Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques

Lenaers, Ian; Boudt, Kris; De Moor, Lieven

Published in:
International Journal of Housing Markets and Analysis

DOI:
10.1108/IJHMA-11-2022-0172

Publication date:
2023

Citation for published version (APA):
Lenaers, I., Boudt, K., & De Moor, L. (2023). Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. *International Journal of Housing Markets and Analysis*. https://doi.org/10.1108/IJHMA-11-2022-0172

Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques

Journal:	*International Journal of Housing Markets and Analysis*
Manuscript ID	IJHMA-11-2022-0172.R2
Manuscript Type:	Research Paper
Keywords:	Rent prediction, Residential real estate, Machine Learning, Interpretable Machine Learning, Black-box, SHapley Additive exPlanations

Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques

Abstract

Purpose – The purpose is twofold. First, this study aims to establish that black-box tree-based Machine Learning (ML) models have better predictive performance than a standard Linear Regression (LR) hedonic model for rent prediction. Second, it shows the added value of analysing tree-based ML models with Interpretable Machine Learning (IML) techniques.

Design/methodology/approach – Data on Belgian residential rental properties were collected. Tree-based ML models, Random Forest Regression (RFR) and eXtreme Gradient Boosting Regression (XGBR), were applied to derive rent prediction models to compare predictive performance with a LR model. Interpretations of the tree-based models regarding important factors in predicting rent were made using SHapley Additive exPlanations (SHAP) Feature Importance (FI) plots and SHAP Summary plots.

Findings – Results indicate that tree-based models perform better than a LR model for Belgian residential rent prediction. The SHAP FI plots agree that asking price, cadastral income, surface livable, number of bedrooms, number of bathrooms, and variables measuring the proximity to Points Of Interest (POIs) are dominant predictors. The direction of relationships between rent and its factors is determined with SHAP Summary plots. In addition to linear relationships, it emerges that non-linear relationships exist.

Originality/value – Rent prediction using ML is relatively less studied than house price prediction. In addition, studying prediction models using IML techniques is relatively new in real estate economics. Moreover, this study is the first to derive insights of driving determinants of predicted rents from SHAP FI and SHAP Summary plots.

Keywords – Rent prediction, Residential real estate, Machine learning, Black-box, Interpretable machine learning, SHapley Additive exPlanations

Paper type – Research paper

Introduction

The ability to accurately predict and at the same time understand the drivers of real estate rent is important to various stakeholders, including but not limited to landlords, tenants, investors, and real estate agents. In practice, hedonic pricing models, mostly Linear Regression (LR), estimated with Ordinary Least Squares (OLS), are applied (Malpezzi, 2002). The literature review by Valier (2020) shows support for LR models due to their ease of interpretation. However, since housing is so heterogeneous and price influences are so complex, it is not easy to achieve accuracy with LR models due to underlying assumptions, such as linearity, among others (Valier, 2020). After all, drivers in the real estate market are characterized by complex non-linear patterns and interaction effects (Krämer *et al.*, 2021).

Due to increasing computing power and digitization, the incorporation of Machine Learning (ML) models is taking place in many domains, including real estate valuation (Breuer & Steininger, 2020; Piegeler & Bauer, 2021). With these models, it is possible to make more accurate predictions than with traditional OLS regressions (Valier, 2020). However, there are also obstacles in the adoption of using ML models when they are so-called black-box. and not straightforward to interpret unlike LR (Molnar, 2022; Surkov *et al.*, 2022). To overcome this obstacle, the use of Interpretable Machine Learning (IML) techniques, also called Explainable Artificial Intelligence, is a solution. With these techniques, it is possible to gain an understanding of the ML model, also known as explanations at the global level. It is also possible to gain an understanding at the level of an individual prediction, known as explanations at the local level (Molnar, 2022). Thus, at the model level, it is possible to determine what factors have a strong influence in determining predictions and what impacts exist. At the individual prediction level, on the other hand, it is possible to indicate why that prediction was given.

The prediction of rent prices for real estate properties is a complex task due to the presence of non-linear relationships between the rent and its determinants. To address this challenge, this paper proposes using tree-based ML models and IML techniques to improve the accuracy and transparency of rent price predictions. The paper aims to demonstrate the greater predictive power of tree-based ML models compared to LR for predicting Belgian rents. Additionally, the criticism on the black-box nature of tree-based models is addressed by applying IML techniques to generate insights into the relationships between rent and its determinants. The objective of this study is therefore twofold. The first objective of this study is to predict rental prices of Belgian residential properties using two black-box tree-based ensemble ML models, eXtreme Gradient Boosting Regression (XGBR) and Random Forest Regression (RFR). The second objective of this study is to interpret these black-box models with two global IML techniques, SHapley Additive exPlanations (SHAP) Feature Importance (FI) plots and SHAP Summary plots, to derive insights into the determinants of Belgian residential real estate rents.

The evaluation metrics on the test set show that the tree-based ML models achieve better predictive performance compared to the LR model. However, it is also pointed out that there are indications that the tree-based ML models exhibit a higher degree of overfitting than the LR model. Furthermore, it becomes clear from the SHAP FI plots that asking price and cadastral income – a fictitious income, corresponding to the average annual net income that a property would yield to its owner in the base year 1975, which is the basis for the calculation of property tax in Belgium (*Cadastral Income | Belgium.Be*, 2022) – have the strongest influence in predicting rental prices of Belgian residential real estate. Furthermore, however to a lesser extent, some location determinants and structural characteristics have influence in the prediction model. Using the SHAP Summary plots, a simple picture of the direction and magnitude of the impact between predictive determinants on predicted rent is provided. In addition, it is shown that the relationships between some determinants and predicted rents are non-linear.

First, the significance of this paper lies in demonstrating that tree-based ML models outperform LR for predicting Belgian residential rent. Second, it is shown that through IML techniques some insight regarding the determinants of rent is gained into the tree-based ML models. In addition, several implications are derived from these conclusions. Overall, this paper adds to the body of knowledge as it is one of the first studies to explain rent prediction models using IML techniques. In this case SHAP-based techniques are applied.

This paper is organized as follows. Section 2 reviews literature on the determinants of real estate prices and papers on rental price prediction models with ML. Section 3 briefly describes the data and the ML prediction models and IML techniques used. Section 4 reports and interprets the results. Finally, in Section 5, the conclusions, implications, and limitations of this study are addressed.

Literature Review

The literature review consists of three parts. First, a section highlights the increased use and advantages of ML in real estate prediction applications. Second, a section is devoted to the criticism of the black-box nature of ML models and the already existing research involving IML techniques in real estate applications. Finally, it is touched upon what type of variables have been used for predicting real estate prices in prior research.

ML models

Most widely used hedonic price model is the LR model, which uses OLS estimation (Malpezzi, 2002). However, this method is mainly useful for explanatory purposes and could be less suitable for predictive purposes in real estate applications compared to ML models (Valier, 2020). In the last decade, ML made its way into predicting real estate prices. ML is concerned with creating algorithms that allow computers to learn from data and improve their performance on a specific task (Hastie *et al.*, 2009). This involves using data to identify patterns and make predictions or decisions.

The literature that follows in this section shows that in almost all studies considered, ML leads to better prediction results than classic LR. This is also concluded by Valier (2020). It is also notable that the amount of research on predicting house prices and rents with ML is increasing. Explanations for this increase in research for prediction with ML in real estate valuation are increasing computing power, digitization and in general the incorporation of ML into new application domains (Breuer & Steininger, 2020; Piegeler & Bauer, 2021). After all, ML models may offer advantages, in many economic applications, over classic LR. The reason is that ML models are more flexible, make fewer assumptions, and are able to capture complex relationships (Antipov & Pokryshevskaya, 2012; Krämer *et al.*, 2021; Lorenz *et al.*, 2022). Thus, the prediction power of ML models can be superior to that of classic LR models.

