Monica Billio | Roberto Casarin | Michele Costola | Veronica Veggente

Learning from Experts: Energy Efficiency in Residential Buildings

SAFE Working Paper No. 403 | October 2023

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

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] *Imperial College Business School*

Abstract

Measuring and reducing energy consumption constitutes a crucial concern in public policies aimed at mitigating global warming. The real estate sector faces the challenge of enhancing building efficiency, where insights from experts play a pivotal role in the evaluation process. This research employs a machine learning approach to analyze expert opinions, seeking to extract the key determinants influencing potential residential building efficiency and establishing an efficient prediction framework. The study leverages open Energy Performance Certificate databases from two countries with distinct latitudes, namely the UK and Italy, to investigate whether enhancing energy efficiency necessitates different intervention approaches. The findings reveal the existence of non-linear relationships between efficiency and building characteristics, which cannot be captured by conventional linear modeling frameworks. By offering insights into the determinants of residential building efficiency, this study provides guidance to policymakers and stakeholders in formulating effective and sustainable strategies for energy efficiency improvement.

Keywords: Energy efficiency, Energy Performance Certificate, Machine learning, Tree-based models, big data

JEL: C10, C53, C50

1. Introduction

The increase in greenhouse gas emissions has a relevant impact on global warming. 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). The actual European building stock consumes 40% of total energy and emits 36% of total CO2 emissions. Overall, only 25% of Europe's building stock is deemed energy efficient, and in this respect, the European Commission (EC) set a 32.5% target

Email addresses: billio@unive.it (Monica Billio), r.casarin@unive.it (Roberto Casarin), michele.costola@unive.it (Michele Costola), v.veggente23@imperial.ac.uk (Veronica Veggente)

as a minimum goal in the 2030 climate and energy framework. Thus, 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).

The present study aims to identify the key factors representing the necessary technical interventions contributing to the potential reduction of energy consumption in residential buildings. This objective aligns with the goals set by the Energy Performance of Buildings Directive (EPBD) issued by the EC in 2002, which seeks to enhance energy efficiency in the real estate sector and ultimately reduce energy consumption.

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. 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. For instance, Billio et al. (2021) find that mortgages on energy-efficient residential buildings in the Dutch market are associated with a lower probability of default, and the relationship is stronger in the low-income group due to savings coming from reduced energy costs. Guin et al. (2022) focus on UK residential mortgages and find those energy-efficient buildings are less frequently in payment arrears than energy-inefficient ones. Ferentinos et al. (2023) show that policies aiming at increasing the energy efficiency level of the stock of buildings can reduce the price of inefficient properties.

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."[1]

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. This approach allows for a comprehensive assessment of energy performance, facilitating comparisons and guiding efforts toward optimizing energy usage and promoting sustainable building practices.

In the European Union, the EPC mechanism was introduced with the 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 (see Arcipowska et al., 2014, for an extensive discussion of the implementation of *Buildings Directive*, 2002/31/EC1 and 2010/91/EU, and EPCs in Europe). As noted in Schuller (2021), EPC procedures differ across countries and are crucial in the measurement of the energetic performance of buildings through the assignment of 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).

[1]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`.

In this paper, we investigate the determinants of energy efficiency in residential buildings and propose a flexible non-parametric approach to the analysis of expert opinions with the aim of providing an effective prediction framework. In the empirical analysis, we considered 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 study, 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: Random Forest (Breiman, 2001), Extreme Gradient Boosting (Chen and Guestrin, 2016), and Bayesian Additive Regression Tree (Chipman et al., 2010). These non-linear models are renowned for their capability to capture non-linear relationships and interactions within

the data, providing a comprehensive understanding of the complexities involved in energy efficiency prediction. 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 with correlated predictors, 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 as important in both cases. We present evidence demonstrating that tree-based models surpass linear models, enabling more precise predictions of potential efficiency improvements. This is particularly notable due to the presence of non-linear relationships between efficiency and building characteristics.

Our study aims to contribute to the field of energy efficiency in residential buildings by providing insights for policymakers to develop targeted policies aimed at reinforcing and boosting energy efficiency. One effective policy measure could be to offer financial incentives, such as tax credits or subsidies, to homeowners who invest in energy-efficient upgrades, such as insulation, high-efficiency heating systems, and energy-saving appliances. Additionally, implementing mandatory energy efficiency standards for new constructions and conducting energy audits for existing buildings could further reinforce sustainable practices and reduce carbon emissions. The prioritization of energy efficiency data disclosure holds paramount importance. By providing transparent and standardized information about the energy performance of buildings, prospective homeowners and investors can make more informed decisions. This not only promotes the adoption of energy-efficient practices but also cultivates a market environment where sustainability is valued and rewarded.

The remainder of the paper is structured as follows. Section 2 outlines the variable of interest for predicting energy efficiency improvement while Section 3 introduces the set of tree-based and linear models. Section 4 presents the empirical analyses for Italian and UK cases with a discussion on the variable selection and non-linear dependencies. Finally, Section 5 concludes the paper.

2. Modelling 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, our variable of interest is built as described in Section 2.1.

2.1. Definition of efficiency improvement

In many countries, energy efficiency is measured by means of ratings (e.g., A-G scale) and numeric indicators. Ratings are, in both cases, derived from the numeric indicator. Denote with W_j the final expected energy performance indicator after interventions, and with V_j the initial energy performance indicator for the building j and $j = 1, \ldots, n$. The indicators satisfy the constrain $0 < W_j \leq Y_j < \infty$, or equivalently $0 < Z_j \leq 1$ where $Z_j = W_j/V_j$.

Since the constraints on the minimum performance value and between the two variables introduce spurious dependence, we apply a transformation. In order to get a variable defined on $(-\infty, +\infty)$ we apply the logistic transform and define

$$Y_j = \varphi^{-1}(Z_j), \tag{1}$$

where $\varphi^{-1}(v) = \log(v) - \log(1 - v))$ is the inverse logistic. Therefore, we have that $Y_j = \log(W_j) - \log(Y_j - W_j)$. In our model, W_j is the inverse current energy efficiency, EE, and Y_j is the inverse potential energy efficiency EE_{pot}. Thus the response variable Y_j

measures the potential variation of the energy performance index and is defined as

$$Y = \log(EE_{pot}) - \log(EE - EE_{pot}), \tag{2}$$

The response variable takes large values when: either the building potential energy is low, that is, the variable EE takes large values, or when the efficiency improvement is modest, that is, the difference $EE - EE_{pot}$ takes small values. It implies that whether the improvement is fixed, a high (small) value for Y is justified by a low (high) potential energy efficiency. On the other hand, keeping the final energy efficiency level stable, Y will be higher (lower) when the potential improvement is smaller (larger). Thus, a high (low) value for Y is associated with a 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 factors influencing energy efficiency 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 efficacy in handling high-dimensional datasets with correlated predictors. Among the non-linear models, we investigate three tree-based models, namely Random Forest (Breiman, 2001), Extreme Gradient Boosting (Chen and Guestrin, 2016), and Bayesian Additive Regression Tree (Chipman et al., 2010), 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 $j = 1, \ldots, n$, that is, the energy efficiency increase for the j-th building in the sample presented in Section 2, and let $\mathbf{x}_j = (x_{1j}, \ldots, x_{mj}) \in \mathbb{R}^m$ 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 the function $f(\cdot)$ 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. The following section introduces the three primary methods: i) Bayesian additive regression trees; ii) Random Forest; iii) Extreme Gradient Boosting; and iv) Penalized regression models.

3.1. 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 nonlinear relationships and interactions among covariates and accounts for uncertainty in the estimates and prediction. The model uses a set of random trees $\mathcal{T}_j$, $j = 1, \ldots, J$ 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:

$$f(\mathbf{x}) = \sum_{j=1}^{J} g_j(\mathbf{x}). \tag{4}$$

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 $\mathcal{M}_j = \{\mu_{j1}, \ldots, \mu_{jL_j}\}$:

$$g(\mathbf{x}; \mathcal{T}_j, \mathcal{M}_j) = \sum_{l=1}^{L_j} \mu_{jl} \mathbb{I}(\mathbf{X} \in \mathcal{X}_{jl}), \tag{5}$$

where $\mathbb{I}(x \in A)$ is the indicator function which takes value 1 if x is in the set A and 0 otherwise.

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 k-th variable is below a threshold c_j, that is $X_{ik} \leq c_j$ and a right node when the k-th 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. Figure 1 provides an illustration of a simple tree with 3 leaf nodes and 4 edges, where the complete partition of the real line, $\mathbb{R} = \mathcal{X}_{11} \cup \mathcal{X}_{12} \cup \mathcal{X}_{13} \cup \mathcal{X}_{14}$, $\mathcal{X}_{11} = (-\infty, 1.36]$, $\mathcal{X}_{12} = (1.36, 13.52]$ and $\mathcal{X}_{13} = (13.52, +\infty)$, is obtained by applying two thresholds to the value of two covariates.

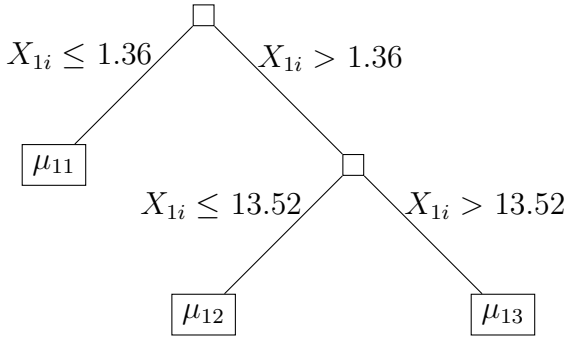

Figure 1: Example of a simple tree with three leaf nodes (square boxes) and four edges determined by thresholding two covariates X_{1i} and X_{2i} (lines).

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

$$\pi(\mathcal{T}_1, \ldots, \mathcal{T}_J, \mathcal{M}_1, \ldots, \mathcal{M}_J, \sigma^2) = \pi(\sigma^2) \prod_{j=1}^{J} \pi(\mathcal{M}_j | \mathcal{T}_j) \pi(\mathcal{T}_j). \tag{6}$$

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 $y_{\min} = \min\{y_1, \ldots, y_m\}$ and $y_{\max} = \max\{y_1, \ldots, y_m\}$. 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 Appendix A.3.

3.2. 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 (see Liaw and Wiener, 2002)[2] and perform cross-validation on the maximum number of terminal nodes in both the Italian and UK cases.

