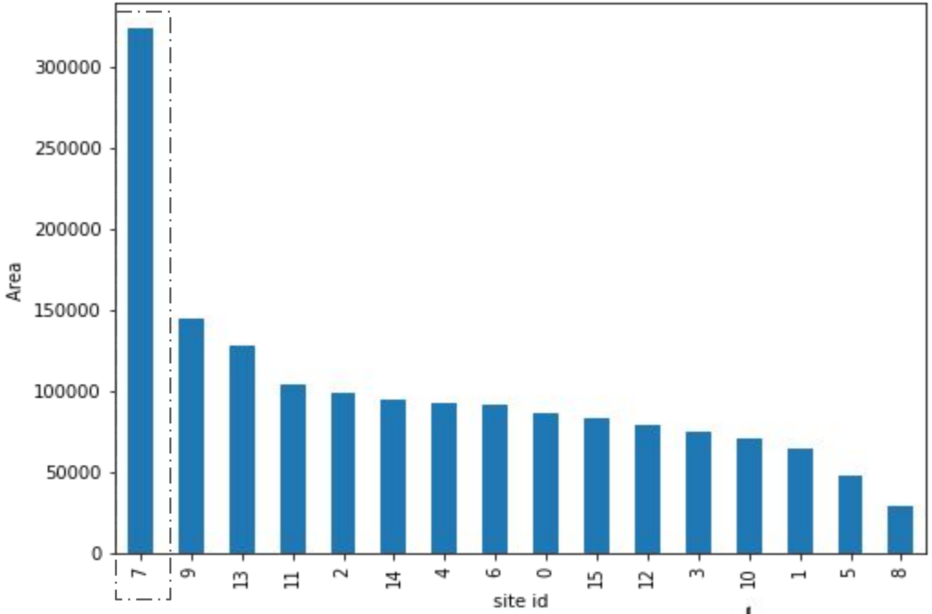
Electricity meter is omnipresent while Steam consumes the most energy



