Opening the Black Box: Explainable AI for Greater Energy Savings

Biodun Iwayemi, PhD., Jeffrey Hullstrung, Nick Neverisky, VEIC, Winooski, VT
Dan Fredman, PhD., Oracle Utilities Opower, Arlington, VA

ABSTRACT

Regression model interpretability is essential in regulated environments such as energy efficiency. This restricts the use of un-interpretable, black box machine learning strategies like gradient-boosted regression trees, even though they significantly outperform ordinary least squares regression methods on hourly or 15-minute interval data. Explainable artificial intelligence (XAI) methods promise the best of both worlds: highly accurate and interpretable models.

To determine if XAI can make advanced regression models accessible to energy analysts, facility managers, and regulators, this paper explores their use with gradient-boosted regression trees for building energy models, using data from several large grocery stores in the American Northeast. The authors convened focus groups to evaluate different XAI charts to determine which visualizations effectively communicate energy modeling results to energy efficiency utility staff and customers. The paper highlights potential pitfalls in using XAI methods and provides recommendations and a list of best practices for using XAI for communicating model results to stakeholders—to drive wider adoption of these methods. The analysis team has found that the promise of XAI for energy efficiency is real, and recommends ways to assess XAI feasibility for use with energy efficiency programs.

Introduction

Machine learning is the process by which computer algorithms learn patterns and relationships from large amounts of data, and then use the learned patterns to create models that make predictions on new or unseen data. Data professionals are using machine learning to such an extent that it is being woven into the fabric of everyday life and work. These developments come with benefits and challenges. Their power and flexibility drive their use for many purposes, ranging from mundane tasks like asking a voice assistant to turn off living room lights, to extremely complex actions like automatically tagging pictures of friends and family on social media, helping autonomous vehicles navigate city roads while avoiding pedestrians, or translating text or speech from one language to another.

These more powerful models are capable of breathtaking accomplishments, but with their rapidly growing capabilities come increasing model complexity and opacity into how these models make their decisions. They are black boxes capable of giving very accurate answers, but opacity is a challenge in regulated sectors like energy efficiency, where program evaluators prefer interpretable and explainable models. In these sectors, it is important to have correct answers to evaluation queries, and to know what factors influenced the decisions or outputs of a model. This is especially important when the models learn spurious relationships, contain hidden biases, or fail in unexpected ways. Advanced and powerful machine learning models like gradient-boosting trees or deep learning methods cannot be widely adopted in such sectors unless users find them trustworthy. That trust hinges on our ability to explain or understand how these models make their decisions.

Rather than stay with the status quo and default to using less powerful but more explainable machine learning models, a research team at VEIC set out to learn if and how XAI strategies can eliminate the trade-off between model complexity and interpretability. Addressing this issue has allowed the team

to gain the best of both worlds: The team could use the most powerful and accurate black box models, and could explain the correlations they learned or why they made a particular prediction.

Interpretable models promise multiple benefits, the most important of which making it easier to adopt cutting-edge machine learning models for energy modeling at customer facilities. Achieving these outcomes requires demonstrating that these models provide consistently superior performance than conventional models, and that their outputs can be easily communicated to energy modelers, evaluators, and customers in simple but more insightful ways than can the current visualizations. To that end, the researchers posed the following questions:

- Are black box models ready for broader adoption in regulated environments?
- To what extent do advanced, black box machine learning models such as gradient-boosting machines (GBMs) outperform ordinary least squares (OLS)-based time-of-week and temperature (TOWT) regression models?
- What are the best methods and charts, using these machine learning models, for effectively and simply communicating the outputs of a facility's energy model? And how comprehensible are those outputs to a broad audience with varying levels of visualization expertise, so that the audience can make well-informed, defensible energy efficiency decisions?

The following sections offers an overview of the evolution of regression modeling for whole-building energy modeling; describes the research team's dataset and modeling methods; compares the performance of TOWT methods against black box GBMs; demonstrates how the SHapley Additive exPlanations (SHAP) library can explain the predictions and associations learned by a GBM; discusses feedback from focus groups on the most effective visualizations in explaining models to as broad an audience as possible; and highlights potential pitfalls from using predictive models to infer causal relationships.

Background

Machine learning is defined as "a set of methods that computers use to make and improve on predictions of behaviors based on data" (Molnar 2020). It is used in voice assistants; predictive text in mobile messaging and email applications; computer vision for autonomous vehicles; and other such applications. At a minimum, machine learning comprises models used for classification—that is, labeling or categorizing things; regression (predicting numerical values—for example, forecasting temperature); and clustering (grouping similar items). This paper addresses regression tasks and thus is restricted only to regression modeling algorithms.

