ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Predicting house prices in the Netherlands: Impact of geospatial clusters

Master Thesis (FEM61007)

P.C.D.M. Kemper
(649364)

June 1, 2023

Abstract: This research paper aims to investigate to what extent the inclusion of geospatial clusters enhances the accuracy of a house price prediction model in the Netherlands. The study evaluated three advanced machine learning algorithms: Extreme gradient boosting, Random Forests, and SVM. It used the best-performing technique to investigate the impact of geographic segmentation based on geospatial variables on the house price models. Although integrating geospatial clusters did not increase the accuracy of the house price prediction model, it did improve efficiency. Clustering zip codes based on given characteristics using the K-means method was found to be a reliable approach. The XGBoost algorithm outperforms the other advanced machine learning techniques regarding predictive performance. The study also highlighted the importance of floor area, location, building year, and plot size in forecasting house values. Overall, the findings provide valuable insights for various stakeholders in the real estate industry.

Keywords: House pricing model, Advanced Machine Learning, Geospatial clusters, The Netherlands, Real Estate Market

Supervisor: dr. J.E.M. van Nierop

Table of Content

1. Introduction

In the last couple of years, there has been a high demand for housing in the Netherlands, which has caused house prices to increase and raised concerns about affordability (Statistiek, 2023). However, there has been a slowdown in the continuous rise of house prices in recent months. The increase in the mortgage interest rate is a crucial factor that impacts individuals' decisions not to make a housing purchase, leading to this deceleration in the continuous rise of house prices (Brasington & Sarama Jr., 2008). The ability to predict house prices accurately is of great interest in mortgage lending. The appraisal is vital in determining the mortgage amount (Zaken, 2022). An accurate appraisal is of great interest to the mortgage lender in perceiving the associated risk and to the mortgagee as a safeguard in not obtaining a mortgage that exceeds their financial capacity. In the last decade, the interest in creating an accurate model-based estimation of house prices increased significantly.

The early focus of the existing literature is on the practice of the hedonic pricing models in the housing market, which use regression analysis to evaluate the property price based on specific attributes, including location, size, number of (bed)rooms, and other relevant features (Herath & Maier, 2010). The recent years' remarkable development in computational processing shifted the focus to applying several machine learning techniques. Advanced machine learning techniques, such as artificial neural networks (Kitapci et al., 2017; Peter et al., 2020; Rahman et al., 2019) and decision tree-based approaches (Yang et al., 2017; Zhang et al., 2021), are applied to outperform the estimation performance of the initial hedonic pricing models. The researchers show that house prices can be estimated using these types of models in many cities (e.g., London (Levantesi & Piscopo, 2020), Xiamen (L. Yang et al., 2021), and Melbourne (Das et al., 2020)). However, the focus of much-existing literature on house price estimations is on a single city within a country.

Another essential facet, besides the decision in method, is how to capture the geospatial aspect in the model. The existing literature has an ongoing debate about how to include how the physical location of the house impacts its value. Location is an essential feature of property valuation (Herath & Maier, 2010). Some of the papers capture the geospatial aspect by adding cartesian coordinates as a location feature (e.g., Dubin, 1998), while others include distances to points of interest (POIs) and area demographics (e.g., Clapp et al., 2004; Hurley & Sweeney, 2022). The paper of Bourassa et al. (2010) compares alternative ways to capture spatial dependence and finds the best results with a geostatistical model considering segmented submarkets. Given this ongoing debate and a dominant focus on single cities within a country in current literature, there is a significant interest in developing a model that

accurately estimates house prices of properties spread around an entire nation, incorporating the geospatial aspect of the properties. Therefore, this research will try to find an answer to the following (sub) research question(s):

RQ: To what extent does the inclusion of geospatial segmentation based on socioeconomic and geospatial attributes enhance the accuracy of house price prediction models in the Netherlands?

SUB-RQ (1): How can the zip codes in the Netherlands most effectively be segmented based on geospatial attributes?

SUB-RQ (2): What is the most accurate advanced machine learning technique in house price predictions in the Netherlands?

SUB-RQ (3): Which factors are the most important in predicting house prices in the Netherlands?

The main research objective is to develop an accurate house pricing model, which is of great interest for practical applications for homeowners, buyers, real estate agents, policymakers, and mortgage lenders. The research contains two parts. The first part will focus on segmenting the zip codes in the Netherlands based on socioeconomic, demographic, and geospatial attributes. The second stage will use these segmented sub-markets and house features to develop a housing pricing model.

This report is organized as follows. The next part, the literature review, includes existing research on house price prediction, geographic clustering, and machine learning approaches in house price prediction models. The data section describes the dataset utilized in the study and the strategies used to prepare the data for analysis. The methodology section outlines the study's techniques, including the machine learning algorithms used for prediction and the statistical methods used to assess the effectiveness of the models. The results section presents the study's findings, including the explanations of the different created clusters, the accuracy of the prediction models, and the influence of geographical clustering on the home price prediction. Finally, the discussion part provides a critical analysis of the study, highlighting the research’s limitations and future research possibilities. The conclusion summarizes the study's key findings and discusses their implications for stakeholders.

2. Literature Review

Predicting house prices is a challenging task that attracted many researchers' attention. In the last years, advanced machine learning techniques made his occurrence in developing models that accurately predict house prices. One of the main challenges is capturing the model's geospatial aspect, as a property's location significantly affects its value. This literature

review explores the state of the art in estimating house prices in the Netherlands using advanced machine learning techniques, considering geospatial data.

2.1 Hedonic house price predictions models

The hedonic price model technique is excessively utilized in the early research on the relationship between house attributes and price (Freeman, 1979; Li & Brown, 1980). According to Rosen's hedonic price model developed in 1974, products are marketed as bundles of intrinsic traits. This means that one product's price will differ from another's because one good has more qualities than the other. Therefore, the price difference between houses is related to the different attributes of one house relative to another. A house's relative price can be calculated by summing all its marginal or implicit prices assessed using regression analysis. Freeman (1979) suggests that housing units can be classified based on structural and locational factors. These house characteristics are captured in the hedonic pricing function based on the levels of these attributes. The hedonic pricing function may be derived using observed prices and characteristics of many models, allowing the price of any model to be calculated depending on its characteristics.

Structural attributes are seen as necessary in the prediction of house prices. Several studies have found a positive relationship between the house price and various house attributes, such as the number of rooms, bedrooms, bathrooms, and the floor area (Clauretie & Neill, 2000; Fletcher et al., 2000; Li & Brown, 1980). These qualities considerably affect a house's market worth, where the floor area is solidly found as the most crucial feature. Clark & Herrin (2000) found that the age of a building is negatively connected to its price because of increased care and repair expenses, as well as changes in architecture, electrical and mechanical systems, and decreasing usefulness; older homes are frequently worth less (Clapp & Giaccotto, 1998). However, Li and Brown (1980) discovered that age had the opposite impact on specific structures, which they attributed to historical relevance or vintage influences. Lot size, garage, patio, water heating system, one or more fireplaces, and an air heating system have all been demonstrated to be determinants in influencing house pricing (Garrod & Willis, 1992). Garrod and Willis (1992) discovered that a single garage adds 6.9% to the price of a house, and a double garage adds three times that amount, while central heating adds roughly 6.5% to the price of a home. However, little study on the impact of structural quality on housing prices has been undertaken, most likely because of the difficulties in objectively and accurately quantifying physical and environmental quality (Morris et al., 1972).

Although the number of rooms and floor size are typically consistent between countries, other factors such as building design or environment may impact buyer preferences.

According to Kohlhase's (1991) research, the structural features that house buyers value may not be stable over time or between nations. This suggests that the relative value of structural features varies depending on the property's environment.

Locational considerations are essential in influencing house values. Locational factors are seen as essential considerations in predicting housing values as well. To quantify locational features, surrogate variables such as socioeconomic class, racial composition, aesthetic attributes, pollution levels, and proximity to local amenities have been utilized (Dubin & Sung, 1990). Furthermore, the view is a residential amenity related to a dwelling site's position (So et al., 1997). Higher floors often have better views. So et al. (1997) discovered a tangible link between view and floor level, with higher-floor apartments typically imposing a higher price than lower-floor units. While neighborhood traits cannot be overtly priced in the marketplace, Goodman (1989) argues that hedonic pricing can implicitly be valued when comparing residences with different neighborhood qualities. Failure to incorporate neighborhood features can result in significant inaccuracies when evaluating individual properties and the market as a whole, as Linneman (1980) shows. The three main categories of neighborhood attributes are (1) socioeconomic variables such as the occupation of the inhabitants, (2) local government or municipal services such as schools and hospitals (e.g., Clauretie & Neill, 2000), and (3) externalities such as crime rates, traffic and airport noise, and shopping centers (e.g., Espey & Lopez, 2000). However, as mentioned earlier, implementing the locational and neighborhood (geospatial) aspect in house price modeling is challenging. Section 2.4 elaborates more on how researchers implement the locational and neighborhood aspects in their models.

The current literature review on using hedonic price models in the housing market gives significant insight into the critical aspects to be incorporated into our pricing model.

2.2 Integration of Advanced Machine Learning Techniques

As mentioned in the previous section, the early focus in the existing literature was on the practice of the hedonic pricing models in the housing market. However, implementing advanced machine learning techniques in real estate price predictions experienced enormous interest in recent years. Ho, Tang, and Wong (2021) researched how effective different machine learning techniques are in the prediction of house prices, including gradient boosting machines (GBM), Random Forest (RF), and support vector machines (SVM). The comparison analyses are conducted on a dataset of residential properties from 14 areas in a highly dense district in Hong Kong. The authors conducted three different models to predict the transaction price of residential properties based on various property characteristics, such as the age of the

building, floor level of the property, and accessibility of the residential property. The gradient boosting machines and Random Forest model outperforms the SVM in accuracy and efficiency. According to the results, advanced machine learning techniques can predict house prices accurately. The R-squared value for gradient boosting and random forest models on the test set is around 0.903.

Similarly, Zhang et al. (2022) explore the accuracy and efficiency of implementing Random Forest in predicting second-hand house prices in Beijing (China). The findings suggest that training the dataset using Random Forest is adequate, whereas it gives an R-squared value of 0.826 on the test set.

The paper of Yang et al. (2021) focused on the relationship between Bus Rapid Transit (BRT) and house prices in Xiamen (China). They conducted a gradient-boosting decision tree (GBDT) algorithm on 5,185 properties. The study found that their GBDT algorithm outperforms the hedonic pricing models regarding predictive power. The findings confirmed a positive and non-linear effect of accessibility to BRT stations on the house price. Further, property location, size, and age are seen as the essential features in the prediction of prices.

2.3 House price predictions in the Netherlands

The paper of Guliker et al. (2022) compared different machine-learning techniques in developing hedonic pricing models in the Netherlands. The authors analyzed a dataset of 11.434 properties sold between 2018 and 2020. The houses are established in five municipalities spread across the Netherlands: Rotterdam, Amsterdam, Eindhoven, Amersfoort, and Groningen. The machine-learning techniques Guliker et al. (2022) implemented are linear regression (LR), geographically weighted regression (GWR), and extreme gradient boosting (XGBoost). The authors consider the geospatial aspect by adding features that capture distances to POIs and neighborhood socioeconomic indicators (e.g., household income and urbanization degree). The geospatial features are well represented in the model, whereas the house characteristics are limited to 7 features. The authors trained the models on the five municipalities individually first. The total living area and the valuation of all nearby houses (WOZ-value) are the essential features in predicting house prices in the five municipalities in the Netherlands. The paper suggests that the XGBoost model outperformed the other two models based on average model performance for the five municipalities. The XGBoost model has an average RMSE of €61,028, whereas the LR and GWR models have RMSEs of €94,927 and €65,826, respectively. The RMSE for XGBoost training across all five municipalities, with the municipality name included as a variable, showed a slight increase to €65,312.

De Vor & de Groot (2011) conducted a hedonic pricing analysis in the Netherlands intending to investigate the impact of the distance to industrial sites on residential property value. They found that an increase of 10% in the distance to an industrial site is associated with a rise of 1.3% in property value. This research by De Vor & de Groot (2011) shows the impact of the physical location of a property on its value in the Netherlands.

2.4 Geospatial aspect

In the literature about the prediction of house prices, geospatial analysis is increasingly adopted, and various methods and algorithms have been proposed. One study by Bourassa et al. (2010) compared different methods to include spatial dependence in price-prediction models for nearly 13,000 houses in Louisville, Kentucky. The findings show the best results with a geostatistical model considering segmented submarkets. In this paper, the submarkets are segmented by hierarchical cluster analysis. The hierarchical cluster analysis includes features such as the hedonic house characteristics (e.g., age, # bathrooms) and price. The authors suggest that an increase in segmented submarkets results in a rise in predictive performance. Goodman & Thibodeau (1998, 2007) and Bourassa et al. (2003) are in line with Bourassa et al. (2010). Both found significant importance in segmenting area housing submarkets, using geospatial analysis, in predicting house prices. The paper of Hsieh (2011) created submarkets using the K-means clustering approach. To create homogenous submarkets, they perform K-means on five variables (site area, floor level, dwelling age, width of closest road, and distance to the city center).

Norman et al. (2008) examine the relationship between a geospatial aspect (educational test scores) and house prices, accounting for spatial dependence. The study is conducted for the ten largest cities in Sweden using a spatial autoregressive model. They found a positive and significant effect on educational test scores and house prices. The focus of Cellmer et al. (2019) research is on the impact of the distance of an airport on the prices of houses. The analysis suggests that the vicinity of an airport and the type of airport, in terms of the frequency of flights, significantly impact house prices. Ramírez-Juidías et al. (2022) presented comparable research but focused on the impact of the distance of urban green spaces on housing prices.

In the existing literature, much evidence exists for creating segments of areas. Further, the importance of spatial analysis in the prediction of house prices is highlighted in the literature.

3. Data Collection

This research aims to create a model that accurately predicts house prices in the Netherlands. The analysis is split into two stages. The first stage focuses on segmenting the zip codes, and the second stage uses these clusters to predict the prices of properties in the Netherlands. The analysis is conducted on the base of two data sets. One data set contains features for the spatial clustering approach on the zip code levels, and the other contains house features of properties spread across the Netherlands. This section will describe the data used in the analysis.

