Machine learning approaches for predictive modelling: energy use in the US commercial buildings

Introduction

Understanding energy use in the buildings sector, which accounts for around 40% of the total energy consumption in the US (EIA, 2018) and the EU (EC, 2021), is key to finding ways to improve energy efficiency and mitigate climate change. One way of estimating energy consumption in buildings is using physics-based models (Zhao & Magoulès, 2012). However, the variety of building types, characteristics and principal activities creates uncertainties, making these models impractical on a larger scale. Alternatively, statistical methods and machine learning (ML) offer a way to avoid the physical modelling expenses, while providing reasonable predictions.

This study aims to predict the Energy Use Intensity (EUI) of commercial buildings and analyse the significance of various building features on the EUI. For this purpose, several ML models are trained on the Commercial Building Energy Consumption Survey (CBECS), the most comprehensive publicly available dataset on energy use in commercial buildings in the US (EIA, 2012).

Related work

There is a large body of literature on applying statistical methods to predict the energy consumption in buildings, some of which use the CBECS dataset.

One study using data on office type buildings in the 1999 CBECS dataset found that artificial neural networks (ANN) provide more accurate EUI predictions compared to multiple linear regression (Yalcintas & Ozturk, 2007). Another study trained a random forest model to predict EUI on medium-sized office and school buildings in CBECS 2003 dataset (Kaskhedikar et al., 2015). The study found that none of the building features are influential and the random forest models were only slightly better than the traditional regression-based models.

A study trained eleven ML-based and two linear models on five variables from the 2012 CBECS dataset (Robinson et al., 2017a). The authors validated the models through New York City benchmark dataset and found that ML models and especially extreme gradient boosting perform better at predicting energy use compared to linear models. In the same year, another article compared the total EUI, HVAC EUI, lighting EUI, and plug load EUI prediction performance of six models trained on the office type buildings in the 2012 CBECS dataset (Deng et al., 2018). The random forest and support vector machine models demonstrated slightly better performance on total EUI, although the plug load EUI was better predicted by linear regression models. Moreover, the study tested the importance of variables using random forest, which has shown inadequate predictive power.

Finally, a more recent study trained three ML models and a multiple linear regression model on the 2012 CBECS dataset, excluding industrial and processing related buildings (Mohammadiziazi & Bilec, 2020). The study found that the random forest model provided the best performance.

The reviewed studies show that ML models, particularly random forest, extreme gradient boost, and ANN perform slightly better on the CEBCS dataset. However, the studies are limited either by the few models they consider or the smaller sections of the dataset and variables they use.

This study will explore using the full 2012 CBECS dataset to estimate feature importance and use the important features to train random forest, gradient boosting, k-nearest neighbours (KNN) and ANN models to predict the EUI.

Methodology

Raw data & pre-processing

CBECS 2012 dataset includes 1119 columns with around 500 attribute variables on 6720 commercial buildings, representing the total commercial buildings stock of the US. The variables include physical and occupancy features, types of equipment and use, weather data as well as the energy consumption and types of energy consumed.

Prior to fitting, the dataset required significant pre-processing: the scale of response variables was reduced; data on very small or recently opened buildings as well as some non-informative variables were eliminated (e.g., imputation flags, costs); zero values were assumed for some of the missing values (e.g., number of elevators). Moreover, the various numbered groupings were converted into binned predictors and, subsequentially, all factor columns were converted into one-hot encoders.

Following that, *EUI* (Btu/sqft) feature (Table 1) was engineered by taking the total annual major fuel consumption per square footage ratio and normalising for weather differences using degree days variables. *EUI* is the most widely used building energy performance metric in the US (Chung et al., 2006). The remaining consumption variables were removed, as were outliers and highly correlated variables. The remaining rows with missing values were also removed leaving a total of 1099 features for 6501 buildings.

Table 1. Primary statistics of EUI and normalised EUI.