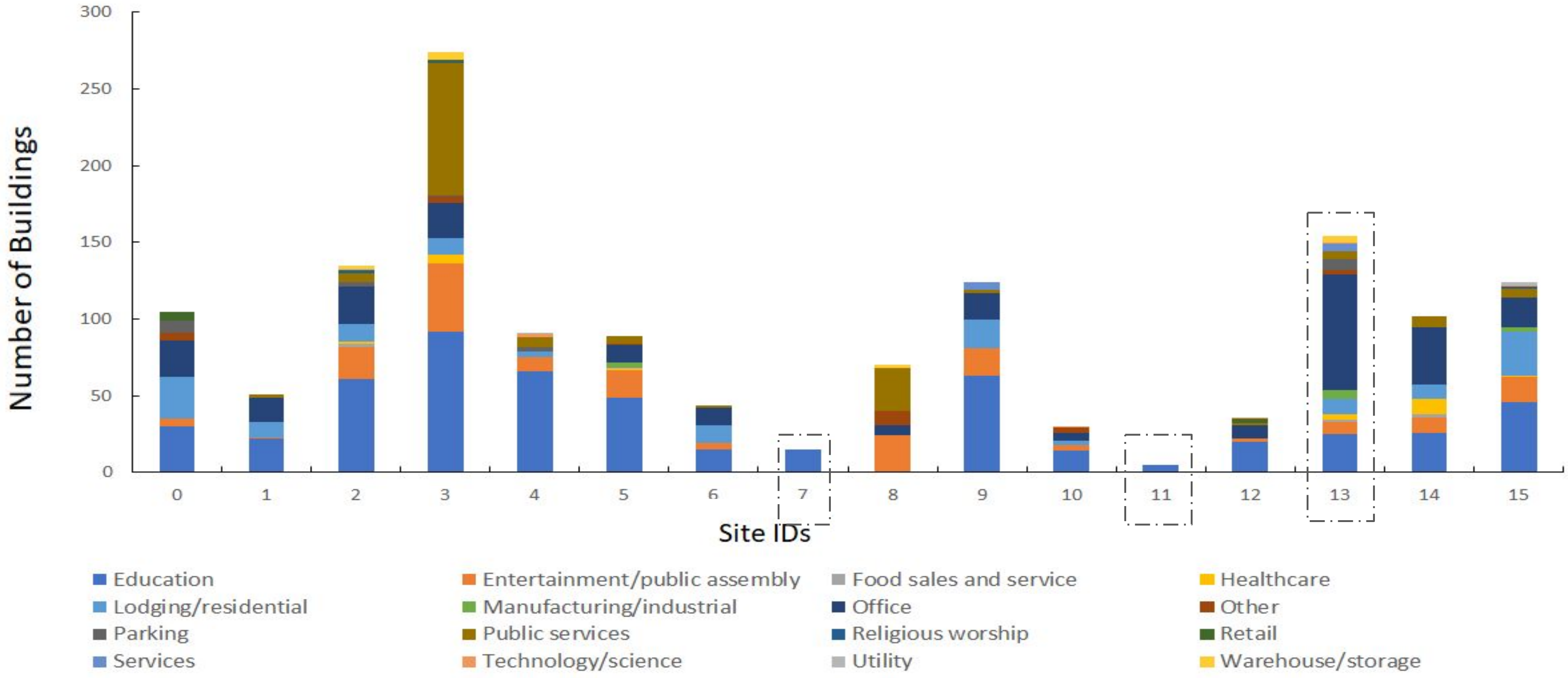


Residential/Lodging are tallest & Site 7 has more spread buildings

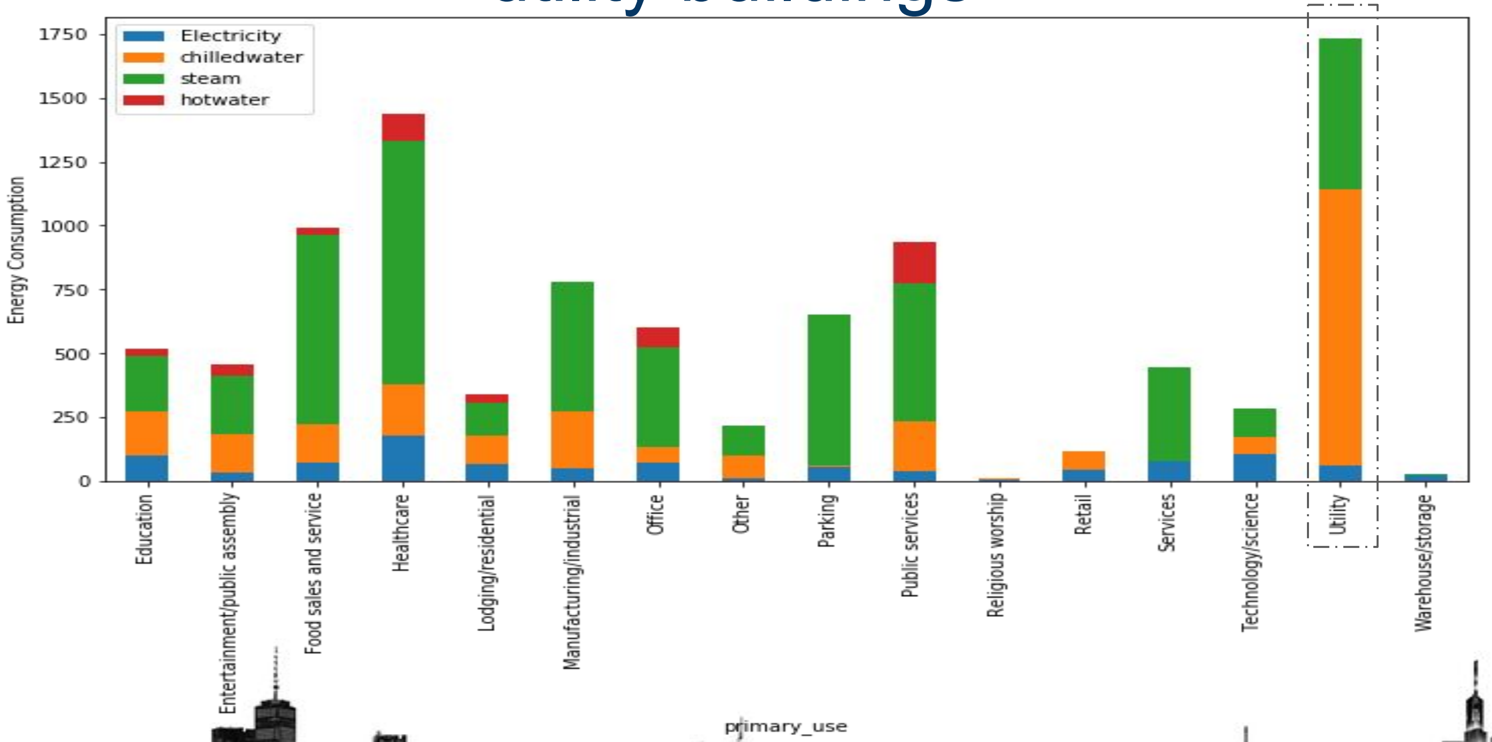