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

Among the specified models with $maxnodes = 10, 50, 100, 200, 500, 1000, 2000, 3000, 4000$ and 5000, we select the model that exhibits the best performance in terms of correlation with the true values, mean absolute error, and mean square error on the in-sample observations within the sub-sample. Our analysis reveals that the optimal model, considering all the metrics, is the one with $maxnodes = 1000$ for the Italian case and $maxnodes = 3000$ for the UK database. Consequently, we utilize the same specifications for both applications on the entire sample.

3.3. Extreme Gradient Boosting (XGBOOST)

The second model we consider is an ensemble model based on a collection of several decision trees $\mathcal{T}_j$ $j = 1, \ldots, J$, a collection of functions $g_j(\cdots; \mathcal{T}_j))$ $j = 1, \ldots, J$ with $g_j \in \mathcal{G}$ and the additive regression function in Eq. 4, 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, with L the number of leaves. The main difference 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

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + g_{J+1}(\mathbf{x}). \tag{7}$$

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 our estimate, we employ the *xgboost* R-package (Chen et al., 2023)[3] and cross-validate the value for the maximum number of iteration in both our exercises, on the full sample and the subsample of 10000 observations.

[3]The *xgboost* R-Package, developed by Jiaming Yuan, is available for download at: `https://cran.r-project.org/web/packages/xgboost/index.html`.

3.4. Lasso, Ridge and Elastic Net

These are linear parametric models that are $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$, 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:

$$||\mathbf{y} - \beta_0 - X\beta||_2^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \right], \tag{8}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)$, $X' = (\mathbf{x}_1', \ldots, \mathbf{x}_n')$, 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.

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 forecast the prospective surge 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 Figure 2, 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

[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`.

efficiency level, factoring in the proposed expert interventions. This visual representation highlights the substantial enhancements achievable through the implementation of expert-guided recommendations.

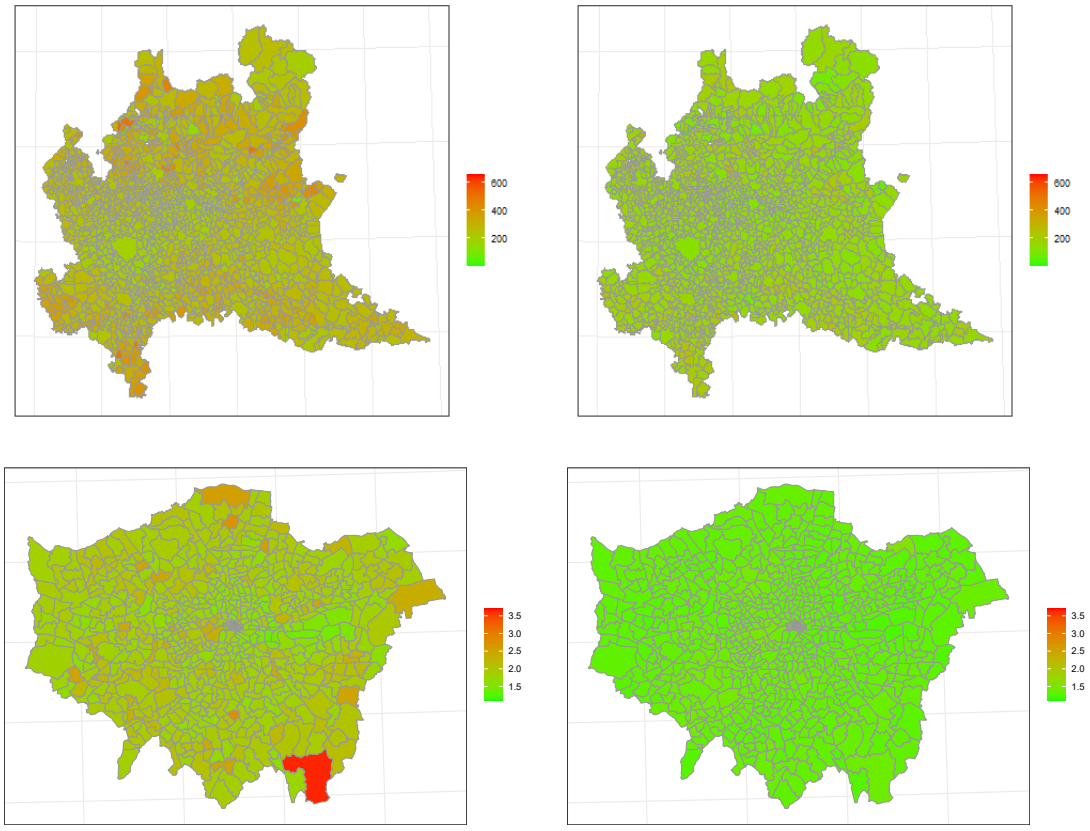

Figure 2: Geographic distribution of the initial (left) and the potential (right) energy efficiency in the Lombardy region of Italy (top) and in the Greater London area of the UK (bottom). In each plot: the colour indicates the efficiency level from high (green) to low (red), and the grey lines provide 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.

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 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 have 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 summarising the estimated potential energy class given the initial conditions of the building and the implementation of one

or more suggested interventions.

In our 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 Equation 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, 2016, to December 31, 2020.[5] The EPBD dataset for the UK encompasses Energy Performance Certificate (EPC) data issued for domestic buildings in England and Wales from January 1, 2008, to December 31, 2021.[6]

4.2. The Italian case

For the Italian case, energy efficiency is measured through the energy performance indicator EP of all the non-renewable sources used in a building (global non-renewable EP) [7], denoted as $EP_{gl,nren}$ and expressed in kWh/m^2.[8] The current and potential energy efficiency in Equation 2 are denoted as $EE = EP_{gl,nren}$ and $EE_{POT} = EP_{riq,gl,nren,ragg}$, respectively.

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 $EE - EE_{pot}$ remains non-negative. Thus, a high value for Y signifies a minimal enhancement rather than a decline in energy efficiency (refer to the upper-left

[5]The full dataset "*CENED+2 Database – Certificazione ENergetica degli EDifici*" is 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).

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

[7]For the detailed description of the methodology behind these indicators, the reader can refer to the "Amendment to the Decree of the Minister of Economic Development, June 26, 2009 - Italian National guidelines for the energy certification of buildings" `https://www.mise.gov.it/index.php/it/normativa/decreti-interministeriali/decreto-interministeriale-26-giugno-2015-adeguamento-linee-guida-nazionali-per-la-certificazione-energetica-degli-edifici`

[8]A description of the original variables is reported in Table A1 in Appendix Appendix A.

panel in Figure 3). 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 Figure 3). 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 Figure 3. Interestingly, the upper bound in the bottom-left panel of Figure 3 highlights that when $EP_{gl,nren}$ 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 an Italian Energy EPC. These recommendations encompass a range of areas including the building's shell, heating and cooling systems, other systems, and renewable sources. Our focus is solely on EPCs containing at least one recommendation from technicians, as each recommendation implies a potential increase in energy efficiency.

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

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

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

[9]The analysis focuses on residential buildings classified as E.1(1) and E.1(2) according to the DPR

information is manually reported in the CENED2+ database, introducing the possibility of typos and inconsistencies. Consequently, we remove outlier observations with initial or potential $EP_{gl,nren}$ values below the 1^{st} or above the 99^{th} percentile. Similarly, we exclude buildings with null potential energy efficiency increases in terms of $EP_{gl,nren}$ or a potential overall decrease in energy class. Hence, records with null or negative potential improvements in energy efficiency are considered erroneous or irrelevant to our study's purpose.

The final dataset of complete cases comprises 205,049 observations and 49 variables (42 of which are used as regressors in the models below) described in Table A1 in Appendix A.[10] Histograms depicting the composition of the full dataset and the subset of complete cases in terms of initial energy efficiency, construction period, and year of EPC issuance can be found in Figures A1 and A2.

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. Consequently, results on the entire dataset are compared with the one obtained in the sub-sample (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 the tree-based regression models and the 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 sub-sample. This improvement in correlation is observed across the in-sample and out-of-sample

classification. Additional information is available in the *Gazzetta Ufficiale*, `https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg`.

