Journal of Business Research 144 (2022) 50–65

Contents lists available at ScienceDirect

Journal of Business Research

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

How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market

Tomasz Potrawa [a], Anastasija Tetereva [b,*]

[a] Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands
[b] Department of Econometrics, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

ARTICLE INFO

Keywords:
Hedonic pricing
Housing market
Unstructured data
Image and text recognition
Model-agnostic methods

ABSTRACT

Understanding the customers' perception of the value of constituent characteristics of a good is among the key questions in any pricing strategy. Hedonic pricing allows such an analysis and is frequently applied in economic fields. Although it is regarded as a benchmark in its original form, the availability of new data sources and the development of machine learning techniques created a space for further improvement. In this study, we propose a general framework for applying machine learning tools to enhance the hedonic pricing model in several directions. We do this, first, by adding image and text sources to conventional data and then by applying an advanced nonparametric prediction model. Lastly, we use model agnostic analysis to uncover new pricing factors and unravel complex relationships that could not be captured by conventional models.

1. Introduction

The analysis of the real estate market has always drawn attention of researchers and practitioners. As a relatively stable, comparable, and easily accessible source of data, the housing market is used as a proxy for unobserved phenomena such as environmental evaluations, macroeconomic fundamentals, and socioeconomic aspects. However, many previous empirical studies showed that estimated prices do not always coincide with the actual sale prices. This bias can be addressed to omitted variables (de Koning et al., 2018; Ghysels et al., 2012) or insufficient flexibility of a modeling procedure (Mason & Quigley, 1996; McMillen & Redfearn, 2007), leading to spurious conclusions in fields that rely on the real estate market and causing wrong policy decisions. Therefore, the ability to model and price real estate correctly is crucial.

Among the most popular methods in pricing theory is hedonic price modeling. According to this approach, a good does not provide any utility by itself. Instead, it consists of characteristics that hold constituted utility. The market price consumer pays for a given good is related to the utility of these characteristics. Therefore, by comparing the prices of goods with different levels of utility for the characteristic of interest, it becomes possible to quantify their value (Lancaster, 1966).

The above-mentioned characteristic of hedonic pricing makes hedonic pricing a perfect tool for estimating the price of such hardly tangible aspects as air quality, green area proximity, a view of the ocean, or the criminality level in the neighborhood (Benson et al., 1998; Bishop & Lange, 2005; Clark & Herrin, 2000; Wolf, 2007). Hedonic models aim at quantifying these individual characteristics while keeping the model interpretable. Consequently, the majority of papers related to hedonic pricing are based on simple parametric models employing structured conventional data. Ordinary least squares (OLS) regression is a perfect example of such a model, as it allows deriving the constituted price of the attribute of interest by simply analyzing the regression coefficients. However, OLS is a parametric model that is subject to several crucial assumptions. Although it can deliver accurate predictions when these assumptions are not met, interpretation of its coefficients might be misleading. In practice, most novel big data generating processes fail to fulfill the usual OLS assumptions and can provide misleading interpretations.

Recent methodological advancements in the areas of machine learning (ML) and artificial intelligence (AI) may be used to improve the performance of marketing research in various ways (Ma & Sun, 2020). Advanced machine learning models such as neural nets (Abidoye & Chan, 2018), random forests (Hong et al., 2020; Neloy et al., 2019), boosted trees (Neloy et al., 2019), or support vector machines (Oladunni & Sharma, 2016) usually offer more predictive power than their conventional counterparts. However, these models are often called black boxes due to their limited interpretation. The majority of studies employing black-box models for a housing market pursued the goal of achieving better predictive accuracy, completely abandoning models' interpretability (Abidoye & Chan, 2018; Hong et al., 2020; Ma & Sun, 2020; Neloy et al., 2019). The inability to uncover patterns and

* Corresponding author.
E-mail address: tetereva@ese.eur.nl (A. Tetereva).

https://doi.org/10.1016/j.jbusres.2022.01.027
Received 8 April 2021; Received in revised form 7 January 2022; Accepted 10 January 2022
Available online 5 February 2022

understand the impact of employed covariates on the housing prices diminished the practicability of the mentioned papers. This inscrutability is especially troublesome outside of the academic world where users may find it difficult to make policy decisions based on conclusions drawn from the model they do not understand. Further, without understanding the logic of a model, it becomes impossible to derive from and invest in factors contributing the most to the prediction. Thus, it is to no surprise that many recent hedonic studies still rely on easily interpretable but often less accurate OLS-based models (Gibbs et al., 2018; Hussain et al., 2019; Zhang & Dong, 2018).

Given the extensiveness in which new AI models are applied in the marketing-oriented areas, the urge of reaching their reliable interpretation is unquestionable. Hence, this demand is addressed by the development of numerous explainable AI (XAI) methods; see Lundberg and Lee (2017), Rai (2019), Ribeiro et al. (2016) and Zhao and Hastie (2019) for more details. If applied properly, the XAI methods may not only reach similar interpretability to the parametric models but can also provide more insights. This improvement leads to a plethora of benefits for a business, e.g., a better understanding of the clients, making conscious decisions, and providing insights on the causality of the uncovered trends (Lipton, 2016; Rai, 2019; Shrestha et al., 2021).

It is estimated that, on average, 80% of the data that companies possess are unstructured and take forms such as images and text (Dayley & Logan, 2015). It is important to note that the volume of these novel sources of information is growing 15 times faster than the structured ones (Nair & Narayanan, 2012). Therefore, unstructured data-based studies usually rely on a considerably wider information set. This can be important in uncovering new factors and dependencies that are not captured by using conventional structured data. The importance of employing rich sources applies especially to the customer-oriented research, which aims at specifying consumers' sentiments and needs (Mustak et al., 2021).

In this paper, we propose a general framework for the application of advanced machine learning methods to reinforce the performance and the interpretability of traditional hedonic pricing models. Besides focusing on these two aspects, we attempt to account for the well-established hedonic theory of Rosen (1974). Particular attention is paid to uncovering the theoretically justified nonlinear dependence of hedonic prices related to the utility of the housing attributes. The empirical illustration considers the Rotterdam housing market and may be divided into three parts. First, we propose an easy and reproducible way of gathering data about any housing market through scraping the information from rental websites and Google Maps. With image recognition and text analysis methods, we present the methodology for defining and extracting the most relevant covariates from unstructured data. We expand much previous research such as Chen et al. (2020), Law et al. (2019) and Zhang and Dong (2018) by combining numerous sources of data into a complete ML framework, imitating the customers' data gathering. Second, we use the collected data next to more conventional variables to create hedonic pricing models for rental prices in Rotterdam. To verify if the application of ML-driven methods significantly increases the performance of hedonic models, we conduct a two-fold comparison. We check whether higher predictive accuracy is associated with the employment of unstructured data, the application of a more advanced black-box model, or both. Such differentiation aims to separately measure the added value of using new sources of information and new modeling techniques. The third part of the study illustrates how specific black-box uncovering XAI methods can be used to study individual attributes' prices. By applying such methods, we aim to show that, due to recent methodological advancements (Lundberg & Lee, 2017; Ribeiro et al., 2016), the ML models do not have to be treated as black boxes anymore. Further, as XAI methods are built on top of complex, nonparametric models, their application may help find new factors and uncover potential nonlinear dependencies that are typically not captured by the conventional hedonic approach.

We view our study as the modernization of hedonic methodology, bridging a gap between two distinct groups of research. The first one, focusing on quantifying the value of good's characteristics via traditional hedonic modeling, and the second one, pursuing the best possible accuracy by employing complex ML models. The findings illustrate the usefulness of incorporating complex data sources in increasing the accuracy of hedonic pricing models. When compared to the traditional methods, the proposed machine-learning-driven approach leads to an increase of 25% in predictive accuracy in the presented settings. Moreover, the applied black-box uncovering methods show that the estimated hedonic prices of housing attributes are not linear. We discover that the marginal price of living area for rented properties is only piecewise linear. Each additional squared meter from 1 to 136 m is estimated to cost 7.26 euros, from 136 to 191 m 1.65 euros, and above 191 m 5.70 euros. Further, we find that increasing the total number of rooms above seven does not impact the rental prices in Rotterdam. Additionally, we observe that income requirements imposed by landlords are related to the drop in rental prices. With the usage of covariates extracted from satellite images, we also aim at accounting for such aspects as traffic, noise pollution, and proximity to green areas and water bodies. Lastly, we discover that the value of housing attributes depends not only on their own quantity and quality but also on the quantity of other attributes. The value of the view on Rotterdam panorama is estimated to cost over 100 euros for large and well-localized properties which is almost 50 euros more than for a less expensive real estate.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive literature review of the theory and application of hedonic pricing models. Section 3 describes data collection. Section 4 presents the methodology and the applied analysis regarding feature extraction, predictive modeling, and augmenting the interpretability of the created regression models. Finally, Section 5 discusses the implications of our findings and their potential usage in real world. Lastly, the limitations of the study are acknowledged and suggestions for future research are provided.

2. Literature

2.1. Hedonic pricing approach in housing market

Hedonic pricing has been extensively applied in real estate appraisal research due to the housing market characteristics. Real estate matches the assumptions of a hedonic approach as it may be treated as a good consisting of multiple individual attributes such as a living area, a number of rooms, or localization. The attributes of a property in the hedonic approach tend to be divided into multiple categories.

The current study considered approach by Chin and Chau (2003), who proposed the division of characteristics into three main groups: the locational group consisting of real estate characteristics, such as the distance to the central business district (CBD) and the type of view from a property, structural attributes describing the living area, the number of bedrooms, or the age of a building, and features characterizing the neighborhood of a real estate. The attributes considered in the current study complement the list by Chin and Chau (2003) and are presented in Table 1.

2.2. Methodology of the previous hedonic studies

2.2.1. Linear approach

The first formal contribution to hedonic price analysis may be granted to Court (1939), who examined automobile price indices. However, the popularization of hedonic pricing took place many years later thanks to Griliches (1961). The considerable response to Griliches (1961) led to the swift development of hedonic pricing (Goodman, 1998), also from a theoretical micro-econometric perspective. Lancaster (1966) and Rosen (1974) contributed greatly to the theory of hedonic

Table 1
Housing attributes commonly used in hedonic pricing models.

Attribute		Reference
Locational	Distance from CBD	McMillan et al. (1992) and Palmquist (1992)
	View on the sea, lakes, rivers, hills etc.	Gillard (1981) and Mok et al. (1995)
	Obstructed view	Benson et al. (1998)
Structural	Number of rooms, bedrooms, bathrooms	Ball (1973) and Garrod and Willis (1992)
	Living area	Ball (1973) and Garrod and Willis (1992)
	Basement, garage, storage and patio	Forrest et al. (1996) and Garrod and Willis (1992)
	Building services (e.g. lift, AC)	Garrod and Willis (1992)
	Floor level	So et al. (1996)
	Structural quality	Kain and Quigley (1970)
	Facilities (e.g. swimming pool, gym)	Garrod and Willis (1992)
Neighborhood	Age of the building	Clark and Herrin (2000) and Li and Brown (1980)
	Income of residents	Kain and Quigley (1970)
	Proximity to good schools	Clark and Herrin (2000)
	Proximity to hospitals	Huh and Kwak (1997)
	Proximity to places of worship	Carroll et al. (1996)
	Proximity to hazardous industrial facilities	Grislain-Letrémy and Katossky (2014)
	Proximity to shopping centers	Des Rosiers et al. (1996)
	Proximity to green areas	Bishop and Lange (2005) and Wolf (2007)
	Proximity to water bodies	Colby and Wishart (2003)
	Crime rate	Clark and Herrin (2000)
	Traffic/airport noise	Espey and Lopez (2000) and Williams (1991)
	Environmental quality (e.g. landscape)	Clark and Herrin (2000)
	Air quality	Smith and Huang (1993) and Zhao and Hastie (2019)

pricing. Although both models are still seen as a benchmark in urban, environmental, or labor economics (Greenstone, 2017), they suffer from several empirical limitations.