Feature	Minimum	Maximum	Median	Mean	Standard Deviation
EUI	0	1340.2	64.4	101.6	122.6
Normalised EUI	0	0.4	0.0124	0.021	0.028

Finally, to account for the skewness of the distributions (Figure 1) and improve the stability of models, a log+1 transformation of continuous variables was performed and all of the data was normalised as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} .$$

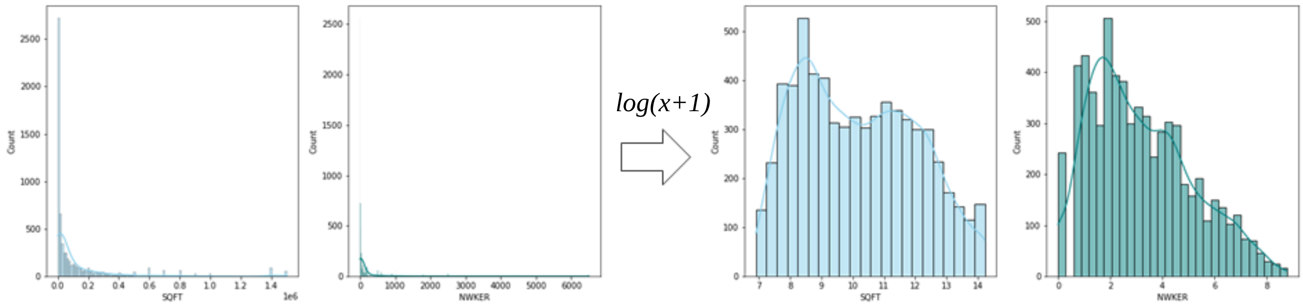

Figure 1. Example distributions (sq. footage and employee number) before and after the log transformation.

Data exploration & feature selection

Considering a very large number of features in the dataset, it is important to select only the significant features to reduce the computational costs and avoid spuriousity without losing important information (Li et al., 2017). One way to reduce the dimension of feature space is Principal Component Analysis (PCA). However, due to the nonlinear nature of the dataset individual components explain very little variance with the first two components accounting for around 17% of the total variance (Figure 2).

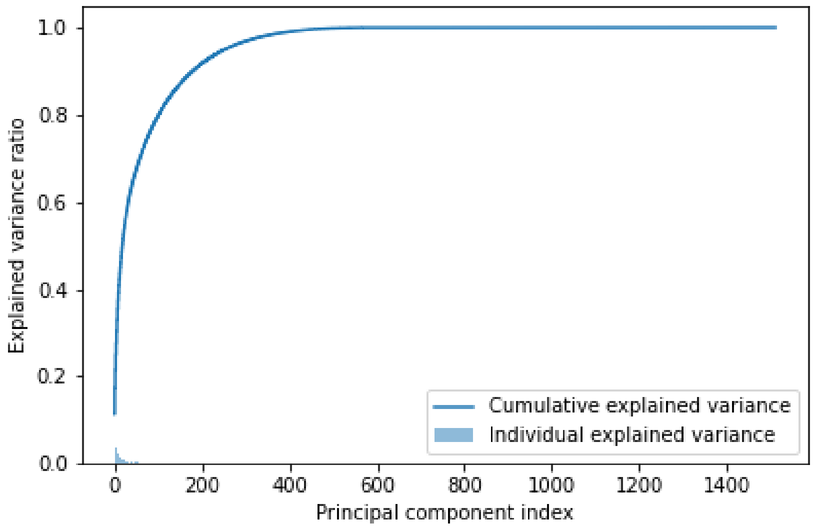

Figure 2. Variance explained by the principal components.

Ordinary least squares (OLS) stepwise regression, included in the US EPA's Energy Star methodology, is a common wrapper-type method to select important variables (EPA, 2011; Sharp et al., 1998).

In the linear regression model, OLS is a method for parameter estimation, which minimises the sum of squared distances between the observations and approximation, with the following form:

$$y = \boldsymbol{X}\beta + \varepsilon$$

where y is the response vector, $\boldsymbol{X}$ is the regressors matrix, β is the coefficient vector and ε is the error vector.

This study implements the OLS forwards selection, assuming the null hypothesis and rejecting variables within a 95% confidence interval. The process begins with a null model and recursively adds the variable with the smallest p-value until there are no variables with a p-value smaller than 0.05.

The stepwise regression selected 112 features, which were examined more closely. Upon examination, more correlated features were identified (e.g., square footage and square footage category) and removed, leaving 97 estimators. The features were then ranked per goodness of fit, using recursive feature elimination (RFE) from the scikit-learn package (Pedregosa et al., 2011) with the random forest model, as this model has shown the best performance in reviewed studies.