ML models for rent and price prediction in the considered literature below include predominantly Gradient Boosting Regression (GBR), K-Nearest Neighbors Regression (KNNR), Neural Network Regression (NNR), RFR, and Support Vector Regression (SVR). GBR also includes the subvarieties such as Light Gradient Boosting Regression (LGBR) and XGBR. Recent literature regarding the ML models that perform best in predicting prices/rents reveals that there is no genuine consensus in terms of favored type of model. McCluskey *et al.* (2013), Oshodi *et al.* (2019) and Shen *et al.* (2022) conclude that NNR offers the best performance. In contrast, Antipov & Pokryshevskaya (2012), Ma *et al.* (2018) and Zhou *et al.* (2019) find that RFR provides the best performance. On the one hand, Zhang *et al.* (2021) conclude that SVR performs best. Finally, on the other hand, Krämer *et al.* (2021), Lorenz *et al.* (2022) and Xu & Li (2021) decide that GBR, especially XGBR, performed best. Notably, most of the studies compare the performance of only a few ML models. However, the publications of Krämer *et al.* (2021) and Lorenz *et al.* (2022) point in the direction that the preferences should tend towards tree-based ensemble models, such as RFR and XGBR, due to their advantages.

The rationale for the preference of tree-based models is because it can map non-linear relationships (Krämer *et al.*, 2021). Additionally, these models are considered robust to outliers by subgrouping the data so that the effect of outliers is isolated in one part of the model and does not affect the entire model. (Antipov & Pokryshevskaya, 2012; Hastie *et al.*, 2009). Moreover, these models do not require any detailed model specification (Hastie *et al.*, 2009). For further argumentation of the use and advantages of tree-based ensemble models, see Antipov & Pokryshevskaya (2012).

IML techniques

Although there is no denying of the better predictive potential with ML models compared to LR models, ML models do face criticism. This is due to the black-box nature of these models, owing to their lack of interpretability and transparency (McCluskey *et al.*, 2013; Piegeler & Bauer, 2021). This criticism is not limited only to applications in predicting real estate prices but is a general criticism of ML models (Surkov *et al.*, 2022). As a result, simpler, more interpretable models that are less accurate in predicting are often preferred. This comes at the expense of more complex, less to non-interpretable, but accurate models (Molnar, 2022).

A way to crack open the black-box nature of complex ML models is through the use of IML techniques (Molnar, 2022; Surkov *et al.*, 2022). These techniques make it possible to gain an understanding of the underlying prediction model, also called explanations at the global level. There are also interpretations at the level of an individual prediction, called explanations at the local level. Thus, at the model level, it is possible to determine what factors have a strong influence in determining predictions and what influences and relations exist. At the local level, on the other hand, it is possible to indicate for each prediction why that prediction was given. These techniques are relatively new in economic applications and certainly in real estate prediction applications. So far there are two works that use global IML techniques, namely studies by Krämer *et al.* (2021) and Lorenz *et al.* (2022), which both take a similar approach by using permutation FI, partial dependence plots and accumulated local effect plots for the interpretation of their prediction models. They both concluded that the XAI techniques improved the transparency of their black-box ML models.

Variable/Feature types

In addition to the debate over which type of ML model is best used in terms of accuracy and interpretability, there is also the debate over which predictive rent determinants should be included for model construction. A plethora of predictive rent determinants have been examined and used in research. However, Zulkifley *et al.* (2020) state that the predictive variables in real estate applications can be divided into four groups, namely structural, location, socioeconomic and macroeconomic variables. Factors that can be associated with these groupings can as such be considered to have some predictive value in predicting rents or property prices. Structural variables include characteristics of the property, such as surface livable, number of bedrooms and bathrooms. Location factors are data regarding geographical location such as proximity to parks, supermarkets, and universities, among others. Socioeconomic factors are considered factors about the neighborhood of the property, such as average income and education level of neighborhoods. However, the boundary between location and socioeconomic factors is thin, so they are often taken together, as also observed in the reviewed studies above. Lastly, the macroeconomic variables that attempt to include the impact of economic cycle, such as the mortgage rate. However, macroeconomic features are less used as observed in the work of Zulkifley *et al.* (2020). According to the research of Chiwuzie *et al.* (2021), one possible explanation for this is the inconsistent and conflicting findings regarding the direction of the relation between the studied macroeconomic variables and property prices. Another argument for this observation is that it can be more difficult to include macroeconomic variables in prediction models since they are often less directly related to real estate prices than location factors and property characteristics are (Warisse, 2017). Among those four groupings, structural and location factors are the two principal categories (Krämer *et al.*, 2021). These two types of variable groupings were used in essentially all the studies that are referenced in the 'ML models' section above.

Data and Methodology

The rental price prediction model is applicable to many countries. What may differ across countries are the features available. Clearly, all houses have structural characteristics such as number of bedrooms and location factors such as proximity to Points Of Interest (POIs). A more specific feature for Belgium is the cadastral income which equals a fictitious income, corresponding to the average annual net income that a property would yield to its owner in the base year 1975, which is the basis for the calculation of property tax in Belgium (*Cadastral Income | Belgium.Be*, 2022). While our analysis is focused on Belgium, the general methodology is applicable to rental price prediction of all countries. First in this section, data and descriptive statistics are discussed. Second, the methodology for constructing the ML prediction models is discussed and tree-based ML models with IML techniques are explained.

Data

The analyses in this study were conducted on part of a private dataset provided by a player in the Belgian real estate landscape. The original dataset contains 22567 residential property observations collected between January 1, 2019, and May 4, 2022. Data were cleaned of unrealistic and erroneous data points in consultation with a real estate expert. This cleaning resulted in 18935 usable observations. The dataset contains the monthly gross rents of Belgian residential properties in EUR, based on estimates from real estate agents. In addition, the dataset includes features regarding structural characteristics of the property. Furthermore, the data were supplemented with location features about proximity of POIs, which were determined by available longitude and latitude using Open Street Map. The location features are divided into two groups: the minimum distance to POIs and the number of POIs within a 5.5 km range. Since coordinates were used to calculate the distances, the minimum distances to POIs are quantified in degrees as the unit of measurement.

Considering that the dataset was collected between the 1st of January 2019 and 4th of May 2022, inflation has an impact on the trend in rents and asking prices of properties over time. In other words, there is an inflation drift. ML models are usually not well suited to account for drifts over time while training models (L'Heureux *et al.*, 2017). Therefore, the objective is to eliminate this inflation drift and put rents and prices on equal footing. This is done by correcting rents and asking prices with the Belgian health index, so that all rents and prices are indexed to January 2019. Cadastral income is not exposed to inflation drift, because it is provided for base year 1975. Neither are the other variables in Table I exposed to inflation drift. The choice of the Belgian health index is based upon it being the basis for the annual indexation of residential property rents in Belgium (*Rent Calculator | Statbel*, 2017).

For variable selection, variables that can be placed in the categories of structural factors, location factors and socioeconomic factors were considered. Furthermore, research by Krämer *et al*. (2021), Lorenz *et al*. (2022), and Oshodi *et al*. (2019) all have characteristics that are roughly similar to those in the used data. In addition, asking price and cadastral income were also included in the feature mix due to their high correlation with rent. Less consideration was given to variable selection because the tree-based models do not require detailed model specifications (Antipov & Pokryshevskaya, 2012). This is because the process of building tree-based models inherently selects which variables are important for making decisions at each split (Hastie *et al.*, 2009). As such, features that are not important will not be used in the trees.

The features, their data type, number of observations, mean, standard deviation, skewness and correlation with target variable are shown in Table I. Average monthly rent indexed to January 2019 is EUR 887 with a standard deviation of EUR 431. Moreover, the average asking price, indexed to January 2019, of a property is EUR 295107 and the average cadastral income is EUR 942. Note that 69.4% of the properties in the data set are apartments, while the remaining 30.6% are houses. Furthermore, a logical intuitive ranking exists between the average number of bedrooms, toilets, and bathrooms, of 2.349, 1.246 and 1.149 respectively. A ranking can also be found for the proximity of POIs by the reader. Interesting is that 12% of the properties are rented out at least partly furnished.