Among the most notable challenges is the employment of hedonic pricing theory in a parametric model. The usual assumption of a linear model is the Gaussian distribution of data which is rarely observed in practice. This is motivated by the fact that, only for Gaussian data, a linear relation is a perfect proxy for a general dependence. Conversely, without the validity of this assumption, linear models might provide misleading results. Moreover, considering all possible interactions among characteristics might lead to multicollinearity problems in OLS and, as a result, unreliable standard errors of estimated coefficients. The usual practical approach is transforming the data closer to Gaussian. Several transformations may be applied in hedonic models, e.g., linear, semi-log, or Box–Cox. Nevertheless, there is little research on how the transformation, or in other words the functional form, should be chosen (Butler, 1982). The most popular technique, Box–Cox transformation, automatically identifies transformation to convert the data as close to Gaussian distribution as possible (Sakia, 1992). However, it does not guarantee that the transformed data will suit the model and fulfill the assumptions.

However, methodological drawbacks of the traditional OLS-based models did not stop researchers from following this approach (Gibbs et al., 2018; Hussain et al., 2019; Zhang & Dong, 2018). The ease of use, combined with the clear interpretability of the created models, often overshadowed the potential inference issues of the models. To this day, numerous recent well-received papers, although still insightful in the context of their study areas, may suffer from the mentioned methodological issues.

2.2.2. Spatial approach

The classic linear models such as OLS regression do not directly consider spatial interactions in the data. The basic assumption of this type of model is the constant relation between dependent and independent variables, unaffected by the geographical space (Cellmer et al., 2020). The use of spatially-dependent variables in the linear settings may result in spatial autocorrelation and spatial non-stationarity, causing the violation of OLS assumptions (Getis, 2011; Kim et al., 2020). Consequently, the linear approach not only fails in capturing the local variations among the predictors (Yoo & Wagner, 2016) but also causes the biased inference of the model (Fotheringham et al., 2002).

Numerous spatial models were developed as an extension of the classic linear approach, usually taking the form of parametric or semi-parametric models such as geographically weighted regression and spatial regression (Basile & Mínguez, 2018). By including the geographical space into the analysis, studies employing spatial approach often surpassed the linear models in terms of both explanatory power and predictive accuracy (Anselin & Lozano-Gracia, 2009). However, most spatial models did not manage to eliminate the restrictions and drawbacks similar to the OLS approach (Hengl et al., 2018).

First, parametric spatial models require the residuals to be stationary and normally distributed. Similarly, as in the linear approach, the data transformations such as Box–Cox are a possibility, although they do not guarantee that the final distribution will meet the model's assumptions (Basile & Mínguez, 2018).

Second, it is difficult to capture complex, nonlinear data patterns, especially in the parametric versions of spatial models. Similarly, as in OLS, some flexibility is offered by transforming the predictors, although it may lead to functional form bias. Some semi-parametric spatial models employing regression spline methods allow for approximating the nonlinear relationships between the dependent variable and the covariates (Basile & Mínguez, 2018). However, the spline method is not expected to work well in the case of adjacent observations, having notably different data values. This limitation is in line with the concept of spatial auto-correlation, claiming that observations in the direct neighborhood should possess similar values. However, in practice, it may not always be the case.

Third, accounting for interaction effects in spatial models is problematic. Usually, the interactions between covariates have to be manually specified. Further, combining numerous interactions with the estimation of separate coefficients based on geographical space leads to an enormous amount of model parameters (Hengl et al., 2018).

2.2.3. Nonparametric approach

The limitations of the parametric approach induced researchers to employ numerous nonparametric models that are not constrained by the form and the predictors' distribution. Mason and Quigley (1996) were among the first to follow this approach by considering a nonparametric procedure to choose optimal functional form for the hedonic pricing model. The authors employed a nonparametric generalized additive model to estimate the hedonic function. Additive structure made it possible to keep the theoretical hedonic pricing framework and relax the constraints imposed by the traditional methodology. Consequently, the nonparametric model uncovered nonlinearities in the data that could not be modeled with the traditional approach.

Throughout the last two decades plethora of studies aimed at increasing the accuracy of predictive models by employing advanced machine learning models. ML models, such as random forest (Hong et al., 2020; Neloy et al., 2019), neural nets (Abidoye & Chan, 2018), and support vector machines (Oladunni & Sharma, 2016) typically provided better results than the traditional hedonic regression in the context of a housing market. Further, comparisons between tree-based models and spatial models showed slight superiority of the former in terms of predictive power (Credit, 2021; Hengl et al., 2018). Hengl et al. (2018) emphasized the natural ability in capturing interaction

effects and nonlinearities of the random forest method, contributing to its advantage over traditional linear and spatial models.

The common denominator of the studies employing machine learning algorithms for the housing market was the focus on predictive performance. This center of attention combined with a difficult interpretation of the ML models often overshadowed the essence of the hedonic approach: the estimation of housing characteristics' value. Contrary to the first attempts by Mason and Quigley (1996), most ML-driven studies overlooked this estimation (Abidoye & Chan, 2018; Hong et al., 2020; Neloy et al., 2019; Oladunni & Sharma, 2016). Only recently, some attention started to be paid to the interpretability of black-box models in the context of a housing market. Zhao and Hastie (2019), through the usage of partial dependence plots (PDP) and individual conditional expectation (ICE), attempted to measure the impact of air pollution on the housing prices. However, as global methods, PDP and ICE did not explain the model's behavior behind every prediction. This methodological challenge was addressed by the development of explainable methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). Nonetheless, due to their recency, there has been a limited number of studies employing these methods in the context of a housing market. As one of the first, Chen et al. (2020) employed SHAP to the XGBoost model, successfully deriving nonlinear patterns in the impact of urban environmental elements on housing prices in Shanghai. However, the study was limited in several ways, which are discussed in detail in the following Section 2.2.4.

2.2.4. Feature extraction

A significant drawback of hedonic pricing is the unrealistic assumption of perfect competition. It implies that the information flow between consumers and suppliers is instant and is not disrupted. Consequently, the model does not consider the delayed reaction of a market to any changes or an imperfect estimation of the value of given attributes. As the assumption of the perfect competition is not unique to hedonic models, mentioned limitations are relatively popular in economic models, although hedonic pricing models face a major limitation. In real life, it is impossible for consumers to have full knowledge of every attribute of a good. Nevertheless, they have to estimate the value/utility of a good based on the knowledge they possess (Kask & Maani, 1992). Therefore, in our eyes, the best possible model has to include multiple sources of information processed by the customers while making their purchases.

Apart from predictive models discussed in Section 2.2.3 above, machine learning models were also used in the hedonic literature to extract information from complex data sources. Poursaeed et al. (2018) used a convolutional neural net to estimate the level of luxury for analyzed real estate, based on the available photos. Some studies did not categorize the data into explicit variables, such as the luxury level. However, they still confirmed the usefulness of image-sourced data in boosting the predictive performance of regression models (Ahmed & Moustafa, 2016; You et al., 2017).

Although the neighborhood stands for 15% up to 50% of the standardized variation of a site evaluation model (Linneman, 1980), this group of external attributes was not an object of special attention in most recent ML-driven housing research. Law et al. (2019) and Zhang and Dong (2018) are two of the few recently published studies that included the neighborhood aspect by successfully using street views and satellite images in the house price evaluation model.

Similar to the studies using ML in predictive modeling, the papers using ML in feature extraction focused on increasing the overall predictive performance. Without evaluating the extracted features in the context of housing prices, these papers did not contribute significantly to the hedonic literature. To our best knowledge, Chen et al. (2020) is the only study that proposed the framework combining ML methods in all three aspects: feature extraction, predictive modeling, and model interpretation.

The framework proposed in the current study is somewhat similar to the one of Chen et al. (2020), although it expands it in multiple ways. First, Chen et al. (2020) suffered from an extremely limited number of housing characteristics used in the model, potentially leading to the omitted-variable bias. We address this issue by presenting a complete research framework, including a variety of housing characteristics originating from rental websites, google maps, images, and text sources. Second, considering the spatiality of the housing data, we use geographical coordinates in ML predictive models instead of simple distances to specific city areas, such as the central business district (CBD). Third, we conduct a full comparison between the ML-driven approach and the traditional hedonic modeling, emphasizing the limitations of the latter in terms of predictive performance and interpretability.

3. Data

Unlike most price-based hedonic pricing models, this study used the rental costs. Such an approach, however, brings a minor drawback related to interpretation. While comparing the findings of this paper to previous research, it has to be acknowledged that the rent:price ratio is not constant. The ratio differs among properties depending on their attributes, e.g., large living area or expensive neighborhood is related to an increase in the price:rent ratio (Bracke, 2014; Clark & Lomax, 2019). The dynamics of real house prices, as well as such macroeconomic policies as low-interest rates, also impact the ratio over time (Sommer et al., 2010). Further, governmental policies aiming at regulating the rental market may affect the rental costs and the rent:price ratio. However, the last aspect is not problematic in the context of this research. In the Netherlands, the rent price regulations such as ceiling costs apply only for social housing, which in 2020 were defined as properties with the total rent cost below 737 euros (Ministry of the Interior Kingdom Relations, 2021). In the gathered dataset properties that fall into this category stand only for 2.2% of the total sample size.

In the conditions of such a hardly regulated market, the strong correlation between the price and rental cost is undeniable. Therefore, the majority of the findings should be applicable to the traditional housing market. Moreover, an advantage of considering rental prices is their higher reliability, as they are rarely a subject of bargaining. Further, due to the rental data accessibility, this research can be reproduced for any other city or region. The data may be collected without reaching third parties such as municipalities or broker agencies.

Based on the problem with the assumptions of perfect competition discussed in Section 2.1, we built the dataset by imitating the process of finding a flat in real life, i.e., the data used for the research reflected the factors that potential tenants mostly pay attention to. The most crucial part of the data describing the basic housing characteristics (BHC) was web-scraped from one of the leading rental websites in the Netherlands. Typical variables of this group were rental price, type of house, or living area of a property. Subsequently, text descriptions of all properties were also collected.

Simultaneously with scraping the data of house characteristics, the images of each rental offer were scraped. With an average of 22 photos per an offer, over 40 000 images were collected. Additionally, a set of over 2000 photos originating from other rental offer websites were collected to serve as a training set for image recognition models. Following this approach assures that the potential increase in the regression accuracy would be caused by the fully-automated feature extraction process. The manual labeling was not performed to avoid potential upward biases in the impact of image-based covariates on the prediction.