Modelling

Four ML models were trained on the cleaned **full dataset**, the **selected features** and the **top quantile** of the ranked selected features: random forest, gradient boosting, KNN and ANN.

Random forest and gradient boosting are tree ensemble algorithms. Random forest uses bootstrapping to construct decision trees parallelly and reduce the bias, then using bagging to average the trees, reducing variance (Breiman, 2001). In turn, gradient boosting grows trees sequentially, improving on the errors of each previous tree (Friedman, 2002).

KNN is a simple algorithm that makes predictions based on the average of closest points that are found with Euclidean distance (Bishop, 2006). These three methods have been used with standard parameters. ANN method builds a network of interconnected "neurons" with hidden layers that allow it to model complex and nonlinear patterns (Haykin & Lippmann, 1994).

The ANN used for this study consists of one input, eight hidden and one output layers, with a rectified linear activation function (ReLU) and adaptive moment estimation (Adam) optimiser (Appendix 2).

The coefficient of determination metric (R^2), which scores predicting the mean as zero, has been utilised to estimate the accuracy of the models.

All of the code was produced in R and Python.

Analysis and results

Data exploration & feature selection

The 97 selected and ranked features (Appendix 1) could be split into four categories (Figure 3) with an exception of degree days, which is not a characteristic of the buildings.

Energy saving, supply and heating properties category is the largest, with variables on *natural gas usage*, *percent heated by boiler* and *variable air volume (VAV) ventilation usage* ranked as the most significant. Despite being the largest, half of the features in this category are in the bottom quantile of the ranking, making it not the most significant.

The second largest category, **operational properties** is by far the most significant, dominating the top quantile of the ranking. Out of the top-five ranked features, three are related to operational properties: *total hours open per week*, *principal building activity* and *number of employees*. The significance of this category could be explained by the heterogeneity of energy consumption between sectors and industries and estimators being relevant only to a specific building activity (Robinson et al., 2017b). The impotance of operational properties in buildings is also widely supported in the literature (Gao & Zhang, 2011; Martani et al., 2012; Masoso & Grobler, 2010).

The top features included in the **equipment and appliance properties** category are related to the *number of refrigerators* and *percentage of cold storage*, as well as *laboratory* and *computing equipment*. These features are also likely to be highly correlated to the principal activity of the building (e.g., offices will have more computers, but less cold storage than the food service buildings).

Finally, the **physical properties** category, including highly ranked features related to *square footage*, *roof tilt and materials*, *percentage of exterior glass* and *wall construction materials*, is the smallest. Although *square footage* feature of this category is ranked as the second most impotant, the remainder of the category is ranked below 20.

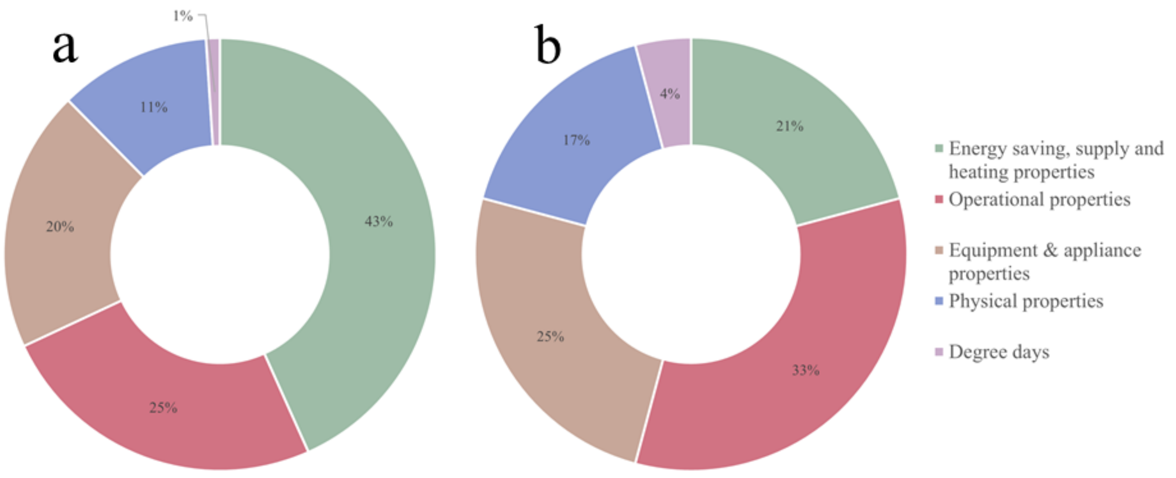

Figure 3. Distribution of categories including selected features in a) full ranking; b) top quantile.