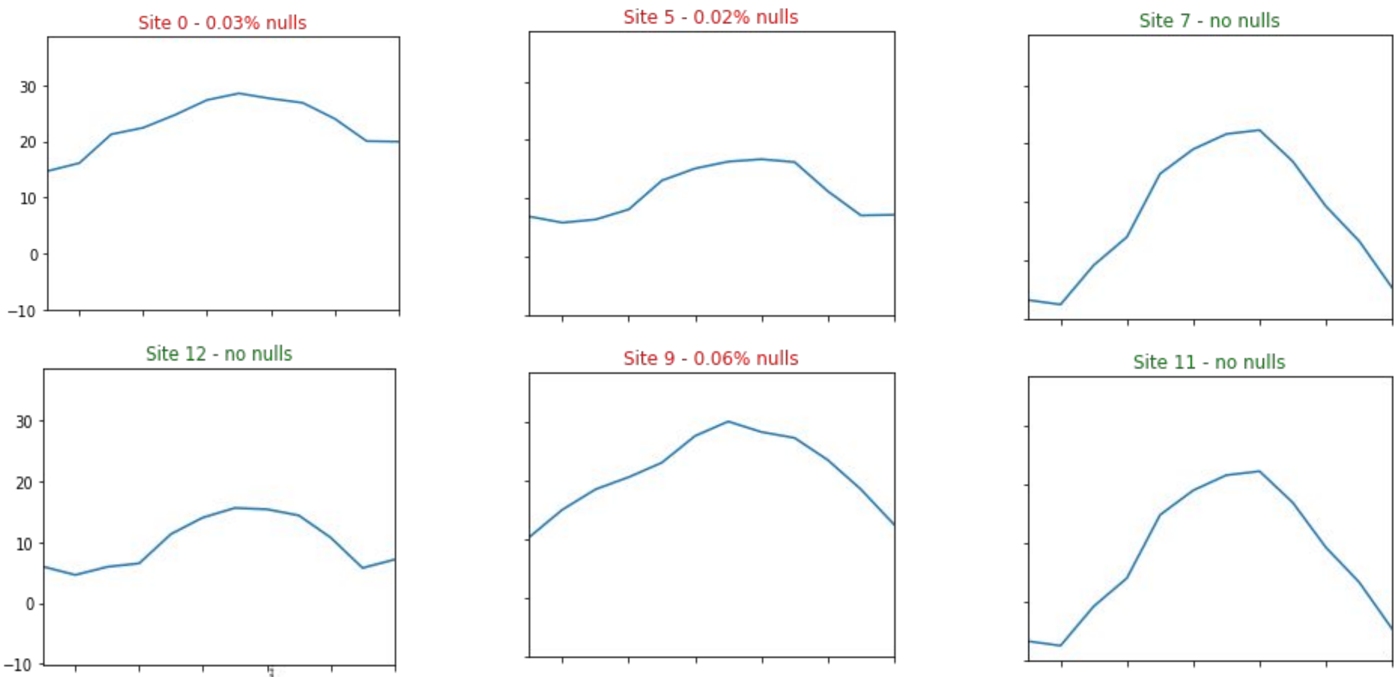
Sites 7 & 11 predominantly have educational buildings while Site 13 mostly has office buildings