A Brief, Opinionated Overview of Machine Learning's Evolution for Whole-Building Energy Modeling

The range of regression algorithms for whole-building energy modeling is relatively broad: OLS regression methods such as the seminal PRInceton Score-keeping Method (PRISM) model (Fels 1986), piece-wise regression models like the Lawrence Berkeley National Laboratory's TOWT models (Mathieu 2011), and gradient-boosted regression tree methods like those described in Touzani et al. (2018).

OLS regression. Regression modeling is the workhorse of building energy modeling. From the development of the PRISM model in 1986 to multivariate regression models with large numbers of parameters, OLS regression has been the premier regression modeling algorithm used by engineers and evaluators in the energy efficiency sector. The reasons are obvious: It is simple, interpretable, and familiar. Unfortunately, these methods deliver results far from the state of the art. And even though they provide

acceptable performance, their underlying assumptions (linearity, normally distributed data, homoscedasticity, and so on) are often violated.

TOWT regression. The Berkeley Lab developed TOWT modeling. It is the designated modeling process for hourly meter data in CalTRACK methods,[1] an industry-wide framework for measurement and verification in estimating avoided energy use from installed energy efficiency measures. TOWT modeling applies a continuous, piece-wise linear regression with 5 temperature segments, or bins. It introduces categorical variables for hour-of-week (168 unique values), day type (regular or holiday), and occupancy. TOWT has been VEIC's primary whole-building energy modeling method since 2018. VEIC has used it for pay-for-performance projects, strategic energy management, and other energy efficiency strategies. TOWT is more flexible than traditional change point models, provides valid and reliable results, and can model non-linearities in the data. The only challenge with this model is that it is difficult to explain to customers and stakeholders, because there are more than 170 parameters in the modeling equation. That is, because time-of-week is one of the indicator variables, and accounts for 168 different values, any other variables add to that basic census of 168.

GBMs. Gradient-boosting machines are an example of ensemble models. They combine "weak" learners—that is, mediocre models that perform slightly better than a coin flip or a 50 / 50 decision—to create a powerful, accurate machine learning model. The key to their accuracy is that each weak learner learns from the mistakes of its predecessor via a process called *boosting.* This creates successively better results. When decision trees are used for the weak learners, the resulting model is called a gradient-boosted regression tree (GBRT). These models have provided excellent results in whole-building energy modeling (Touzani 2018 and Severinsen 2019). But as with TOWT, their performance comes at the expense of model explainability, even more so than with TOWT. With TOWT analysts can examine a model equation, but that is no longer available in GBRTs. They might perform better than TOWT models, but does the increase in performance outweigh their lack of interpretability? Is there a way to have the best of both worlds—to use the most powerful methods available, yet be able to examine what the model has learned and explore the associations in the data that are driving model predictions?

Methods

The VEIC team evaluated TOWT and GBRT models on eight grocery stores in the Northeast. The average store size is 50,000 square feet, and the total annual energy consumption for all stores was approximately 14,000 MWh. The team had analyzed these stores for large multi-measure projects using Efficiency Vermont's[2] standard weather normalization protocols. With the available baseline data, the team knew those stores would be good candidates for this new approach.

Model Development and Evaluation

The mark of a good predictive model is its ability to generalize to new, unseen data. The best ways to achieve this are by cross-validation or out of sample testing. Each requires splitting the data into at least a training set and a test set. The model "learns" on the training set, and is evaluated on the test set. The VEIC team used a 70:30 split between training and test sets, and stratified the data in each set to ensure that all the seasons of the year were represented in each group. The team also chunked the smart meter interval or AMI[3] data into daily groups of 96 contiguous 15-minute readings, to maintain the auto-

[1] See www.caltrack.org, for access to the methods.
[2] VEIC operates Efficiency Vermont, a statewide, regulated energy efficiency utility.
[3] Advanced metering infrastructure (or smart meter) data in 15-minute or hourly intervals.

correlation between adjacent measurements, rather than adopting a naïve, random selection of time series data. Figure 1 shows that the training and test sets are similar in structure and trends, confirming that both groups are representative of the original data.