3.1 Geospatial Data Set

As explained earlier, the first part of the analysis focuses on segmenting the zip codes based on spatial, socioeconomic, and demographic features. The data set is on a six numerical zip code (PC6) level. The motivation behind the decision to do the clustering on a six numerical zip code level is that the WOZ-value per zip code shows us that the house price indices significantly differ among six numerical zip codes. Therefore, performing the clustering algorithm on a five- or four-numerical zip code level gives a potential loss in information. However, the drawback of choosing the six numerical zip code levels is that there is a significant amount of missing socioeconomic and demographic data points. This problem is partially solved by taking the values of some of the features' five numerical zip codes (presented in Table 1). This approach gives us a data set of 444,492 zip codes, representing 96.5% of all the zip codes in the Netherlands. The missing zip codes are primarily in less populated areas and will be no significant problem in the application. Some demographic and socioeconomic variables are converted into percentages of the total inhabitants/households to make comparing the zip codes based on compositions easier. Further, the data set contains distances to points of interest (POIs) in one straight line. The distances are calculated using the `*geosphere*` package in R by taking the average longitude and latitude per zip code and the longitude and latitude of the POI. The explanation and source of each variable in this data set are presented in Table 1. The data on the socioeconomic variables are from 2020. The variables in this data set align with the most commonly used variables for geospatial analysis in other house pricing studies (Bourassa et al., 2010; Cellmer et al., 2019; Guliker et al., 2022; Yang et al., 2021).

TABLE 1: Explanation of the features of the geospatial data set

Variable	Explanation	ZC (Level)	Source
Inhabitants	Number of inhabitants	PC6	CBS (2020)
Houses	Number of houses	PC6	CBS (2020)
under_15	% of inhabitants under an age of 15	PC5	CBS (2020)
b_15_24	% of inhabitants between an age of 15 and 24	PC5	CBS (2020)
b_25_44	% of inhabitants between an age of 24 and 44	PC5	CBS (2020)
b_45_64	% of inhabitants between an age of 45 and 64	PC5	CBS (2020)
above_65	% of inhabitants above an age of 65	PC5	CBS (2020)
origin_NL	% of inhabitants with Dutch heritage	PC5	CBS (2020)
origin_WE	% of inhabitants with a western heritage	PC5	CBS (2020)
origin_NW	% of inhabits with a non-western heritage	PC5	CBS (2020)
hh_one_pers	% of households with one person	PC5	CBS (2020)
hh_one_parent	% of households with one parent	PC5	CBS (2020)
hh_two_parent	% of households with two parents	PC5	CBS (2020)
hh_no_child	% of households with no children and multiple adults	PC5	CBS (2020)
buy_house	% of houses as buy houses	PC6	CBS (2020)
Rent_house	% of houses as renting houses	PC6	CBS (2020)
woz_value	Average WOZ value (in thousands €)	PC6	CBS (2020)
density	House density (addresses/ km^2)	PC5	CBS (2020)
urbanization	Urbanization degree (1 is high urbanization, 5 is low urbanization)	PC5	CBS (2020)
avg_dist_super	Average distance to closest supermarket (m)	PC6	-
avg_dist_pschool	Average distance to closest primary school (m)	PC6	Rijksoverheid (2023)
avg_dist_hschool	Average distance to closest secondary school (m)	PC6	Rijksoverheid (2023)
avg_dist_train	Average distance to closest train station (m)	PC6	Rijdendetreinen.nl (2023)
avg_dist_airport	Average distance to closest airport (m)	PC6	Google maps (2023)
avg_dist_green	Average distance to closest green space (m)	PC6	PDOK (2023)
Longitude	Measurement of place east or west	PC6	CBS (2020)
Latitude	Measurement of place north or south	PC6	CBS (2020)

The summary descriptives of the geospatial data set are shown in Table 2. One feature per category that serves as a reference group is deleted to address potential issues resulting from perfect collinearity between the categories that total to 100%. Additionally, compared to removing one reference feature per group, keeping all of the characteristics for each category in the dataset would not contribute any additional information. For the studies, the dataset is stripped of the over-65 age group, one-parent families, non-Western origins, and the percentage of rented housing. To provide a more practical knowledge of the clusters, the characteristics are introduced once again in the description of the clusters. The data set captures 444,492 six-numerical zip codes. On average, the number of inhabitants is 42, with a

minimum of 5 and a maximum of 2180. Further, the table shows us that the mean percentages of the compositions of the households, except for the one-parent households, are comparable. Further, the mean ratio of buy/rent houses is 60/40, and the zip codes are roughly equally distributed among urban and less urban areas. On average are green spaces closer located than high schools. The primary schools are the closest on average to each zip code, whereas the airports are the furthest. The average WOZ value of the zip codes is €258,840 in 2020, where the minimum and maximum are €3,000 and €5,813,000, respectively. The zip codes corresponding to the minimum of €3,000 represents only garage boxes, which explains the low average WOZ value. The average WOZ value of a home in 2020 is roughly €351,000, which differs from the average WOZ value of the data set since the average WOZ value of a house is computed on an individual level (CBS, 2020). However, the WOZ value of the data set is derived on a zip code level. This discrepancy in calculation results in a disparity since the number of dwellings in a zip code is not considered in calculating the data set's mean.

TABLE 2: Summary Descriptives for the features of the geospatial data set (n = 444,492)

Features	Mean	Sd	Min	Max	Range
inhabitants	41.58	31.92	5.00	2180.00	2175.00
houses	19.76	17.08	5.00	1050.00	1045.00
woz_value (in thousands €)	285.84	163.86	3.00	5813.00	5810.00
householdsize	2.24	0.58	1.00	6.70	5.70
age_under_15	0.15	0.05	0.00	0.70	0.70
age_15_24	0.12	0.05	0.00	1.00	1.00
age_25_44	0.24	0.09	0.00	1.00	1.00
age_45_64	0.28	0.06	0.00	1.00	1.00
origin_NL	0.77	0.18	0.00	1.00	1.00
origin_WE	0.11	0.07	0.00	0.70	0.70
hh_one_pers	0.36	0.15	0.00	1.00	1.00
hh_two_parent	0.26	0.12	0.00	1.00	1.00
hh_no_child	0.31	0.09	0.00	1.00	1.00
buy_house	0.61	0.26	0.00	1.00	1.00
Density (addresses/km^2)	1817.18	1758.15	3.00	12474.00	12471.00
urbanization	2.89	1.44	1.00	5.00	4.00
avg_dist_pschool (m)	533.82	514.46	0.00	10865.32	10865.32
avg_dist_hschool (m)	1946.39	2069.06	0.00	17452.48	17452.48
avg_dist_train (m)	3852.59	4181.26	7.03	35666.02	35658.98
avg_dist_airport (m)	38011.76	26497.87	388.14	116432.01	116043.87
avg_dist_green (m)	864.55	1277.40	0.97	28297.74	28296.78
avg_dist_super (m)	647.03	732.84	0.24	10179.88	10179.64
Longitude	5.34	0.79	3.36	7.22	3.86
Latitude	52.11	0.54	50.75	53.50	2.75

(A) Kralingen Oost (Rotterdam)

(B) Pierik (Zwolle)

(C) Binnenstad (Maastricht)

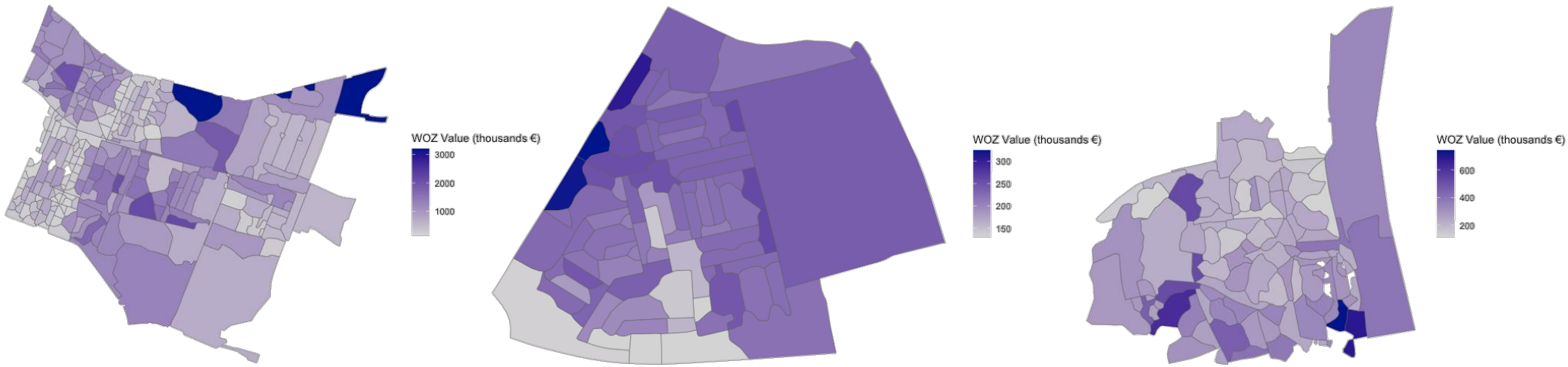


FIGURE 1. Visualization of the average WOZ Value on zip code level in the neighborhoods Kralingen Oost (Rotterdam), Pierik (Zwolle), and Binnenstad (Maastricht)

(A) Kralingen Oost (Rotterdam)

(B) Pierik (Zwolle)

(C) Binnenstad (Maastricht)

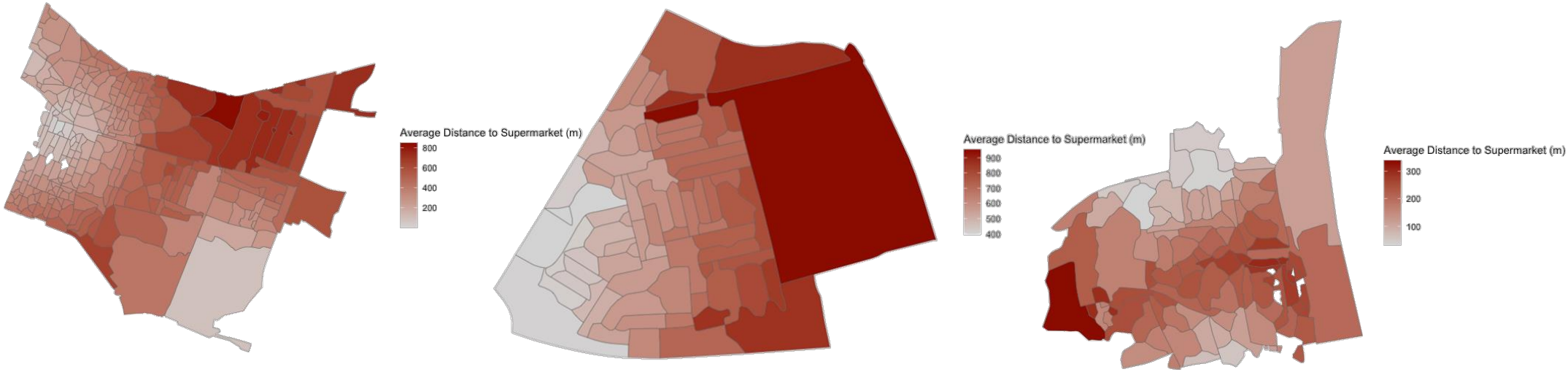


FIGURE 2. Visualization of the average distances to supermarkets on zip code level in the neighborhoods Kralingen Oost (Rotterdam), Pierik (Zwolle), and Binnenstad (Maastricht)

To illustrate the spatial distribution of WOZ values and minimum distance to supermarkets for three neighborhoods the `*ggplot2*` package in R is used. The average WOZ values and distances to supermarkets for the neighborhoods Kralingen-Oost (Rotterdam), Pierik (Zwolle), and Binnenstad (Maastricht) on a zip code level are shown in Figures 1 and 2. As can be observed, these variables vary significantly between zip codes, indicating the prevalence of zip code differences.

3.2 House Characteristic Data Set

The second data set contains house features that are collected from VBO.nl with the use of the web scraping package '*rvest*' in R (VBO, 2023). VBO.nl is a large real estate platform providing detailed house features for sale in the Netherlands. Vereniging Bemiddeling Onroerend Goed (VBO) is informed and has approved using their data. After removing missing and unreasonable high/low values and running through data pre-processing steps (e.g., text mining of house descriptions and cleaning of the data), the data set contains 10.840 houses (observations) for sale between the beginning of February and the end of April (2023). Figure 3 shows the locations of the properties in the data set and confirms that the houses are spread all over the Netherlands, with a higher density in the "Randstad".

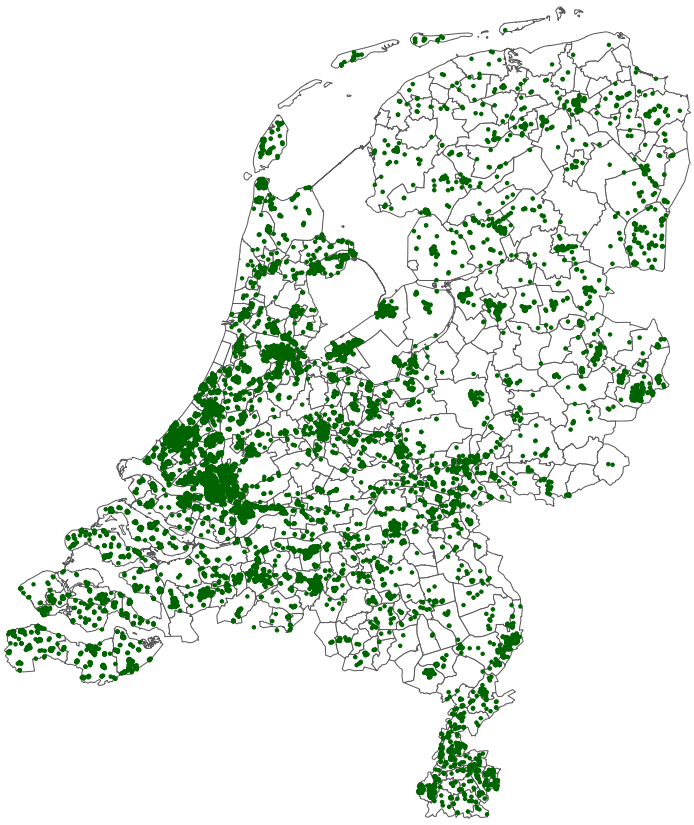


FIGURE 3. The allocation of the 10,840 houses across the Netherlands

The features of a balcony, air conditioning, solar panels, garage, fireplace, bath, pool, jacuzzi, and heat-related, are created using text mining of the house text description. In the house description scraped from VBO.nl, different terms related to the different features are searched with the *'grepl'* function in R, given a boolean variable of the occurrence of these related terms. This text-mining approach has considered negations to create accurate values for the different features. Since the house descriptions are in Dutch, the negations taken into account are "*geen", "niet" and "mogelijkheid tot".* The negations "*geen*" and "*niet*" are the most common negations in the Dutch language. The phrase *"mogelijkheid tot"* is often used in the real estate industry to indicate that a feature can be added in the future, even though it does not currently exist. We look at the two words before the focal word when considering negations. In order to confirm the text mining method, a random sample of 50 observations was reviewed and confirmed. Table 3 shows the included house features in the data set, separated by the features obtained from VBO directly and the text-mined features.