Next to structural attributes represented by BHC, the process of data-scraping also included two remaining categories of predictors: location and neighborhood. Next to the usage of Google Maps API, a set of variables were collected: the distance from a house street to Central Business District (CBD) and the time it takes to travel the distance by walking, biking, and public transport. Similarly, the geographical

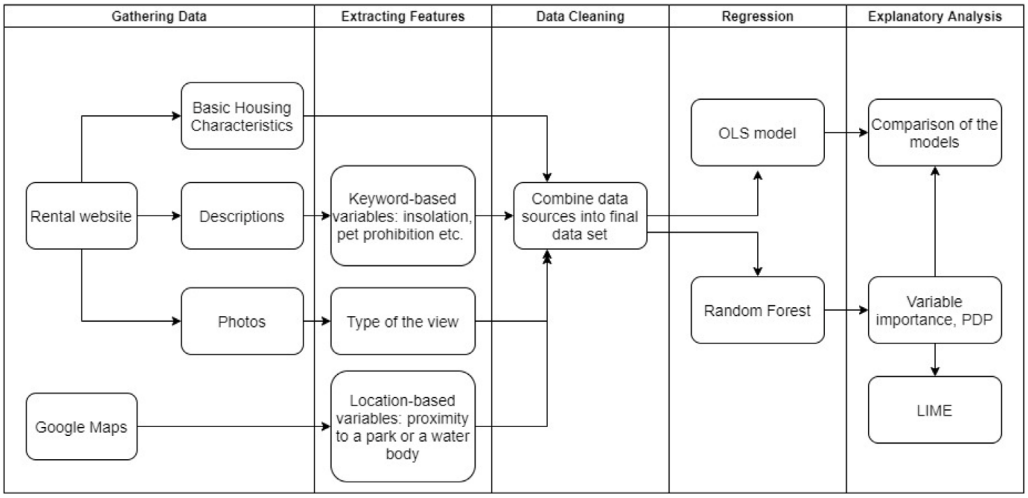

Fig. 1. Workflow diagram of the study.

coordinates of each street appearing in the initial data set in the form of longitude and latitude were gathered. Lastly, the satellite photos of each street's neighborhood at various zooms were saved.

The process of gathering data was fully automated with the usage of R. Packages that allowed scrapping the data based on HTML and XML code of the website were `rvest`, `selectr`, and `xml2`. Consequently, over 2000 real estate rental offers were found and crawled from the beginning of April till the end of June 2020. After removing the duplicates and fake advertisements, the final data set consisted of 1844 unique observations.

4. Methodology and application

The current study was conducted in three main steps: (1) extraction of covariates from complex data sources such as images and text and transforming them into low-dimensional tabular data; (2) modeling and predicting the rental costs of properties in Rotterdam based on conventional and big data; (3) interpretation of the created predictive models. Fig. 1 provides a workflow diagram of all steps that are discussed in detail in the next subsections.

4.1. Feature extraction

4.1.1. Image recognition: Methodology

Among the main challenges of the study was to extract meaningful and useful predictors from images and texts. The majority of recent studies agreed on the fact that the best-performing image recognition models were based on convolutional neural networks (CNN) (Khan et al., 2020; Simonyan & Zisserman, 2014; Szegedy et al., 2015).

CNNs are constructed similarly to regular neural networks. The main difference lies in the architecture of the layers. For a better understanding of CNN, we provide the example of a model's dataflow in Fig. 2, with steps which we describe below and enumerate from 1 to 9. In image recognition models, the input usually takes the form of A × B × 3 matrices, where A stands for the picture width, B for the picture height, and 3 represents color channel values (RGB). To analyze such complex input types, CNN introduces three-dimensional hidden layers, whose size is reduced in subsequent processing. The first layer in the model is called the convolutional layer. In this part, the model analyzes one part of the picture, multiplies its values by a pre-defined smaller matrix known as a filter or kernel (1), and then convolves to a new part until the whole picture is scanned (2). The size of the part of a picture scanned at one moment is equal to the kernel size, which is one of the parameters that may be tuned. However, the sizes of 3 × 3 × 3 and 5 × 5 × 3 pixels are the most popular (Simonyan & Zisserman, 2014; Szegedy et al., 2015).

The output of numerous multiplications is stored in a two-dimensional matrix known as an activation map (3). The calculated activation map is subsequently passed to an activation function (4). The goal of the function is to activate and use in subsequent processing only those parts of the picture that generate patterns we try to recognize with the model. In the case of the CNN, a ReLU function is usually used. ReLU is defined as $f(x) = \max(0, x)$, where x is an input value. By applying this function element-wise, more complex patterns can be captured by the model; see Li et al. (2019) for more details.

The number of filters used in the CNN is the parameter of their tuning. As filters may be seen as feature detectors, increasing their number often leads to model's better accuracy. On the other hand, a separate activation map has to be calculated for each added filter, which drastically increases the computational power needed for training a model. Similarly, the number of convolutional layers used in the CNN may be adjusted. Generally, the lower convolutional layers are responsible for recognizing simple features. The deeper convolutional layers use these features as an input in detecting more sophisticated characteristics of an image (Zhu et al., 2020).

In the next step, the processed activation maps are stacked along the depth dimension (5). The created three-dimensional matrix is used as an input for the next layer used in CNN: a pooling layer, whose main function is reducing the size of convolved features and, consequently, the number of parameters and computational time. This goal is accomplished by analyzing the input matrix by parts. In the case of the most popular approach known as max pooling, only the maximum value is returned for each submatrix, whose size depends on the pre-defined kernel (6). Then, the process is repeated until the whole image is traversed and the original matrix is transformed into a less-dimensional one.

Next, the matrix is transferred to the fully connected layer (also known as a dense layer). At this point, the data are transformed into a column vector, which is subsequently used as an input for a regular feed-forward neural net (7). Each of the input values, also known as a neuron, (8) is then multiplied by a weight (9) and becomes an addend of neurons in the next layer.

Contrary to ReLU activation function used in three dimensional layers, softmax function is defined as:

$$\delta(z_l) = \frac{e^{z_l}}{\sum_{j=1}^{K} e^{z_j}}, \tag{1}$$

where z_l, $l = 1, \cdot, K$, values are the elements of the input vector and K is the number of classes in the classifier. This is a commonly used function

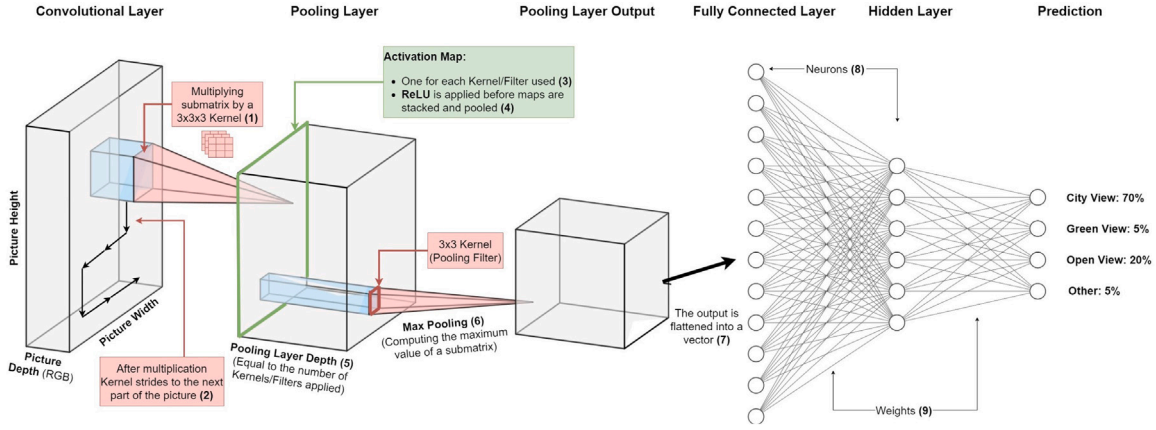

Fig. 2. The example of CNN architecture.

for the output layer, whose main purpose is to provide probabilities for each of the output classes (Nwankpa et al., 2020). In the current study, classes are related to four different types of views from a window: city view, green area view, open view, and others.

After producing predictions, CNN calculates the prediction error through the usage of a loss function. Among the most popular approaches used in deep neural nets is the cross-entropy function, due to its robustness and high accuracy (Zhang & Sabuncu, 2018). In our case, the softmax function is used to determine the output class probabilities. Cross entropy is defined as:

$$CE = -\sum_{l=1}^{K} t_l log(\delta(z_l)), \tag{2}$$

where t_l and $\delta(z_l)$, $l = 1, \cdot, K$, are the target values and the probabilities calculated by softmax function for each class. Lastly, the model iteratively updates the weights and filters values in order to minimize the prediction error. The provided description of CNN is an example of a simplified model architecture. The number of layers and methods used, as well as their order, is usually more complex for real-life modes. For a more detailed explanation of the model's technical attributes and architecture, see Yamashita et al. (2018).

4.1.2. Image recognition: Application for Rotterdam housing market

The majority of the housing pricing research that included image recognition methods focused on investigating the interior of real estate (Poursaeed et al., 2018; You et al., 2017). This study considered the external factors that may impact the renting price, especially the type of view from the property.

The image recognition process for different types of views was divided into two sequential CNN models. The first model aimed to filter the property images and classify them as outside or inside. The goal of the second model was to analyze the outside photos and classify them into four categories:

- **View on the city:** category featuring photos with a relatively unbroken view of the city panorama;
- **Green view:** category featuring images with a view of a park, a forest, or other green areas;
- **Open view:** category featuring images with an unbroken view, e.g., photos presenting an open view of a neighborhood or a river;
- **Other:** category featuring all the other images.

Apart from the usage of property photos, the collected maps of neighborhoods were also analyzed. Following the conclusions drawn from the literature review and the specific nature of Rotterdam's urban layout, we decided to focus on two categories. For each of the categories, a separate image recognition model was created. As a result, the following dummy variables were created:

Table 2
Accuracy of the image recognition models.

Variable	Accuracy
Water body	94%
Park	86%
City view	90%
Green view	86%
Open view	76%

- **Parks:** category indicating whether a given property lies, in a direct neighborhood of a park or other green areas.
- **Water bodies:** category indicating whether a given property lies, in a direct neighborhood of a river or a large lake.

The accuracy of the feature extraction process, measured on the test set, is presented in Table 2.

4.1.3. Text sources

The initial analysis of property descriptions showed that most of them were written similarly. When combined with relatively small sample size, it was no surprise that the attempted sentiment analysis did not bring additional value to the research. The distribution of features based on natural language processing ended up being extremely skewed as the majority of descriptions were written using unnatural positive language. In contrast, the content of descriptions in terms of the mentioned house characteristics differed notably. Therefore, the text analysis part of the research was based on information extraction using keywords. The first step of the analysis was to identify the most commonly used words in rental offer descriptions.

Following Vijayarani et al. (2015), the pre-processing part included the transformation of all the descriptions into lowercase, applying stemming and removing stop words obtained from a pre-compiled list from R package `stopwords`. Subsequently, the bag-of-words model was applied. In the first step, the model learned vocabulary from all descriptions. In the second part, it counted the number of times each word appeared in each description, disregarding the order in which they appeared.

Unsurprisingly, most of the frequently appearing words were related to the variables mentioned on the rental website, such as the number of rooms or location. Nonetheless, several additional keywords, potentially carrying information about rental cost, were found, e.g., "pets" or "sunny". Each description was scanned for these keywords, leading to the creation of dummy variables, indicating whether a property is exposed to a given factor. Additionally, for each keyword, its surrounding was scanned for such phrases impacting the original meaning as "not", "no", etc. In cases when such phrases were found, the value of the dummy variable was adjusted accordingly. In the end, three variables were built based on text sources:

Table 3
Variables present in the initial data set.

Variable	Type	Description
Street	Character	Street
Price	Numeric	Price in euros
Living_area	Numeric	Living area in squared meters
Rooms	Numeric	Number of rooms
House_type	Factor (3 levels)	Is a property a house, a room or a flat?
Bedrooms	Numeric	Number of bedrooms
Bathrooms	Numeric	Number of bathrooms
Balcony	Factor (2 levels)	Does a property have a balcony?
Garden	Factor (2 levels)	Does a property have a garden?
Storage	Factor (2 levels)	Does a property have a storage?
Garage	Factor (2 levels)	Does a property have a garage?
Bath	Factor (2 levels)	Does a property have a bath?
Lift	Factor (2 levels)	Does a property have a lift?
Toilet	Factor (2 levels)	Does a property have a separate toilet?
Furnished	Factor (2 levels)	Is a property fully furnished?
Description	Character	Description of a property
Time_walking	Numeric	Travel time by walking in seconds
Time_biking	Numeric	Travel time by biking in seconds
Time_public	Numeric	Travel time by public transport in seconds
Longitude	Numeric	Street's longitude coordinate
Latitude	Numeric	Street's latitude coordinate

Table 4
Descriptive statistics of numeric basic housing characteristics.

