Friends don't let friends over-fit: Benchmarking open-source prediction methods on building electrical meter data

Clayton Miller

Building and Urban Data Science (BUDS) Lab, Dept. of Building, School of Design and Environment (SDE), National University of Singapore (NUS)

Abstract

Prediction is a common machine learning (ML) models used on sub-hourly building energy consumption data. This process is valuable for anomaly detection, load profile-based control, energy plant systems control, and measurement and verification procedures. Literally hundreds of building energy prediction techniques have been developed over the last three decades, yet there is still no consensus on which techniques are the most effective for various building types. In addition, many of the techniques developed are proprietary and unavailable to the general research community. This paper outlines a library of open source regression techniques from the *Scikit-Learn* Python library and describes the process of applying them to open hourly electrical meter data from 482 non-residential buildings from data from the *Building Data Genome Project*. The results illustrate that there is no *one size-fits-all* modeling solution and that various types of temporal behavior are difficult to capture using machine learning. This framework and methodology is designed to be a *baseline* implementation for other building energy data prediction methods developed by commercial providers or the wider research community. The benchmark data set can also be expanded with numerous other building performance data from a wider representation of buildings from around the world. The use of a baseline data set in future prediction research results in comparability and reproducibility of techniques in the built environment domain.

*Corresponding author email: clayton@nus.edu.sg, Phone: +65 81602452

Keywords: Building energy prediction, Building performance prediction, Performance prediction, Machine learning, Smart meters, Artificial neural networks, Support vector machines, Deep Learning

1. Introduction

Machine learning prediction models are highly impacting all facets of industry and science. They are being developed to diagnose illnesses, drive cars, suggest purchases to potential customers and mine the human genome. The built environment has the opportunity to leverage the same algorithms and techniques to improve efficiency and to create new business models [1]. Building performance analysis has dozens of uses for temporal prediction of electricity, heating and cooling energy. Prediction is often made both on the short-term (hours or days ahead) or long-term (weeks, months or years ahead). Short-term prediction are generally used for real-time HVAC control and efficiency of upcoming hours [2], scheduling and management of power stations and demand response schemes [3], and the analysis of residential metering and sub-metering [4], in addition to many other applications. Long-term prediction is used for the evaluation of energy conservation measures through a baseline model generation [5] and capacity expansion and planning [cite here]. Figure 1 illustrates the measurement and verification procedure using long-term energy prediction models. A period of baseline energy consumption is used to create a machine learning prediction model to evaluate how much energy a building would use in a *status quo* baseline mode. A energy conservation measure (ECM) is installed and the difference between the baseline is the avoided energy consumption or demand. This process is crucial for the implementation of energy savings implementations as it gives building owners, designers, and contractors a means of evaluating success of such measures.

1.1. Contemporary building energy prediction

An updated comprehensive review of building performance prediction studies is available that describes the models, techniques, input features, and uses for energy prediction in buildings [7]. This study reviewed research that implemented the most common prediction modelling techniques as applied to building energy prediction. These models include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Ensemble Methods, and various

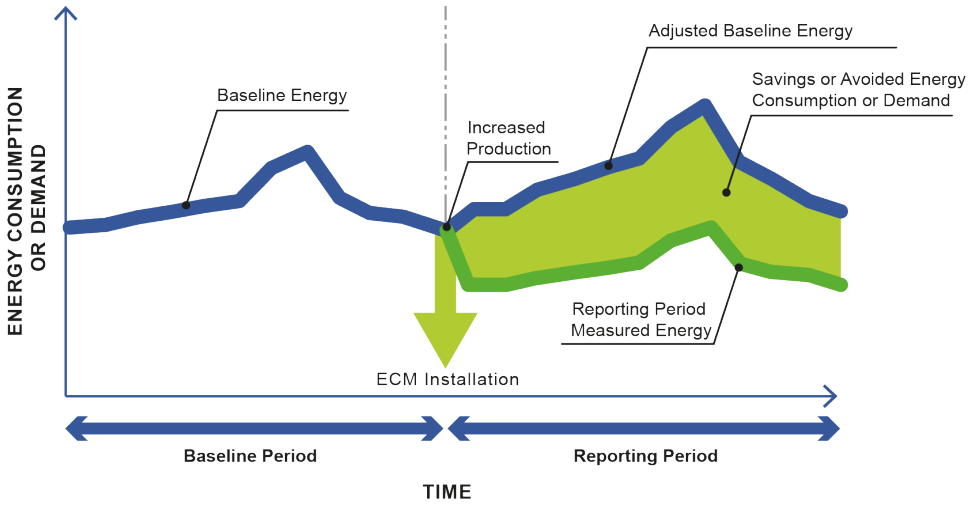

Figure 1: Prediction models as a comparison to energy savings interventions (Used with permission from the EVO - IPMVP [6])

other methods. A majority of the literature focuses on single building or a small set of buildings case studies.

The results of the studies with a low number of samples are problematic from a machine learning standpoint as the techniques are mostly applied to a single building or a small set of buildings. Training a machine learning model on such a small set of data results in a solution that is grossly over-fitted for a specific scenario. Such over-fitting results in models that are inappropriate to apply to data from the wider building stock. Over-fitting is a well-understood challenge in the machine learning community. This concept is related the bias-variance trade off that is illustrated in Figure 3. A model with high bias suffers from a lack of complexity to capture the behavior that is occurring in reality in a meaningful way. This model would be considered *underfitting* the data. On the other hand, a model that is overly complex suffers from high variance; this model is able to capture all the detailed behavior in the data that is used to train the model, but the model does not perform as well with new data sets or future data.