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

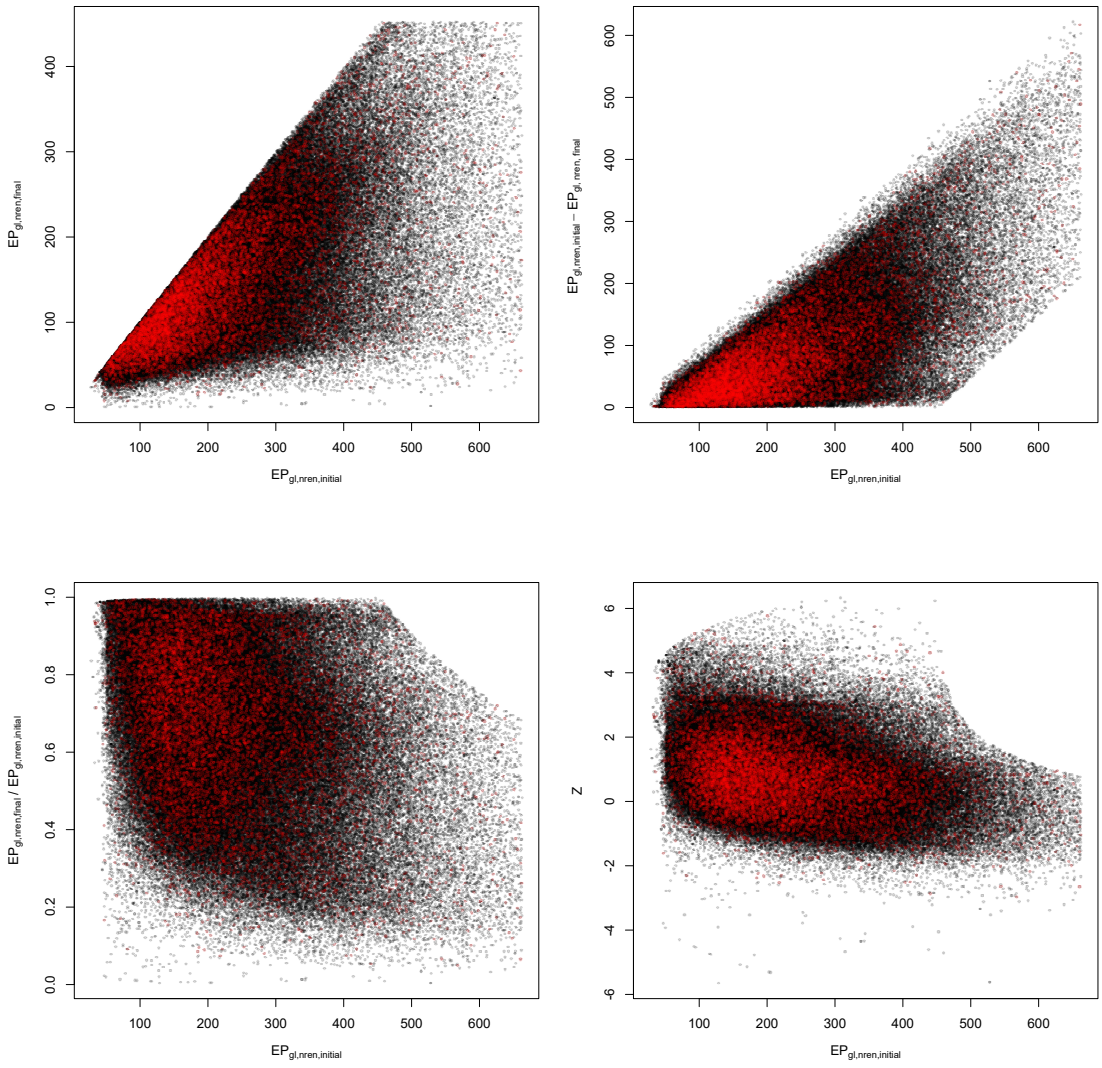

Figure 3: The case of Lombardy (Italy). The initial non-renewable energy performance index $EP_{gl,nren,initial}$ (horizontal axis) versus the final index $EP_{gl,nren,initial}$ (vertical axis, top left), the expected performance difference $EP_{gl,nren,initial} - EP_{gl,nren,final}$ (vertical axis, top right), the expected performance ratio $EP_{gl,nren,final}/\ EP_{gl,nren,initial}$ (vertical axis, bottom left), and the response variable $Z = \log(EP_{gl,nren,final}) - \log(EP_{gl,nren,initial} - EP_{gl,nren,final})$ (vertical axis, bottom right). The entire sample involves 205,049 buildings (gray dots) and a sub-sample of 10,000 buildings (red dots).

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 0.675.6 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 RANDOM FOREST, XGBOOST, and BART, 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. However, in the subsample, the performance of XGBOOST and BART 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.

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.

	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
RANDOM FOREST	0.7028	0.7052	0.6879	0.6831
XGBOOST	0.7705	0.7292	0.7976	0.6684
BART	0.7259	0.7139	0.7236	0.6766
		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
RANDOM FOREST	0.6583	0.6613	0.6716	0.7087
XGBOOST	0.5275	0.6126	0.4778	0.7347
BART	0.6125	0.6415	0.6085	0.7197
		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
RANDOM FOREST	0.6184	0.6197	0.6251	0.6413
XGBOOST	0.5513	0.5914	0.5315	0.6580
BART	0.5956	0.6071	0.5997	0.6524

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

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 for the top 15 variables across a range of models, including linear models like LASSO, RIDGE, and ELASTIC NET, as well as tree-based models like RANDOM FOREST, XGBOOST, and BART. 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 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 RANDOM FOREST, XGBOOST, and BART – show correlation values ranging from 0.60 to 0.80. This suggests that these models consistently 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.

Linear Models		
LASSO	**RIDGE**	**ELASTIC NET**
R1: Opaque Shell	R1: Opaque Shell	R1: Opaque Shell
Domestic Water System: Heat Pump	Main Heating System Main Fuel: Solar Heating	Domestic Water System: Heat Pump
Surface/Volume Ratio	Main Heating System: Solar Heating	Number of Recommendations
R6: Renewables	Main Heating System Year: 1930-1945	Surface/Volume Ratio
Number of Recommendations	Number of Recommendations	R6: Renewables
Winter Energy Efficiency (Red)	Winter Energy Efficiency (yellow)	Winter Energy Efficiency (Red)
Age Band: >2006	Winter Energy Efficiency (Red)	Age Band: >2006
Cooling System Flag	Transaction Type: New Dwelling	Cooling System Flag
Transaction Type: New Dwelling	Winter Energy Efficiency (green)	Transaction Type: New Dwelling
Age Band: 1993-2006	Main Heating System: Joule Effect Generator	Age Band: 1993-2006
Main Heating System: Joule Effect Generator	Domestic Water System Year: 1930-1945	Summer Energy Efficiency (green)
Main Heating System: Biomass Generator	Goods and People Transport System Flag	Main Heating System: Biomass Generator
R4: New Cooling System	Age Band: >2006	R4: New Cooling System
Summer Energy Efficiency (green)	Domestic Water System: Heat Pump	Domestic Water System Energy Efficiency
Domestic Water System Energy Efficiency	R6: Renewables	Domestic Water System: District Heating
Tree-based models		
RANDOM FOREST	**XGBOOST**	**BART**
R1: Opaque Shell	R1: Opaque Shell	R1: Opaque Shell
Number of Recommendations	Number of Recommendations	R3: Heating System
Thermal Efficiency	Thermal Efficiency	Thermal Efficiency
Age Band	Surface/Volume Ratio	Number of Recommendations
CO2 Emissions	Winter Energy Efficiency (Red)	R6: Renewable Sources
Difference of EE with Similar Building when New	Thermal Transmittance	Surface/Volume Ratio
Surface/Volume Ratio	Heating System Efficiency	R2: Transparent Shell
Thermal Transmittance	Cooling System Flag	Winter Energy Efficiency (yellow)
Dispersing Surface	CO2 Emissions	Domestic Water System Main Fuel: Natural Gas
Heating System Efficiency	Difference of EE with Similar Building when New	Solar Heating Flag
Effective Heated Surface	Number of Residential Units	Cooling System Flag
Natural Gas Consumption	Current Energy Efficiency Renewable	Age Band: 1993-2006
Domestic Water System Efficiency	Dispersing Surface	Current Energy Efficiency Renewable
Current Energy Efficiency Renewable	Domestic Water System Efficiency	Main Heating System: Joule Effect Generator
R2: Transparent Shell	Age Band: 1993-2006	Summer Energy Efficiency (green)

Table 3: Variable Importance Rankings for the Top 15 Variables across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART models for the Italian case.

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

Table 4: Rank correlation between variable importance ranking across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART Models for the Italian case.

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 to delve deeper into 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, Figure 4 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 $200kWh/m^2$. For the sake of completeness, 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 2014 and 2021.[11] 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.[12] 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

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

[12]See `https://www.gov.uk/guidance/standard-assessment-procedure` for a detailed description of the methodology used to compute these indices and Table A2 in Appendix A for a description of the variables included in this analysis.

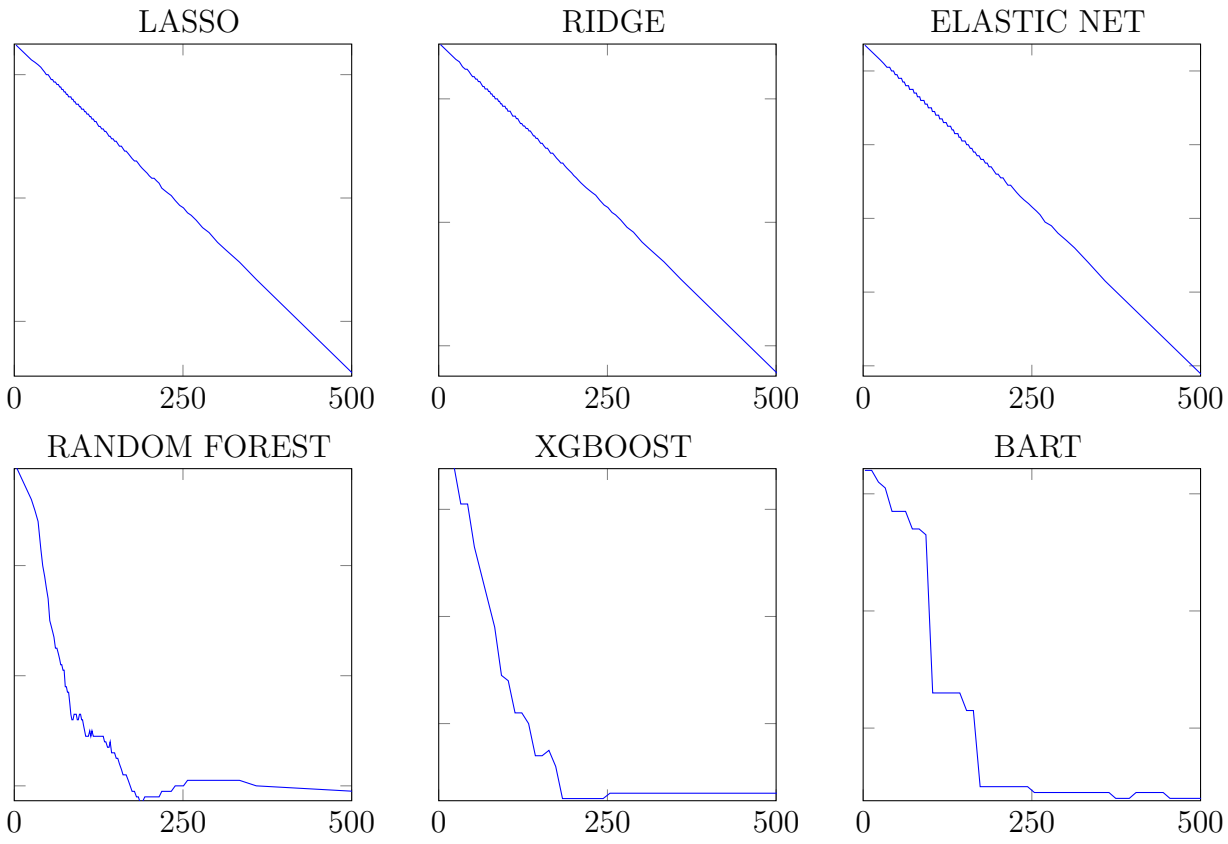

Figure 4: Partial Dependence Plots (PDPs) in the Italian case for the "THERMAL_EFFICIENCY" characteristic across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART. For each model, the plot illustrates the relationship between thermal efficiency (measured in kWh/m^2) on the x-axis and the predicted potential variation of energy performance on the y-axis. The limits of the latter are set by series min-max values.

the energy efficiency indicator is indicative of lower energy efficiency. Conversely, in the UK, elevated 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. $EE = (1/\text{CURRENT\_ENERGY\_EFFICIENCY}) \cdot 100$ and $EE_{POT} = (1/\text{POTENTIAL\_ENERGY\_EFFICIENCY}) \cdot 100$ to compute the target variable as in Equation 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 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.[13] The comprehensive list of variables considered for our study can be found in Table A2 of Appendix A. The dataset initially contains 1,041,806 rows, which is reduced to 445,661 complete records after cleaning. A subsample of 10,000 units is randomly selected from this complete set, mirroring the approach undertaken in the Italian case. All other aspects not addressed in this paper adhere to the guidelines provided by the data owner without modification.

