CLIMA 2022
Eye on 2030 | Towards digitalized, healthy, circular and energy efficient HVAC
REHVA 14th HVAC World Congress
22nd – 25th May, Rotterdam, The Netherlands

Data-driven statiscal and machine learning modelling of real building stock energy use

Matthias Van Hove[1], Jelle Laverge[1], Arnold Janssens[1], Marc Delghust[1]

[1]Department of Architecture and Urban Planning, Ghent University, Belgium, matthias.vanhove@ugent.be

Abstract. One of today's major challenges is to become climate neutral by 2050. Large potential for energy reduction is found in the building sector (which accounts for 40% of Europe's total primary energy use). To compare energy reduction strategies, Building-Stock Energy Models are vital instruments. Yet, the regulatory energy performance calculation (which is currently used by EU policy makers) poorly predicts the real building energy use in residential buildings and largely overestimates the potential energy savings. Promising data-driven black-box models are gaining considerable traction in a wide range of applications. This paper evaluates whether data-driven linear regression and gradient boosting machine models provide better predictions of the real total building energy use at large scale as compared to the current regulatory white-box building energy calculation method. Compared to the performance of the regulatory method, both the linear regression models and the gradient boosting regression trees perform better (gradient boosting regression trees slightly worse than multiple linear regression). Yet, a large part of the variance in the linear regression models is left unexplained and also for the gradient boosting trees, there is room for improvement. At individual building level, it is clear that both the linear regression model performance and the gradient boosting regression tree performance is too poor for inference. At stock level, however, both types of models seem promising and can be a useful tool to inform big housing owners (*e.g.*, financial institutions, governments, housing companies *etc.*) or for policy making.

Keywords. Data-driven modelling, Linear regression, Gradient boosting trees, Building stock energy use
DOI: https://doi.org/10.34641/clima.2022.379

1. Introduction

One of today's major challenges is to become climate neutral by 2050. Large potential for energy reduction is found in the building sector, which accounts for 40% of Europe's total primary energy use and 36% of the CO_2-emissions (EU, 2020). To compare competing energy reduction strategies, Building-Stock Energy Models (BSEMs) are vital instruments. Yet, the current regulatory energy performance calculation (which is currently used by EU policy makers) poorly predicts the real building energy use in residential buildings and consequently poorly informs current and future home owners and tenants about their energy use, largely overestimates the potential energy savings and therefore undermines policy making (Van Hove *et al.*, 2021).

Since the introduction of the regulatory energy performance calculation methods in 2009 (*i.e.*, specified by the Energy Performance of Buildings Directive (EPBD)), national building energy registries have emerged and vastly increased ever since. These data registries consist of aggregated building characteristic data which are used for the regulatory energy performance calculation (white-box) (EPC, 2015). With this data being available and the fact that the regulatory methods only poorly estimate the real building energy use (Macjen *et al.*, 2013; Van Hove *et al.*, 2021), the question rises whether data-driven statistical models and/or machine learning models can replace the regulatory methods for predicting the real annual building energy use.

Statistical models on the one hand, such as the Ordinary Least Squares (OLS) linear regression (Zdaniuk, 2014), have been around for a couple of

decades and have already been tested in other EU-countries, though not for Flanders and often not including socio-demographic variables. Machine learning models on the other hand, such as gradient boosting regression trees (de Queiroz *et al.*, 2016), are gaining considerable traction in a wide range of applications and have only been scarcely used for predicting the annual building energy use (also not for Flanders).

In this paper, we aim to study the predictive performance of data-driven linear regression models and data-driven gradient boosting regression trees for predicting the real annual total building energy use. Then, the results are being compared to the predictive performance of the regulatory calculation methods and there is being evaluated wether these data-driven black-box models can potentially replace the current regulatory white-box models for predicting the annual building energy use at individual building level as well as at stock level.

2. Research Methods

2.1. Data set

The data analysed for this paper were collected from a former study together with Flemish Energy and Climate Agency (Van Hove *et al.*, 2021). In total, it comprises 122,680 cases from the Flemish EPC registry (*i.e.*, one centralised database with data from the regulatory energy performance certificates of all registered existing buildings constructed before 2006). The data provide information about the building characteristics, technical systems and some detailed building geometry data. Also, there are annual real meter data available from the Belgian distribution system operator Fluvius and there are some socio-demographic and climate variables as well. After data cleansing, filtering and coupling, 56,930 cases were excluded from the sample based on the following five criteria:

(i) Coupling of data from various databases (*e.g.*, Fluvius real energy use data and data from the energy performance database).
(ii) Inconsistencies in the PV-data.
(iii) Doubt about the reliability of the real energy consumption data.
(iv) Single-family houses with energy sources other than natural gas and electricity for space heating (SH) and/or domestic hot water (DHW).