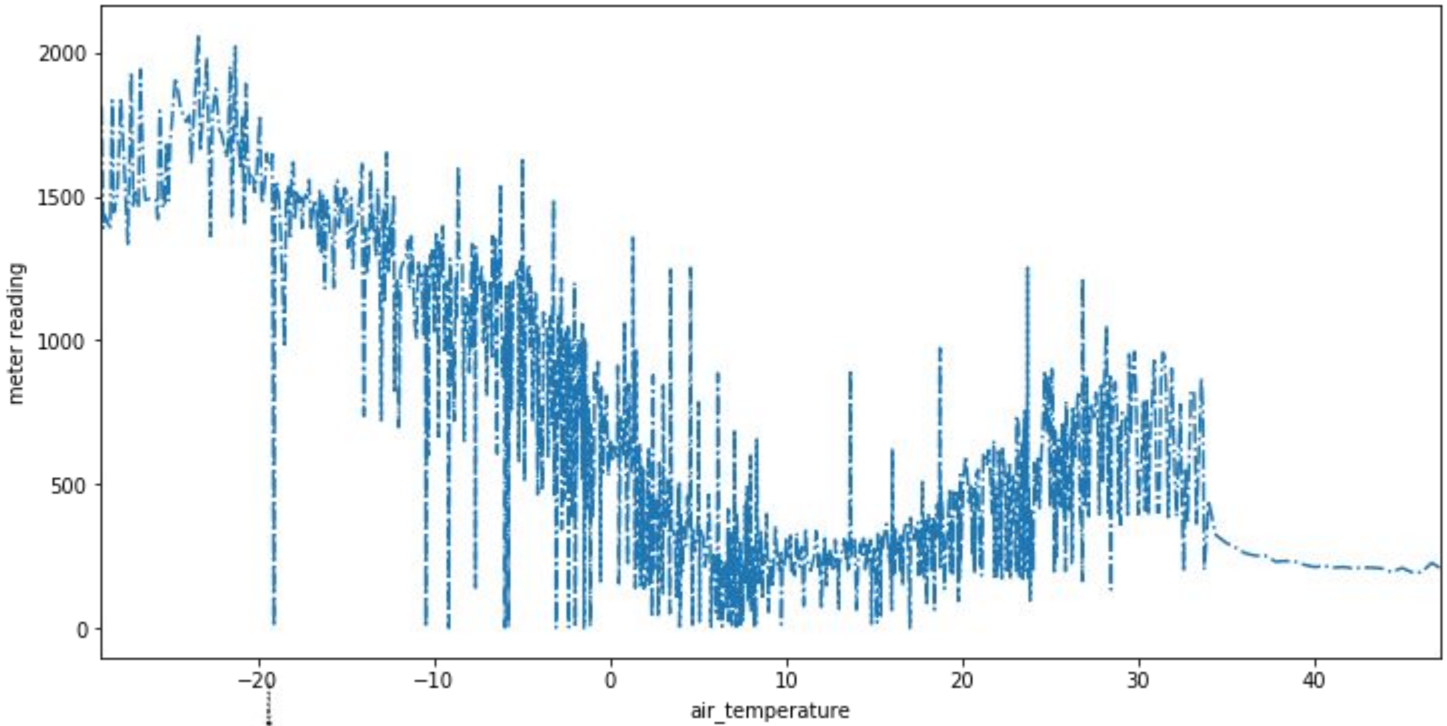
Chilled water was the most consumed energy in utility buildings



Monthly Weather patterns (Air Temperature) across Sites

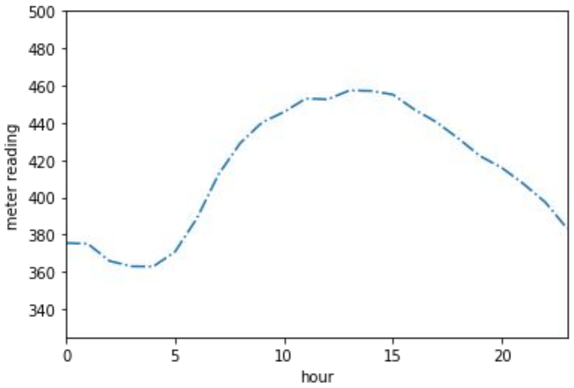


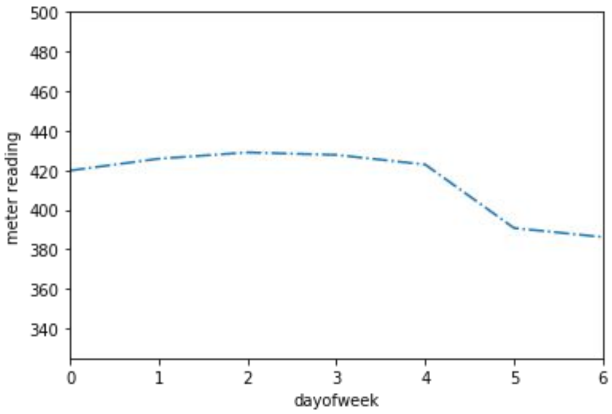
Air Temperature is a good indicator of energy consumption