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

[13]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.

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

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.

models display correlation coefficients 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, XGBoost and BART 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.

	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
RF	0.9469	0.9462	0.9405	0.9411
XGBOOST	0.9816	0.9745	0.9880	0.9565
BART	0.9681	0.9661	0.9686	0.9555
		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
RF	0.1166	0.1185	0.1393	0.1478
XGBOOST	0.0402	0.0556	0.0268	0.1000
BART	0.0690	0.0735	0.0689	0.1026
		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
RF	0.2497	0.2514	0.2697	0.2768
XGBOOST	0.1383	0.1613	0.1177	0.2231
BART	0.1895	0.1945	0.1961	0.2302

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

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's 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 millimeters. 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, the RANDOM FOREST, XGBOOST, and BART 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 RIDGE and BART, exhibiting a correlation of 0.49.

Similarly to the Italian case, Figure 5 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

Linear Models		
LASSO	**RIDGE**	**ELASTIC NET**
Main Fuel: Dual	Main Fuel: Dual	Main Fuel: Dual
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: >2012	R40: High heat retention storage heaters
R13: Upgrade heating controls	Main Fuel: Liquid Biomass	R13: Upgrade heating controls
R26: Change room heaters to condensing boiler	R40: High heat retention storage heaters	R26: Change room heaters to condensing boiler
R7: 50 mm internal or external wall insulation	R26: Change room heaters to condensing boiler	R7: 50 mm internal or external wall insulation
R30: Flat roof insulation	R7: 50 mm internal or external wall insulation	R30: Flat roof insulation
Construction Period: >2012	R30: Flat roof insulation	Construction Period: >2012
Main Fuel: Liquid Biomass	Main Fuel: LPG	Main Fuel: Liquid Biomass
Main Fuel: Oil	R20: Replacement warm air unit	Main Fuel: Oil
Solar Water Heating Flag	Solar Water Heating Flag	Solar Water Heating Flag
Transaction Type: Renewable Heat Incentive	R9: Secondary glazing to single glazed windows	Transaction Type: Renewable Heat Incentive
R20: Replacement warm air unit	Transaction Type: Renewable Heat Incentive	R12: Time and temperature zone control
R12: Time and temperature zone control	R13: Upgrade heating controls	R20: Replacement warm air unit
Tree-based models		
RANDOM FOREST	**XGBOOST**	**BART**
Current Environmental Impact	Current Environmental Impact	Current Environmental Impact
Number of Recommendations	Number of Recommendations	R30: Flat roof insulation
Current Heating Cost	Main Heating System Energy Efficiency	R7: 50 mm internal or external wall insulation
Current Energy Consumption	Current Heating Cost	Main Heating System Energy Efficiency
Local Authority Label	R23: Low energy lighting for all fixed outlets	Current Heating Cost
Main Heating System Energy Efficiency	R7: 50 mm internal or external wall insulation	Number of Recommendations
R7: 50 mm internal or external wall insulation	Current Energy Consumption	R40: High heat retention storage heaters
Current CO2 Emissions	Main Heating System Environmental Efficiency	R39: HHRS heaters - dual immersion cylinder
Walls Energy Efficiency	R30: Flat roof insulation	Walls Energy Efficiency
Hot Water Energy Efficiency	Walls Energy Efficiency	R5: Increase loft insulation to 270 mm
Walls Environmental Efficiency	Hot Water Energy Efficiency	R15: Replace boiler with new condensing boiler
Current Cost Hot Water System	Current Cost Hot Water System	R6: Cavity wall insulation
Construction Period	R39: HHRS heaters - dual immersion cylinder	Current Cost Hot Water System
R23: Low energy lighting for all fixed outlets	R34: Heat recovery system for mixer showers	Main Heating System Environmental Efficiency
Low Energy Lighting	Low Energy Lighting	Current CO2 Emissions

Table 7: Variable Importance Rankings for the Top 15 Variables across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART Models for the UK case.

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

Table 8: Rank correlation between variable importance ranking across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART Models for the UK case.

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.

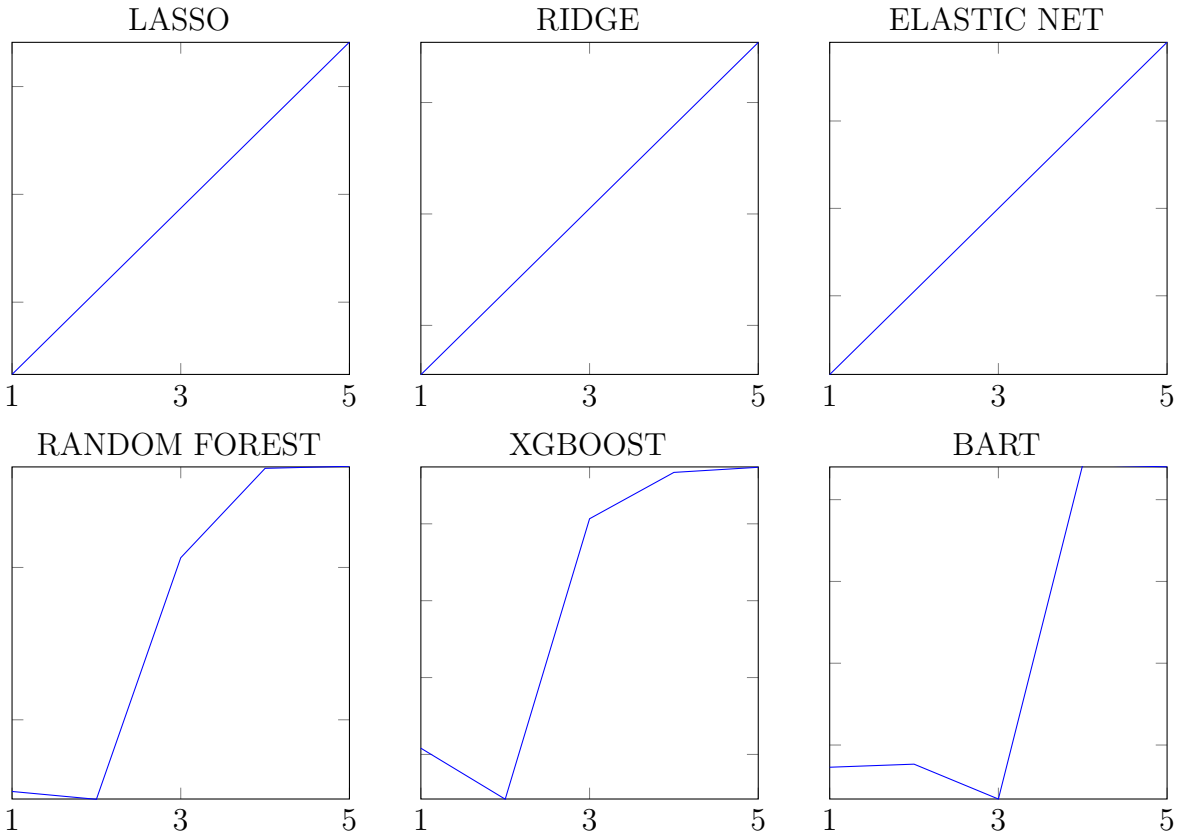

Figure 5: Partial Dependence Plots (PDPs) in the UK case for the "WALLS_ENERGY_EFF" characteristic across LASSO, RIDGE, ELASTIC NET, RANDOM FOREST, XGBOOST, and BART. For each model, the plot illustrates the relationship between the energy efficiency the rating of a building's walls efficiency on the 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.

5. Conclusion and Policy implications

In this study, we explored the determinants of energy efficiency in residential buildings by utilizing large and comprehensive datasets of Energy Performance Certificates (EPCs) from Lombardy, Italy, and London, UK. The primary objective is to gain insights into the

factors influencing energy efficiency and to develop accurate forecasts of potential efficiency improvements using the buildings' characteristics and the recommendations made by the experts in the ECP.

Our study has several policy implications for policymakers and stakeholders interested in improving energy efficiency in residential buildings. First, our findings demonstrated that tree-based models have the ability to more accurately capture the complexities and non-linear dependencies present in the EPC data. This finding suggests that the non-linear relationships observed between predictors and target variables in both countries can be effectively modeled using tree-based approaches. By leveraging these models, we can improve our understanding of the determinants of energy efficiency and make more accurate predictions regarding potential efficiency improvements in residential buildings. Therefore, policymakers can use these models to develop targeted policies to improve energy efficiency in residential buildings by identifying the key factors that contribute to (energy) inefficiency.

Second, this modeling framework can be extended to include the cost of the transition resulting from specific government green policies. This can help policymakers develop cost-effective policies and achieve the desired outcomes in terms of improving energy efficiency in residential buildings. For example, our model can be used to estimate the cost of transitioning from traditional heating systems to more energy-efficient systems, such as heat pumps or solar panels. Billio et al. (2022) highlight a historical lack of attention by policymakers towards insulation policies. This can be particularly helpful in estimating the cost-effectiveness of implementing these policies to improve energy efficiency in residential buildings. It can also be used to assess the effectiveness of combining renewable energy sources, like solar rooftop photovoltaic systems and heat pumps, with other sustainable solutions, such as green roofs, in improving energy efficiency in buildings, particularly in specific climate and architectural conditions.

