Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis
Data Science and Marketing Analytics

Hedonic Pricing Model for Rotterdam Housing Market Empowered by Image Recognition and Text Analysis Methods

Student name: Tomasz POTRAWA
Student ID number: 537444

Supervisor: Dr. Anastasija TETEREVA
Second assessor: Dr. Bas DONKERS

July 22, 2020

ERASMUS UNIVERSITEIT ROTTERDAM

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction 3

2 Theoretical background 6

2.1 Hedonic pricing theory 6

2.2 Application of hedonic pricing for housing market 8

2.3 Methodology 12

3 Methodology 15

3.1 Data extraction 15

Image analysis 15

Text analysis 17

3.2 Regression analysis 18

Hedonic linear model 18

Random Forest 19

3.3 Model-Agnostic methods 20

Variable Importance 20

Partial Dependence Plots 21

Local Interpretable Model-Agnostic Explanations (LIME) 22

4 Data and feature extraction 24

4.1 Initial data 25

4.2 Image recognition 25

4.3 Text analysis 30

4.4 Summary 31

5 Analysis and results 35

5.1 Regression 35

5.2 Explanatory analysis 42

6 Conclusions 49

7 Appendix 50

List of Figures

1 The accuracy of the model versus the number of epochs 28

2 Correlation plot for numerical variables. 36

3 Residuals versus Fitted Values 40

4 The Partial Dependence Plot of Living Area 44

5 The absolute value of a view on the city versus predicted rental cost 47

6 The absolute value of insolation versus predicted rental cost 47

7 The absolute value of water body proximity versus predicted rental cost 48

8 Auto-correlation plot for residuals of linear model 53

9 Normal QQ plot for linear model 53

10 Residuals versus fitted values after Box-Cox transformation 54

11 Variable importance of the final random forest model 55

List of Tables

1 List of Commonly Used Housing Attributes in Hedonic Price Models (Chin & Chau, 2003). 10

2 Variables present in the initial data set 26

3 Accuracy of the image recognition models 29

4 Accuracy of the view image recognition model on the property level 30

5 Descriptive statistics 1 32

6 Descriptive statistics 2 33

7 The coefficients of the linear model empowered by image and text features 38

8 The coefficients of the final linear model 39

9 The summary of regression models 42

10 The comparison of variables' impact on the prediction between the models 43

11 Frequency of the words appearing in the descriptions 50

12 Generalized Variance Inflation Factor for linear regression 52

Abstract

This study primarily aimed to address the question of whether using relevant images and descriptions of rental apartments increases the performance of a hedonic pricing model for the Rotterdam housing market. Secondarily, an attempt on deriving the hedonic prices of housing attributes and their dependence on the used regression model was made. With the usage of convolutional neural nets and text analysis methods, features related primarily to the external attributes of properties were extracted and transformed into tabular data. Two models were applied, the OLS regression and the random forest. In both cases using the extracted variables in addition to the more traditional predictors led to increases in accuracy and explained variation. A comparison of accuracy metric revealed significant superiority of the random forest over its linear counterpart. On the contrary to the previous research conducted in the field, the comparison between both methods was not limited solely to the accuracy metric and included the comparison of models' coefficients acquired with the model agnostic methods. It was found that the OLS regression model, when compared to the random forest, is more likely to overestimate the value of nonessential structural attributes such as garden or furnishings. This may be a reflection of the model's reduced capability of capturing locational aspects. Lastly, based on the obtained data, and the usage of local interpretable model-agnostic explanations, it was concluded that the hedonic price of a particular housing attribute is unlikely to be constant and its variability is dependent on the total value of a property.

Keywords: hedonic price model, housing attributes, image recognition, model-agnostic methods

1 Introduction

The analysis of the real estate market has always been an object of remarkable interest among researchers. The reason behind this trend is the fact that household prices affect directly the level of life of a significant part of society. Furthermore, the fluctuations in the real estate market reflect accurately the macroeconomic phenomena happening in the world, thus making them an interesting benchmark in the analysis of other social and economical aspects. There have been many different approaches taken in explaining the household prices so far. One of the most popular methods, hedonic price modeling, claims that a good which in case of a housing market would be a real estate, does not possess any utility by itself. Instead the good consists of characteristics that posses constituted utility (Lancaster, 1966). As such, by comparing the prices of goods that do possess a given attribute and the goods that do not, it becomes possible to derive the value of the attribute in the eyes of consumers.

This characteristic of hedonic pricing made it a perfect tool in estimating the value of hardly tangible assets. In numerous research, applying hedonic pricing on the housing market allowed measuring the value of aspects such as air quality, green area proximity, a view on the ocean, or the criminality level in the neighborhood. Hedonic models aim to quantify the importance of these less tangible assets, therefore making the interpretability of the model a top priority. Consequently, the majority of papers related to hedonic pricing are based on simple statistical models such as linear regression where the constituted value of the attribute of interest may be easily derived by analyzing the regression coefficients. Models that could offer more predictive power like neural network or random forest have remained relatively unpopular in this area of study, due to the complexity of their interpretation, leaving a space for further research.

Additionally, most of the hedonic price modeling research have been based on simple, tabular data, obtained often from institutions such as municipalities or real estate agencies. Only recently the interest in more complex data extraction from sources such as texts and images started growing. While the automation of the mentioned data extraction is definitely not an easy task, in the long run, mastering it will allow decreasing the costs of gathering complex data and increase the versatility and accuracy of future research.

Lastly, in comparison to the traditional housing market, there is relatively little research done on the rental prices. It applies especially for the measurement of external and locational factors impacting the rental price of a property such as the proximity of green areas, view from a property, nearby services etc. While it may be argued that the mentioned type of variables has probably lower impact on the rental prices than on the selling prices of real estates, it is rather unlikely that these attributes do not contribute at all to the former.

As such, it has been concluded, that there are a few gaps in the hedonic pricing literature that this paper aims to fill. First and foremost, the study addresses the functionality of images and texts in the hedonic pricing of Rotterdam's rental market. Therefore, the main research question may be formulated in the following way:

Does including the features that are automatically extracted from the relevant image and text sources significantly increase the accuracy of the hedonic pricing model of Rotterdam's rental housing market?

Additionally, the study focuses on measuring the impact of external factors such as the neighborhood characteristics on the rental prices in Rotterdam. The analysis extends the previous research of Law et al. (2019) as in order to address the question above, not only maps but also rental offer photos serve as a source of information for the study.

Furthermore, the comparison between the accuracy of traditional hedonic linear regression and complex machine learning methods is carried out. While this kind of comparison is definitely not new in the research area of the housing market, to the best knowledge of the author none of the previous papers studied the differences in the impact of external variables on the prediction across the models. The lack of such an approach in the literature may be caused by the cumbersomeness and the moderate reliability of the methods aiming at explaining the black box model behaviour. As such it is understandable that the previous research opted to focus on the easily measurable and comparable accuracy of the models.

Nevertheless in the eyes of the author the blind pursuit of the best possible accuracy without understanding how the prediction is being made is not completely relevant in the context of hedonic pricing. One could argue that the ability to quantify the constituent value of goods' characteristics is the essence of the traditional hedonic model. Once this ability is lost it becomes infeasible to use the model in deriving the value of assets such as previously mentioned air quality or green area proximity. Therefore, with the usage of model agnostic methods the paper aims to prove that the more complex, but often more accurate black box models may also be used to derive the value of goods' attributes. The importance of housing attributes in advanced machine learning algorithms is compared with the traditional hedonic OLS regression coefficients. Furthermore, based on a non-linear regression model, it is measured how hedonic prices of given attributes change, depending on the price of a property.

It is worth mentioning here that the study has been designed in a way that for the most part its methodology and findings should be relatively easily applicable to other cities. Rotterdam is a major European city with diverse ethnic and religious groups, highly-developed services as well as the transportation sector and reputable educational institutions. As such, the phenomena specific for the Rotterdam's housing market may at least partially be true for many other cities in the world. Moreover, the data used in the study has been acquired from the rental websites and for the most part, represents features and attributes valid for any other housing market. This further facilitates conducting similar research for other cities.

2 Theoretical background

In order to prepare for addressing the questions and problems defined for this research, proper groundwork in the form of a broad literature review was laid. The following section provides an in-depth overview of what has been already done in the literature with respect to hedonic methods and real estate appraisal. The first part aims to provide the theoretical background and assumptions of the hedonic price modeling. Subsequently, the summary of the most relevant real estate appraisal papers and the insights gathered from them is presented. The last part of this chapter focuses on the methodology used in particular research, and the comparison between statistical models commonly used in the area of house market pricing.

2.1 Hedonic pricing theory

Hedonic pricing is a method rooted in the theory of consumer demand developed by Lancaster (1966). In his work, the author breaks away from the back-then traditional assumption claiming that a good by itself possesses utility. Lancaster's innovative approach may be summarised by three main statements. Firstly, a good, per se, does not provide any utility. Instead, each good consists of characteristics that posses constituted utility. Secondly, a good is usually characterized by many attributes, and a single attribute is most of the time shared by multiple goods. Lastly, according to the author, the combination of goods may possess different characteristics than the goods would have, if evaluated separately. The value of each attribute in Lancaster's work is based on the relationship between the observed prices of the goods and the number of attributes associated with them (Chin & Chau, 2003).

Even though hedonic pricing is heavily influenced by Lancaster's theory of demand from 1966, it was officially formulated a few years later by Rosen (1974). Despite the fact that both Rosen's and Lancaster's models are based on similar assumptions, they do have some significant differences. In the Lancastrian model, it is claimed that goods, and consequently their characteristics, are aggregated in groups, which subsequently are purchased by the customer. On the other hand, Rosen states that goods are usually not acquired in combinations. Instead, for each good customer makes a separate decision after analyzing the spectrum of brands offering the given good. While both models do not necessarily contradict each other, they do seem to fit better different kinds of products. Chin & Chau (2003) argues, that Rosen's hedonic model is more suited to durable goods while the Lancastrian model may explain better phenomena related to consumer goods.

Another dimension on which both authors do not come to similar conlusion is the type of relationship between the price of a good and its characteristics. In Lancaster's work it is presumed that the mentioned relation is linear, whereas Rosen postulated that it is more likely for this relationship to be non-linear. The latter, implies that the implicit price of an attribute is non-constant and according to Rosen depends on the interaction between the supply and the demand. More precisely, in Rosen's model the marginal implicit price of an attribute is presented as the partial derivative of the price function over the given attribute, which always has to be equal to the marginal cost of an extra unit of the attribute in the market and the individual consumer's marginal willingness to pay for the attribute. The latter aspect, which could also be simply called a demand price, is assumed to be a function of the buyer's income, the utility level and other variables describing particular consumer such as tastes and preferences, age or education (Chin & Chau, 2003).

Despite the fact, that Rosen's model is still seen as a landmark paper, allowing a better understanding of multiple fields of economics such as urban, environmental, or labor economics (Greenstone, 2017), it does suffer from multiple empirical issues. The first is the choice of the functional form. Throughout the years, multiple types of forms have been applied to hedonic models e.g. linear, semi-log, or Box-Cox transformation-based. Nevertheless, there is little research on how the functional function should be chosen (Butler, 1982). Moreover, even with applying more complex techniques such as Box-Cox transformation aiming at normalizing the distribution of the data, it is not guaranteed that the transformed data will suit the model and its assumptions. On the other hand, applying more advanced machine learning methods with less strict requirements may lead to a situation where measuring the marginal implicit price of an attribute may be challenging or sometimes even impossible.

Another issue related to hedonic pricing models is the problem of incorrect choice of the attributes, also knows as the misspecification of the variables. In cases when the non-relevant predictor is included in the model we refer to it as an over-specification, whereas the opposite situation, when a key variable is not included in the analysis is known as an under-specification. While both cases are obviously not welcome, it may be argued which one causes more concern and consequently bias more the results of the hedonic model. Chin & Chau (2003) argues that as a result of an over-specification predictors are still unbiased and consistent, however, some of them are inefficient. An under-specification on the other hand results in an estimated biased and inconsistent coefficients. In the opinion of Butler (1982) as all the estimates of a hedonic price model are in some degree misspecified, in general, the models with a small number of key attributes should perform well enough. Overall, Butler suggests to use in the regression only the predictors that are expensive in producing and are likely to yield additional value in the form of utility. To similar conclusion came Mok et al. (1995) who concurred that the bias introduced to a model due to missing attribute of a good is usually small, and does not cause a significant drop in model's performance in the terms of prediction and explanatory power. In addition, the author discourages using proxy predictors in a hedonic price model as they result in biased and inconsistent outcome of an analysis.

The mentioned drawbacks of hedonic pricing are not the only issues related to this method. The description of each one of them is not in the scope of this paper, however, the last one worth mentioning is the unrealistic assumption of perfect competition. It implies among the others, that the information flow between consumers and suppliers is instant and not disrupted in any way. Consequently, the model does not take into account the delayed reaction of a market to any changes or an imperfect estimation of a value (or in this case also utility) of given attributes. As far as the assumption of the perfect competition is not unique to hedonic models, and therefore, mentioned limitations are relatively popular in economical models, hedonic pricing models distinguish themselves with one major limitation. In real life, it is completely possible that consumers do not have a full knowledge of every attribute of a good they are considering to purchase. Nevertheless, they still have to estimate the value/utility of a good on the basis of the knowledge they possess. Therefore, even the best hedonic model may fail in estimating the constituent price of an attribute if the consumers fail to perceive it.

All the mentioned limitations did not stop, hedonic pricing from becoming a popular tool in multiple aspects of economics. It is still widely used in numerous research especially among those focusing on deriving the value of hardly tangible environmental and sociological aspects such criminality rate or air pollution. Even though the above summary may seem like an oversimplification of a complex phenomenon which is hedonic pricing, hopefully, it is sufficient to prove that a hedonic model is not just a regression analysis aiming at measuring the magnitude of attributes' coefficients, but also a theory-based marketing analysis.

2.2 Application of hedonic pricing for housing market

Hedonic pricing has been extensively applied in the real estate appraisal research due to the characteristics of a housing market. The fact that real estates may be treated as goods consisting of multiple separate attributes such as a living area, number of rooms or localization, perfectly matches the requirements of a hedonic approach. Furthermore, the changes in the housing market prices accurately reflect the macro-economical phenomena happening in the society, which puts a housing market in the position of a popular benchmark used in other types of analysis. Moreover, the combined versatility of a housing market and hedonic approach in addition to the easily accessible data, makes it a perfect tool e.g. in estimating the values of environmental features such as air quality over the world.