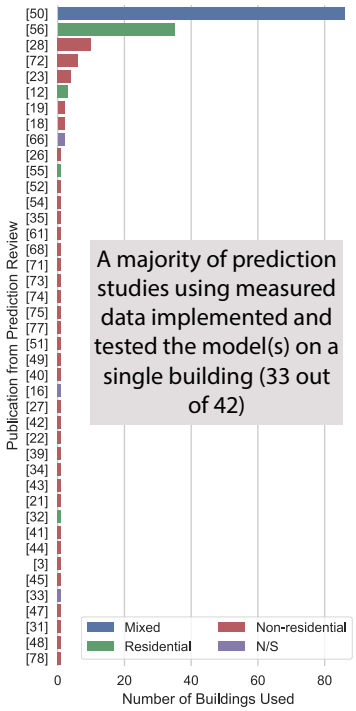

Figure 2: In the most recent prediction review study [7], a majority of publications developed and tested a machine learning modeling framework on data from a single building. These studies have the tendency to create very complex models for those particular buildings

Three key studies were completed in the last four years which finally took the process of machine learning for prediction to a diverse set of buildings to attempt to tackle the issue of overfitting. The first used 400 randomly selected buildings and applied six common, open prediction models for the purpose of creating a measurement and verficiation baseline [5]. The next study used 537 buildings and applied ten prediction models to further understand which model performed best among a large set of building [9]. The final study focused on the evaluation of what percentage of the building stock are appropriate to be used with *automated methods* as developed in the previous two studies [10].

Overall, these studies tested a large set energy prediction methods on a large set of buildings and this method was a step in the direction of general-

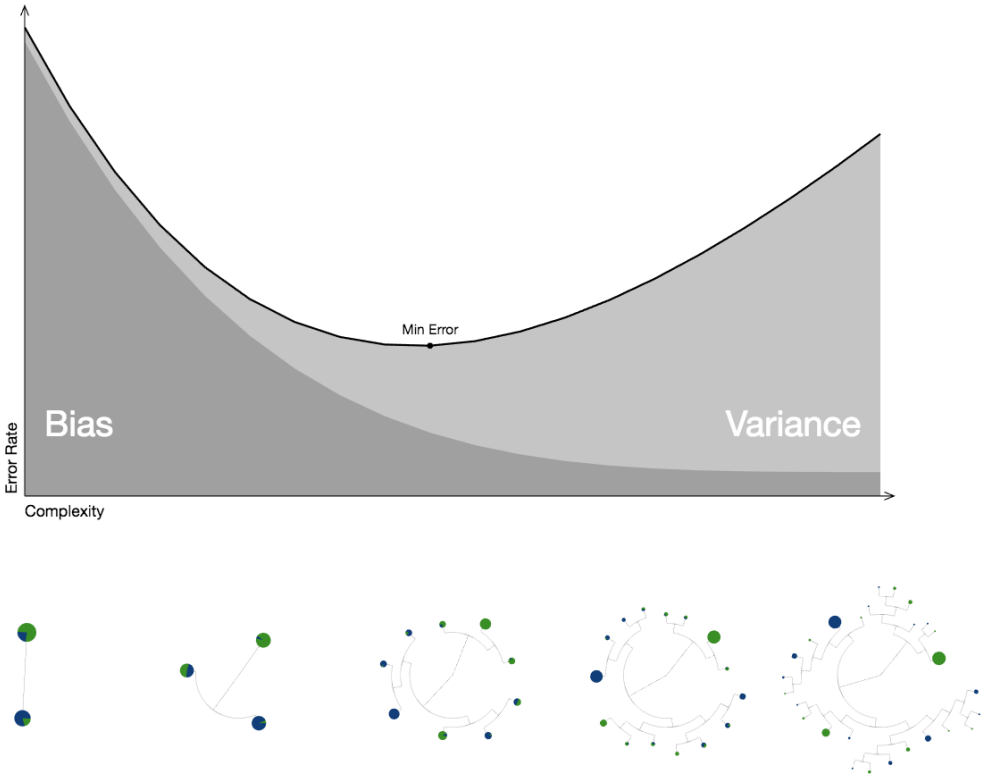

Figure 3: Bias-variance trade-off - The graph shows how as complexity increases the *bias* of the model drops as does the error rate. At a certain point the model becomes so complex that variance, or over-fitting becomes the key issue and the error rates increase again. The tree-based models below the chart show the increasing complexity of models of that type. Graphic adapted with permission from [8]

izability of energy prediction models. However, these studies are problematic due to the lack of access to the exact models or data the researchers used. Access to these aspects of their studies would allow future techniques to be compared directly through application of the old machine learning methods on new data or application of new machine learning methods on their old data. These studies give the community an understanding of what models work better than others in the context of the bias-variance trade-off, but they do not provide the ability to test new models and techniques.