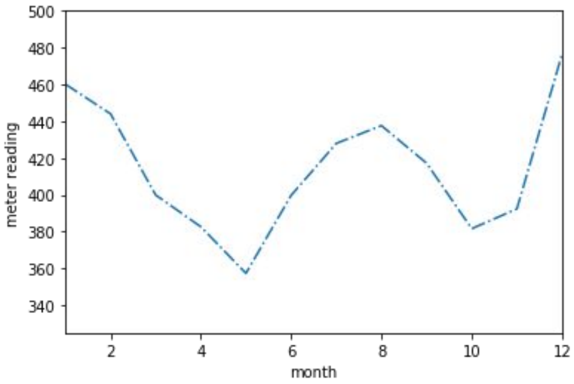


Time Series Analysis of Energy Consumption

Hour of the Day	Day of the Week	Month of the Year



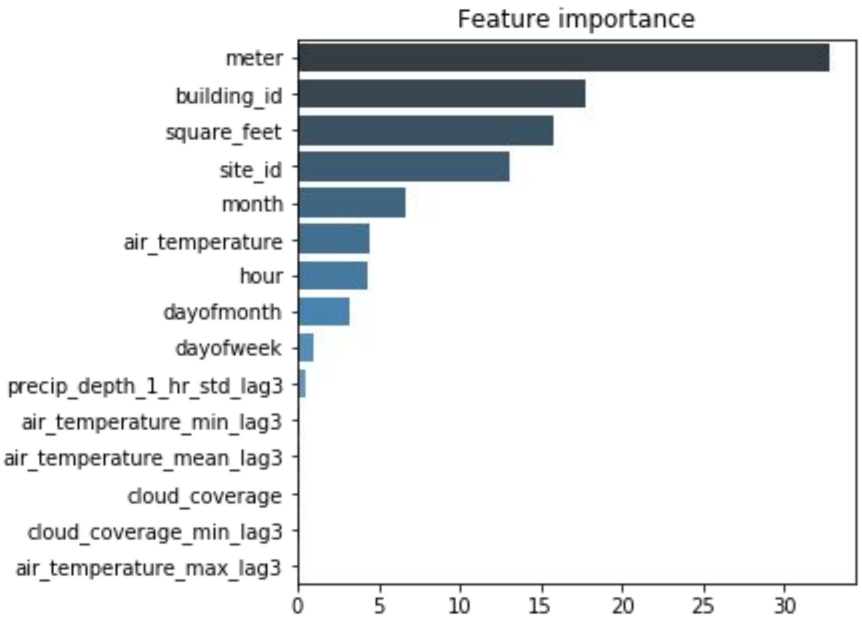




Higher energy consumption during day time	Higher energy consumption during the weekdays	Higher energy consumption during Dec-Jan

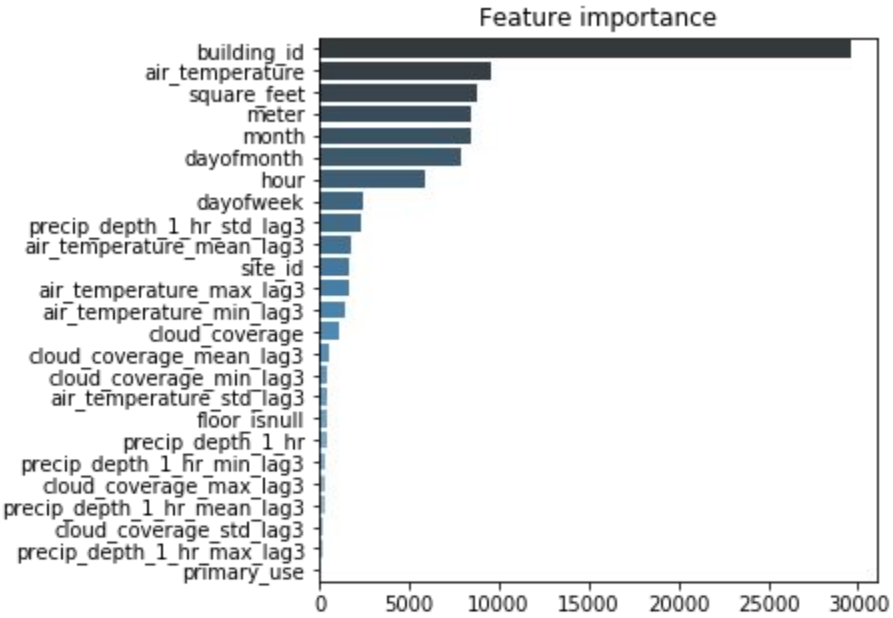
Modeling - CatBoost

- Train and Validation
  - H1 vs H2
  - 3 fold CV
- 500 boosting trees
- Parameters
  - Learning rate: 0.05
  - Max depth: 8
- Train Score: 0.84
- Validation Score: 1.29
- Test Score: 1.34