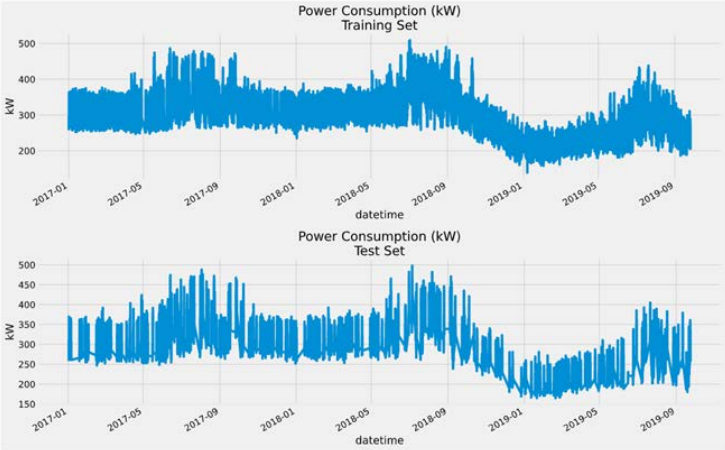


Figure 1. Comparison of training and test set data on the power consumption of a sample building.

The team evaluated the implemented GBRTs, using the Python version of LightGBM[4], Microsoft Research's open-source GBM software library. LightGBM (Ke et al, 2017) is significantly faster than XGBoost[5], the premier open-source GBM library, while providing comparable accuracy, and lower memory requirements. LightGBM also shares XGBoost's ability to work well with large datasets. We used the following features as inputs to our LightGBM model:

- Temperature (outdoor dry bulb)
- Relative humidity (outdoor)
- Hour of day (categorical, 24 possibilities)
- Day of week (categorical, 7 possibilities)
- Month (categorical, 12 possibilities)
- Day type (categorical, 2 possibilities: regular or holiday)

Evaluation Metrics

The team compared GBRTs against a TOWT benchmark, using the coefficient of determination or R^2 (the proportion of the variance accounted for or explained by the model) and the coefficient of variation of the root-mean-square error, CV-RMSE[6] as the team's metrics.

Results

The VEIC team found that GBRTs consistently outperformed TOWT on the grocery store dataset. Of the 8 stores, the GBM model had 15 percent lower CV-RMSE (that is, higher precision) on the test set, and a 3.7 percent increase in R^2 (greater accuracy), compared to the TOWT method. Tables 1 and 2 provide a breakdown of model performance across the 8 stores.

[4] The LightGBM framework can be found at https://github.com/microsoft/LightGBM.
[5] The XGBoost library: https://github.com/dmlc/xgboost
[6] CV-RMSE is a measure of the difference between a model's predicted results and ground truth.

Table 1. LightGBM vs. TOWT model performance comparison in stores 1 -4

Parameter			Percent difference
	Store 1		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.0%	4.5%	
R^2 on training set	94.8%	88.0%	
CV-RMSE on test set	3.4%	4.3%	22%
R^2 on test set	92.0%	87.0%	6%
	Store 2		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.5%	8.0%	
R^2 on training set	97.5%	87.2%	
CV-RMSE on test set	5.6%	7.7%	28%
R^2 on test set	93.4%	87.4%	7%
	Store 3		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.8%	5.7%	
R^2 on training set	92.6%	83.2%	
CV-RMSE on test set	4.4%	5.3%	18%
R^2 on test set	89.3%	83.9%	6%
	Store 4		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.4%	5.1%	
R^2 on training set	96.3%	91.5%	
CV-RMSE on test set	3.9%	4.5%	13%
R^2 on test set	94.4%	92.7%	2%

Table 2. LightGBM vs. TOWT model performance comparison in stores 5 - 8

Parameter			Percent difference
	Store 5		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.1%	4.3%	
R^2 on training set	95.3%	91.0%	
CV-RMSE on test set	3.3%	3.5%	6%
R^2 on test set	94.2%	93.4%	1%
	Store 6		
Model type	GBRT	TOWT	
CV-RMSE on training set	3.3%	4.2%	
R^2 on training set	96.8%	94.7%	
CV-RMSE on test set	3.8%	4.1%	7%
R^2 on test set	95.5%	94.8%	1%
	Store 7		
Model type	GBRT	TOWT	
CV-RMSE on training set	2.2%	3.0%	
R^2 on training set	96.9%	94.6%	
CV-RMSE on test set	4.2%	4.3%	2%
R^2 on test set	88.7%	88.2%	1%
	Store 8		

Parameter			Percent difference
Model type	GBRT	TOWT	
CV-RMSE on training set	2.6%	3.9%	
R^2 on training set	97.6%	94.7%	
CV-RMSE on test set	3.1%	3.7%	16%
R^2 on test set	96.4%	94.9%	2%

To get a complete view of model performance, it is important to look beyond aggregate metrics such as R^2 and CV-RMSE, and create diagnostic plots that show the distribution of the data and model predictions. The VEIC team used two diagnostic plots to do that: a prediction error plot (Figure 2) and a residuals plot (Figure 3). The team generated them both with Yellowbrick[7], a visual diagnostics tool for models compatible with Scikit-Learn[8], a machine learning library in Python. The prediction error plot helped us to quickly determine if there were outliers in the data or bias in the model (did the model consistently under- or over-predict a value?), while the residuals plot showed us whether or not the training and test set residuals had Gaussian distributions. The diagnostic plots do not show anything abnormal, so the team is confident that the trained models were good.