The attributes of a property in the hedonic approach tend to be divided into multiple categories in the literature. In their review of hedonic pricing papers Chin & Chau (2003) propose the division of characteristics into three main groups. The first one, the locational group consists of real estate characteristics such as the distance to the central business district (CBD) and the type of view that a given location has to offer e.g. view on the lake or the golf course. The accessibility to the center of a city has been found to have an impact on the price of a house in multiple research (McMillan, 1992; Palmquist, 1992). Similarly, numerous types of views from a property have been proven to influence the price of an estate (Gillard, 1981; Mok et al., 1995). Moreover, according to Benson et al. (1998) not only the type of view affects the value of a house, but the quality of it also plays a crucial role. As an example, in their work, Benson et al. (1998) concluded that the ocean frontage adds 147% to the property's sales price, while the obstructed, partial view on the ocean only contributes with an increase of 10% of the price.

The second category, namely the structural attributes, describes the traditional characteristics of a house such as a living area, the number of bedrooms, or the age of a building. Throughout the years a plethora of research aimed to quantify the importance of particular construction aspects of a building on a price. Numerous papers agree on the fact that one of the most important structural attributes is a floor area followed by the number of rooms, bedrooms, and bathrooms (Ball, 1973; Garrod, 1992). While the age of building according to multiple research such as D. E. Clark & Herrin (2000) is negatively related to the estate price, in some analysis (H. J. Li M M Brown, 1980) it has been proved to have an opposite effect, probably due to a historical value older buildings may possess. Lastly, additional areas of an estate such as a garage, a basement, a patio, or a storage in multiple research were also classified as significant predictors in the regression analysis of house prices (Forrest, 1996; Garrod, 1992).

The last class of attributes relates to a broad range of features characterizing the neighborhood of a real estate. According to Linneman (1980), these attributes stand for 15% up to 50% of the standardized variation of a site evaluation model. Previous research divide further the neighborhood characteristics into three subgroups: socio-economic variables, local government/municipality services, and externalities such as crime rate or traffic noise (Chin & Chau, 2003).

In the past studies, the level of income in the neighborhood (Kain, 1970) and the dominating ethnicity in the area (Ketkar, 1992) have been found as significant variables impacting the predicted housing price. In terms of municipality services, the proximity to good schools (D. E. Clark & Herrin, 2000) and places of worship such as mosques (Carroll, 1996) proved to positively impact the value of a house. On the contrary, the high criminality rate in an area undoubtedly lowers the average value of an estate (D. E. Clark & Herrin, 2000). The slightly modified summary of the most popular variables used in hedonic price models, and their impact on the real estates' prices prepared by Chin & Chau (2003) is presented in Table 1 below.

The external factors impacting the housing price have been much more widely covered by Karanikolas et al. (2011) where the authors provide a summary of insights gathered from numerous research in that area. In general, the features analyzed in this paper may be divided into four categories: green areas, water areas, topography, and environmental risks. The first category was a matter of interest for Bishop (2005) where it was concluded that there is an evident correlation between the existence of green spaces and the market value of houses. It was estimated that the houses located nearby green areas may be even up 20% more expensive than their counterparts without that access.

Similarly, in Wolf (2007) the author found out that potential buyers are willing to pay much more for a house if it is located nearby an existing park. Moreover, the type of a green area had a significant impact on the estate price. The traditional parks located in a distance of up to 400

Table 1: List of Commonly Used Housing Attributes in Hedonic Price Models (Chin & Chau, 2003).

	Attribute	**Expected effect on housing price**
Locational	Distance from CBD	negative
	View of the sea, lakes or rivers	positive
	View of hills/valley/golf course	positive
	Obstructed view	negative
	Length of land lease	positive
Structural	Number of rooms, bedrooms, bathrooms	positive
	Floor area	positive
	Basement, garage, storage and patio	positive
	Building services (e.g. lift, AC)	positive
	Floor level	positive
	Structural quality	positive
	Facilities (e.g. swimming pool, gym, tennis court)	positive
	Age of the building	rather negative
Neighbourhood	Income of residents	positive
	Proximity to good schools	positive
	Proximity to hospitals	unknown
	Proximity to places of worship (e.g. churches or mosques)	positive
	Crime rate	negative
	Traffic/airport noise	negative
	Proximity to shopping centers	unknown
	Proximity to forest	unknown
	Environmental quality (e.g. landscape, garden, playground)	positive

meters increased on average the price by 10% while the existence of nearby parks that were not free for the public lead to an increase of 20%.

In terms of water areas, their impact on household prices in Arizona have been investigated by Colby (2003). In general, the prices of estates were higher in cases where water area such as a lake or a river were located up to 2 kilometers from a house. An estimated increase in the price of 5.9% was the highest for properties located around 160 meters from the water. On the other hand, it is worth mentioning that the properties located at a distance of less than 150 meters were cheaper, probably due to the risk of floods. Nevertheless, the area of Arizona for which the research was made, is characterized by dry and desert climate giving the paper's result a specific context.

Another type of external features impacting the price of households refers to the topography of the real estate's neighborhood. While the natural landform is not an object of primary focus in this research, the impact of anthropogenic features of the neighborhood cannot be omitted. According to Klein (2003) the proximity of a highway negatively impacts the price of real estates by approximately 8 to 10%. It could mean, that the easiness of communication is not enough to compensate for noise pollution caused by high-speed roads. Interestingly, the impact of railway proximity has been estimated to be smaller in Brinckerhoff (2001) and was equal to 6.7% decrease in the market value of a property.

Due to a much wider range of literature, the gathered insights mostly relate to the value of real estate, not its rental price. As far as these two are obviously strongly correlated, they may differ significantly in some aspects, especially in the context of hedonic pricing. In the end, the hedonic method implies that the value of a good comes from the consumers' willingness to pay for its particular attributes. Therefore, the total utility of the same property may be different for a potential tenant and a potential buyer. As an example, a poorly equipped, small flat close to the university may be worth much more in the eyes of a student interested in renting a property than of a working person looking for a property to buy.

The ratio between house price and rental cost has been an object of analysis for S. Clark & Lomax (2019). The authors reach in the paper quite a few remarkable conclusions on the rent/price ratio. Firstly, the ratio is the highest for flats, followed by terraced houses, leaving (semi)detached houses behind. Secondly, most probably due to a handful of rental offers with large living areas, the ratio has been found to raise with the increase in the surface and the number of bedrooms. In terms of neighborhood attributes, the ratio turns out to be lower for properties being in proximity to "healthy retail environment (away from fast-food restaurants, tobacconists and gambling) and access to health services" (S. Clark & Lomax, 2019). On the contrary, the neighborhood with high-quality environmental features such as low air pollution and access to green areas, raises the mentioned price/rent ratio. Finally, S. Clark & Lomax (2019) concluded, that some attributes that increase the price of a real estate, do not impact its rental price e.g. the distance to a good school or proximity of amenities.

2.3 Methodology

Most of the studies analyzed in the previous section have been based on a traditional, linear hedonic pricing regression model. While the simplification in the form of linearity assumption, allows an easy determination of the importance of each attribute characterizing a given good (by analyzing the coefficients of an OLS regression), it disallows the model to capture more complex, nonlinear patterns in the data. While it is difficult to distinguish the complexity of patterns in a given data set prior to the analysis, in his work Rosen (1974) argues that in general, the nonlinearity between the price of goods and their inherent attributes is likely to happen. The problem can be addressed by applying advanced machine learning methods. This part of the literature review provides a summary of research where this type of approach has been used in the context of the housing market.

One of the attempts to compare the performance between the traditional hedonic pricing regression model and more complex algorithms has been made by Limsombunc et al. (2004). According to the results, the artificial neural network outperforms the hedonic price model in terms of sheer predictive power. Moreover, the hedonic model has been criticized by authors, due to its assumptions and common problems such as data multicollinearity and heteroscedasticity or inability to capture non-linear patterns.

On the other hand, the authors emphasized the complexity of neural network interpretation and did not perform any type of variable importance analysis. Furthermore, the paper includes references to previous research such as Lenk et al. (1997) and Do & Grudnitski (1992) which argue that the results of a neural network model may be inconsistent and not always outperform regression models.

Another comparison between the performance of hedonic regression and machine learning algorithms has been done by Neloy et al. (2019). On contrary to most of the research in the area of housing pricing, the authors opted to base their study on rental cost data. The comparison between linear regression, penalized linear regression, support vector machine, neural network, and multiple versions of decision tree-based ensemble learning methods, shows the empirical superiority of the latter. While the differences in the average RMSE between tree-based methods and other machine learning models were of high magnitude, the differences between variations of random forest in the form of Bagged Trees, Gradient Boosted and XGBoosted were rather negligible.

Recently Hong et al. (2020) applied a random forest model in price evaluation of Seoul households. The results of the research were surprisingly good as the average percentage deviation between the predicted and actual market price was equal to 5.5 (out-of-sample). In comparison, applying the traditional OLS-based hedonic regression on the same data led to an average percentage deviation of almost 20. As such, it can be concluded, that decision tree models can be more successful in predicting house prices, than their traditional linear regression counterparts. Nevertheless, it may be worth mentioning that the data used in the research was of extremely good quality as it consisted of 40% of all the transactions made in the area within the last 10 years. Moreover, the data was limited only to one, not diversified district of Seoul. Therefore, achieving such good results with the usage of a decision tree model only, may not be feasible in most of the cases.

The choice of a regression technique for this research is not an easy task as both, probably most popular advanced machine learning techniques neural networks and random forest have proved to be effective in the previous research. Nevertheless, decision tree-based methods seem more reasonable. As non-linear models, they can capture more complicated patterns in the data than the traditional hedonic regression. On the other hand, the regression will be based on a relatively simple tabular data where extremely complicated relations between variables are not expected. In his study Rossbach (2018) provides empirical results of a comparison between the performance of 179 neural net and random forest models. In most of the cases, random forest did not come in short in comparison with its deep learning counterparts. Moreover, the author emphasizes the advantage of tree-based methods over neural nets in terms of robustness, benefits in cost and time, and especially the easiness of interpretability.

The superiority of artificial neural networks is usually visible in areas such as text or image analysis where connections between different predictors may be extremely abstruse. Therefore, random forest seems like a well-suited model for this research, due to fairly advanced pattern recognition and the ease of use. The latter will be of most benefit in the later stage of research, where multiple models will be compared to see if including features gathered from images and other sources significantly improves the prediction accuracy. On the other hand, neural networks are probably the best method to be used in the mentioned image recognition part of the study.

The number of applications where image recognition is being used has been constantly growing over the last decade. The ability of software to identify patterns, people, and objects from the image is not only directly used in areas such as driver assistance systems but has also proved to be helpful in increasing the accuracy of statistical models. In their research Bajari et al. (2019) approached the topic of measuring inflation based on quality-adjusted prices of different products available at Amazon. The features of the mentioned goods have not been limited just to conventional ones as both text and image analysis have been performed. The authors have found that including features extracted from images as well as texts leads to a significant improvement in model's performance.

The usage of image recognition in the analysis of the real estate market has been a popular area of study in recent years. Law et al. (2019) successfully used the Street View and satellite images to improve the performance of house price evaluation models. Nonetheless, the authors opted to focus on the accuracy of pricing by using black box model, while the relative importance of different predictors has not been their top priority. Moreover, Law et al. (2019) work could also be extended by using text analysis, environmental attributes usually present in hedonic pricing models and images of property interior.

The last aspect has been studied in papers such as Poursaeed et al. (2018) and You et al. (2017). In the former, the convolutional neural network has been trained to rank the photos on the scale of one to eight based on the luxury of a property. In the latter, the authors have taken quite an innovative approach of using recurrent neural network LSTM model in order to predict the price of the house solely on the basis of photos and location. Typical data e.g. size or number of rooms has not been used in the research which did not stay in the way of reaching promising results.

Unfortunately, there is one major shortcoming of all the papers cited in the methodological part of the literature review. All the mentioned research, do apply advanced machine learning algorithms to the housing market data, nevertheless, their focus is solely put on the prediction accuracy of the created models. Almost no attention is given to the estimation of the housing attributes importance and their utility. Therefore, it could be argued that the comparison between the hedonic linear regression and machine learning methods presented in the cited papers is not fully appropriate, as the essence of the hedonic price modelling, the value of each attribute of a given good, is being lost in the pursuit of achieving the best possible accuracy of a prediction.

3 Methodology

The methodology used in the research may be divided into three main categories. The first one, consists of methods used to extract features from complex data sources such as images and text, and transform the gathered insights into low-dimensional tabular data. The second class includes the methodology describing the regression models used in the research. The purpose of the methods present in the last group is to explain the behaviour of the black box model described in the second category of the section, and determine the importance as well as the impact of particular predictors on the prediction.

3.1 Data extraction

Image analysis. Using a convolutional neural network (CNN) is a standard approach while building an image recognition model (F. Li et al., 2019). Convolutional neural nets are constructed similarly as regular neural networks, they also consist of the input layer, hidden layers, and the output. The main difference lies in the architecture of the layers. In image recognition models, the input usually takes the form of A x B x 3 matrices, where A stands for the picture width, B for the picture height while 3 represents color channel values (RGB). Let's assume that the pictures fed to the model have a small size of 100x100 pixels. In regular neural net each neuron in the first fully connected hidden layer would already have 30000 weights. It is easy to imagine how the number would scale for larger images. As a result of the regular net architecture, not only the computational time of the model would be enormous but the number of adjustable weights could easily lead to overfitting of the algorithm.

Convolutional neural nets deal with this problem by introducing three-dimensional hidden layers, which size is being reduced in a subsequent processing. The first layer in the model is called convolutional layer. Its input size is equal to the input image A (width) x B (height) x 3 (depth), which in consequence usually requires all the images to be scaled to the same size. At 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, and then moves to the new part until the whole picture is scanned. The size of the part of a picture being 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 x 3 x 3 and 5 x 5 x 3 are the most popular, mostly due to the fact that they are used in some of the best performing CNNs: GoogleNet and VGG (Szegedy et al., 2015; Simonyan & Zisserman, 2014). The process of multiplying the input values with a filter and moving to a new part is in the literature referred to as convolving the filter with the image (F. Li et al., 2019). The number of times the filter has to convolve with the picture depends on the kernel size and a parameter known as a stride, which indicates by how many pixels the filter should "slide".

As the output of the numerous multiplication, a two-dimensional activation map is being created for each filter used. The output of a convolutional layer, created by stacking activation maps along the depth dimension takes a form of a D x E x C sized matrix, where C substitutes the previous input image depth of 3, with the value equal to the number of activation maps. D and E again stand for the width and height of a matrix, however on the contrary to A and B they do not always match the image size. Instead, their value depends on the number of times the kernel may fit in the input matrix. Therefore, D and E tend to have smaller values than A and B unless the technique known as padding is being used. The classic way of padding called "same padding", augments the input matrix by adding column(s) and row(s) on the matrix borders with imputed values of 0. The aim of this approach as the name suggests is to equalize the sizes of input and output matrices, so no data would be lost in the process of transferring data to consecutive layers.