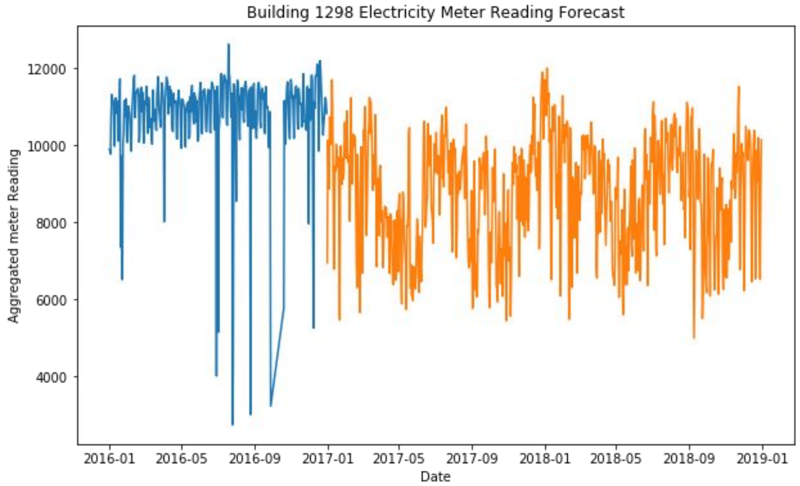


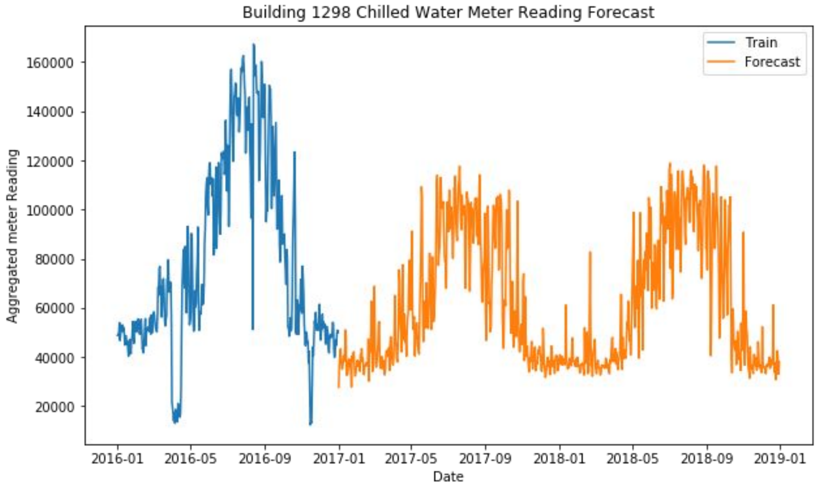
Modeling - LGBM

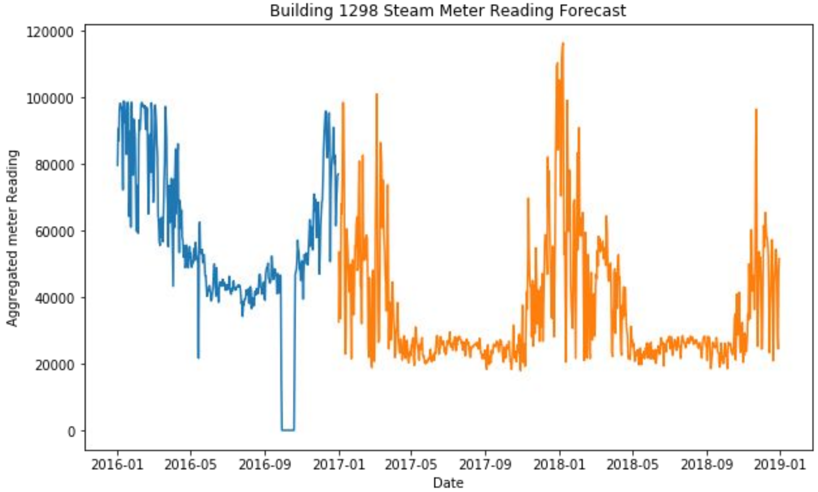
- 3 fold cross validation
- Tuned Parameters
  - Learning rate: 0.03
  - Feature Fraction: 0.85
  - Reg Lambda: 0.2
  - Max depth: 10
- Train Score: 0.55
- Validation Score: 1.02
- Test Score: 1.11 (Kaggle Top 40%)
- Test Score Post-Processing: 1.01 **(Kaggle Top 20%)**