TABLE 3: House features and data type

VBO features	**Data Type**	**Text mining features**	**Data Type**
House Type (20 categories)	Categorical	Balcony	Boolean
Garden	Boolean	Patio	Boolean
Plot Size (m^2)	Integer	Airconditioning	Boolean
Floor area (m^2)	Integer	Solar panels	Boolean
External Storage (m^2)	Numerical	Garage	Boolean
Building Year	Integer	Fireplace	Boolean
Number of floors	Integer	Pool	Boolean
Floor Location	Integer	Jacuzzi	Boolean
Price (in €)	Numerical	Bath	Boolean
Number of Bedrooms	Integer	Central Heater Boiler (NL: CV-Ketel)	Boolean
Number of Bathrooms	Integer	Block Heater (NL: Bloksverwarming)	Boolean
Latitude (North)	Numerical	City Heater (NL: Stadsverwarming)	Boolean
Longitude (East)	Numerical	Heater Pump (NL: Warmtepomp)	Boolean
		Under Floor Heater (NL: Vloerverwarming)	Boolean

The summary descriptives of the numerical features are presented in Table 4. The table shows that the mean price in the data set is €443,854, whereas the most and less expensive houses are €6,900,000 and €104,000, respectively. According to Centraal Bureau voor de Statistiek (CBS, 2023), the average sale price of a property in March 2023 is €415,100, which is slightly lower than our data set. The maximum building year is 2025, representing a house that will be delivered in 2025 but is already available for sale. Based on the coordinates, the

midway point of the dwellings in the data set is around Utrecht. The summary statistics of the factor variables are presented in Appendix A.

The data set is divided into train and test data in the modeling process, with an 80:20 ratio. The models are trained using the training data, while the test data is used to evaluate the model.

TABLE 4: Summary Descriptives for the features of the house data set (n = 10,840)

Features	Mean	Sd	Min	Max	Range
Plot Size (m^2)	512.50	6019.90	26.00	372000.00	371974.00
Floor Area (m^2)	121.29	55.45	24.00	735.00	711.00
External Storage (m^2)	12.13	36.90	0.00	996.00	996.00
Building Year	1966.76	42.02	1363.00	2025.00	662.00
Number of Floors	2.38	0.96	1.00	7.00	6.00
Floor Location	1.55	1.98	1.00	51.00	50.00
Bedrooms	3.24	1.24	1.00	14.00	13.00
Bathrooms	1.10	0.36	1.00	11.00	10.00
Price (€)	443,853.67	271,289.57	104,000.00	6,900,000.00	6,796,000.00
Longitude	4.97	0.77	3.38	7.16	3.78
Latitude	51.99	0.49	50.76	53.48	2.72

4. Methodology

As discussed earlier, the research consists of two stages. The first stage will segment the zip codes in the Netherlands using socioeconomic, demographic, and geospatial variables. The other stage will create a house prediction model by combining the cluster of the first stage with a data set including house features. The following part will discuss the proposed methods for both parts.

4.1 Clustering analysis

The K-means clustering approach is implemented to identify submarkets in the Netherlands. K-means is a prominent clustering approach that is utilized in a variety of disciplines. The K-means algorithm is an unsupervised clustering approach that identifies k centers in the feature space to split a dataset into k groups. The k-means technique is based on doing a series of local searches to find the best solution to a clustering issue (Likas et al., 2003). This strategy aims to find homogenous k clusters within each cluster but diverse between them. The algorithm is divided into three phases. The required number of clusters (k) is first provided. Second, the initial centroids are determined, and each observation is allocated to the nearest centroid by calculating the Euclidean distance between each observation and the centroids. Finally, the new mean of each cluster updates the initial centroid, and the procedure is repeated until the clustering solution is stable (Hamzah et al.,

2017). In other words, each observation is allocated to the closest center, and the center points are changed iteratively to reduce within-cluster variance while maximizing variance among centers. The mathematics behind k-means is that the algorithm iterates over a series of required conditions in order to minimize the k-means objective function, indicated as

$$\mathrm{J}(\mathrm{z},\mathrm{A}) = \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik}(\|x_i - a_k\|)^2 \qquad (1)$$

Where $\mathrm{J}(\mathrm{z},\mathrm{A})$ is the k-means objective function, c is the number of clusters, z_{ik} indicates whether an observation x_i belongs to the k[th] cluster, a_k is the cluster center in cluster k, and $(\|x_i - a_k\|)$ is the Euclidian distance (Sinaga & Yang, 2020). One of the most challenging aspects of the k-means clustering technique is that the number of clusters must be predetermined. Widely used approaches to determine the optimal number of clusters are the elbow rule (calculating the total within-cluster sum of square (WSS)), the average silhouette method, and the Gap-statistic.

The K-means algorithm has considerable advantages in that it works well with large data sets and performs better than the Hierarchical Clustering Algorithm, resulting in a higher quality of clusters (Kaushik & Mathur, 2014). This study uses K-means to group zip codes based on standard socioeconomic and demographic features and average distances to points of interest (POIs). Segmenting the zip codes based on the features and distances to POIs with the use of K-means is a new approach in the field of house price predictions. The paper of Hsieh (2011) has a similar idea of implementing K-means but uses the clustering approach to create submarkets based on structural and locational house attributes. However, the main drawback of this approach is that it is only easily applicable to a house in the original data set. In that case, the clustering process must be started again with the new properties. Using the K-means algorithm on a zip code level within a country allows one to easily connect any house in a country to a geospatial cluster using the zip code. Segmenting zip codes based on socioeconomic attributes is often performed in other fields (e.g., medical research). For example, Akbilgic et al. (2021) investigated the influence of socioeconomic and environmental variables on racial differences in children's preoperative physical condition. The authors implemented the K-means algorithm to segment the 29 zip codes from Memphis based on socioeconomic attributes. The paper of Khmaissia et al. (2020) aims to discover the underlying characteristics associated with a rise in the number of new COVID-19 cases in New York City. The study aims to identify the significant elements related to the virus's proliferation throughout the city. The authors successfully use the K-means algorithm to split the 177 zip codes in New York into 6 clusters based on demographic, socioeconomic, and mobility features and connect this back to the daily COVID-19 case increase rates.

Shokoohyar et al. (2019) implemented K-means similarly to this report. The authors investigate the impact of clusters, based on socioeconomic and transportation features, on the Uber waiting time. The results suggest that Uber tends to be more accessible in locations with a more significant population density, male inhabitants, better access to public transit, and a lack of facilities within walking distance.

Literature confirms that K-means helps find comparable zip codes based on demographic, socioeconomic, and distance data. Therefore, this report will use the K-means algorithm to segment the zip codes in the Netherlands and use this to create an accurate house price model.

4.2 Machine Learning Modeling

Advanced machine learning techniques are used to predict house prices in the Netherlands. The Extreme Gradient Boosting (XGboost), Random Forest, and Support Vector Machines (SVM) approaches are implemented to create an accurate house pricing model. The main advantage of these techniques is that they can deal with complex, mixed, and high-dimensional data. Advanced machine learning techniques are recommended since they provide overall accurate prediction models (Ho et al., 2021). Further, another advantage of the advanced machine learning techniques is that multicollinearity does not influence the models' predictive performance. However, the main disadvantage is that these models are often called "black box" models since it is difficult to interpret them. The three advanced machine learning techniques are frequently utilized for predicting house prices (Adetunji et al., 2022; Levantesi & Piscopo, 2020; L. Yang et al., 2021; W. Zhang et al., 2021; Y. Zhang et al., 2022). The advanced machine learning techniques are compared to an ordinary least square (OLS) model as a benchmark. Together with the home characteristics, the K-means method clusters are included in the model as a factor feature. Compared to SVM and OLS, which consider the cluster variable as one-hot encoding, the Random Forest and XGBoost technique can directly handle a factor feature. The following sections discuss the three advanced machine learning techniques and the benchmark method.

4.2.1 Extreme Gradient Boosting (XGboost)

Extreme gradient boosting (XGBoost) is a recent advancement in gradient-boosting machine learning. Friedman presented the initial gradient-boosting technique in 2001. Boosting is an ensemble learning approach of K-classification and regression trees that is an iterative procedure that creates numerous decision trees, one after the other. Boosting is based on the idea that each model concentrates on the poorly handled data in the preceding model, culminating in a powerful learning model with less bias. XGBoost is a more efficient and

adaptable version of Friedman's (2001) initial boosting strategy. One benefit of XGBoost over gradient boosting is that it manages overfitting better by including a regularization term, resulting in higher model performance.

Chen and He (2016) discuss the XGBoost algorithm in their paper. The objective functions in the XGBoost must be decreased for the boosting method to stop. In this paradigm, the regularized objective function is defined as follows:

$$\phi = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

Where n represents the distinct data samples and the loss function is detailed as $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ by the fit of that model. $\gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is defined as the regularization term in the objective function, which penalizes the function for the model's complexity. Because the trees are formed sequentially, each objective function of the next tree is based on the preceding tree's loss function or residuals in the predictions. This method is repeated until the objective function is reduced and the best model is discovered. In summary, XGBoost is a robust machine learning method that combines gradient boosting's benefits with extra features such as regularization to control overfitting. The XGBoost technique is a widely used modeling tool in real estate price predictions (Avanijaa & Al, 2021; Guliker et al., 2022; Zaki et al., 2022). The authors confirm that this approach gives accurate house price predictions and outperforms other advanced machine learning techniques. Finally, the model's accuracy is improved by performing a 10-k fold cross-validation and using a grid search for hyperparameter tuning, using RMSE as the accuracy metric on the training set.

4.2.2 Random Forest

Random forest is a commonly used ensemble learning approach described by Ho (1995) and used for classification and regression applications. Breiman (2001) lengthened this approach by extending the original. The first stage in a random forest uses bootstrap resampling to divide the original sample into several samples. A decision tree is generated for each of these samples. Each observation class will be identified by aggregating the majority vote of each decision tree in classification problems and averaging the predictions of each decision tree in regression problems. This underlines the critical distinction between random forest and other ensemble approaches, such as XGBoost, in that the latter grows trees sequentially while the former produces trees independently.

The Random Forest algorithm entails constructing several decision trees and merging their results to provide a final forecast. Each decision tree in this approach is built using a random subset of the features and a random subset of the training data. The combination of these trees

aids in the reduction of overfitting and the improvement of generalization performance. The average of the projections from each tree in the forest is used to get the final prediction. The prediction formula of the random forest algorithm is as follows:

$$Y = \frac{1}{n}\sum_{i=1}^{n} f_i(X) \quad (3)$$

Here, Y acts as the predicted outcome, n represents the number of trees, f_i is the prediction provided by the mth tree, and X represents the value of the data point being tested at a particular node. To maximize the model's performance, the algorithm's hyperparameters, such as the number of trees, the maximum depth of each tree, and the number of features examined at each split are tuned using a grid search. Further, the analysis conducted a 10-k fold cross-validation to improve its predictive power.

Random Forest is a popular approach in different fields, including real estate, because of its high predictive performance and reliable feature importance estimations (Ho et al., 2021). However, the main disadvantage of this approach is the relatively high computational time and the necessary amount of vector memory compared to a machine learning technique such as the Support Vector Machine (SVM).

4.2.3 Support Vector Machine (SVM)

The Support Vector Machine algorithm was initially introduced by Cortes & Vapnik in 1995. The Support Vector Machine (SVM) technique is commonly employed for classification and regression issues. SVM tries to identify a hyperplane in a higher-dimensional space to separate samples belonging to various categories in classification efficiently. SVM may be built for multi-class classification by merging two classifiers. SVM regression, on the other hand, attempts to predict continuous values based on numerous sample properties (Mohandes et al., 2004). The regression function of SVM is presented in equation (4):

$$y = (z^* * x_t) + l \quad (4)$$

Where the input data is expressed by x_t and the corresponding output data by y. z^* captures the weight vector and the l is a normal. The goal of SVM is to identify the function that minimizes the sum of the errors between the predicted and actual values, given a specific margin constraint. In order to get this done, the SVM uses the idea of "support vectors," which are data points closest to the hyperplane. The "margin" acts as the distance between the hyperplane and a support vector, and the best hyperplane is the one that maximizes this margin. The weight vector determines the margin. The SVM approach can be extended by implementing a Kernel function to solve the problem of handling non-linearly data.

Implementing a Kernel function reduces the complexity of calculations and, therefore, the computational time. Like other advanced machine learning techniques, the choice of each variable's value in the model significantly impacts the model's performance, such as the kind of kernel function and the related kernel parameters. Therefore, comparable to the other advanced machine learning techniques, the parameters are tuned using grid search, and 10-fold cross-validation is performed.

The support vector machine (SVM) is an essential component in the deployment of advanced machine-learning techniques due to its short processing time and exceptional prediction performance. In the existing literature on house price prediction, the results of SVM are seen as accurate and have less error compared to other advanced machine learning techniques, for example, Neural Networks (Sarip & Hafez, 2015; Sasaki & Yamamoto, 2018). Consequently, SVM may be used with other advanced machine learning algorithms to improve overall efficiency and effectiveness.

4.2.4 Ordinary Least Square (OLS) (Benchmark)

As explained in the literature review, Lancaster (1966) and Rosen (1974) laid the theoretical groundwork for the hedonic pricing model. As a baseline model in our analysis, we used a typical hedonic pricing model calculated using ordinary least squares (OLS) regression. This benchmark model is applied to investigate the need for advanced machine learning techniques in creating house price models because OLS has interpretability and computational benefits over advanced machine learning techniques. In order to compare the results of advanced machine learning techniques and ordinary least squares, the same explanatory variables are used in all analyses. This results in the creation of dummy variables for categorical variables and boolean variables for OLS.

4.3 Model evaluation

The predictive performance of the models is measured based on several metrics, such as root means squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and R-squared. These are the commonly used metrics for advanced machine-learning techniques (Guliker et al., 2022). The metrics are applied to assess the models' accuracy and precision in the house price predictions.