Hence, the total sample size was 69,870 cases which formed the basis for all the analyses carried out in this paper.

2.2. (In)dependent variables

The dependent variable in all analyses, the OLS regression models and the gradient boosting regression tree models is the annual real primary total energy use [kWh/y]. An overview of the independent variables (or predictors) is given in *Table 1* and *Table 2*. *Table 1* shows the general building variables used and their frequencies or summary statistics (M means Mean, SD means standard deviation for the continuous variables). *Table 2* shows descriptive information for the socio-demographic and weather variables.

Tab. 1 - Overview of general building variables and their frequencies (bold = reference category).

Variable	Categories (N)
Energy score	n/a (cont.: M=363kWh/m²·y, SD=173)
Construction year	n/a (cont.: M=1966, SD=28)
Latitude	n/a (cont.: M=51.05, SD=0.17)
Longitude	n/a (cont.: M=4.17, SD=0.72)
Usable floor space	n/a (cont.: M=169.6m², SD=61.7m²)
Building volume	n/a (cont.: M=512.1m³, SD=193.9m³)
Dwelling type	Detached (39.8%), **semi-detached (31.2%)**, terraced (29.0%)
Number of floors	n/a (cont.: M=2.3, SD=1.6)
Basement?	Yes (43.4%), **no (56.6%)**
Roof insulation?	Yes (25.4%), **no (74.6%)**
Floor insulation?	Yes (18.9%), **no (81.1%)**
Wall insulation?	Yes (38.4%), **no (61.6%)**
DHW on gas? (no=elec)	Yes (78.4%), **no (21.6%)**
SH on gas? (no=elec)	Yes (93.4%), **no (6.6%)**
Ventilation system	**A (96.8%)**, B (0.2%), C (2.1%), D (0.9%)
Condensing boiler?	Yes (49.4%), **no (50.6%)**
DHW storage vessel?	Yes (30.6%), **no (69.4%)**
SH Floor heating?	Yes (3.6%), **no (96.4%)**
SH Radiator/ Convector?	Yes (20.9%), **no (79.1%)**
SH Air heating?	Yes (1.1%), **no (98.9%)**
PV-panels?	Yes (3.6%), **no (96.4%)**
Heat pump?	Yes (0.3%), **no (99.7%)**
Solar collector?	Yes (1.5%), **no (98.5%)**
Space cooling?	Yes (1.5%), **no (98.5%)**
Social housing?	Yes (5.8%), **no (94.2%)**

Tab. 2 - Overview of socio-demographic and weather variables and their frequencies (bold = reference category).

Variable	Categories (N)
Number of occupants	n/a (cont.: M=2.63, SD=1.33)
Children 00-04yr?	Yes (28.4%), **no (71.6%)**
Children 05-12yr?	Yes (25.1%), **no (74.9%)**
Children 13-18yr?	Yes (14.4%), **no (85.6%)**
Adults 19-29yr?	Yes (26.8%), **no (73.2%)**
Adults 30-44yr?	Yes (53.0%), **no (47.0%)**
Adults 45-64yr?	Yes (38.0%), **no (62.0%)**
Adults 65+yr?	Yes (16.1%), **no (83.9%)**
Number of 65+-adults	n/a (cont.: M=0.21, SD=0.53)
Number of Children	n/a (cont.: M=0.95, SD=1.12)
Number of adults	n/a (cont.: M=1.92, SD=0.72)
HH composition	1Ad, 0Ch (17.3%), 1Ad, 1Ch (2.5%), 2Ad, 0Ch (24.1%), 2Ad, 1Ch (15.1%), 2Ad, 2Ch (18.2%), 2Ad, 3Ch (5.2%), 3Ad, 0Ch (4.7%), 3Ad, 1Ch (2.2%)
Length residency	n/a (cont.: M=7.19y, SD=10.4y)

weighted annual average temperature (Δ_Tav)	n/a (cont.: M=0.04°C, SD=0.19°C)
weighted annual DegreeDays (Δ_DD)	n/a (cont.: M=-26.0, SD=42.9)
weighted annual Global Horizontal Irradiance (Δ_GHI)	n/a (cont.: M=-1.16W/m^2, SD=17.56W/m^2)
weighted annual SolarHours (Δ_SH)	n/a (cont.: M=226.4, SD=73.3)