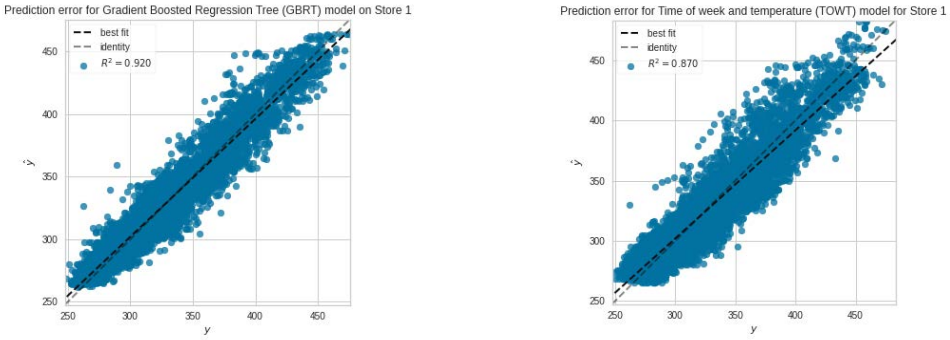


Figure 2. Comparison of prediction error between GBRT and TOWT models on Store 1.

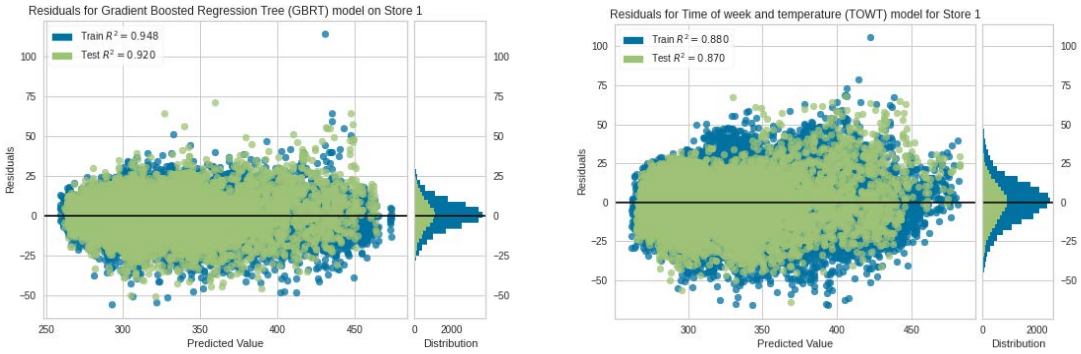


Figure 3. Distribution of residuals for GBRT (left) and TOWT models (right).

[7] District Data Labs has created a Github site for Yellowbrick, at https://github.com/DistrictDataLabs/yellowbrick.
[8] Scikit-Learn https://scikit-learn.org/

So, What Is Driving These Great Predictions?

The team found that both TOWT and GBRTs provide excellent predictions of future energy use, but they remain black boxes that lack both the elegant simplicity and the interpretability of OLS regression models. Black box results are hard to explain to end users or facilities managers. OLS regression models can provide customers with simple equations that relate model inputs to model outputs. The coefficients of these equations show which inputs have the greatest impact, as well as their direction (either increasing or decreasing energy use). How can these features be replicated with black box models? How can end users understand which model inputs have the greatest effect on the output (predicted energy use), and to what extent? And why did a model make a particular decision?

Explaining Machine Learning Models using SHAP

One of the most popular strategies for explaining machine learning model predictions is SHAP[9] (Lundberg and Lee, 2017). This framework was developed by researchers at the University of Washington, and is based on Nobel-prize winning work on cooperative game theory by Lloyd Shapely. SHAP takes a compatible machine learning model that has been trained on a specific dataset, and creates a post-hoc explanatory model for it, allowing us to fairly attribute impacts to model inputs. It shows how the value of each input parameter drove the prediction up or down.

For example, Figure 4 demonstrates how SHAP is applied to a machine learning model used for loan predictions and shows the correlations the model has learned. In this example, the applicant's job status as an employee at a startup (lowest red bar) negatively affects the likelihood of loan approval, whereas the checking account balance (long blue bar) increases the likelihood of loan approval.