In addition, global interpretability methods, such as feature importance analysis and partial dependency plots (PDP), and local interpretability methods, such as LIME, were conducted to interpret the model better. The feature importance analysis identifies the most crucial features or variables contributing to the model's overall performance. Additionally, the PDP provides

information on the link between a specific characteristic and the model's output across all observations. In order to make the PDPs easier to understand, we remove outliers when making the plots. LIME (Local Interpretable Model-agnostic Explanations) helps explain machine learning models' predictions by using an easy-to-understand model for each instance. These are helpful tools since the advanced machine learning techniques are difficult to interpret and seen as black boxes.

As discussed earlier, the created models will be compared to an ordinary least square (OLS) as a robustness check. Further, to answer the research question, to what extent the inclusion of geospatial segmentation based on socioeconomic and geospatial attributes enhance the accuracy of house price prediction models in the Netherlands, the models are created as well, including all socioeconomic and geospatial attributes, instead of the cluster. When these findings are compared to the results of the models that include the clusters, the influence of employing geographic clusters rather than all variables can be isolated. In line with the dataset of K-means, one reference feature will be removed per category to overcome potential problems regarding perfect collinearity.

5. Results

This section will discuss the results following the outline of the entire report by reviewing the results per stage. Therefore, the first part of the results focuses on interpreting the clusters created using K-means on the geospatial data set. The second part compares the predictive performance of the advanced machine-learning techniques and the OLS benchmark model. The results aim to give an excellent fundamental to answer the research questions in the conclusion.

5.1 Geospatial clustering analysis

An essential aspect of the K-means clustering analysis is predetermining the number of clusters. However, the traditional methods, mentioned in the methodology section, do not provide a clear outcome in determining the number of clusters. The total within the sum of square plots shows a slowly continuous decrease, where no elbow can be interpreted. The other two measurements suggest 1 and 2 clusters, which is not meaningful in this context of house price predictions. Since the clustering stage is an aid in achieving the main objective of creating an accurate house price prediction model, the number of clusters will be determined based on the predictive performance of advanced machine learning and the clusters' interpretability. This process involves training three advanced machine learning techniques using default settings and varying the number of clusters (10, 20, 25, 30, 40, 50) to determine

the best predictive performance. Afterward, the clusters' interpretability will be taken into consideration. This approach shows us that the model with 25 clusters outperforms the other number of clusters based on the metrics RMSE, MSE, MAE, and R-squared for the XGBoost and RF approach. The SVM technique achieves the best performance for 20 and 25 clusters. Further, the 25 clusters have favorable interpretability properties. Therefore, we choose to continue with 25 clusters. The RMSE per model and cluster is presented in Appendix B.

Understanding the characteristics of the different clusters formed with the zip codes is essential to understand each cluster's impact on the house price in the end model. This section will focus on elaborating on the most interesting clusters. Table 5 shows this section's most important information about the explained clusters. The complete information on each cluster can be found here. Figure 4 shows the allocation of the clusters in six neighborhoods.

Table 5 shows us that clusters 4 and 22 represent zip codes with, on average, the lowest WOZ value, whereas cluster 2 has the highest WOZ value. The average WOZ value differs by roughly €875,000. Clusters 2 and 22 are mainly located in moderate urban areas, whereas cluster 4 is in less urban areas. Figure 5 shows that the zip codes in cluster 4 are mostly located in the Northern part of the Netherlands, in contrast to cluster 9, which mostly are in the south of the Netherlands. Table 5 shows that the main differences between the high and low average WOZ value clusters are the percentage of buying houses and average household size. Whereas cluster 2 consists primarily of buying houses, cluster 22 is the inverse. Cluster 4 shows an equal division of buying and renting houses. The average household size of the zip codes in clusters 4 and 22 are 2.05 and 1.80, respectively, whereas cluster 2 has a significantly higher average household size of 2.62. According to the map in Figure 4, a substantial portion of Kralingen-Oost refers to cluster 2, while Pierik corresponds to cluster 22.

Cluster 15 has a high average number of inhabitants, where one zip code in Kralingen Oost (Rotterdam) and a great part of Oosterhout (Nijmegen) belongs to. The highest average household size can explain this. Cluster 3 shows a comparable characteristic. These clusters can be categorized as "Family-friendly zip codes" based on family size, the high proportion of families with two parents, and the age distribution. The biggest difference between the two is in the number of residents, with cluster 3 having a larger proportion of people aged 45 to 64, while cluster 15 has a higher proportion of people aged 25 to 44. Therefore, cluster 15 can be interpreted as zip codes housing younger families than Cluster 3. In contrast to clusters 3 and 15, cluster 7 segmented the more elderly zip codes. In these zip codes, on average, 45% of the inhabitants are older than 65, 56% of households are one-person households on average, and

TABLE 5: Average Information per most interesting cluster

This table shows the most important average values of the socioeconomic, demographic, and distance attributes per elaborated cluster. The complete information per cluster can be found here. The meanings of the abbreviations are: Inhab = inhabitants, WOZ = WOZ Value, HH-Size = Household size, NL = Netherlands, WE = Western, NW = Non-Western, One Person = One Person Household, Two Parent = Two Parent Household, No Child = More adults but no children household, Buy = Buying houses vs. renting houses ratio, Urban = Urbanization Degree, Prim. School = Distance to Primary School, and Supermarket = Distance to Supermarket.

				Age groups					Origin			Household Composition						Distance (m)	
Cluster	Inhab	WOZ	HH-size	<15	15-24	25-44	45-64	> 65	NL	WE	NW	One person	Two Parent	No Child	Buy	Density	Urban	Prim. School	Supermarket
2	31.60	1046.60	2.62	0.18	0.12	0.17	0.30	0.23	0.74	0.18	0.07	0.32	0.33	0.30	0.76	2305	2.44	501.33	620.99
3	43.18	351.08	2.96	0.19	0.15	0.20	0.32	0.15	0.93	0.03	0.02	0.18	0.45	0.31	0.85	552	4.38	675.17	910.10
4	33.45	170.87	2.05	0.15	0.11	0.22	0.26	0.26	0.85	0.07	0.08	0.42	0.20	0.29	0.51	1245	3.15	423.58	442.31
5	38.35	312.20	2.34	0.14	0.11	0.19	0.32	0.24	0.88	0.09	0.02	0.26	0.30	0.38	0.79	671	4.19	519.90	662.23
6	60.60	363.91	2.82	0.19	0.14	0.22	0.32	0.14	0.80	0.10	0.10	0.22	0.42	0.29	0.82	1687	2.49	397.27	530.30
7	36.83	217.93	1.61	0.08	0.07	0.17	0.24	0.45	0.79	0.11	0.10	0.56	0.10	0.28	0.36	2162	2.02	415.45	334.16
9	38.34	185.37	2.02	0.13	0.11	0.23	0.29	0.24	0.71	0.18	0.11	0.42	0.20	0.29	0.50	1614	2.54	436.83	458.09
10	38.39	250.35	2.33	0.16	0.11	0.21	0.29	0.23	0.86	0.08	0.06	0.30	0.29	0.34	0.69	1303	3.03	395.75	494.65
11	42.83	235.51	1.45	0.05	0.38	0.36	0.13	0.08	0.64	0.19	0.16	0.76	0.05	0.15	0.24	4124	1.21	447.34	258.08
12	43.05	385.42	1.72	0.12	0.12	0.42	0.23	0.12	0.48	0.26	0.26	0.58	0.13	0.22	0.33	7650	1.01	284.25	185.93
15	69.03	358.10	3.00	0.29	0.09	0.36	0.20	0.06	0.81	0.08	0.11	0.17	0.51	0.23	0.76	1086	3.45	524.13	701.25
16	32.69	266.28	2.44	0.16	0.12	0.21	0.32	0.20	0.93	0.05	0.01	0.26	0.33	0.35	0.79	118	5.00	634.74	3278.09
19	42.02	254.68	2.20	0.17	0.11	0.27	0.27	0.18	0.68	0.12	0.20	0.40	0.24	0.25	0.50	2512	1.71	349.87	360.44
20	26.59	329.70	2.49	0.14	0.12	0.17	0.36	0.21	0.92	0.05	0.01	0.24	0.33	0.37	0.83	100	4.98	2639.31	3003.04
21	37.01	230.57	1.63	0.11	0.14	0.38	0.23	0.15	0.63	0.16	0.21	0.60	0.11	0.22	0.36	3533	1.22	383.27	280.25
22	32.48	177.54	1.80	0.13	0.13	0.28	0.26	0.20	0.72	0.11	0.17	0.53	0.15	0.22	0.35	2210	1.90	378.65	346.82
23	36.79	303.55	2.22	0.15	0.11	0.21	0.30	0.24	0.83	0.10	0.08	0.31	0.28	0.34	0.72	1579	2.61	393.57	447.59
Average	**47.48**	**302.79**	**2.23**	**0.15**	**0.13**	**0.24**	**0.28**	**0.20**	**0.78**	**0.11**	**0.11**	**0.37**	**0.26**	**0.29**	**0.61**	**1769**	**3.08**	**595.60**	**756.43**

28% comprise just adults and no children. These numbers suggest that the zip codes in this cluster are the more “elderly” zip codes. Clusters 11, 12, and 21 are urban areas near city centers with low urbanization degrees and relatively low distances to points of interest. Cluster 12 has the highest housing density of 7,650 houses/km^2, while clusters 11 and 21 have 4,124 and 3,533 houses/km^2, respectively. Cluster 12 is responsible for a significant portion of the Zuidas Zuid neighborhood in Amsterdam and has the highest average WOZ value among the urban area clusters. Cluster 11 is interpreted as a student neighborhood with many one-person households aged 15-24 living in rented houses. The map of the Binnenstad (Maastricht) and Kralingen-Oost (Rotterdam) in Figure 4 shows that the complete neighborhood of Binnenstad (Maastricht), and a couple of zip codes in Kralingen-Oost (Rotterdam) belongs to cluster 11. This aligns with the interpretation of a “student” neighborhood since Rotterdam and Maastricht have large universities and many students. In contrast, the least urban zip codes are represented by cluster 16 and 20, represented in the neighborhood Zalk in Kampen. Based on the locations of the zip codes, low house density, and an urbanization degree of nearly five, the zip codes in these clusters are interpreted as the farmer areas. Further, the inhabitants of these clusters are of Dutch or Western origin. The cluster information shows us that the points of interest are located at a significant distance from the location of the zip codes. On average, the houses in cluster 20 have a higher WOZ value compared to cluster 16.

Most zip codes are grouped into clusters 5, 6, 10, 19, and 23. These clusters account each for more than 7% of all zip codes, with Clusters 19 and 23 accounting for most of the zip codes with 9%. Cluster 6 has a relatively high number of occupants, with an average of 61 people, and is located in moderate urban regions. Further, Cluster 23 has an average WOZ value of €303,549, whereas cluster 19 has a significantly lower value of €254,678. Cluster 5 represents the least urban areas among the most representing clusters, while cluster 19 represents the most urban areas. Table 5 shows that most of the inhabitants have a Dutch origin, whereas only cluster 19 shows a higher percentage of non-western inhabitants. Cluster 10 contains the zip codes furthest from a commercial airport. Figure 4 shows that some of the zip codes in Pierik (Zwolle) are clustered in cluster 10. The zip codes in Pierik have a distance to the airport of Groningen of around 76 km (in a straight line). Additionally, Figure 6 displays that Cluster 10 primarily comprises zip codes located in the eastern region of the Netherlands. In contrast, Cluster 6 and 23 consist mostly of zip codes in major cities and the "Randstad" area.

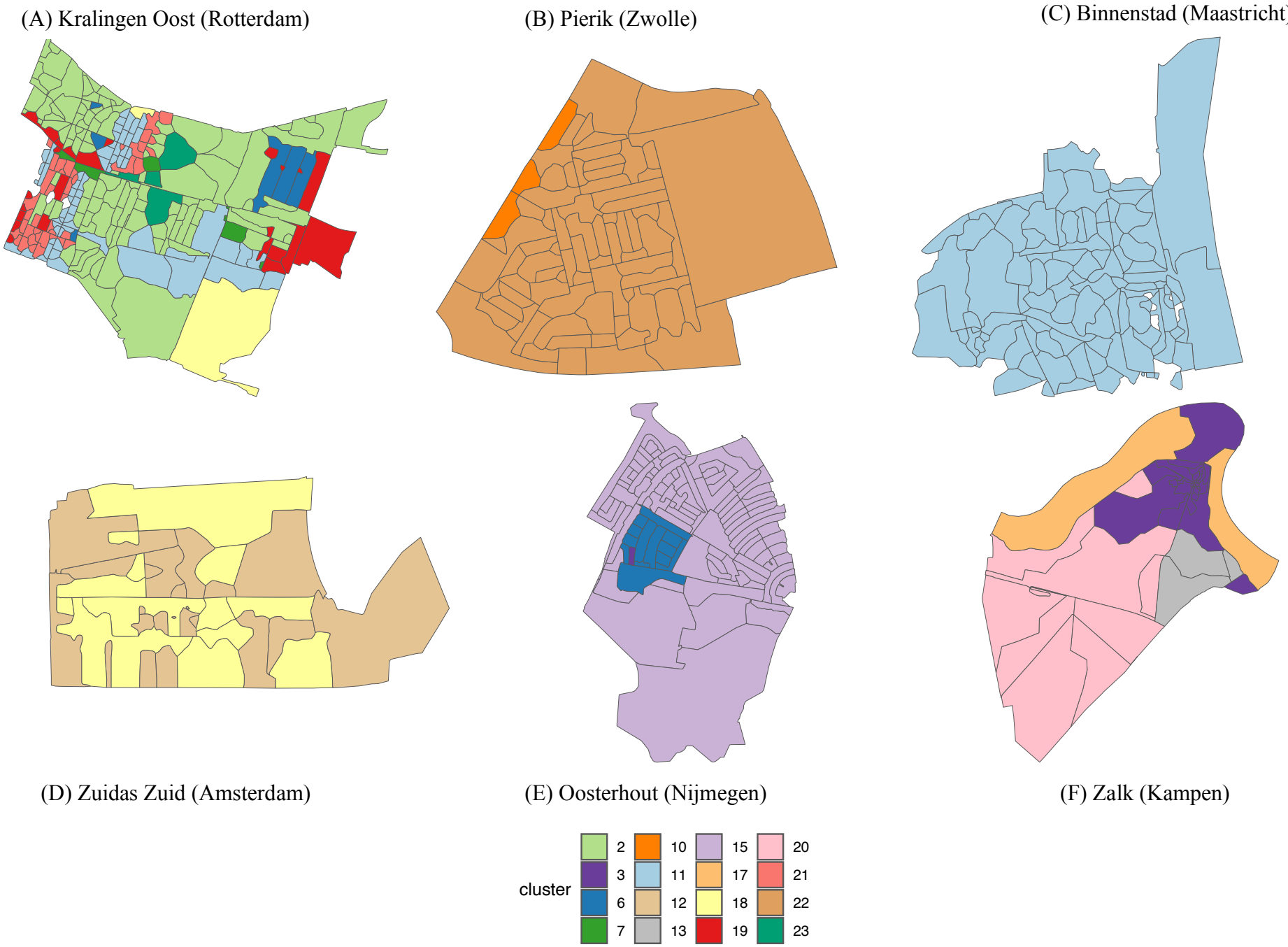


FIGURE 4. Cluster allocation on zip code level in the Neighborhoods Kralingen Oost (Rotterdam), Pierik (Zwolle), and Binnenstad (Maastricht)

The clusters created with K-means are well interpretable. Some clusters are quite similar and have minor differences, while others are distinct and dependent on factors such as the WOZ value. K-means performed an excellent job of grouping the zip codes and avoiding a single cluster with most of the zip codes. In the following stage, the clusters are employed as a categorical variable based on the zip codes of the properties for sale.