All property types were subject to the modeling process. However, the feature mix includes a binary variable 'Property type' that distinguishes between houses and apartments. For modeling purposes, a log transformation is applied to both inflation indexed rents and inflation indexed asking prices. This is standard practice in the literature, as seen in studies of Lorenz *et al.* (2022) and Steurer *et al.* (2021). This is done to mitigate the impact of outliers, especially on the upside of the distributions. In addition, to reduce dimensionality, a Principal Component Analysis (PCA) was performed on the location features regarding minimum distance to POIs and number of nearby POIs, respectively. These PCAs resulted in five Principal Components (PCs) for the former location factors and three PCs for the latter location factors. In both cases PCs explained at least 75% of the variance of the original data. The PCs are as such variables that measure the proximity to POIs.

For training the models, 80% of the data were used. 20% of the data were used for evaluation, which is in accordance with the classically named train-test split. As observed in Table I, there are missing values for a portion of the variables. To deal with this, the missing values were imputed with the MissForest algorithm that was proposed by Stekhoven & Bühlmann (2012). According to Waljee *et al.* (2013), the algorithm gives good results compared to other imputers and is robust against noisy data. Finally, to prepare the data for modeling, categorical variables were encoded to continuous variables with CatBoostEncoder. This is done to prevent the creation of many dummy variables – and as such an increase in dimensionality – if one-hot encoding were to be used. For clarity, for the purposes of descriptive statistics, they were treated as categorical. For modeling and interpretation with IML techniques, the categorical variables were treated as continuous. Then the variables were scaled with StandardScaler.

Table I: Summary statistics for the target variable and features

Feature name	Data type	Nr. of observations	Mean	Standard deviation	Skewness	Correlation with target
Monthly rent (indexed to 2019) in EUR	Continuous	18935	887.384	430.668	4.270	/
Asking price (indexed to 2019) in EUR	Continuous	18831	295106.971	155702.016	3.409	0.812
Cadastral income (in base year 1975) in EUR	Continuous	6270	941.815	745.955	5.579	0.614
Energy Prestation Certificate (EPC; A = 0, B = 1, C = 2, D = 3, E = 4, F = 5)	Ordinal	8302	/	/	/	/
Furnished (0 = no, 1 = yes)	Binary	5896	0.120	0.325	2.341	-0.019
Garage (0 = no, 1 = yes)	Binary	16067	0.264	0.441	1.066	0.079
Garden (0 = no, 1 = yes)	Binary	14052	0.320	0.467	0.771	0.098
Land area in m^2	Continuous	11054	478.771	1088.858	7.319	0.087
Min. distance to doctor (in degrees)	Continuous	18935	0.023	0.029	1.843	-0.126
Min. distance to hospital (in degrees)	Continuous	18935	0.160	0.113	0.965	-0.058
Min. distance to mall (in degrees)	Continuous	18935	0.196	0.155	0.795	-0.172
Min. distance to park (in degrees)	Continuous	18935	0.133	0.080	0.748	-0.040
Min. distance to restaurant (in degrees)	Continuous	18935	0.008	0.010	2.238	-0.050
Min. distance to school (in degrees)	Continuous	18935	0.023	0.025	1.916	-0.101
Min. distance to station (in degrees)	Continuous	18935	0.024	0.027	2.315	-0.057
Min. distance to supermarket (in degrees)	Continuous	18935	0.017	0.020	1.718	-0.057
Min. distance to university (in degrees)	Continuous	18935	0.256	0.153	-0.126	-0.0127
Nr. of bathrooms	Integer	16924	1.149	0.451	3.905	0.506
Nr. of bedrooms	Integer	15817	2.349	1.012	0.689	0.434
Nr. of doctors in 5.5km range	Integer	18935	18.669	26.950	1.461	0.226
Nr. of hospitals in 5.5km range	Integer	18935	0.388	1.062	3.600	0.094
Nr. of malls in 5.5km range	Integer	18935	0.293	0.552	1.736	0.191
Nr. of parks in 5.5km range	Integer	18935	0.231	0.522	2.213	0.020
Nr. of restaurants in 5.5km range	Integer	18935	136.421	208.078	2.291	0.253
Nr. of schools in 5.5km range	Integer	18935	11.213	13.706	1.296	0.244
Nr. of stations in 5.5km range	Integer	18935	5.234	9.899	4.720	0.184
Nr. of supermarkets in 5.5km range	Integer	18935	41.407	68.017	2.201	0.253
Nr. of universities in 5.5km range	Integer	18935	0.368	1.358	7.967	0.089
Number of toilets	Integer	11296	1.246	0.745	1.074	0.395
Property type (0 = house, 1 = apartment)	Binary	18935	0.694	0.461	-0.844	-0.202
Surface livable in m^2	Continuous	17263	122.999	70.087	3.884	0.616
Surface of garden in m^2	Continuous	6827	28.857	163.532	15.004	0.093
Terrace (0 = no, 1 = yes)	Binary	14698	0.775	0.418	-1.317	-0.012
Zip code	Nominal	18935	/	/	/	/

Source: own elaboration of the provided data

ML models

Tree-based (ensemble) models seem to have a slight edge in the literature considered, therefore RFR and XGBR were selected as ML models for this study. Both are compared on predictive performance to a LR model, estimated with OLS, which serves as a baseline model.

Tree-based regression techniques are based on individual Regression Trees (RT), which can identify complex patterns such as non-linear relationships and interactions (Antipov & Pokryshevskaya, 2012).

However, individual RTs are prone to misspecification, overfitting and there is as such little generalization (Provost & Fawcett, 2013). Therefore, tree-based ensemble techniques are used, which address these drawbacks while still being able to capture the complex relationships. These ensemble models create several sub models based on RTs. With the predictions of the sub models, the ensemble models combine the different predictions, of rent in this study, to obtain a final prediction (Hastie *et al.*, 2009). RFR and XGBR both belong to different groups of ensemble models, which are called bagging and boosting, respectively.

RFR is a bagging ensemble method that combines predictions from several individual RTs into a final prediction. In parallel RTs are constructed. Random variation is introduced by selecting a subset of the data along one side for constructing the RT and along the other side by randomly selecting features for constructing branches of the RT (Hastie *et al.*, 2009). From the predictions of different RTs, the average is taken to provide a final prediction. For more technical information on RFR, see Hastie *et al.* (2009).

Unlike RFR, which builds the forest of RTs in parallel, XGBR will build a sequence of trees where the next RT tries to improve the errors of the previous RTs. In its core concept, GBR builds an initial RT, calculates the errors of predictions with the initial RT and then builds a new RT to incrementally lower the prediction error and thus improve the final prediction model (Hastie *et al.*, 2009). Several frameworks have been developed for implementing the boosting algorithm, including XGBR. For more technical details on XGBR, see Chen & Guestrin (2016).

Hyperparameter tuning is done with the Tree-structured Parzen Estimator (TPE), a Bayesian optimizer. This method is chosen since the tuning process is considerably faster and more efficient than a grid search. Furthermore, the outcome is superior to a random search of hyperparameters (Yang & Shami, 2020). Models are then trained via 10-fold cross-validation (CV) on the train set.

The evaluation metrics are calculated on both train and test set, so that an impression can be obtained about the degree of overfitting of the models. Considering there are many evaluation metrics (Steurer *et al.*, 2021), a selection of metrics, which were also considered in the studied literature, is chosen. These metrics are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE) and are defined according to the following formulas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - f_i|, \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - f_i)^2}, \tag{2}$$

$$MAPE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\frac{y_i - f_i}{y_i}\right|, \qquad (3)$$

with n the number of observations (in train and test set for evaluation of models respectively on train and test set), y_i the actual rent for property i and f_i the predicted rent for property i.