2.3. OLS linear regression

Statistical modelling with the data are all conducted in Python with the statistical packages 'scikit-learn' (Pedregosa *et al.*, 2011) and 'statsmodels' (Skipper *et al.*, 2010) in combination with the data analysis and visualisation package 'pandas' (McKinney, 2010). Initially, a linear ordinary least squares (OLS) regression model (Zdaniuk, 2014) has been built for both a set of 'general building variables' and a set of 'socio-demographic and weather variables'. Given the suspected issue of multicollinearity, the variance inflation factors (VIF) have been inspected. VIF indicates how much the variance of an estimated regression coefficient increases if the explanatory variables are correlated. If uncorrelated, VIF=1. In this paper a threshold of 5 has been used, as suggested in literature (Roberts *et al.*, 2009; Chan *et al.*, 2012). If VIFs greater than 5 are found in the OLS regression, then one (or more) of those correlated variables are excluded one-by-one stepwise depending on the regression coefficient and the individual VIF until a model is obtained with no collinearity issues.

Further, all *p*-values and bootstrapped 95% confidence intervals (CI) of the predictors are checked for irregularities. Only the predictors for which the *p*-values are <.05 and the confidence intervals do not include nil are significant and thus kept in the model. The exclusion of explanatory variables from the model is once more done one by one stepwise. After building the individual models (*i.e.,* 'general building variables' and 'socio-demographic and weather variables'), the models are combined until resulting in a final model encompassing all explanatory variables, tested and adjusted for multicollinearity and significant regression coefficients. As all model input variables are normalised, the magnitude of the regression coefficients gives an indication of the parameters relative importance in the regression model.

In order to fulfil the necessary assumptions for linear regression models, the model input variables are checked for linearity, autocorrelation and multicollinearity; the residuals are checked for independency, homoscedasticity and normality.

2.4. Gradient boosNng regression trees

The gradient boosting machine (GBM) is part of a class of powerful ensemble machine learning algorithms based on the concept that a "strong learner" (de Queiroz *et al.*, 2016), having high prediction accuracy, can be obtained by iteratively combining several less complex models, called "weak learners". Such ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimisation algorithm. This gives the technique its name, "gradient boosting", as the loss gradient is minimised as the model is fit, much like a neural network. Gradient boosting is an effective machine learning algorithm and is often the main, or one of the main, algorithms used to win machine learning competitions (like Kaggle) on tabular and similar structured datasets, which is why we test its performance in this paper.

The algorithm provides hyperparameters that must be tuned for a specific dataset. Such parameters include the number of trees or estimators in the model, the learning rate of the model, the maximum tree depth, the minimum tree weight *etc.* The gradient boosting implementation that we are using in this paper is XGBoost (Extreme Gradient Boosting) with the Python package 'xgboost' (Chen *et al.*, 2016). First a baseline model is made with a 80%-20% training-set split without hyper parameter tuning. Then hyper parameter tuning is performed with the Bayesian optimisation algorithm (HYPEROPT) (*N.B.*, see *Table 3* for the used hyperparameter space) (Bergstra *et al.*, 2015), with 80%-20% training-set split and a 5-fold cross-validation. Similar to the OLS regression models, first initial models are built for a set of 'general building variables' and a set of 'socio-demographic and weather variables'. Then models are combined until resulting in a final model encompassing all explanatory variables.

Tab. 3 - List of hyperparameters of the GBM models and tuning range.

name	tuning range
learning rate	[0.001, 0.1]
hessian regularisation	[1, 10]
loss regularisation	[1, 12]
column sampling by tree	[0.5, 1]
column sampling by level	[0.5, 1]
maximum depth	[3, 25]
number of iterations	[100, 1800] (with early stopping at 50)

We further use SHAP (Shapley Additive exPlanations) values to interpret the outputs (Lundberg *et al.*, 2017)(Fig. 1). The x-axis of a SHAP summary plot shows the impact off features on the outcomes, based on the SHAP values. Features are sorted based on their impact, thus the variable at the top has the highest impact. The color represents the feature values, with red for high values and blue for low values. The vertical dispersion for each feature corresponds to the data

points with the same SHAP values. If red points are mostly present on the positive side of SHAP values, it means by increasing the value of the independent variable, the dependent variable increases as well.

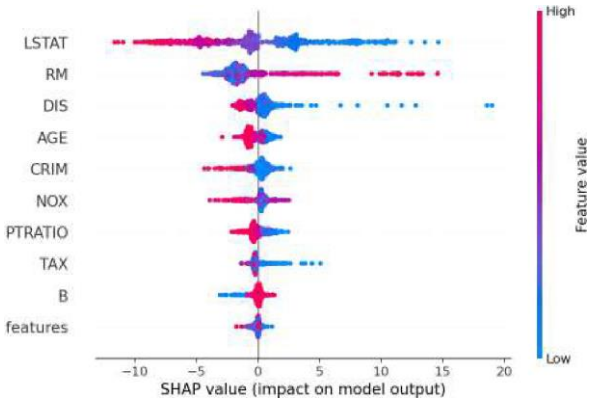

Fig. 1 - SHAP reference graph.