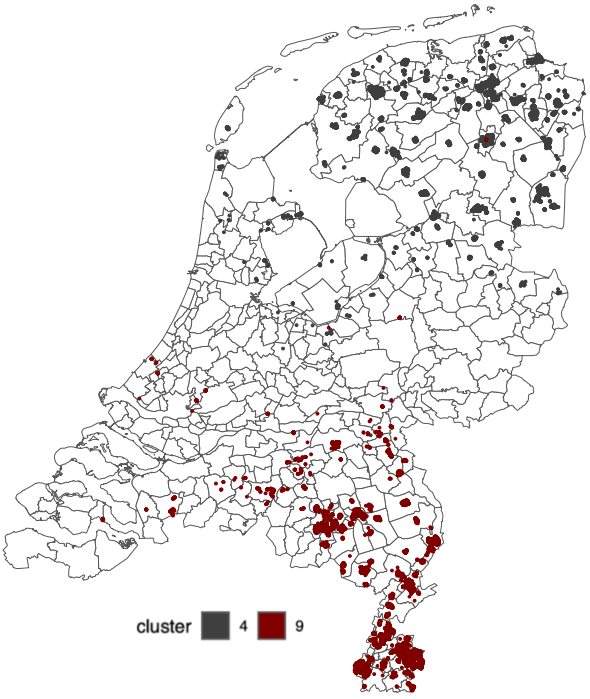


FIGURE 5. The allocation of zip codes in cluster 4 and 9

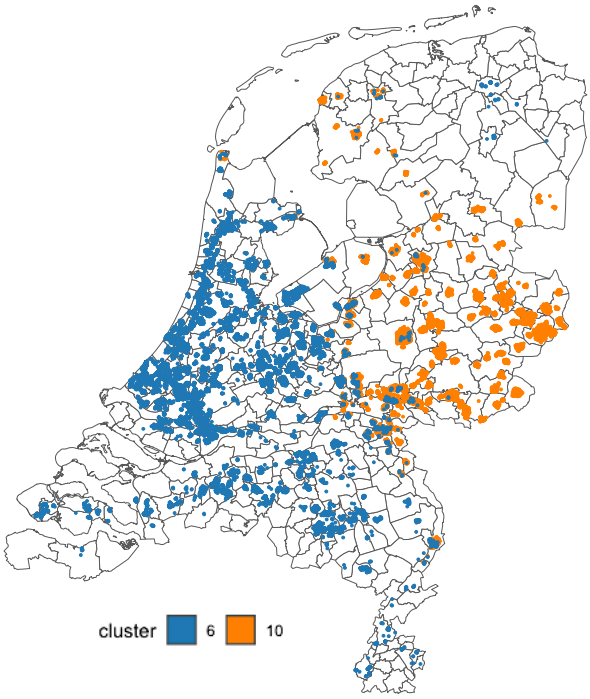


FIGURE 6. The allocation of zip codes in cluster 6 and 10

5.2 House price predictions

In this section, we provide a summary of the results obtained from the XGBoost, Random Forest, and SVM models that were trained. The evaluation of each model is based on the quantitative metrics outlined in the methodology section. The hyperparameters of each model are tuned and cross-validated to achieve the best possible results. The hyperparameters settings are presented in Appendix D. The results are compared to a hedonic house price benchmark model computed with an ordinary least square regression. To assess their performance, the models undergo training using a training data set and then undergo evaluation using a separate test set. The evaluation metrics of each model are displayed in Table 6 and are based on the test data set. The three advanced machine learning models exceed the performance of the benchmark model. However, the SVM model has the weakest

performance, as indicated by significantly lower evaluation metrics. The results suggest that the XGBoost and RF models have a strong R-squared value of 0.831 and 0.795, respectively. This means that 83.1% and 79.5% of the variation in house price can be attributed to the house's features and the cluster it belongs to. Based on quantitative evaluation metrics, the XGBoost model performs better than the RF model. This is evident by its lower error rates, as indicated by the lowest RMSE, MAE, and MAPE. The XGBoost has a mean absolute error (MAE) of €60,002. This means that, on average, there is a difference of €60,002 between predicted and actual house prices. However, the Random Forest approach performs better than XGBoost in terms of computational time.

TABLE 6: Predictive Performance Models

Model	R^2	RMSE	MAE	MAPE	Time
XGBoost	0.831	€ 111,543	€ 60,002	13.11 %	235 min
Random Forest	0.795	€ 122,366	€ 65,403	14.72%	198 min
SVM	0.764	€ 135,205	€ 79,161	17.69 %	54 min
OLS	0.690	€ 149,352	€ 94,861	22.39 %	< 1 sec

*Note: Time refers to the total time needed for training the model, which involves hyperparameter tuning and cross-validation.

Figure 7 shows the actual price versus the predicted price of the XGBoost (A) and RF (B) models. Based on the plot, there is a significant difference between the predicted and actual values of houses with higher worth. The plot shows that most predicted prices in the higher price range are underpriced, especially in the Random Forest model. The houses above €750,000 are filtered out to see if the predictive performance increases. Houses worth more than €750,000 may potentially not be a good representation of the overall population of houses. This approach has filtered out 8.7% of the total houses in the data set.

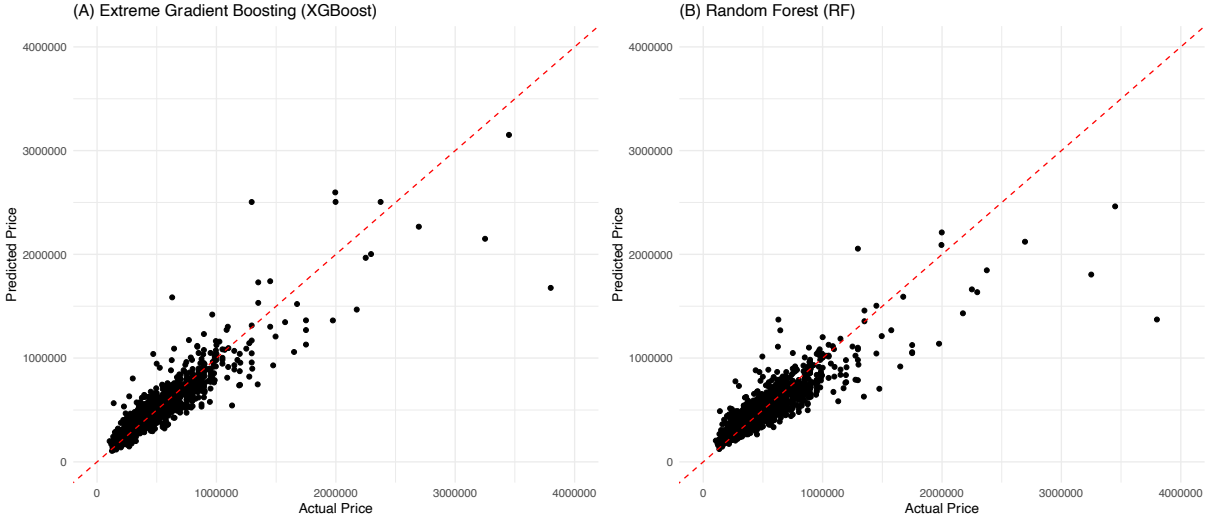


FIGURE 7. XGBoost and RF actual vs. predicted values

TABLE 7: Predictive Performance Models with only houses worth < € 750,000

Model	R^2	RMSE	MAE	MAPE
XGBoost	0.794	€ 60,792	€ 43,228	11.93 %
Random Forest	0.754	€ 67,011	€ 48,023	13.34 %
SVM	0.648	€ 80,082	€ 58,759	16.03 %
OLS	0.549	€ 89,946	€ 68,437	19.48 %

The results of filtering out houses worth more than €750,000 are shown in Table 7. It is observed that the RMSE and MAE metrics significantly decrease when this filter is applied to the models. The MAPE assessment measure is the most essential when evaluating forecasts' improvement since it focuses on percentage changes. In contrast, MAE utilizes absolute numbers, which logically drop if it eliminates residences worth more than € 750,000. According to the results, the MAPE improves slightly with approximately 1% when high-value houses are removed compared to when they are not removed. However, the value of R-squared decreases noticeably.

In order to determine to what extent geospatial clusters influence the accuracy of house price prediction models, we used the most successful method, XGBoost, to train a model that incorporates all socioeconomic, demographic, and distance attributes rather than just the cluster itself. Table 8 presents the results, indicating that a model which includes all socioeconomic, demographic, and distance attributes delivers better predictive performance than one that only incorporates the cluster. Including the features result in an increase in variance explained (R^2) of 3.9%. Additionally, the validation metrics for error rates indicate a decrease. The mean absolute error (MAE) for the XGBoost benchmark and XGBoost without high-value houses both decreased by €13,943 and €3,613, respectively. One weakness of including all features is the increased variables, leading to a significant rise in computational time. The K-means algorithm took around 3 minutes to compute, which makes a total of 238 minutes to compute the XGBoost algorithm with the cluster variable. Therefore, the model that did not include clusters took 55% longer to compute than the one that did.

TABLE 8: Predictive Performance XGBoost with all socioeconomic, demographic, and distance attributes

Model	R^2	RMSE	MAE	MAPE	Time
XGBoost (Benchmark)	0.870	€ 97,600	€ 55,454	12.50 %	368 min
XGBoost (Benchmark) *Only houses < € 750,000*	0.818	€ 57,179	€ 40,114	11.07 %	343 min

5.2.1 Global Interpretability

Advanced machine learning models have a major drawback: They are often referred to as "black box models" because they are difficult to interpret. To improve our understanding of the XGBoost model's predictions and the factors contributing to its accuracy, it is necessary to enhance its interpretability.

The first technique is the feature importance plot, which is a global model-independent way of determining the increase in the model's prediction error after permuting the feature. It shows which features are the most important in predicting the house price. Figure 8 shows the feature importance plots for both XGBoost models, including all houses (A) and omitting high-value houses (B). The graph clearly illustrates that the living area (m2) is the most essential element in both models. Both models place significant importance on the longitude and latitude coordinates feature. When considering the initial XGBoost, the vital structural features to take into account are the plot size, building year, floor location, and balcony existence. If we remove high-value houses from the XGBoost model, the presence of an underfloor heating system becomes more important than the house's floor location or a balcony's existence. The plots reveal that the number of bedrooms did not have much importance in the initial XGBoost model, but it does in the modified XGBoost model. Two clusters stand out as the most important: clusters 2 and 12 in the original model and only cluster 12 in the revised one. Cluster 2 indicates the pricier neighborhoods, while Cluster 12 signifies the areas with the highest concentration of houses close to city centers.
To enhance comprehension of the impact of each feature, partial dependency plots (PDPs) are utilized on a global level for interpretability. These plots provide insight into each feature's marginal influence and complement feature significance plots.

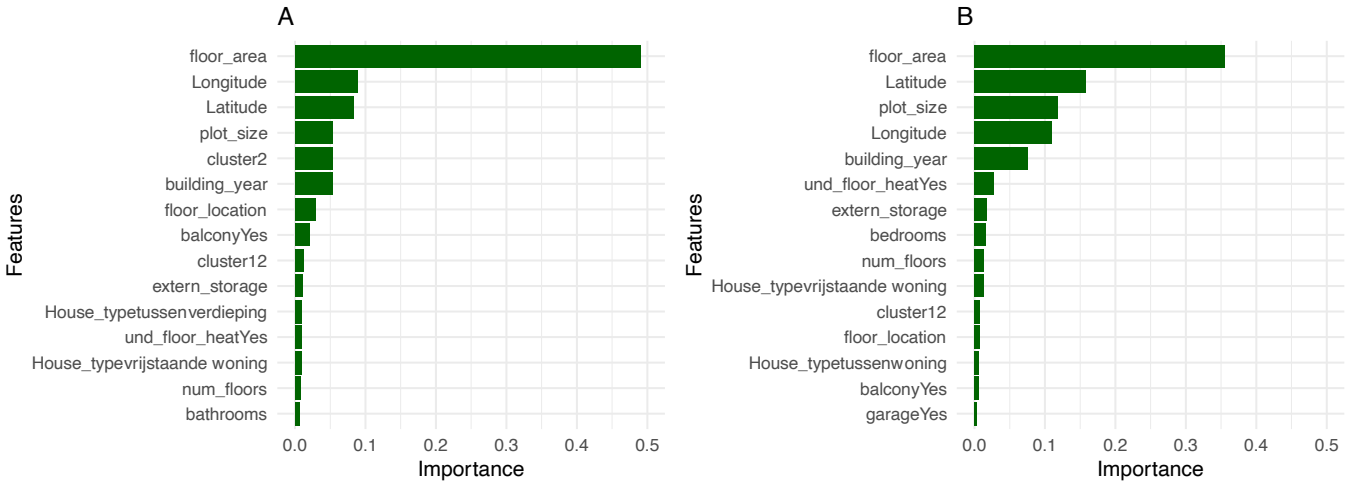


FIGURE 8. Feature Importance Plot XGBoost Models

Figure 9 shows the partial dependency plots of the most important structural house features. Floor area (m^2) is seen as the most important feature in the XGBoost model. The partial dependency plot of floor area suggests a positive relationship between floor area and house price. The plot size also exhibits a positive relationship in line with the floor area. Based on the partial dependency plot of the building year, there appears to be a U-shaped relationship between the building year and house price. Initially, newer homes are priced lower up to a certain point, after which there is a significant positive relation. A similar relationship can be observed for the external storage house characteristic. However, the relationship between external storage and house price is more complex. Overall, a negative relationship may be seen first, followed by a positive one. The model's features that capture the property's location are longitude, latitude, and the created cluster. The partial dependency plots of latitude and longitude in Appendix E illustrate how the house's location affects its price in terms of north/south and east/west directions. Overall, there is a positive association between latitude and housing price, with a negative relationship between 52.25 and 52.85. This negative pattern is observed to the north of the "Randstad". Further, the more eastern the house's location, the lower the house price, as indicated by the negative relationship of the longitude. In Appendix F, the two-dimensional spatial dependency plot shows a positive impact in the "Randstad" area when examining longitude and latitude. Figure 10 shows the partial dependency plot for the created cluster variable, which is important to understand each cluster's effect on the house's price. The plot shows that the two most important clusters in the model, 2 and 12, have the highest positive impact on house prices. Cluster 16 has the least