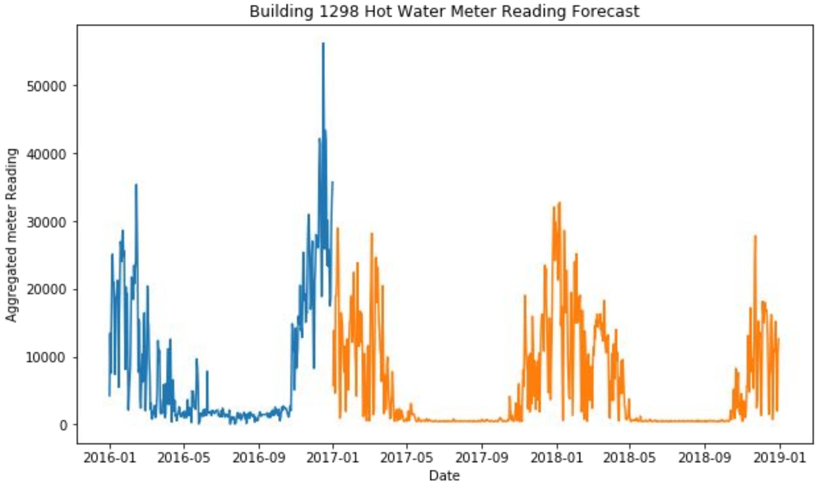


Model Prediction (for Building 1298)