1.2. The importance of benchmarking - A remedy for a crisis of over-fitting

Despite the advancement of machine learning and prediction for performance data for buildings, a major barrier to wide-spread dissemination is that the techniques are not easily reproducible. Engineers, data scientists and researchers should be really asking themselves *does my machine learning technique actually scale across hundreds of buildings? And is it actually faster or more accurate? How do we actually compare, each individual technique against previously created methods?*

The time-series data mining community identified this problem as early as 2003: "Much of this work has very little utility because the contribution made"..."offer an amount of improvement that would have been completely dwarfed by the variance that would have been observed by testing on many real world data sets, or the variance that would have been observed by changing minor (unstated) implementation details." [cite keogh]

Figure 4 illustrates the conventional way that machine learning is generally implemented in the built environment domain. Most of the existing building performance data science studies rely on each individual researcher creating their own methods, finding a case study data set and determining efficacy on their own. Not surprisingly, most of those researcher find positive results. However, the ability to compare those results to other publications and techniques is limited.

[here go into the efforts in thi benchmarking in computer science and machine learning]

Using a large, consistent benchmark data set from hundreds (or thousands) of buildings, a researcher can determine how well their methods actually perform across a heterogeneous data set. If multiple researcher use the same data set, then there can be meaningful comparisons of accuracy, speed and ease-of-use. The purpose of this paper is to establish an example of machine learning benchmarking for the purposes of energy forecasting and prediction.

The creation of benchmarking data sets and methods sets the research community towards an environment in which there is a measure of accountability in the claims made by new algorithms. These new techniques will likely be implemented on the test case(s) developed in the course of a research study, but will need to also be applied as much as possible to benchmarking methods in order to show which type of improvement is occurring in the literature and the quantification of improvement.

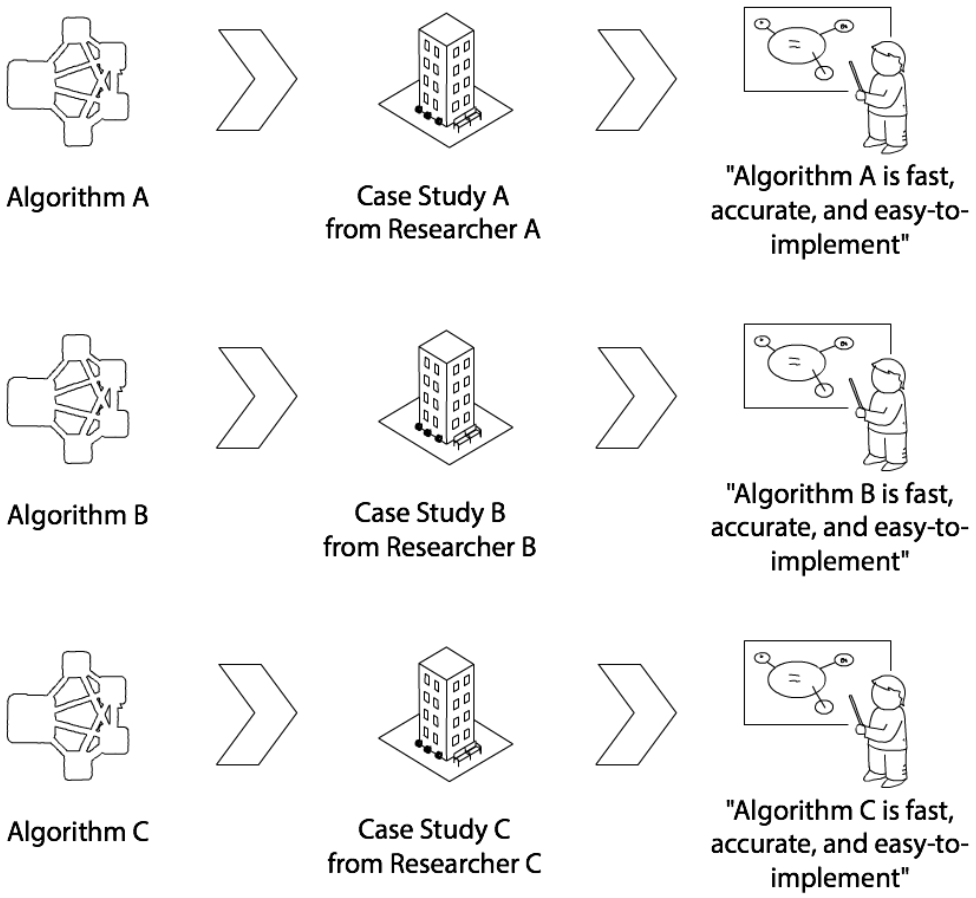

Figure 4: Old anecdotal machine learning testing and implementation methodology that results in a lack of comparative opportunities for new techniques

1.2.1. The original benchmark - Great Energy Predictor Shootout I and II

In the non-residential building research domain, there is a single set of examples of a benchmarking data set that was utilized for several machine learning studies. This set is the *Great Energy Predictor Shootout* competitions held in the early 1990's. The first competition included the use of a single building's electrical, cooling, and heating meters in a competition where the participants were asked predict a single month of data using three months of training data [11]. This competition resulted in several publication based on the top set of performers in the competition. A second competition