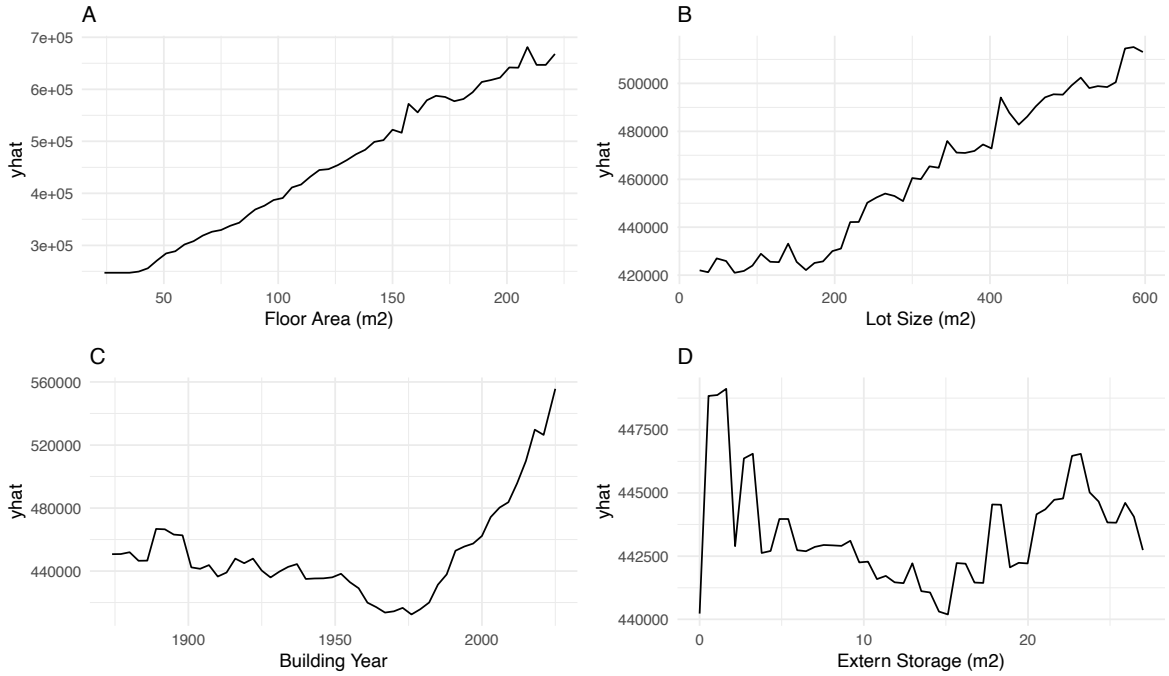


FIGURE 9. Partial Dependency Plots most important structural features

beneficial influence on home prices, implying that houses in less urban regions negatively impact its house price. Based on the plot, it is interesting to note that the cluster identified as the least expensive neighborhood (cluster 22) actually has a somewhat positive effect on the house prices.

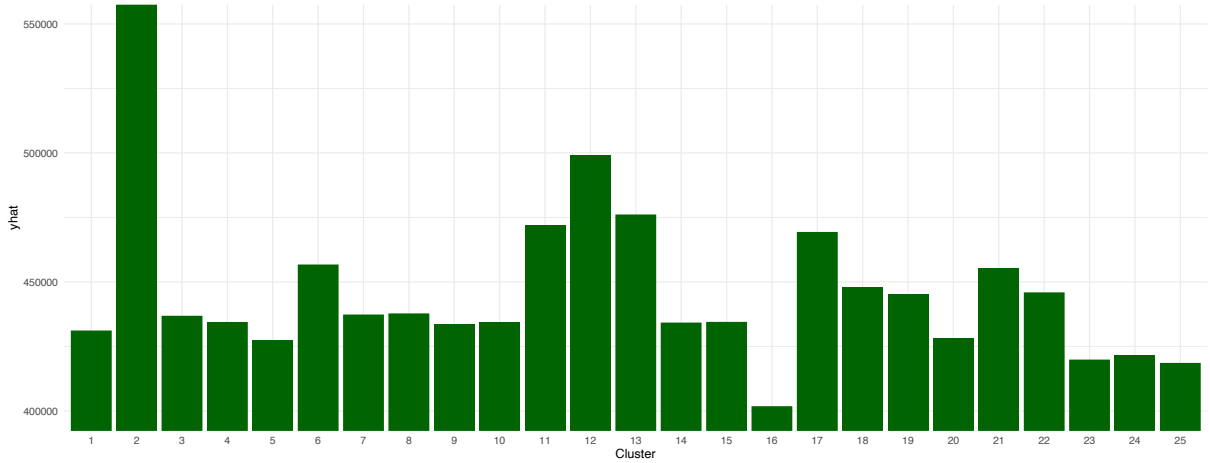


FIGURE 10. Partial Dependency Plot of the cluster feature

5.2.2 Local Interpretability

Global interpretability approaches were utilized in the preceding section to better comprehend the XGBoost model on a global basis. Local interpretability approaches can be used to understand better how a house price prediction is raised on an individual house level. The local interpretability LIME is used in this study, as mentioned in the methods section. To enhance comprehension of the XGBoost model's inaccurate predictions, this section will analyze two houses where one is under-predicted and the other over-predicted. The house attributes of each specific case can be found in Appendix I.

Figure 11 presents the LIME plots of the two cases (5962 and 3806), where case 5962 is overpredicted and case 3806 underpredicted. The property price of case 5962 is predicted at € 442,298, which is actually € 749,500. The most important supporting feature is the latitude between 52 and 52.3, which represents the middle part of the Netherlands. The main negative feature of the price prediction is the non-existence of a balcony. Additionally, the property's plot size is 121 square meters and has four bedrooms which negatively impacts the predicted price. By examining the details of the house, we can see that it underwent a complete renovation in early 2022. This may have led to an undervaluation of the property. Case 3808 is valued at € 417,247, which is actually € 269,000. The house features that positively affect the price prediction are that the houses' location belongs to cluster 12 and a latitude between 51.7 and 52.0. The floor area, house type and building year, all have a negative influence on the prediction, but not enough to compensate for a considerable overestimation. Upon

examining this particular case, it becomes evident that the house is a standard fixer-upper, potentially leading to the over-estimation.

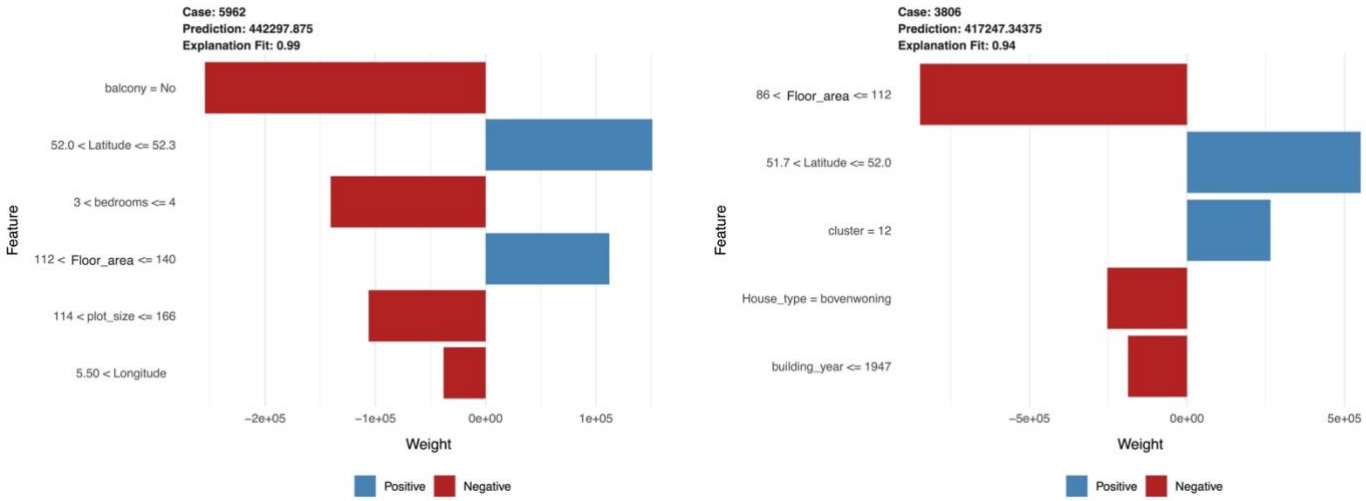


FIGURE 11. LIME Plots for cases 5962 and 3806

6. Conclusion

This paper investigates if including geospatial clusters improves the accuracy of the house pricing model. We conducted a study comparing three machine-learning techniques' predictive performance and efficiency. Based on our findings, we used the best-performing technique to investigate the impact of geospatial clusters. The conclusion section of this study focuses on answering to what extent the inclusion of geospatial segmentation based on socioeconomic, demographic, and geospatial attributes enhances the accuracy of house price prediction models in the Netherlands. Additionally, this research will address three sub-questions to enhance the foundation of the study.

After conducting extensive research, it can be concluded that geospatial segmentation based on socioeconomic, demographic, and geospatial factors does not improve the accuracy of house price prediction models in the Netherlands. The model's ability to make predictions is only slightly affected when comparing the one that includes clusters to the one that does not, in terms of predictive performance. The model without clusters has a MAPE of 0.61% lower and an R-squared of 3.9% higher. However, the model's training process, including clusters, has an efficiency advantage, where the training process takes 55% less time. Therefore, there is a trade-off between predictive performance and efficiency.

The first stage of this study segmented the zip codes using K-means. This approach could group approximately 445,000 zip codes into 25 distinct clusters. The 25 clusters are easily understandable, and the small increase in model fit by 3.9% shows that they capture

significant information. The loss of information caused by creating clusters is minimal. Therefore, it can be concluded that K-means is an accurate method to cluster zip codes based on socioeconomic, demographic, and geospatial features.

The study showed that advanced machine-learning techniques give better results than the traditional ordinary least square approach. The extreme gradient boosting model (XGBoost) outperformed the other models, such as Random Forest and SVM, in terms of predictive performance and is therefore recommended for use in future house price prediction models in the Netherlands. The biggest disadvantage of the XGBoost method is its comparatively long calculation time.

The findings demonstrate that the floor area is the most important house feature in the XGBoost model and has a positive relationship with the house price, which is consistent with the current literature (Guliker et al., 2022; Ho et al., 2021). The location in terms of coordinates (Longitude and Latitude) also shows high importance. The cluster feature created is quite significant in the model, particularly clusters 2 and 22, as they positively affect the house price. Building year and plot size are the most important structural house features besides the floor area. In line with floor area, the feature plot size shows a positive relationship with the house price, whereas the building year has a U-shaped relationship.

Overall, this study provided insight into the factors that impact home prices in the Netherlands and how locational segmentation and sophisticated machine-learning techniques influence the accuracy of house price prediction models. These findings can be valuable for various stakeholders, including homeowners, buyers, real estate agents, policymakers, and mortgage lenders.

7. Discussion

In the end, the XGBoost model outperforms the other advanced machine learning techniques. However, the mean absolute error on the test set is € 60,002, which is a significant amount in the real estate market. If the model uses only houses of under € 750,000 because especially the higher value houses cause a large increase in variance, the error rates decreases. However, excluding the houses worth over € 750,000 still shows the most variance in the higher segment. The problem here is that the higher segment houses are scarce in the training data set, potentially resulting in a higher error rate in this segment. This shows the challenge of the valuation of the most expensive houses.

Another limitation of this research is that it takes asking prices as house prices instead of the official sale price. Asking prices are chosen because they are easily accessible data

compared to actual sale prices. However, it was frequently observed in 2022 that buyers offered higher bids than the asking price. The first half of the year saw 80% of purchases being outbid, while in the second half, the figure was 44%. The beginning of 2023 shows a further decrease in the outbidding on house prices (*Hypotheek.nl*, 2023). Predicting asking prices is useful for various stakeholders, and there is a decreasing trend in outbidding. Therefore, using asking prices is not an issue. Furthermore, because all of the properties utilize asking prices, this technique allows for property comparisons, which might be useful for mortgage lenders.

We use text mining techniques on the house description to obtain many of the house features. Negations are also taken into consideration in order to make the Boolean variables as precise as possible. However, it is important to note that the accuracy of the obtained data can be impacted by incomplete descriptions. Inaccuracies may affect the model and its predictions.

The last limitation of this study is that it does not capture an energy label, house quality, and customer preferences. It is possible that approximately 17% of the unexplained variance is due to missing variables that can explain the condition of the house, including energy labels and house quality, as well as customer preferences. The two cases explained with the LIME interpretability method show the importance of house conditions. The energy label is left out of the analysis since text mining on the house description for the energy label resulted in many missing values.

This study shows that one of the biggest challenges in predicting house prices is obtaining precise data on the house's features, accurately valuing high-end properties, and incorporating the house's condition into the model.

Based on the previous conclusions and discussions, it is recommended that future research focuses on how to capture the house's condition into the model. Using a neural network technique to analyze the house images from real estate websites to assess the house’s condition and customer preferences would be interesting. Adopting this approach can improve the model fit as it provides a better understanding of the house.

Additionally, future research could also consider developing localized house price prediction models on a city or province level. The accuracy of predictions can be improved by using a geospatial clustering technique on a localized house price prediction model. It would be beneficial to explore the impact of incorporating geospatial clusters in these localized types of models for predicting house prices.