Finally, policymakers could use the proposed models to simulate the impact of different climate scenarios on energy efficiency in residential buildings. This can help them identify

the most effective policies for mitigating the impact of climate change on energy consumption in residential buildings. For example, one could simulate the impact of alternative climate scenarios based on the energy consumption in residential buildings and identify the most effective policies for reducing energy consumption in these scenarios.

Overall, our study provides valuable insights into the determinants of energy efficiency in residential buildings and highlights the potential of tree-based models for forecasting potential efficiency improvements. The presented model can support policymakers and stakeholders in developing effective and sustainable strategies for improving energy efficiency in residential buildings, ultimately reducing carbon emissions and energy costs.

Acknowledgments

The authors acknowledge the support from the European Union - Next Generation EU - Project 'GRINS - Growing Resilient, INclusive and Sustainable'; the National Recovery and Resilience Plan (NRRP) Spoke 4. 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) and the 10th Italian Congress of Econometrics and Empirical Economics (ICEEE 2023) for their valuable discussions and comments.

References

Arcipowska, A., Anagnostopoulos, F., Mariottini, F., and Kunkel, S. (2014). Energy performance certificates across the EU. *A mapping of national approaches*, 60.

Baek, C. and Park, S. (2012). Policy measures to overcome barriers to energy renovation of existing buildings. *Renewable and Sustainable Energy Reviews*, 16(6):3939–3947.

Barbeito, I., Zaragoza, S., Tarrío-Saavedra, J., and Naya, S. (2017). Assessing thermal comfort and energy efficiency in buildings by statistical quality control for autocorrelated data. *Applied Energy*, 190:1–17.

Basel Committee on Banking Supervision (2021). Climate-related financial risks—measurement methodologies. Technical report.

Billio, M., Costola, M., Hristova, I., Latino, C., and Pelizzon, L. (2021). Inside the ESG ratings:(Dis) agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5):1426–1445.

Billio, M., Gianni, C., and Hristova, I. (2022). Technical Report on relevant public support actions in relation to EEM. Technical report, Energy Efficient Mortgages Initiative.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Carvalho, D., Cardoso Pereira, S., and Rocha, A. (2021). Future surface temperatures over Europe according to CMIP6 climate projections: an analysis with original and bias-corrected data. *Climatic Change*, 167:1–17.

Casarin, R., Facchinetti, A., Sorice, D., and Tonellato, S. (2021). Decision trees and random forests. chapter 10. Routledge, Taylor & Francis.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2023). *xgboost: Extreme Gradient Boosting*. R package version 1.7.3.1.

Chipman, H. and McCulloch, R. (2016). *BayesTree: Bayesian Additive Regression Trees*. R package version 0.3-1.4.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Danish, M. S. S., Senjyu, T., Ibrahimi, A. M., Ahmadi, M., and Howlader, A. M. (2019). A managed framework for energy-efficient building. *Journal of Building Engineering*, 21:120–128.

Fan, C., Xiao, F., Li, Z., and Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159:296–308.

Ferentinos, K., Gibberd, A., and Guin, B. (2023). Stranded houses? The price effect of a minimum energy efficiency standard. *Energy Economics*, page 106555.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

García, S., Luengo, J., and Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

Gómez-Omella, M., Esnaola-Gonzalez, I., Ferreiro, S., and Sierra, B. (2021). k-Nearest patterns for electrical demand forecasting in residential and small commercial buildings. *Energy and Buildings*, 253:111396.

Grolinger, K., L'Heureux, A., Capretz, M. A., and Seewald, L. (2016). Energy forecasting for event venues: Big data and prediction accuracy. *Energy and Buildings*, 112:222–233.

Guin, B., Korhonen, P., and Moktan, S. (2022). Risk differentials between green and brown assets? *Economics Letters*, 213:110320.

Guzhov, S. and 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)*, pages 1–5. IEEE.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6:1–15.

Liaw, A. and 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. *Journal of the American Statistical Association*, 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 and Buildings*, 202:109383.

Pratola, M. T. (2016). Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3):885–911.

Schuller, M. (2021). Sustainable Bonds: Bothered by buildings? techreport, ING.

Semple, S. and Jenkins, D. (2020). Variation of energy performance certificate assessments in the European Union. *Energy Policy*, 137:111127.

Skomski, E., Lee, J.-Y., Kim, W., Chandan, V., Katipamula, S., and Hutchinson, B. (2020). Sequence-to-sequence neural networks for short-term electrical load forecasting in commercial office buildings. *Energy and Buildings*, 226:110350.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1):1–66.

Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the general BART model. *Statistics in Medicine*, 38(25):5048–5069.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tronchin, L. and Fabbri, K. (2012). Energy Performance Certificate of building and confidence interval in assessment: An Italian case study. *Energy Policy*, 48:176–184.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Appendix A. Supplementary Insights: Italian and UK Dataset Details

In this section, we provide additional information on the datasets used in our analysis for Italy (Section Appendix A.1) and the UK (Section Appendix A.2). For each country, we present histograms that compare the original databases with the subsets of complete cases considered in our study to demonstrate the absence of selection bias in our identification strategy. Specifically, we examine variables such as construction period[14], year of EPC issuance, initial energy rating, and climatic area (for Italy only). Additionally, we provide a comprehensive list of variables included as covariates in the estimated models for both countries, as presented in Tables A1 and A2. Additionally, we focus on the various recommendations assessors can suggest to increase energy efficiency, as shown in Table 1. Lastly, in Section Appendix A.3, we present supplementary graphical results for both the Italian and UK analyses.

Appendix A.1. The Italian dataset

The section provides a comprehensive overview of the Italian dataset, featuring key figures and tables for detailed analysis. Figure A1 showcases the composition of the original Italian dataset, with insights into the construction period, inspection year, energy rating, and climatic area. Building on that, Figure A2 delves into the composition of the 205,049 complete cases within the Italian dataset, providing a closer examination of the same variables. Moreover, Table A1 presents essential information, including original and English labels, descriptions, and variable types, for the 49 variables analyzed in the Italian case. For further clarity, Table 1 illustrates the original and English labels for recommendation identifiers in the Italian dataset.

[14]Following discussions with the EPC Data Team, observations falling under the "2007 onwards" class in the UK database were reclassified as "2007-2011".

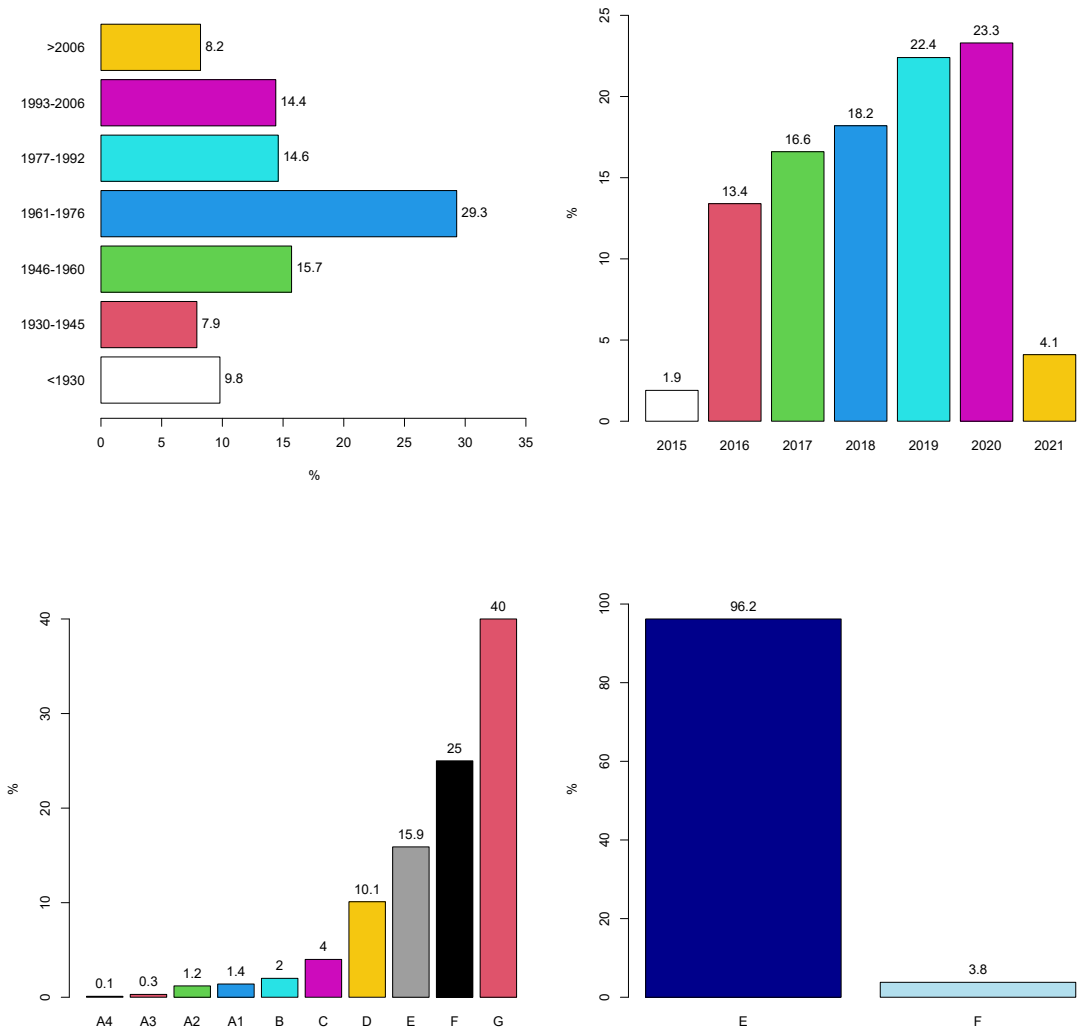

Figure A1: Composition in terms of the construction period (top left), inspection year (top right), energy rating (bottom left), and climatic area (bottom right) of the original Italian dataset.