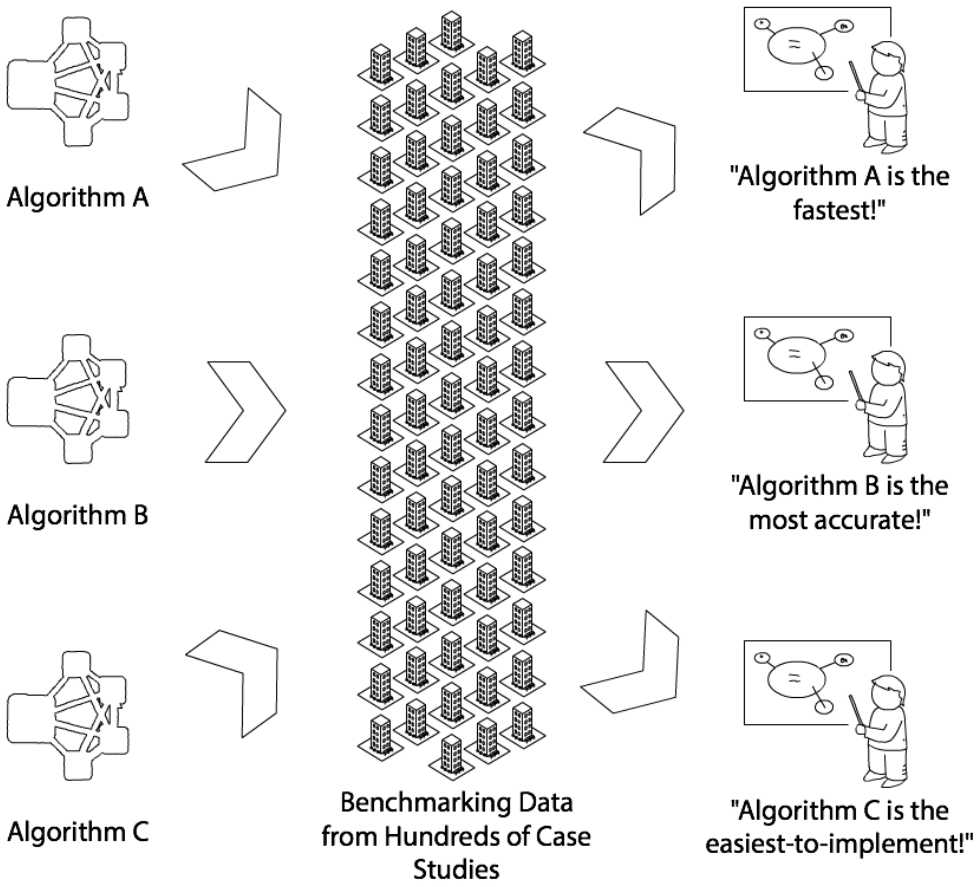

Figure 5: Using a common bench-marking data set enables the comparison of algorithms developed in the research community

was held, the aptly-named *Great Energy Predictor Shootout II* that asked participants to predict the energy savings for a energy savings project [12]. These data sets have been since been used as a benchmarking comparison data set on two studies; one focused on residential modelling using SVM [ref] and the other to predict consumption of office buildings in Greece [ref].

1.3. Tackling the over-fitting problem - open benchmarking data sets and open models

This paper outlines a demonstration of how benchmarking of performance prediction models can be accomplished on an open data set using open-source prediction techniques. Initially, a review of prediction considerations for building energy performance is reviewed. Next, a framework for implementation is outlined using the Building Data Genome Project data set. The results are showcased and interpreted in the context of the different primary use types and operations considerations. Finally, a discussion of the limitations and guidelines for implementation of future techniques is presented.

2. Review of building energy prediction input considerations

To initiate the prediction model discussion, an overview of the conventional energy performance prediction considerations is covered in this section. These aspects of machine learning-based modelling for buildings are most prominent in non-residential buildings such as offices, educational facilities, laboratories, and health-care. These categories are the key considerations when developing prediction models for buildings.

2.1. Daily, weekly and seasonal schedules

Buildings operate on several types of schedules. A majority of non-residential building have *occupied* and *unoccupied* periods that usually coincide with daytime and nighttime. These diurnal cycles are generally one of the best indicators of the building use type - i.e.: offices are open from 9am to 6pm and hotels are most active from 6pm to 10am. Most buildings also have a weekly schedule; the default being certain behavior on weekdays and a different behavior on weekends. Finally, many buildings have seasonal changes such as when educational buildings have certain behavior during a regular session versus during breaks or holiday seasons. The concept of scheduling in non-residential building is most often related to the predefined schedules programmed into the automation systems in the building, unlike the human behavior that is discussed later. These schedules are determined usually by the operations and maintenance policy or by the energy management group within an organization. Many buildings have very predictable schedules, while others are much more volatile. Modelling such behavior can often be done using time-series methods that find auto-correlation behavior or by using date/time features as inputs to prediction models.

[insert the references from the temporal mining - uses dayfilter and other promienent daily and weekly clustering]

2.2. Human behavior

The concept of *human behavior* as an influence is similar to the previously-discussed schedules, however there is a more stochastic element to these behavior. Buildings that are more influenced by occupant behavior generally have demand response-based control systems that use sensors, cameras or other detection methods to modulate systems only when humans are present or using the space for a specific purpose. Sometimes humans even have the ability to control spaces using various types of interfaces with the building, although this is less common in non-residential buildings. Modelling occupant behavior is considered more complex than schedule as human behavior can be less systematic, thus auto correlation-based time-series methods are less effective.

2.3. Weather

A major energy consuming component for most buildings are heating, ventilation, and air-conditioning (HVAC) systems. Intuitively, these systems tend to use more energy as the outdoor conditions get hotter (in the case of cooling) or colder (in the case of heating). Often this relationship is linear, but it can also be non-linear based on the HVAC system type and operation policy. The degree of influence of weather varies greatly among the building stock. It is influenced by the percentage of internal load vs. envelope-based load, HVAC system type, climate, and other factors.

