ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Modelling the Dutch housing market: understanding how the asking price is formed

Name student: Diëgo Mahabier

Student ID number: 430494

Supervisor: dr. Erjen Van Nierop

Second assessor: dr. Carlo Cavicchia

Date final version: October 21, 2023

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

Table of Contents

1 Introduction

The Dutch housing market is overheated. For the past five years, the prices of houses increased with 8% per year (De Nederlandsche Bank, sd). Compared to 2013, existing housing prices increased with 91% in 2022 (Centraal Bureau voor de Statistiek, 2022). This makes it hard for households to buy a house, and if they find a house, household must get a much bigger loan compared to five years ago (Langenberg & Jonkers, 2022).

But what causes this overheated housing market? According to De Nederlandsche Bank, the prices of houses are overheated due to a shortage of houses. In 2021, the Netherlands had a shortage of 279.000 houses (Ollongren, 2021). Another factor for the overheated housing market is the low interest rate, which makes it easier for households to get a cheaper mortgage from the banks. Households tend to bid higher than usual for a house, and this drives up the prices for the houses. Finally, the overheated housing market is caused by tax benefits and broad conditions for the mortgages (De Nederlandsche Bank, sd). The Dutch government gives a tax benefit on household that buy their house. This tax benefit translates into financing costs that are deductible from their income tax as well as the interest payments for the mortgage. Compared to other countries, in the Netherlands, households can take up a mortgage for 100% of the value of the house, which makes it easier for households to bid above the value and/or asking price of a property (De Nederlandsche Bank, sd).

According to Centraal Bureau voor de Statistiek (2022) the housing prices increased due to an increase in the population of people of 11% with the age between 25 and 35 years old over the time span from 2014 up to 2021. These are the people that are the most active in trying to buy a house. This is due to a birth wave in the ’90’s after a high unemployment rate and a lot of uncertainty in the ‘80’s. Also, elderly are less willing to go into an elderly’s home and rather stay longer in their own house, which decreases the supply of houses. Finally, real estate became a good investment for private investors due to lower interest rates and a higher demand for private rental properties. Private investors got a “boost” from housing corporations because housing corporations were not building more properties for the Dutch population, so buying a house and renting it became a lucrative business.

Buying houses involves the capability of a household to either pay the asking price or finance the asking price to become a property owner. Of course, when buying a house, an important, if not the most important factor is the asking price. In times of an economic boom, households tend to bid higher than the asking price, however, in times of an economic downturn households are more likely to meet the asking price or bid under the asking price. From this, we can conclude that the asking price plays an important role for the market price of a property; the price of a house when the purchase has been made. The asking price thus can be seen as a baseline for the market value of a property.

In other words, the asking price is a vital part in estimating the market value of a property. The advantage is that the asking price is mostly estimated by a realtor that inspects the whole house. A realtor investigates multiple factors of a property when assigning a value. External factors such as the living area can raise or lower the value of a property but also the selling price of properties in the area.

This has also been proven in the research of Zhang, Zhang, & Miller (2021) for the city of Toronto, Canada. Their findings showed that when predicting the price of a property in Toronto, an important factor is the social environment. Truong, Nguyen, Dang, & Mei (2022) showed the importance of the geographical characteristics. Characteristics of the property do also have an impact on the price, such as having a dormer that was created for an extra room, or a newly built kitchen. This can be confirmed by the research of Zhang et al. (2021) and Truong et al. (2022). These two papers also showed that characteristics of the property are important in estimating a predicted value for the property, such as the number of bedrooms or the age of the house. By comparing different sales and multiple input factors, the realtor estimates the selling price of a property (Hoe wordt de vraagprijs van mijn woning bepaald?, 2022)

Personal situations also matter when selling a house. If the household wants to sell their house as fast as possible because of a divorce for example, the realtor will have a different marketing strategy compared to a household that sells and are waiting for a newly built house (Hoe wordt de vraagprijs van mijn woning bepaald?, 2022).

Finally, there are also psychological asking prices on the real estate market. Psychological prices work as follows, in the grocery store products are not sold for a round number like € 20, but most likely for a psychological price of € 19.99. This psychological price tries to make it seem like the product is cheaper for the customer. This also happens for properties, more in the sense of € 500,000 compared to € 499,000 (Hoe wordt de vraagprijs van mijn woning bepaald?, 2022).

Simply put, the prices for properties are built up from different factors. In this research, we want to investigate housing prices, more specifically, asking prices of properties. That is why the research question of this paper is as follows:

“What drives the asking price of the Dutch housing market and how can we predict the housing prices?”

The following subquestions arise:

- What are important characteristics of a house for the asking price?
- How does the neighborhood influence house prices?
- How do models of house prediction perform compared to each other?
- How do different machine learning methods differ in predicting housing prices?