Modelling

Results in Table 2 indicate that the accuracies of the models are low, with Gradient Boosting and Random Forest outperforming KNN and ANN. This is likely due to the strong non-linearity of the features, which was previously shown by the principal components, as well as unreported occupant behaviour patterns, gaps in the collected physical parameters and overfitting.

Table 2. The coefficient of determination metric (R^2) of EUI prediction in testing sets by different models.

Algorithm	Full Dataset	Selected Features	Top Quantile
Gradient Boosting	0.64	0.63	0.61
Random Forest	0.63	0.60	0.60
KNN	0.37	0.48	0.40
ANN	0.50	0.49	0.52

Although the overall accuracy of the models in the reviewed works is similar, the top-performing models in these studies have achieved results that are marginally better. This might be due to the consideration of smaller subsets of the dataset in the papers. The subsets were based on the principal activity in the buildings, which reduces the heterogeneity of the energy use and therefore the complexity of the problem. This suggests that energy use and efficiency in commercial buildings should be analysed separately with regard to the principal building activity.

Table 2 also shows that training the models on the selected features, as well as on the top quantile of the selected features, leads to only a minor decrease in performance in Gradient Boosting and Random Forest, and increases the accuracy in KNN and ANN. As such, feature reduction according to the proposed methodology allows lowering the computational costs, without losing significant information.

The EUI error plots (Figure 4) show that all models, but especially ANN, are systematically underestimating higher consumption values. This could be due to data becoming much sparser at higher values, making the models biased towards the lower values.