The number of filters used in the convolutional neural net is yet another parameter that may be tuned. As one could think of filters as feature detectors, it could be argued that the more complex the patterns that the model aims to capture, the more filters it will need to do it properly. On the other hand, each filter drastically increases the output of a convolutional layer, thus, increasing the computational power needed to train a model. As the training process continues, the network learns filters that are being activated when some specific features of the image are being captured. The first convolutional layer is responsible for extracting low-level features of an image such as edges or colors. With each convolutional layer added, the higher-level features start to be captured by the model. The output of a convolutional layer is subsequently passed to an activation function layer which in case of a CNN is usually a ReLU function defined as $f(x) = max(0, x)$ which applies elementwise non-linearity (F. Li et al., 2019). However, in cases of too complex architecture of a model, the problem with overfitting may appear. Srivastava et al. (2014) propose a relatively simple way of addressing that problem with a technique called dropout. The key idea of this approach is to drop out a random set of activations, by setting them to zero. As counter-intuitive as the method may seem, it forces the model to learn how to correctly classify an image, even when some of its key attributes are lost.

The next type of layer used in convolutional neural nets is called a pooling layer. Its main function is to reduce the size of convolved features, and therefore, the number of parameters and computational time. This goal is again accomplished by analyzing the input matrix by parts. In case of the most popular approach known as max pooling, for each submatrix which size depends on the pre-defined kernel, only the maximum value is returned. Then, the process is repeated until the whole image is traversed and the original matrix is transformed into less-dimensional one.

In each model depending on the needs of the analyst, multiple convolutional, ReLU, and pooling layers may be used. No matter how many of them are included in the architecture of the model and what is their order, the final output is being transferred to the fully connected layer (also known as a dense layer). At this point, the data is being transformed into a column vector, which subsequently is fed to a regular feed-forward neural net, which in the end returns probability values for each class.

Similarly, as in traditional, artificial neural network, the loss and activation functions have to be set separately for the dense layer.

Convolutional neural networks are also analogous to artificial neural nets in terms of data requirements. Both methods usually require a large amount of training data to perform well. However, in case of CNN the process of gathering additional data in the form of images is often more cumbersome than collecting extra tabular data for ANN. Image augmentation is one of the techniques allowing artificially increasing the training sample size. The additional observations are created on the basis of various transformations applied to original photos such as random rotation, shifts in width and height, zooming or flips. The choice of transformations used in a particular model should always be adjusted individually, as their functionality depends on the context of images and the patterns that the model is aiming to recognize.

Text analysis. After the initial analysis of apartment search sites and housing market research, it has been concluded that there is a considerable amount of information gathered in rental offers' description that would be difficult to extract from other sources such as images or maps. Therefore, the text analytics methods are used in the research, however to a limited degree. It appears that most of the text analytics method e.g. sentiment analysis would probably not bring any additional value to the study, due to a distinctive nature of the real estates' descriptions and the type of language they are written in. As such, the focus of this part of the research is put on a simple information extraction from the rental offer description.

In general, a rental offer description may contain a lot of useful information about the neighborhood, close services, and potential rental restrictions. On the other hand, the narrative is expected to be biased as owners do not mention the negative characteristics of their real estates. Moreover, the descriptions differ significantly between each other, and the information mentioned in some is missing in the others. Therefore, where possible, the extraction is done in a way that gathered attributes are easy to categorize or have a dummy form in which, in case of missing data, a default value may be set. The examples of features and data extraction methodology should help in understanding that logic:

- Rental restrictions: some landlords do not allow pets, children, students, or smoking persons in their properties. In order to find if it is a case for a particular property, the description is searched for the mentioned keywords (smoking, pets, etc.). After finding the position of a keyword in a description, the surrounding words are searched for contradictions. The range of the search, usually referred to as a window, is set after the initial analysis of the description examples, and depends on the most popular grammar structures used. The size of a window and its (a)symmetry may also vary between the keywords searched for.

- Nearby services: as the type of services mentioned in the descriptions differs, and it is less probable to assume that e.g. if the owner does not mention nearby restaurants there are indeed no restaurants in the neighborhood, creating dummy variables may not be the best idea. To overcome this problem, a set of keywords based on the literature review, and common sense such as hairdresser, market, restaurant, etc. are created. Then, the number of existing keywords in each description are summed up.

- Additional information: sometimes additional requirements have to be met to rent a property. Similarly, with the usage of keywords, sentences with the words “guarantor”, “minimal income” etc. are found and the window of keywords is searched for the amount of required money/income.

Additionally, in order not to miss any popularly mentioned features in the descriptions that could be specific for the Netherlands or the area of Rotterdam, the frequency analysis of words is performed. Firstly, the popular stop words are removed from the descriptions. Subsequently, the most popular words are manually screened to see if any potentially important characteristic of a property or neighborhood has not been omitted.

3.2 Regression analysis

Hedonic linear model. Most of the research using hedonic pricing are based on an Ordinary Least Squared regression, most commonly called linear regression. The OLS regression model takes the form of::

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

where:
y_i is the value of a dependent variable for i_{th} observation,
β_0 is the intercept term,
β_p is the regression coefficient of the p_{th} independent variable and
ϵ_i is the model's error term.

Let's recall that in the hedonic model presented by Rosen (1974) the marginal implicit price of an attribute is presented as the partial derivative of the price function over the given attribute, which in case of an OLS regression is equal to the value of an attribute's coefficient. Therefore, following the linearity of the model the total value of an attribute p for an i_{th} observation (good) is a product of the observation's value of p and β_p coefficient.

However, in order to rely on a linear model's results five main assumptions have to be met:

- The linearity of the data: the relationship between the dependent and independent variables has to be linear. This assumption can be checked by using scatterplot comparing residuals of the model with fitted values.

- Multivariate normality: the residuals of the model should be normally distributed which can be checked with a Q-Q plot or with a goodness of fit test such as the Kolmogorov-Smirnov test.

- Little to no multicollinearity in the data: the independent variables should not be highly correlated with each other. This assumption can be tested by calculating Pearson's correlation coefficients for each pair of predictors and with a Variance Inflation Factor (VIF).

- Little to no auto-correlation: the residuals of the model should not be correlated with each other which can be checked with the Durbin-Watson test. However, the mentioned test only checks the correlation between directly neighboring residuals, therefore auto-correlation function should also be applied to assure that there are no higher-order effects.

- Homoscedasticity: the variance of error terms should be similar at all levels of the independent variable. Any significant indications of heteroscedasticity may be checked by analyzing the plot of standardized residuals versus predicted values.

While the interpretation and the ease of use are undoubtedly substantial advantages of an OLS regression, it is rather a rare case that data in their original form meet all the above assumptions. Instead, the data transformation using methods such as Box-Cox transformation is often needed. Nevertheless, the more transformed the data, the less interpretable the results of the model become. Moreover, it is argued that Box-Cox transformation reduces the accuracy of any single coefficient (Cassel, 1985). Lastly, no data transformation technique guarantees that the relationship between variables will become linear, thus, bending or breaking the assumptions of an OLS regression is sometimes inevitable. If this is the case, applying an OLS regression to such data leads to the situation where the forecasts, confidence intervals and the insights provided by the model become inefficient and misleading. To conclude, capturing non-linear patterns in the data with an OLS regression is a difficult or sometimes even an impossible task. This is problematic in the context of hedonic price modeling as even in his early work Rosen (1974) claims that some variables such as budget constraints are likely to be nonlinear.

Random Forest. As argued in the methodological section of the literature review, the decision tree-based method, random forest seems to be a suitable method for this research. The decision tree is a non-linear model consisting of multiple conditional statements that separate the data into smaller nodes. The variable on which the split is performed in a given part of the regression model is based on the decrease in RSS it would cause. Overall, at each moment the algorithm chooses the split that leads to the greatest decrease in RSS, then repeats the process for newly created nodes until a stopping rule is met, or the number of observations in each node is equal to one.

The common problem with decision trees is their tendency to grow deep and therefore overfit the data. Random forest is one of the methods that help in dealing with this problem with the usage of bootstrapping. Bootstrap method simulates N new data sets by randomly drawing observations with replacement from the original data set. Subsequently, for each N_{th} bootstrapped sample decision tree model is built. The prediction of the random forest model is an average prediction value among all N regression trees. The fact that the final prediction is the average score of multiple predictions helps in dealing with overfitting by reducing the overall variation of the model in comparison to a single tree (James et al., 2013).

One of the imperfections of this approach is the fact that by default bootstrapped-based trees are highly correlated. Random forest 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 it also disallows the most important variables to completely dominate the final model. As each variable appears only in a subset of trees, predictors which would seem to have less impact on the prediction has more space to act in the final model. The optimal number of predictors used in each tree is relatively easy to define by applying the grid search and choosing the model with the best performance in a chosen metric such as RMSE.

The quite important flaw of traditional random forest is its tendency to rank continuous variables or categorical variables with multiple levels as more important than the others. That is because these types of variables have more possible options to split the data and therefore are more likely to be chosen at higher level splits than e.g. binary variable. In case of this research, a significant part of the data set, especially the features extracted from images have a form of dummy variables. Therefore, in order to limit the mentioned bias, conditional decision tree-based random forest are also used. In contrast to regular trees, conditional trees base the splits in the model on the results of permutation-based significance tests performed for each variable.

3.3 Model-Agnostic methods

In numerous research random forest has proved its superiority over an OLS regression in terms of prediction accuracy. Nevertheless, this advantage comes at the cost of much more complicated architecture of a model which in consequence disallows a simple determination of the variables importance and their impact on the prediction. As previously mentioned in this paper, the issue is especially troublesome for hedonic price modeling, in which the determination of the marginal implicit price of an attribute may be treated as the essence of the method. Fortunately, there is a wide range of methods aiming to uncover the functioning of black box models.

Variable Importance. One of the most standard approaches in determining the importance of variables used in a black box model is the analysis of 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 more the accuracy of the new model decreases in comparison to its original counterpart, the more important the predictor is deemed. As such by repeating the procedure for each predictor in the data set, their relative importance can be gathered. While the method may be criticised for its moderate robustness and unreliability in case of strongly correlated data, it is still an easy and a fast way to acquire the basic overview of the black box model.

Partial Dependence Plots. The partial dependence plot (PDP) shows the marginal effect a feature has on the predicted outcome of a machine learning model (Friedman, 2001). In addition, it allows us to measure in approximation the type of relation the independent and dependent variables have e.g. linear, monotonic, or complex. PDP is based on a partial dependence function, which in case of regression is defined as:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) dP(x_C), \tag{2}$$

where:
x_S is the feature for which partial dependence function should be calculated, and
x_C are all the other predictors featuring in the machine learning model.

While the set S may consist of multiple variables, usually it is limited only to one, as the PDP visualization in more than two dimensions is difficult to analyze. Partial dependence works by marginalizing the machine learning model output over the distribution of the features in set C (Molnar, 2019). That way, the function estimates the relationship between the x_S feature and the predicted outcome of the machine learning model. According to Friedman's approach, the partial dependence plots are obtained by calculating the following average and plotting it over a range of x_S values:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}), \tag{3}$$

where:
n is the number of observations in the train set, and
$x_C^{(i)}$ is the value of the i_{th} sample for the x_C features.

An important assumption of a partial dependence function is no serious correlation between feature(s) x_S and features x_C. If this assumption is not satisfied, the averages calculated in the formula 3 will include data points that are unlikely or even impossible to reach (Molnar, 2019). Another drawback of partial dependence plots is the fact that they do not contain feature distribution. Therefore, the analysis of PDP may be misleading in some parts where not enough data is available.

The partial dependence method may be described as a global function. Based on all the observations that were fed to a machine learning model, it provides a global relationship between their predictions and a particular feature for which function was estimated.

Local Interpretable Model-Agnostic Explanations (LIME). While the globality of a partial dependence function may be seen more as a method's characteristic and not necessarily as a disadvantage or limitation, it does not allow us to understand the reasoning of a machine learning method behind a single prediction. One of the methods which allows this type of analysis is known as Local Interpretable Model-agnostic Explanations or simply LIME. Originally introduced by Ribeiro et al. (2016), LIME aims to explain the functionality of a black box method for a specific observation by fitting a simpler, easier to interpret model also known as a glass-box model at a local scale.

LIME is rooted in the assumption that even the most complex model is linear on a local scale. In consequence, the assumption implies the key idea of LIME which may be formulated in the following way: if two observations possess very similar characteristics, they should behave similarly in a machine learning model. Therefore, if multiple similar observations behave similarly in a black box model it is possible to fit a prediction model on their basis, that would mimic and consequently explain how the original model behaves at that locality (Pedersen & Benesty, 2019).

The surrogate model may take numerous forms such as linear regression, LASSO or decision tree. The only limitation for the chosen type of a model is to be easily interpretable. The accuracy with which the surrogate model explains the black box algorithm's behaviour is known in the literature as a local fidelity (Molnar, 2019). Local fidelity may be used as a metric on the basis of which the optimal type of local model is chosen.

Mathematically, the local approximation of a black box model may be defined as:

$$\hat{g} = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{4}$$

where:

G represents the class of interpretable models,

g is an interpretable model belonging to the class G,

L is the fidelity measure measured with a chosen metric e.g. mean squared error,

f is the black box model for which the local approximation is sought,

π_x is the proximity measure defining a neighborhood of an instance x in which the approximation is sought, and

$\Omega(g)$ is the g model's penalty of complexity (Biecek & Burzykowski, 2020).

While the above formula already shed some light on the theory behind LIME, the application side has been described in an approachable way by Pedersen & Benesty (2019). Firstly for each prediction to explain the observation is permuted n times. Then, the prediction for each permuted observation is run with the usage of the original black box model. Subsequently, the distances from all permutations to the original observation are calculated and converted into similarity score. After selecting the number of features best describing the black box model with $\Omega(g)$, a surrogate model is fitted to the permuted data. While training the local explanatory model the outcome is weighted for a permuted observation by its similarity to the original observation. Finally, by extracting the feature weights from the simple model it becomes possible to use these as explanations for the complex model's local behavior.

In theory, following the above procedure should allow us to explain the mechanics of any black box model for every single prediction. Nevertheless, LIME is still a method in a development phase and as promising as the above statement sounds, it does not always end up being that functional. As a relatively new method, LIME suffers from a number of issues that have not been yet resolved. Firstly, there is not much literature available on the tuning of LIME parameters, especially the width of a smoothing kernel π_x, which for now has to be tuned with a trial and error approach.

Another problem, the robustness of explanatory models created with LIME causes even more concern. In Alvarez-Melis & Jaakkola (2018) the authors show that the explanation provided with LIME for two similar observations may vary significantly. Molnar (2019) also criticizes the stability of LIME, claiming that the results may change after repeating the sampling process.