Variable	Min.	Mean	Median	Max	St. dev.
Price	295.00	1347.00	1295.00	4995.00	542.38
Living_Area	6.00	75.61	73.00	935.00	41.29
Rooms	1.00	2.72	3.00	11.00	1.13
Bedrooms	1.00	1.20	1.00	5.00	0.81
Time_walking	128.00	2275.00	1754.00	9037.00	1512.49
Time_biking	32.00	704.70	540.00	2530.00	458.37
Time_public	109.00	1085.70	996.50	3085.00	530.86
Longitude	4.41	4.48	4.48	4.58	0.03
Latitude	51.87	51.92	51.92	51.98	0.02

Table 5
Descriptive statistics of binary basic housing characteristics.

Variable	Not present	Present
Balcony	1152	692
Garden	1638	206
Storage	1382	462
Garage	1759	85
Bath	1371	473
Lift	1456	388
Toilet	1053	791
Furnished	562	1282
Park	1359	485
View_on_the_city	1617	227
Enjoyable_view	1204	640
Green_view	1589	255
Water_body	892	952
Pets_not_allowed	1696	148
Income	1693	151
Insolation	1171	673

- **Pets:** category indicating whether pets are allowed in the property;
- **Insolation:** category based on words indicating that a property is exposed to sunlight, e.g., bright, sunny;
- **Income:** category indicating whether a financial requirement is mentioned in the description based on the words "guarantor" and "income".

4.1.4. Feature selection

The feature extraction process combined with the initial data set directly scrapped from the rental websites and Google API resulted in the final dataset consisting of 21 variables. Their description and descriptive statistics are presented in Tables 3, 4, and 5.

Apart from the described variables, the distance to arterial roads was measured based on Google Maps images and used as a proxy for traffic and noise pollution. Services mentioned in properties' descriptions were used as a proxy for the neighborhood's utility . However, the application of these factors did not provide insightful results. This can be explained by considerable diversity in the quality of descriptions.

Previous studies identified several other external aspects impacting the value of the real estate, e.g., air pollution (Smith & Huang, 1993), proximity to hazardous industrial facilities (Grislain-Letrémy & Katossky, 2014), or neighborhood characteristics such as crime rate (Dubin & Goodman, 1982) and the quality of schools (Clark & Herrin, 2000). We decided not to include these factors in the analysis.

First, we find Rotterdam as a rather homogeneous region in terms of most of the mentioned characteristics. As a windy city located near the North Sea, Rotterdam does not suffer severely from air pollution (The World Air Quality Project, 2021). Areas with slightly better air quality are parks and water bodies, which are already accounted for by other predictors in the dataset. Further, the majority of Rotterdam heavy industry is located nearby the port, far away from the residential areas.

Second, including aspects specific to Rotterdam would diminish the flexibility of the research framework, which is applicable to most cities in the world in the presented setting. Moreover, according to the hedonic pricing theory, the customers' willingness to pay for a particular good is based only on the attributes they may perceive. Therefore, aspects such as air quality or environmental risks seemed less likely to affect the rental price, compared to more easily noticeable features, such as the view from the window.

Third, the detailed neighborhood characteristics could suffer from endogeneity bias. As an example, many papers find that proximity to good schools (Clark & Herrin, 2000) and shopping malls (Wilhelmsson & Long, 2020) positively impact the housing prices. However, such services are also more likely to be opened and operate in rich neighborhoods. This problem of symbiotic relation between dependent and independent variables, known as simultaneity, may cause biased inference of the predictive model. On the other hand, completely omitting relevant predictors that impact the predicted variable also causes an endogenous issue known as omitted-variable bias.

If addressed properly, the variables omitted in our research could lead to augmented interpretability and better predictive performance of the created models. However, rental market prices are mostly driven by housing characteristics and the distance to the city center. In that context, we expect the added value of incorporating a few missing variables to be rather negligible. Therefore, employing full statistical endogeneity analysis, which would significantly enlarge an already wide methodology part of the paper, was not a priority in the current study. For more details regarding causal inference and controlling for endogeneity by employing instrumental variables, see Bascle (2008) and Chernozhukov et al. (2018).

Instead, we decided to follow a simplified proxy variable approach. By employing geographical coordinates in a nonlinear machine learning model, the area of Rotterdam was divided into multiple regions. Therefore, the omitted neighborhood characteristics were indirectly mirrored in the locational parts of the models described in the following section. As for the rest of the variables used in the study, we found their relation to rental prices to be rather one-sided and not simultaneous, e.g., increased living area is likely to increase the rent, although increasing the rent does not increase the living area.

It is important to note that, in the current study, we were unable to employ demand-sided attributes such as buyer characteristics. We found the process of gathering tenants-oriented data without cooperating with real estate brokers a futile task. In the hedonic literature, when individual-level data were not available, a common approach was to use demographic data published by municipalities in public reports and censuses (Day et al., 2003). Unfortunately, we were not able to reach district-level demographic data for Rotterdam.

Even though buyers' characteristics played a significant role in the original hedonic theory of Rosen (1974), in practice, their impact on the inference of first-stage hedonic models was not crucial. In their study, Kim (1992) built a series of OLS-based hedonic models aiming at predicting monthly rental costs of properties in Sacramento, California. The first model consisted solely of housing characteristics. The second model used only demand-based characteristics such as income, age, and commuting time of tenants. The third model combined both housing and demand characteristics. Interestingly, the comparison of housing characteristics coefficients between the first and the third model showed minimal differences. The clearest contrast between both models lied in the constant value which was significantly lower in the third model. This, however, was compensated by demand-based coefficients absent in the first model. As such, it may be concluded that incorporating demand-based characteristics into the model did increase its overall interpretability but barely changed the inference of housing characteristics.

The lack of demand-based attributes in the current study may be seen as a drawback in the context of hedonic pricing. However, demographic characteristics, similar as in the case of other omitted variables discussed above, were indirectly reflected in the locational parts of the created models. This, combined with the empirical results of Kim (1992), prompts us to conclude that the missing demand attributes did not interfere with the overall analysis and results presented in the paper.

4.2. Predictive modeling

4.2.1. Linear regression

Most studies conducted in hedonic pricing were based on a parsimonious linear regression model or its variations (Owusu-Ansah, 2013). The current study also followed this approach to build an initial predictive model that served as a point of reference in the evaluation of the accuracy and interpretability of more advanced machine learning models.

Before estimating an OLS model, an initial analysis of the available data was performed, and some multicollinearity issues were observed. To diminish them, the variable describing the number of bedrooms was dropped as it was strongly correlated with the number of rooms as well as living area. Another adjustment that had to be made was the choice of the locational predictor. Out of three variables extracted from Google Maps API, biking time was used due to biking popularity in the city.

Although the longitude and latitude were not linearly correlated with the biking time, it was reasonable to assume that these variables were strongly related in a nonlinear way. In linear conditions, geographical coordinates were only able to indicate a part of the city (e.g., west or east in the case of longitude) with higher rental costs. Therefore, measuring the location of property through biking time to CBD was more proper for the traditional hedonic model. Nonetheless, for completeness, both locational approaches were used for prediction.

Several linear models were estimated and compared in terms of predictive accuracy and reliability. For all the models, the same train and test sets were used, where the former consisted of 70% and the latter of 30% of observations. Although 70% / 30% splitting is nonstandard in the OLS setting, it was employed to perform a fair out-of-sample comparison to the ML technique discussed below. The chosen train:test ratio was based on the empirical results of Gholamy et al. (2018). Following Chai and Draxler (2014), the models were compared in terms of root mean square error (RMSE)

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}, \tag{3}$$

calculated on the test set. In (3), $\hat{y}_i$ is a predicted value for i_{th} observation, y_i is an observed value of an i_{th} observation, and n is the number of observations.

Table 6
The coefficients of the final linear model and corresponding p-values for t-test.

Coefficients	Estimate	Pr(>|t|)	Significance
(Intercept)	433.37	<0.001	***
Living_Area	6.09	<0.001	***
Rooms	89.80	<0.001	***
House_Type House	26.52	0.589	
House_Type Room	−192.16	<0.001	***
Bathrooms	238.77	<0.001	***
Balcony Present	−5.44	0.813	
Garden Present	111.29	0.002	**
Storage Present	−33.27	0.234	
Garage Present	256.53	<0.001	***
Bath Present	64.62	0.011	*
Lift Present	38.26	0.188	
Toilet Present	−9.34	0.678	
Furnished 1	87.34	<0.001	***
View_on_the_city 1	115.21	<0.001	***
Water_body 1	46.30	0.025	*
Income 1	−165.19	<0.001	***
Insolation 1	117.83	<0.001	***
Time_biking	−0.26	<0.001	***

The variables were found significant at different levels:
*p < 0.05.
**p < 0.01.
***p < 0.001.

Two linear BHC-based models were estimated and compared to select the best locational proxy. Both models contained *Living_Area, Rooms, House_Type, Bathrooms, Balcony, Garden, Storage, Garage, Bath, Toilet, Lift* and *Furnished* as covariates. Model (4) used biking time as a locational proxy, while Model (5) included geographical coordinates.

$$\begin{aligned} y_i^{OLS^A} &= \alpha^A + \beta_{LA}^A * LivingArea_i + \beta_R^A * Rooms_i + \cdots + \beta_F^A * Furnished_i \\ &\quad + \beta_{TB}^A * Time\ biking_i + \epsilon_i^A \end{aligned} \tag{4}$$

$$\begin{aligned} y_i^{OLS^B} &= \alpha^B + \beta_{LA}^B * LivingArea_i + \cdots + \beta_F^B * Furnished_i \\ &\quad + \beta_{LON}^B * Longitude_i + \beta_{LAT}^B * Latitude_i + \epsilon_i^B \end{aligned} \tag{5}$$

In (4), (5) and (6), α is an intercept, and $\boldsymbol{\beta} = (\beta_{LA}, \beta_R, \ldots, \beta_{LAT})$ is the set of coefficients to be estimated y, and ϵ, are the price and the error term for the i_{th} observation, $i = 1, \ldots, n$.

The comparison of both models proved a slight superiority of biking time over geographical coordinates in terms of RMSE and the proportion of the variance explained (R^2) in linear conditions. The difference in the accuracy of the models was further confirmed by the conducted paired Wilcoxon test on models' residuals. It resulted in a p-value of 1.496^{-3}, indicating that the median prediction error of Model (4) was significantly lower than the one of Model (5) Consequently, Model (4) was enriched with images and text data, leading to the creation of the final Model (6)

$$\begin{aligned} y_i^{OLS^C} &= \alpha^C + \beta_{LA}^C * LivingArea_i + \cdots + \beta_I^C * Insolation_i \\ &\quad + \beta_{TB}^A * Time\ biking_i + \epsilon_i^C \end{aligned} \tag{6}$$

However, many variables originating from rich data sources were found insignificant and were omitted to avoid over-specification of the model; see Butler (1982) and Mok et al. (1995) for more details. The p-value of 1.447^{-5} for the paired Wilcoxon test showed the significant advantage of the final Model (6) over Model (4) in terms of predictive accuracy. The coefficients of the Model (6) are presented in Table 6 below.