3. Results

3.1. General building characterisNcs OLS model

General building variables explained (adjusted) R^2=38.1% of the variability in the real total energy use. Two variable showed VIF values above the chosen threshold criterion so some of those had to be excluded one-by-one stepwise (*i.e.*, building volume). Further, five variables had confidence intervals including nil and a *p*-value >.05 meaning that some of those had to be excluded one-by-one stepwise as well (*i.e.*, type of ventilation system, construction year, roof insulation?, number of floors and SH energy carrier). After exclusion of those variables, an OLS regression rerun on the remaining variables resulted in a model that explained R^2=41.3% of the variability in the real total energy use. *Table 5* shows the coefficients of the reduced OLS model (*i.e.*, standardised coefficients β_{OLS}, 95% confidence intervals of the standardised coefficients β_{OLS} and the p-values of the standardised coefficients β_{OLS}). Five variables are significant: A larger dwelling size is associated with higher total energy use, the presence of renewable systems (*i.e.*, PV-panels, heat pump and solar collector) is associated with using less total energy use, floor and air heating is associated with more total energy use compared to radiator space heating, having space cooling is associated with having higher total energy use and detached houses are associated with having higher total energy use compared to semi-detached houses while terraced houses are associated with having less total energy use compared to semi-detached houses.

3.2. Socio-demographic & weather OLS model

The socio-demographic and weather model explained (adjusted) R^2=11.7% of the variability in the real total energy use. Six variables showed VIF values above the chosen threshold criterion so some of those had to be excluded one-by-one stepwise (*i.e.*, the number of children, average age of HH, HH with children, the number of 65+ per HH). Further, six variables had confidence intervals including nil and a *p*-value >.05 meaning that some of those had to be excluded one-by-one stepwise as well (*i.e.*, HH with children 5-12 and 13-18, HH with adults 30-44 and 65+, HH composition 2Ad 1Child and weighted annual GHI). After exclusion of those variables, an OLS regression rerun on the remaining variables resulted in a model that explained R^2=13.6% of the variability in the real total energy use. *Table 6* shows the coefficients of the reduced OLS model. Three variables are significant: A larger household size and a larger number of adults is associated with higher total energy use, singles are associated with less total energy use and a more annual heating degree days at the dwelling's location are associated with higher total energy use.

3.3. Combined OLS model

In the next step, the two different individual models are combined together for increments in explanatory power through adding additional variables. For the building and socio-demographic and weather model, only the variables that had remained after VIF- and CI-checks have been included. The combined 'general building and socio-demographic and weather' model explained (adjusted) R^2=46.6% of the variability in real total energy use. No variables showed VIF values above the chosen threshold criterion and no variables had confidence intervals including nil. An OLS regression rerun on the remaining variables was thus not necessary. This 5.3% increase in R^2 compared to the model with general building characteristic variables only is significant (p<.01). Also in comparison to the socio-demographic and weather OLS model, the increase in R^2 is significant (p<.01). *Table 7* summarises the coefficients for all variables that remained after VIF- and CI-checks.

Table 4 shows the adjusted R^2 of the individual OLS models and the combined OLS model as well as results for the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. As expected, building characteristic variables explain by far the most of the variability in real total energy use, on their own, and also when added to building and socio-demographic and weather variables. Socio-demographic variables play a lesser but still significant role in explaining real total energy use. Extra detailed building variables don't add much information to the model to predict more of the variability in the real total energy use. Therefore, in explaining more of the variability (and avoiding collinearity problems), the modeller can better gather more different types of variables (*i.e.*, building characteristics, socio-demographics, incomes, weather, appliance ownerships *etc.*) rather than more detailed variables within the same type of variables (*e.g.*, extra detailed variables related to

building characteristics). Further, as a model with general building characteristics (*i.e.*, building features that inhabitants can easily fill in themselves) performs equally well compared to a model including detailed building characteristics and intermediate results from the regulatory calculation, the final model can be used (*e.g.*, in an online tool) as a primary model to inform inhabitants, tenants and home owners about their energy use, without needing extra inputs from the EPC-registry to improve performance.

Tab. 4 - Adjusted R^2, MAE and RMSE for the two individual models and the combined model.

model	gen_build	socio	final
(adj) R^2	41.3%	13.6%	46.6%
MAE	5227 kWh/y	6211 kWh/y	5011 kWh/y
RMSE	6469 kWh/y	7832 kWh/y	6214 kWh/y

Tab. 5 - Coefficients of OLS regression model, general building variables.