4 Data and feature extraction

The research has been based not on the prices of properties but the rental costs of real estates instead. There are numerous reasons for choosing this approach. Firstly, the number of research on the prices of house rents is relatively limited in comparison to the house prices, while the results of both may differ significantly (S. Clark & Lomax, 2019). Secondly, obtaining the data in English on the Dutch market, which is a primary interest of the study, is much easier in case of rentals. In terms of data, it has been collected only from rental websites and Google Maps, on the contrary to most of the research mentioned in the literature review, which are based on data received from real estate agencies. One could argue that the web-scraped data may be of lower quality in some aspects. For example, there is no certainty that the rental price shown in the offer has not been negotiated, and is the final rent that the customer agreed to pay. Nevertheless by taking into account the relatively fast-moving nature of the rental market in large cities limiting the negotiating space, making this kind of assumptions seems reasonable. On the other hand, gathering data from the rental websites allows for a more flexible approach in its usage, as the research would not be limited to predefined data set. Moreover, basing the study on accessible data leads to a situation where the model can be easily applied to any other city or area.

Another reason for using only data from rental websites is linked to the theory of hedonic models. One of the most often mentioned characteristics (or even limitations) of hedonic models is the customers' willingness to pay more or less for a product depending on the utility of its attributes, but only the ones that they can perceive. It may be argued that in case of renting an estate, potential customers do less in-depth research than they would if they wanted to buy it. Therefore, it makes a rental offer a primary source of information for customers not only in terms of house characteristics but external factors as well. Nevertheless, it would be naive to assume that people willing to rent a house do not make any research on their own. As one of the most important features of a real estate is its location, Google Maps seems like an obvious tool that in a short amount of time, allows us to gather relatively a lot of knowledge about the neighborhood of a property. Therefore, extracting features from satellite images plays a significant role in this research.

The core of the data gathered for the research comes from one of the leading rental websites in the Netherlands. The main reason for limiting the data source only to one web page is the fact that all rental offers there are presented in English, which is rather uncommon for the Dutch market. Moreover, the website is well structured and allows for a relatively easy process of web scraping. Furthermore, as the web page is administrated by real estate brokers the quality of the data presented in the rental offers is on a very high level. Therefore, the estimated values of the properties advertised on the website are assumed to be accurate and in consequence, should limit the number of outliers in the data. Last but not least, it appears that most rental offers are unique, which is quite a phenomenon when compared to other websites where duplicates appear quite often.

4.1 Initial data

The most crucial part of the data describing the main house characteristics has been web-scraped from the mentioned rental website. Typical variable examples of such data are rental price, type of a house or living area of a property. As far as most of the data in this section according to the nomenclature proposed by Chin & Chau (2003) would be defined as structural, some information such as the name of a street or postal code would match more the Chin's category locality.

Nevertheless, after investigating multiple rental offers it has been found out, that many offers differ in the amount of information fields presented to a visitor of the website e.g. energy rating field is present only in a relatively small number of offers. Therefore, in order to create a reliable data set some variables have been excluded from the process of web scraping.

The process of gathering data has been fully automated with the usage of R and respective packages allowing scrapping the data based on HTML and XML code of the website. Firstly, the links to all the rental offers have been scrapped from the search page with the only filter of the city being Rotterdam. Over 2000 real estate rental offers have been found and crawled since the beginning of April till the end of May 2020.

After performing data cleaning the data set consisting of 1844 unique observations has been created. Moreover, as one of the most important factors in predicting the rental cost of a house is its location, with the usage of Google Maps API, a set of variables have been collected: the distance from a house street to the Rotterdam Central Station and the time it takes to travel the distance by walking, by biking and by public transport. Similarly, the geographical coordinates of each street appearing in the initial data set in the form of longitude and latitude have been gathered.

Table 2 contains a detailed description of the variables present in the initial data set. Out of 1844 observations, the only missing values appear in variables *Construction_ year*, *Bedrooms* and *Bathrooms*. For the last two variables, they do miss only in case of house type being a single room. Therefore, it has been decided to substitute the missing values in these cases with 1. In terms of construction year 721 observations out of 1844 miss this information.

4.2 Image recognition

Simultaneously with scraping typical house characteristics data, images of each rental offer have been collected. With an average of 22 photos per one offer, over 40000 images have been gathered in total. Moreover, as one of the goals of the research is to use image recognition models in order to extract features from the mentioned photos, additional images have been gathered to use them in training classification models. Around 2000 images from other rental offer websites have been gathered, this time however, the data came not only from Rotterdam but other Dutch cities as well. Such an approach allows limiting the potential bias in classification models' accuracy, which would be otherwise caused by manual labeling the part of collected photos.

Table 2: Variables present in the initial data set

Variable	**Type**	**Description**
House.ID	Numeric	Unique ID of a house
Street	Character	Street
URL	Character	URL
Postal_code	Character	Postal code
District	Factor (64 levels)	District
Price	Numeric	Price in euros
Living_area	Numeric	Living area in squared meters
Rooms	Numeric	Number of rooms
Construction_year	Numeric	Construction year
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?
Shower	Factor (2 levels)	Does a property have a shower?
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?
Service_cost	Factor (2 levels)	Are service costs included in the price?
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

Subsequently, the process of extracting features that according to the previous research may turn out significant in predicting housing price has been started. Firstly, the rental offers' photos have been initially analyzed. Although unsurprisingly a majority of photos present the inside of a given estate, it has been found out that most of the offers do feature photos showing the outside of the building and view from the property as well. Therefore, as this study focuses mostly on the external factors impacting the rental price in Rotterdam, the decision to analyze the view from properties has been made. It may be argued to which degree the view from a property is an external, an internal, or a locational factor. Nevertheless, it is rather safe to assume that view is affected by external factors such as the height and density of the buildings in the direct neighborhood, thus making it at least partially externally dependant.

The image recognition process for different types of views has been divided into two sequential convolutional neural net models. The first model aims to filter the property images and classify them whether as outside or inside. The model has been trained in a way that outside label is given to photos of balconies, views from the windows, street images, etc. All the other photos including not only interior but e.g. graphics illustrating the layout of a flat are classified as inside. Even though the added value this classification model brings to the data set and the future prediction model is low, it allows us to simplify the subsequent model.

The goal of a second model is to analyze the outside photos and classify them into four categories:

1. **View on the city:** category featuring photos with relatively unbroken view on the charismatic panorama of Rotterdam
2. **Green view:** category featuring images with a view on a park, a canal or another green area
3. **Enjoyable view:** arbitrary category featuring images with a pleasant, above the average view e.g. photos presenting an unbroken view on a neighborhood, river or open areas
4. **Other:** category featuring all the other images.

Both models have been built with the usage of Keras neural network library with the TensorFlow backend and are based on the same architecture. Firstly, images were scaled to the same size and divided into three sets: train, validation, and test. The first model has been trained with 1500 images while the validation and training sets had 300 photos. The second model was trained on 900 observations while the test and validation sets consisted of 200 images. Subsequently, the train images were augmented by applying width and height shifts, horizontal flips, and shear transformations.

The overall architecture of the models is fairly simple. The first convolutional layer consists of 16 filters and the kernel size applied to the layer has a standard size of 3 x 3 pixels. Additionally to preserve the data same padding has been used. The number of filters has been set to a rather small number as it was expected of the model to capture mostly low-level features of an image e.g. sky in the first model or vegetation in the second one.

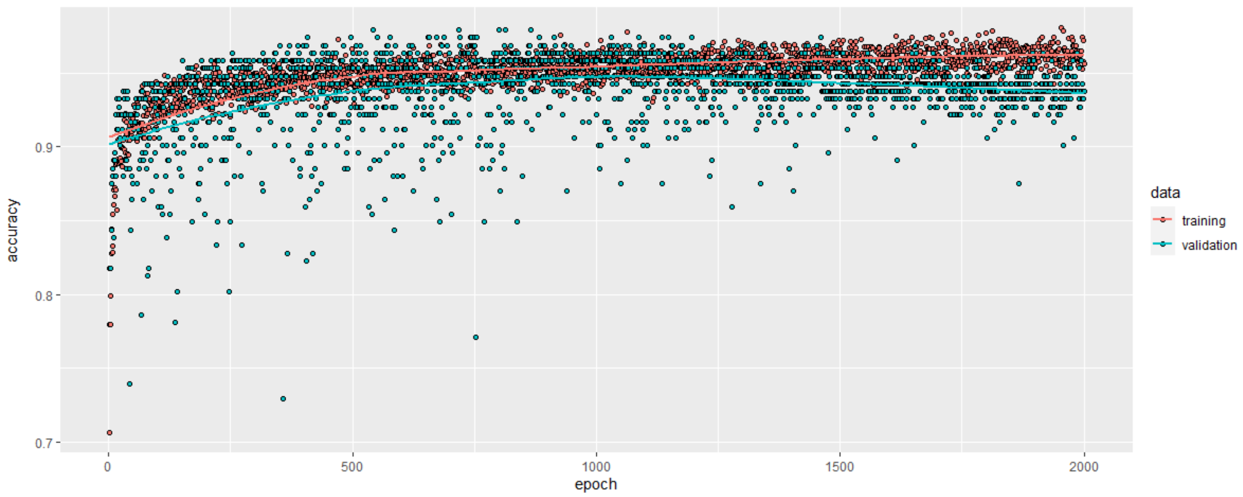

Figure 1: The accuracy of the model versus the number of epochs

Subsequently, the output of the first layer is transferred to the ReLU layer, which after applying the activation function transfers the data to the next convolutional layer. The second convolutional layer differs from the first one only in terms of the number of filters, which according to the best practice has been increased to 32. The reasoning behind this practice is the fact that it is more profitable to increase the number of filters in the deeper parts of the model, where calculations are performed on the initially filtered, less noisy data.

Afterward, the outcome is passed to the next ReLU layer, which subsequently transfers the data to the pooling layer. Except for reducing the size of an input matrix, the pooling layer additionally applies the dropout method with the parameter equal to 0.25 which stands for 25% chance of deactivating each active node. Lastly, the output of the pooling layer is transferred to the dense layer for which commonly used functions have been set in the form of activation function being softmax and loss function being categorical cross-entropy. To ensure the stability of the prediction, the number of epochs which is the number of times the entire dataset is passed forward and backward through the neural net has been set to 2000. With each epoch added, the model has an additional opportunity to adjust its weights aiming at improving the accuracy. Therefore, the higher number of epochs usually leads to better predictions, nevertheless at the cost of increased computational time. As such a compromise between these two aspects had to be found leading to the chosen number of epochs where the quality of predictions became acceptable. The progress of training the first model is presented in Figure 1.

After performing the analysis of the rental offers' photos, the Google Maps have been put to use. Out of 1844 observations in the data set, 578 unique street names have been filtered. For each street, two types of Google Maps images have been collected. The first one, the map at the zoom of 17 (scale used in Google Maps) presents the street and its neighborhood in a 300m radius. The second map at the zoom of 15 shows the surrounding of a street with a radius of 1km. Similarly as

Table 3: Accuracy of the image recognition models

Model	Validation	Test
Outside/Inside	97%	95%
Water body	95%	94%
Park	87%	86%
View	78%	76%

with rental offers' photos, to decrease the accuracy bias additional maps have been scrapped from other cities in the Netherlands.

Following the conclusions drawn from the literature review, it seemed appropriate to analyze the impact of neighborhood features on rental property prices. By taking into account the specific nature of Rotterdam's urban layout, it has been concluded to focus on two categories:

1. **Parks:** category indicating whether a given property lies in a direct neighborhood (300m) of a park or other green area.
2. **Water bodies:** category indicating whether a given property lies in a direct neighborhood (1km) of the Meuse river (the main river in Rotterdam, flowing through its city center) or a large lake.

For each of the aspects with the usage of Google Maps images, a convolutional neural net model has been created. Due to a limited amount of time and resources, the architecture of the model has been copied from the previous models. The models analyzing park and water body proximity have been trained on 500 maps while the test and the validation sets had 100 observations. Moreover, the data augmentation was more limited in case of these models as all the map images from Google Maps keep the same formatting and orientation. The accuracy of all the models described above is presented in Table 3.

While the accuracy of the "Water body", “Outside” and “Park” models is relatively high, the score for the “View” model may cause some concern. However, it has to be remembered that the accuracy is calculated on the photo level, while the accuracy that is the most important for the research lies on the property level. The classification thresholds of the model have been set in a way, that the model is more likely to underestimate the view e.g. by categorizing the “Enjoyable view” image as “Other” than to overestimate it. As each house has multiple photos of the outside, it is sufficient if the model confidently classify just one photo instead of risking the error in case of images being on the verge between classes. The accuracy on the property level is represented in Table 4.

Table 4: Accuracy of the view image recognition model on the property level

Variable	**Test**
City	90%
Green	86%
Enjoyable	76%

4.3 Text analysis

The text analysis part of the research as mentioned in the methodological section, has been limited to a fairly simple information extraction. The first step of the analysis was to identify the most commonly used words in rental offer descriptions. The preparation part included the transformation of all the descriptions into lowercase and removing stop words. Subsequently, the words with the frequency of above 150 appearances have been analyzed. The frequency count is presented in Table 11 in the appendix.

Unsurprisingly, most of the words are related to the previously created variables based on the featured fields on the rental website such as the number of rooms, location, or the surface of an estate. Nevertheless, three additional categories of words which do not depend completely on the structural attributes of a property have been distinguished:

- **Services:** words such as restaurant, shop, cinema,
- **Owner's restrictions:** words indicating the restrictions under which the rent is allowed e.g. pets, income, deposit,
- **Insolation:** words indicating that a property is exposed to sunlight e.g. bright, sunny, south.

Undoubtedly, insolation is an attribute strongly related to the structural attributes of a house. Nonetheless, similar to the view from a property, it is dependent on the height and density of the buildings in the direct neighborhood thus making the variable at least partially external, which is why it has been included in the research. For each of the featured categories slightly different approach of variable creation has been applied.

- **Services** On the basis of Table 11 and common sense, a list of keywords relating to various types of services has been created. The list featured the following strings: shop, restaurant, hairdres, barber, supermarket, bar, school, university, erasmus, markthal, market, cafe, food, gym, and sport. Subsequently, the description of each property has been scanned for the mentioned keywords. The final value for a given observation is the sum of unique services found in its description.

- **Owner's restrictions** based on this category two variables have been created. The first one, *Income* refers to the financial requirements for renting a property. In most of the observations containing keywords income and/or guarantor, the financial stipulation is set to three times the amount of rental cost. Therefore, the final variable income takes the form of a dummy, indicating whether a financial condition is mentioned or not.

  The second variable in this category refers to pet restriction. Each description has been scanned for the presence of keyword pet. After finding the position of the keyword, the window of one word before and two words after has been searched for strings "no", "n't", "aren't". As such the most popular grammar forms "No pets", "pets are not allowed" and "pets aren't allowed" have been included. If in the window no contradictions are found, the variable takes the form of 0, otherwise is given the value of 1. For observations in which the keyword "pet" is not present, the value is set to 0.

- **Insolation** Each description has been scanned for keywords light, bright, sun, and south. In case of the presence of at least one of the first three keywords, the variable insolation takes the form of 1. However, in case of the word south, additional requirements have been set to distinguish descriptions relating to the south in the context of location and direction of the flat. As in the context of the former, the word south is usually in the direct neighborhood of the word Rotterdam ("south Rotterdam", "Rotterdam south", "south of Rotterdam"), the window of one word before and two words after has been applied. If the word Rotterdam is present in the window, the variable takes the form of 0, otherwise it is set to 1.