IML techniques

The use of tree-based ML models enables post-hoc analyses with IML techniques to interpret ML models and as a result derive insights in the models and make them more transparent. Key determinants and an initial grasp of their impact in predicting Belgian residential rents are determined with two post-hoc model agnostic IML techniques, namely SHAP FI plots and SHAP Summary plots.

SHapley Additive exPlanations (SHAP) values, developed by Lundberg & Lee (2017), is an interpretation method extended from the concept of Shapley value in game theory, which tries to fairly divide player contributions when they collectively achieve a given outcome. In ML, Shapley values are used to measure how much each feature in the model contributes to the prediction as a whole (Molnar, 2022).

SHAP's main contribution is to generate locally additive feature attribution. According to Lundberg *et al.* (2019), additive feature attribution methods have an explanation model g that is a linear function of binary variables:

$$g(z) = \phi_0 + \sum_{i=1}^{M}\phi_i z_i, \qquad (4)$$

where $z \in \{0,1\}^M$, M is the number of input features, and $\phi_i \in \mathbb{R}$.

The z_i variables typically represent a feature being observed ($z_i = 1$) or unknown ($z_i = 0$), and the $\phi_i's$ are the feature attribution values.

To calculate SHAP values, Lundberg *et al.* (2019) define

$$f_x(S) = E[f(x) \mid x_S], \qquad (5)$$

where S is the set of non-zero indexes in z, and $E[f(x)|x_S]$ is the expected value of the function/model $f(x)$ conditioned on a subset S of the input features x. SHAP values combine these conditional expectations with the classic Shapley values to attribute ϕ_i values to each feature:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!}[f_x(S \cup \{i\}) - f_x(S)], \qquad (6)$$

where N is the set of all input features.

In this way, the sum of all SHAP values is equal to the difference between the actual prediction and the average prediction. So, SHAP values aim to explain the difference between actual prediction with the model and the average prediction with the model. It is thus in its core used for local explanations. However, global interpretations are obtained by aggregating the SHAP values. In this study, the SHAP FI plot and the SHAP Summary plot are used as two of these global interpretations.

In the first global interpretation method considered, the SHAP FI plot, the average absolute SHAP values per feature are calculated across the data and then visualised (Molnar, 2022). SHAP FI plots is an alternative to the PFI used in Krämer *et al.* (2021) and Lorenz *et al.* (2022). SHAP FI considers in a highly compressed manner which features have the most impact, globally, in making predictions (Molnar, 2022). The SHAP FI plot then ranks the relative importance of the variables in the prediction models, in this study RFR and XGBR. A higher value of a feature corresponds to a greater effect the feature has in the prediction model. A disadvantage associated with SHAP FI is that there may be bias if unrealistic observations are included. Furthermore, if correlated features are included, this may result in a lower importance of one or more of the correlated features (Molnar, 2022).

Although SHAP FI plots are useful for determining important features, they do not provide more information. Therefore, secondly, the SHAP Summary plot, that combines FI with feature effects, are created (Molnar, 2022). This plot shows all SHAP values for all features of all data represented in a structured way. Through color coding, the plot attempts to map the relationships between feature values, SHAP values and impact on the model's predictions. The plot provides in this manner initial indications of the relationships between values of features and impacts on predictions. For more technical details of SHAP, SHAP FI plot and SHAP Summary plot, see Lundberg *et al.* (2019) and Molnar (2022).

Implementation of MissForest, CatboostEncoder and StandardScaler were performed using scikit-learn in Python 3.9. Implementation of the ML models (XGBR and RFR) and IML techniques (SHAP FI plot and SHAP Summary plot) were done with PyCaret 2.3.6 in Python 3.8.

Results & Discussion

ML models

The evaluation metrics on LR, tuned RFR and tuned XGBR models on both train and test set are shown in Table II. Notice that XGBR on test set scores best for all evaluation metrics with EUR 104.68 for MAE, EUR 210.71 for RMSE and 10.33% for MAPE. Relative to RFR which scores slightly worse with EUR 110.07, EUR 221.95 and 10.83% for MAE, RMSE and MAPE respectively. Notice that LR has the poorest performance of the three models on the test set with EUR 120.82, EUR 246.47 and 11.59% for MAE, RMSE and MAPE. When looking at the relative (percentage) improvement in performance on the test set, it will show for RFR that MAE, RMSE and MAPE are respectively 8.90%, 9.95% and 6.56% lower compared to LR. For XGBR this is 13.26%, 14.51% and 10.87% respectively. Therefore, it is stated that XGBR is the best performing model in this study.

However, note that the absolute differences between evaluation of train set and test set for XGBR are quite large, being EUR 83.26 for MAE, EUR 176.77 for RMSE and 7.92% for MAPE. These large differences may point toward overfitting of the XGBR model on the train data. These differences are less explicit, but still quite sizable for RFR with EUR 45.30, EUR 102.51, and 3,94% for MAE, RMSE and MAPE, respectively, which may indicate less overfitting with this model. However, remark that LR has the smallest differences between evaluation metrics for test and train set with EUR 8.17 for MAE, EUR 19.12 for RMSE and 0.44% for MAPE, indicating that LR overfits the least.

Table II: Results of the LR, RFR and XGBR models

Model	MAE	RMSE	MAPE
Train set			
LR	112.65	227.35	11.15%
RFR	64.77	119.44	6.89%
XGBR	21.42	33.94	2.41%
Test set			
LR	120.82	246.47	11.59%
RFR	110.07	221.95	10.83%
XGBR	104.68	210.71	10.33%

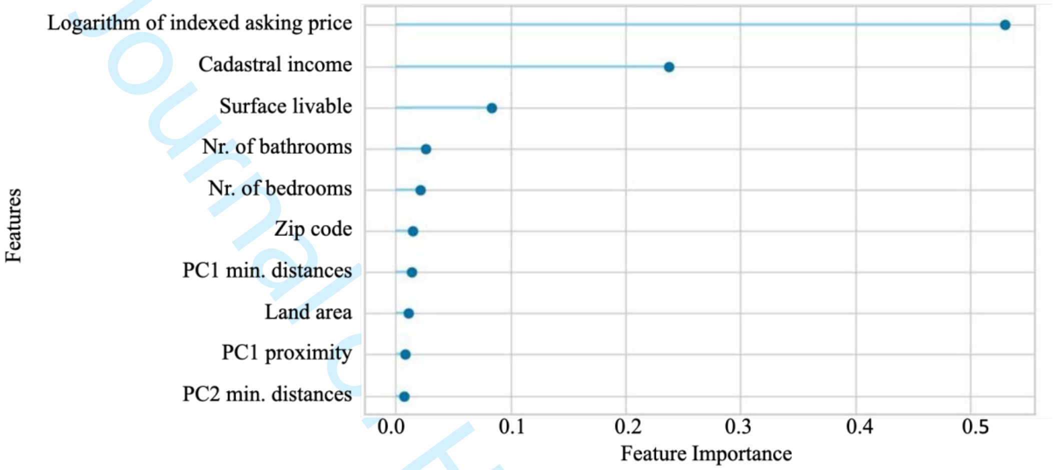
Target variable for evaluation is in terms of monthly rent indexed to 2019 in EUR. Features used are cadastral income, EPC, furniture, garage, garden, land area, logarithm of indexed asking price, number of bathrooms, number of bedrooms, number of toilets, 5 PCs on minimum distance to POI's, 3 PCs on number of POIs in neighbourhood, surface garden, surface livable, terrace, property type, zip code.

IML techniques

The tuned RFR model and XGBR model allow doing post-hoc analyses with the proposed IML techniques. For the IML results, everything is expressed in SHAP values for the logarithm of predicted indexed rents, since the internal workings of the prediction models work with the logarithm of the indexed rents. However, this does not limit the interpretations of the IML techniques used.