Out of BHC, living area, the number of rooms and bathrooms, the presence of a garden, garage, and bath, as well as the furnishing turned out to be significant positive predictors of a rental cost prediction. These findings combined with the fact that the distance to CBD lowers the rental price match the previous research provided in Table 1. The

Fig. 3. The example view on Rotterdam panorama.

findings got more peculiar in terms of features extracted from images and descriptions. Income requirements for tenants were related to the average drop in a rent price of 165 euros. While the direction of causality may be an object of discussion, we found it as an interesting finding, which, to the best of our knowledge, had not been studied previously. The pet permission turned out to be an insignificant factor. On the contrary, both the proximity of a water body and insolation positively impacted the real estate's rental price. The same was true for the view out of the window, although only one featuring a city panorama.

Numerous studies indicated that customers were willing to spend up to 9.2% more money on properties with good views such as lakes or green areas (Gillard, 1981; Mok et al., 1995). This study did not confirm the significant impact of these types of views on the rental costs of properties. Similarly, the proximity to parks did not seem to affect the renting price. Conversely, Bishop and Lange (2005) and Wolf (2007) reported that parks located in the distance of up to 400 m, depending on their characteristics, increased on average the price of properties by 10% or even 20%. Further, as opposed to Ming and Hian (2005), who found obstructed view to depress property value by 8%, we did not find any indications of such relation.

Some of these findings might be specific to the area of Rotterdam due to its distinctive nature. As seen in Fig. 5, parks and green areas are equally distributed throughout Rotterdam and are easily accessible for the majority of citizens. On the other hand, water bodies are specific to the central (Meuse river) and northern parts (two lakes) of the city. Therefore, it is rather unsurprising that the proximity to these recreational areas was related to the increased rental costs. The fact that the only type of view significantly impacting rental prices was a city view might be connected to the remarkable beauty of Rotterdam panorama, presented in Fig. 3.

It is important to note that the current study employed OLS as a prediction technique. While the linearity assumption enabled easy interpretation, the OLS model suffered from several issues related to its assumptions. Firstly, the residuals of the model were not normally distributed. Secondly, the problem with heteroscedasticity was found. To overcome these problems, semi-log and Box–Cox transformations were applied according to (Butler, 1982). Nevertheless, this procedure did not solve the issue while the accuracy of the model dropped. Therefore, the estimated coefficients and the standard errors of the created OLS model were not fully reliable, which consequently diminished its main advantage over more advanced machine learning models: ease and reliability of interpretation and availability of the testing procedure.

4.2.2. Random forest

Due to the limitations of linear regression, we decided to take advantage of a nonparametric approach. Previous research conducted in the area of the housing market showed better performance of decision tree models, especially one of its ensemble variations, random forest, over traditional hedonic regression (Hong et al., 2020; Neloy et al., 2019), as well as the spatial approach (Credit, 2021; Hengl et al., 2018).

As a nonparametric method, the decision-tree-based models do not require any assumptions regarding the distribution of the data and can capture more complex patterns than their linear counterparts. Moreover, as opposed to spatial models, they are able to capture interactions between covariates naturally and can be easily applied to a combination of numerical and categorical variables without creating huge sets of dummy variables. Both features are important for hedonic pricing models. First, in the original hedonic pricing theory, the importance of interactions between goods and their quantity is emphasized (Rosen, 1974). Further, typical real estate data contain many categorical variables.

The decision tree is a nonlinear model consisting of sequential conditional statements that separate the data into smaller subsets known as nodes. Geometrically, this corresponds to separating the observations by hyperplanes parallel to one axis of the feature space (Ho, 1995). At each step, the algorithm chooses the split that leads to the greatest drop in the applied loss function. The type of applied loss function depends mostly on the type of the dependent variable. The standard choice for regression problem is the residual sum of squares (RSS), defined as

$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{7}$$

where $\hat{y}_i$ is a predicted value for i_{th} observation, y_i is an observed value of an i_{th} observation, and n is the number of observations. The splitting process is repeated for newly created nodes until a stopping rule is met, e.g. when the number of observations in newly created nodes would fall below a specified threshold.

In Fig. 4, we provide an example of the simplified decision tree model using geographical coordinates and living area to predict rental price, where final nodes were specified to consist of at least 5% of the total number of observations.

The common weakness of decision trees is their tendency to grow deep and thus overfit the data. One of the approaches that help in dealing with this problem is the usage of bootstrap-based ensemble method such as bagging, defined as:

$$f(x) = \frac{1}{N} \sum_{n=1}^{N} f_n(X_n^B), \; n = 1, \ldots, N \tag{8}$$

Bootstrap method simulates N new data sets X_n^B by randomly drawing observations with replacement from the original data set. Subsequently, for each n_{th} bootstrapped sample, a decision tree model $f_n(\cdot)$ is built. The prediction of the bagged model is an average predicted value among all N regression trees, which reduces the overall variation

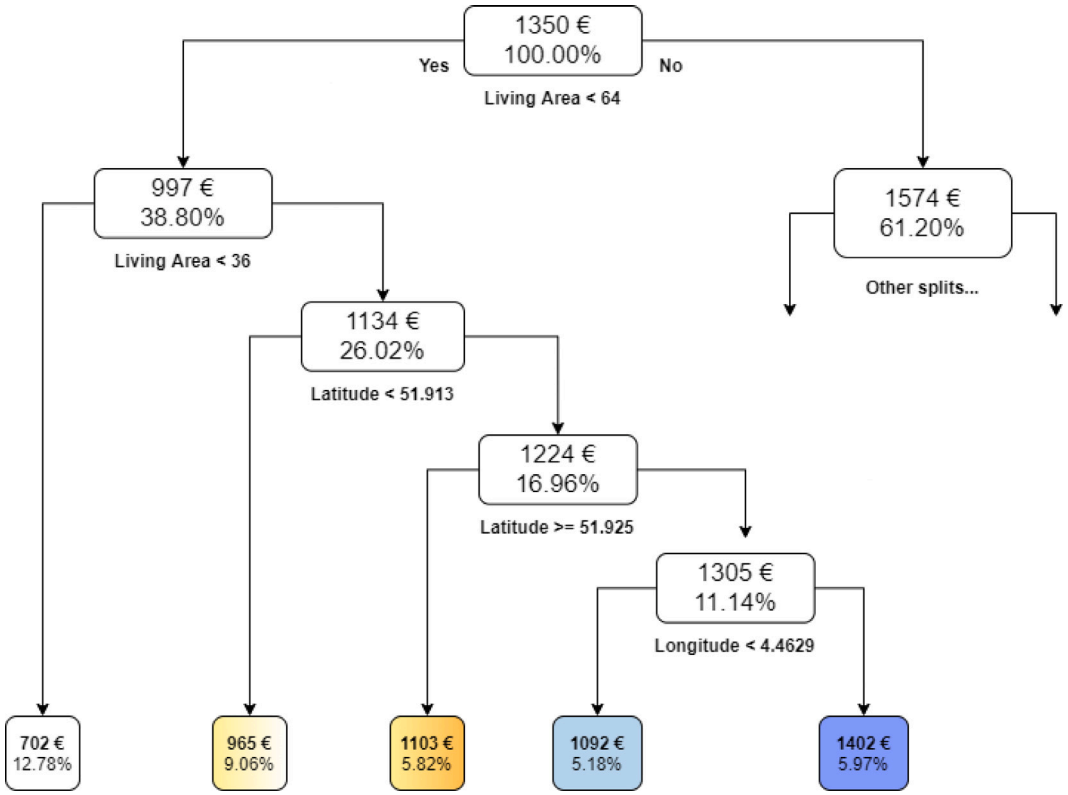

Fig. 4. The example of decision tree using geographical coordinates and living area to predict rental price.

in comparison to a single tree and results in better out-of-sample performance. For more details, we suggest (James et al., 2013).

One of the weaknesses of bagging is the fact that bootstrapped-based trees are highly correlated by construction. Random forest, which is a variation of bagging, addresses this problem by allowing only a random subset of original variables to be used in a single bootstrapped tree. This approach not only decorrelates the trees but also prevents numerical variables from completely dominating over categorical ones. The number of predictors used in each tree is a parameter that may be tuned via grid search with cross-validation.

Random forest models were built similar to the OLS models (4), (5) and (6):

$$y_i^{RF^A} = f^{RF^A}(LivingArea_i, \dots, Furnished_i, Time\ biking_i) + \epsilon_i^{RF^A} \quad (9)$$

$$y_i^{RF^B} = f^{RF^B}(LivingArea_i, \dots, Furnished_i, Longitude_i, Latitude_i) + \epsilon_i^{RF^B} \quad (10)$$

$$y_i^{RF^C} = f^{RF^C}(LivingArea_i, \dots, Insolation_i, Longitude_i, Latitude_i) + \epsilon_i^{RF^C} \quad (11)$$

The first two models aimed to distinguish which locational approach is more suitable. Next to BHC, Model (9) employed biking time to CBD, while Model (10) used geographical coordinates. In (9), (10), and (11), $f(\cdot)$ stands for the function estimated by random forest, y and ϵ are the price and the error term for the i_{th} observation, where $i = 1, \dots, n$.

The latter approach was shown to be superior to its biking counterpart. Both RMSE and R^2 scores for Model (10) improved when compared with Model (9). The conducted paired Wilcoxon test on models' residuals with the p-value of 1.482^{-2} further confirmed the better performance of Model (10). This difference may be explained by the fact that by using such a nonlinear model as random forest, it became possible to fully use the potential laying in geographical coordinates. Not only the mixture of both variables contained all the information about the distance to CBD, but also served as a proxy for unobserved characteristics such as criminality level or noise pollution of the neighborhood. Fig. 5 provides a simplified visualization of this phenomenon. Splitting the observations even by the limited number of lines parallel to the axes of longitude and latitude led to the creation of heterogeneous regions of properties in Rotterdam.

Therefore, the final Model (11) consisted of BHC, geographical coordinates, and covariates extracted from images and text, which turned out significant in the linear model. Similar to Models (9) and (10), the number of predictors used in each tree was set to 10 according to the results of a tuning process based on a grid search with cross-validation. Model (11) turned out to be the most accurate in terms of RMSE and the proportion of the variance explained (R^2). This is further confirmed by the conducted nonparametric Wilcoxon test between the predictions of the random forest model that used unstructured data sources, and those that did not. The test resulted in a p-value of 3.986^{-2}, implying that there was a significant difference between the accuracy of both models. Lastly, in comparison to the final OLS Model (6), Model (11) also proved superiority in terms of predictive accuracy according to Wilcoxon test, resulting in a p-value of 3.706^{-11}. The performance of all models is presented in Table 7.