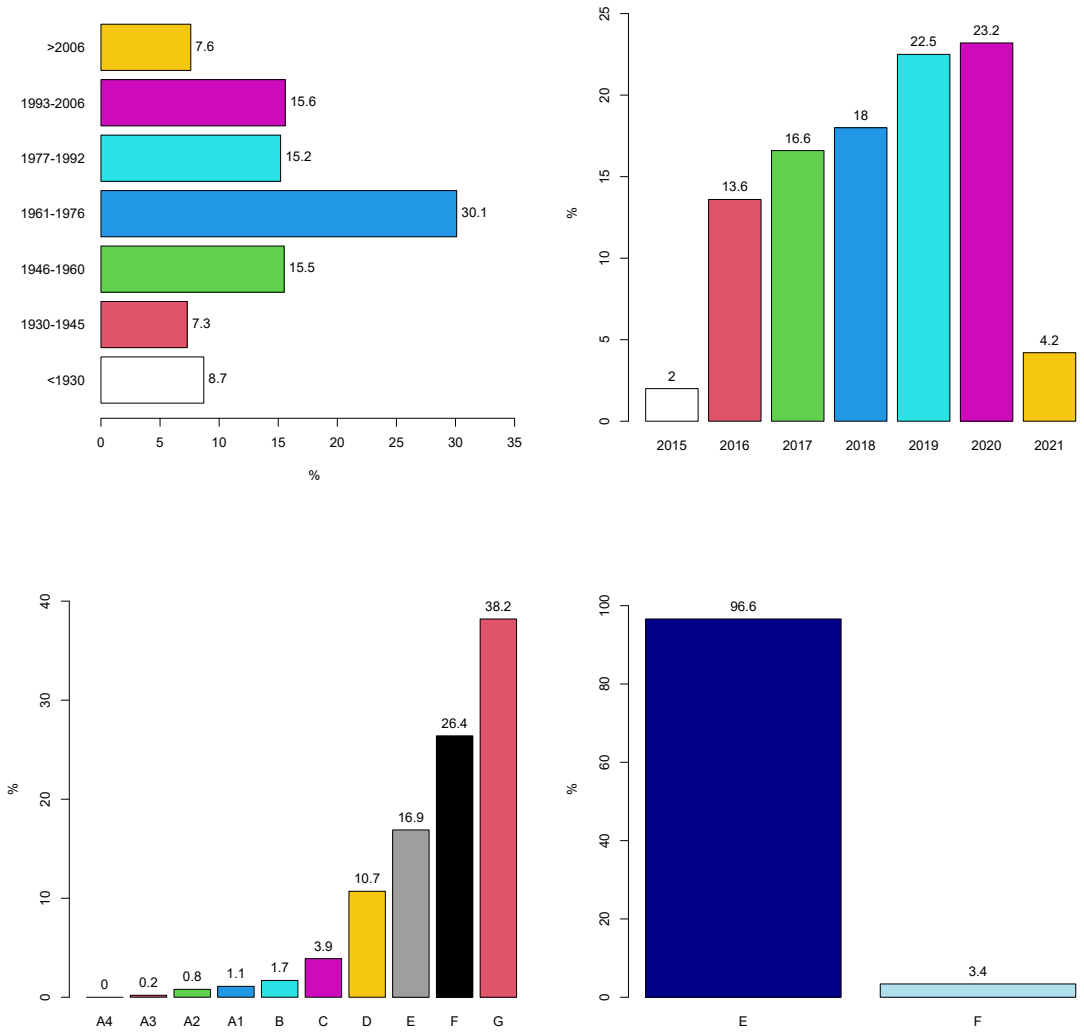

Figure A2: Composition in terms of the construction period (top left), inspection year (top right), energy rating (bottom left), and climatic area (bottom right) of the 205,049 complete cases in the Italian dataset.

Appendix A.2. The UK dataset

The appendix presents essential visualizations and tables related to the UK dataset, shedding light on crucial aspects of the analysis. Figure A3 provides a comprehensive composition overview, highlighting the construction period, inspection year, and energy rating of the original UK dataset. Building on that, Figure A4 focuses on a sub-sample, comprising 445,661 complete cases, and offers valuable insights into the distribution of the

	Variable name	Description	Variable type	Original variable
1	MUNICIPALITY	Municipality name	Categorical	COMUNE
2	CLIMATIC_AREA	Climatic area (D or E)	Categorical	ZONA_CLIMATICA
3	AGE_BAND	Construction period	Categorical	ANNO_COSTRUZIONE
4	POTENTIAL_ENERGY_RATING	Potential energy rating (A4-G)	Categorical	RIQ_CLASSE_RAGG
5	CURRENT_ENERGY_RATING	Current energy rating (A4-G)	Categorical	CLASSE_ENERGETICA
6	MAIN_HEATING_SYSTEM	Main heating system type	Categorical	CI_TIPO_IMPIANTO_1
7	MAIN_HEATING_SYSTEM_YEAR	Installation year of main heating system	Categorical	CI_ANNO_INSTALLAZIONE_1
8	MAIN_HEATING_SYSTEM_MAIN_FUEL	Main fuel of main heating system	Categorical	CI_VETTORE_ENERGETICO_1
9	DOMESTIC_WATER_SYSTEM	Domestic water system type	Categorical	PA_TIPO_IMPIANTO_1
10	DOMESTIC_WATER_SYSTEM_YEAR	Installation year of domestic water system	Categorical	PA_ANNO_INSTALLAZIONE_1
11	DOMESTIC_WATER_SYSTEM_MAIN_FUEL	Main fuel of domestic heating system	Categorical	PA_VETTORE_ENERGETICO_1
12	year	Year of inspection	Categorical	DATA_INS
13	TRANSACTION_TYPE	Reason for EPC issuance	Categorical	NUOVA_COSTRUZIONE, PASSAGGIO_PROPRIETA, LOCAZIONE, RISTRUTTURAZIONE_IMPORTANTE, RIQUALIFICAZIONE_ENERGETICA
14	date	Inspection date	Date	DATA_INS
15	COOLING_SYSTEM_FLAG	Dummy for cooling system	Dummy	CLIMATIZZAZIONE_ESTIVA
16	MECHANICAL_VENTILATION_FLAG	Dummy for mechanical ventilation	Dummy	VENTILAZIONE_MECCANICA
17	GOODS_PEOPLE_TRANSPORT_FLAG	Dummy for goods and people transport	Dummy	TRASPORTO_PERSONE_COSE
18	r_1	Recommendation: opaque shell	Dummy	DS_TIPO_INTERVENTO_1
19	r_2	Recommendation: transparent shell	Dummy	DS_TIPO_INTERVENTO_2
20	r_3	Recommendation: new heating system	Dummy	DS_TIPO_INTERVENTO_3
21	r_4	Recommendation: new cooling system	Dummy	DS_TIPO_INTERVENTO_4
22	r_5	Recommendation: change other systems	Dummy	DS_TIPO_INTERVENTO_5
23	r_6	Recommendation: renewables	Dummy	DS_TIPO_INTERVENTO_6
24	LPG_FLAG	Dummy for LPG	Dummy	CONSUMI_GPL
25	DISTRICT_HEATING_FLAG	Dummy for district heating	Dummy	CONSUMI_TELERISCALDAMENTO
26	DIESEL_FLAG	Dummy for Diesel	Dummy	CONSUMI_GASOLIO
27	BIOMASS_FLAG	Dummy for biomass	Dummy	CONSUMI_BIOMASSE_GASSOSE, CONSUMI_BIOMASSE_LIQUIDE, CONSUMI_BIOMASSE_SOLIDE
28	SOLAR_PHOTOVOLTAIC_FLAG	Dummy for solar photovoltaic	Dummy	CONSUMI_SOLARE_FOTOVOLTAICO
29	SOLAR_HEATING_FLAG	Dummy for solar heating	Dummy	CONSUMI_SOLARE_TERMICO
30	N_RESIDENTIAL_UNIT	Number of residential units in the building	Numeric	NUMERO_UNITA_IMMOBILIARI
31	EFFECTIVE_HEATED_SURFACE	Effective heated surface (square meters)	Numeric	SUPERFICIE_UTILE_RISCALDATA
32	EE_WINTER	Energy efficiency in winter	Numeric	PI
33	EE_SUMMER	Energy efficiency in summer	Numeric	PE
34	CURRENT_ENERGY_EFFICIENCY	Current energy efficiency (the lowest the better)	Numeric	EP_GL_NREN
35	CURRENT_ENERGY_EFFICIENCY_REN	Current energy efficiency for renewables (the lowest the better)	Numeric	EP_GL_REN
36	POTENTIAL_ENERGY_EFFICIENCY	Potential energy efficiency (the lowest the better)	Numeric	RIQ_EP_GL_NREN_RAGG
37	CO2_EMISSIONS	CO2 emissions	Numeric	EMISSIONI_CO2
38	ELECTRICITY_CONSUMPTION	Electricity consumption	Numeric	CONSUMI_ENERGIA_ELETTRICA
39	NATURAL_GAS_CONSUMPTION	Natural gas consumption	Numeric	CONSUMI_GAS_NATURALE
40	DISPERSING_SURFACE	Dispersing surface	Numeric	SUPERFICIE_DISPERDENTE
41	SV_RATIO	Surface to volume ratio	Numeric	RAPPORTO_SV
42	THERMAL_EFFICIENCY	Thermal efficiency	Numeric	EP_H_ND
43	SUMMER_EQ_SOLAR_AREA	Summer equivalent solar area	Numeric	A_SOL_EST_A_SUP_UTILE
44	THERMAL_TRANSMITTANCE	Thermal transmittance	Numeric	Y_IE
45	HEATING_SYSTEM_EFFICIENCY	Heating system energy efficiency	Numeric	CI_EFFICIENZA_MEDIA
46	DOMESTIC_WATER_SYSTEM_EFF	Domestic water system energy efficiency	Numeric	PA_EFFICIENZA_MEDIA
47	N_R	Number of recommendations	Numeric	DS_TIPO_INTERVENTO_1, DS_TIPO_INTERVENTO_2, DS_TIPO_INTERVENTO_3, DS_TIPO_INTERVENTO_3, DS_TIPO_INTERVENTO_4, DS_TIPO_INTERVENTO_5, DS_TIPO_INTERVENTO_6
48	DIFF_ENERGY_EFF_NEW_BUILD	Difference in energy efficiency with respect to similar building when new	Numeric	Y
49	Z	Energy efficiency increase (logistic transformation)	Numeric	EP_GL_NREN, RIQ_EP_GL_NREN_RAGG

Table A1: Original (fourth column) and English (first column) labels, description (second column), and variable type (third column) of the 49 variables included in the analysis for the Italian case.

construction period, inspection year, and energy rating within this subset.

Furthermore, Table A3 contains information about the 82 variables analyzed in the UK case, including their labels, descriptions, and variable types. This table serves as a valuable reference for understanding the dataset's characteristics. Moreover, Table A4 features both the original and re-coded recommendation identifiers, along with detailed descriptions of the interventions within the UK dataset.