Predictor	β_{OLS}	95% CI_{OLS}	p_{OLS}
(constant)	22694.2	[22526.9, 24373.3]	<.001
Energy score	1698.9	[1549.1, 1927.9]	<.001
Latitude	193.0	[20.4, 342.7]	<.013
Longitude	695.6	[550.6, 876.8]	<.001
Usable floor space	5034.9	[4854.8, 5212.1]	<.001
Detached?	2214.8	[1834.0, 2443.4]	<.001
Terraced?	-1800.9	[-2096.4, -1579.6]	<.001
Basement?	745.6	[506.5, 967.8]	<.001
Floor insulation?	-553.3	[-850.4, -247.2]	<.002
Wall insulation?	-517.2	[-772.3, -228.3]	<.001
DHW on gas?	717.4	[306.3, 1018.2]	<.001
Condensing boiler?	-1404.5	[-1611.4, -1098.0]	<.001
DHW storage vessel?	1719.0	[1525.3, 2085.8]	<.001
SH Floor heating?	3928.5	[3244.1, 4740.9]	<.001
SH Radiator?	2145.3	[1884.9, 2687.9]	<.001
SH Air heating?	3589.5	[2581.5, 4933.7]	<.001
PV-panels?	-3475.2	[-3991.1, -3080.9]	<.001
Heat pump?	-4613.1	[-6561.9, -813.4]	<.003
Solar collector?	-1865.6	[-2488.6, -752.7]	<.001
Space cooling?	2236.3	[1374.4, 2679.9]	<.002
Social housing?	1154.5	[775.8, 1399.9]	<.001

Tab. 6 - Coefficients of OLS regression model, socio-demographic variables.

Predictor	β_{OLS}	95% CI_{OLS}	p_{OLS}
(constant)	24760.7	[24326.2, 25237.3]	<.001
Number of occupants	1247.5	[1027.9,1606.9]	<.001
Number of adults	1397.5	[1045.1,1703.2]	<.001
Children 00-04yr?	-1028.6	[-1398.4,-630.6]	<.001
Adults 19-29yr?	-1143.6	[-1463.4,-809.2]	<.001
Adults 45-64yr?	870.8	[517.1,1099.1]	<.001
Length residency	158.2	[127.2,186.3]	<.001
HH: 1Ad, 0Child	-3133.4	[-3789.8,-2416.8]	<.001
HH: 1Ad, 1Child	-1227.0	[-2282.9,-505.5]	<.001
HH: 2Ad, 0Child	-1190.3	[-1688.9,-720.7]	<.001
HH: 2Ad, 2Child	1238.4	[701.7,1464.8]	<.001
weigh. ann. Tav	703.2	[326.9,1028.9]	<.001
weigh. DegreeDays	1413.0	[981.2,1783.4]	<.001
weigh. SolarHours	-659.7	[-825.1,-490.8]	<.001

Tab. 7 - OLS coefficients for the final combined regression model.

Predictor	β_{OLS}	95% CI_{OLS}	p_{OLS}
(constant)	11427.4	[21995.9, 23892.4]	<.001
Energy score	1619.9	[1483.9, 1852.8]	<.001
Usable floor space	3780.8	[3199.2, 4138.1]	<.003
Detached?	2334.4	[1967.6, 2544.4]	<.001
Terraced?	-1846.7	[-2111.1, -1619.2]	<.001
Basement?	659.4	[435.4, 894.8]	<.001
Floor insulation?	-399.9	[-698.8, -98.6]	<.001
Wall insulation?	-451.8	[-669.5, -198.2]	<.001
DHW on gas?	546.2	[169.9, 857.1]	<.001
Condensing boiler?	-1168.1	[-1387.5, -883.1]	<.001
DHW storage vessel?	1607.6	[1435.4, 2009.9]	<.001
SH Floor heating?	4010.4	[3448.2, 4875.4]	<.001
SH Radiator?	2024.9	[1767.9, 2508.9]	<.001
SH Air heating?	3546.6	[2499.6, 4855.8]	<.001
PV-panels?	-3913.3	[-4436.5, -3477.3]	<.017
Heat pump?	-4981.3	[-6854.2, -1449.2]	<.011
Solar collector?	-1553.6	[-2321.5, -465.6]	<.002
Space cooling?	2237.1	[1427.8, 2752.5]	<.001
Social housing?	527.2	[199.2, 830.8]	<.001
Children 00-04yr?	-642.3	[-829.8, -89.3]	<.001
Children 05-12yr?	654.2	[207.2, 1007.5]	<.001
Children 13-18yr?	701.2	[393.1, 1218.1]	<.001
Adults 45-64yr?	454.9	[265.9, 773.9]	<.001
Number of occupants	1177.8	[842.0, 1401.1]	<.001
Number of adults	840.2	[548.9, 1048.2]	<.001
Length residency	78.8	[62.2, 112.7]	<.001
HH: Average Age	841.2	[492.8, 1049.8]	<.001
HH: 1Ad, 0Child	-1660.9	[-2218.1, -1069.6]	<.001
HH: 2Ad, 0Child	-1132.6	[-1408.9, -560.5]	<.001
weigh. ann. Tav	564.8	[302.3, 851.8]	<.001
weigh. DegreeDays	720.3	[471.6, 1123.4]	<.001
weigh. SolarHours	-201.1	[-319.5, -42.3]	<.001