2.4. Non-routine events

Non-routine events are disruptions in the systematic operation of a building due to events such as a change in the operational schedule of the building, equipment breakdowns, and events or human behavior that is highly irregular as compared to past behavior. These events could be planned or unplanned by the operations and facilities management staff. Non-routine events are hard to predict with conventional input variables, but models can be tuned to quickly adapt to the change using approaches related to change-point prediction or other types of behavior change detection techniques.

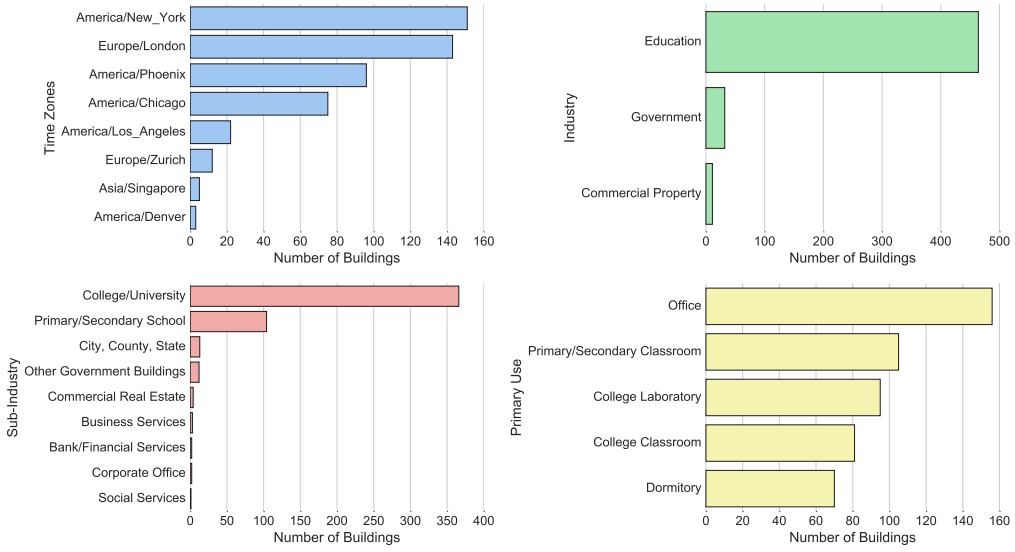

Figure 6: Meta-data breakdown of the *Building Data Genome Project* benchmarking data set used in this publication. Used with permission from (cite)

3. Benchmarking open source prediction methods on open building energy data

The key to this benchmarking analysis is to use a diverse enough data set to illustrate the benefits of scalabiliy and generalizabiliy amongst the building stock. The regression testing framework for this paper is outlined in this section. The open hourly data from the Building Data Genome Project is used in this paper as a starting point. This data set includes data from over 500 buildings, mostly from educational institutions. Figure 6 illustrates the breakdown of four meta-data points from this data-set.

3.1. Machine learning input variables

The input variables available as independent predictors of energy consumption are outlined in this section. Table 1 outlines the basic set of machine learning input variables used in this benchmarking approach.

3.2. Training and test data set scenario

For the purposes of the comparison of various open source techniques, a simplified training and testing scenario is utilized for the comparison. One

Category	Variable	Behavior Targeted
Temporal	Time of Day	Daily schedules
	Day of Week	Weekly schedules
	Public Holiday Schedule	Holidays
	Schedule Type	Seasonal schedules (summer, etc.)
Weather	Outdoor Air Temp.	Sensible heating and cooling
	Outdoor Air Hum.	Latent heating and cooling
Meta	Industry Sector	General category of use
	Primary Use Type	Specific category of use
	Floor area	Size of building
	Number of floors	Height of building
	Climate zone	Type of climate
	Maximum occupancy	Total number of people
	Cooling system type	Typical cooling efficiency
	Heating system type	Typical heating efficiency
	Performance rating	Comparison to its peers

Table 1: Independent input features used in this benchmarking process

year of whole building hourly electrical meter data is available for all 482 buildings. Four different training and testing data scenarios are tested as seen in Figure 3.2. Scenarios 1-3 are made up of 3, 6, and 9 month training windows and a 3 month continuous testing window. Scenario 4 is made up of 3 month training windows followed by a 1 month test window that repeats itself three times. These scenarios provide a certain level of cross-validation that is realistic in the context of medium to long-term energy prediction in the built environment.

3.3. Open source regression models from Sci-kit Learn Python Library

The regression model catalogue from the Sci-kit Learn Library is used in this benchmarking process. We use these models as a starting point for the benchmarking process as many of these models have been developed for decades and are some of the most often used prediction models in the machine learning community. The larger array of more advanced models (e.g.: deep learning, specific energy-focused prediction models) is left outside the scope of this paper as these will be the *improved techniques* that the wider research community would likely be testing.