4.4 Summary

The final dataset consists of 27 variables which descriptive statistics are presented in Tables 5 and 6. According to the data, an average property rented in Rotterdam has the living area of 73 squared meters, divided into three separate rooms. While the average rental price for such a property equals 1295€, the variable price is characterized by its wide range with a minimum of 295€, the maximum of 4995€ with the standard deviation of 542€. Another continuous variable that draws attention is the variable *Services* extracted from the descriptions of rental offers. By taking into account the mean of less than two services mentioned in the average offer and the possible maximum of fifteen services, it seems that the owners do not share information about the neighborhood as often as it was initially deemed. Consequently, the variable can be seen as of rather low quality and may not be particularly useful in the modeling process at least in its original interpretation. As numerous descriptions do not feature information about the nearby services, the potential significance of the variable could be only interpreted directly; as the number of services mentioned in the description, not as the approximated number of services in the property's neighborhood.

In terms of dummy variables presented in Table 6 the majority of features represented by them are rather uncommon for an average property in Rotterdam. The only characteristics which appear in the majority of real estates are furnishing and proximity to some water body. The overall frequency of the variables is little wonder in most of the cases and generally seems reasonable. Nevertheless, a problem has been noticed with variables *Shower* and *Bath*, which even if summed up, do not exceed the total number of observations. This implies that there are real estates without access to any of these objects. As such a situation seems extremely unlikely, it has been concluded that the owners simply did not treat shower as a characteristic worth mentioning in numerous rental offers. Therefore, the variable *Shower* has been excluded from further analysis while the variable *Bath* has been kept in its original form.

Table 5: Descriptive statistics 1

Variable	**Min.**	**Mean**	**Median**	**Max**	**St. dev.**
Price	295	1347.00	1295	4995	542.38
Living_Area	6	75.61	73	935	41.29
Rooms	1	2.72	3	11	1.13
Bedrooms	1	1.20	1	5	0.81
Time_walking	123	2275.00	1754	9037	1512.49
Time_biking	32	704.70	540	2530	458.37
Time_public	109	1085.70	996	3085	530.86
Longitude	4.41	4.48	4.48	4.58	0.027
Latitude	51.87	51.92	51.92	51.98	0.017
Services	0.00	1.82	1.00	9.00	1.863

Table 6: Descriptive statistics 2

Variable	**Not present**	**Present**
Balcony	1152	692
Garden	1638	206
Storage	1382	462
Garage	1759	85
Shower	1291	553
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

Before moving to the analysis part, this seems like a proper moment to provide the reader with the reasoning behind the choice of variables used in the research. Let's recall that in the literature review a handful of external aspects impacting the value of a real estate has been described e.g. air and noise pollution, while they did not find their place in the final data set. The initial analysis of air pollution shown, that as a windy city located nearby a sea, Rotterdam in general does not suffer greatly from air quality problems. Moreover, air quality in the city seems to be rather homogenous, without any districts standing out notably from the average. Furthermore, as stated in the theoretical part of the study, according to the hedonic modeling theory, the customers' willingness to pay for a particular good is based only on the attributes that they may perceive. Therefore aspects such as air quality or environmental risks seemed unlikely to affect the rental price. In the end, it is rather uncommon for the potential tenants to conduct research on these areas. Instead, it is much more probable that the evaluation of a property is usually based on more easily noticeable features such as the proximity to a park or the view from the window.

Nonetheless, as in any other big city, the problem of noise pollution should not be omitted. In this study, an attempt of addressing this problem has been made with the usage of Google Maps. An image recognition model has been trained to derive if a given property lies in a direct neighborhood of any of Rotterdam's busy roads. Unfortunately, the accuracy of the initial model has not been satisfying, and therefore, with the limited time and resources taken into account, the idea has been abandoned.

Due to limited data sources, it is a difficult task to include in the study all the structural and locational attributes studied in the previous research. However, numerous aspects are indirectly included in the analysis through the usage of geographical coordinates which serve as a proxy variable. On the one hand, it is not possible to quantify with this approach the value of the neighborhood characteristics such as ethnicity in the area or criminality rate. On the other hand, it is rather safe to assume that if these characteristics do impact the rental price in some of the Rotterdam's districts it will be reflected in the regression model through locational variables.

5 Analysis and results

The analysis has been divided into two parts. Firstly, multiple regression models have been created. After comparing the accuracy of the models, the one with the best performance served as a base for the explanatory analysis of the housing market in Rotterdam, which is presented in the second part of the section.

5.1 Regression

Before proceeding with the regression analysis, the correlation of continuous variables in the data set have been checked. The result in the form of a correlation plot is presented in Figure 2. At the first glance, it may be seen that there are two aggregations of linearly correlated variables. Living area, the number of rooms and bedrooms are unsurprisingly highly correlated with each other. In order to diminish the problems caused by multicollinearity of the explanatory variables in the regression model, the variable *Bedrooms* has been dropped. Nevertheless, even with the high correlation taken into account, it is safe to assume that the number of rooms and living area are crucial characteristics of a property and should not be excluded from the analysis.

The second aggregation of correlation consists of features describing the localization of a property. While there is no point in using all three measures of time needed to travel the distance to Rotterdam Central Station in the regression, their correlation provides some valuable insights. As all the variables are similarly correlated it may be assumed that the biking infrastructure and public transport are evenly distributed over the area of Rotterdam. Therefore, after additionally taking in to account the extraordinary popularity of biking in Rotterdam, the variables *Time_ walking* and *Time_ public* have been excluded from further analysis.

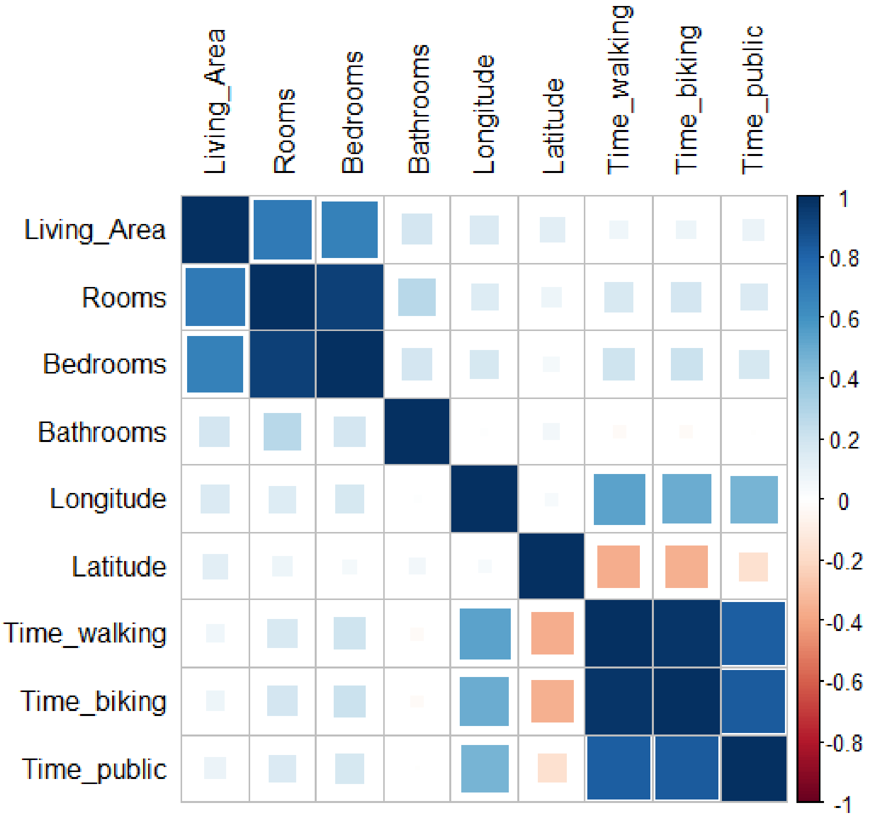

Figure 2: Correlation plot for numerical variables.

Lastly, even though the longitude and latitude are not linearly correlated with the variable *Time_ biking* it is safe to assume that in the non-linear model, the multicollinearity problem would appear. Therefore, the decision on choosing only one of the locational approaches had to be made. Overall, in case of classic hedonic models, the variable *Time_ biking* seems more appropriate than the geographical coordinates, which are not expected to perform well in linear conditions. The linear model allows geographical variable only to indicate through the sign of the coefficient in which part of the city (e.g west or east in case of longitude) the rental costs are on average higher. However, it is completely possible that while the majority of expensive estates are located on the western side of a city, eastern suburbs feature some luxurious districts. Such a phenomenon could not be captured through the usage of geographical coordinates in the traditional OLS regression.

On the contrary, in non-linear models such as decision trees, the combination of longitude and latitude is expected to capture such patterns for all the possible geographical directions. On the other hand, by including the geographical coordinates in the non-linear model, the variables extracted from Google Maps: *Water_ body* and *Park* would coincide, which consequently could decrease their importance and impact on prediction accuracy.

As such multiple models have been created and compared to determine the most effective one in terms of prediction accuracy and reliability. For all the models the same train and test sets have been used, where the former consisted of 70% and the latter of 30% of observations. The comparison between the models has been based on the RMSE metric calculated on the test set.

The initial models have been based on a classic hedonic pricing approach in the form of an OLS regression. The first two models aimed to determine which locational approach performs better in linear conditions: *Time_ biking* or longitude and latitude. The regression models except for locational variables consisted of the following predictors: *Living_ Area*, *Rooms*, *House_ Type*, *Bathrooms*, *Balcony*, *Garden*, *Storage*, *Garage*, *Bath*, *Lift*, *Toilet* and *Furnished*. It may be recalled that all the mentioned variables are part of the initial data set, therefore they have been directly scrapped from featured fields of rental offers website. As such, this set of variables has been used as a benchmark in addressing the research question asking if image and text features impact the accuracy of the hedonic model. For the simplicity of visualization, the set of variables used will be referred to as basic housing attributes (BHC) in the following parts of the analysis.

The comparison of the models proved the slight superiority of *Time_ biking* over geographical coordinates in linear conditions. The model with variable *Time_ biking* surpassed its counterpart in terms of both RMSE (321 vs 329) and R^2 (0.59 vs 0.56). Therefore, it has been decided that the *Time_ biking* model will serve as a benchmark for more complex linear models. Subsequently, an OLS model featuring all the BHC and features extracted from images and descriptions have been built. The coefficients of the model are presented in Table 7. Out of basic housing attributes living area, the number of rooms and bathrooms, presence of garden, garage and bath as well as the furnishing turned out to be significant, positive predictors of a rental cost prediction. Moreover, it has been found out that the single rooms are 194 euros cheaper than the flats, leaving all the other variable values constant. Lastly, the farther from the city center the property is located, the less expensive its rent is.

Table 7: The coefficients of the linear model empowered by image and text features

Variable	**Estimate**	**Pr(>|t|)**
(Intercept)	420.62	< 0.001
Living_Area	6.06	< 0.001
Rooms	88.88	< 0.001
House_Type House	39.58	0.422
House_Type Room	-194.04	< 0.001
Bathrooms	238.40	< 0.001
Balcony Present	-5.36	0.817
Garden Present	108.09	0.002
Storage Present	-39.66	0.158
Garage Present	252.32	< 0.001
Bath Present	56.31	0.027
Lift Present	36.65	0.211
Toilet Present	-11.57	0.608
Furnished Yes	81.46	< 0.001
View_on_the_city 1	122.41	< 0.001
Park 1	5.26	0.822
Pets_not_allowed 1	-60.26	0.160
Enjoyable_view 1	8.29	0.711
Green_view 1	29.48	0.314
Water_body 1	47.67	0.022
Income 1	-159.12	< 0.001
Services	13.22	0.125
Insolation 1	106.76	< 0.001
Time_biking	-0.26	< 0.001

In terms of features extracted from images and descriptions, the view on Rotterdam, the proximity of a water body (either the Meuse river or a large lake), and insolation positively impact the rental price of a house. On the other hand, the financial requirements for renting a property are linked with an average drop in rental price of 159 euros. With regards to the overall performance of the model, RMSE decreased in comparison to the previous OLS model to 308, while R^2 increased by 0.02 to the level of 0.61. Therefore, the initial analysis provides strong foundations to claim that the additionally extracted features do positively impact the accuracy of a hedonic model. Nevertheless let's recall that according to Butler (1982) and Mok et al. (1995) the under-specification of variables is more welcome than the over-specification in case of hedonic regression models. Therefore, the last linear model has been built, this time however the insignificant features extracted from images and descriptions have been dropped. The performance of the last model, almost did not change in comparison with its predecessor as its R^2 and RMSE scores were equal to 0.61 and 309 respectively. Moreover, only slight changes in the size of coefficients has been found which may be seen in Table 8.

Table 8: The coefficients of the final linear model

Coefficients	**Estimate**	**Pr(>|t|)**
(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 Yes	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

Nevertheless, as mentioned previously in the methodological part of the study, an OLS model is not fully reliable if it does not meet its five main assumptions. Therefore, the verification of the linear requirements has been performed. Firstly, the assumption of linearity of the data has been checked with the usage of plot comparing residuals of the model with fitted values presented in Figure 3. For smaller predictions, the residuals are fairly equally located around the 0 line. Nevertheless, by following a red line (showing the average values of residuals at given fitted values) the decreasing trend may be noticed. This proves that the relationships in the data are not perfectly linear, as for larger predictions the model underestimates the values of a property. Moreover, on the basis of Figure 3 it may be concluded that the model suffers from a heteroscedasticity, as the larger the fitted values are, the larger the spread of residuals becomes.