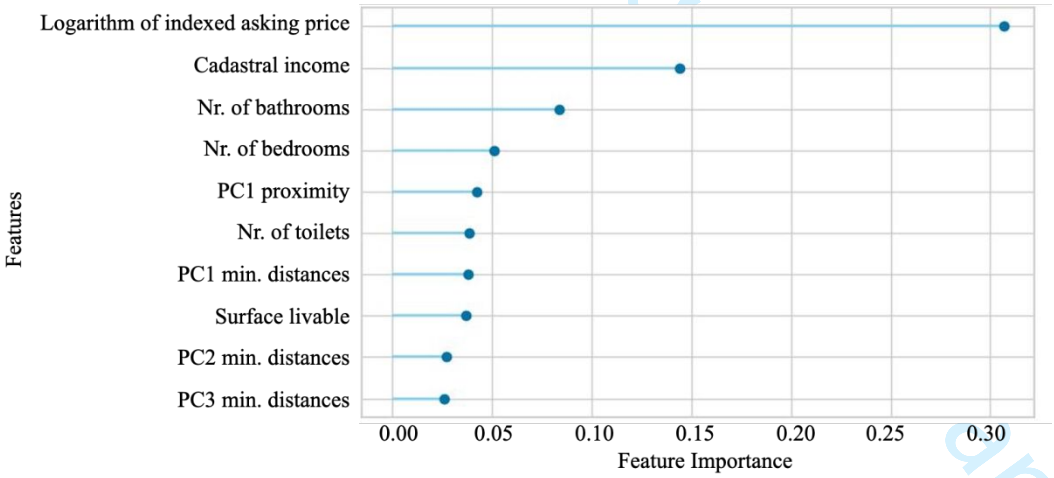
The SHAP FI plot ranks the features of the RFR model and the XGBR model according to impact on the predicted target, namely the logarithm of indexed rent in this study. Figures 1 and 2 show the SHAP FI plot with the 10 most important predictors respectively from the RFR model and the XGBR model. The plots show that for both models, the top 10 predictors are broadly similar. The results are also consistent with what was expected. The plots show that asking price and cadastral income are crucial factors that have a substantial effect in the predictive models. Indeed, these two variables are strongly correlated with rent. Of these two variables, asking price is more important than cadastral income, which serves as an approximation of the average net rental value in the base year 1975. Furthermore, not surprisingly, structural characteristics such as surface livable, number of bathrooms and number of bedrooms are also important. In addition, the importance of location parameters also becomes apparent, albeit to a lesser extent, as some of the variables measuring the proximity to POIs are in the top 10. An important remark here is that asking price and cadastral income probably cover a large part of the information regarding structural characteristics and location factors.

Figure 1: SHAP Feature Importance plot for RFR model



This SHAP FI plot depicts the importance of each feature in the RFR predictive model. Values on X-axis are measured as mean absolute SHAP values for the logarithm of predicted indexed rent. Thus, the SHAP value given here is the difference between the logarithm of the prediction of the indexed rent for an observation and the mean logarithm prediction of indexed rent. With PC1 min. distances the first PC of features of minimum distance to POIs, PC1 proximity the first PC of features of nr. of POIs in 5.5km range and PC2 min. distances the second PC of features of minimum distance to POIs.

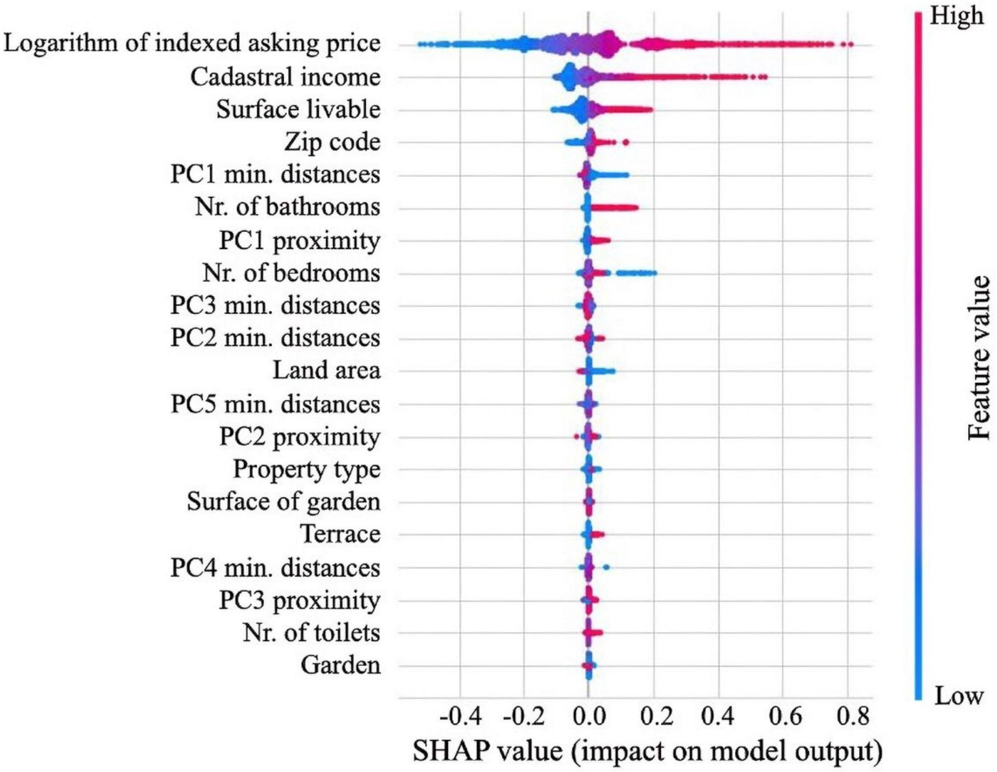
Figure 2: SHAP Feature Importance plot for XGBR model



This SHAP FI plot depicts the importance of each feature in the XGBR predictive model. Values on X-axis are measured as mean absolute SHAP values for the logarithm of predicted indexed rent. Thus, the SHAP value given here is the difference between the logarithm of the prediction of the indexed rent for an observation and the mean logarithm prediction of indexed rent. With PC1 proximity the first PC of features of nr. of POIs in 5.5km range, PC1 min. distances the first PC of features of minimum distance to POIs, PC2 min. distances the second PC of features of minimum distance to POIs and PC3 min. distances the third PC of features of minimum distance to POIs.

To get an idea of the direction and shape of the relationships, the SHAP Summary plots for the RFR model and the XGBR models are examined, which are shown in Figure 3 and 4. These plots depict how high and low feature values are related to the SHAP values in the training data set. Here, by using the SHAP values as a building block, associations are made between features and higher or lower predicted indexed rents in logarithmic terms relative to the average predicted indexed rent in logarithmic terms in the train data. In these SHAP Summary plots, non-linearities can be recognized in addition to linear associations.