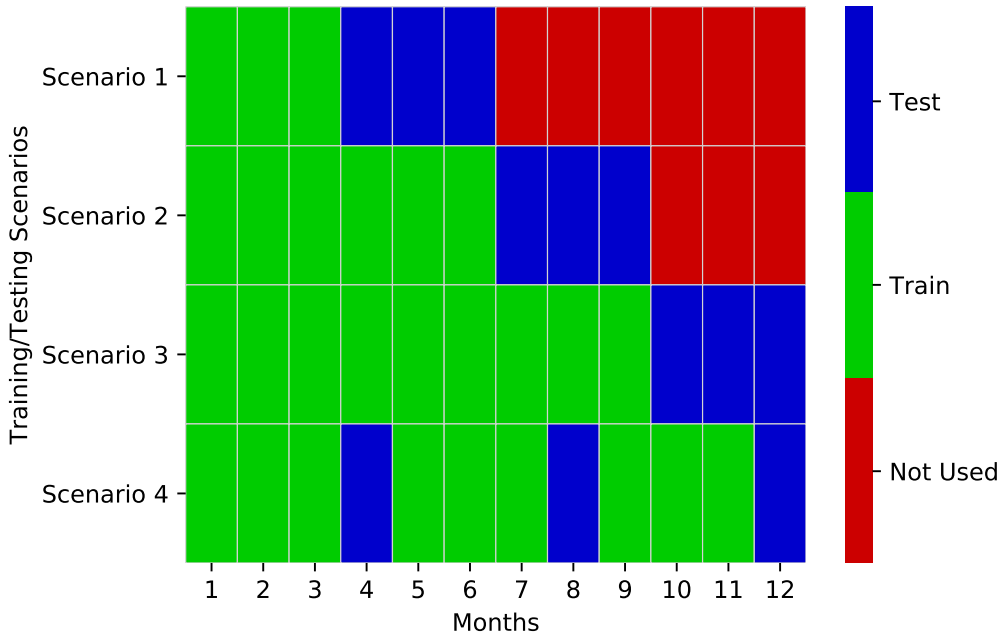

Figure 7: Overview of the four training/testing scenarios

This study focuses on general forecasting methods as a foundation for comparison for more building domain specific regression or prediction models. These models do not take into consideration the auto-correlation aspect of prediction or the building context-specific nature of the built environment.

3.4. Accuracy metrics

The three metrics used in this analysis to evaluate model fit are the MAPE, NMBE, and CVRSME metrics.

4. Results

4.1. Model fit overview

Figure 8 illustrates 12 Sci-Kit learn models applied to all the building use types using the MAPE and CVRSME metrics. In general, laboratories have the highest accuracy across all of the models and in general. This situation is due to laboratories being the most *systematically schedule-driven* of all

the building types. Laboratories often have large equipment that is operated continuously or in set time schedules throughout the course of an entire year. University classrooms and offices behave in similar ways across the models as these two use types are often similar and many of these buildings are mixed use types. University dormitories have the fourth highest accuracy for most of the models, however they are better fits for models such as the Huber Regressor or the TheilSen Regressor. Finally, the Primary School buildings are the worst performers among the building use types. These buildings are dependent more on human behavior within their annual schedule phases, resulting in the lack of predictability using the methods and models outlined in this simple example.

Appendix A provides a much more detailed breakdown of each of the building use types according to all three error metrics as well as an analysis of each of the four training/testing scenarios.

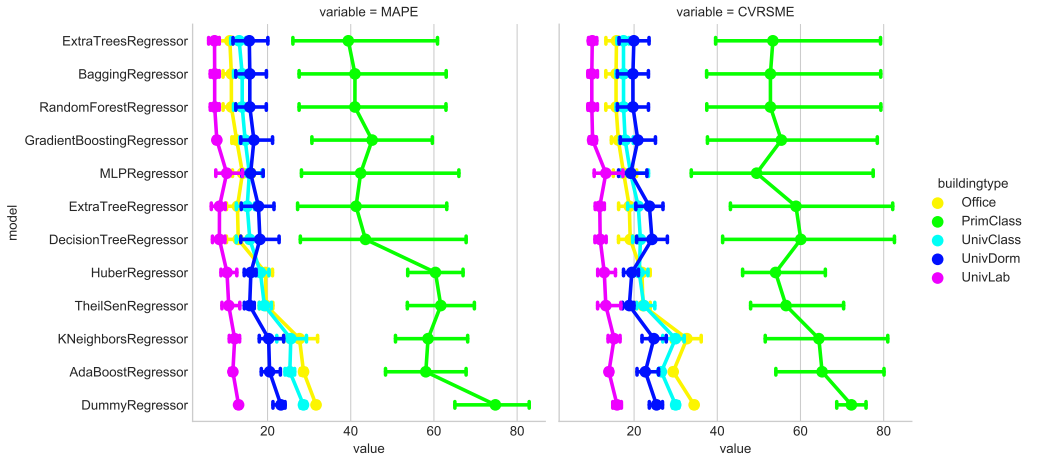

Figure 8: Overview of models fit metrics MAPE and CVRSME for each building use type across 12 Sci-kit Learn models

5. Conclusion

This publication gives readers a simple example of mainstream prediction techniques applied to a large, diverse data set. This example is designed to illustrate how future techniques can and should be applied to open data sets.

5.1. Limitations

The key limitation of this analysis is congruent with the limitation of machine learning in general: the models developed and the resulting metrics are limited by the range of training data utilized to build the models. The conclusions of the model comparison in this paper are only relevant to office, laboratories, classrooms, and dormitories from university campuses in the context of the geographical and operational environments of the Building Data Genome Project data set. Thus, the primary goal of this analysis is not to be comprehensive in capturing the behavior of the building stock, but to provide an example and data set that can be built upon with a more diverse set of data. *It is crucial that additional data from thousands (or millions) of other buildings are added over time for the methodology to achieve the main goal of a comprehensive benchmarking process.*

New algorithms coming into the public domain could be tested against this growing set of building data sets to quantify what improvements are being made in the domain.

5.2. Open Source Repository of Data and Implementation

To keep in the spirit of reproducibility, this publication is fully reproducible using the Building Data Genome Project.