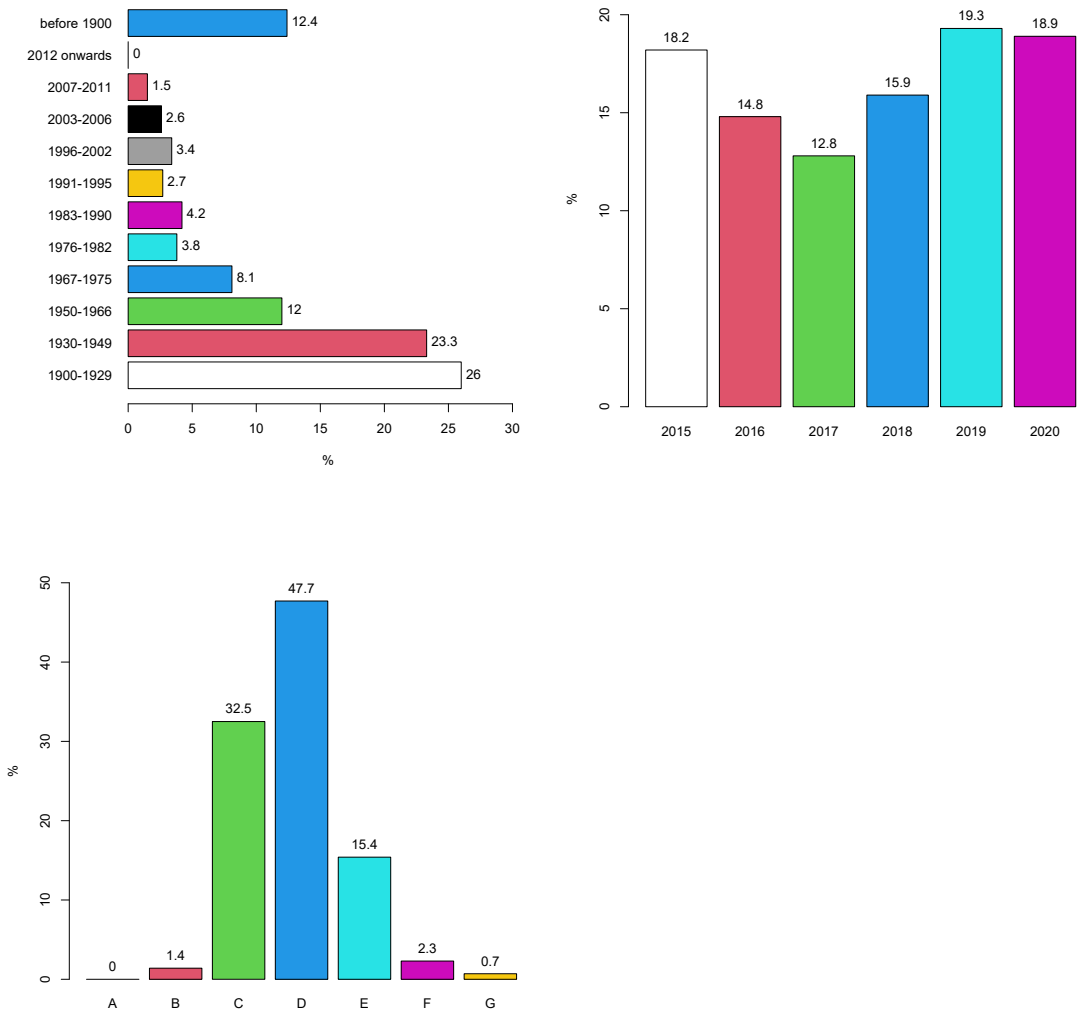

Figure A3: Composition in terms of the construction period (top left), inspection year (top right), energy rating (bottom) of the original UK dataset.

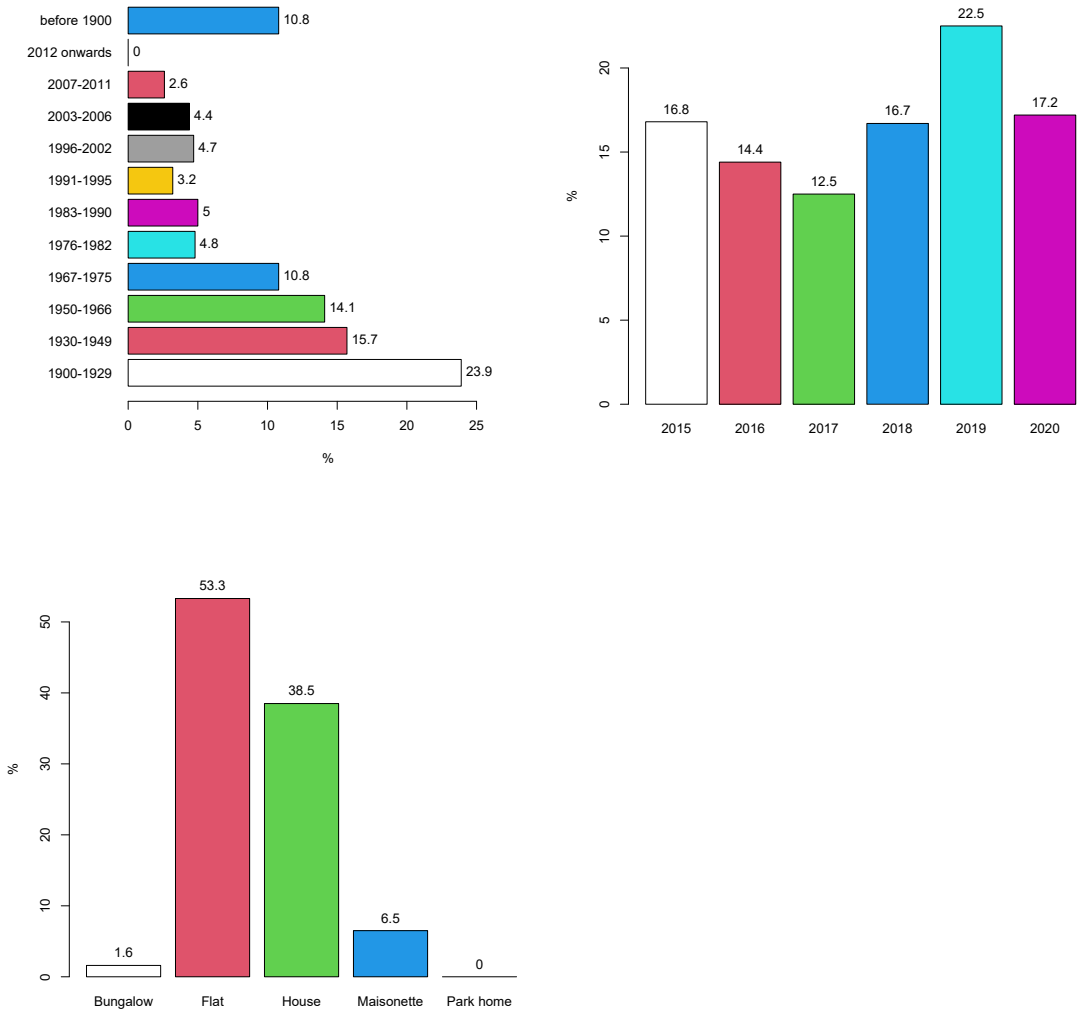

Figure A4: Sub-sample composition in terms of the construction period (top left), inspection year (top right), energy rating (bottom) of the 445,661 complete cases in the UK dataset.