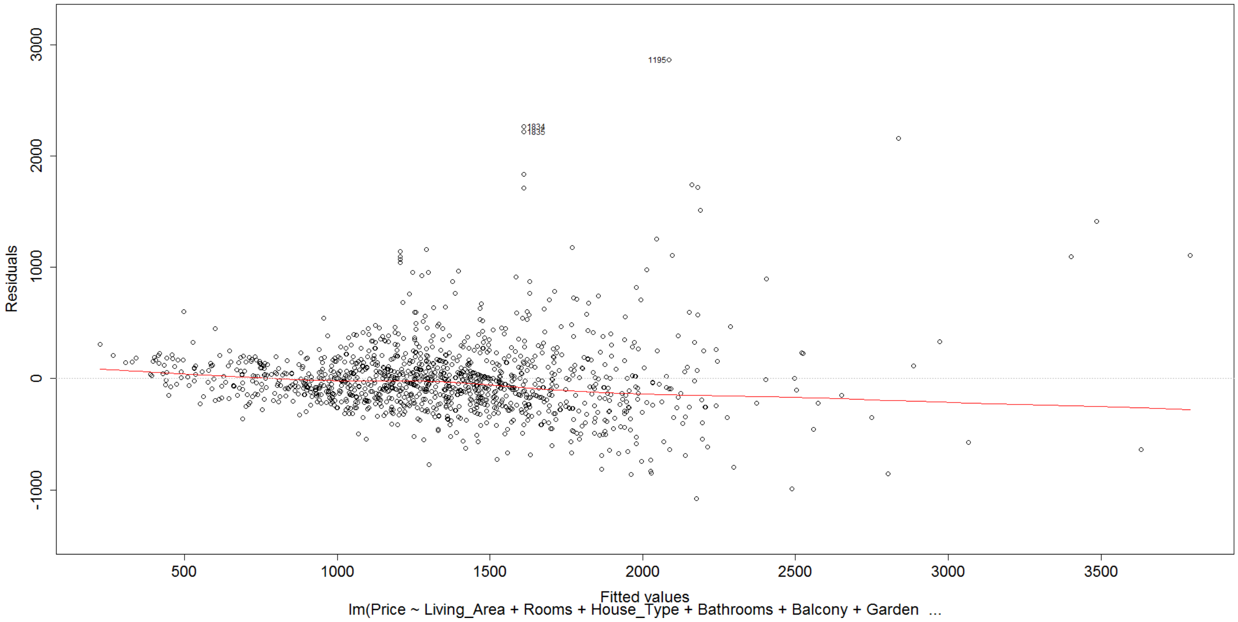

Figure 3: Residuals versus Fitted Values

Afterward, the multicollinearity in the model has been checked with the usage of Generalized Variance Inflation Factor (GVIF), a variation of VIF, which is applicable also for categorical variables. While the rule of thumb in interpreting the VIF is not to have a single value exceeding 5, in case of generalized metric it is $GVIF^{(1/(2*Df))}$ that should not reach that threshold (Buteikis, 2019). Therefore, the results presented in Table 12 in the appendix show no serious problem with multicollinearity in the model. The next assumption, lack of auto-correlation in residuals of the model is met, as seen in Figure 8 in the appendix. The last assumption, claiming that residuals should be roughly normally distributed is violated in this particular model. In the QQ plot presented as a Figure 9 in the appendix the heavy tails may be noticed, indicating that the distribution of residuals is not of gaussian type.

Two main approaches may be taken to overcome the problems with violated linearity assumptions. The first one, the data manipulation using methods such as Box-Cox transformation has been frequently applied in other hedonic research papers. While this approach is not a primary focus of the study, some standard data transformations have been applied in order to verify if they would help with meeting the above assumptions. Nonetheless, neither logarithmic nor Box-Cox transformation led to a notable improvement in the area. In both cases there were still assumptions violated which example may be seen in Figure 10 in the appendix. Moreover, the accuracy of the transformed models have dropped in comparison to the original counterpart.

As such, more attention has been given to the second approach: addressing the above problems through the usage of a non-parametric model, that does not have such strong assumptions as the OLS regression. Moreover, as discussed in the previous sections of the paper, applying random forest or any other advanced machine learning method, allows capturing more complex patterns in the data which consequently, improves the model's performance in terms of accuracy. However, as the linear models did perform relatively well in this study, and their assumptions were not extremely violated, the random forest models have been built in a similar manner.

In total five random forest models have been built. Before proceeding to their description, it is crucial to recall, that random forest is a method based on bootstrapping. This method, brings an essential problem, in the form of its randomness. By default in R, setting the random number generator does not impact the results of bootstrapping which is always random. Consequently, even with random number generator set, the results of random forest may slightly change in each run. Therefore, all the results shown below are presented in the form of average model's accuracy, calculated on the basis of 20 runs.

The first two models, similarly as in case of an OLS regression, aimed to distinguish which locational approach is more suited: *Time_biking* or geographical coordinates. On the contrary to linear models, the latter approach indisputably proved to be superior over its counterpart. The third model, consisted of basic housing attributes (BHC), geographical coordinates and features extracted from images and text, that turned out significant in the linear conditions. The fourth model, shared the same predictors with the third one, however instead of traditional random forest, conditional random forest was used. The fifth model was created with the sole goal of being used as a benchmark and consisted only of three predictors: Living_area, Longitude and Latitude. The performance of all the created models is presented in Table 9.

Overall, the random forest model including the significant variables extracted from images and texts turned out to be the most accurate model. Nevertheless in order to formally address the main research question, the difference in accuracy between this model and the random forest that did not use the mentioned predictors has been statistically tested. As both samples, consisted only of twenty observations with each one representing model's accuracy in a given run, the non-parametric Mann-Whitney U test was performed. It resulted in a p-value of 1.451e-11 implying, that there is a significant difference between the accuracy of both models that is not caused by the randomness of the random forest method.

Table 9: The summary of regression models

Model	**R Squared**	**RMSE**
OLS: BHC + Long|Lat	0.56	329
OLS: BHC + Time_biking	0.59	321
OLS: BHC + Time_biking + Image and Text Variables	0.61	308
OLS: BHC + Time_biking + Significant Image and Text Variables	0.61	309
RF: BHC + Time_biking	0.70	258
RF: BHC + Long|Lat	0.71	246
RF: BHC + Long|Lat + Significant Image and Text Variables	0.74	240
CRF: BHC + Long|Lat + Significant Image and Text Variables	0.65	290
RF: Living_area + Long|Lat	0.69	249

5.2 Explanatory analysis

The results presented in the previous section, provide us with a fairly reasonable conclusion that the (semi)external factors do impact the rental price of a property in Rotterdam to a small extent. Before proceeding with the more in-depth explanatory analysis of the random forest model, its variable importance based on the mean decrease in accuracy has been calculated. The results presented in Figure 11 in the appendix indicate that the model's accuracy is mostly based on the variables living area, the number of rooms and the geographical coordinates of the real estate. While such results are definitely not surprising in the context of the housing market and especially its rental subset, the variable importance does show that the other predictors, both external and structural, also contribute to the final prediction of the model.

Therefore, the last step in the research aims to quantify this impact and determine how the estimation changes depending on the model used. In case of linear model, estimating the attributes' effect on the price of a good may be easily done by analyzing the coefficients of an OLS regression. This cannot be said about the random forest model, where in order to determine the effect of a single predictor on the prediction, slightly more advanced methods have to be used. As such, in order to derive the magnitude of the mentioned effects, partial dependence function has been calculated for the majority of variables used in the model. Table 10 presents the comparison between the values obtained with partial dependence functions and the coefficients of the final linear model.

Table 10: The comparison of variables' impact on the prediction between the 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

The results of the comparison are somewhat surprising. While both the random forest and the OLS regression models agree on the sign of variables' coefficients, there are notable differences in their magnitude. For the majority of predictors, especially for *Garden*, *Garage*, and *Income*, their impact on the prediction is much lower in case of a random forest model. The only variable which the random forest model evaluated as more influential than the OLS regression did is the *House_ Type Room*.

Apart from the categorical variables presented in Table 10 the partial dependence has been also applied to the numerical variables. Probably the most interesting conclusion may be drawn from the analysis of the living area presented in Figure 4. On the contrary to the OLS regression where each additional squared meter of the property has been connected to the increase in rental price of 6.09€, the dependence between both variables in the random forest has been found as not constant. For the properties with the living area between 6 and 136 meters, the average value of each additional squared meter has been estimated to 7.26€. Surprisingly, this value drops drastically to 1.65€ after reaching the threshold of 136 squared meters, in order to eventually rise again to 5.54€ after passing the threshold of 191 squared meters. However, the average price of each squared meter calculated based on the whole set is equal to 5.70€ which is not that different from the OLS coefficient.

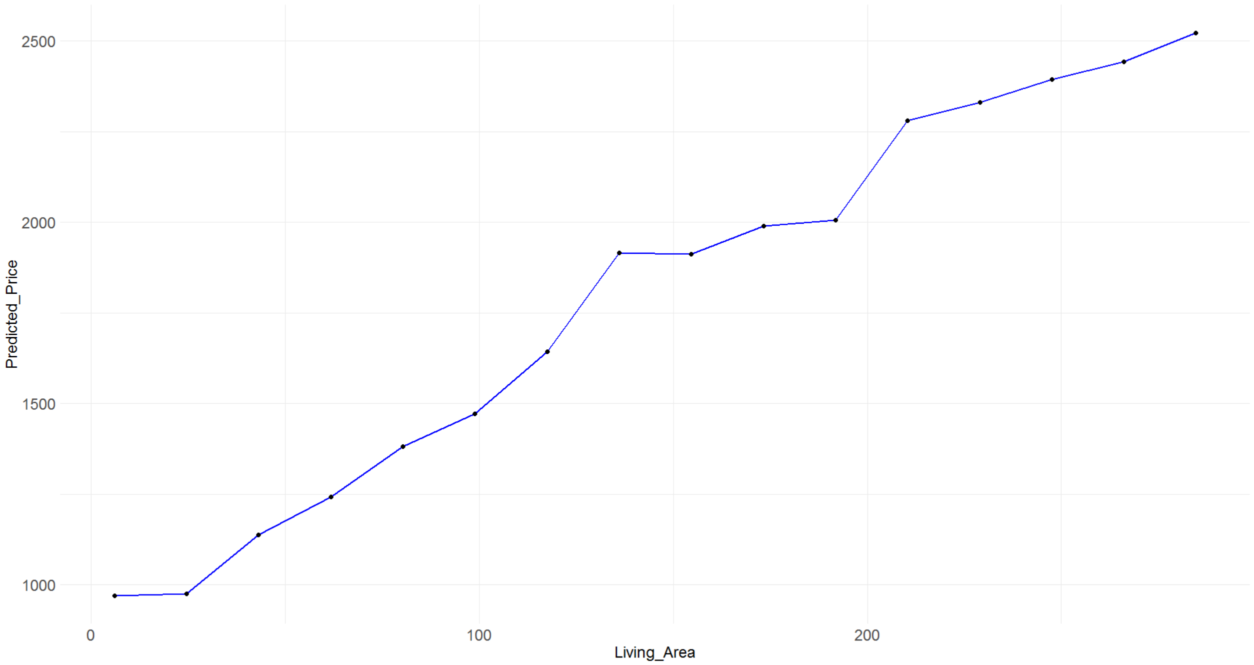

Figure 4: The Partial Dependence Plot of Living Area

Similarly to the living area, the number of rooms also has a comparable value between two models. On the basis of partial dependence function it has been concluded that each additional room increases on average the rental price of a property by 87.16€. However it only applies until the threshold of seven rooms after which no further increase in the price has been noticed. When compared to the OLS coefficient of 89.80€ the difference again is quite negligible and is definitely much smaller than in case of categorical variables. This phenomena suggests that the differences between both models may be partially caused by the nature of random forest which in general favors the continuous variables in terms of their impact and importance. Nonetheless, if both living area and the number of rooms do share similar coefficients across the models, the drop in the importance of categorical variables in the random forest has to be compensated by other predictors.

The difference in the locational approach of both models at least partially explains the presented behavior. In the random forest model longitude and latitude have been used, on the contrary to the OLS model where *Time_ biking* variable is present. As shown in the previous section, in non-linear conditions geographical coordinates perform much better than the distance to the city center. As such, it may be suspected that as longitude and latitude are much more functional in the prediction (as proxy variables they indirectly reflect numerous aspects that *Time_ biking* does not) they affect the other predictors in the model to a further degree than the *Time_ biking* does in the OLS case.

There is no certainty in establishing which of the models describes better the real-life values of properties' characteristics. However, the 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 do prompt to conclude, that the obtained coefficients may be closer to reality than the ones of the OLS regression, even with the drawbacks of the partial dependence function taken into account. This claim however implies that the OLS model may have a tendency to overestimate the value of the less important structural attributes which is caused by its incapability of capturing the locational aspects well.

For the most part, the results of the random forest regression model come to similar conclusions as previous research presented in the literature review part. The most important attributes of a property in terms of its rental price are the living area, the number of rooms, and the general location in the city. Nevertheless, on the contrary to numerous studies, the proximity of parks or other green areas does not seem to affect the rental prices in Rotterdam. Moreover, only a small increase in the price of 10.88€ has been associated with the proximity of any major water body. As such it may be concluded that the importance of location is caused by some other aspects contained in the geographical coordinates which is an interesting aspect to be studied in the future.

Additionally, it has been found out that the decisions made by landlords are not important in the prediction. Neither pet prohibition, nor the number of services mentioned in the property description impact the rental price, while the income requirement has only a small negative impact of -15.44€.

In regards to additional structural attributes of a property, a garage seems like a desirable perk among Rotterdam's residents. On average they are willing to pay for it 104.96€. Similarly, the existence of a bath in a property is associated with an increase in rent by 37.59€. However many attributes which in accordance with the summary provided in Table 1 should theoretically increase the rental price, have been found as insignificant in case of Rotterdam. These attributes are storage room, lift, separate toilet, and balcony. Moreover, gardens have been evaluated with only 9.53€ which stands in strong contrast with the results of the OLS approach. Furthermore, it seems that for the tenants in Rotterdam the furnishing has very little value as on average the rent of fully furnished properties is only 57.12€ more expensive.

Previous research indicate that the view from a property impacts the housing price positively as long as the view is unbroken and features some enjoyable objects such as sea, river, or hill. On the other hand, obstructed views do have a tendency to lower the value of a property (Chin & Chau, 2003). The insignificance of variables *Green_view* and *Enjoyable_view* shows that even the view that subjectively has been rated as above the average, it does not suffice for an average Rotterdam's tenant to pay a premium for it. In case of this study it appears that in order for a view to have an impact on the predicted rental price, it has to feature a spectacular panorama of Rotterdam. Even though the value of the view on Rotterdam has been derived with the usage of partial dependence function, the more detailed analysis of the variable, together with *Insolation* and *Water_body* is presented down below in the paper.

Even in his early work Rosen (1974) argued that the marginal willingness to pay for the attribute of a good changes for consumers, depending on their (nonlinear) budget constraints and preferences. Therefore, not only it further supports the decision of using a non-linear machine learning model which is the random forest, but it also encourages checking how the hedonic price of household attributes varies in Rotterdam. Nevertheless, due to the globality of a partial dependence function, the analysis is not possible with this particular method. Instead, Local Interpretable Model-Agnostic Explanations (LIME) have been applied.

As mentioned in the methodological part of the paper, there is not much literature on parameter tuning for LIME. As LIME is a relatively computationally expensive method, the tuning has been performed with a trial and error approach, resulting in the following parameters:

- Local interpretable model type: decision tree
- The width of smoothing kernel π_x: 2
- Distance function: Manhattan
- Number of explained features: 18.

For each observation in the train set, the local surrogate model has been built, allowing an approximation of the impact of each random forest's predictor on each prediction separately. However, as for some observations, the explanatory model featured extremely low R^2 value, only the observations with the R^2 of above 0.3 were taken into account in the further analysis. Figures 5, 6 and 7 below represent the changes in the value of housing attributes depending on the rental cost of properties.