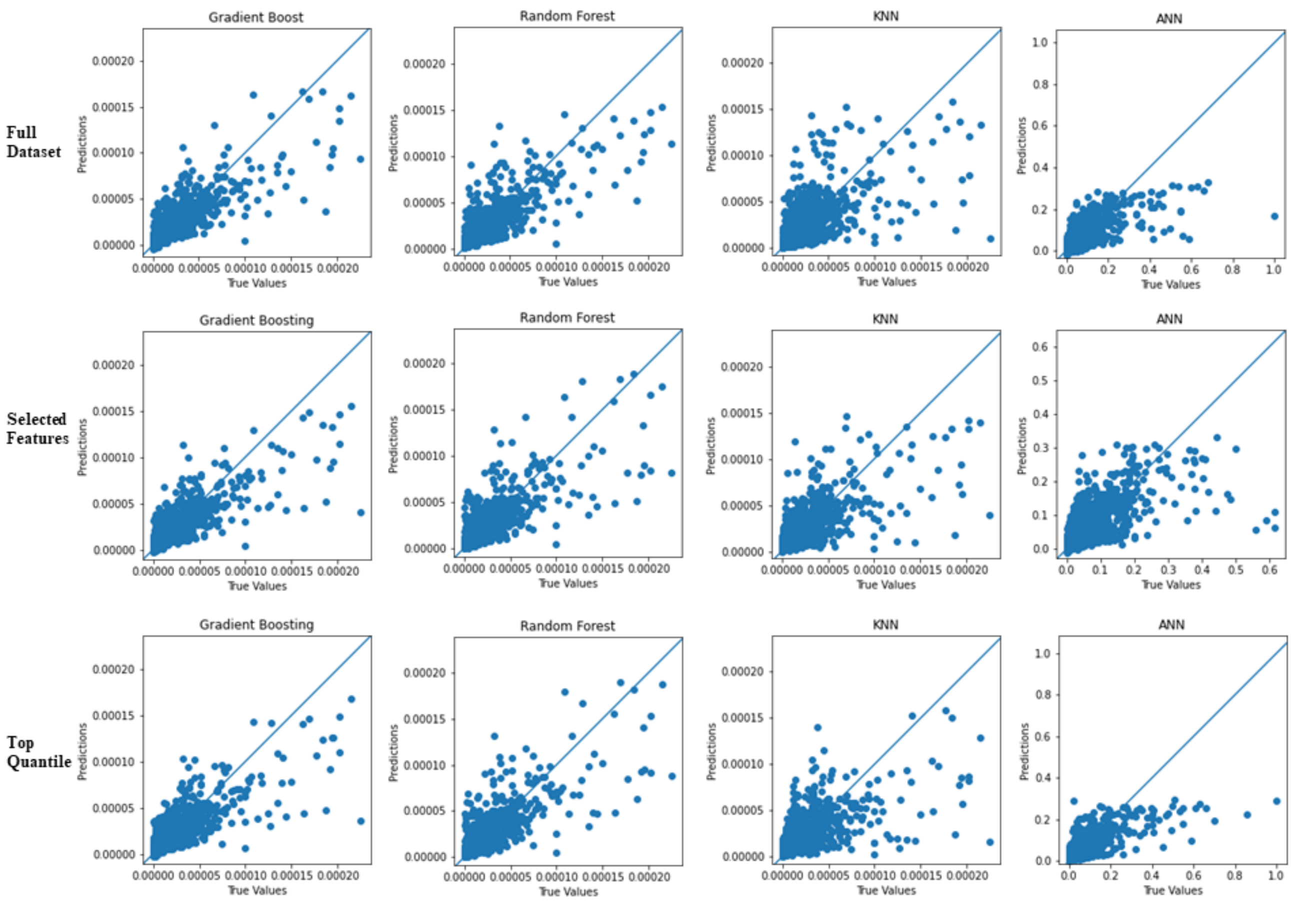

Figure 4. Error plots of predicted EUI values compared against the true values for the full dataset, the full ranking of selected features and the top quantile of the ranking.

Conclusion

This study analyses the importance of features in the CBECS dataset for predicting the EUI and trains machine learning models to estimate the energy consumption in commercial buildings.

It was found that through stepwise OLS and RFE significant feature reduction can be achieved without any major accuracy losses. Furthermore, the results reveal the importance of features related to operational properties as well as the non-linear and heterogeneous nature of the dataset. It is suggested that consideration of the building energy use in categories based on principal building activity could lead to better model performance. Feature selection and ranking further highlights the importance of targeting not only the physical properties of buildings, but also the way that they are used to improve the efficiency of the energy use.

The sophisticated machine learning algorithms used in the study could not capture the uncertainties and complex interaction of energy consumption patterns, despite their ability to offer nonlinear and sparse solutions.

The performance gap between the ground truth and EUI prediction could be accredited to imputed values, outliers and complex interactions between the estimators. The physical diversity of building types, discrepancies in occupant behaviour and strong energy consumption heterogeneity between principal building activities could also be negatively impacting the accuracy of the models (Azar & Menassa, 2012). As such, the study could be improved by predicting the energy use in different categories based on the principal building activities as well as addition of the thermal performance and occupant behaviour variables.

Further improvements could include hyperparameter tuning for the models using grid search to achieve better accuracy. ANN could benefit from additional layers as deeper networks tend to perform better with nonlinear data (Ogunmolu et al., 2016). Moreover, a shrinkage parameter for ANN can prevent overfitting by penalising larger weights (Hastie et al., 2009).

References

Azar, E., & Menassa, C. C. (2012). A comprehensive analysis of the impact of occupancy parameters in energy simulation of office buildings. *Energy and Buildings*, *55*, 841–853. https://doi.org/10.1016/J.ENBUILD.2012.10.002

Bishop, C. M. (2006). *Prml*. https://link.springer.com/book/9780387310732

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chung, W., Hui, Y. v., & Lam, Y. M. (2006). Benchmarking the energy efficiency of commercial buildings. *Applied Energy*, *83*(1), 1–14. https://doi.org/10.1016/J.APENERGY.2004.11.003

Deng, H., Fannon, D., & Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. *Energy and Buildings*, *163*, 34–43. https://doi.org/10.1016/J.ENBUILD.2017.12.031

EC. (2021). *Energy performance of buildings directive*. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en