	Variable name	Description	Variable type
1	BUILT_FORM	Building type	Categorical
2	CONSTRUCTION_PERIOD	Construction period	Categorical
3	FLOOR_LEVEL	Floor level	Categorical
4	GLAZED_TYPE	Glazed type	Categorical
5	LOCAL_AUTHORITY_LABEL	Local authority	Categorical
6	MAIN_FUEL	Main fuel type	Categorical
7	MECHANICAL_VENTILATION	Mechanical ventilation system type	Categorical
8	PROPERTY_TYPE	Property type	Categorical
9	TENURE	Tenure type	Categorical
10	TRANSACTION_TYPE	Transaction type	Categorical
11	YEAR	Year of inspection	Categorical
12	MAINS_GAS_FLAG	Mains gas flag	Dummy
13	r_1	Recommendation: Insulate hot water cylinder with 80 mm jacket	Dummy
14	r_2	Recommendation: Increase hot water cylinder insulation	Dummy
15	r_3	Recommendation: Add additional 80 mm jacket to hot water cylinder	Dummy
16	r_4	Recommendation: Hot water cylinder thermostat	Dummy
17	r_5	Recommendation: Increase loft insulation to 270 mm	Dummy
18	r_6	Recommendation: Cavity wall insulation	Dummy
19	r_7	Recommendation: 50 mm internal or external wall insulation	Dummy
20	r_8	Recommendation: Replace single glazed windows with low-E double glazing	Dummy
21	r_9	Recommendation: Secondary glazing to single glazed windows	Dummy
22	r_10	Recommendation: Draughtproof single-glazed windows	Dummy
23	r_11	Recommendation: Upgrading heating controls	Dummy
24	r_12	Recommendation: Time and temperature zone control	Dummy
25	r_13	Recommendation: Upgrade heating controls	Dummy
26	r_14	Recommendation: Solar water heating	Dummy
27	r_15	Recommendation: Replace boiler with new condensing boiler	Dummy
28	r_16	Recommendation: Replace boiler with biomass boiler	Dummy
29	r_17	Recommendation: Wood pellet stove with boiler and radiators	Dummy
30	r_18	Recommendation: Fan assisted storage heaters and dual immersion cylinder	Dummy
31	r_19	Recommendation: Fan assisted storage heaters	Dummy
32	r_20	Recommendation: Replacement warm air unit	Dummy
33	r_21	Recommendation: Change heating to gas condensing boiler	Dummy
34	r_22	Recommendation: Solar photovoltaic panels, 2.5 kWp	Dummy
35	r_23	Recommendation: Low energy lighting for all fixed outlets	Dummy
36	r_24	Recommendation: Replace heating unit with condensing unit	Dummy
37	r_25	Recommendation: Install condensing boiler	Dummy
38	r_26	Recommendation: Change room heaters to condensing boiler	Dummy
39	r_27	Recommendation: Replace heating unit with mains gas condensing unit	Dummy
40	r_28	Recommendation: Condensing oil boiler with radiators	Dummy
41	r_29	Recommendation: Wind turbine	Dummy
42	r_30	Recommendation: Flat roof insulation	Dummy
43	r_31	Recommendation: Room-in-roof insulation	Dummy
44	r_32	Recommendation: Floor insulation	Dummy
45	r_33	Recommendation: High performance external doors	Dummy
46	r_34	Recommendation: Heat recovery system for mixer showers	Dummy
47	r_35	Recommendation: Flue gas heat recovery device in conjunction with boiler	Dummy
48	r_36	Recommendation: Replacement glazing units	Dummy
49	r_37	Recommendation: Suspended floor insulation	Dummy
50	r_38	Recommendation: Solid floor insulation	Dummy
51	r_39	Recommendation: High heat retention storage heaters and dual immersion cylinder	Dummy
52	r_40	Recommendation: High heat retention storage heaters	Dummy
53	r_41	Recommendation: Party wall insulation	Dummy
54	SOLAR_WATER_HEATING_FLAG	Solar water heating flag	Dummy
55	CO2_EMISSIONS_CURRENT	Current CO2 emissions	Numeric
56	ENERGY_CONSUMPTION_CURRENT	Current energy consumption	Numeric
57	ENVIRONMENT_IMPACT_CURRENT	Current environmental impact	Numeric
58	EXTENSION_COUNT	Number of extensions	Numeric
59	GLAZED_AREA	Glazed area	Numeric
60	HEATING_COST_CURRENT	Current heating cost	Numeric
61	HOT_WATER_COST_CURRENT	Current hot water cost	Numeric
62	HOT_WATER_ENERGY_EFF	Current hot water energy efficiency	Numeric
63	HOT_WATER_ENV_EFF	Current hot water environmental efficiency	Numeric
64	LIGHTING_COST_CURRENT	Current lighting cost	Numeric
65	LIGHTING_ENERGY_EFF	Current lighting energy efficiency	Numeric
66	LIGHTING_ENV_EFF	Current lighting environmental efficiency	Numeric
67	LOW_ENERGY_LIGHTING	Proportion of low energy lighting	Numeric
68	MAINHEAT_ENERGY_EFF	Current main heating system energy efficiency	Numeric
69	MAINHEAT_ENV_EFF	Current main heating system environmental efficiency	Numeric
70	MAINHEATC_ENERGY_EFF	Current main heating system control energy efficiency	Numeric
71	MAINHEATC_ENV_EFF	Current main heating system control environmental efficiency	Numeric
72	MULTI_GLAZE_PROPORTION	Proportion of multi glaze	Numeric
73	N_RECOMMENDATION	Number of recommendations	Numeric
74	NUMBER_HABITABLE_ROOMS	Number of habitable rooms	Numeric
75	NUMBER_HEATED_ROOMS	Number of heated rooms	Numeric
76	NUMBER_OPEN_FIREPLACES	Number of open fireplaces	Numeric
77	TOTAL_FLOOR_AREA	Total floor area	Numeric
78	WALLS_ENERGY_EFF	Current walls energy efficiency	Numeric
79	WALLS_ENV_EFF	Current walls environmental efficiency	Numeric
80	WIND_TURBINE_COUNT	Number of wind turbines	Numeric
81	WINDOWS_ENERGY_EFF	Current windows energy efficiency	Numeric
82	WINDOWS_ENV_EFF	Current windows environmental efficiency	Numeric

Table A2: Labels (first column), description (second column), and variable type (third column) of the 82 variables included in the analysis for the UK case.

Appendix A.3. Selecting k: Cross-validation on BART

In this section, we present the process for selecting the number of trees, denoted as k, in the BART model. We estimate eight variations of the BART(k) model using both sampling schemes, with k values of 1, 5, 10, 20, 50, 100, 150, and 200. The posterior distribution is approximated via a 2500-iteration MCMC procedure. Figure A5 displays the MCMC trace plot for the error variance σ^2. The results show that BART(150) (in black) converges faster to the posterior distribution compared to BART(1) (in grey), both computed on the full sample. A vertical line marks the end of the burn-in samples, which are excluded when computing posterior quantities of interest.

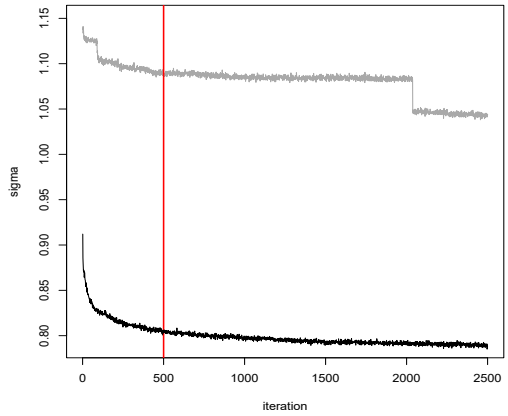

Figure A5: MCMC trace plot for the error variance σ^2 for BART(150) (black) and BART(1) (grey) fitted computed on the full sample. The vertical line indicates the end of the burn-in samples.

The predictive performance exhibits a noticeable improvement for smaller numbers of trees and reaches stability when the number of trees is 100 or greater (refer to Figure A6). These findings align with previous literature on BART models (see, for instance, Tan and Roy, 2019; Chipman et al., 2010). It is worth noting that the predictions generated by BART with 100, 150, and 200 trees are almost equivalent. In Figure A7, we present the distribution of MSE for each posterior draw obtained from the BART model with 200 trees (indicated by the grey bars). Additionally, the MSE of the final estimate is depicted as a blue line, accompanied by its 95% Highest Posterior Density (HPD) region, indicated by blue dots.

Comparing the results, we observe that the MSE of the LASSO model (red dashed line) differs significantly from that of the BART models in both the in-sample (left panel) and the out-of-sample (right panel) analyses. This difference in MSE underscores the distinct predictive performance between the LASSO model and the BART models, making the latter a more promising and accurate choice for the considered analysis.

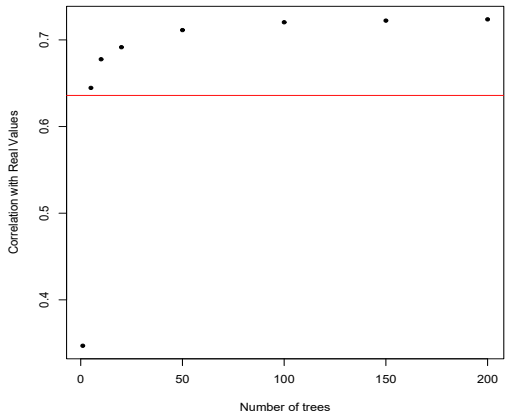

Figure A6: Correlation between Z and BART estimates for each fitting (k = 1, 5, 10, 20, 50, 100, 150, 200)(black dots) and correlation between Z and Linear Regression estimated values (red line). In sample fitting on the whole sample.

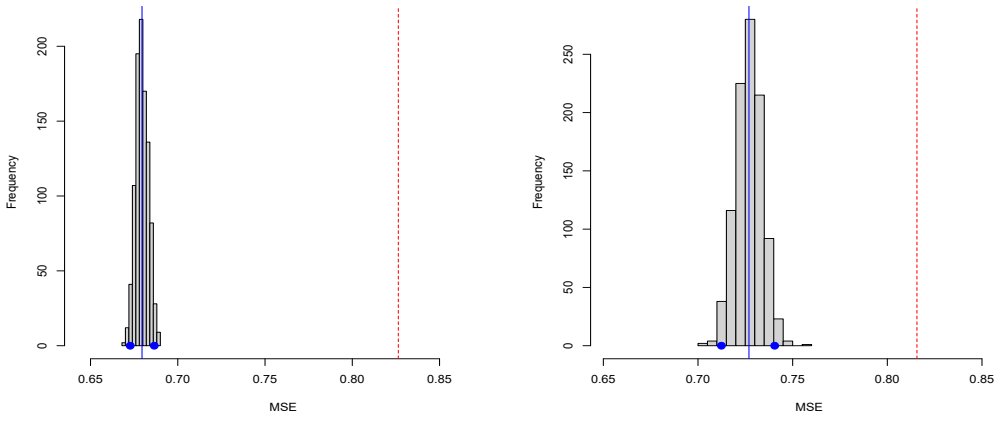

Figure A7: MSE posterior distribution for the BART(200) fitted on the sub-sample (grey bars), the average MSE (blue line), 95% HPD region (blue dots) and the MSE for the LASSO model (red dashed line). Left panel: in-sample. Right panel: out-of-sample.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 402	Julian Detemple, Michael Kosfeld	Fairness and Inequality in Institution Formation
No. 401	Kevin Bauer, Oliver Hinz, Michael Kosfeld, Lena Liebich	Decoding GPT's Hidden 'Rationality' of Cooperation
No. 400	Andreas Hackethal, Philip Schnorpfeil, Michael Weber	Households' Response to the Wealth Effects of Inflation
No. 399	Raimond Maurer, Sehrish Usman	Dynamics of Life Course Family Transitions in Germany: Exploring Patterns, Process and Relationships
No. 398	Pantelis Karapanagiotis, Marius Liebald	Entity Matching with Similarity Encoding: A Supervised Learning Recommendation Framework for Linking (Big) Data
No. 397	Matteo Bagnara, Milad Goodarzi	Clustering-Based Sector Investing
No. 396	Nils Grevenbrock, Alexander Ludwig, Nawid Siassi	Homeownership Rates, Housing Policies, and Co-Residence Decisions
No. 395	Ruggero Jappelli, Loriana Pelizzon, Marti Subrahmanyam	Quantitative Easing, the Repo Market, and the Term Structure of Interest Rates
No. 394	Kevin Bauer, Oliver Hinz, Moritz von Zahn	Please Take Over: XAI, Delegation of Authority, and Domain Knowledge
No. 393	Michael Kosfeld, Zahra Sharafi	The Preference Survey Module: Evidence on Social Preferences from Tehran
No. 392	Christian Mücke	Bank Dividend Restrictions and Banks' Institutional Investors
No. 391	Carmelo Latino, Loriana Pelizzon, Max Riedel	How to Green the European Auto ABS Market? A Literature Survey
No. 390	Kamelia Kosekova, Angela Maddaloni, Melina Papoutsi, Fabiano Schivardi	Firm-Bank Relationships: A Cross-Country Comparison