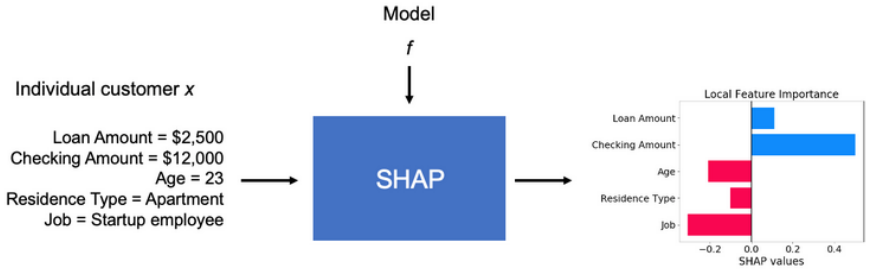


Figure 4. SHAP attribution of effects from model inputs.[10]

SHAP creates two types of explanations: local and global models. Figure 4 is an example of a local model and provides an explanation for single prediction. Global models provide explanations over multiple rows / readings, and show the correlations learned over the entire dataset. The team used them to visualize what the model has learned over a quarter of a year, or over a full year. Figure 5 shows a global model, whereas Figure 6 is a bee-swarm version of a global chart. Figure 5 indicates that temperature has the greatest effect on predicted power draw, increasing the predicted power over the average value by 41.34 kW, whereas season has the smallest effect. Figure 6 shows how individual local models can be combined to create a global model with a comprehensive view of how different input values influenced the models' predictions.

SHAP visualizations allowed the team to open the black box and explain how model inputs influenced predictions. But an open question was how easy and interpretable these charts might be for a

[9] The software package is available at: https://github.com/slundberg/shap

[10] A fuller explanation of how SHAP and another model, SAGE, can explain machine learning can be found at Ian Covert's website: https://iancovert.com/blog/understanding-shap-sage/.

wider audience. What are the most effective charts for communicating modeling output to a broad range of stakeholders? The team used focus groups to answer these questions.

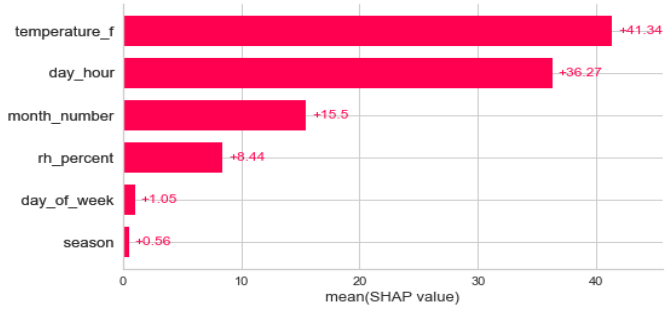


Figure 5. Global SHAP model: Correlations learned across the entire dataset.

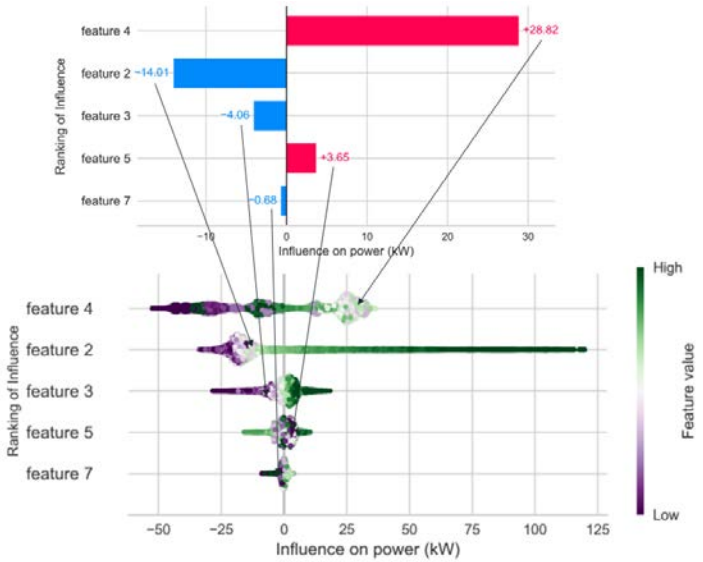


Figure 6. Mapping from local to global models.