EIA. (2012). *Commercial Buildings Energy Consumption Survey (CBECS) Data*. https://www.eia.gov/consumption/commercial/data/2012/

EIA. (2018). *U.S. energy flow*. https://www.eia.gov/totalenergy/data/monthly/pdf/flow/total_energy.pdf

EPA. (2011). *ENERGY STAR ® Performance Ratings Technical Methodology*.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gao, X., & Zhang, D. (2011). Analysis of the rule of influence of hotel occupancy ratio on energy consumption. *2011 International Conference on Electric Technology and Civil Engineering, ICETCE 2011 - Proceedings*, 1009–1014. https://doi.org/10.1109/ICETCE.2011.5774470

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7

Haykin, S., & Lippmann, R. (1994). Neural networks, a comprehensive foundation. *International Journal of Neural Systems*, *5*(4), 363–364.

Kaskhedikar, A., Reddy PhD PE, T. A., & Runger PhD, G. (2015). Use of Random Forest Algorithm to Evaluate Model-Based EUI Benchmarks from CBECS Database. *ASHRAE Transactions*, *121*, 17–28. https://www.proquest.com/scholarly-journals/use-random-forest-algorithm-evaluate-model-based/docview/1725205963/se-2?accountid=14511

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6). https://doi.org/10.1145/3136625/FORMAT/PDF

Martani, C., Lee, D., Robinson, P., Britter, R., & Ratti, C. (2012). ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings*, *47*, 584–591. https://doi.org/10.1016/J.ENBUILD.2011.12.037

Masoso, O. T., & Grobler, L. J. (2010). The dark side of occupants’ behaviour on building energy use. *Energy and Buildings*, *42*(2), 173–177. https://doi.org/10.1016/J.ENBUILD.2009.08.009

Mohammadiziazi, R., & Bilec, M. M. (2020). Application of Machine Learning for Predicting Building Energy Use at Different Temporal and Spatial Resolution under Climate Change in USA. *Buildings 2020, Vol. 10, Page 139*, *10*(8), 139. https://doi.org/10.3390/BUILDINGS10080139

Ogunmolu, O., Gu, X., Jiang, S., & Gans, N. (2016). *Nonlinear Systems Identification Using Deep Dynamic Neural Networks*. https://doi.org/10.48550/arxiv.1610.01439

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017a). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, *208*, 889–904. https://doi.org/10.1016/J.APENERGY.2017.09.060

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017b). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, *208*, 889–904. https://doi.org/10.1016/J.APENERGY.2017.09.060

Sharp, T. R., Sharp, T. R., Ridge, O., & Laboratoy, N. (1998). Benchmarking Energy Use in Schools. *PROCEEDINGS OF THE ACEEE 1998 SUMMER STUDY ON ENERGY EFFICIENCY IN BUILDINGS (3*, 305--316. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.172.1234

Yalcintas, M., & Ozturk, U. A. (2007). An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *International Journal of Energy Research*, *31*(4), 412–421. https://doi.org/10.1002/ER.1232

Zhao, H. X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. https://doi.org/10.1016/J.RSER.2012.02.049

Appendicies

Appendix 1. Selected features ranked by significance.