4.3. Explainable AI

4.3.1. Explainable AI: Methodology

The results presented in the previous section provided us with two main conclusions. Firstly, including unstructured data in the research, even in a relatively simple form, allowed us to significantly increase the performance of the predictive models. Secondly, applying advanced machine learning techniques also significantly increased the accuracy when compared to the traditional OLS regression. However, in the context of hedonic pricing, superior forecasting ability may be seen as a success only after reaching a certain level of interpretability. Therefore,

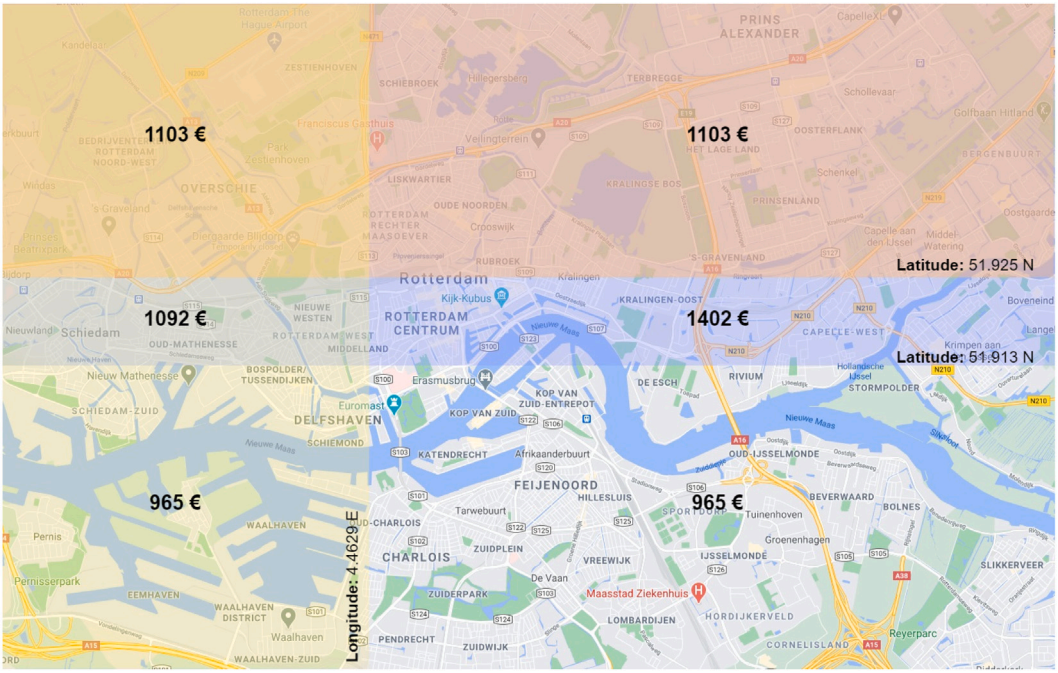

Fig. 5. Mapped results of the decision tree presented in Fig. 4. The regressed rental prices of properties with living area between 36 and 64 squared meters visibly differ depending on the location.

Table 7
The performance of OLS and random forest models.

Model	R^2	RMSE
(5) OLS: BHC + Longitude | Latitude	0.56	329
(4) OLS: BHC + Time_biking	0.59	321
(6) OLS: BHC + Time_biking + Significant image and text variables	0.61	309
(9) RF: BHC + Time_biking	0.70	258
(10) RF: BHC + Longitude | Latitude	0.71	246
(11) RF: BHC + Longitude | Latitude + Significant image and text variables	0.74	240

in contrast to the previous research conducted in the area, e.g., Hong et al. (2020), Law et al. (2019), and Neloy et al. (2019), this study aimed to tackle the problem of interpretability by applying explainable AI methods. To uncover the functioning of black-box models, three different model agnostic methods were applied: variable importance, partial dependence analysis, and local interpretable model-agnostic explanations (LIME).

Variable importance analysis is based on the mean decrease in accuracy. The idea of the method is to permute one predictor in order to decouple its relation with the dependent variable. Subsequently, a new model with the permuted predictor is fitted and its accuracy is measured. The higher reduction of accuracy corresponds to the higher importance of the predictor.

In contrast to variable importance, partial dependence analysis is a graphical technique. It shows the marginal effect of a feature on the predicted outcome of a machine learning model (Friedman, 2001). In practice, the partial dependence function is estimated by averaging values observed in the training set. It works by marginalizing the machine learning model output over the distribution of all the features except the feature of interest (Molnar, 2019). This way, the function estimates the relationship between the feature of interest and the predicted outcome of the machine learning model. It allows approximating the type of relation between the independent and dependent variables, e.g., linear, monotonic, or complex, which was a particular point of interest of this research. In the case of categorical predictors, the partial dependence is calculated in two steps. First, all observations are forced to have the same category. Second, the predictions for these modified observations are computed and averaged. The process is then repeated for each level of the categorical predictor. Consequently, the average marginal effect on the prediction for each level is obtained.

However, both methods function on the global scale of a model, meaning that they do not allow understanding the reasoning of an algorithm for a single prediction. LIME approach, first introduced by Ribeiro et al. (2016), aims to tackle this issue. LIME is rooted in the assumption that even the most complex model is linear on a local scale. This assumption implies the key idea of LIME: if two observations possess very similar covariates, they should have similar responses in a machine learning model. Therefore, if multiple similar observations behave similarly in a black-box model, it is possible to fit a local prediction (surrogate) model on their basis. A surrogate model is aimed at mimicking and consequently explaining the local behavior of the original model.

In the first step of LIME, the observations of interest are permuted N times. Then, the prediction for each permutation is obtained by applying the original black-box model. Subsequently, the distances between the original observation and the prediction scores for permuted observations are calculated based on the chosen distance measure. The distances are then converted to similarity scores using an exponential smoothing kernel of a width equal to 0.75 times the square root of the number of features (Ribeiro et al., 2016).

In the next step, a surrogate model is fitted to the permuted data. The surrogate model may take numerous forms, e.g., OLS regression or decision tree. The only requirement for the chosen type of model is its interpretability. While training the local explanatory model, the outcome is weighted for a permuted observation by its similarity to the original observation. Finally, the feature weights of the surrogate model may be extracted and used as a proxy for the complex model's

Table 8
The comparison of variables' impact on the prediction between the OLS and random forest models.

Variable	OLS coefficient	RF partial dependence
House_Type room	−192.16	−281.73
Garden	111.29	9.53
Garage	256.53	104.96
Bath	64.62	37.59
Furnished	87.34	57.12
View_on_the_city	115.21	62.56
Water_body	46.30	10.88
Income	−165.19	−15.44
Insolation	117.82	51.61

local behavior. The type of obtained feature weights depends on the applied surrogate model, e.g., in case of OLS regression, they take the form of regression coefficients. As a result, LIME allows to analyze the regression model on an observation level (Alvarez-Melis & Jaakkola, 2018; Molnar, 2019). This aspect may help uncover different impacts of the same predictor on observations, potentially leading to a better understanding of the hedonic prices of an analyzed good.

4.3.2. Explainable AI: Application

The explanatory analysis of the final random forest model started with measuring variable importance based on the mean decrease in accuracy. The analysis presented in Fig. A0 indicated that the model's accuracy was mostly based on the living area's size, the localization of the real estate, and the number of rooms. These findings concurred well with the simplified decision tree model presented in 4, where the model based the splits on the same covariates. These results mainly coincided with the conclusions drawn from the OLS model. However, after considering the magnitude of the biking time coefficient and the average value of this variable, we concluded that it impacted linear model predictions to a smaller degree than the geographical coordinates in random forests.

To get an insight into the form of nonparametric function produced by the random forest model, partial dependence functions were calculated. Table 8 presents the comparison between the values obtained with partial dependence functions and the coefficients of the final linear model. The latter may be directly interpreted as the partial dependence scores of the OLS model. When applied to a linear model, the partial dependence always shows a linear relationship between the analyzed variable and a prediction, with a magnitude equal to the predictor's coefficient.

The results showed substantial differences between the models. While both models agreed on the sign of coefficients, there were notable differences in their magnitude. For the majority of predictors, especially for *Garden, Garage,* and *Income*, their impact on the prediction was much lower in the case of a random forest model. The only variable which the random forest model evaluated as more influential than the OLS regression was the *House_Type Room*. Apart from the categorical variables presented in Table 8, the partial dependence was also applied to the numerical variables. Probably the most interesting conclusion was drawn from the analysis of the living area presented in Fig. 6. In the OLS regression, it was only possible to derive the average value of 6.09 euros for each additional squared meter of the property.

On the contrary, the dependence between both variables in the random forest was nonlinear. For the properties with a living area between 6 and 136 m, the average value of each additional squared meter was estimated at 7.26 euros. Surprisingly, this value dropped to 1.65 euros after reaching the threshold of 136 square meters, to eventually rise again to 5.54 euros after passing the threshold of 191 square meters. However, the average price of each square meter calculated based on the whole set was equal to 5.70 euros, which was close to the OLS-estimated coefficient.

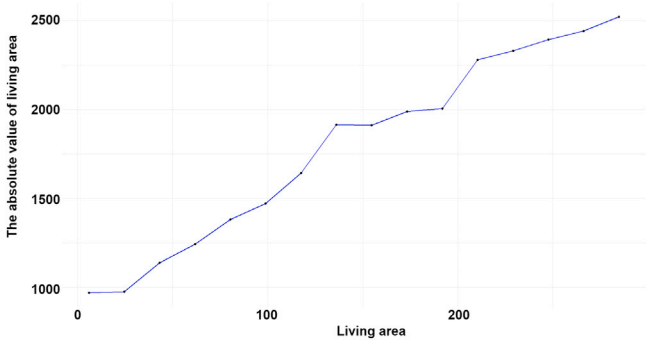

Fig. 6. The partial dependence plot of living area.

Similar to the living area, the number of rooms had a comparable value between the two models. Based on the partial dependence function, it was concluded that each additional room up to the threshold of seven rooms increased on average the rental price of a property by 87.16 euros. When compared to the OLS coefficient of 89.80, the difference was quite negligible and was much smaller than in the case of categorical variables.

This phenomenon suggests that the differences between both models might have been partially caused by the nature of random forest which, generally, favors the continuous variables in terms of their impact and importance. Nonetheless, if both living area and the number of rooms shared similar coefficients across the models, the drop in the importance of categorical variables in random forest had to be compensated by other predictors. Apart from methodology, the only difference between the models lay in the chosen locational approach. In our opinion, this factor partially explained the presented behavior. Longitude and latitude as proxy variables indirectly reflected numerous aspects that linear distance to CBD did not. Therefore, it may be suspected that the geographical coordinates affected the other predictors in the model to a further degree than the biking time in the OLS case.

There was no certainty in establishing the model describing the real-life values of properties' characteristics better. However, a much better performance of a random forest model in terms of accuracy and the variance explained together with the violated assumptions of the linear model prompted us to conclude that the obtained coefficients were more valid for prediction than those of the OLS regression. This conclusion implied that the OLS model may tend to overestimate the value of less important structural attributes, which were caused by its incapability to capture the locational aspects.

The explanatory analysis provided us with strong indications that housing attributes such as the size of the living area did not possess a constant marginal price for each additional unit and therefore did not have a linear relation to the predicted rental cost. Another aspect we were interested in is related to the original hedonic theory of Rosen (1974). Rosen (1974) argued that the marginal willingness to pay for the attribute of good changes for consumers depending on their (nonlinear) budget constraints and preferences. This statement further supported the decision to use a nonlinear machine learning model and encouraged checking how the hedonic price of household attributes varied among properties. To answer this question, LIME with a decision tree as a surrogate model, and Manhattan as a distance function were applied.

The explanatory model featured an extremely low R^2 value for some observations. After examining the distribution of variance explained among observations, we decided to analyze only the ones with the R^2 above 0.3. This arbitrary threshold was based on the trade-off between a reliable inference on observation and global levels. The results of the LIME analysis indicated that the hedonic price of an attribute was not constant. For example, the value of the city view presented in Fig. 7 for the flats/rooms with rent below 1000 euros stayed at the level of

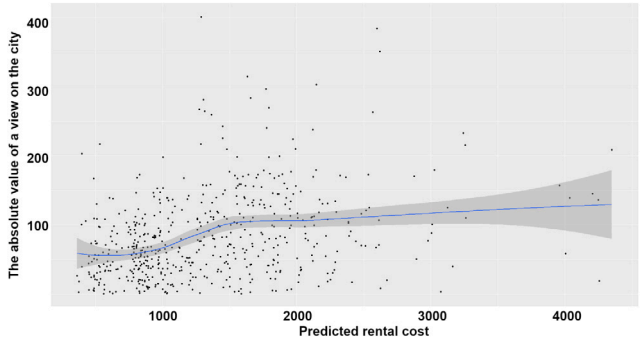

Fig. 7. View value versus predicted rental cost.