5.3. Future work: the Great Building Energy Predictor Shootout 2019

A machine learning competition is in the planning phase to further extend the benchmark data set to over 4000 energy data streams from 1200+ buildings from around the world. This competition will put potentially thousands of machine learning and building systems experts head-to-head to come up with the best set of methods to predict.

Key areas of potential improvement include:

1. Next generation time-series features - using the autocorrelated patterns in more advanced ways
2. Multi-variate learning - using a building's *peers* to predict future behavior
3. Advanced models - using deep learning to leverage new and diverse types of behavior without feature engineering
4. Advanced ensembles - a voting ensemble that could significantly improve prediction for different building types
5. White or Gray Box model convergence - using physics-based models to inform the machine learning algorithm

6. References

[1] A. Agrawal, J. Gans, A. Goldfarb, Prediction Machines: The simple economics of artificial intelligence, Harvard Business Press, 2018.

[2] D. M. Solomon, R. L. Winter, A. G. Boulanger, R. N. Anderson, L. L. Wu, Forecasting energy demand in large commercial buildings using support vector machine regression, Department of Computer Science, Columbia University, Tech. Rep. CUCS-040-11 (2011).

[3] C. E. Borges, Y. K. Penya, I. Ferńandez, J. Prieto, O. Bretos, Assessing tolerance-based robust short-term load forecasting in buildings, Energies 6 (2013) 2110–2129.

[4] R. K. Jain, K. M. Smith, P. J. Culligan, J. E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Applied Energy 123 (2014) 168–178.

[5] J. Granderson, P. N. Price, D. Jump, N. Addy, M. D. Sohn, Automated measurement and verification: Performance of public domain whole-building electric baseline models, Applied Energy 144 (2015) 106–113.

[6] Efficiency Valuation Organisation, International performance measurement and verification protocol, Technical Report EVO 10000-1;, 2012.

[7] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, 2018.

[8] T. C. Yee, Stephanie, Model Tuning and the Bias-Variance Tradeoff, ????

[9] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, S. Fernandes, Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings, Applied Energy 173 (2016) 296–308.

[10] J. Grandersona, S. Touzani, S. Fernandes, C. Taylor, Application of automated measurement and verification to utility energy efficiency program data, Energy and Buildings 142 (2017) 191–199.

[11] J. F. Kreider, J. S. Haberl, Predicting hourly building energy use: The great energy predictor shootout – Overview and discussion of results (????).

[12] J. S. Haberl, S. Thamilseran, Great energy predictor shootout II: Measuring retrofit savings–overview and discussion of results, Technical Report, American Society of Heating, Refrigerating and Air-Conditioning Engineers˜. . . , 1996.

Appendix A. Detailed Model Comparison Breakdowns

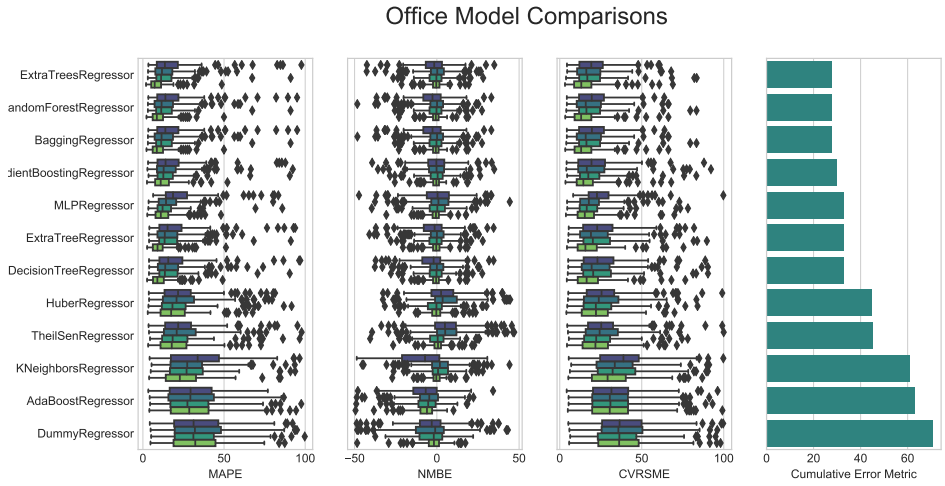

Figure A.9: Detailed Breakdown of Benchmarking Models on Office Buildings

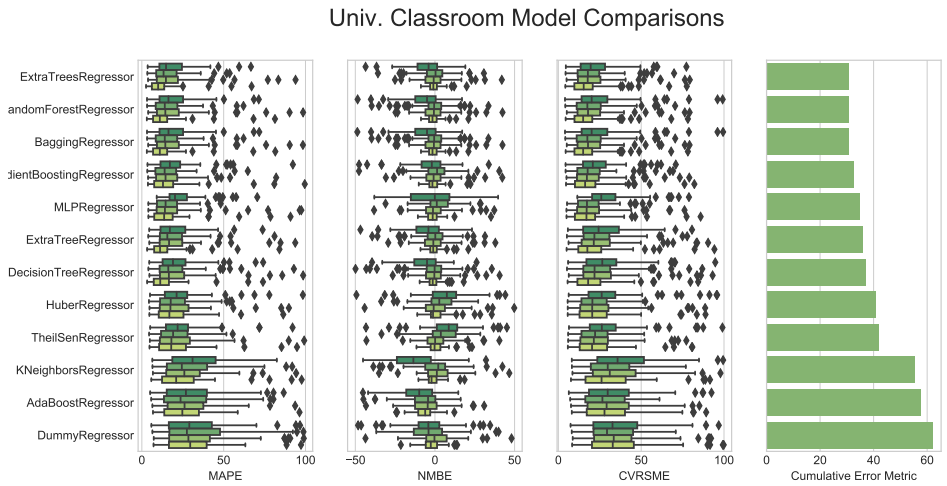

Figure A.10: Detailed Breakdown of Benchmarking Models on University Classroom Buildings

