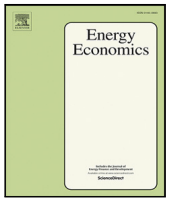
Energy Economics 136 (2024) 107650

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneeco



Learning from experts: Energy efficiency in residential buildings

Monica Billio [a], Roberto Casarin [a,*], Michele Costola [a], Veronica Veggente [b]

[a] Department of Economics, University Ca' Foscari Venezia, Venezia, Italy
[b] Department of Finance, Imperial College Business School, London, UK

ARTICLE INFO

JEL classification:
C10
C53
C50

Keywords:
Energy efficiency
Energy performance certificate
Machine learning
Tree-based models
Big data

ABSTRACT

Reducing energy consumption is a key policy focus for mitigating climate change. This study investigates the determinants of residential building energy efficiency, leveraging expert insights from Energy Performance Certificates (EPCs) to develop a machine learning prediction framework. Datasets from countries at distinct latitudes, the UK and Italy, are analyzed to identify potential regional variations in the factors influencing energy efficiency. Findings reveal the crucial role of factors related to heating systems and insulation materials in the determination of the building's efficiency. Also, there is evidence of the superior ability of non-linear machine learning models to capture complex relationships between building characteristics and efficiency. A scenario analysis further demonstrates the cost-effectiveness of policies informed by machine learning recommendations.

1. Introduction

The increase in greenhouse gas emissions has a relevant impact on global warming as extensively documented in the literature (e.g., see Lashof and Ahuja, 1990; Shine et al., 2005; Hijioka et al., 2006; Yoro and Daramola, 2020). Commercial and residential buildings are responsible for more than 40% of the world's resource and energy consumption and around 33% of the total CO2 emissions (Baek and Park, 2012). Energy efficiency is receiving increasing attention from government and international institutions and represents one of the key policy actions for mitigating global warming and fossil fuel usage (see, e.g. Danish et al., 2019). As an example, in March 2024, the EU Parliament approved the revised Energy Performance Buildings Directive, also known as the "EU Green Homes Directive", aimed at decreasing the environmental impact of Europe's building stock, with targets for zero-emission new residential buildings by 2030 and climate neutrality for all buildings by 2050. The directive emphasizes the significance of financing extensive renovations, encouraging member states to allocate resources to initiatives that guarantee minimal energy savings. Consequently, member states must enact measures to achieve a reduction of at least 16% in average primary energy consumption by 2030 and a reduction of at least 20 to 22% by 2035 in residential buildings.[1] Therefore, policymakers aim to reduce greenhouse gas emissions to decrease the environmental impact of production and consumption activities at the national level and meet the treaties' targets. In this context, green building has emerged as a relevant goal to alleviate the impacts of the building stock on the environment, society, and the economy. The four pillars of green building include minimizing impacts on the environment, improving occupant health conditions, preserving the return on investment for owners and the community, and accounting for the life cycle in the planning and development process. Energy efficiency and greenhouse gas emission reduction represent the key drivers of the environmental impact, along with water efficiency and resource efficiency (Zuo and Zhao, 2014; Robichaud and Anantatmula, 2011).

Along with the environmental impact, reduced energy consumption has relevant consequences also in financial risk management. First, greenhouse gas emissions are one of the main drivers of transition risk (see Basel Committee on Banking Supervision, 2021, for a detailed description of physical and transition risk drivers). Second, recent findings on the mortgage credit market have shown that energy-efficient buildings are associated with a lower solvency risk (Billio et al., 2022; Guin et al., 2022; Ferentinos et al., 2023).

* Corresponding author.
E-mail addresses: billio@unive.it (M. Billio), r.casarin@unive.it (R. Casarin), michele.costola@unive.it (M. Costola), v.veggente23@imperial.ac.uk (V. Veggente).

[1] Please refer to https://www.europarl.europa.eu/news/en/press-room/20240308IPR19003/energy-efficiency-of-buildings-meps-adopt-plans-to-decarbonise-the-sector for more information.

https://doi.org/10.1016/j.eneco.2024.107650
Received 12 October 2023; Received in revised form 15 April 2024; Accepted 17 May 2024
Available online 22 May 2024

The present study aims to identify the key factors representing the necessary technical interventions that could reduce energy consumption in residential buildings. Several definitions of energy efficiency have been provided by policymakers and public policy institutes (Semple and Jenkins, 2020). The European Union defines energy efficiency as "the ratio of the output of performance, service, goods or energy, to the input of energy". The Environmental and Energy Study Institute (EESI) defines energy efficiency as "using less energy to perform the same task – that is, eliminating energy waste".[2] Within the Energy Performance Certificate (EPC) framework, the quantification of a building's energy efficiency is contingent upon its utilization of non-renewable energy sources. In essence, the lower the consumption of non-renewable energy, the higher the level of efficiency attributed to the building.

In the European Union, the EPC mechanism was introduced with the Energy Performance of Buildings Directive (EPBD) to monitor the building stock, and in 2010 new requirements were further added to improve the usability of EPCs in the real estate market.[3] As noted in Schuller (2021), EPC procedures differ across countries and are crucial in measuring the energetic performance of buildings by assigning an overall grade based on the characteristics of the services installed. EPCs contain specific information on the structural characteristics of buildings and services installed, such as heating systems, cooling systems, and domestic water production, with energy sources and consumption measures. Furthermore, it is widely recognized that the opinions about energy efficiency and the effect of hypothetical retrofitting can vary consistently across experts issuing EPCs, even within the same country (Tronchin and Fabbri, 2012).

Recently, the usage of big data in building energy efficiency has been applied to (i) forecast energy demand in residential and commercial buildings (Gómez-Omella et al., 2021; Skomski et al., 2020; Grolinger et al., 2016), (ii) forecast energy efficient enhancement on buildings (Mehmood et al., 2019; Fan et al., 2018), and (iii) evaluate the effectiveness of retrofitting measures (Guzhov and Krolin, 2018) also taking into account the thermal comfort of environmentally friendly constructions (Barbeito et al., 2017).

We investigate the determinants of energy efficiency in residential buildings, proposing a flexible approach to expert opinion analysis, with the aim of establishing an effective predictive framework. Specifically, we consider two geographical areas from the mid-latitude zone (35°–55°) but with different thermal gradients: (i) the Lombardy region in Italy and (ii) the Great London region in the UK. The two areas are expected to experience different extreme climate conditions, such as an increase in the number of hot days and tropical nights, according to most recent climate projections (see, e.g. Carvalho et al., 2021). The public availability of big datasets for the two areas constitutes a unique opportunity to study the effectiveness of machine learning techniques in predicting energy efficiency and providing support to public policies aimed at climate change adaptation and mitigation. The first dataset is the Italian EPCs data, also known as *APE* (*Attestato di Prestazione Energetica*) and focuses on the Lombardy Region that has made publicly available the CENED (Certificazione ENergetica EDifici) database. Beyond energy ratings, the information available in the CENED database relates to the location of certified buildings, the energy demand associated with the services present in the building, the characteristics of buildings, energy systems, and the use of renewable energy sources. The second dataset considers the UK EPCs data, focuses on the London area's residential buildings, and includes information such as average energy efficiency ratings, energy use, carbon dioxide emissions, location, and characteristics of the buildings.

Understanding the relationships between building features and potential energy efficiency improvements is challenging, given the large number of variables involved. In this respect, we employ a comprehensive set of linear and non-linear approaches to delve into these relationships and enhance our understanding of the factors influencing energy efficiency. Among the nonlinear and nonparametric methods, we explore three tree-based models: Bayesian Additive Regression Tree (Chipman et al., 2010), Random Forest (Breiman, 2001), and Extreme Gradient Boosting (Chen and Guestrin, 2016). These non-linear models are widely used to capture non-linear relationships and interactions between variables. In addition, a comparison with benchmark linear models is considered, which includes Lasso (Tibshirani, 1996), Ridge (Hoerl and Kennard, 1970), and Elastic Net (Zou and Hastie, 2005). These models have demonstrated their effectiveness in handling high-dimensional datasets, making them suitable candidates for examining the relationship between building features and energy efficiency potential.

Our findings demonstrate non-linear relationships between building features and efficiency improvements. Specifically, we provide evidence that a set of interventions, such as installing internal or exterior insulation and improving heating systems, as well as the characteristics of buildings, can lead to an improvement in the energy efficiency of a property. We discuss the results obtained from variable importance and partial dependence analyses for Italy and the UK and compare the determinants identified in both cases. The findings of the study reveal that tree-based models exhibit superior predictive performance compared to linear models. The better accuracy in forecasting potential efficiency improvements is attributed to the tree models' ability to capture non-linear relationships between efficiency and building characteristics. Among the tree-based models employed, the Extreme Gradient Boosting model consistently outperforms its counterparts in both in-sample and out-of-sample analyses for the Italian and UK cases.

We conduct a scenario analysis to assess the costs associated with achieving potential energy efficiency, considering two alternative green policies. The first policy leverages technical suggestions generated by the Extreme Gradient Boosting model, selected for its demonstrated superior forecasting capabilities. The second policy is intentionally structured without a specified order of preferences for the recommendations, prioritizing the enhancement of energy efficiency in residential buildings by implementing all provided suggestions outlined in the EPC. To gauge policy effectiveness, we examine payback periods for implemented recommendations in the Italian case and direct costs in the UK case. The results indicate that, in both the Italian and UK contexts, the machine learning-recommended policy shows a greater ability to produce cost-efficient and economically viable outcomes in achieving energy efficiency improvements. In the Italian case, the machine learning policy features an average payback period that is 19.41% lower than that of the alternative policy, while in the UK case, it results in a reduction in carbon dioxide emissions nearly 2.5 times greater. These findings underscore the relevance of machine learning approaches for the formulation and implementation of policies within the realm of energy efficiency initiatives.

The remainder of the paper is organized as follows: Section 2 outlines the variable of interest, and Section 3 introduces tree-based and linear models. Section 4 presents empirical analyses for Italian and UK cases, including discussions on variable selection and non-linear dependencies. Section 5 focuses on scenario analysis for green energy financing policies, and the final section concludes.

2. Modeling energy efficiency

In this section, we present the predicted variable that measures the potential energy efficiency gain following the implementation of the recommendations in the EPC reported by the technicians who conducted the inspection to release the energy certificate. Specifically, the variable of interest is built as described in Section 2.1.

[2] Further information can be found in the following sources: https://www.europarl.europa.eu/RegData/etudes/BRIE/2015/568361/EPRS_BRI(2015)568361_EN.pdf and https://www.eesi.org/topics/energy-efficiency/description.

[3] For an in-depth discussion on the implementation of the Buildings Directive, refer to 2002/31/EC1 and 2010/91/EU, and the EPCs framework in Europe (Arcipowska et al., 2014).

2.1. Definition of efficiency improvement

Energy efficiency is commonly measured using numerical performance indicators of energy consumption, which are then converted into ratings for enhanced interpretability, as exemplified by scales like A-G. Let $\texttt{EE\_POT}_j$ denote the potential final energy performance indicator expected after interventions and $\texttt{EE}_j$ denote the initial energy performance indicator for building j, where $j = 1, \dots, n$. An improvement in energy efficiency implies $0 < \texttt{EE\_POT}_j \leq \texttt{EE}_j < \infty$, reflecting a decrease in energy consumption requirements. Note that if no interventions are required, signifying that the house has reached its maximum potential, we have $\texttt{EE\_POT}_j = \texttt{EE}_j$.