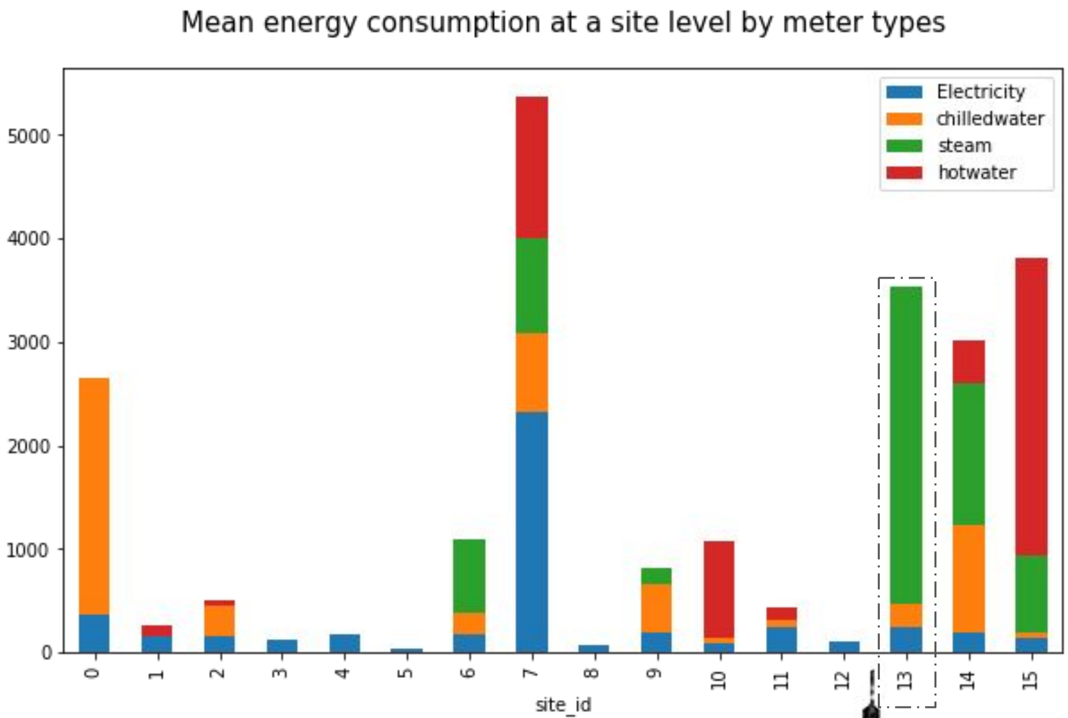
Recommendation

Insights

Site 13 consumes high amount of steam

Recommendation

Upgrade the steam equipment used in the offices in Site 13



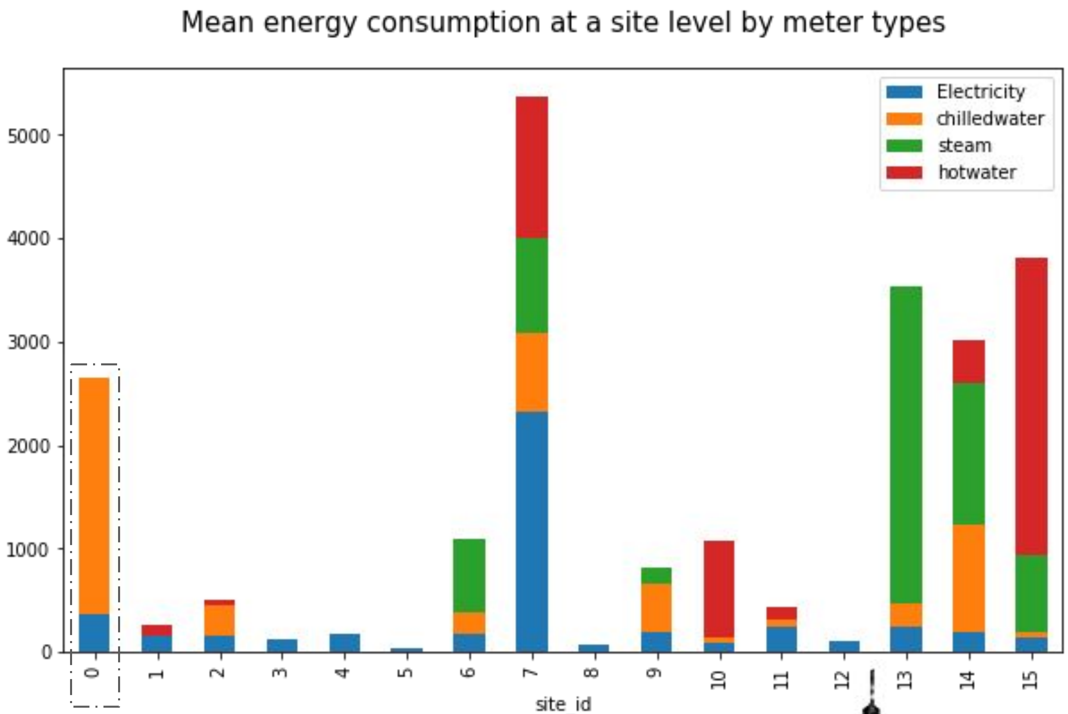
Recommendation

Insights

Site 0 consumes most amount of chilled water, at least 6 times more than others

Recommendation

Upgrade the chilled water system in the educational and residential buildings



Recommendation

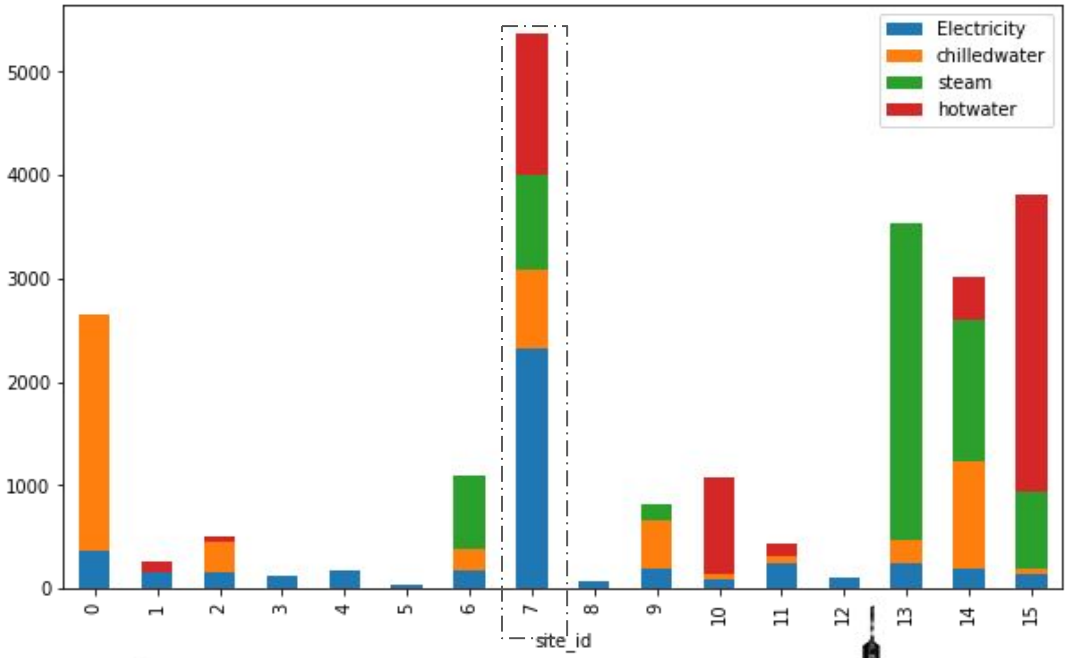
Insights

Site 7 has the highest energy consumption across all sites

Recommendation

Upgrade all the old equipments in the educational buildings in this site

Mean energy consumption at a site level by meter types



Next Steps

- Different approaches yet to be tried:
    - Train for each meter
    - Train for each site
- Try to utilize external data
- Tune parameters further
- Try stacking CatBoost and LGBM outputs
- Try H20.ai Auto ML algorithm for benchmarking

Questions?