Rank	ID	Name	Category
1	DD	Degree days	Weather
2	SQFT	Square footage	Physical properties
3	WKHRS	Total hours open per week	Operational properties
4	PBA.15	Principal building activity = "Food service"	Operational properties
5	NWKER	Number of employees	Operational properties
6	PBA.16	Principal building activity = "Inpatient health care"	Operational properties
7	RFGWIN	Number of walk-in units	Equipment & appliance properties
8	NGUSED	Natural gas used (binary)	Energy saving, supply and heating properties
9	RFGICN	Numer of ice makers	Equipment & appliance properties
10	RFGCOMPN	Number of compact refrigerators	Equipment & appliance properties
11	HWTRM	Large ammounts of hot water (binary)	Operational properties
12	BOILP	Percent heated by boiler	Energy saving, supply and heating properties
13	PBA.6	Principal building activity = "Food sales"	Operational properties
14	NGWATR	Natural gas used for water (binary)	Energy saving, supply and heating properties
15	RFGSTP	Percent cold storage	Equipment & appliance properties
16	NGSRC..1	How natural gas is purchased = "Bought from the local utility"	Energy saving, supply and heating properties
17	PBA.5	Principal building activity = "Nonrefrigerated warehouse"	Operational properties
18	LABEQP	Laboratory equipment (binary)	Equipment & appliance properties
19	CLVVAV	Cooling ventilation: Central air-handling unit with VAV (binary)	Energy saving, supply and heating properties
20	RFGRES.2	Full-size residential-type refrigerator (binary)	Equipment & appliance properties
21	RFTILT.1	Roof tilt = "Flat"	Physical properties
22	RFCNS.6	Roof construction material = "Plastic, rubber, or synthetic sheeting"	Physical properties
23	PBA.91	Principal building activity = "Other"	Operational properties
24	GLSSPC.4	Percent exterior glass = "25 to 50 percent"	Physical properties
25	MAINT	Regular HVAC maintenance (binary)	Energy saving, supply and heating properties
26	BLDPLT.2	Central heating (binary)	Energy saving, supply and heating properties
27	GLSSPC.5	Percent exterior glass = "51 to 75 percent"	Physical properties
28	RDHTNF	Heating reduced during 24 hour period (binary)	Operational properties
29	STRLZR	Sterilizers or autoclaves (binary)	Equipment & appliance properties
30	STHW	District steam or hot water piped in (binary)	Energy saving, supply and heating properties
31	DCNTRSFC.5	Data centre or server farm sqft category = "Over 10,000 sqft"	Equipment & appliance properties
32	HT2	Energy used for secondary heating (binary)	Energy saving, supply and heating properties

33	WLCNS.6	Wall construction material	Physical properties
34	LOHRPC.4	Lit when open ="76 to 100 percent"	Operational properties
35	LTEXPC.4	Percent of exterior lighted = "More than 50 percent"	Operational properties
36	ACWNWP	Percent cooled by individual room A/C	Operational properties
37	SUNGLS.1	Glass on the sides with most sunlight = "More glass area"	Physical properties
38	GLSSPC.3	Percent exterior glass = "11 to 25 percent"	Physical properties
39	OWNTYPE.3	Building owner = "Individual owner(s)"	Operational properties
40	LAPTPC.1	Number of laptops category = "1 to 4"	Equipment & appliance properties
41	HWHT1	District hot water used for main heating (binary)	Energy saving, supply and heating properties
42	COPIER	Photocopiers (binary)	Equipment & appliance properties
43	FKHT2	Fuel oil used for secondary heating (binary)	Energy saving, supply and heating properties
44	TRNGRM	Computer-based training room (binary)	Equipment & appliance properties
45	PBA.12	Principal building activity = "Religious worship"	Operational properties
46	PKGFURN	Packaged heating component: Furnace (binary)	Equipment & appliance properties
47	CHLDUCT	Chiller system: Duct reheat (binary)	Energy saving, supply and heating properties
48	WHRECOV	Waste heat recovery (binary)	Energy saving, supply and heating properties
49	FACDST	Plant produces district steam (binary)	Operational properties
50	RENRDC	Reduction of floorspace (binary)	Physical properties
51	FURNAC	Furnaces that heat air directly (binary)	Energy saving, supply and heating properties
52	FASTFD	Fast food or small restaurant (binary)	Operational properties
53	RGSTR	Cash registers (binary)	Equipment & appliance properties
54	ELEVTR	Elevators (binary)	Equipment & appliance properties
55	SHUNIT	Individual heater: Unit heater (binary)	Energy saving, supply and heating properties
56	DHRAD	District heat system: Radiators (binary)	Energy saving, supply and heating properties
57	MONUSE	Months in use	Operational properties
58	DCNTRSFC.2	Data centre or server farm sqft category = "501 to 1,500 sqft"	Equipment & appliance properties
59	BOOSTWT	Booster water heaters (binary)	Energy saving, supply and heating properties
60	PBSEAT	Assembly seating capacity	Physical properties
61	PRCOOK	Propane used for cooking (binary)	Energy saving, supply and heating properties
62	EVAPCL	Evaporative or swamp coolers (binary)	Energy saving, supply and heating properties
63	OPEN24	Open 24 hours a day (binary)	Operational properties
64	MRI	MRI machines (binary)	Equipment & appliance properties
65	OPNMF	Open during week (binary)	Operational properties
66	OWNOCC.3	Owner occupied or leased to tenant(s) = "Combination occupied and leased"	Operational properties
67	FDPREP	Commercial or large kitchen (binary)	Equipment & appliance properties
68	FACACT.5	Type of complex = "Retail complex"	Operational properties