The results of LIME analysis indicate that indeed the hedonic price of an attribute may not be linear in many cases. For example, the value a view on the city presented in Figure 5 for the flats/rooms with rent below 1000€ stays at the level of around 60€, in order to rise above 100€ for the properties costing more than 1500€ per month.

A similar growing tendency has been noted for the insolation of a property. In this case however, the hedonic price of an attribute lowers after reaching the rental cost threshold of 1500€. Such a behavior may be explained by the fact that in the data set detached houses and larger apartments start appearing after that threshold. As it may be expected from such properties to be exposed to more than one geographical direction, the more sunny such places usually are by default.

The implicit price of proximity to a water body presented in Figure 7 also has a rising pattern. While similarly to the partial dependence results for the most common, relatively cheap properties the value is expected to be below 20€, the value of the trait reaches 50€ for more expensive estates.

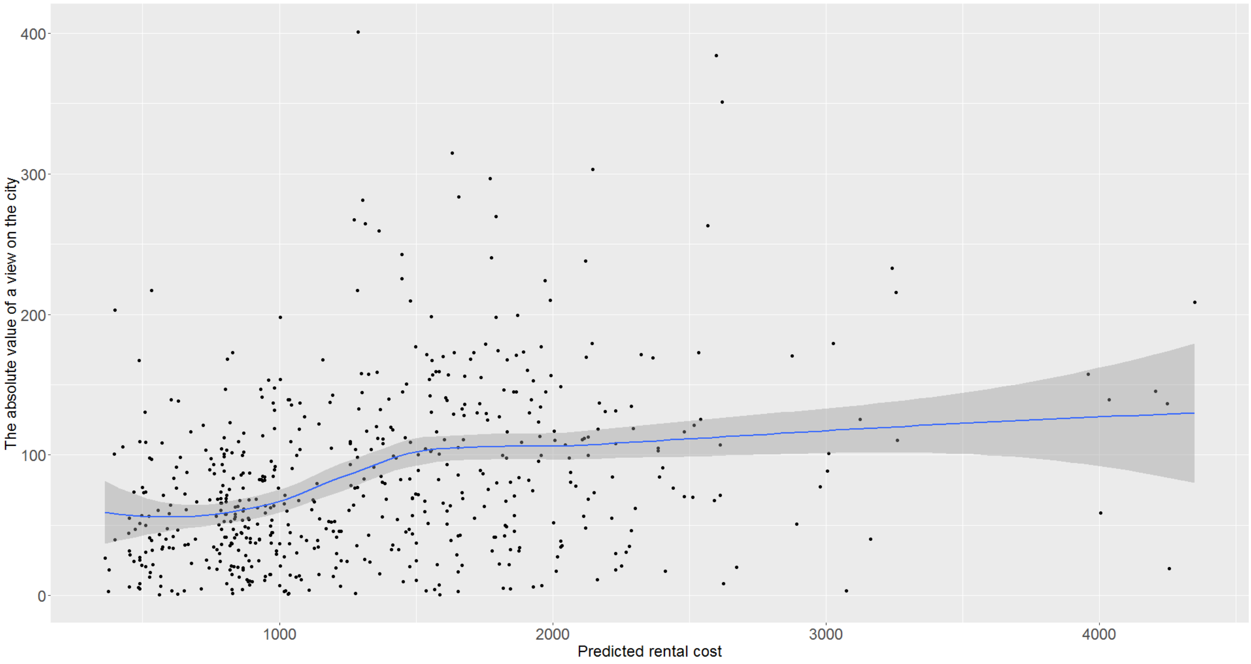

Figure 5: The absolute value of a view on the city versus predicted rental cost

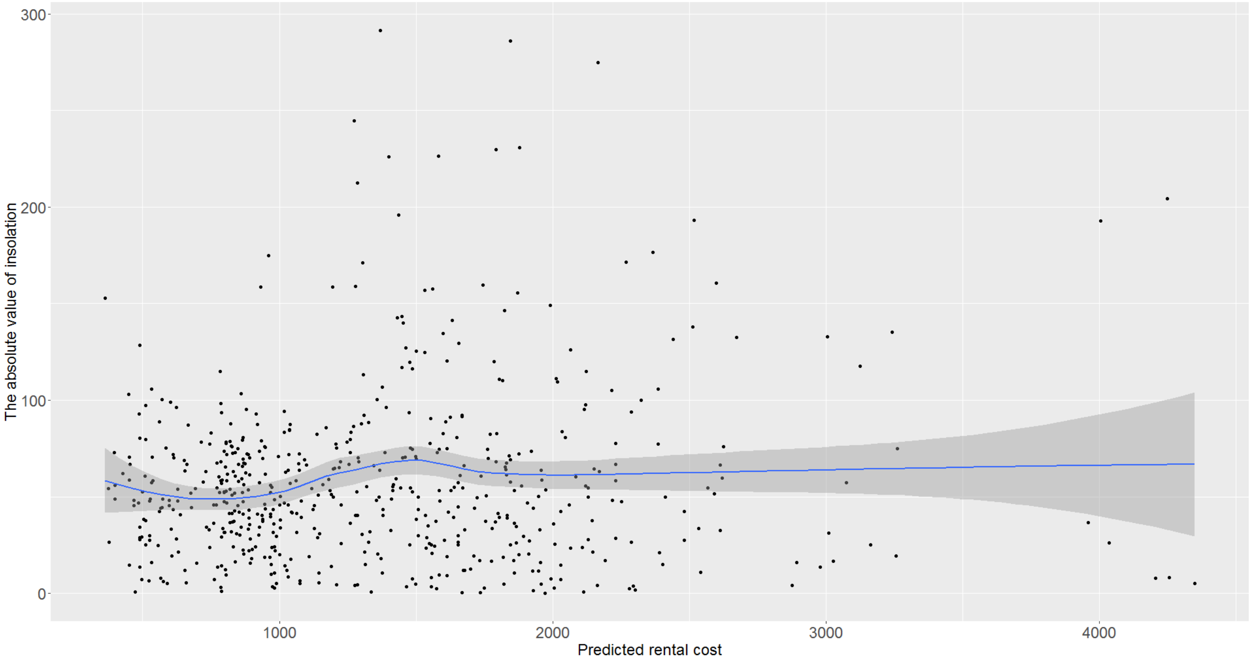

Figure 6: The absolute value of insolation versus predicted rental cost

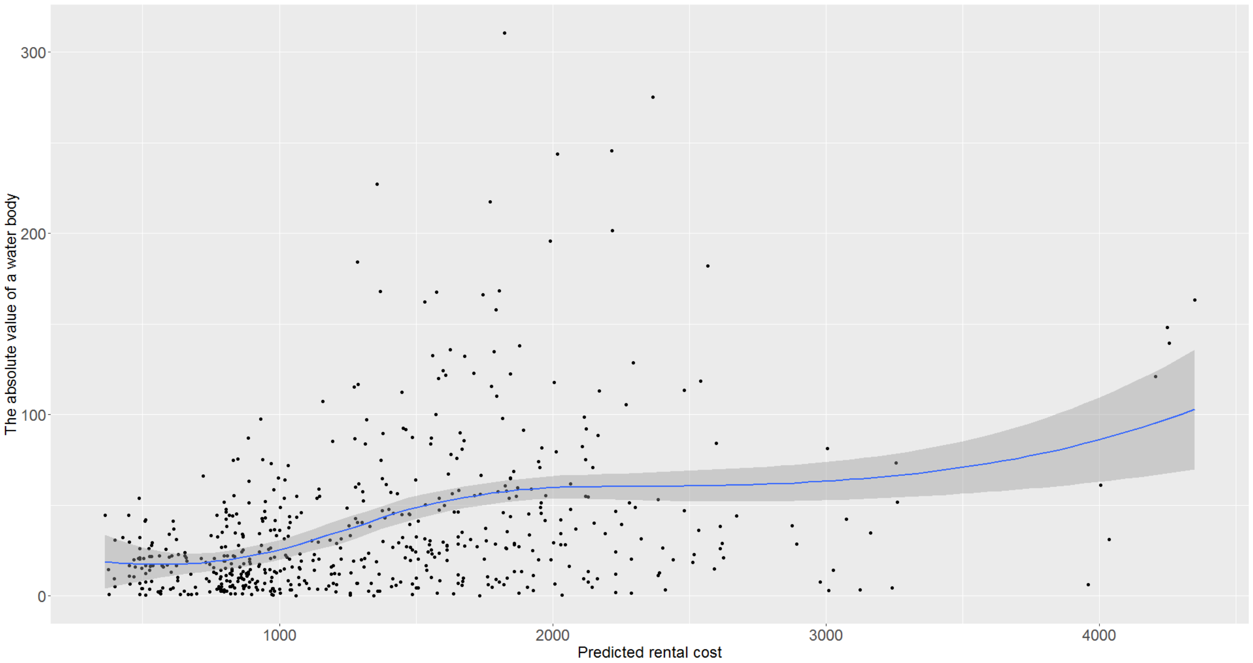

Figure 7: The absolute value of water body proximity versus predicted rental cost

Nevertheless, if we treated the rental cost of a property as the value of a budget constraint function, the interpretation of the figures could be linked to the hedonic pricing theory of Rosen (1974). As according to Rosen, hedonic (implicit) price of an attribute depends on the function of utility level, budget constraint, consumers preferences and other variables such as age or education, what the graphs above represent, is in fact a simplified example of such function with just one predictor used. Unfortunately, it is nearly impossible to gather data relating to matters such as age or education level of tenants without cooperating with real estate brokers or municipality. Therefore, deriving a more advanced hedonic price function for the given attributes was not feasible in this research, however it is an interesting aspect to be analysed in the future studies.

6 Conclusions

For almost 50 years hedonic price models have been extensively used in numerous research. However, despite their undeniable popularity, in most cases, the hedonic models have been based on relatively simple regression methods and tabular data. This study proves that such an approach can be successfully developed with the usage of complex data sources and advanced machine learning algorithms. The variables extracted from images and descriptions of rental offers allowed to improve the performances of both traditional hedonic regression and random forest. Moreover, an accuracy comparison between the models revealed an undeniable superiority of the decision tree-based method.

The problem with the interpretability of the black box model has been addressed with model-agnostic methods which allowed the comparison of variables' importance among the models. As long as the models agreed on some of the most important predictors namely the living area and the number of rooms, the locational aspect has been more emphasized by the random forest model. The results have also revealed that the OLS regression model, when compared to the random forest, is more likely to overestimate the value of nonessential structural attributes. Lastly, with the innovative usage of LIME, it has been found out that the hedonic prices of housing attributes are not likely to be constant and depend on the total value of a property. Such results seem to be in accordance with the original hedonic theory published by Rosen (1974), who argued that in general, the nonlinearity between the price of goods and their inherent attributes is likely to happen.

7 Appendix

Table 11: Frequency of the words appearing in the descriptions

Word	Frequency	Word	Frequency	Word	Frequency
apartment	3052	space	424	shared	221
floor	2194	month	420	contact	220
rotterdam	2043	property	416	windows	219
bedroom	1922	facilities	400	short	217
kitchen	1919	ground	395	erasmus	215
located	1752	income	389	service	213
living	1715	building	386	neighborhood	211
spacious	1551	shopping	371	furniture	210
rent	1440	m2	370	lovely	210
bathroom	1401	hob	368	3rd	207
equipped	1265	metro	365	students	207
furnished	1206	tram	363	master	205
access	1150	layout	357	accessible	204
toilet	1072	microwave	355	upholstered	203
balcony	1006	light	354	south	202
shower	993	price	352	terrace	201
center	917	dining	344	requirement	199
entrance	908	rear	342	basin	198
house	893	period	340	combi	197
city	806	deposit	337	derived	194
bedrooms	799	street	333	maximum	194
distance	780	corner	327	details	192
beautiful	764	garden	321	rights	192
public	725	de	312	extra	191
double	717	bright	307	free	186
walking	715	luxury	307	tiled	185
hall	688	excluding	306	bike	183
washing	673	electricity	302	popular	183
transport	661	private	302	person	180
bed	655	completely	295	rented	180
shops	647	dryer	295	offers	179

Continued on next page

Table 11 – *Continued from previous page*

Word	Frequency	Word	Frequency	Word	Frequency
storage	641	extractor	293	staircase	179
station	619	apartments	290	sunny	179
district	601	laminate	286	connection	178
rental	591	studio	286	stay	178
central	581	minimum	276	cozy	177
front	580	cupboard	275	hallway	177
machine	579	washbasin	273	information	177
separate	564	hood	272	accommodation	172
minutes	563	situated	270	kralingen	170
modern	555	including	264	glazing	169
view	536	wardrobe	264	basement	168
walk	524	monthly	263	bus	168
gas	509	door	260	renting	168
months	509	centre	256	cleaning	162
water	506	built	255	entire	162
fridge	504	ca	254	mailboxes	162
renovated	493	offer	252	design	161
sink	493	close	249	pets	161
parking	488	costs	247	text	161
tv	484	bath	246	reached	160
heating	457	stairs	243	refrigerator	160
restaurants	457	mirror	240	west	160
freezer	449	reach	228	quiet	159
dishwasher	444	appliances	225	home	158
oven	440	luxurious	225	allowed	157
nice	436	suitable	223	radiator	156
approx	433	closet	222	elevator	155
location	428	complex	221	closed	154
internet	425	roads	221	doors	154

Table 12: Generalized Variance Inflation Factor for linear regression

Variable	**GVIF**	**Df**	**GVIF(̂1/(2*Df))**
Living_Area	2.233044	1	1.494337
Rooms	2.496504	1	1.580033
House_Type	2.359207	2	1.239343
Bathrooms	1.136956	1	1.066282
Balcony	1.303721	1	1.141806
Garden	1.22419	1	1.106431
Storage	1.524984	1	1.234902
Garage	1.150035	1	1.072397
Bath	1.277762	1	1.130381
Lift	1.433762	1	1.197398
Toilet	1.270875	1	1.127331
Furnished	1.165545	1	1.079604
View_on_the_city	1.066187	1	1.032564
Water_body	1.086581	1	1.042392
Income	1.074908	1	1.036778
Light	1.064132	1	1.031568
Time_biking	1.237806	1	1.112567