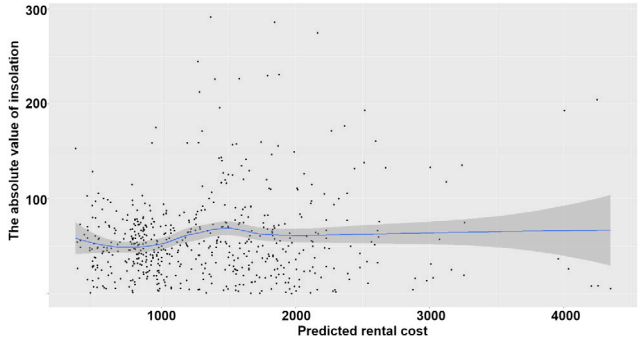

Fig. 8. Insolation value versus predicted rental cost.

around 60 euros, in order to rise above 100 euros for the properties that cost more than 1500 euros per month. A similar growing tendency was noted for the insolation of property in Fig. 8. In this case, however, the hedonic price of an attribute lowered after reaching the rental cost threshold of 1500 euros. Such behavior may be explained by the fact that detached houses and larger apartments started appearing after that threshold. As it may be expected, such properties are exposed to more than one geographical direction and are sunny by default.

The analysis shows that the value of the same attribute may not be constant and impacts different real estates to a different degree. The presented results provide clear trends of hedonic prices which can be directly applied to the real estate evaluation. However, the estimation of hedonic prices is only the first part of the two-stage procedure envisioned by Rosen (1974). In the second stage, the estimated hedonic prices may be used to uncover the customer preferences and an inverse demand function for the attribute of interest.

The observation-level results of LIME presented above can be used to create a function explaining the drivers of the estimated hedonic prices. Additionally, by incorporating tenants' information such as education, ethnicity, or age into the formula, it would become possible to shed some light on how the estimated prices of housing attributes depend on the demand characteristics. Formula (12) serves as a theoretical example of how the hedonic price of the city view may be explained by other predictors:

$$P_{view_i} = f(view_i, P_{property_i}, HC_i, TI_i) + \epsilon_i. \quad (12)$$

In (12), P_{view} is the hedonic price of the view for i_{th} observation estimated by LIME, $view$ is a dummy variable indicating the existence of the city view, $P_{property}$ is the total price of the i_{th} property, HC are other housing characteristics, and TC are tenants characteristics.

However, as pointed out by Brown and Rosen (1982), no information about customer preferences can be obtained directly in the second stage of the model as it would only reproduce the information provisioned by the first-stage estimation (Mei et al., 2017). This pitfall is most often addressed in the literature by exploiting the nonlinearity of the marginal hedonic price (Ekeland et al., 2004) or by extending the analysis to multiple markets (Bishop & Timmins, 2018; Palmquist, 1984). Additionally, to reach reliable estimates in (12), the clear endogeneity problem would have to be addressed, for example, by employing instrumental variables; see Bartik (1987) and Palmquist (1984) for more details.

Incorporating the second stage of the analysis envisioned by Rosen (1974) would require a wide methodological toolkit, significantly enlarging an already wide scope of the paper. Further, as mentioned in Section 4.1.4, we were unable to gather data regarding tenants and detailed demographics of Rotterdam districts. Consequently, this would diminish the potential insights from the analysis we envisioned. However, we leave this short theoretical discussion in hope of encouraging researchers to include the aspect of consumer preferences by employing model agnostic methods such as LIME in their future studies.

5. Conclusions

For over 50 years, hedonic pricing models have been extensively researched. Despite their undeniable popularity, in most cases, the hedonic models have been based on simple linear regression methods employing structured conventional data. This study confirms that such an approach can be successfully improved without significant drawbacks through the usage of advanced machine learning techniques.

First, the methodology in extracting covariates from property images, satellite maps, and text descriptions and incorporating them into a hedonic pricing model was provided. Most previous research (Law et al., 2019; Zhang & Dong, 2018) focused on measuring the impact of adding singular feature types, e.g., google street images to the basic analysis. On the contrary, we built a more complete framework, combining multiple sources of data. With this approach, we attempted to mimic the information-gathering process of tenants before making the purchase decision.

Second, an accuracy-based comparison between the OLS and random forest models indicated the undeniable superiority of the advanced machine learning approach over the traditional hedonic regression. Employing complex data sources made it possible to increase the predictive accuracy even further. When compared to the OLS model, machine learning approach in the presented setting led to a drop of 25% in RMSE and an increase of 0.15 in R^2.

The majority of previous machine learning-driven studies for a housing market did not include interpretation and evaluation of covariates employed in the models (Abidoye & Chan, 2018; Hong et al., 2020; Law et al., 2019; Neloy et al., 2019; Zhang & Dong, 2018). In this study, the problem of interpretability of the black-box model was successfully addressed by explainable AI methods. The provided empirical results show that the more complex but at the same time more accurate ML methods may be as interpretable as traditional hedonic models. Additionally, the results revealed that the marginal prices of some housing attributes such as living area were not constant, which could not be captured by a conventional OLS approach. Such results provided empirical evidence for the original hedonic theory published by Rosen (1974), who argued that, generally, the nonlinearity between the price of goods and their inherent attributes is likely to happen. Further, it was observed that the value of a view on the city panorama differed between 60 and 100 euros, even though the quality of the views was on at similar level among all the properties. A clear relation between the total rental cost and view value was found. It showed that the average value of the view rose with the total rental cost of a property. On the contrary, insolation was evaluated higher for mid-ranged price properties than for cheap or expensive real estate.

The results of the study showed limited predictive ability for the distance to the central business district as a measure of locational

aspect. According to our findings, this commonly used approach (Chen & Hao, 2008; Gaolu, 2015) performed worse than the geographical coordinates in terms of accuracy and may have contributed to the overestimation of other predictors in the model. The evaluation of the living area by both models resulted in a similar average price of around 6 euros per one additional squared meter. However, the nonlinearities uncovered by the random forest model showed that this price level did not fully apply to large properties. The drop to 1.65 euros between 136 and 191 square meters implied the low utility of the living area in this interval. A similar conclusion was drawn for the number of rooms, which did not seem to provide extra utility after reaching the threshold of seven. The majority of covariates extracted from satellite images, aiming at accounting for aspects such as traffic, noise pollution, and proximity to parks, did not bring insightful results. However, the proximity to water bodies was found significant and related to a slight increase in rental prices. On the contrary, the income requirements for tenants, extracted from rental offer descriptions, were observed to be negatively correlated with the rental prices.

The shift from the traditional linear estimation of hedonic prices to the one presented in this study might be of great value for the academic as well as the business world. Applying the proposed framework in other research areas and settings is expected to show promising results in terms of the predictive accuracy and interpretability of created models. Apart from a more complex methodology, we do not find significant drawbacks of the presented machine learning approach when compared to traditional hedonic approach. Therefore, we hope that the study will inspire future studies to follow a similar direction. We find the proposed research framework and analysis beneficial to landlords, real estate brokers, and municipalities. Better estimation of the optimal real estate value, as well as related environmental and social factors, may lead to more accurate and conscious decision-making. The conclusions drawn from the analysis could also be used to create an application allowing future tenants to estimate a "fair" rental cost for a property of interest depending on its attributes.

However, this study was limited in several ways. First, the sample size was relatively small and could be increased in other settings. Due to the text analysis part, only offers containing information in English were included in the research. Additionally, the data were gathered only for three months during the first wave of COVID-19, which could have impacted the rental market prices. Second, we were unable to capture demand-based characteristics of tenants and consequently conduct a full, two-staged hedonic procedure, as envisioned by Rosen (1974). We leave the theoretical discussion on that topic to inspire future studies to take full advantage of the model-agnostic methods in the context of hedonic pricing. Third, the potential endogeneity in the data could be addressed better through methods made specifically to serve this purpose. Lastly, even though the employed random forest model performed well in the study, other modern ML models, such as geographical random forest (Georganos et al., 2019), may provide better results. Given the flexibility of our research framework, largely based on model-agnostic methods, the application of other modeling approaches should be straightforward.

Appendix

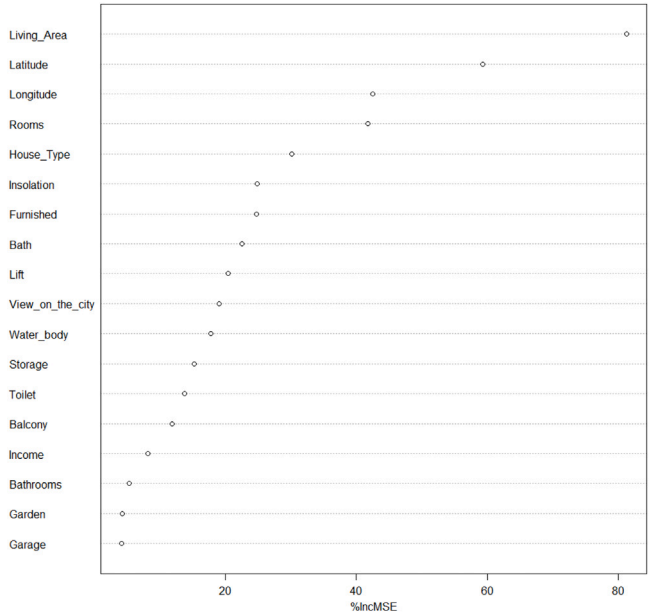

Fig. A.1. Variable importance of the final random forest model.

References

Abidoye, R., & Chan, A. (2018). Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal*.

Ahmed, E., & Moustafa, M. (2016). House price estimation from visual and textual features.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. In *Proceedings of the 2018 icml workshop on human interpretability in machine learning (whi 2018)*.

Anselin, L., & Lozano-Gracia, N. (2009). Spatial hedonic models. *2*, (pp. 1213–1250).

Ball, M. (1973). Recent empirical work of the determinants of relative house prices. *Urban Studies*, *10*, 213–233.

Bartik, T. (1987). The estimation of demand parameters in hedonic price models. *Journal of Political Economy*, *95*, 81–88.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, *6*, 285–327.

Basile, R., & Mínguez, R. (2018). Advances in spatial econometrics: Parametric vs. Semiparametric spatial autoregressive models. In P. Commendatore, I. Kubin, S. Bougheas, A. Kirman, M. Kopel, & G. I. Bischi (Eds.), *The economy as a complex spatial system* (pp. 81–106). Springer International Publishing.

Benson, E. D., Hansen, J. L., Schwartz, A. L., & Smersh, G. T. (1998). Pricing residential amenities: The value of a view. *Journal of Real Estate Finance and Economics*, *16*(1), 55–73.

Bishop, I., & Lange, E. (2005). *Visualization in landscape and environmental planning*. Taylor & Francis.

Bishop, K. C., & Timmins, C. (2018). Identification and estimation of hedonic models. *Journal of the Association of Environmental and Resource Economists*, *5*, 517–543.

Bracke, P. (2014). House prices and rents: Micro evidence from a matched dataset in central London. *Real Estate Economics*, *42*(2), 403–431.

Brown, J. N., & Rosen, H. S. (1982). On the estimation of structural hedonic price models. *Econometrica*, *50*(3), 765–768.

Butler, R. V. (1982). The specification of hedonic indexes for urban housing. *Land Economics*, *58*(1), 94–108.

Carroll, T. M., Clauretie, T. M., & Jensen, J. (1996). Living next to godliness: Residential property values and churches. *Journal of Real Estate Finance and Economics*, *12*(1), 319–330.

Cellmer, R., Cichulska, A., & Beł, M. (2020). Spatial analysis of housing prices and market activity with the geographically weighted regression. *International Journal of Geo-Information*, *9*, 380.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*, 1247–1250. http://dx.doi.org/10.5194/gmd-7-1247-2014.