69	PCTRMC.10	Number of computers category = "2,500 or more"	Equipment & appliance properties
70	DAYHARV	Daylight harvesting (binary)	Energy saving, supply and heating properties
71	HCBED_bin.2	Licensed bed capacity = "More than 250"	Operational properties
72	MAINCL.7	Main cooling equipment = "Swamp or evaporative coolers"	Energy saving, supply and heating properties
73	FKHT1	Fuel oil used for main heating (binary)	Energy saving, supply and heating properties
74	FACACT.13	Type of complex = "Industrial complex"	Operational properties
75	LODGRM	Number of guest rooms	Physical properties
76	WOOTH	Wood for some other use (binary)	Energy saving, supply and heating properties
77	WOUSED	Wood used (binary)	Energy saving, supply and heating properties
78	WOGENR	Wood used for electricity generation (binary)	Energy saving, supply and heating properties
79	WOCOOK	Wood used for cooking (binary)	Energy saving, supply and heating properties
80	HTRCHLR	Heater chiller (binary)	Energy saving, supply and heating properties
81	COGENR	Coal used for electricity generation (binary)	Energy saving, supply and heating properties
82	WOWATR	Wood used for water heating (binary)	Energy saving, supply and heating properties
83	COHT1	Coal used for main heating (binary)	Energy saving, supply and heating properties
84	DRYCL	Dry cleaning onsite (binary)	Equipment & appliance properties
85	WOMANU	Wood used for manufacturing (binary)	Energy saving, supply and heating properties
86	COOTH	Coal for some other use (binary)	Energy saving, supply and heating properties
87	COMANU	Coal used for manufacturing (binary)	Energy saving, supply and heating properties
88	COUSED	Coal used (binary)	Energy saving, supply and heating properties
89	COWATR	Coal used for water heating (binary)	Energy saving, supply and heating properties
90	WOAMT..1	Amount wood burned = "Less than 1 cord"	Energy saving, supply and heating properties
91	COCOOK	Coal used for cooking (binary)	Energy saving, supply and heating properties
92	ELUSED	Electricity used (binary)	Energy saving, supply and heating properties
93	COHT2	Coal used for secondary heating (binary)	Energy saving, supply and heating properties
94	COURT	Food court (binary)	Operational properties
95	WOHT1	Wood used for main heating (binary)	Energy saving, supply and heating properties
96	WOSRC..1	How wood was obtained = "Purchased all wood"	Energy saving, supply and heating properties
97	WOHT2	Wood used for secondary heating (binary)	Energy saving, supply and heating properties

Appendix 2. Artificial neural network architecture used in the study.

Input layer	input:	(None, x_length)	Parameters
	output:	(None, x_length)	0

↓

Hidden layer 1		input:	(None, x_length)	Parameters
Dense	ReLu	output:	(None, 2048)	200704

↓

Hidden layer 2		input:	(None, 2048)	Parameters
Dense	ReLu	output:	(None, 1024)	2098176

↓

Hidden layer 3		input:	(None, 1024)	Parameters
Dense	ReLu	output:	(None, 512)	524800

↓

Hidden layer 4		input:	(None, 512)	Parameters
Dense	ReLu	output:	(None, 256)	131328

↓

Hidden layer 5		input:	(None, 256)	Parameters
Dense	ReLu	output:	(None, 128)	32896

↓

Hidden layer 6		input:	(None, 128)	Parameters
Dense	ReLu	output:	(None, 64)	8256

↓

Hidden layer 7		input:	(None, 64)	Parameters
Dense	ReLu	output:	(None, 32)	2080

↓

Hidden layer 8		input:	(None, 32)	Parameters
Dense	ReLu	output:	(None, 16)	528

↓

Output layer		input:	(None, 16)	Parameters
Dense	Linear	output:	(None, 1)	17