3.4. General building characterisNcs XGB model

For a XGB baseline model with general building variables, we obtained a MAE and RMSE of respectively 6041 kWh/y and 7856 kWh/y. After hyperparameter tuning with 5-fold cross-validation, a XGB rerun resulted in a model with a MAE and RMSE of respectively 6020 kWh/y and 7828 kWh/y (learning rate = 0.0118, n estimators = 1450, colsample bytree = 0.510, gamma = 0.451, min child weight = 1, max depth = 3, colsample bylevel = 0.563). *Fig. 2* shows the SHAP summary plot for predicting the real total energy use based on general building variables. The most important general building variables are the usable floor space, the building volume, the dwelling type and the calculated energy performance score.

3.5. Socio-demographic & weather XGB model

A XGB baseline model with socio-demographic and weather variables resulted in a MAE and RMSE

of respectively 6993 kWh/y and 9088 kWh/y. After hyperparameter tuning, a XGB rerun resulted in a model with a MAE and RMSE of respectively 6990 kWh/y and 9100 kWh/y (learning rate = 0.0468, n estimators = 370, colsample bytree = 0.817, gamma = 0.352, min child weight = 9, max depth = 4, colsample bylevel = 0.563). *Fig. 3* shows the SHAP summary plot for predicting the real total energy use based on socio demographic and weather variables. The most important socio-demographic and weather variables are the number of occupants, the number of adults, the length of residency and the normalised number of solar hours and degree days.

3.6. Combined XGB model

For the combined 'general building and socio-demographic and weather' XGB baseline model, we obtained a MAE and RMSE of respectively 5747 kWh/y and 7499 kWh/y. After hyperparameter tuning, a XGB rerun resulted in a model with a MAE and RMSE of respectively 5740 kWh/y and 7505 kWh/y (learning rate = 0.0290, n estimators = 1360, colsample bytree = 0.507, gamma = 0.966, min child weight = 8, max depth = 3, colsample bylevel = 0.501). *Fig. 4* shows the SHAP summary plot for predicting the real total energy use for the final combined XGB model. The most important variables are the dwelling type, the usable floor space, the building volume and the number of occupants.

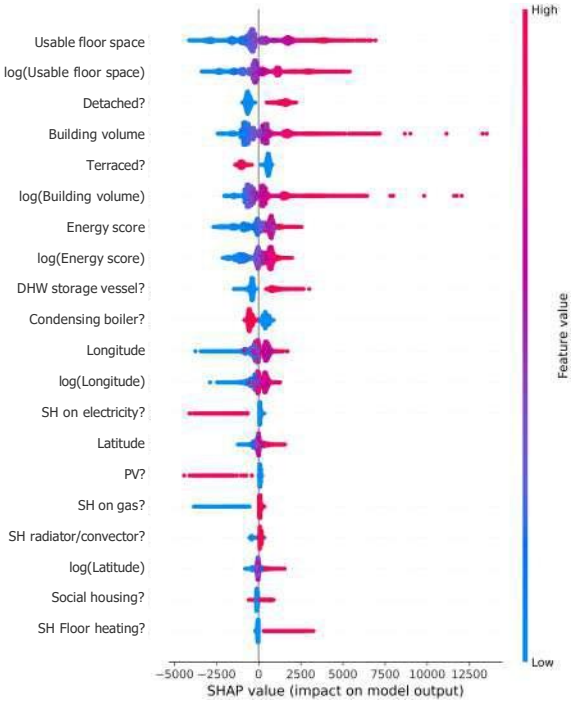

Fig. 2 - Contribution of the 20 most important features in the 'general building variables model'.