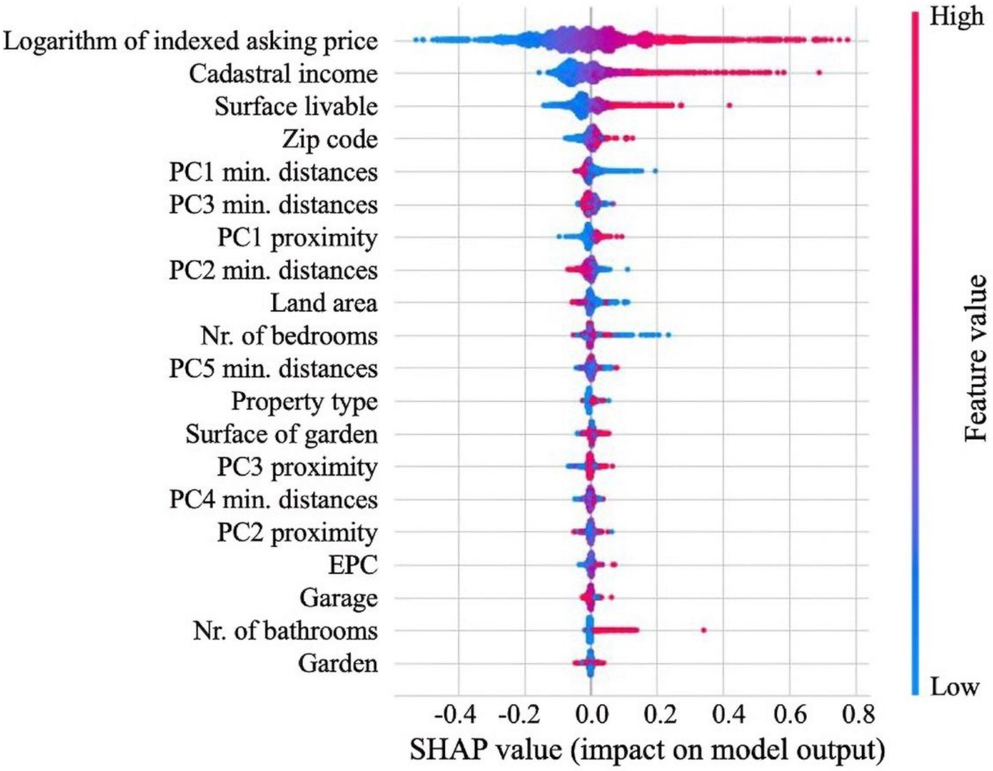
Figure 3: SHAP Summary Plots for RFR model



SHAP Summary plot of the top 20 features of the RFR model. The higher the SHAP value of a feature, the higher the predicted rent. A dot is created for each feature value of each property, and thus one property is allocated one dot on the line for each feature. Dots are colored according to the values of features for the respective property and accumulate vertically to provide insight into the distribution of SHAP values per feature. Red represents higher feature values, and blue represents lower feature values. SHAP values here are measured as mean absolute SHAP values for the logarithm of predicted indexed rent. Thus, the SHAP values given here are the difference between the logarithm of the prediction of the indexed rent for an observation and the mean logarithm prediction of indexed rent. With PC1 min. distances the first PC of features of minimum distance to POIs, PC1 proximity the first PC of features of nr. of POIs in 5.5km range, PC3 min. distances the third PC of features of minimum distance to POIs, PC2 min. distances the second PC of features of minimum distance to POIs, PC5 min. distances the fifth PC of features of minimum distance to POIs, PC2 proximity the second PC of features of nr. of POIs in 5.5km range, PC4 min. distances the fourth PC of features of minimum distance to POIs and PC3 proximity the third PC of features of nr. of POIs in 5.5km range.

An example of such non-linearity for the RFR model is the feature “Nr. of bathrooms”. When looking at the SHAP Summary plot for the RFR model, it is observed that low feature values for 'Nr. of bathrooms', i.e., low number of bathrooms, shown with blue dots, tends to correspond to SHAP values close to zero. SHAP values close to zero have little impact on the model output, here the predicted indexed rent in logarithmic terms. So, this suggests that a low number of bathrooms has a low impact on the predicted rent. This is in contrast when the feature value for “Nr. of bathrooms” is higher, that is, the number of bathrooms is higher, which is indicated by red dots. These higher feature values correspond to larger positive SHAP values. This, in turn, has a positive impact on the predicted indexed rent in logarithmic terms. Thus, this results in higher predicted rents. It can be argued that non-linearity exists in this case. The logical reasoning of the non-linearity between predicted rent and the number of bathrooms can be intuitively deduced. After all, if a property has a low number of bathrooms, it seems reasonable that this would have a lower effect on the rent. In that case, having a bathroom is more of a basic requirement. However, a higher number of bathrooms could indicate properties with more luxury. As such, it may be linked to higher rents.

Figure 4: SHAP Summary Plots for XGBR model



SHAP Summary plot of the top 20 features of the XGBR model. The higher the SHAP value of a feature, the higher the predicted rent. A dot is created for each feature value of each property, and thus one property is allocated one dot on the line for each feature. Dots are colored according to the values of features for the respective property and accumulate vertically to provide insight into the distribution of SHAP values per feature. Red represents higher feature values, and blue represents lower feature values. SHAP values here are measured as mean absolute SHAP values for the logarithm of predicted indexed rent. Thus, the SHAP values given here are the difference between the logarithm of the prediction of the indexed rent for an observation and the mean logarithm prediction of indexed rent. With PC1 min. distances the first PC of features of minimum distance to POIs, PC3 min. distances the third PC of features of minimum distance to POIs, PC1 proximity the first PC of features of nr. of POIs in 5.5km range, PC2 min. distances the second PC of features of minimum distance to POIs, PC5 min. distances the fifth PC of features of minimum distance to POIs, PC3 proximity the third PC of features of nr. of POIs in 5.5km range, PC4 min. distances the fourth PC of features of minimum distance to POIs and PC2 proximity the second PC of features of nr. of POIs in 5.5km range.

Pronounced non-linear relationships are also clearly visible in Figures 3 and 4 for cadastral income, some variables measuring the proximity to POIs, number of bedrooms and land area in the SHAP Summary plot of the RFR model. For the XGBR model, these are largely the same non-linear relationships. It can be noted that in the latter model the non-linearity of the variables measuring the proximity to POIs appears to be less explicit than in the RFR model. Similar intuitions can be formed for these other non-linear relationships.

Furthermore, the association can be made with the superior evaluation metrics of the tree-based models in Table II. The non-linear relationships between rents and their determinants in the SHAP Summary plots demonstrate the benefits of using ML models such as tree-based ensemble models that can capture these non-linear relationships. This contrasts with LR models that capture only linear relationships.

Conclusions

The purpose of this study was twofold. First, it aims to demonstrate the stronger predictive power of tree-based ML models for predicting Belgian residential rental prices compared to a LR model. Second, it interprets the tree-based models with global IML techniques to understand the relationship between predicted rent and feature values. Based on a cleaned data set of 18935 data points, LR, RFR and XGBR models are trained and evaluated, followed by interpretations for the tree-based models with the SHAP FI plots and SHAP Summary plots.

Both tree-based ensemble models, RFR and XGBR, perform relatively well in predicting indexed monthly rent of Belgian residential real estate compared to LR. However, XGBR has a slight advantage given its evaluation metrics on the test set are slightly better than RFR. Regardless, it should be kept in mind that LR seems to be overfitting less than RFR and XGBR, when evaluation metrics on train and test data are compared for the three models. The models face the risk of not generalizing well because of this possibility of overfitting. This implies that the results of the IML techniques should be viewed with caution, as the SHAP FI and SHAP Summary plot might be influenced by the overfitting.

SHAP FI plots demonstrate the importance of the 10 most important variables. The results show that asking price, cadastral income, surface livable, number of bedrooms and number of bathrooms are important predictors in the RFR model. Similar results for the XGBR model, where the top 5 predictive determinants are asking price, cadastral income, number of bedrooms, number of bathrooms and a variable that measures the proximity to POIs. Asking price and cadastral income are very important in both prediction models, not surprisingly as they are highly correlated with monthly rent indexed to 2019.

In its own way, the SHAP Summary plot additionally displays the relationships between feature value, SHAP values and predictions. Here, it is visually observed that there are linear relationships between certain features and rents, but that there are also some non-linear relationships. This, in addition, shows why previous research favors ML models to the disadvantage of classic hedonic models such as LR, given that (some) black-box ML models can capture non-linear relationships. Thereby, it can also be suggested that this is a possible reason why tree-based models prevail in the more recent literature, as the basis ingredient of these tree-based ensemble models used, namely RTs, is good at capturing non-linearities.