Since the constraint $\texttt{EE\_POT}_j \leq \texttt{EE}_j$ is always satisfied, small values of $\texttt{EE\_POT}_j$ may be attributed to small values of $\texttt{EE}_j$, indicating a high initial efficiency level. To mitigate reliance on initial conditions inherent in the efficiency evaluation process, we consider the energy variation $\texttt{EE\_POT}_j/\texttt{EE}_j$, which falls within the range of $(0, 1)$. The closer the value is to zero, the greater the energy efficiency enhancement.

In order to obtain a variable defined on $(-\infty, +\infty)$, we apply the logistic transform and define:

$$Y_j = \varphi^{-1}(\texttt{EE\_POT}_j/\texttt{EE}_j), \tag{1}$$

where $\varphi^{-1}(v) = \log(v) - \log(1 - v)$ is the inverse logistic. Hence, the response variable Y_j measures the potential variation of the energy performance index, defined as:

$$Y_j = \log(\texttt{EE\_POT}_j) - \log(\texttt{EE}_j - \texttt{EE\_POT}_j). \tag{2}$$

The target variable may exhibit large values under two scenarios: when the building potential energy is low (i.e., the variable $\texttt{EE}_j$ takes large values), or when the efficiency improvement is modest (i.e., the difference $\texttt{EE}_j - \texttt{EE\_POT}_j$ takes small values). In summary, a high (low) value for Y_j is indicative of generally poor (good) energetic performance or potential improvement.

3. Methods

Motivated by the intricate relationships that may exist between the building features and the potential energy efficiency increase, we employ a comprehensive set of linear and non-linear modeling techniques. This approach allows us to gain a deeper understanding of the energy efficiency determinants and facilitates more accurate predictions of the potential energy efficiency increase. Among the linear models, we consider Lasso (Tibshirani, 1996), Ridge (Hoerl and Kennard, 1970), and Elastic Net (Zou and Hastie, 2005), which have demonstrated their effectiveness in handling high-dimensional datasets.

Lasso utilizes ℓ_1 regularization and allows for feature selection, resulting in sparse models. Ridge, employing ℓ_2 regularization, specializes in addressing correlated features. Elastic Net combines both ℓ_1 and ℓ_2 regularization, achieving an approach that considers feature selection and exploits feature correlation. Among the non-linear models, we investigate three tree-based models, namely Bayesian Additive Regression Tree (Chipman et al., 2010), Random Forest (Breiman, 2001), and Extreme Gradient Boosting (Chen and Guestrin, 2016), known for their ability to capture non-linear relationships and interactions in the data.

Let Y_j be the dependent variable measured for the statistical unit j, with $j = 1, \dots, n$, that is, the energy efficiency increase for the jth building in the sample presented in Section 2, and let $\mathbf{x}_j = (x_{1j}, \dots, x_{mj}) \in \mathbb{R}^m$ be a vector of covariates, that are the building and intervention features. The following relationship is assumed

$$Y_j = f(\mathbf{x}_j) + \varepsilon_j, \quad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2), \tag{3}$$

where $f(\cdot)$ is an unknown and possibly nonlinear function. In many applications, the function $f(\cdot)$ may not be smooth, but it could exhibit discontinuities in certain regions of its support.

3.1. Lasso, ridge and elastic net

These are linear parametric models that are $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$, with a penalization that shrinks the coefficients estimates to reduce the overall model complexity (Tibshirani, 1996). Lasso sets a subset of coefficients to zero using an ℓ_1 penalization, Ridge reduces the impact of the features on the response variable, using an ℓ_2 penalization, and Elastic Net combines ℓ_1 and ℓ_2 penalizations. In particular, consider the general minimization problem:

$$\|\mathbf{y} - \beta_0 - X\beta\|_2^2 + \lambda \left[\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2\right], \tag{4}$$

where $\mathbf{y} = (Y_1, \dots, Y_n)$ and $X' = (\mathbf{x}_1', \dots, \mathbf{x}_n')$ are the collections of the target variables and features, respectively. β_0 is the constant, $\boldsymbol{\beta}$ is the m-vector of the regularized coefficients, λ is the regularization parameter, $\alpha \in (0, 1)$ represents the weight for the Lasso component, and $1 - \alpha$, the weight for the Ridge one. We employ **`glmnet`** R-package (see Friedman et al., 2010)[4] to fit three different specifications: (i) Lasso ($\alpha = 1$); (ii) Ridge ($\alpha = 0$); and (iii) Elastic Net ($\alpha = 0.5$). As suggested in Krstajic et al. (2014), λ is validated using the largest value for which the error is within one standard error of the minimum found for λ.

3.2. Bayesian additive regression tree (BART)

The BART model is a flexible inference framework that combines non-parametric regression and ensemble learning (Chipman et al., 2010). BART is a probabilistic framework that captures possible non-linear relationships and interactions among covariates and accounts for uncertainty in the estimates and prediction. The model uses a set of random trees $\mathcal{T}_j$, $j = 1, \dots, J$ to define a flexible functional form for the conditional mean of the variable Y_i. The regression function $f(\cdot)$ is given by a sum of J piece-wise constant functions, $g_j(\cdot)$, called simple functions:

$$f(\mathbf{x}) = \sum_{j=1}^{J} g_j(\mathbf{x}). \tag{5}$$

The simple functions $g_j(\cdot) = g(\cdot; \mathcal{T}_j, \mathcal{M}_j)$ are parametrized by a random tree $\mathcal{T}_j$ and a set of tree-specific coefficients $\mathcal{M}_j = \{\mu_{j1}, \dots, \mu_{jL_j}\}$:

$$g(\mathbf{x}; \mathcal{T}_j, \mathcal{M}_j) = \sum_{l=1}^{L_j} \mu_{jl}\mathbb{I}(\mathbf{x} \in \mathcal{X}_{jl}), \tag{6}$$

where $\mathbb{I}(x \in A)$ is the indicator function, which takes the value 1 if x is in the set A and 0 otherwise, and $\mathcal{X}_{jl} \in \mathbb{R}^m$.

Each random tree $\mathcal{T}_j$ contains a set of internal and terminal nodes (leaves). Each internal node is associated with a binary splitting rule such that the node is connected to two child nodes: a left node when the kth variable is below a threshold c_j, that is $X_{ik} \leq c_j$ and a right node when the kth variable is above, that is $X_{ik} > c_j$. A leaf node, say l, has no splitting rule and is assigned to a parameter μ_{jl}. The tree is random since the choice of the splitting variable and the value of the parameter at the terminal nodes are random, which adds flexibility to the model.

Each tree generates a partition $\mathcal{X}_{j1}, \dots, \mathcal{X}_{jL_j}$ of the covariate space $\mathbb{R}^m$ such that $\mathcal{X}_{jl} \cap \mathcal{X}_{jl'} =$ for $l' \neq l$ and $\mathcal{X}_{j1} \cup \cdots \cup \mathcal{X}_{jL_j} = \mathbb{R}^m$. In the BART model, the parameter μ_{jl} represents the contribution given by the jth tree to the conditional expected value of Y_i when $\mathbf{X}_i$ in the lth element of the partition, given the random partition induced by the jth tree.

The specification of the BART model includes the prior distribution on the tree structures, the leaf parameters, and the variance of the error term

$$\pi(\mathcal{T}_1, \dots, \mathcal{T}_J, \mathcal{M}_1, \dots, \mathcal{M}_J, \sigma^2) = \pi(\sigma^2)\prod_{j=1}^{J}\pi(\mathcal{M}_j|\mathcal{T}_j)\pi(\mathcal{T}_j). \tag{7}$$

[4] The **`glmnet`** R-package, developed by Trevor Hastie and Rob Tibshirani, is available for download at https://cran.r-project.org/web/packages/glmnet.

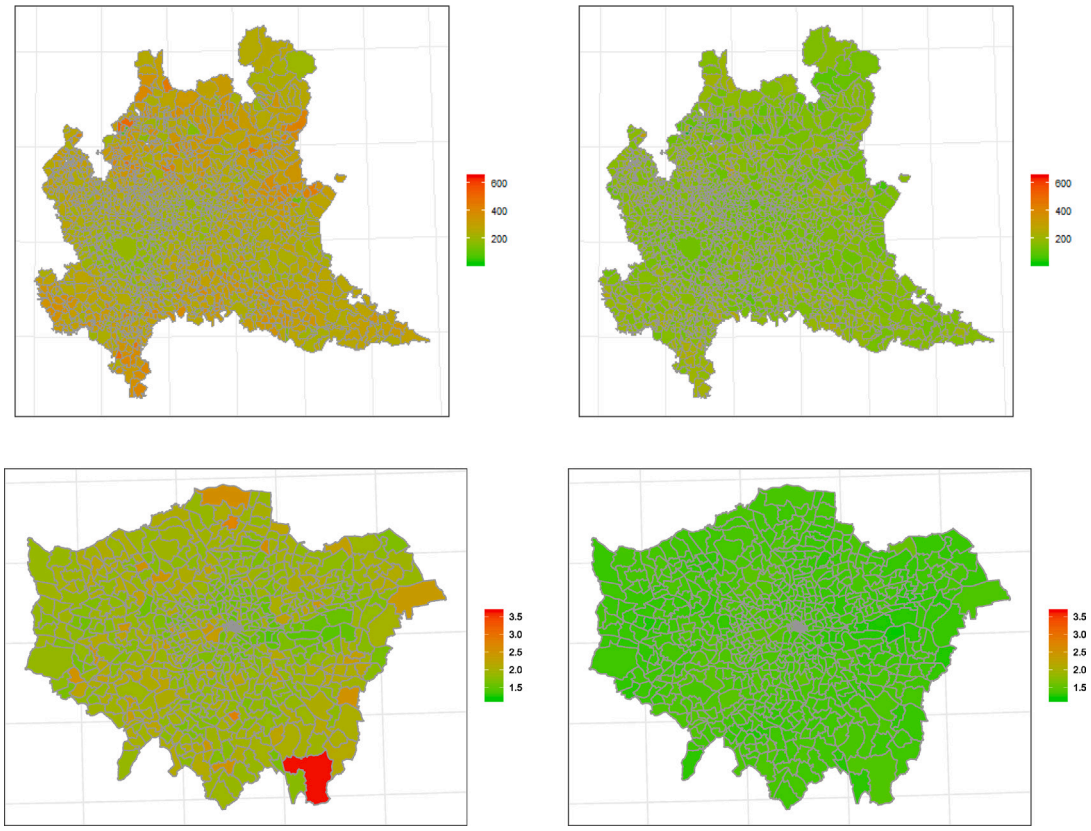


Fig. 1. Geographic distribution of the initial (left) and the potential (right) energy efficiency in the Lombardy region of Italy (top) and in the Greater London area of the UK (bottom). In each plot: the color indicates the efficiency level from high (green) to low (red), and the gray lines provide the limits of the administrative units in longitude (horizontal axis) and latitude (vertical axis) coordinates. The red area in the UK map refers to the ward of Darwin, where around 6% of the postcode areas have lower energy efficiency than the average of the least efficient 0.001% postcode areas in Greater London. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We consider here the choice for $\pi(\mathcal{T}_j)$, which is given by the product of the following prior distributions: (i) a prior distribution $\alpha(1+d)^{-\beta}$ for the depth $d \in \{0,1,2,\ldots\}$ of the tree with $\alpha \in (0,1)$ and $\beta \in [0,\infty)$; (ii) independent normal distributions $\mathcal{N}(m_\mu, \sigma_\mu)$ for the leaf parameters μ_{jl}; and (iii) the conjugate scaled inverse Chi-square prior distribution $\nu\lambda\mathcal{X}^2(\nu)$ for σ^2. Regarding the splitting rule, at each internal node, each splitting covariate has an equal prior probability of being chosen, i.e. $1/n$ (see, for instance, Chipman et al., 2010; Pratola, 2016; Linero, 2018). The posterior distribution is not tractable and following the standard practice in Bayesian analysis, it has been approximated numerically via a Markov Chain Monte Carlo (MCMC) algorithm that generates samples from the parameter and tree posteriors and from the posterior predictive. In the application we considered the following hyper-parameter setting: $\alpha = 0.95$, $\beta = 2$, $m_\mu = 0$ and $\sigma_\mu^2 = (y_{\max} - y_{\min})/(2k\sqrt{J})$, $k = 2$, $\nu = 3$, and $\lambda = 0.1468$, where $y_{\min} = \min\{y_1,\ldots,y_n\}$ and $y_{\max} = \max\{y_1,\ldots,y_n\}$. See also (Sparapani et al., 2021) for further discussion on the prior choice. We use the R implementation of the MCMC algorithm included in the packages `BayesTree` (Chipman and McCulloch, 2016) and `BART` (Sparapani et al., 2021). To select the number of trees k, we perform a cross-validation exercise as reported in 3.