This research is relevant for households that are interested in buying and selling a house. Since the asking price is a vital part of when selling or buying a house, households might be better off investing time in understanding the importance of asking prices and their underlying features when bidding on or selling houses. Also, this research is relevant for realtors when estimating the value of a property compared to the models and methods that are already used by them.

From the perspective of the researcher, not much about predicting housing prices in the Netherlands has been done yet. Liu (2013) did research spatial and temporal dependence in predicting house prices, but the location of interest was only the Dutch Randstad and not the whole country. The research its main focus is to improve the power of house price prediction with integration of the spatial and temporal dependence in hedonic models compared to traditional hedonic models that do not consider these effects. This research thus does not focus on finding the optimal model for predicting prices on the Dutch housing market.

Since there is not much research done in predicting the Dutch housing market, we can speak of a gap in the literature. I will try to fill this gap by providing research that tries to predict prices of the Dutch housing market as good as possible. This makes the research that will be done more relevant.

2 Literature review

The literature review is divided into two parts. First, we will look at the literature of spatial dependence. Next, price predictions will be discussed. Finally, we will look at housing characteristics.

2.1 Spatial dependence

Looking into the literature for predicting housing prices, we find that spatial dependence is a key factor in predicting house prices. Spatial dependence in the housing market means that the house prices are influenced by the houses nearby. In automated valuation models, controlling spatial dependence is an issue (Steven, Cantoni, & Hoesli, 2010). Steven et al. (2010) researched the market of Louisville, Kentucky of the United States of America. In their research, they created 3 different types of models: an ordinary least squared regression (hereafter: OLS), an OLS with 10 nearest neighbor residuals variables and a geostatistical model. When predicting, they added dummy variables and used estimated separate equations for each submarket to model spatial dependence as accurate as possible. The research of Steven et al. (2010) highlights the advantages of including spatial dependency into the error term. The residuals OLS with 10 nearest neighbor residuals variables appears less helpful compared to the geostatistical model. A geostatistical model without taking into account submarkets, performs about as well as an OLS model that considers disaggregated submarkets. Disaggregated submarkets are groups of houses that differ from other groups and thus, are not aggregated with each other. The best results were found in the geostatistical model with dummy variables for disaggregated submarkets.

Dubin (1998) showed that correlations between prices of neighboring houses can be combined when estimating the coefficients of an OLS regression. In the research of Dubin (1998), data of house listings in Baltimore, Maryland of the United States of America was used. For the correlation, correlograms were used based on the distance (measured in feet and measured in number of houses) and functional form (negative exponential or gaussian). Comparing the OLS results with the results of the OLS combined with the correlogram, the correlograms provides improvement over the OLS model. The negative exponential correlogram shows the most improved results. Considering distance, feet was more useful than houses.

When predicting housing prices, socioeconomic environment has a high explanatory power for the city of Toronto, Canada (Zhang, Zhang, & Miller, 2021). This study found out that the prices of houses are determimed by the following attributes: social environment, the distance and accessibility of the neighborhood and the age and physical condition of the house. However, the density and diversity of the surrounding area have relatively little to no impact on predicting housing prices.

In this paper, the country of interest is the Netherlands. An insightful paper about a part of the country of interest was written by Liu (2013). Liu (2013) did research on spatial and temporal dependence in predicting house prices. The research was based on data from the Dutch Randstad. This research mainly focuses on improving the power of house price prediction with integration of the spatial and temporal dependence in hedonic models compared to traditional hedonic models that do not consider these effects. The models that were used are the STAR model and the Parsimonious model. They found that the prediction error decreases when the effect of spatial and temporal dependence is taken into consideration when creating a prediction model. To tackle this problem, postal area codes were used to take the spatial dependence into consideration.

Steven et al. (2010), Liu (2013) and Dubin (1998) show that including neighbors houses improve the results of the predictions and make more accurate predictions. Therefore, this research will also include consideration of neighbors housing. As mentioned previously, realtors estimate market values partially based on houses sold on the same street. This is in line with the studies of Steven et al. (2010), Liu (2013) and Dubin (1998) which showed that neighboring houses are correlated with each other.

2.2 Housing characteristics

In the study of Park & Bae (2015), they developed several prediction models for housing prices. The data that was used had 5,359 houses with their characteristics from different sources. The houses are based in Fairfax County, in the state of Virginia in the United States of America. In this study, they created machine learning models such as C4.5, RIPPER, Naïve Bayesian and AdaBoost. Form their study, it is concluded that the RIPPER model outperforms the other models. Park & Bae (2015) even call this model superior compared to the other models. In their study they demonstrated how a machine learning algorithm may improve housing price forecasting and considerably contribute to the accurate assessment of real estate pricing.