These findings show that IML techniques can make black-box ML models more transparent. Where, as stated by Valier (2020), ML can be more effective at predicting than LR and the latter had more capacity for explanation, IML can be used to ensure that better predictions from ML models are enriched by interpretations and explanations. Despite their simplicity, the IML techniques used are an effective tool for visualizing the impact of determinants in tree-based rent prediction models. This transparency may persuade wait-and-see parties to adopt and accept ML. After all, using ML can be beneficial to several stakeholders. Just to mention real estate agents who can explicitly say why the property is worth less than the sellers hope or banks and insurance companies who can get a better picture of the property and its valuation.

This paper provides some empirical and managerial implications. First, guidance is provided on the use of IML techniques to obtain some insights into rent and its determinants. In fact, IML provides insights into the black-box tree-based ML models that are used to predict the rental price of Belgian residential real estate. Moreover, several factors of predicted Belgian residential rents and their relationship with rents are derived. These determinants include asking price, cadastral income, surface livable, number of bedrooms, number of bathrooms and a variable measuring the proximity to POIs. This information can be used by policy makers, real estate professionals and investors to understand rent and its associated determinants. This may be useful in decision-making as knowing which variables have the highest impact on rent prices can help property managers and investors make informed decisions on pricing, marketing, and property improvements. These factors can also be used in models to predict and understand trends in the Belgian rental market. Next to that, the findings reinforce the widely accepted idea that both structural and location factors are important in modelling rental prices. Furthermore, the existence of non-linear relationships between predicted rent and some of its determinants can be confirmed. This information implies that stakeholders, who want to get a grasp of the relationships between rent and its determinants, should take these non-linearities into account for policy and strategic decisions. Furthermore, stakeholders that aim to create a model for the Belgian residential rental market should use models capable of capturing these non-linear relationships. However, since there are indications of overfitting of the tree-based ML models, the results from the models may not generalize well. Hence, it is important to be cautious when using these models. It is thus recommended for stakeholders, who aim to build rent prediction models, to minimize overfitting by applying cross-validation, regularization, pruning or early stopping at the modeling stage. Additionally, it is important to remember that while predictive models can be useful tools for making predictions, they should not be relied upon exclusively for decision-making, and should be used in combination with other sources of information and expertise.

Looking at limits, it is worth noting that in the feature set indexed asking price and cadastral income are used. These two features are highly correlated with indexed monthly rent. To better understand the effects of structural and location factors, it is useful to train the models without these two features. After all, part of the content of other factors will be contained in asking price or cadastral income. As such the predictive value of the structural and location features will be carried by asking price and cadastral income and thus lead to suboptimal interpretations with IML, as also stated by Molnar (2022).

In addition, the sample size is also relatively small to make a prediction model for Belgium as a whole. Furthermore, missing values are imputed, and categorical variables are encoded, which can lead to introduction of noise. Thus, it would be an opportunity to take full advantage of two other strengths of tree-based models, namely the ability of working with missing values and categorical variables (Antipov & Pokryshevskaya, 2012).

Looking at further extensions, the possibilities seem ample. Other global interpretation techniques can be considered, such as permutation FI, partial dependence plots and accumulated local effect plots that Krämer *et al.* (2021) and Lorenz *et al.* (2022) already applied in their research. Krämer *et al.* (2021) applied it for seven major cities in Germany. Lorenz *et al.* (2022) applied it for Frankfurt am Main, which is also a major German city. These global techniques are interesting for deriving and decomposing drivers, as well as for studying trends and evolutions in real estate markets. Furthermore, analyses with local interpretation techniques, such as LIME, SHAP values and counterfactuals can also be considered to provide explanations of individual predicted rents. The latter interpretations are of interest to landlords, tenants, and real estate agents, for example. In addition, as Molnar (2022) states, the field of IML is still young and developing. Thus, techniques will be created that lead to different, complementary, and possibly better interpretations of the ML models on both global and local levels.

References

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Breuer, W., & Steininger, B. I. (2020). Recent trends in real estate research: A comparison of recent working papers and publications using machine learning algorithms. *Journal of Business Economics*, *90*(7), 963–974. https://doi.org/10.1007/s11573-020-01005-w

Cadastral income | Belgium.be. (2022, March 23). Belgium.Be. https://www.belgium.be/en/housing/buying_or_selling_home/cadastral_income

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chiwuzie, A., Dabara, D., Omotehinshe, O., Prince, E., & Aiyepada, E. (2021). Changing macroeconomic indicators and the rental values of residential properties in. *YBL Journal of Built Environment*, *5*, 1–18. https://doi.org/10.33796/ajober.5.1.01

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Krämer, B., Stang, M., Nagl, C., & Schäfers, W. (2021). *Explainable AI in a Real Estate Context—Exploring the Determinants of Residential Real Estate Values* (SSRN Scholarly Paper No. 3989721). https://doi.org/10.2139/ssrn.3989721

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable machine learning for real estate market analysis. *Real Estate Economics*, *0*(0), 1–31. https://doi.org/10.1111/1540-6229.12397

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). *Consistent Individualized Feature Attribution for Tree Ensembles* (arXiv:1802.03888). arXiv. https://doi.org/10.48550/arXiv.1802.03888

Ma, Y., Zhang, Z., Ihler, A., & Pan, B. (2018). Estimating Warehouse Rental Price using Machine Learning Techniques. *International Journal Of Computers Communications & Control*, *13*(2), Article 2.

Malpezzi, S. (2002). Hedonic Pricing Models: A Selective and Applied Review. In *Housing Economics and Public Policy* (pp. 67–89). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470690680.ch5

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Molnar, C. (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Independently published.

Oshodi, O. S., Thwala, W. D., Odubiyi, T. B., Abidoye, R. B., & Aigbavboa, C. O. (2019). Using neural network model to estimate the rental price of residential properties. *Journal of Financial Management of Property and Construction*, *24*(2), 217–230. https://doi.org/10.1108/JFMPC-06-2019-0047

Piegeler, T., & Bauer, S. (2021). *Knowing what others don't: Gaining a competitive edge in real estate with AI-driven geospatial analytics*. Deloitte. https://www2.deloitte.com/ce/en/pages/real-estate/articles/gaining-a-competitive-edge-in-real-estate.html

Provost, F., & Fawcett, T. (2013). *Data science for business: What You Need to Know about Data Mining and Data-Analytic Thinking* (1st ed.). O'Reilly.

Rent Calculator | Statbel. (2017). https://statbel.fgov.be/en/themes/consumer-prices/rent-calculator

Shen, H., Li, L., Zhu, H., & Li, F. (2022). A Pricing Model for Urban Rental Housing Based on Convolutional Neural Networks and Spatial Density: A Case Study of Wuhan, China. *ISPRS International Journal of Geo-Information*, *11*(1), Article 1. https://doi.org/10.3390/ijgi11010053

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, *38*(2), 99–129. https://doi.org/10.1080/09599916.2020.1858937

Surkov, A., Srinivas, V., & Gregorie, J. (2022, May 17). *Unleashing the power of machine learning models in banking through explainable artificial intelligence (XAI)*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/financial-services/explainable-ai-in-banking.html

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*, *38*(3), 213–225. https://doi.org/10.1108/JPIF-12-2019-0157

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. https://doi.org/10.1136/bmjopen-2013-002847