3.3. Random forest

This nonparametric model can capture non-linearity in predicting the energy efficiency gain. Similarly to BART, it relies on the notion of a decision tree given in the previous section. The Random Forest model, introduced by Breiman (2001), is based on a combination of single decision trees trained in parallel on random subsets of the data. At each node, a subset of the total number of features is selected as candidates to define the splitting rule. This ensures that the model can handle the correlation between features and grows somewhat uncorrelated trees. See Casarin et al. (2021) for an introduction to random forests with applications.

We employ the `randomForest` R-package (Liaw and Wiener, 2002)[5], setting the number of trees equal to 150, and leaving all other parameters to the default values. In the application to the full sample, we set the maximum number of nodes equal to 500 for computational reasons. We do not restrict this parameter in the application to the subsample.

3.4. Extreme gradient boosting (XGBOOST)

The second model we consider is an ensemble model based on a collection of decision trees $\mathcal{T}_j$, $j = 1,\ldots,J$, a collection of functions $g_j(\cdot;\mathcal{T}_j)$, $j = 1,\ldots,J$ with $g_j \in \mathcal{G}$ and the additive regression function in Eq. (5), where $\mathcal{G} = \left\{g(\mathbf{x}) = w_{q(\mathbf{x})},\ q : \mathbb{R}^m \to 1,\ldots,L, w \in \mathbb{R}^L\right\}$ is the space of regression trees and L is the number of leaves. The main difference compared to BART is that trees, in this case, are grown sequentially on a modified version of the original dataset. At the iteration t, given a set of trees $g_1,\ldots,g_J$, a new tree $g(\mathbf{x}) \in \mathcal{F}$ is included to obtain a new regression function

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + g_{J+1}(\mathbf{x}). \tag{8}$$

[5] The `randomForest` R-package, developed by Andy Liaw and Matthew Wiener, is available for download at https://cran.r-project.org/web/packages/randomForest.

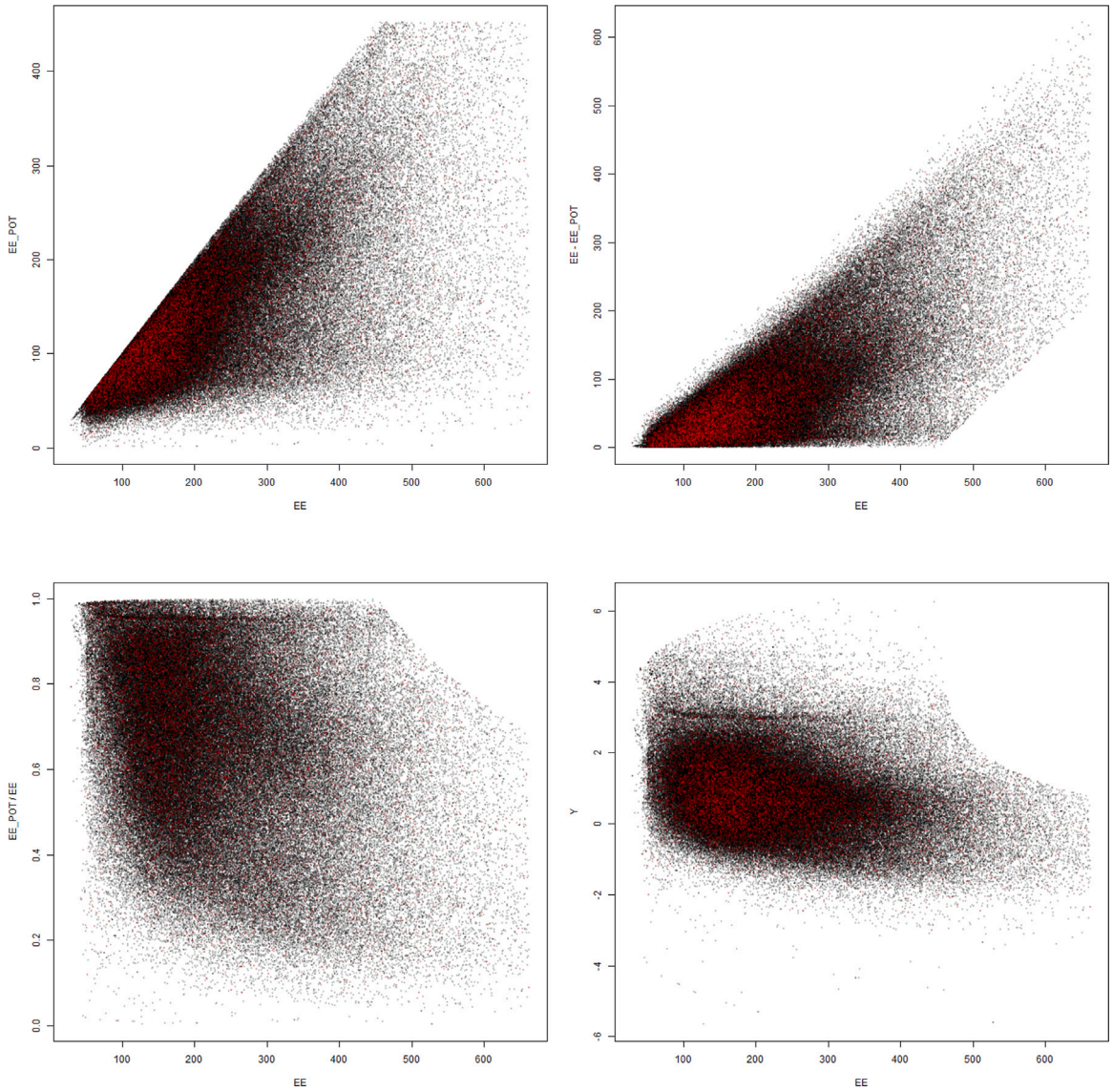


Fig. 2. The case of Lombardy (Italy). The initial non-renewable energy performance index EE (horizontal axis) versus the final index EE_POT (vertical axis, top left), the expected performance difference EE - EE_POT (vertical axis, top right), the expected performance ratio EE_POT/EE (vertical axis, bottom left), and the response variable $Y = \log(\text{EE\_POT}) - \log(\text{EE} - \text{EE\_POT})$ (vertical axis, bottom right). The entire sample involves 205,049 buildings (gray dots) and a subsample of 10,000 buildings (red dots).

The newly added tree g_{J+1} is chosen based on the errors produced by the trees of the previous iteration (Chen and Guestrin, 2016). This algorithm is designed to learn slowly from the data, which helps avoid overfitting. For the estimation, we utilize the **xgboost** R-package (Chen and He, 2023)[6] and cross-validate the value for the maximum number of iterations, using both the full sample and a subsample of 10,000 observations.

4. Empirical analysis

In this section, we apply the presented models to the EPC data for the two geographical areas with different latitudinal temperature gradients and climate conditions: the Lombardy region in the north of Italy and the Greater London area in the UK.

To predict the potential increase in energy efficiency, we leverage the technical specifications outlined in the EPC, coupled with the expert-recommended interventions derived from the assessment process. Illustrated in Fig. 1, the spatial distribution delineates the energy efficiency measure across both datasets. The left column portrays the current energy-efficient status, while the right column projects the potential energy efficiency level, factoring in the proposed expert interventions. This visual representation highlights the substantial enhancements achievable through the implementation of expert-guided recommendations.

In the first part of each country analysis, we present and compare results obtained between tree-based and linear models. The models are applied to predict the energy efficiency potential improvement, leveraging on granular information on the initial characteristics of the stock of buildings, their energy services, and the interventions recommended by technicians. In order to ensure a comprehensive analysis, we conduct the applications on both the full sample and a subsample of the data by selecting a random sample without replacement of 10,000 observations (see, for instance, García et al., 2015). On one hand, the inclusion of the full sample allows us to capture the overall trends and patterns present in the dataset, providing a broader perspective on the relationship between the predictors and the target variable. On the

[6] The **xgboost** R-Package, developed by Jiaming Yuan, can be downloaded from https://cran.r-project.org/web/packages/xgboost/index.html.

other hand, the subsample analysis is particularly valuable in making the data-driven framework computationally feasible and applicable to real-time decision-making scenarios. By examining a smaller subset of the data, we can ensure its robustness and evaluate its ability to generalize across different data distributions.

In the second part, we focus on the variable importance to gather a comprehensive understanding of how the two modeling methods assign significance to the different features and expert opinions under scrutiny.

4.1. Data source and description

Variables of interest in the EPC databases can be grouped as follows: (a) initial characteristics of the building, its services, and consumption levels; (b) current energy efficiency; (c) suggested interventions; (d) potential energy efficiency once the interventions are implemented.

1. **Initial characteristics:** these include data related to the (i) general characteristics of the building (such as intended use, location, age of the building, size of the real estate unit, number of real estate units in the building,...); (ii) energetic services installed in the building such as heating system, cooling system, production of domestic water;
2. **Current energy efficiency:** consumption and energetic performance, expressed using a number of different indicators such as thermal efficiency, global energetic performance of renewable and non-renewable energetic sources, consumption level for different fuel types, energetic class, and similar;
3. **Suggested interventions:** in the EPC, experts are required to report one or more possible interventions to increase the energy efficiency level of the building;
4. **Potential energetic performance:** variables summarizing the estimated potential energy class given the initial conditions of the building and the implementation of one or more suggested interventions.

In the analysis, features falling in the first and third categories are used as predictors to forecast the potential increase in energy efficiency. As described in Section 2, the measure is computed using energy performance indicators included in the second and fourth categories and is obtained from Eq. (2). The two datasets for different geographical areas are the CENED+2 dataset for Lombardy (Italy) and the EPBD UK dataset (UK). The former pertains to EPCs issued for buildings in the Lombardy region from January 1, 2015.[7] The EPBD dataset for the UK encompasses Energy Performance Certificate (EPC) data issued for domestic buildings in England and Wales from January 1, 2008.[8]

4.2. The Italian case

In the original Italian dataset, energy efficiency is measured through the energy performance indicator of all the non-renewable sources used in a building. Panel C in Table A1 presents the original Italian labels for initial (EE) and potential (EE_POT) energy efficiency and a description of the indicator.