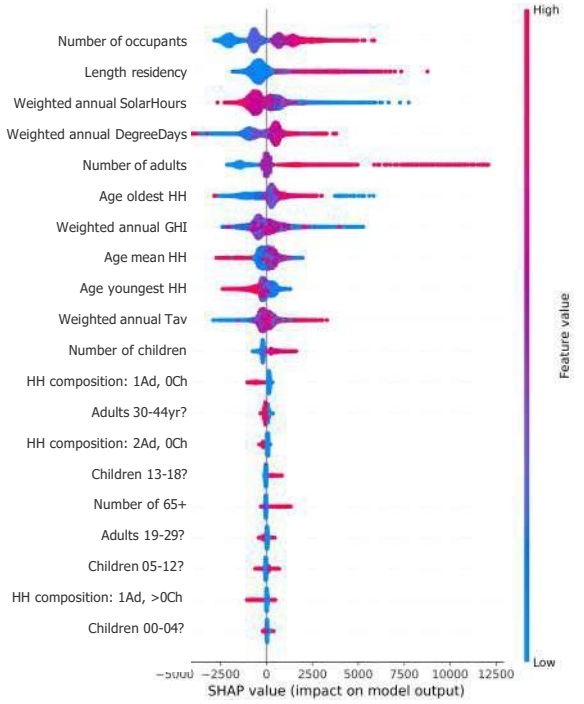

Fig. 3 - Contribution of the 20 most important features in the 'socio demographic and weather variables model'.

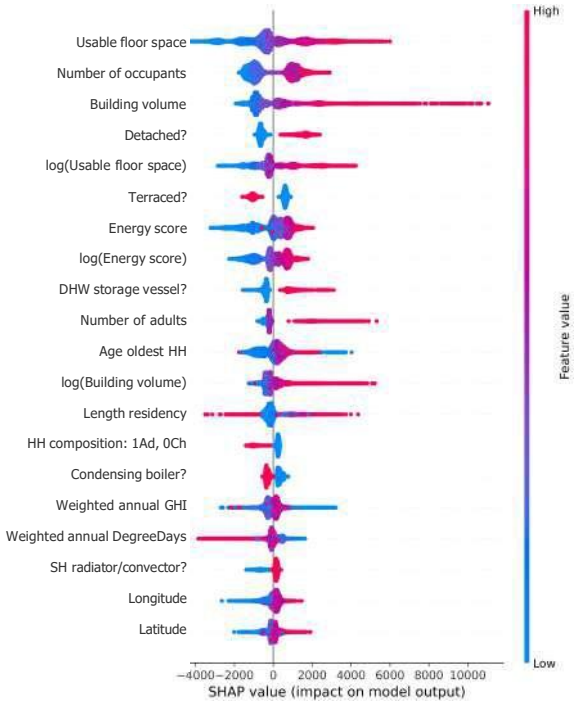

Fig. 4 - Contribution of the 20 most important features in the 'combined model'.

3.7. Regulatory calculaNon method

Fig. 5 shows a scatter plot of the real and regulatory calculated total energy use [kWh/y]. In an ideal scenario, a linear function should closely describe the relationship between both parameters. As expected, an ideal relation is not obtained. In fact, a negative R^2 (R^2=-15%) between both (*N.B.*, R-Squared can be negative only when the chosen model does not follow the trend of the

data, so fits worse than a horizontal line (Hastie *et al.*, 2009; James *et al.*, 2013) indicates that the average real total energy use of the stock is, on average, a better prediction of the real total energy use of an arbitrary single-family house than the annual regulatory calculated total energy use. Furthermore, the RMSE and MAE between both variables are respectively 45808 kWh/y and 35325 kWh/y. When considering that the average real total energy use of the studied sample is 27286 kWh/y, the MAE and RMSE results are 1.29 to 1.68 times larger.

Compared to the performance of the regulatory method, both the linear regression models and the gradient boosting regression trees perform better with results for the MAE respectively 6.05 and 5.15 times better (smaller) and the results for the RMSE respectively 6.37 times and 5.10 times better. Gradient boosting regression trees perform slightly worse than multiple linear regression.

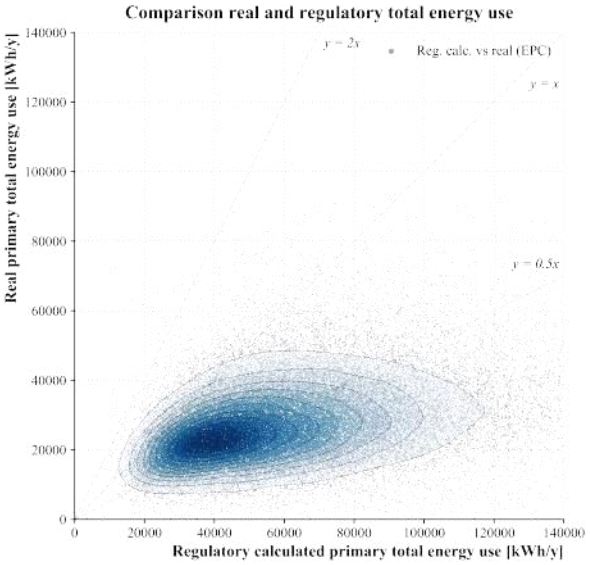

Fig. 5 - Scatter plot of the real and regulatory calculated annual total primary energy use in Flemish single-family houses.