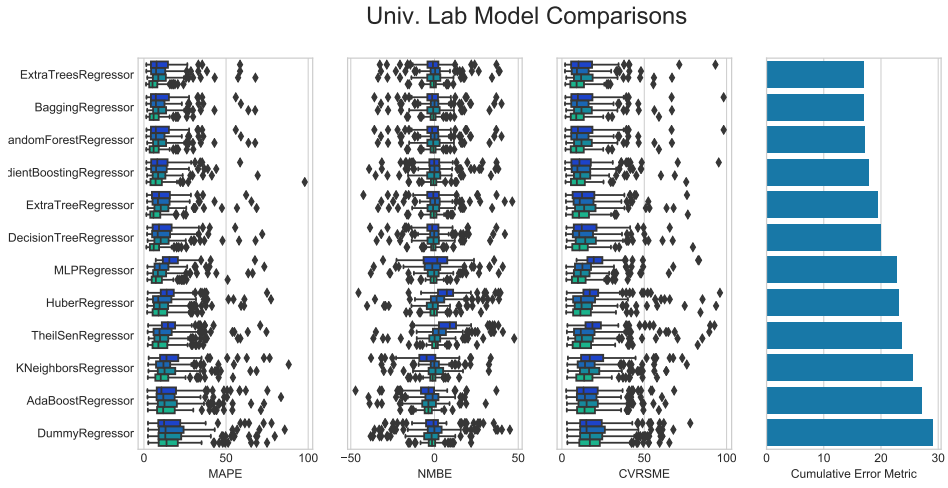

Figure A.11: Detailed Breakdown of Benchmarking Models on University Laboratory Buildings

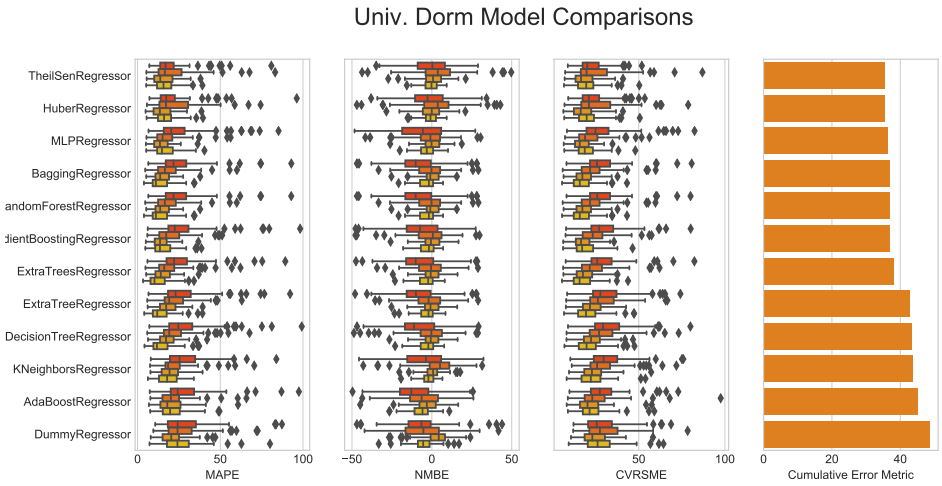

Figure A.12: Detailed Breakdown of Benchmarking Models on University Dormitory Buildings

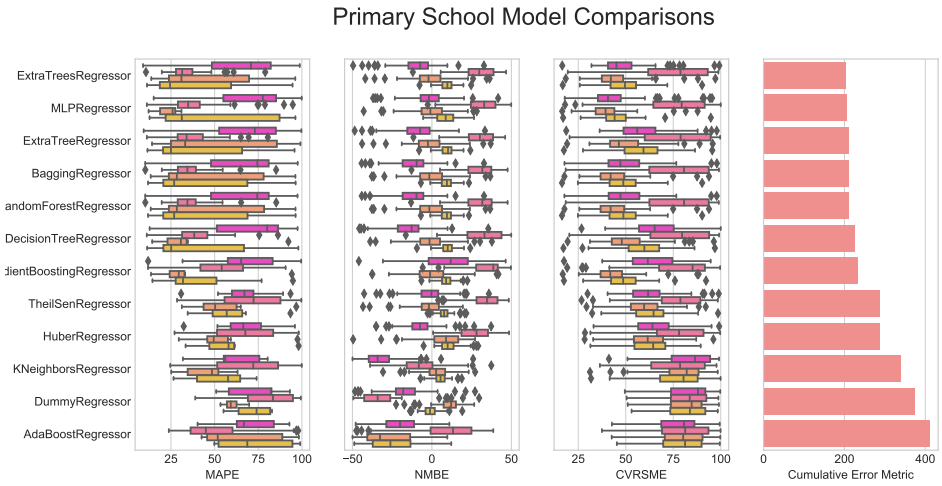

Figure A.13: Detailed Breakdown of Benchmarking Models on Primary School Buildings