The response variable Y, as defined in Section 2, exhibits an inverse relationship with the enhancement in energy efficiency and solely captures positive enhancements. In simpler terms, the value of (EE - EE_POT) is non-negative. Thus, a high value for Y signifies a minimal enhancement rather than a decline in energy efficiency (refer to the upper-left panel in Fig. 2). Intriguingly, buildings with lower energy efficiency are anticipated to experience more substantial improvements (as evident from the lower limit in the upper-right panel of Fig. 2). This observation implies that when a building's initial energy efficiency is exceedingly low, any proposed intervention is likely to yield some degree of enhancement. Conversely, highly efficient buildings cannot exhibit significant performance increments.

Generally, the maximum potential improvement in energy efficiency tends to decrease as the initial energy efficiency of the building increases, as depicted in the upper bound of the upper-right panel of Fig. 2. Interestingly, the upper bound in the bottom-left panel of Fig. 2 highlights that when EE is approximately below 450 kWh/m^2, the attainable improvement remains below 1. This underscores a technological constraint within the building that hinders a complete elimination of inefficiency, preventing it from reaching the highest energy class.

Table 1 provides an overview of the six distinct structural interventions that experts can recommend for improving a building's energy efficiency within the context of the Italian EPC. These recommendations encompass a range of areas including the building's shell, heating and cooling systems, other systems, and renewable sources. The focus is solely on EPCs containing at least one recommendation from technicians, as each recommendation implies a potential increase in energy efficiency.

From the initial dataset, we exclude observations that do not pertain to private residential, single-unit, and non-publicly used buildings.[9] It is worth noting that EPC information is manually reported in the CENED2+ database, introducing the possibility of typos and inconsistencies. Consequently, we remove outlier observations with initial or potential EE values below the 1st or above the 99th percentile. Similarly, we exclude buildings with null potential energy efficiency increases or a potential overall decrease in energy class. Hence, records with null or negative potential improvements in energy efficiency are deemed irrelevant or erroneous to the purpose of the study.

A comprehensive description of the database can be found in the online Appendix A1. The dataset used in subsequent predictive analysis comprises 205,049 complete observations and 42 covariates in total, described in Panel A of Table A1.[10] Histograms depicting the composition of the full dataset and the subset of complete cases in terms of initial energy efficiency, construction period, year of EPC issuance and climatic area can be found in Figures A2 and A3.

As discussed in the previous section, we also consider a subsample to reduce computational costs in real-time scenario analyses. Working on a subsample allows for a relevant decrease in execution time and computational costs leading to almost unchanged results in terms of predictive accuracy.[11] Consequently, results on the entire dataset are

Table 1
Labels for recommendation identifiers in the Italian dataset, presented in both the original Italian and translated English forms.

IMPROVEMENT_ID	English description	Italian description
1	Opaque shell	Involucro Opaco
2	Transparent shell	Involucro Trasparente
3	Heating System	Impianto climatizzazione Inverno
4	Cooling System	Impianto climatizzazione Estate
5	Other Systems	Altri Impianti
6	Renewable Sources	Fonti Rinnovabili

[7] The full dataset "*CENED+2 Database – Certificazione ENergetica degli EDifici*" is publicly available at https://www.dati.lombardia.it/Energia/Database-CENED-2-Certificazione-ENergetica-degli-E/bbky-sde5 and can be used under the Creative Commons Licence Zero (CC0 1.0 universal).

[8] The full dataset "*Energy Performance of Buildings Data: England and Wales*" is publicly available at https://epc.opendatacommunities.org/.

[9] The analysis focuses on residential buildings classified as E.1(1) and E.1(2) according to the DPR classification. Additional information is available in the *Gazzetta Ufficiale*, https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg.

[10] For a detailed overview of the dataset, please refer to https://www.dati.lombardia.it/Energia/Database-CENED-2-Certificazione-ENergetica-degli-E/bbky-sde5.

[11] In the Online Appendix A4, we delve into the execution time of the proposed model specifications on both the full sample and the subsample.

compared with the one obtained in the subsample (about 4.9% of the whole sample). In both sampling schemes, the whole and the thinned sample, we split the dataset into a training set (in the sample, 70% of the observations) and a test set (out-of-sample, 30%).

4.2.1. Forecasting results in the Italian case

The comparison of predictive performance between tree-based regression models and linear models, including LASSO, RIDGE, and ELASTIC NET, is depicted in Table 2. In terms of the correlation between the predicted and actual Y indicator, the tree-based regression model consistently outperforms the linear models for both the full sample and subsample. This improvement in correlation is observed across the in-sample and out-of-sample analyses. In the full sample, the tree-based regression models demonstrate a correlation above 0.71 (0.68) for the out-of-sample in the full sample (subsample), whereas the linear models exhibit correlations around 0.63 (0.57). Notably, while RANDOM FOREST and XGBOOST models show slightly better performance in the out-of-sample results, the correlation levels of the tree-based regression model remain consistently aligned.

The table presented here offers an insightful comparison of predictive model performance, focusing specifically on the Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics. For the full sample, we observe interesting patterns in terms of MSE and MAE values.

In both in-sample and out-of-sample scenarios, linear models – such as LASSO, RIDGE, and ELASTIC NET – exhibit varying levels of performance, indicating a consistent alignment between the two samples. In the out-of-sample case, LASSO and ELASTIC NET produce nearly identical MSE and MAE values, with MSE at 0.779 and MAE at 0.6755. RIDGE exhibits slightly higher MSE (0.8552) and MAE (0.7139) values. Transitioning to the subsample analysis, we observe a continuation of consistent trends. Here, the performance of the linear models remains steady, with both LASSO and ELASTIC NET displaying similar values of MSE and MAE, both of which outperform the RIDGE case. This observation emphasizes the stability of these models, particularly in the context of the subsample.

The tree-based models, specifically BART, RANDOM FOREST, and XGBOOST, demonstrate remarkable performance superiority in both the full and subsample datasets compared to the linear models. Within the full sample, XGBOOST showcases the highest accuracy in terms of MSE and MAE, trailed by the BART and RANDOM FOREST models. In the subsample, the performance of XGBOOST and BART remains the highest but experiences a noticeable decline between in-sample and out-of-sample scenarios, while RANDOM FOREST maintains greater stability in its performance across the two situations. Given its consistently superior performance in both in-sample and out-of-sample analyses within the full - and sub-samples, XGBOOST emerges as the preferred choice for predicting energy efficiency improvements.

The divergence in performance between linear and tree-based models can be attributed to the non-linearities inherent in the EPC data incorporating the building characteristics, the energetic performance, and technicians' recommendations to reduce the building's energy consumption. The ability of tree-based models to better accommodate these complexities underscores their utility in accurately representing and predicting energy efficiency improvements.

4.2.2. Most relevant variables for the Italian case

Variable importance holds a crucial significance in comprehending the individual contributions of features to the predictive outcomes of machine learning models. Within the context of EPCs, this analysis assumes even greater relevance as it sheds light on the relative influence of each feature, encompassing building characteristics and technician recommendations, in predicting energy efficiency improvements.

Table 3 presents the ranking of variable importance, including the bar chart weights for the top 15 variables, across the considered range of models encompassing linear methods (LASSO, RIDGE, and ELASTIC NET), as well as tree-based approaches (BART, RANDOM FOREST, and XGBOOST).[12] These models encompass selected variables derived from both building characteristics and technician recommendations, denoted by the label "`R_:`". Notably, the preeminent variable across all models is the "`R1: Opaque Shell`" recommendation, representing one of the six potential suggested implementations. This recommendation involves applying insulating materials to the solid structural components, aimed at enhancing the building's thermal performance by reducing heat loss in colder periods and heat gain in hotter periods. This practice significantly contributes to energy efficiency by diminishing the necessity for heating and cooling systems, resulting in decreased energy consumption and utility bills. Another feature consistently present in all models, albeit with varying levels of importance, is the number of recommendations. The interpretation is straightforward: the greater the count of suggested interventions proposed by experts, the greater the potential enhancement in energy efficiency. Other building characteristics encompass factors such as "`EE_WINTER`" (Energy Efficiency in Winter) and "`AGE_BAND`" (Construction period). When examining the tree-based models, in addition to R1 and the number of recommendations, it becomes clear that frequently chosen variables include "`THERMAL_EFFICIENCY`", "`SV_RATIO`" (Surface/Volume ratio), and "`CURRENY_ENERGY_EFFICIENCY_REN`" (Current energy efficiency for renewables) which further emphasizes the significance of current structural attributes in determining a building's potential energy efficiency gain. Regarding other recommendations made by the experts in the EPC, "`R2: Transparent Shell`" is selected both

Table 2
The Italian case. Correlation (top panel), Mean Square Error (mid panel), and Mean Absolute Error (bottom panel) between actual and predicted values estimated by Lasso, Ridge, Elastic Net, BART, Random Forest, and XGBoost. In-sample and out-of-sample results for the whole sample (first and second column) and a random subsample (third and fourth column).

	Full sample		Subsample	
	In sample	Out of sample	In sample	Out of sample
	Correlation			
LASSO	0.6351	0.6364	0.6150	0.6135
RIDGE	0.6219	0.6231	0.4927	0.4951
ELASTIC NET	0.6354	0.6367	0.6111	0.6113
BART	0.7259	0.7139	0.7236	0.6766
RANDOM FOREST	0.7028	0.7052	0.6879	0.6831
XGBOOST	0.7705	0.7292	0.7976	0.6684
	Mean Square Error			
LASSO	0.7732	0.7795	0.7998	0.8350
RIDGE	0.8471	0.8552	1.2682	1.3210
ELASTIC NET	0.7729	0.7792	0.8095	0.8429
BART	0.6125	0.6415	0.6085	0.7197
RANDOM FOREST	0.6583	0.6613	0.6716	0.7087
XGBOOST	0.5275	0.6126	0.4778	0.7347
	Mean Absolute Error			
LASSO	0.6731	0.6755	0.6867	0.7057
RIDGE	0.7096	0.7139	0.8794	0.9049
ELASTIC NET	0.6730	0.6755	0.6912	0.7102
BART	0.5956	0.6071	0.5997	0.6524
RANDOM FOREST	0.6184	0.6197	0.6251	0.6413
XGBOOST	0.5513	0.5914	0.5315	0.6580

[12] Feature importance is obtained for each model as follows: (i-ii-iii) LASSO, RIDGE, and ELASTIC NET: (`caret::varImp`) – the absolute value of the t-statistic for each model parameter; (iv) BART: (`BART::wbart$varcount.mean`) – mean of the total count of the number of times that a variable is used in a tree decision rule (over all trees); (v) RANDOM FOREST (`randomForest::importance`) – total decrease in node impurities from splitting on the variable, averaged over all trees, measured by the residual sum of squares; and (vi) XGBOOST: (`xgboost::xgb.importance`) – the fractional contribution of each feature to the model determined by the cumulative gain from the splits involving that particular feature.

Table 3
The Italian case. Variable importance rankings, displaying bar chart weights for the top 15 variables across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST models.