Evaluating Chart Effectiveness and Interpretability, Using Focus Groups

The high-level goal of the focus groups was to identify methods for communicating outputs of the machine learning algorithms to several audiences with experience in energy management and modeling. Sub-goals were to identify:

- Chart types that could effectively communicate inputs correlated with increased/decreased power use
- Chart elements that might aid in communication
- Chart elements that caused confusion

The focus groups represented VEIC's marketing, engineering, customer service, and account management departments (Figure 7), with 21 participants spread across two groups. Two focus group sessions involved asking participants to play the role of facility managers participating in an Efficiency Vermont program that would help them better understand how different factors might influence energy use at their facilities. The team presented participants with local and global SHAP model visualizations. Local models represent factors correlated with increased (or decreased) power demand for a specific snapshot in time

(e.g. outdoor air temperature, humidity, and season of year a particular hour of a specific day), while global models represent the same factors and learned correlations over multiple time periods (e.g. several months).The team asked focus group participants to interpret the charts (e.g., identify which inputs were correlated with the largest increase or decrease in predicted power use). The accuracy with which participants interpreted each chart provided a metric of the ease with which that chart type could be understood. and rank the level of difficulty they had in interpreting each of the charts. To prevent any subject matter expertise from influencing their answers, the team replaced the model input names (outdoor air temperature, humidity, and so on) with feature identifiers such as *Feature 1* or *Feature 3.* The results are in Figures 8, 9, 10, and 11.

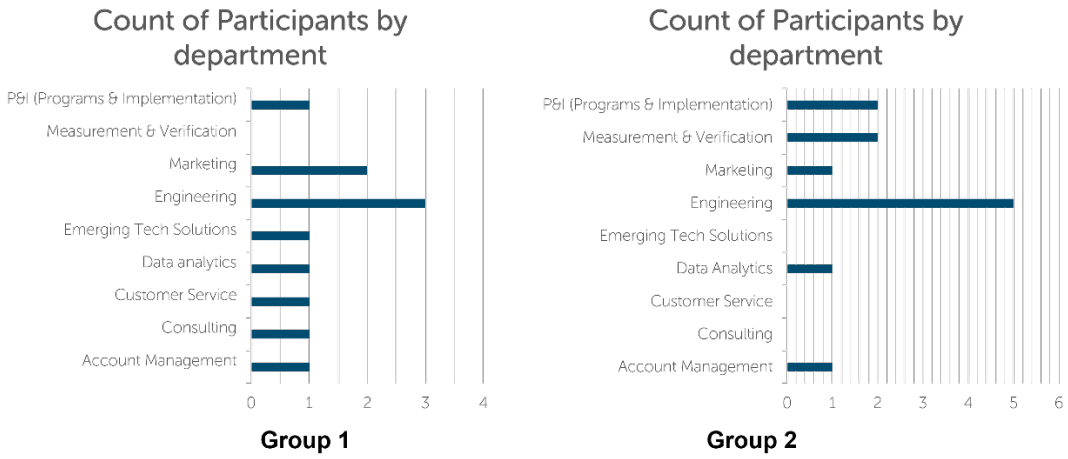
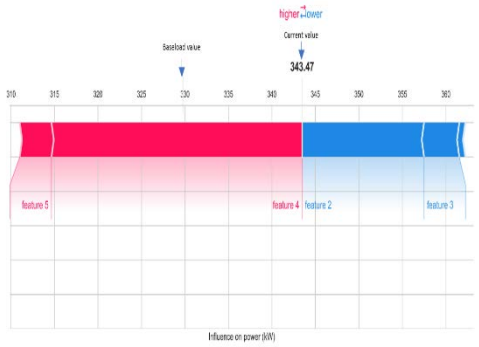


Figure 7. Subject matter expertise among each of the two focus groups.

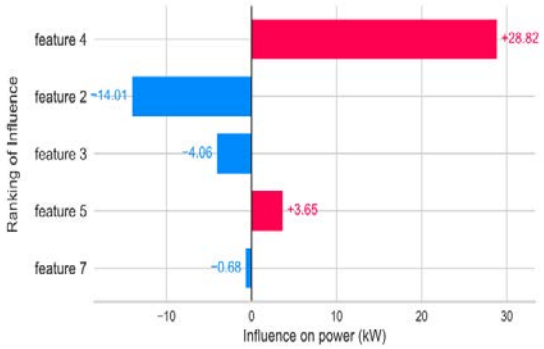


	All participants	Session #1	Session #2
Feature most increasing power use	78%	80%	77%
Feature most decreasing power use	63%	56%	69%
Feature with largest effect on power use	92%	100%	85%
Power use	75%	55%	92%
Average	77%	73%	81%
% indicating difficulty	66%	90%	46%

Values other than "% indicating difficulty" are the percent of respondents who answered correctly

Figure 8. SHAP force plot showing correlations learned from a snapshot of specific model inputs (given specific values for each feature, what are the learned correlations, relationships between inputs and outputs, as well as predicted power demand). Participants found the force plot the hardest chart to interpret.