Lastly, house price per square foot is a common term used in the real estate market in the Netherlands. Therefore, it would be interesting to model the average house price per square foot in a zip code based on socioeconomic, demographic, and geospatial attributes. The average square footage, determined by zip code, can potentially enhance the accuracy of the house price prediction model. This approach will be another way to include the geospatial aspect of a property.

Appendix

Abbreviations:

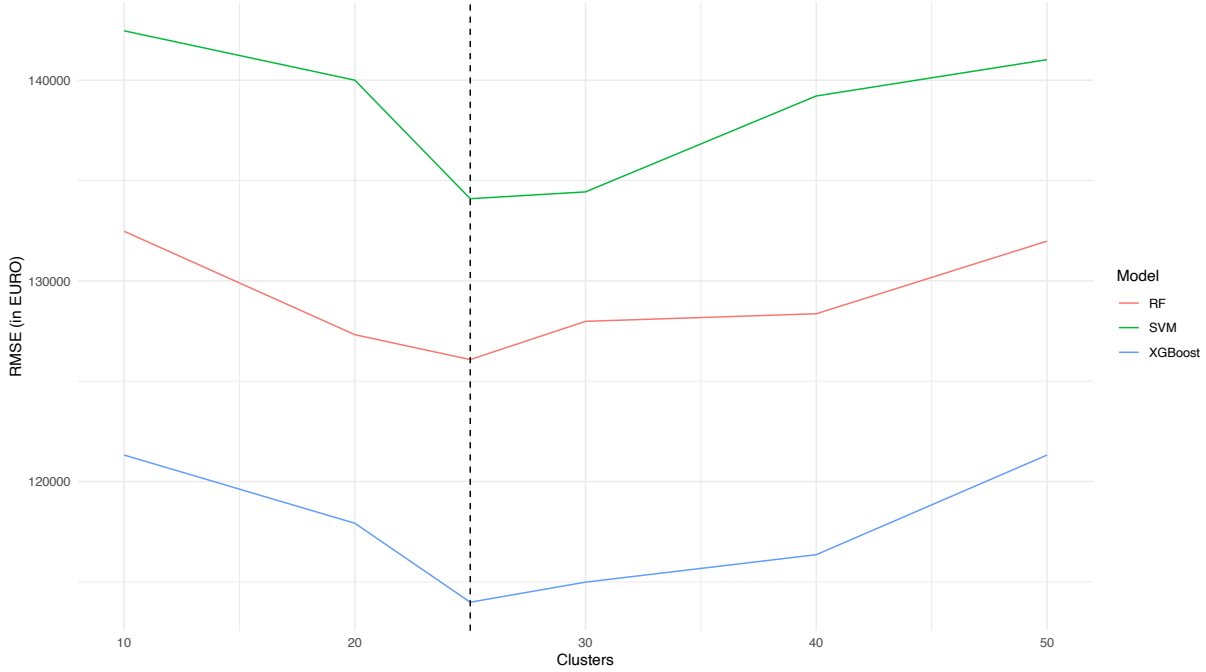
NL	Dutch language (Netherlands)
XGBoost	Extreme Gradient Boosting
RF	Random Forest
SVM	Support Vector Machine
ML	Machine Learning
OLS	Ordinary Least Square
CBS	"Centraal Bureau Voor de Statistiek" (In English: Central Agency For Statistics)
PDP	Patial Dependency Plot
LIME	Local Interpretable Model-Agnostic Explanations

Appendix A. Summary Statistics Categorical House Features

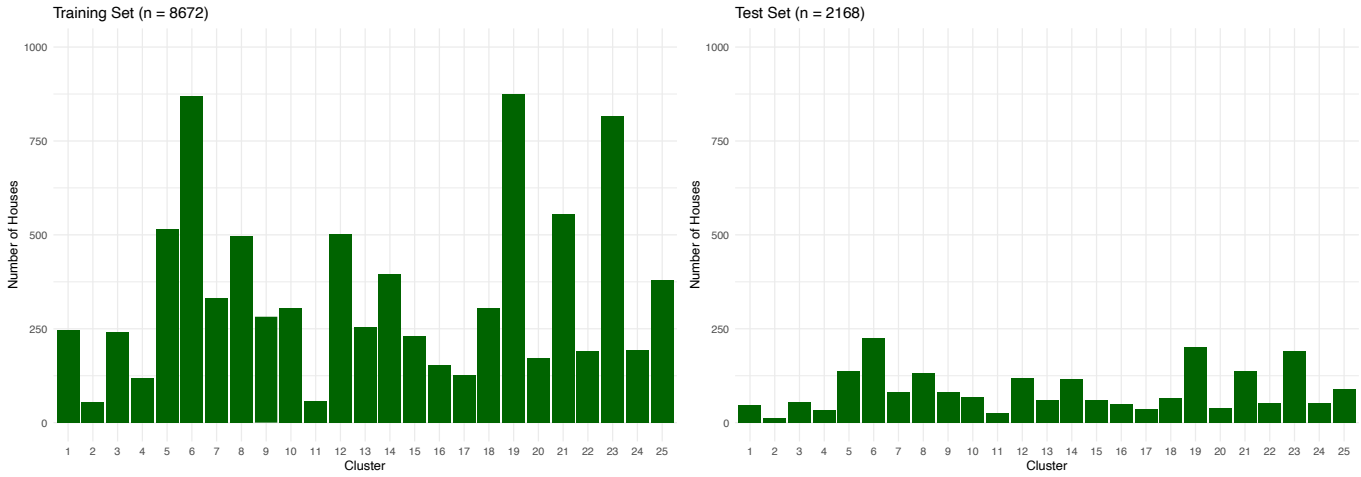
Feature	N = 10,840	Feature	N = 10,840
House_type		balcony	4,267 (39%)
2-onder-1-kapwoning	1,145 (11%)	garden	8,239 (76%)
beneden + bovenwoning	33 (0.3%)	Patio	248 (2.3%)
benedenwoning	444 (4.1%)	airconditioning	1,063 (9.8%)
bovenwoning	719 (6.6%)	solar_panels	1,989 (18%)
dubbel benedenhuis	37 (0.4%)	garage	2,421 (22%)
eindwoning	190 (1.8%)	fireplace	261 (2.4%)
galerijflat	499 (4.6%)	pool	757 (7.0%)
geschakelde 2-onder-1-kapwoning	164 (1.5%)	jacuzzi	138 (1.3%)
geschakelde woning	298 (2.7%)	und_floor_heat	2,897 (27%)
halfvrijstaande woning	234 (2.2%)	cent_heat_boiler	8,411 (78%)
hoekwoning	886 (8.2%)	block_heat	53 (0.5%)
maisonnette	249 (2.3%)	city_heat	63 (0.6%)
penthouse	77 (0.7%)	heat_pump	491 (4.5%)
portiekflat	888 (8.2%)	bath	3,755 (35%)
portiekwoning	206 (1.9%)		
Recreatiewoning	142 (1.3%)		
tussenverdieping	120 (1.1%)		
tussenwoning	2,483 (23%)		
vrijstaande woning	2,026 (19%)		

*Note: The boolean variables are set to “Yes”

Appendix B. Determining number of clusters using RMSE (€)



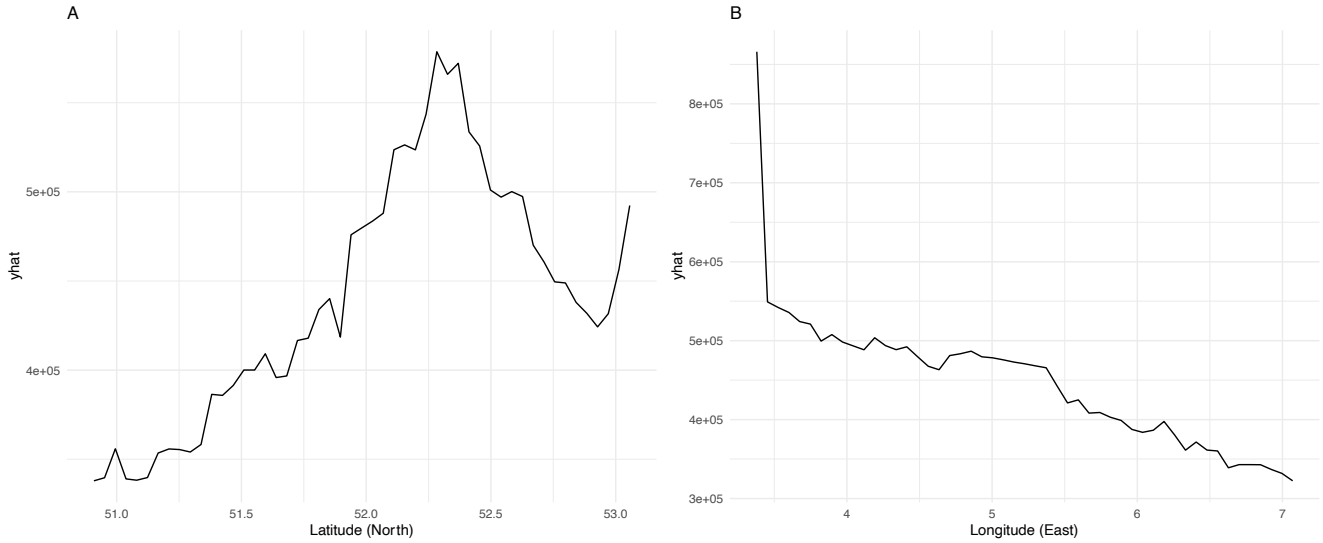
Appendix C. Allocation of the clusters within the training and test data set



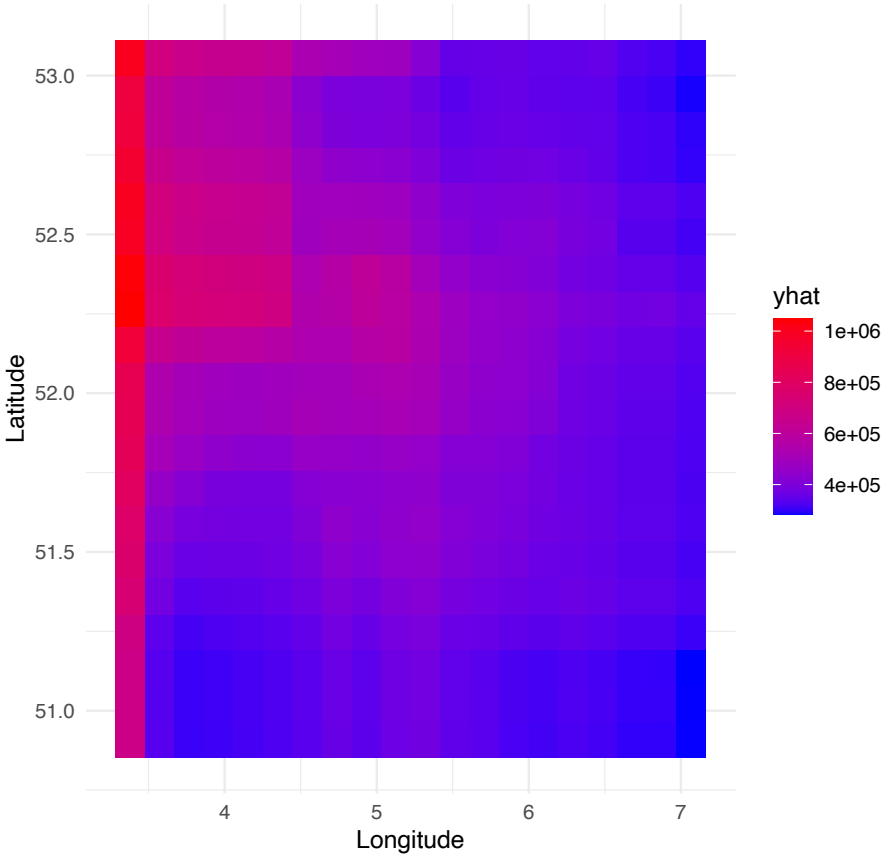
Appendix D. Parameters settings

Model	Parameters
XGBoost	nrounds = 500, max_depth = 7, eta = 0.1, gamma = 0, colsample_bytree = 1, min_child_weight = 1, and subsample = 0.65
XGBoost < €750,000	nrounds = 500, max_depth = 7, eta = 0.05, gamma = 0, colsample_bytree = 0.8, min_child_weight = 3, and subsample = 0.65
RF	ntree = 500, mtry = 27, splitrule = variance and min.node.size = 1
RF < €750,00	ntree = 500, mtry = 27, splitrule = variance and min.node.size = 1
SVM	Kernel function = Radialcost, Cost = 10
SVM < €750,000	Kernel function = Radialcost, Cost = 10
XGBoost (Excl. Clusters)	nrounds = 500, max_depth = 5, eta = 0.1, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, and subsample = 0.8
XGBoost (Excl. Clusters) < € 750,000	nrounds = 500, max_depth = 7, eta = 0.05, gamma = 0, colsample_bytree = 0.8, min_child_weight = 7, and subsample = 0.65

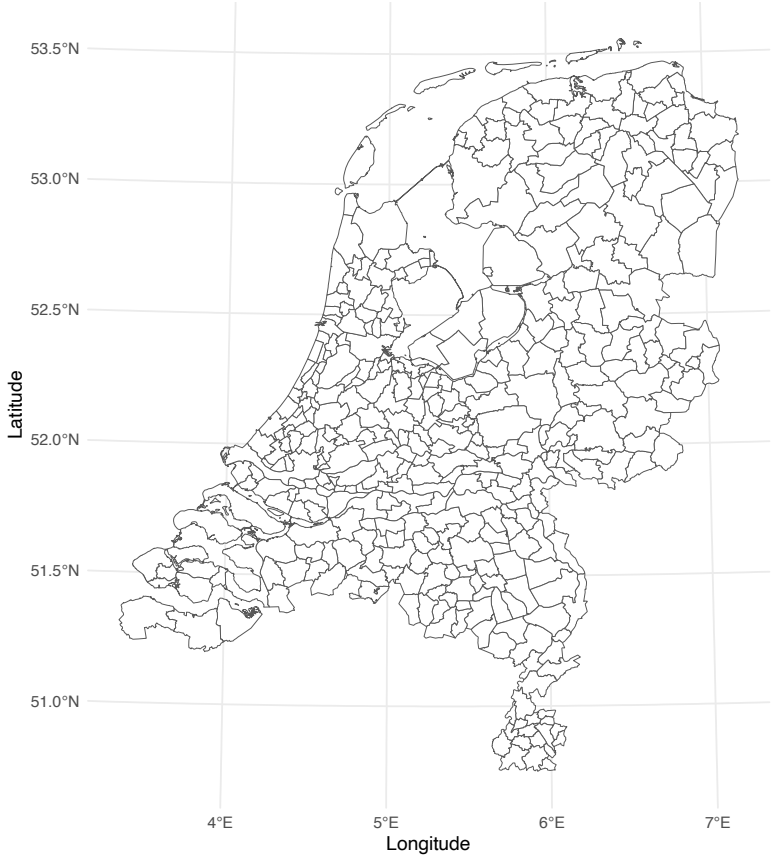
Appendix E. Partial Dependency Plot Latitude (A) and Longitude (B)