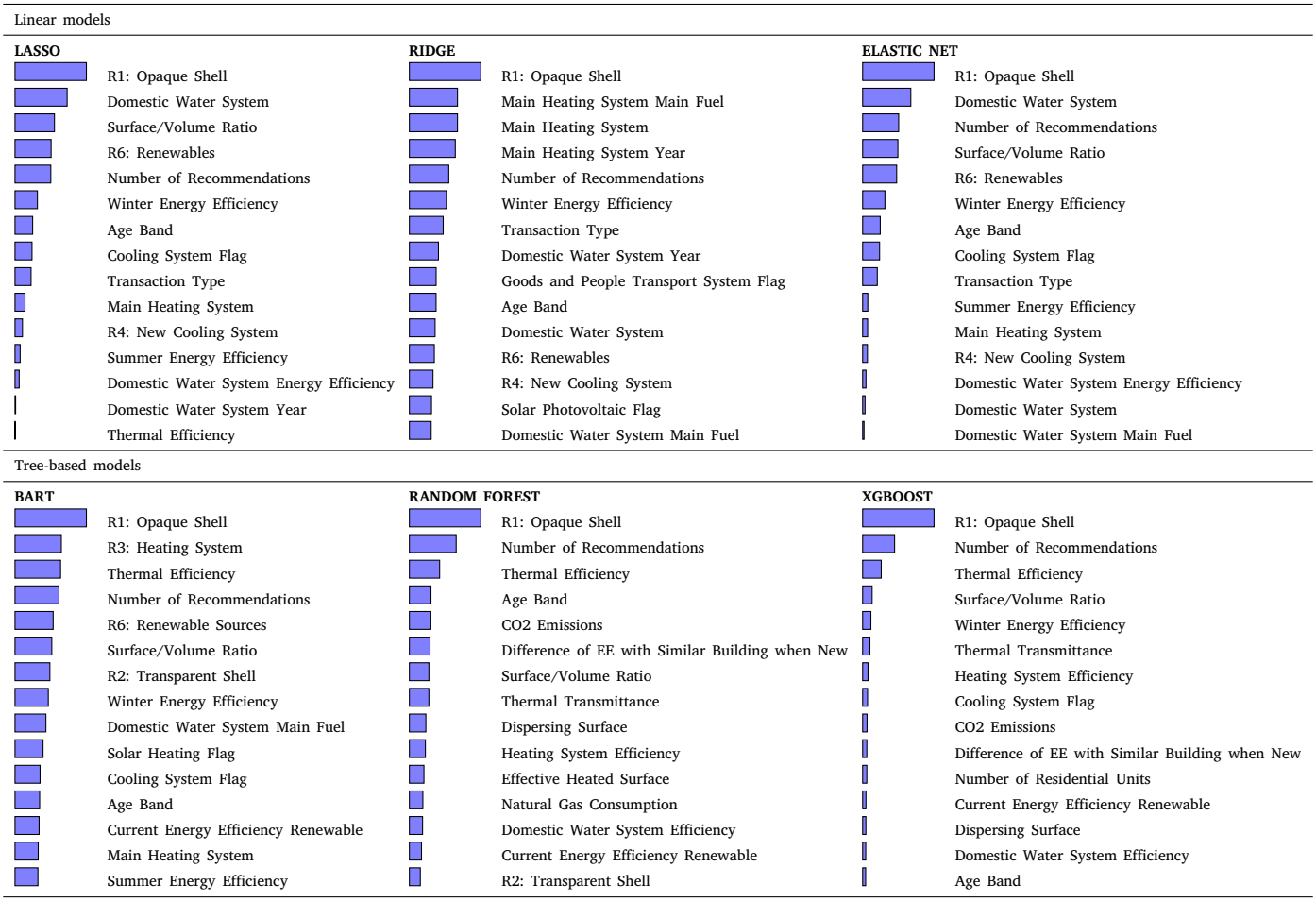


Table 4
The Italian case. Rank correlation between variable importance ranking across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST models.

	LASSO	RIDGE	ELASTIC NET	BART	RANDOM FOREST	XGBOOST
LASSO	–					
RIDGE	0.42	–				
ELASTIC NET	0.86	0.51	–			
BART	0.09	0.08	0.18	–		
RANDOM FOREST	−0.10	−0.02	0.02	0.69	–	
XGBOOST	0.02	−0.08	0.13	0.60	0.80	–

by the RANDOM FOREST and the BART models. Transparent shells impact energy consumption through several mechanisms. Firstly, they allow natural light to penetrate indoor spaces, reducing the need for artificial lighting during daylight hours. This contributes to energy savings and decreases electricity consumption. Secondly, transparent shells influence the thermal performance of a building. While they allow solar radiation to enter, they can also lead to heat gain, especially during warmer periods. To mitigate this, advanced glazing systems with low solar heat gain coefficients are often employed, diminishing the influence of solar radiation on indoor temperatures and cooling systems. Efforts to enhance energy efficiency in buildings encompass the utilization of double or triple glazing, low-emissivity coatings, and insulated frames to curtail heat transfer through windows and mitigate thermal bridging. The BART model selects two additional recommendations, namely the "`R3: Heating System`" and "`R6: Renewable Sources`". The former typically involves upgrading or optimizing components such as boilers, radiators, and heat pumps to reduce heat loss during colder months, improve heat distribution, and enhance overall energy performance. The latter entails harnessing solar energy, wind power, hydropower, and geothermal energy to generate electricity or heat, thereby reducing reliance on non-renewable energy sources.

Additionally, the rank correlation presented in Table 4 provides valuable insights into the consistency of variable importance rankings among different machine learning models. This analysis sheds light on the robustness and stability of the feature selection process, offering a deeper understanding of which variables consistently contribute to the predictive performance across diverse modeling techniques. Notably, the linear models – LASSO, RIDGE, and ELASTIC NET – exhibit a varying correlation with each other, with values ranging from 0.42 to 0.86. The highest correlation for linear models is in the case of LASSO and ELASTIC NET. On the other hand, the tree-based models – including BART, RANDOM FOREST, and XGBOOST – show correlation values ranging from 0.60 to 0.80. This suggests that these models consistently

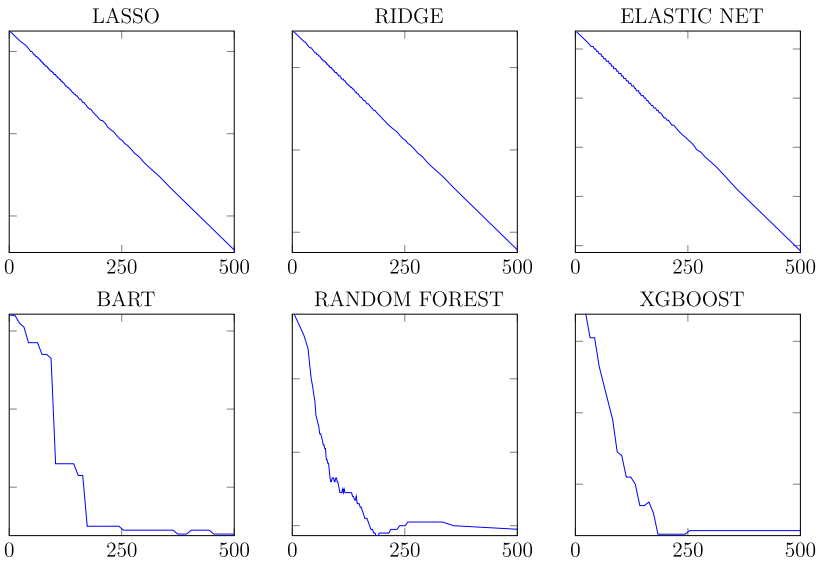


Fig. 3. Partial Dependence Plots (PDPs) in the Italian case for the "THERMAL_EFFICIENCY" characteristic across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST. For each model, the plot illustrates the relationship between thermal efficiency (measured in kWh/m^2) on the x-axis and the predicted potential variation of energy performance on the y-axis. The limits of the latter are set by series min–max values.

agree on the relative importance of variables for predicting energy efficiency improvement Comparatively, the linear and tree-based models show almost no correlation, indicating a lack of commonalities in the feature importance rankings. The highest correlation values are observed in the case of ELASTIC NET with XGBOOST and BART, having values of 0.13 and 0.18, respectively.

The previous findings highlight the inherent non-linear nature of the data, which makes tree-based models more effective at revealing patterns that linear models cannot capture. A valuable tool for further exploration of this point is the Partial Dependence Plot (PDP), derived from the partial dependence function (Friedman, 2001), which illustrates the dependency of the potential variation of energy performance on a specific building's characteristics. As an illustrative example, Fig. 3 provides for each considered model the PDP for the "THERMAL_EFFICIENCY" characteristic which measures in kilowatt-hours per square meter (kWh/m^2) how efficiently a building can be heated during the winter. Tree-based models reveal a negative non-linear relationship and suggest that the potential improvement in energy efficiency plateaus beyond 200 kWh/m^2. As a reference, we also include in the figure the PDP for the linear models, which is indeed linear. This inherent non-linearity is a key factor contributing to the superior forecasting abilities of tree-based models. Buildings exhibit intricate relationships between their characteristics and potential energy efficiency improvements. This complexity appears to be better captured by tree-based models compared to linear models.

4.3. The UK case

In the analysis, we focus on residential buildings in the London area (local area codes from E09000001 to E09000033) and consider EPCs issued between 2015 and 2021.[13] We select the current and potential energy efficiency indicators, CURRENT_ENERGY_EFFICIENCY and POTENTIAL_ENERGY_EFFICIENCY, which account for the cost of energy required for space and water heating and lighting multiplied by fuel costs.[14] This indicator considers the cost of energy and is expressed in £/m^2/year, where cost is derived from kWh.

Importantly, it should be noted that in the context of Italy, a higher value of the energy efficiency indicator is indicative of lower energy efficiency. Conversely, in the UK, large values of the energy efficiency index correspond to heightened energy performance. For consistency of the two cases, we consider the inverse of the above indicators, i.e. EE = (1/CURRENT_ENERGY_EFFICIENCY) · 100 and EE_POT = (1/POTENTIAL_ENERGY_EFFICIENCY) · 100 to compute the target variable as in Eq. (2).

Table 5 shows 63 structural interventions that experts can recommend for improving a building's energy efficiency. The recommendation identifiers in the table have been re-encoded to address duplicates in the dataset (shown in bold) and ensure that the labels accurately represent each unique intervention. For instance, improvement IDs 11, 12, 13, 14, 15, 17, and 18 were consolidated into a single intervention labeled "Upgrade heating controls", which emphasizes the same recommended action of enhancing heating controls across multiple instances. Other examples of re-encoded recommendations include "Replace boiler with new condensing boiler" (IDs 20 and 21) and "Wood pellet stove with boiler and radiators" (IDs 23 and 39), reflecting the consolidation of similar interventions under standardized labels. The total number of interventions after the re-encoding process is 41.

In contrast to the Italian dataset, which encompasses a more limited set of 6 recommendations, the UK dataset exhibits a higher level of granularity and diversity in the types of interventions such as upgrading heating controls, insulation enhancements for different building components, replacement of heating systems with more efficient alternatives, installation of renewable energy sources like solar panels and wind turbines, as well as improvements in lighting and glazing.

Finally, we consider a record complete when data points are provided for all 82 features involved.[15] The dataset initially contains 1,041,806 rows, which is reduced to 445,661 complete records after cleaning. Histograms showing the composition of the full dataset and the subset of complete cases in terms of construction period, year of EPC issuance and initial energetic class can be found in Figures A5 and A6. A subsample of 10,000 units is randomly selected from this complete set, mirroring the approach undertaken in the Italian case. The comprehensive description of the database can be found

[13] For a detailed description of the variables, the reader can refer to the guidance page available at https://epc.opendatacommunities.org/docs/guidance.

[14] See Table A6 in the online Appendix A1 for a description of the variables included in this analysis.