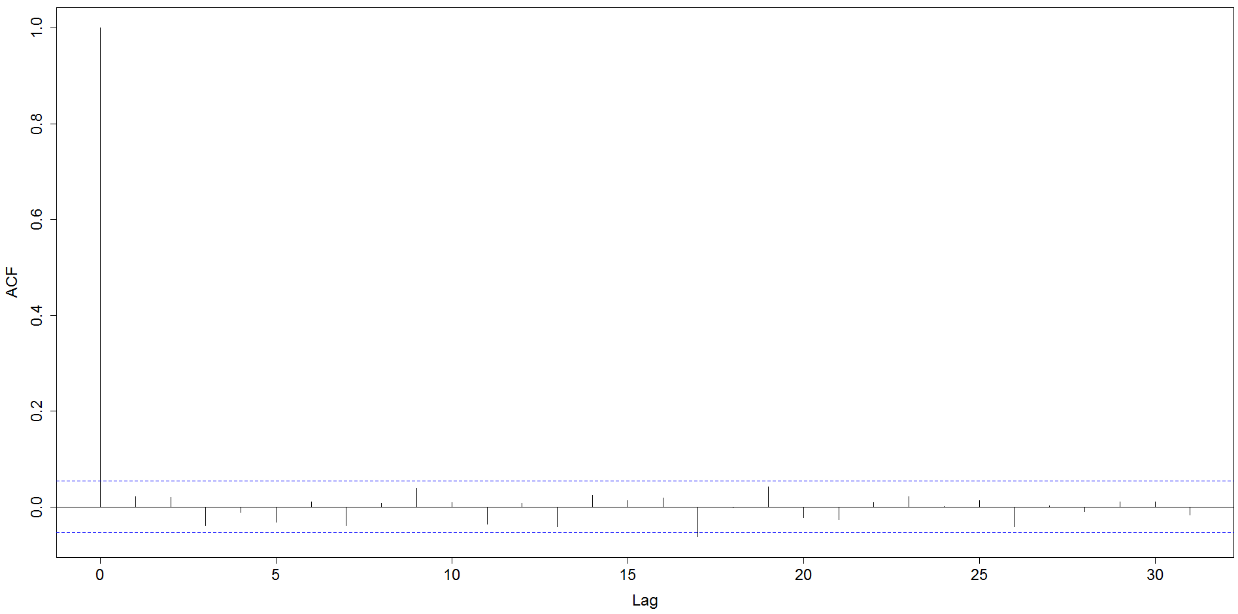

Figure 8: Auto-correlation plot for residuals of linear model

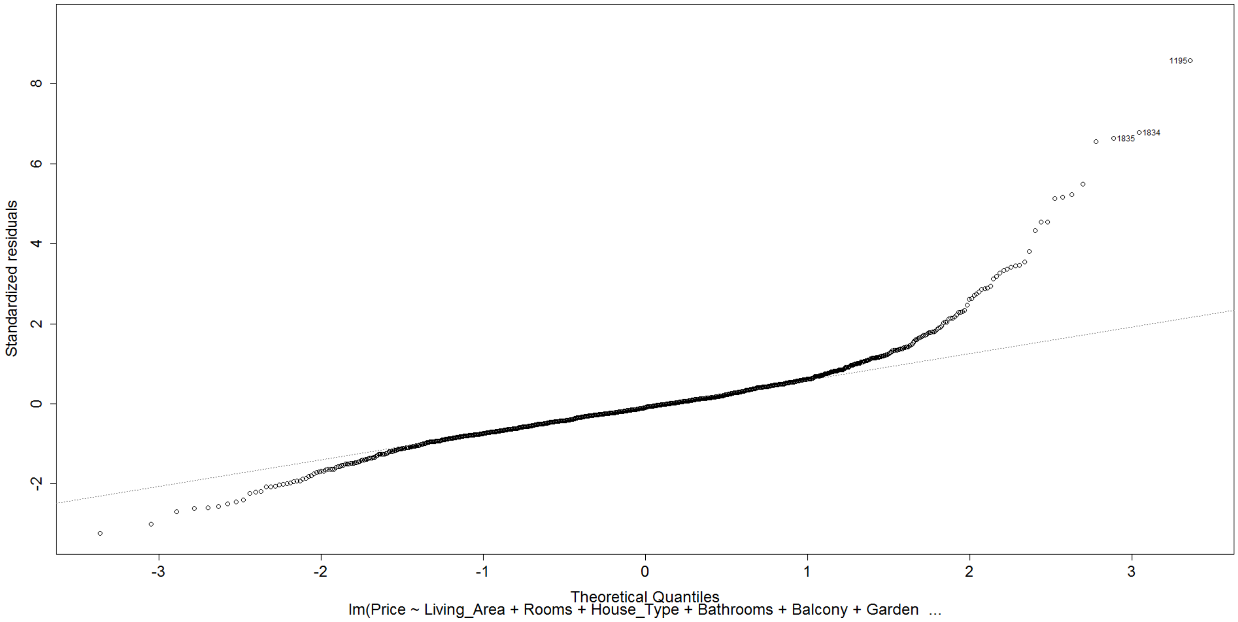

Figure 9: Normal QQ plot for linear model

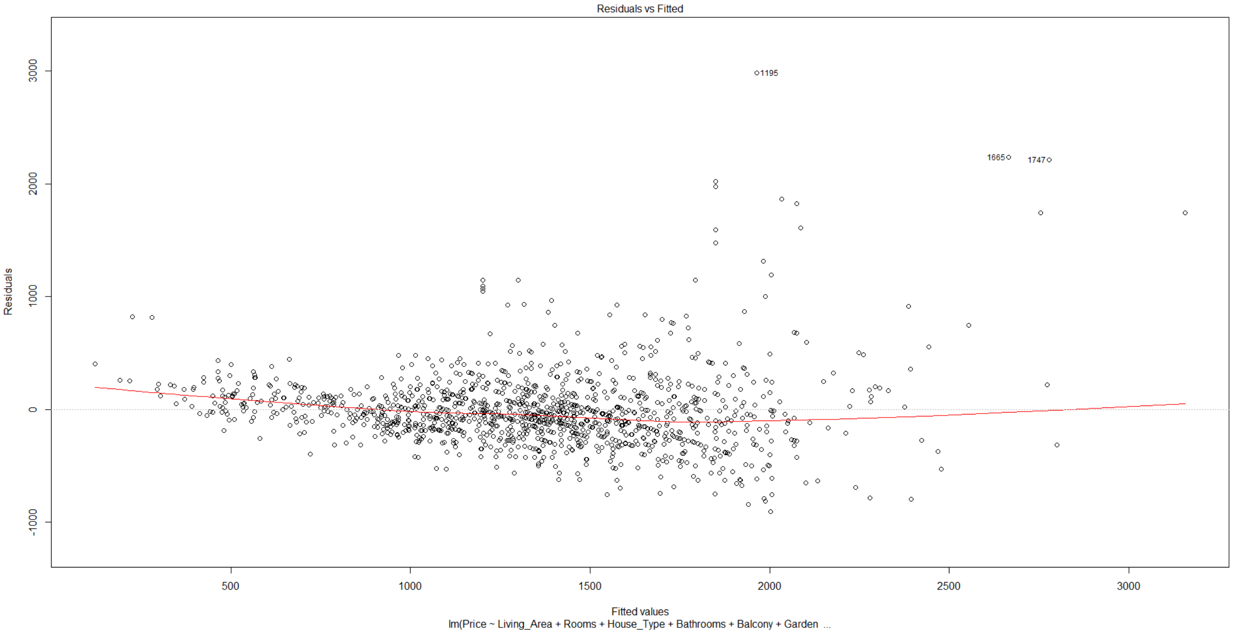

Figure 10: Residuals versus fitted values after Box-Cox transformation

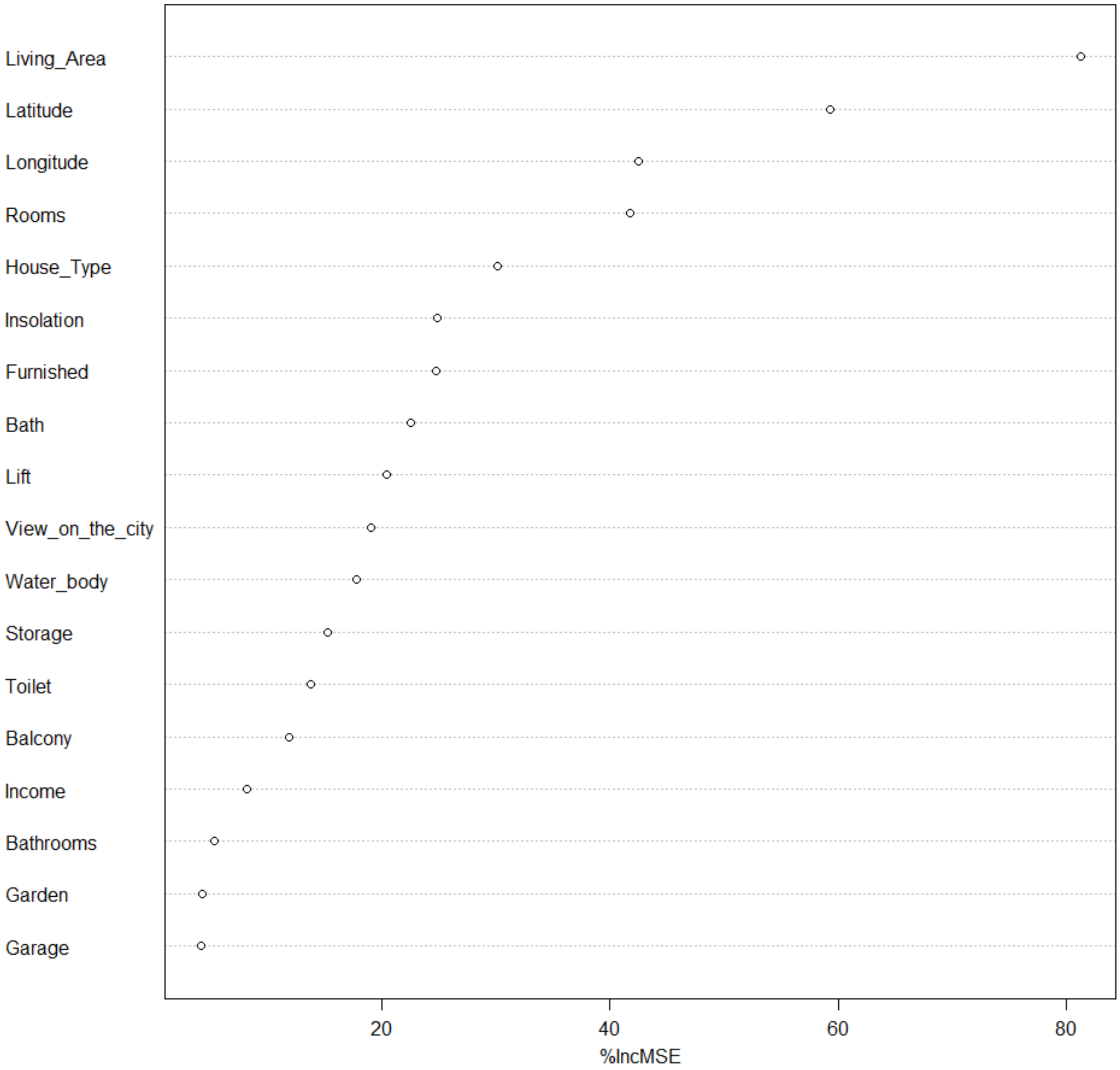

Figure 11: Variable importance of the final random forest model

References

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the robustness of interpretability methods.* Retrieved 2020-06-08, from https://arxiv.org/abs/1806.08049

Bajari, P., Chernozhukov, V., Huerta, R., Monokrousos, G., Manukonda, M., Mishra, A., & Schoelkopf, B. (2019). *Quality-adjusted price indices powered by ml and ai.* Amazon Core AI Science-Engineering. Retrieved 2020-06-04, from https://www.census.gov/content/dam/Census/about/about-the-bureau/adrm/FESAC/meetings/Chernozhukov%20Presentation.pdf

Ball, M. (1973). Recent empirical work of the determinants of relative house prices. *Urban Studies*, *10*(1), 213-233.

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.

Biecek, P., & Burzykowski, T. (2020). *Explanatory model analysis.* Retrieved 2020-06-08, from https://pbiecek.github.io/ema/

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

Brinckerhoff, P. (2001). *The effect of rail transit on property values: Summary of studies.* Neorail. Retrieved 2020-06-04, from http://www3.drcog.org/documents/archive/The_effect_of_Rail_Transit_on_Property_Values_Summary_of_Studies1.pdf

Buteikis, A. (2019). *Practical econometrics and data science.* Vilnius University. Retrieved 2020-06-12, from http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/index.html

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

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

Cassel, E. . M. R. (1985). The choice of functional forms for hedonic price equations: Comment. *Journal of Urban Economics*, *18*(2), 135-142.

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

Do, A., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. *Real Estate Appraiser*, *58*(3), 38.

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

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

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

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

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.

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.

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

Kain, J. M., J F Quigley. (1970). Some empirical results on the nature of the hedonic price function for the urban housing market. *Journal of the American Statistical Association*, *65*(1), 532-548.

Karanikolas, N., Vagiona, D., & Xifilidou, A. (2011). Real estate values and environment: A case study on the effect of the environment on residential real estate values. *International Journal Of Academic Research*, *3*(1), 861-868.

Ketkar, K. (1992). Hazardous waste sites and property values in the state of new jersey. *Applied Economics*, *24*(1), 647-659.

Klein, R. (2003). *How to win land development issues.* Owings Mills, Md.: Community and Environmental Defense Services.

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*, *10*(5), 1-19.

Lenk, M., Worzala, E., & Silva, A. (1997). High-tech valuation: should artificial neural networks bypass the human valuer? *Journal Of Property Valuation And Investment*, *15*(1), 8-26.

Li, F., Krishna, R., & Xu, D. (2019). *Convolutional neural networks for visual recognition.* Stanford University. Retrieved 2020-07-13, from `https://cs231n.github.io/convolutional-networks/`

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

Limsombunc, V., Gan, C., & Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network. *American Journal Of Applied Sciences*, *1*(3), 193-201.

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.

McMillan, R. . T. P., D Jarmin. (1992). Selection bias and land development in the monocentric model. *Journal of Urban Economics*, *31*(1), 273-284.

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/`)

Neloy, A., Haque, H., & Islam, M. (2019). Ensemble learning based rental apartment price prediction model by categorical features factoring. In (p. 350-356). doi: 10.1145/3318299.3318377

Palmquist, R. B. (1992). Valuing localized externalities. *Journal of Urban Economics*, *31*(1), 59-68.

Pedersen, T. L., & Benesty, M. (2019). *Understanding lime.* The Comprehensive R Archive Network. Retrieved 2020-06-08, from `https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html`

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision And Applications*, *29*(4), 667-676.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In (p. 97-101). doi: 10.18653/v1/N16-3020

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 35-55.

Rossbach, P. (2018). *Neural networks vs. random forests – does it always have to be deep learning?* Frankfurt of Finance and Management. Retrieved 2020-06-05, from `https://blog.frankfurt-school.de/neural-networks-vs-random-forests-does-it-always-have-to-be-deep-learning/`

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. *arXiv:1409.4842*, 1-9.

Wolf, K. L. (2007). City trees and property values. *Arborist News*, *16*(4), 34-36.

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions On Multimedia*, *19*(12), 2751-2759.