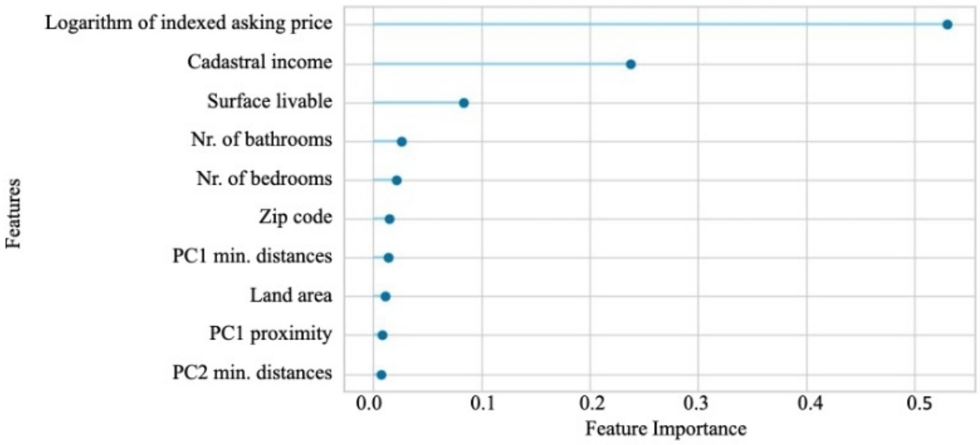
Xu, L., & Li, Z. (2021). A New Appraisal Model of Second-Hand Housing Prices in China's First-Tier Cities Based on Machine Learning Algorithms. *Computational Economics*, *57*(2), 617–637. https://doi.org/10.1007/s10614-020-09973-5

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zhang, P., Hu, S., Li, W., Zhang, C., Yang, S., & Qu, S. (2021). Modeling fine-scale residential land price distribution: An experimental study using open data and machine learning. *Applied Geography*, *129*, 102442. https://doi.org/10.1016/j.apgeog.2021.102442

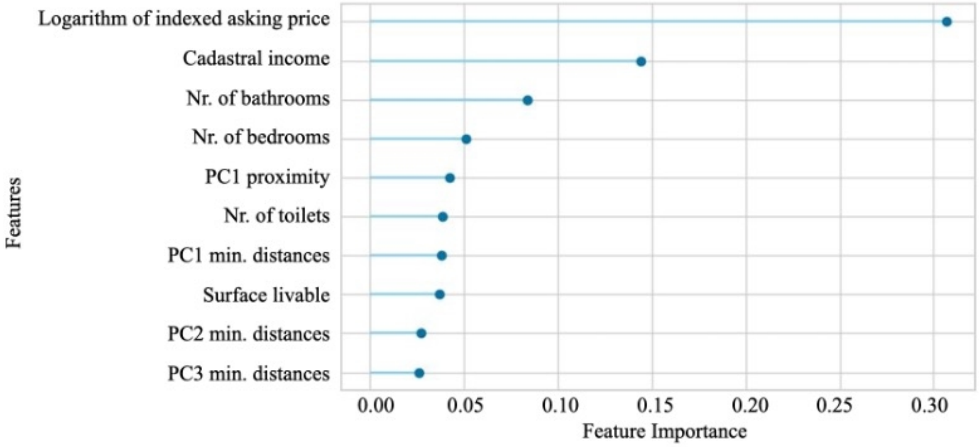
Zhou, X., Tong, W., & Li, D. (2019). Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. *ISPRS International Journal of Geo-Information*, *8*(8), Article 8. https://doi.org/10.3390/ijgi8080349

Zulkifley, N., Rahman, S., Nor Hasbiah, U., & Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *International Journal of Modern Education and Computer Science*, *12*, 46–54. https://doi.org/10.5815/ijmecs.2020.06.04



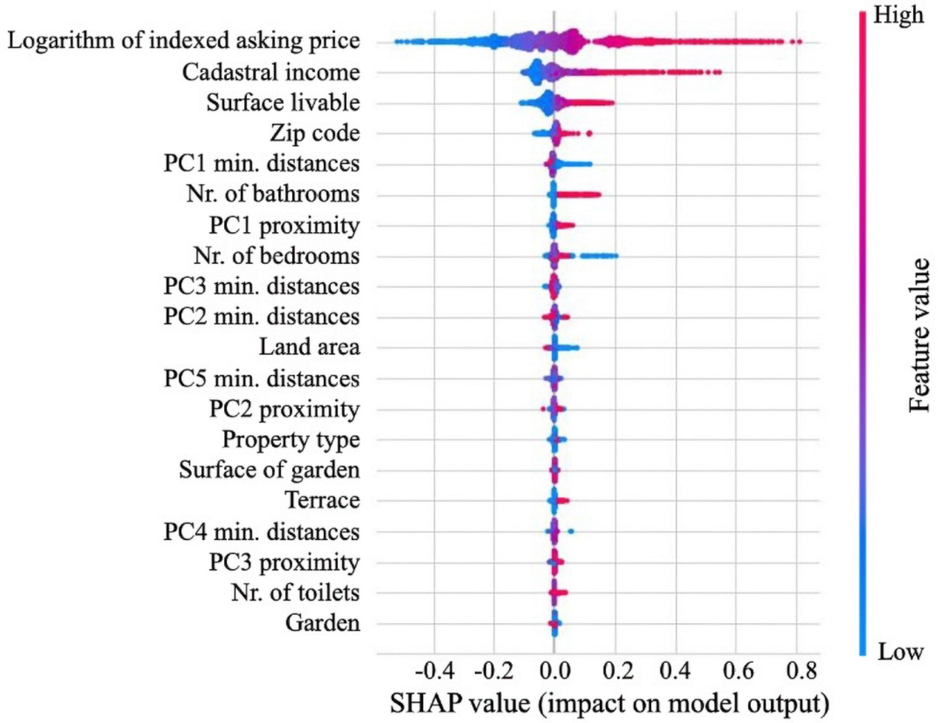
SHAP FI for RFR model

140x63mm (144 x 144 DPI)



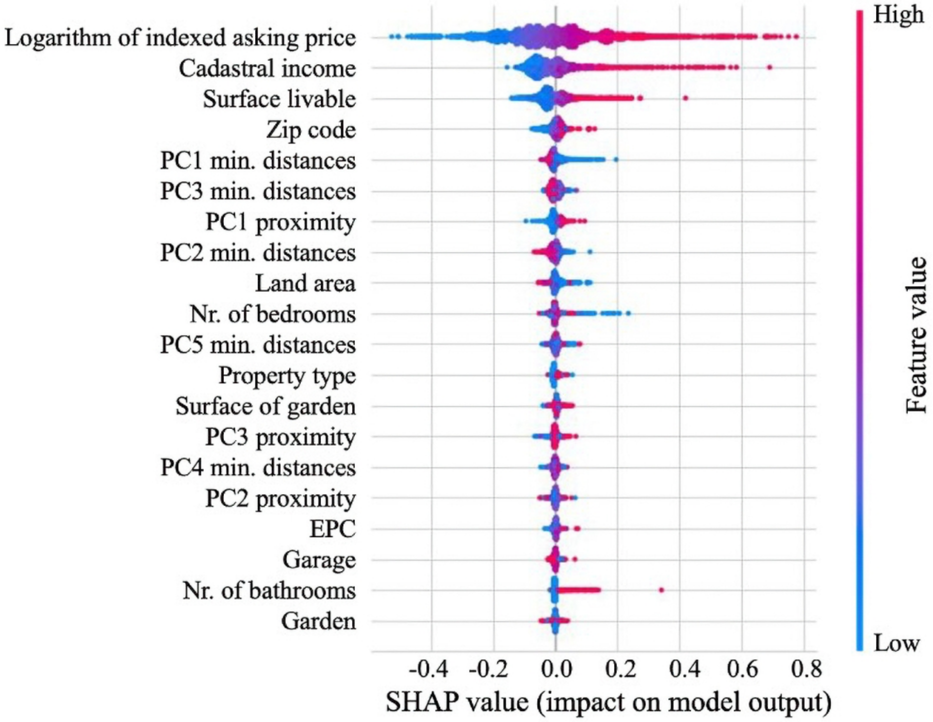
SHAP FI for XGBR model

140x64mm (144 x 144 DPI)



SHAP Summary plot for RFR model

168x125mm (240 x 240 DPI)



SHAP Summary plot for XGBR model

168x125mm (240 x 240 DPI)