[15] To handle missing values, we remove all the records including "NA", "N A", "N/A", "N/A", "N/A", "NO DATA!", "INVALID!", "Not recorded", "Not applicable", or empty data points.

Table 5
Original (first column) and re-coded (second column) recommendation identifiers, along with detailed descriptions of the interventions (third column), in the context of the UK dataset. Duplicates are highlighted in bold.

Improvement ID	New improvement ID	Description
1	1	Insulate hot water cylinder with 80 mm jacket
2	2	Increase hot water cylinder insulation
3	3	Add additional 80 mm jacket to hot water cylinder
4	4	Hot water cylinder thermostat
5	5	Increase loft insulation to 270 mm
6	6	Cavity wall insulation
7	7	50 mm internal or external wall insulation
8	8	Replace single glazed windows with low-E double glazing
9	9	Secondary glazing to single glazed windows
10	10	Draughtproof single-glazed windows
11	**11**	**Upgrade heating controls**
12	**11**	**Upgrade heating controls**
13	**11**	**Upgrade heating controls**
14	**11**	**Upgrade heating controls**
15	**11**	**Upgrading heating controls**
16	12	Time and temperature zone control
17	**13**	**Upgrade heating controls**
18	**13**	**Upgrade heating controls**
19	14	Solar water heating
20	**15**	**Replace boiler with new condensing boiler**
21	**15**	**Replace boiler with new condensing boiler**
22	16	Replace boiler with biomass boiler
23	**17**	**Wood pellet stove with boiler and radiators**
39	**17**	**Wood pellet stove with boiler and radiators**
24	**18**	**Fan assisted storage heaters and dual immersion cylinder**
30	**18**	**Fan assisted storage heaters and dual immersion cylinder**
25	**19**	**Fan assisted storage heaters**
31	**19**	**Fan-assisted storage heaters**
26	20	Replacement warm air unit
27	**21**	**Change heating to gas condensing boiler**
29	**21**	**Change heating to gas condensing boiler**
32	**21**	**Change heating to gas condensing boiler**
34	22	Solar photovoltaic panels, 2.5 kWp
35	23	Low energy lighting for all fixed outlets
36	24	Replace heating unit with condensing unit
37	**25**	**Install condensing boiler**
38	**25**	**Install condensing boiler**
40	**26**	**Change room heaters to condensing boiler**
41	**26**	**Change room heaters to condensing boiler**
42	27	Replace heating unit with mains gas condensing unit
28	**28**	**Condensing oil boiler with radiators**
43	**28**	**Condensing oil boiler with radiators**
44	29	Wind turbine
45	30	Flat roof insulation
46	31	Room-in-roof insulation
47	32	Floor insulation
48	33	High performance external doors
49	34	Heat recovery system for mixer showers
50	35	Flue gas heat recovery device in conjunction with boiler
56	36	Replacement glazing units
57	37	Suspended floor insulation
58	38	Solid floor insulation
59	**39**	**High heat retention storage heaters and dual immersion cylinder**
61	**39**	**High heat retention storage heaters and dual immersion cylinder**
60	**40**	**High heat retention storage heaters**
62	**40**	**High heat retention storage heaters**
63	41	Party wall insulation

in the online 1. The dataset used in subsequent predictive analysis comprises 445,661 complete observations and 82 covariates in total, described Panel A of Table A6. All other aspects not addressed in this paper adhere to the guidelines provided by the data owner without modification.[16]

[16] See https://epc.opendatacommunities.org/docs/guidance.

4.3.1. Forecasting results in the UK case

As for the Italian case, we evaluate linear and tree-based models in terms of correlation, MSE, and MAE, as detailed in Table 6. Beginning with the correlation analysis, the table illustrates the degree of linear association between predicted and actual values. For both the full sample and subsample scenarios, the models consistently exhibit robust correlation values, surpassing those observed in the Italian case. In the out-of-sample scenario, linear models display correlation coefficients

Table 6
The UK case. Correlation (top), Mean Square Error (mid), and Mean Absolute Error (bottom) between actual and predicted values estimated by LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST. In-sample and out-of-sample results for the whole sample (first and second column) and a random subsample (third and fourth column).

	Full sample		Subsample	
	In sample	Out of sample	In sample	Out of sample
	Correlation			
LASSO	0.9226	0.9222	0.9241	0.9218
RIDGE	0.9196	0.9192	0.9211	0.9185
ELASTIC NET	0.9227	0.9223	0.9230	0.9211
BART	0.9681	0.9661	0.9686	0.9555
RF	0.9469	0.9462	0.9405	0.9411
XGBOOST	0.9816	0.9745	0.9880	0.9565
	Mean Square Error			
LASSO	0.1635	0.1650	0.1630	0.1771
RIDGE	0.1706	0.1721	0.1705	0.1860
ELASTIC NET	0.1633	0.1647	0.1655	0.1789
BART	0.0690	0.0735	0.0689	0.1026
RF	0.1166	0.1185	0.1393	0.1478
XGBOOST	0.0402	0.0556	0.0268	0.1000
	Mean Absolute Error			
LASSO	0.3085	0.3092	0.3112	0.3192
RIDGE	0.3151	0.3160	0.3172	0.3269
ELASTIC NET	0.3083	0.3090	0.3139	0.3213
BART	0.1895	0.1945	0.1961	0.2302
RF	0.2497	0.2514	0.2697	0.2768
XGBOOST	0.1383	0.1613	0.1177	0.2231

hovering around 0.92, whereas all tree-based models demonstrate even stronger correlation, notably XGBOOST, and BART, with correlation coefficients ranging from 0.96 to 0.97. This pattern persists across the MSE and MAE metrics in all the investigated cases. Once more, BART and XGBOOST stand out by achieving the lowest MSE and MAE values, underscoring their robust prediction accuracy. When scrutinizing performance within model types, it becomes clear that tree-based models outperform their linear counterparts across all metrics. This reaffirms, for the UK case as well, that the inclusion of non-linear characteristics captured by the tree-based models significantly enhances their predictive capabilities in comparison to the linear models.

4.3.2. Most relevant variables for the UK case

The comparison of variable importance rankings across different machine learning models provides valuable insights into the significance of the top 15 features for predicting energy efficiency improvements, as demonstrated in Table 7. As for the Italian case, this discussion focuses on the similarities observed within linear models, the tree-based models, and the selection of features between linear and non-linear models.

Starting with the linear models, there is a consistent emphasis on factors related to heating systems, particularly the recommendation of using dual-fuel systems and the incorporation of "`High Heat Retention Storage (HHRS)`" heaters with dual immersion cylinders. It is noteworthy that several recommendations made by technicians are included in the most important selected features. For instance, upgrading heating controls and changing to gas-condensing boilers also emerge as important features across all three linear models. These similarities underscore the agreement of linear models on energy efficiency improvements as shown by the rank correlation in Table 8. Specifically, LASSO and ELASTIC NET provide approximately the same variable selection as reported by the correlation value of 0.98.

In contrast, the tree-based models exhibit a broader range of variables in their top importance rankings. As for the Italian case, all the considered models include a recommendation for wall insulation. Specifically, recommendation "`R7: 50 mm internal or external wall insulation`" entails the application of insulation materials to either the interior or exterior walls of a building with a thickness of 50 mm. This practice aims to enhance the energy efficiency of the building by reducing heat loss through its walls. Insulating walls can lead to improved thermal comfort and lower energy consumption for heating and cooling, as it helps maintain a more stable indoor temperature.

While there is some overlap with the linear models, the tree-based models exhibit a higher degree of complexity and granularity in identifying relevant features. Notably, the tree-based models place a strong emphasis on variables related to the current environmental impact, energy efficiency of heating systems, and heating costs. Additionally, these models highlight the significance of factors such as low-energy lighting and hot water energy efficiency, which were not as prominently featured in the linear models. Interestingly, BART, RANDOM FOREST, and XGBOOST models incorporate the "`number of recommendations`" feature made by technicians. As observed for the Italian case, this inclusion underscores the idea that a greater number of suggested interventions corresponds to a potentially higher level of energy-efficient improvement.

When comparing the feature selection process between linear and non-linear models, it becomes evident that tree-based models tend to encompass a broader spectrum of variables within their feature importance rankings. This observation is substantiated by the rank correlation values, which exhibit lower scores (ranging from 0.58 to 0.73) for the tree-based models in contrast to the linear models (ranging from 0.79 to 0.98). Additionally, the correlation between the linear and non-linear models is more varied in the Italian case, ranging from 0.07 to 0.49. Notably, the highest correlation for the linear models is observed with BART and RIDGE, exhibiting a correlation of 0.49.

Similarly to the Italian case, Fig. 4 provides an illustrative example of the Partial Dependence Plot (PDP) using the "`WALLS_ENERGY_EFF`" characteristic, which pertains to the energy efficiency rating of a building's walls. This rating is categorized as "`very poor`", "`poor`", "`average`", "`good`", or "`very good`", and is typically represented on energy certificates using a one to five-star scale. This assessment helps evaluate the energy performance of the building's walls, contributing to efforts aimed at enhancing energy conservation. Interestingly, the RANDOM FOREST and XGBOOST models reveal a skewed U-shaped non-linear trend, where the most substantial improvement occurs from the "`poor`" to the "`average`" category, and a slight decline in the potential energy efficiency improvement is observed from the "`very poor`" to the "`poor`" category. The BART model exhibits a similar trend, with the most significant improvement occurring from the "`average`" to the "`good`" category, and a minor decline from the "`poor`" to the "`average`" category. The unexpected trend of decreasing potential energy efficiency gains in buildings with improved "`WALLS_ENERGY_EFF`" ratings from the lowest categories might be attributed to the potential miscategorization of the two types that could be very close from a technical standpoint. If this is the case, an overlap between categories could blur the distinction between them, potentially leading to instances of incorrect classification.

5. Scenario analysis of Green energy financing policies

In this section, we conduct a scenario analysis aimed at testing two alternative green policies to increase energy efficiency in residential buildings within the considered geographical areas.[17] The first policy builds upon the technical recommendations made by the experts identified by the machine learning model, as discussed in Section 4. In this regard, we have selected XGBOOST as the reference model due to its demonstrated superior forecasting abilities in both Italian and UK cases

[17] We are grateful to an anonymous reviewer for their valuable suggestions.

Table 7
The UK case. Variable importance rankings, displaying bar chart weights for the top 15 variables across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST models.