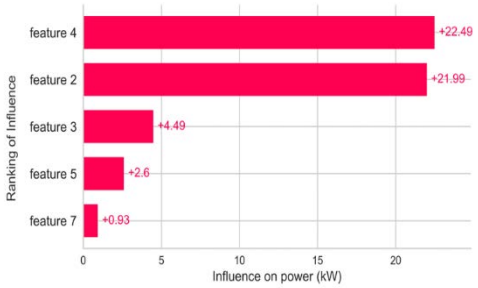
The team's findings were that SHAP bar charts (Figures 9 and 10) were the most interpretable chart for participants, whereas force plots (Figure 8) where the hardest plots to read. In fact, 66 percent of both focus groups ranked this plot as *difficult*. Bee swarm plots were also challenging for users, but qualitative feedback indicated that these were the most information-rich and engaging charts, and that with adequate labeling, explanation, and possible pairing with bi-directional charts, they could communicate substantial amounts of helpful information.



	All participants	Session #1	Session #2
Feature most increasing power use	87%	82%	92%
Feature most decreasing power use	92%	82%	100%
Feature with largest effect	96%	91%	100%
Average	92%	85%	97%
% indicating difficulty	5%	10%	0%

Values other than "% indicating difficulty" are the percent of respondents who answered correctly

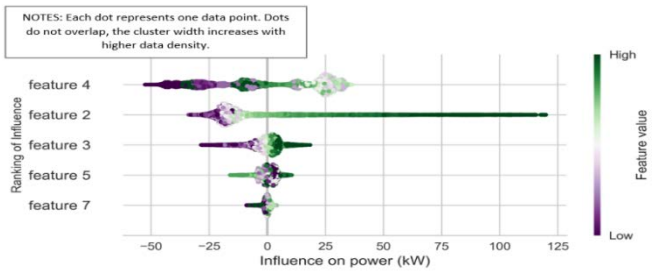
Figure 9. Bar chart (local model): this chart shows correlations learned from a snapshot of specific model inputs (given specific values for each feature, what are the learned correlations as well as predicted power demand). Users found the bi-directional bar charts the most intuitive of all the chart types.



	All participants	Session #1	Session #2
Feature with largest effect on power use	96%	91%	100%
Feature with smallest effect on power use	100%	100%	100%
Average	98%	96%	100%
% indicating difficulty	5%	10%	0%

Values other than "% indicating difficulty" are the percent of respondents who answered correctly

Figure 10. Bar chart (global model) showing correlations learned over several months of data. Users found this chart easy to interpret. It also had the highest rate of accurate interpretations across all the charts.



	All participants	Session #1	Session #2
Feature with largest potential to affect power use	70%	45%	92%
Feature with smallest potential to affect power use	92%	91%	92%
Identify features with potential to not affect power use			
Average	81%	68%	92%
% indicating difficulty	52%	50%	54%

Values other than "% indicating difficulty" are the percent of respondents who answered correctly

Figure 11. Bee swarm chart (global model): This chart was the most complex visualization, but also the one that most captured participants' interest. User feedback was that this chart would provide the most information if users were guided on how to use/interpret it.

The key points that the team learned from the quantitative and qualitative focus group feedback were:

- Visually easy-to-understand charts offer good representations of data about power use. The charts were especially useful for (1) facilities managers who wanted a tool to understand what their energy model said about their facility; and (2) Efficiency Vermont technical and account management staff seeking information about a customer site's power use.

- Clarity of chart labels and design elements, such as helper text and color-blind-friendly color schemes, were important for ensuring that charts could be easily understood and accessible to as broad an audience as possible.
- Presenting power use in dollars terms (rather than energy use in kWh) makes it easier for the customer to relate the data to operating costs and financial impacts, which might motivate action on energy efficiency investments.

Putting It All Together

Based on focus group feedback, the team designed the infographic in Figure 12 and Figure 13, for visualizing energy model results for customers.

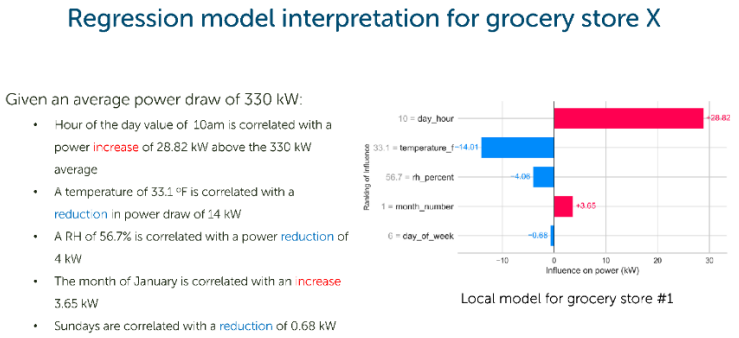


Figure 12. Model explanation chart including the elements learned from focus group interviews.