Manasa, Narahari, & Gupta (2020) focused on regression models when predicting house prices. In their study they construct a predictive model for evaluating the price based on the factors that affect the price the house prices in Bengaluru, India. Their dataset included 9 variables such as the area type, the availablity, the price, number of bedrooms, bathrooms, kitchens and hallways, the society it belongs to, the size of the property and the location in the city. The dataset that was used contained 12,680 observations. The folliowing methods were used: OLS regression, Lasso regression, Ridge regression, Support Vector regression and XGBoost regression model. They concluded that “an optimal model does not necessarily represent a robust model. (Manasa et al. 2020)” This leads to the implication that a model that performs well when predicting housing prices does not nessecarily means that this model will be consistent in its prediction. The advanced regression models performed better compared to the basic linear regression. However, the advanced regression models all performed in a similar matter. Any of the advanced models is better in predicting houses and thus more useful compared to the basic linear regression (Manasa, Narahari, & Gupta, 2020). However, Imran, Zaman, Waqar, & Zaman (2021) also used several machine learning methods in their research of predicting house prices for Islamabad, Pakistan. They used a Linear regression, Support Vector regression, Ridge regression, Lasso regression, Gradient Boosting regression, Random Forest, Stochastic Gradient Descent regression and a Passive-aggressive regression to model the housing market of Islamabad. For this, they collected a dataset with 23 different variables and 44,647 observations. The variables include psychical features of the properties, such as the size, number of bed- and bathrooms, and if there is a swimming pool. Geograpical and environmental features were also included in the dataset, such as the if there is a school or hospital nearby. Their results showed that SVM performs the best across all the methods that they used. Compared to Manasa et al. (2020), Imran et al (2021) have found that SVM predicts the best from all the regression models, where Manasa et al. (2020) concluded that any advanced regression model is suitable for housing predictions.

Anh Le My (2016) compared a SVM with a Neural Network for predicting housing prices. In this study it was found that parameter tuning is essential when creating models since the research shows that the SVM outperforms the Neural Network, even though the Neural Network is gets at predicting the housing prices when the number of nodes increases.

Truong, Nguyen, Dang, & Mei (2022) also used various regression models to predict housing prices in Beijing, China. Their dataset had more than 300,000 entries with 26 variables. The data had housing prices traded between 2009 and 2018. The data that was used had variables such as the number of bedrooms, building structure, age of the house, and geographical characteristics. The research was conducted with the following models: XGBoost, LightGBM, Random Forest, Hybrid regression and Stacked Generalization regression. Concluded was that the models were all desirable for price predictions, but the Random Forest model was prone to overfitting (Truing et al. 2022). When comparing accuracy, the XGBoost and LightGBM. Due to its generality, the Hybrid regression approach is straightforward yet significantly more effective than the three other methods. The Stacked Generalization regression approach is the best option when accuracy is the main concern, despite its complex construction.

Quantile regression over the prices of residential properties confirms that most variables do not show a statistical significance under ordinary least squares or two-stage least squares (Ziets, Zietz, & Simans, 2008). Interesting insights from Quantile regression are from the explanatory variables for the selling price. Ziets et al. (2008) found that some variables have a greater impact as the selling price increases, such as square footage or lot size, where other variables stay relatively constant over the selling price across the different quartiles, such as having a garage or the distance to the city center.

In the study of Steven et al. (2010), several housing characteristics were included in their models. From this, we see that the land area and the floor area are significant at 1%. The coefficients of these variables in their OLS estimation are positive. This means that if the land and floor area increases, the price also increases. The same applies for the number of bathrooms, this variable is significant at the 1% level and has a positive coefficient. The age of a house is also significant at the 1% level, but the coefficient is negative. Thus, from these estimations it could be concluded that the older the property is, the negative the impact on the house price is. In this study, they also found that the variables about having a basement, having an AC unit, having a fireplace and the number of garages is all significant at the 1% level and have a positive effect on the housing price.

The study of Dubin (1998) also included housing characteristics. The study shows that having a patio, a fireplace, an AC unit, the number of basements, the land area, the floor area, and the number of garage spaces all are significant and have a positive coefficient. These variables have a positive effect on the housing prices according to the study. However, just like Steven et al. (2010), the study found that the age of the property has a negative coefficient and thus if the age of the property increases, the value of the property decreases. For the variable age, in their models a new variable age squared was also included. This variable had a positive effect on the price of a property ‘’because house prices are expected to decrease with age at a decreasing rate” (Dubin, 1998). The number of rooms and bathrooms also had a positive coefficient. The variable that is contained if the house is a detached house also has a positive coefficient, but the interaction variable between a detached house and the number of stories has a negative coefficient. Reasoning behind this is that people do not like to climb stairs, but three-story row houses are more appealing than two-story row houses. The coefficient of the variable that shows the interaction between the number of bathrooms and the number of bedrooms is positive because people tend to like to have more bathrooms when more bedrooms are included in that house.