Linear models

LASSO	RIDGE	ELASTIC NET
Main Fuel	Main Fuel	Main Fuel
R39: HHRS heaters - dual immersion cylinder	R39: HHRS heaters - dual immersion cylinder	R39: HHRS heaters - dual immersion cylinder
R21: Change heating to gas condensing boiler	R21: Change heating to gas condensing boiler	R21: Change heating to gas condensing boiler
R40: High heat retention storage heaters	Construction Period	R40: High heat retention storage heaters
R13: Upgrade heating controls	R40: High heat retention storage heaters	R13: Upgrade heating controls
R26: Change room heaters to condensing boiler	R26: Change room heaters to condensing boiler	R26: Change room heaters to condensing boiler
R7: 50 mm internal or external wall insulation	R7: 50 mm internal or external wall insulation	R7: 50 mm internal or external wall insulation
R30: Flat roof insulation	R30: Flat roof insulation	R30: Flat roof insulation
Construction Period	R20: Replacement warm air unit	Construction Period
Solar Water Heating Flag	Solar Water Heating Flag	Solar Water Heating Flag
Transaction Type	R9: Secondary glazing to single glazed windows	Transaction Type
R20: Replacement warm air unit	Transaction Type	R12: Time and temperature zone control
R12: Time and temperature zone control	R13: Upgrade heating controls	R20: Replacement warm air unit
R1: Insulate hot water cylinder with 80 mm jacket	R12: Time and temperature zone control	R1: Insulate hot water cylinder with 80 mm jacket
R6: Cavity wall insulation	R31: Room-in-roof insulation	R6: Cavity wall insulation

Tree-based models

BART	RANDOM FOREST	XGBOOST
Current Environmental Impact	Current Environmental Impact	Current Environmental Impact
R30: Flat roof insulation	Number of Recommendations	Number of Recommendations
R7: 50 mm internal or external wall insulation	Current Heating Cost	Main Heating System Energy Efficiency
Main Heating System Energy Efficiency	Current Energy Consumption	Current Heating Cost
Current Heating Cost	Local Authority Label	R23: Low energy lighting for all fixed outlets
Number of Recommendations	Main Heating System Energy Efficiency	R7: 50 mm internal or external wall insulation
R40: High heat retention storage heaters	R7: 50 mm internal or external wall insulation	Current Energy Consumption
R39: HHRS heaters - dual immersion cylinder	Current CO2 Emissions	Main Heating System Environmental Efficiency
Walls Energy Efficiency	Walls Energy Efficiency	R30: Flat roof insulation
R5: Increase loft insulation to 270 mm	Hot Water Energy Efficiency	Walls Energy Efficiency
R15: Replace boiler with new condensing boiler	Walls Environmental Efficiency	Hot Water Energy Efficiency
R6: Cavity wall insulation	Current Cost Hot Water System	Current Cost Hot Water System
Current Cost Hot Water System	Construction Period	R39: HHRS heaters - dual immersion cylinder
Main Heating System Environmental Efficiency	R23: Low energy lighting for all fixed outlets	R34: Heat recovery system for mixer showers
Current CO2 Emissions	Low Energy Lighting	Low Energy Lighting

Table 8
The UK case. Rank correlation between variable importance ranking across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST models.

	LASSO	RIDGE	ELASTIC NET	BART	RANDOM FOREST	XGBOOST
LASSO	–					
RIDGE	0.79	–				
ELASTIC NET	0.98	0.80	–			
BART	0.32	0.49	0.35	–		
RANDOM FOREST	0.18	0.23	0.19	0.67	–	
XGBOOST	0.07	0.27	0.12	0.73	0.58	–

(see Tables 2 and 6, respectively).[18] We label this strategy as the "ML Recommendations Policy". The second policy is based on a "Bottom-Up" approach, aiming to finance the improvement of residential buildings starting from the lowest EPC class in each country. In the latter case, energy efficiency improvement is attained by implementing all the provided recommendations by technicians for each EPC of a building within that EPC class, with the observation that the policymaker lacks a specific order of preferences for the suggested recommendations. A direct application of the Bottom-Up policy involves selecting the lowest EPC class and funding all recommended interventions to enhance the energy efficiency of all buildings within that class.

To evaluate the effectiveness of the presented policies, we utilize specific metrics available in the datasets, considering the latest available EPC for each building. In the Italian case, we analyze the payback periods for the implemented recommendations, while in the UK case, we focus on the direct cost.

5.1. The Italian case scenario analysis

In the Italian case, the ML Recommendations Policy is devised based on interventions identified by the XGBOOST model with the highest variable importance, as detailed in Table 3. Consequently, we formulate the policy, where the government subsidizes intervention "`R1 Opaque shell`", the sole recommendation appearing in the top 15 variables.[19] In contrast, the Bottom-Up policy allocates funds to all recommended interventions for buildings in lower EPC classes, effectively targeting the least efficient units in the residential building stock.

To ensure comparability in terms of policy preferences for social welfare, the scenario analysis has been structured in a manner whereby the policy actions target an equivalent number of buildings in both strategies. In the ML Recommendations policy, the policymaker subsidizes the implementation of opaque shells for all buildings that received that recommendation in the EPC, totaling 164,210 buildings.

[18] Additionally, Table D1 in the online Appendix indicates that XGBOOST outperforms other non-linear models in terms of execution time when applied to both the full sample and the subsample.

[19] The feature "`R1 Opaque shell`" demonstrates a fractional contribution of 34% to the total gain, noticeably surpassing the 2.81% cumulative weight of the other 5 recommendations.

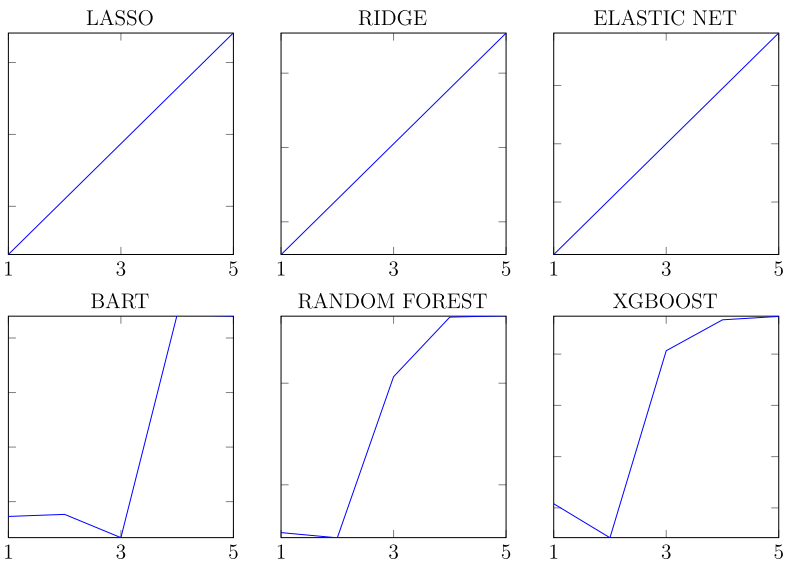


Fig. 4. Partial Dependence Plots (PDPs) in the UK case for the "WALLS_ENERGY_EFF" characteristic across LASSO, RIDGE, ELASTIC NET, BART, RANDOM FOREST, and XGBOOST. For each model, the plot illustrates the relationship between the rating of a building's wall efficiency on the x-axis and the predicted potential variation of energy performance on the y-axis. The limits of the latter are set by series min–max values.

Differently, the Bottom-Up policy focuses on buildings with initial energy classes E, F, and G, subsidizing all interventions suggested by the technician during the inspection, accounting for 167,333 buildings. The difference in size between the two groups is less than 2%.

The Italian dataset includes information on expected energy efficiency improvements and associated payback periods (measured in years) for each recommended intervention in a given residential building.[20] The significance of the payback period parallels that of financial duration, as it represents the duration influenced by intervention costs and technical efficiency. This metric depends on the accumulation of monetary savings over time, gradually offsetting the initial investment. Consequently, it emerges as a pivotal indicator encapsulating the economic viability and efficiency of an investment endeavor. In Table 9, it is observed that under the ML Recommendations policy, out of a total of 164,210 subsidized buildings (second column), the average payback period for the funded interventions is 12.04 years (third column). In contrast, the Bottom-Up policy, covering 167,333 buildings, exhibits an average payback period of 14.94 years. The ML Recommendations policy features an average payback period that is 19.41% lower than that of the Bottom-Up policy, demonstrating the ML model's ability to select the most effective recommendations.

Finally, the Average Δ EE (last column in the table) showcases the average increase in energy efficiency per intervention in kWh/m^2 per year. This is calculated as the average reduction in the EE indicator following subsidized interventions under each policy. Results show that the ML Recommendations policy achieves a superior increase in energy efficiency at 83.31 kWh/m^2 compared to the Bottom-Up policy's 59.44 kWh/m^2, representing a boost of 40.16%.[21]

5.2. The UK case scenario analysis

The UK dataset provides detailed information on the cost of each technician-recommended intervention, accompanied by an assigned cost band indicative of the associated intervention expenses. Analogously to the Italian case, the ML Recommendations Policy in the UK relies on interventions identified by the XGBOOST model, as outlined in Table 7. The recommendations include: (i) "`R23: Low energy lighting for all fixed outlets`"; (ii) "`R7: 50 mm internal or external wall insulation`"; (iii) "`R30: Flat roof insulation`"; (iv) "`R39: High heat retention storage heaters and dual immersion cylinder`"; and (v) "`R34: Heat recovery system for mixer showers`".[22]

Considering the selection of multiple recommendations by XGBOOST, we have structured the policy to assign weights to each intervention based on their respective relevance identified by the variable importance.[23] Then, we determine the number of interventions that can be funded within the allocated budget, along with the estimated cost per intervention. Finally, for each building recommended with a given intervention, we estimate the expected increase in energy efficiency and reduction in CO2 emissions. In contrast to the Italian scenario, where the two policies targeted a comparable number of buildings due to the only available payback period information, in the UK case, both policies have been allocated a monetary budget, as the implementation costs of the recommendations are provided.

Specifically, we assume that the government allocates a budget of £10 million. The Bottom-Up policy targets class-G buildings (i.e., the least efficient), with the total expenditure to implement all recommendations for these low-class properties estimated at £109 million. With a budget limit of £10 million, the government can cover approximately 9% of the expenses. Therefore, the ML Recommendations policy subsidizes 115,742 interventions, while the Bottom-Up policy targets 2868 actions.

Table 10 shows that the average cost of renovations (second column) subsidized under the ML Recommendations policy is approximately 97.52% lower than those financed under the Bottom-Up policy. Interestingly, the first recommendation proposed by the ML model is the replacement of lighting systems (`R23`), which account for 42% of the budget, proving to be a relatively low-cost action with a substantial impact on overall energy efficiency improvements. In terms of energy efficiency (third column), the performance at the aggregated level is measured as the reduction in the `ENERGY_CONSUMPTION` indicator,

[20] The specific feature is denoted as `PAYBACK_PERIOD_n`, and its description can be found in Panel C of Table A1 within the online Appendix.

[21] The result also holds when considering the median payback period.

[22] Also for the UK case, the cumulative weight of the remaining 36 recommendations is negligible since it represents the 2.44% of the total gain.

[23] Weights allocated to recommendations are subsequently normalized to fulfill the budget constraint. Specifically, the following weights are assigned: R23 = 0.42, R7 = 0.31, R30 = 0.12, R9 = 0.08, and R34 = 0.07.

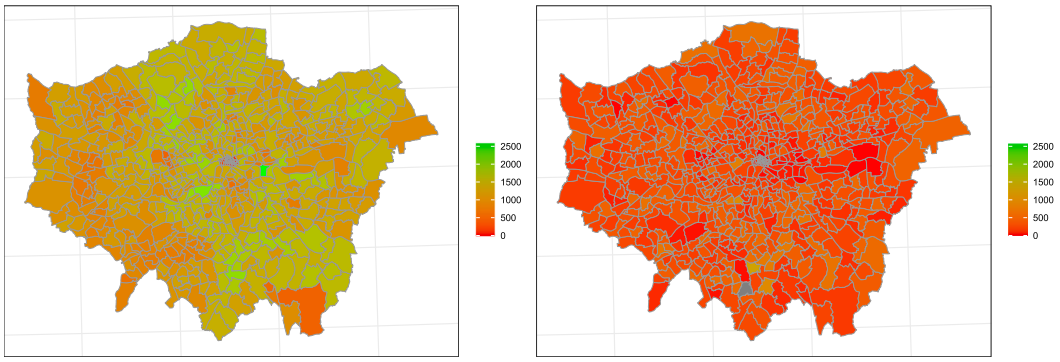


Fig. 5. The UK case. Total energy efficiency increase (in kWh/m^2 per year) resulting from the ML Recommendations Policy (left figure), and the Bottom-Up subsidizing interventions for buildings with energy class G only (right panel).