Chen, J., & Hao, Q. (2008). The impacts of distance to CBD on housing prices in Shanghai: a hedonic analysis. *Journal of Chinese Economic and Business Studies*, *6*, 291–302. http://dx.doi.org/10.1080/14765280802283584.

Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., & Chi, T. (2020). Measuring impacts of urban environmental elements on housing prices based on multisource data—A case study of Shanghai, China.. *International Journal of Geo-Information*, *9*, 106.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.. *The Econometrics Journal*, *21*, 1–71.

Chin, T. L., & Chau, K. W. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing and its Applications*, *27*(2), 145–165.

Clark, D. E., & Herrin, W. E. (2000). The impact of public school attributes on home sale price in California. *Growth and Change*, *31*(1), 385–407.

Clark, S., & Lomax, N. (2019). Rent/price ratio for English housing sub-markets using matched sales and rental data. *Area*, *52*(1), 136–147.

Colby, B., & Wishart, S. (2003). Riparian areas generate property value premium for landowners. *Arizona Review*, *1*(1), 12–16.

Court, A. (1939). Hedonic price indexes with automotive examples. *The Dynamics of Automobile Demand*, 98–119.

Credit, K. (2021). Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in los angeles. *Geographical Analysis*.

Day, B., Bateman, I., & Lake, I. (2003). *What price peace? A comprehensive approach to the specification and estimation of hedonic housing price models*: *Cserge working paper edm*.

Dayley, A., & Logan, D. (2015). Organizations will need to tackle three challenges to curb unstructured data glut and neglect. https://www.gartner.com/en/documents/3077117/organizations-will-need-to-tackle-three-challenges-to-cu (Accessed: 2021-03-30).

de Koning, K., Filatova, T., & Bin, O. (2018). Improved methods for predicting property prices in hazard prone dynamic markets.. *Environmental and Resource Economics*, *69*, 247–263.

Des Rosiers, F., Lagana, A., Theriault, M., & Beaudoin, M. (1996). Shopping centres and house values: An empirical investigation. *Journal of Property Valuation & Investment*, *14*(4), 41–62.

Dubin, R. A., & Goodman, A. C. (1982). Valuation of education and crime neighborhood characteristics through hedonic housing prices. *Population and Environment*, *5*, 166–181.

Ekeland, I., Heckman, J. J., & Nesheim, L. P. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, *112*, 60–109.

Espey, M., & Lopez, H. (2000). The impact of airport noise and proximity on residential property values. *Growth and Change*, *31*, 408–419.

Forrest, D., Glen, J., & Ward, R. (1996). The impact of a light rail system on the structure of house prices. *Journal of Transport Economics and Policy*, *31*(4), 15–29.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression the analysis of spatially varying relationships*. Wiley.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gaolu, Z. (2015). The effect of central business district on house prices in chengdu metropolitan area: A hedonic approach.

Garrod, G., & Willis, K. (1992). Valuing the goods characteristics – an application of the hedonic price method to environmental attributes. *Journal of Environmental Management*, *34*(1), 59–76.

Georganos, S., Grippa, T., Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2019). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*.

Getis, A. (2011). Reflections on spatial autocorrelation. *Regional Sciences and Urban Economics*, *37*(4), 491–496.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics*, *11*(2), 105–111.

Ghysels, E., Plazzi, A., Torous, W., & Valkanov, R. (2012). Forecasting real estate prices. *Handbook of Economic Forecasting*, *2*, 509–580.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, *35*(1), 46–56.

Gillard, Q. (1981). The effect of environment amenities on house values: The example of a view lot. *Professional Geographer*, *33*(1), 216–220.

Goodman, A. C. (1998). Andrew court and the invention of hedonic price analysis. *Journal of Urban Economics*, *44*(2), 291–298.

Greenstone, M. (2017). The continuing impact of sherwin rosen's "hedonic prices and implicit markets: Product differentiation in pure competition". *Journal of Political Economy*, *125*(6), 1891–1902.

Griliches, Z. (1961). *Hedonic price indexes for automobiles: An econometric of quality change* (pp. 173–196). National Bureau of Economic Research, Inc.

Grislain-Letrémy, C., & Katossky, A. (2014). The impact of hazardous industrial facilities on housing prices: A comparison of parametric and semiparametric hedonic price models. *Regional Science and Urban Economics*, *49*, 93–107.

Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., & Graeler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition, Vol. 1* (pp. 278–282).

Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140–152.

Huh, S., & Kwak, S. J. (1997). The choice of functional form and variables in the hedonic price model in Seoul. *Urban Studies*, *34*(7), 989–998.

Hussain, T., Abbas, J., Wei, Z., & Nurunnabi, M. (2019). The effect of sustainable urban planning and slum disamenity on the value of neighboring residential property: Application of the hedonic pricing model in rent price appraisal. *Sustainability*, *11*(4), 1144.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in r*. Springer.

Kain, J. F., & Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association*, *65*, 532–548.

Kask, S. B., & Maani, S. A. (1992). Uncertainty, information, and hedonic pricing. *Land Economics*, *68*(2), 170–184.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455–5516.

Kim, S. (1992). Search, hedonic prices and housing demand. *The Review of Economics and Statistics*, *74*(3), 503–508.

Kim, J.-W., Yoon, S., Yang, E., & Thapa, B. (2020). Valuing recreational beaches: A spatial hedonic pricing approach. *Coastal Management*.

Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy*, *74*(2), 132–157.

Law, S., Paige, B., & Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(5), 1–19.

Li, M. M., & Brown, H. J. (1980). Micro-neighbourhood externalities and hedonic housing prices. *Land Economics*, *56*(2), 125–141.

Li, F., Krishna, R., & Xu, D. (2019). Convolutional neural networks for visual recognition. https://cs231n.github.io/convolutional-networks/ (Accessed: 2021-03-30).

Linneman, P. (1980). Some empirical results on the nature of the hedonic price function for the urban housing market. *Journal of Urban Economics*, *8*(1), 47–68.

Lipton, Z. (2016). The mythos of model interpretability. *Communications of the ACM*, *61*, http://dx.doi.org/10.1145/3233231.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, Vol. 30*. Curran Associates, Inc..

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504.

Mason, C., & Quigley, J. M. (1996). Non-parametric hedonic housing prices. *Housing Studies*, *11*(3), 373–385.

McMillan, D., Jarmin, R., & Thorsnes, P. (1992). Selection bias and land development in the monocentric model. *Journal of Urban Economics*, *31*, 273–284.

McMillen, D., & Redfearn, C. (2007). Estimation and hypothesis testing for nonparametric hedonic house price functions. *Journal of Regional Science*, *20*.

Mei, Y., Hite, D., & Sohngen, B. (2017). Demand for urban tree cover: A two-stage hedonic price analysis in California. *Forest Policy and Economics*, *83*, 29–35.

Ming, Y. S., & Hian, C. C. (2005). Obstruction of view and its impact on residential apartment prices. *Pacific Rim Property Research Journal*, *11*(3), 299–315.

Ministry of the Interior Kingdom Relations (2021). Rented housing. https://www.government.nl/topics/housing/rented-housing (Accessed: 2021-07-15).

Mok, H. M. K., Chan, P. P. K., & Cho, Y. S. (1995). A hedonic price model for private properties in Hong Kong. *Journal of Real Estate Finance and Economics*, *10*(1), 37–48.

Molnar, C. (2019). Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/ (Accessed: 2021-03-30).

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, *124*, 389–404.

Nair, R., & Narayanan, A. (2012). Benefitting from big data leveraging unstructured data capabilities for competitive advantage. https://motamem.org/wp-content/uploads/2019/01/PWC-Big-data-Definition.pdf (Accessed: 2021-03-30).

Neloy, A., Haque, H., & Islam, M. (2019). Ensemble learning based rental apartment price prediction model by categorical features factoring. In *Proceedings of the 2019 11th international conference on machine learning and computing* (pp. 350–356).

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2020). Activation functions: Comparison of trends in practice and research for deep learning. In *Proceedings of 2nd international conference on computational sciences and technologies*.

Oladunni, T., & Sharma, S. (2016). Hedonic housing theory – A machine learning investigation.

Owusu-Ansah, A. (2013). A review of hedonic pricing models in housing research. *A Compendium of International Real Estate and Construction Issues*, *1*, 17–38.

Palmquist, R. B. (1984). Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics*, *66*, 394–404.
Palmquist, R. B. (1992). Valuing localized externalities. *Journal of Urban Economics*, *31*, 59–68.
Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, *29*(4), 667–676.
Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141. http://dx.doi.org/10.1007/s11747-019-00710-5.
Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 97–101).
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 35–55.
Sakia, R. (1992). The box-cox transformation technique: A review. *The Statistician*, *41*(2), 169–178.
Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd international conference on learning representations (iclr2015)*.
Smith, V. K., & Huang, J. C. (1993). Hedonic models and air pollution: Twenty-five years and counting. *Environmental and Resource Economics*, *3*, 381–394.
So, H. M., Tse, R. Y. C., & Ganesan, S. (1996). Estimating the influence of transport on house prices: Evidence from Hong Kong. *Journal of Property Valuation & Investment*, *15*(1), 40–47.
Sommer, K., Sullivan, P., & Verbrugge, R. (2010). *Run-up in the house price-rent ratio: How much can be explained by fundamentals?*: *Working papers 441*, U.S. Bureau of Labor Statistics.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the 2015 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1–9).
The World Air Quality Project (2021). Air pollution in rotterdam. https://aqicn.org/map/rotterdam/ (Accessed: 2021-07-15).
Vijayarani, S., Ilamathi, J., & Nithya, S. (2015). Preprocessing techniques for text mining - An overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.
Wilhelmsson, M., & Long, R. (2020). Impacts of shopping malls on apartment prices: the case of stockholm. *Nordic Journal of Surveying and Real Estate Research*, *5*, 29–48.
Williams, A. (1991). A guide to valuing transport externalities by hedonic means. *Transport Review*, *11*(4), 311–324.
Wolf, K. L. (2007). City trees and property values. *Arborist News*, *16*(4), 34–36.
Yamashita, R., Nishio, M., Do, R., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, *9*, 611–629.
Yoo, S., & Wagner, J. E. (2016). A review of the hedonic literatures in environmental amenities from open space: A traditional econometric vs. spatial econometric model. *International Journal of Urban Sciences*, *20*, 141–166.
You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, *19*(12), 2751–2759.
Zhang, Y., & Dong, R. (2018). Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in Beijing. *ISPRS International Journal of Geo-Information*, *7*(3), 104.
Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of advances in neural information processing systems, Vol. 31 (neurips 2018)*.
Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, *39*(1), 272–281.
Zhu, J. J., Chang, Y.-C., Ku, C.-H., Li, S. Y., & Chen, C.-J. (2020). Online critical review classification in response strategy and service provider rating: Algorithms from heuristic processing, sentiment analysis to deep learning. *Journal of Business Research*.

Tomasz Potrawa received his B.S. in Quantitative Economics and Information Systems from Warsaw School of Economics and M.S. in Data Science and Marketing Analytics from Erasmus School of Economics. His research interests include consumer behavior, digital marketing and information retrieval. He has participated in research studying aspects such as marketing mixmodeling, propensity modeling and customer success management.

Anastasija Tetereva holds Diploma in Mathematics from the University of Latvia, Master's Degree in Statistics from Humboldt University of Berlin, Germany, and Ph.D. in Economics and Finance from the University of St. Gallen, Switzerland. Prior to joining Erasmus University Rotterdam as an Assistant Professor, she had postdoctoral appointment at the Swiss Institute for Empirical Economic research. Her research interests are mainly in the field of Financial Econometrics with a focus on machine learning methods for modeling high-dimensional and high-frequency time series. Her second area of interest is the incorporation of novel data sources into econometric models.