4. Conclusions

This study investigated the predictive performance of data-driven linear regression models and data-driven gradient boosting regression trees for predicting the real annual total building energy use. Also, the results are being compared with the predictive performance of the regulatory calculation methods and there is being evaluated wether these data-driven black-box models can potentially replace the current regulatory white-box models for predicting the annual building energy use at individual building level as well as at stock level.

A total of 46.6% of the variability in total energy use is explained by the final linear regression model based on a combined set of predictors (general building variables, socio-demographic and weather variables) and we obtained MAE and RMSE results of respectively 5011 kWh/y and 6214 kWh/y. For a final XGB model with hyperparameter tuning, based on a combined set of predictors (general building variables, socio-demographic and weather variables), we obtained MAE and RMSE results of respectively 5432 kWh/y and 6543 kWh/y.

Compared to the performance of the regulatory method, both the linear regression models and the gradient boosting regression trees perform better (gradient boosting regression trees slightly worse than multiple linear regression). Yet, a large part of the variance in the linear regression models is left unexplained and also for the gradient boosting trees, there is room for improvement. This means that a large portion of evidence/information has to be attributed either to parameters that are not listed among the variables of the EPB registry (*e.g.*, occupant behaviour, appliance ownership, income) or that the values of the parameters listed are inaccurate (*e.g.*, inaccurate default values).

At individual building level, it is clear that both the linear regression model performance and the gradient boosting regression tree performance is too poor for inference. At stock level, however, both types of models seem promising and can be a useful tool to inform big housing owners (*e.g.*, financial institutions, governments, housing companies *etc.*) or for policy making.

5. Acknowledgements

The authors want to thank the Flemish Energy and Climate Agency (VEKA) and the Belgian grid operator Fluvius for data collection and helpful feedback during the course of the study.

6. Funding

This study and the work by the first author were supported by the VEKA and by a PhD scholarship granted by Ghent University (BOF).

7. References

(1) EU. (2020). Energy performance of buildings directive. European Commission Department of Energy.

(2) Van Hove, M., Delghust, M., Janssens, A. (2021). Analyse naar de haalbaarheid van statistische modellen die energiegebruik in woningen kunnen voorspellen op basis van gebouwparameters. https://www.energiesparen.be/marktonderzoek

(3) EPC Formulestructuur. (2015). Flemish Energy and Climate Agency.

(4) Majcen, D., Itard, L. C. M., & Visscher, H. (2013). Theoretical vs.actual energy use of labelled dwellings in the Netherlands: Discrepancies and policy implications. *Energy Policy*, 54, 125-136.

(5) Van Hove, M., Deurinck, M., Lameire, W., M., Janssens, A., Delghust, M. (2021). Data-driven statistical modelling of real energy use for spatial heating and DHW in modern Flemish single-family houses: a feasibility study. *Proceedings of the 17th International Conference of IBPSA (BS2021)*, Brugge, Belgium.

(6) Zdaniuk, B. (2014). Ordinary Least-Squares (OLS) Model. In: Michalos A.C. (eds) *Encyclopedia of Quality of Life and Well-Being Research.* Springer, Dordrecht.

(7) de Queiroz, G., Vaidyanathan, R. (2016). useR! Machine Learning Tutorial. https://github.com/ledell/useR-machine-learning-tutorial/blob/master/gradient-boosting-machines.Rmd

(8) Hastie ,T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

(9) James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer, New York.

(10) Pedregosa *et al*. (2011). Scikit-learn: Machine Learning in Python. *JMLR*. 12, 2825-2830.

(11) Skipper, Seabold, Perktold, J. (2010). Statsmodel: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*.

(12) McKinney (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*. 445.

(13) Roberts, N., Thatcher, J.B. (2009). Conceptualising and testing formative constructs: tutorial and annotated example. *Data Base for Advances in Information Systems*, 40, 9-39.

(14) Chan, S.H., Chen, J.H. Li, Y.H., Tsai, L.M. (2012). Gly1057Asp polymorphism of insulin receptor substrate-2 is associated with coronary artery disease in the Taiwanese population. *Journal of Biomedical Science*, 19, 1-8.

(15) Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM.

(16) Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., Cox, D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, 8(1), 014008.

(17) Lundberg, S.M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.