Table 9
The Italian case. Results for the ML Recommendations and Bottom-Up policies include the number of subsidized buildings (second column), the average payback period of funded interventions in years (third column), and the average energy performance increase per intervention in kWh/m^2 per year (last column). The last row indicates the percentage deviation of the ML Recommendations policy from the Bottom-Up policy.

Policy	N.	Average payback period	Average Δ EE
ML Recommendations	164,210	12.04 years	83.31 kWh/m^2
Bottom-Up	167,333	14.94 years	59.44 kWh/m^2
Percentage deviation	−1.87%	−19.41%	40.16%

Table 10
The UK case. Results for the ML Recommendations and Bottom-Up policies based on the average cost per intervention in £ (second column), the average energy performance increase at aggregate levels (third column) in kWh/m^2 per year, and aggregate reduction in CO_2 emissions in tonnes per year (fourth column). The last row indicates the percentage deviation of the ML Recommendations policy from the Bottom-Up policy.

Policy (Budget £10M.)	Average cost	Average Δ EE	Average Δ CO2
ML Recommendations	£86.40	1,043,702.06 kWh/m^2	10,512.95 tonnes
Bottom-Up	£3486.34	196,651.2 kWh/m^2	2982.71 tonnes
Percentage deviation	−97.52%	+469.26%	+252.46%

expressed in kWh/m^2 per year.[24] Also in this case, interventions subsidized by the ML Recommendation policy prove to be more effective in improving the energy efficiency of residential buildings, exhibiting a remarkable increase of 469.26% compared to the Bottom-Up policy. Based on the available `CO2_EMISSIONS` information in the UK dataset, the ML Recommendation policy achieves better results, yielding a reduction in carbon dioxide emissions nearly 2.5 times greater than the reduction achievable under the Bottom-Up policy, as can be seen in the last column of the table.

Finally, with the same allocated budget for the two potential government actions, we illustrate in Fig. 5 the improved energy performance associated with both the ML Recommendations Policy (left) and the Bottom-Up approach (right). Notably, the energy efficiency gains stemming from the ML-based policy are higher and exhibit a more balanced spatial distribution across locations, contrasting with the outcomes of a policy solely focused on improving low-efficiency buildings. While the ML Recommendations Policy targets a mix of renovations and impacts across the entire Greater London area, the Bottom-Up approach only influences zones with a high concentration of properties exhibiting poor energy efficiency levels.

[24] The average increase in energy efficiency is obtained as the sum of energy efficiency increases due to the financed interventions, considering the budget constraint, the weights associated with each code, and the unitary energy efficiency enhancement resulting from the implementation of each intervention code. The latter is calculated as the average potential energy efficiency increase of buildings for which only the specific intervention code was recommended.

6. Conclusion and policy implications

This study investigates the determinants of residential building energy efficiency, leveraging extensive datasets of EPCs from Lombardy, Italy, and London, UK. The focus is on the identification of key factors influencing efficiency and forecasting potential improvements based on building characteristics and EPC recommendations.

The findings demonstrate the superior ability of tree-based models to capture the complex, non-linear dependencies inherent in EPC data. These approaches outperform the capabilities of traditional linear models, with the XGBOOST, in particular, achieving the highest accuracy in identifying relationships between predictors and target variables across both countries. We believe that our study may provide valuable insights for policymakers and stakeholders aiming to enhance energy efficiency within the residential building sector. In this respect, we conducted a scenario analysis to assess the costs of achieving potential efficiency improvements under two alternative green policies. The first policy prioritized expert technical suggestions derived from the XGBOOST model, selected for its superior forecasting abilities. The second policy aimed to enhance energy efficiency by implementing all recommendations outlined in the EPCs without prioritizing interventions. Results consistently demonstrate that the machine learning-recommended policy delivers more cost-efficient outcomes for energy efficiency improvements in both Italy and the UK.

Overall, this framework has direct applications for legislators and interested parties seeking to develop effective, sustainable strategies for enhancing residential building energy efficiency by designing cost-effective policies tailored to achieve desired outcomes. In advancing targeted energy efficiency policies, machine learning approaches can support informed decision-making and accelerate progress toward the climate goals, as required by the EU Green Homes Directive.

CRediT authorship contribution statement

Monica Billio: Conceptualization, Funding acquisition, Investigation, Methodology, Supervision. **Roberto Casarin:** Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Michele Costola:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Veronica Veggente:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

We are grateful to Editors and the two anonymous reviewers for their comments, which helped improve the earlier draft of the manuscript. The authors acknowledge the support from the European Union - Next Generation EU - Project 'GRINS - Growing Resilient, INclusive and Sustainable' project (PE0000018); the National Recovery and Resilience Plan (NRRP) Spoke 4. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. Michele Costola also acknowledges research support from the Leibniz Institute for Financial Research SAFE. This research used the SCSCF and HPC multiprocessor cluster systems and is part of the Venice Center for Risk Analytics (VERA) project at Ca' Foscari University of Venice. We thank Rebeca Cristina Cabrera Rivera for her excellent research assistance. The authors also thank Nicolas Bianco, Filippo Pellegrino, and Chiara Vergeat, as well as the participants of the 12th European Seminar on Bayesian Econometrics (ESOBE 2022, Salzburg, Austria), the 10th Italian Congress of Econometrics and Empirical Economics (ICEEE 2023, Cagliari, Italy), the 25th Workshop on Quantitative Finance (Bologna, Italy) and the Workshop on Econometrics of the Energy Transition (EET 2024, Milan, Italy) for their valuable discussions and comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eneco.2024.107650.

References

Arcipowska, A., Anagnostopoulos, F., Mariottini, F., Kunkel, S., 2014. Energy performance certificates across the EU. Mapp. Natl. Approaches 60.
Baek, C., Park, S., 2012. Policy measures to overcome barriers to energy renovation of existing buildings. Renew. Sustain. Energy Rev. 16 (6), 3939–3947.
Barbeito, I., Zaragoza, S., Tarrío-Saavedra, J., Naya, S., 2017. Assessing thermal comfort and energy efficiency in buildings by statistical quality control for autocorrelated data. Appl. Energy 190, 1–17.
Basel Committee on Banking Supervision, 2021. Climate-Related Financial Risks–Measurement Methodologies. Technical report, Bank for International Settlements (BIS) Basel, Switzerland.
Billio, M., Costola, M., Pelizzon, L., Riedel, M., 2022. Buildings' energy efficiency and the probability of mortgage default: The Dutch case. J Real Estate Financ. 65 (3), 419–450.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Carvalho, D., Cardoso Pereira, S., Rocha, A., 2021. Future surface temperatures over Europe according to CMIP6 climate projections: An analysis with original and bias-corrected data. Clim. Change 167, 1–17.
Casarin, R., Facchinetti, A., Sorice, D., Tonellato, S., 2021. In: Abedin, M.Z., Hassan, M.K., Hajek, P., Uddin, M.M. (Eds.), Decision Trees and Random Forests. Routledge, Taylor & Francis.
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.
Chen, T., He, T., 2023. Xgboost: Extreme Gradient Boosting. R package version 1.7.3.1.
Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: Bayesian Additive Regression Trees. Ann. Appl. Stat. 4 (1), 266–298.
Chipman, H., McCulloch, R., 2016. BayesTree: Bayesian Additive Regression Trees. R package version 0.3-1.4.
Danish, M.S.S., Senjyu, T., Ibrahimi, A.M., Ahmadi, M., Howlader, A.M., 2019. A managed framework for energy-efficient building. J. Build. Eng. 21, 120–128.
Fan, C., Xiao, F., Li, Z., Wang, J., 2018. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy Build. 159, 296–308.
Ferentinos, K., Gibberd, A., Guin, B., 2023. Stranded houses? The price effect of a minimum energy efficiency standard. Energy Econ. 106555.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 1189–1232.
Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33 (1), 1–22.
García, S., Luengo, J., Herrera, F., 2015. Data Preprocessing in Data Mining. Springer.
Gómez-Omella, M., Esnaola-Gonzalez, I., Ferreiro, S., Sierra, B., 2021. k-Nearest patterns for electrical demand forecasting in residential and small commercial buildings. Energy Build. 253, 111396.
Grolinger, K., L'Heureux, A., Capretz, M.A., Seewald, L., 2016. Energy forecasting for event venues: Big data and prediction accuracy. Energy Build. 112, 222–233.
Guin, B., Korhonen, P., Moktan, S., 2022. Risk differentials between green and brown assets? Econom. Lett. 213, 110320.
Guzhov, S., Krolin, A., 2018. Use of big data technologies for the implementation of energy-saving measures and renewable energy sources in buildings. In: 2018 Renewable Energies, Power Systems & Green Inclusive Economy. REPS-GIE, IEEE, pp. 1–5.
Hijioka, Y., Masui, T., Takahashi, K., Matsuoka, Y., Harasawa, H., 2006. Development of a support tool for greenhouse gas emissions control policy to help mitigate the impact of global warming. Environ. Econ. Policy Stud. 7, 331–345.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6, 1–15.
Lashof, D.A., Ahuja, D.R., 1990. Relative contributions of greenhouse gas emissions to global warming. Nature 344 (6266), 529–531.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.
Linero, A.R., 2018. Bayesian regression trees for high-dimensional prediction and variable selection. J. Amer. Statist. Assoc. 113 (522), 626–636.
Mehmood, M.U., Chun, D., Han, H., Jeon, G., Chen, K., et al., 2019. A review of the applications of artificial intelligence and big data to buildings for energy-efficiency and a comfortable indoor living environment. Energy Build. 202, 109383.
Pratola, M.T., 2016. Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models. Bayesian Anal. 11 (3), 885–911.
Robichaud, L.B., Anantatmula, V.S., 2011. Greening project management practices for sustainable construction. J. Manage. Eng. 27 (1), 48–57.
Schuller, M., 2021. Sustainable Bonds: Bothered by buildings?. Technical report, ING.
Semple, S., Jenkins, D., 2020. Variation of energy performance certificate assessments in the European Union. Energy Policy 137, 111127.
Shine, K.P., Fuglestvedt, J.S., Hailemariam, K., Stuber, N., 2005. Alternatives to the global warming potential for comparing climate impacts of emissions of greenhouse gases. Clim. Change 68 (3), 281–302.
Skomski, E., Lee, J.-Y., Kim, W., Chandan, V., Katipamula, S., Hutchinson, B., 2020. Sequence-to-sequence neural networks for short-term electrical load forecasting in commercial office buildings. Energy Build. 226, 110350.
Sparapani, R., Spanbauer, C., McCulloch, R., 2021. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. J. Stat. Softw. 97 (1), 1–66.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Tronchin, L., Fabbri, K., 2012. Energy performance certificate of building and confidence interval in assessment: An Italian case study. Energy Policy 48, 176–184.
Yoro, K.O., Daramola, M.O., 2020. CO2 emission sources, greenhouse gases, and the global warming effect. In: Advances in Carbon Capture. Elsevier, pp. 3–28.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.
Zuo, J., Zhao, Z.-Y., 2014. Green building research–current status and future agenda: A review. Renew. Sustain. Energy Rev. 30, 271–281.