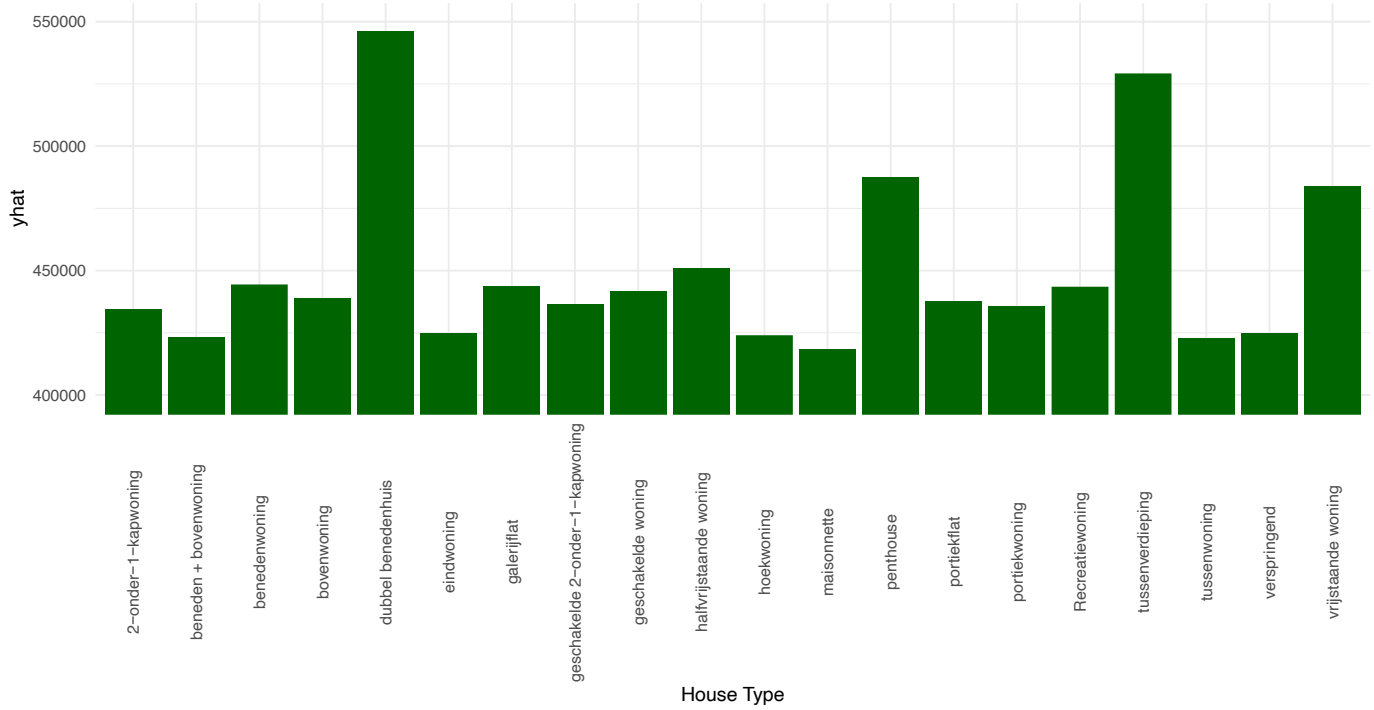
Appendix F. Two-dimensional Partial Dependency Plot Latitude and Longitude



Appendix G. Coordinates (Longitude/Latitude) Map of The Netherlands



Appendix H. Partial Dependency Plot House Type



Appendix I. Features cases 3806 and 5963 for LIME

Note: * are the heating related variables.

Case	House_type	balcony	garden	Patio	plot_size	living_space	extern_storage	airconditioning	building_year
3806	bovenwoning	No	No	No	105	90	0	No	1887
5962	tussenwoning	No	Yes	No	121	121	5	Yes	2021
Case	**# Floors**	**floor_location**	**solar_panels**	**garage**	**fireplace**	**bedrooms**	**bathrooms**	**pool**	**jacuzzi**
3806	2	2	No	No	No	2	1	No	No
5962	3	1	No	No	No	4	1	No	No
Case	**und_floor***	**Boiler***	**Block***	**City***	**Pump***	**bath**	**Longitude**	**Latitude**	**cluster**
3806	No	Yes	No	No	No	Yes	4.47912	51.93279	12
5962	Yes	Yes	No	No	No	No	5.565662	51.95599	22
Case	**Price**	**Predicted Price**							
3806	€ 269,000	€ 417,247							
5962	€ 749,500	€ 442,298							

References

1. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100
2. Akbilgic, O., Shin, E. K., & Shaban-Nejad, A. (2021). A Data Science Approach to Analyze the Association of Socioeconomic and Environmental Conditions With Disparities in Pediatric Surgery. *Frontiers in Pediatrics*, *9*, 620848. https://doi.org/10.3389/fped.2021.620848
3. Avanijaa, J., & Al, E. (2021). Prediction of House Price Using XGBoost Regression Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2), Article 2. https://doi.org/10.17762/turcomat.v12i2.1870
4. Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *The Journal of Real Estate Research*, *32*(2), 139–159.
5. Brasington, D. M., & Sarama Jr., R. F. (2008). Deed Types, House Prices and Mortgage Interest Rates. *Real Estate Economics*, *36*(3), 587–610. https://doi.org/10.1111/j.1540-6229.2008.00223.x
6. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
7. Cellmer, R., Bełej, M., & Konowalczuk, J. (2019). Impact of a Vicinity of Airport on the Prices of Single-Family Houses with the Use of Geospatial Analysis. *ISPRS International Journal of Geo-Information*, *8*(11), Article 11. https://doi.org/10.3390/ijgi8110471
8. Clapp, J., Case, B., Dubin, R., & Rodríguez, M. (2004). Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models. *The Journal of Real Estate Finance and Economics*, *29*, 167–191. https://doi.org/10.1023/B:REAL.0000035309.60607.53
9. Clapp, J. M., & Giaccotto, C. (1998). Residential Hedonic Models: A Rational Expectations Approach to Age Effects. *Journal of Urban Economics*, *44*(3), 415–437. https://doi.org/10.1006/juec.1997.2076
10. Clark, D. E., & Herrin, W. E. (2000). The Impact of Public School Attributes on Home Sale Prices in California. *Growth and Change*, *31*(3), 385–407. https://doi.org/10.1111/0017-4815.00134
11. Clauretie, T. M., & Neill, H. R. (n.d.). *Year-Round School Schedules and Residential Property Values*.
12. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
13. Das, S. S. S., Ali, M. E., Li, Y.-F., Kang, Y.-B., & Sellis, T. (2020). *Boosting House Price Predictions using Geo-Spatial Network Embedding* (arXiv:2009.00254). arXiv. http://arxiv.org/abs/2009.00254
14. de Vor, F., & de Groot, H. L. F. (2011). The Impact of Industrial Sites on Residential Property Values: A Hedonic Pricing Analysis from the Netherlands. *Regional Studies*, *45*(5), 609–623. https://doi.org/10.1080/00343401003601925
15. Dubin, R. (1998). Predicting House Prices Using Multiple Listings Data. *The Journal of Real Estate Finance and Economics*, *17*, 35–59. https://doi.org/10.1023/A:1007751112669
16. Espey, M., & Lopez, H. (2000). The Impact of Airport Noise and Proximity on Residential Property Values. *Growth and Change*, *31*(3), 408–419. https://doi.org/10.1111/0017-4815.00135
17. Fletcher, M., Gallimore, P., & Mangan, J. (2000). Heteroscedasticity in hedonic house price models. *Journal of Property Research*, *17*(2), 93–108. https://doi.org/10.1080/095999100367930
18. Freeman, A. M. (1979). Hedonic Prices, Property Values and Measuring Environmental Benefits: A Survey of the Issues. *The Scandinavian Journal of Economics*, *81*(2), 154. https://doi.org/10.2307/3439957
19. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.
20. Garrod, G. D., & Willis, K. G. (1992). Valuing goods' characteristics: An application of the hedonic price method to environmental attributes. *Journal of Environmental Management*, *34*(1), 59–76. https://doi.org/10.1016/S0301-4797(05)80110-0
21. Goodman, A. C. (1989). Topics in Empirical Urban Housing Research. In *The Economics of Housing Markets*. Routledge.
22. Goodman, A. C., & Thibodeau, T. G. (1998). Housing Market Segmentation. *Journal of Housing Economics*, *7*(2), 121–143. https://doi.org/10.1006/jhec.1998.0229

23. Goodman, A. C., & Thibodeau, T. G. (2007). The Spatial Proximity of Metropolitan Area Housing Submarkets. *Real Estate Economics*, *35*(2), 209–232. https://doi.org/10.1111/j.1540-6229.2007.00188.x
24. Guliker, E., Folmer, E., & van Sinderen, M. (2022). Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. *ISPRS International Journal of Geo-Information*, *11*(2), Article 2. https://doi.org/10.3390/ijgi11020125
25. Hamzah, N. A., Kek, S. L., & Saharan, S. (2017). The Performance of K-Means and K-Modes Clustering to Identify Cluster in Numerical Data. *Journal of Science and Technology*, *9*(3), Article 3. https://publisher.uthm.edu.my/ojs/index.php/JST/article/view/2038
26. Herath, S., & Maier, G. (n.d.). *The hedonic price method in real estate and housing market research: A review of the literature*.
27. Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558
28. Hurley, A. K., & Sweeney, J. (2022). Irish Property Price Estimation Using A Flexible Geo-spatial Smoothing Approach: What is the Impact of an Address? *The Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-022-09888-y
29. Kaushik, M., & Mathur, B. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. *Internatioonal Journal of Software and Hardware Research in Engineering*, *2*, 93–98.
30. Khmaissia, F., Haghighi, P. S., Jayaprakash, A., Wu, Z., Papadopoulos, S., Lai, Y., & Nguyen, F. T. (2020). *An Unsupervised Machine Learning Approach to Assess the ZIP Code Level Impact of COVID-19 in NYC* (arXiv:2006.08361). arXiv. http://arxiv.org/abs/2006.08361
31. Kitapci, O., Tosun, O, Tuna, M.F., & Turk, T. (2017). The Use of Artificial Neural Networks (ANN) in Forecasting Housing Prices in Ankara, Turkey. *Journal of Marketing and Consumer Behaviour in Emerging Markets*, *5*(1), 4–14.
32. Levantesi, S., & Piscopo, G. (2020). The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. *Risks*, *8*(4), Article 4. https://doi.org/10.3390/risks8040112
33. Li, M. M., & Brown, H. J. (1980). Micro-Neighborhood Externalities and Hedonic Housing Prices. *Land Economics*, *56*(2), 125–141. https://doi.org/10.2307/3145857
34. 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. https://doi.org/10.1016/0094-1190(80)90055-8
35. Mohandes, M. A., Halawani, T. O., Rehman, S., & Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, *29*(6), 939–947. https://doi.org/10.1016/j.renene.2003.11.009
36. Morris, E. W., Woods, M. E., & Jacobson, A. L. (1972). The Measurement of Housing Quality. *Land Economics*, *48*(4), 383–387. https://doi.org/10.2307/3145315
37. NJ Peter, HI Okagbue, ECM Obasi, & AO Akinola. (2020). Review on the Application of Artificial Neural Networks in Real Estate Valuation. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 2918–2925. https://doi.org/10.30534/ijatcse/2020/66932020
38. *Overbieden huis: Hoeveel moet je bieden? | Hypotheek.nl*. (n.d.). Retrieved May 8, 2023, from https://www.hypotheek.nl/kennisbank/huis-kopen/overbieden-huis/
39. Rahman, S. N. A., Maimun, N. H. A., Razali, M. N. M., & Ismail, S. (2019). THE ARTIFICIAL NEURAL NETWORK MODEL (ANN) FOR MALAYSIAN HOUSING MARKET ANALYSIS. *PLANNING MALAYSIA*, *17*. https://doi.org/10.21837/pm.v17i9.581
40. Ramírez-Juidías, E., Amaro-Mellado, J.-L., & Leiva-Piedra, J. L. (2022). Influence of the Urban Green Spaces of Seville (Spain) on Housing Prices through the Hedonic Assessment Methodology and Geospatial Analysis. *Sustainability*, *14*(24), Article 24. https://doi.org/10.3390/su142416613
41. Sina Shokoohyar, Anae Sobhani, & Saeed R. Ramezanpour Nargesi. (2019). *On the determinants of Uber accessibility and its spatial distribution: Evidence from Uber in Philadelphia*. https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1362
42. Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796
43. Statistiek, C. B. voor de. (2023). *Woningmarkt* [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/visualisaties/dashboard-economie/woningmarkt

44. Yang, L., Liang, Y., Zhu, Q., & Chu, X. (2021). Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. *Annals of GIS*, *27*(3), 273–284. https://doi.org/10.1080/19475683.2021.1906746
45. Yang, L.-C., Chou, S.-Y., Liu, J.-Y., Yang, Y.-H., & Chen, Y.-A. (2017). *Revisiting the problem of audio-based hit song prediction using convolutional neural networks* (arXiv:1704.01280). arXiv. http://arxiv.org/abs/1704.01280
46. Zaken, M. van A. (2022, June 20). *Hoeveel kan ik maximaal lenen voor mijn koopwoning? - Rijksoverheid.nl* [Onderwerp]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/onderwerpen/huis-kopen/vraag-en-antwoord/onderwerpen/huis-kopen/vraag-en-antwoord/maximaal-bedrag-lenen-koopwoning
47. Zaki, J., Nayyar, A., Dalal, S., & Ali, Z. H. (2022). House price prediction using hedonic pricing model and machine learning techniques. *Concurrency and Computation: Practice and Experience*, *34*(27), e7342. https://doi.org/10.1002/cpe.7342
48. Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007
49. Zhang, Y., Huang, J., Zhang, J., Liu, S., & Shorman, S. (2022). Analysis and prediction of second-hand house price based on random forest. *Applied Mathematics and Nonlinear Sciences*, *7*(1), 27–42. https://doi.org/10.2478/amns.2022.1.00052