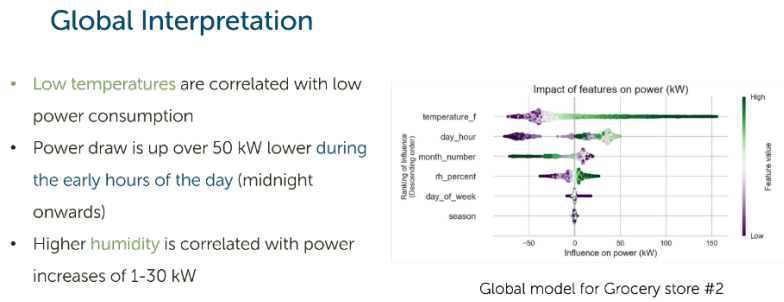


Figure 13. Global model interpretation for Store 2.

Provisos and Pitfalls: Predicting Is Not Explaining

SHAP charts make it easy to assume that a team can know the causal drivers of a model's prediction. But that is not the case. The SHAP plots show only the correlations learned by the model; and as the maxim goes: Correlation is not causation. The models the VEIC team created are predictive models, not explanatory models in the strict causal sense. As Galit Shmueli makes clear in her paper, explaining and predicting are very different endeavors (Shmueli 2010). These predictive models cannot be used to say that adjusting model input X by 30 percent should result in an output of Y. The learned models are valid only for relationships learned within the dataset the models have seen. These correlations could change as modifications are made to the building, or if there are changes to how the building is operated. Determining causal relationships between model inputs and the model output requires more rigor, and is best achieved through causal modeling, along with the requisite experiments and causal hypotheses.

Putting XAI into Production

To maximize the likelihood of a successful deployment of GBRTs with explainable AI, the VEIC Team recommends the following steps:

- **Determine if you have sufficient data**. These methods require 15-minute or hourly interval data. The team also recommends at least one year of data for modeling, with data that capture all the seasons experienced at the building's location.
- **Benchmark your models.** Modelers should compare the performance of existing modeling strategies with GBRTs. The VEIC team advises switching to GBRTs only if the performance gain on the data is significant.
- **Customize your plots.** The team recommends using simple SHAP plot types, with added helper text and color-blind-friendly colors.
- **Be clear about model limitations.** These are predictive, not causal, models.

Conclusions

The goals of this project were to determine if advanced machine learning models such as GBRTs could produce better energy models than OLS methods, without sacrificing model interpretability. The VEIC team's results indicate that when these models are combined with an XAI framework like SHAP, and when the model interpretation plots are carefully chosen for simplicity and well-labeled chart elements, the answer is a resounding *Yes*. Care must be taken to communicate to stakeholders that these are predictive, not causal models. But XAI for energy efficiency allows teams to have the best of both worlds—powerful yet interpretable models. XAI for energy modeling is ready for prime time.

References

Fels, M. 1986. "PRISM: An Introduction," *Energy and Buildings 9*: 5-18. https://www.sciencedirect.com/science/article/abs/pii/0378778886900034.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. 2017. "LightGBM: A highly efficient gradient boosting decision tree." *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS 2017). Red Hook, NY: Curran Associates: 3149–57.

Lundberg, S. M., and S. I. Lee 2017. "A unified approach to interpreting model predictions." *Proceedings of the 31st International Conference,* NIPS 2017. Red Hook, NY: Curran Associates*:* 4768-47.

Mathieu, J.L., P.N. Price, S. Kiliccote, and M.A. Piette. 2011. "Quantifying changes in building electricity use, with application to demand response." *IEEE Transactions on Smart Grid 2.3*: 507-18. https://www.researchgate.net/publication/220592806_Quantifying_Changes_in_Building_Electricity_Use_With_Application_to_Demand_Response.

Molnar, C. 2019. "Interpretable Machine Learning. A Guide for Making Black Box Models Explainable." https://christophm.github.io/interpretable-ml-book/.

Severinsen, A., and R. Hyndman. 2019. "Quantification of energy savings from energy conservation measures in buildings using machine learning." *ECEEE Summer Study Proceedings.* https://robjhyndman.com/publications/energy-savings/.

Shmueli, G. 2010. "To Explain or To Predict?" *Statistical Science*, *25*(3), 289-310. https://www.galitshmueli.com/publication/explain-or-predict-0.

Touzani, S., J. Granderson, and S. Fernandes, S. 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158, 1533-43. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=NhpEsTMAAAAJ&citation_for_view=NhpEsTMAAAAJ:9yKSN-GCB0IC.