Park et al. (2015) also had physical characteristics of the houses for modelling with Ripper and C4.5. Similar to Dubin (1998) and Steven et al. (2010), the variables for the size of the living area and the size of the property is important, as well as the garage space and the number of bedrooms and bathrooms. In their study, they also had data for the structure of the building (concrete or wooden walls for example) and heating resources (hot water from natural gas or electricity for example) that were important for their models. Their research, however, does not tell in what way these variables were important for the estimation.

The study of Ziets et al. (2008) concluded that positive significant variables for their models are the lot size, the size of the living area, the number and type of bathrooms and the type of floor. The regression coefficient of these variables increases as the selling price increases. Variables with a significant negative regression coefficient while the selling price increases are the age of the property and if the house had a mountain view. Coefficients that were relatively constant as the selling price increases were the type of exterior, the garage, having a sprinkler system and having an AC unit. The number of bedrooms did not show a pattern in the regression coefficient when the selling price increased. This contradicts the results of other papers that were mentioned previously.

Houses in Beijing, China shows a relationship between building type and the price (Truong et al., 2022). The research has shown that bungalows are correlated with a higher price compared to tower, tower and plate and plate buildings. Similar to Dubin (1998) does this study shows that the building type is an important variable in predicting housing prices. In the study of Manasa et al. (2022) it showed that the price was distributed normally over the number of balconies and the number of bathrooms. The research of Liu (2013) also concluded that the building type has a significant effect in the predicting models. The study of Imran et al. (2021) however, showed that there was not a high correlation between housing characteristics and the price. The highest correlation was found between the price and the number of bathrooms with a correlation value of 0.061.

Housing characteristics in the study of Zhang et al. (2021) were important in predicting prices, especially for the number of rooms and the age of the house. These variables were significant under the 1% level. The age of the house has a negative effect on the price of the property and the number of bedrooms has a positive effect on the price of the property. The condition of the house, however, did not have a significant effect on the regression model.

From the studies mentioned above, we can conclude that housing characteristics do have an impact on predicting housing prices, mostly with significant values for modelling house prices. We find that the size of the property and living area, as well as the number of bedrooms and bathrooms highly affects the price. Also, the building type seems important concluding from the studies mentioned above.

3 Data

For the data, three sources were used. The main dataset was found on Kaggle.com that includes houses for sale all over the Netherlands. The second dataset included the geographical references of the Netherlands. This was found on opendatasoft.com. This database is free to use. Also, the shapefiles of the Netherlands were found from the same website. Finally, the Google Maps API was used to combine the previously mentioned dataset together.

This main dataset that was used contains lots of usable information on houses for sale on the website funda.nl. This is an online website for houses that are for sale in the Netherlands. The dataset contains data from the first 8 months of 2022. The dataset has 5,555 observations with 16 variables. Compared to the statistics of funda.nl itself, the number of houses for sale for this period is around 80,000 (Funda Index, 2023) https://www.funda.nl/funda-index/juni-2023/. This dataset contains about 7% of the total number of houses that were available in the first 8 months of 2022.

The variables include the name of the street, the city where the house resides, the price of the house, the lot size given in square meters, the living space size given in square meters, the build year of the house, the build type of the house (existing or newly build), the type of house (bungalow, condo, etc.), the type of roof, the number of rooms, the number of bathrooms, the type of floor, the energy label, the geographical position of the house (for example: located in the city center or sea view) and the estimated price per square meter for the neighborhood.

The dataset containing the geographical references included the postal codes and their province and municipality. Geographical coordinates were also included in this dataset.

The Google Maps API was used to combine the datasets together. This API makes it possible to export information that can be found on Google Maps for multiple addresses. This API makes it easier to retrieve data from Google Maps. Keeping in mind that spatial dependence might have an impact on the prediction of the price, postal area codes could be insightful when predicting prices.

The final dataset will be split into a train- and test dataset for 75% and 25% respectively.

3.1 Combining the dataset

Making a complete dataset would include variables such as the postal area code to the dataset. The postal area codes would give an impression of a cluster of houses so that the prices will be in the same range in this postal area code. The dataset from Kaggle.com does not include postal area codes. For this, the Google Maps API was used to retrieve the address of its postal area code from the dataset its street and city. This gave us a new variable in the dataset, the postal area code in 4 digits. A typical Dutch postal area code consists of 4 digits and 2 letters and has the following format: 1234 AB. The letters were left out because of the specificity. Including the letters in the postal area code makes the group of houses smaller and the range of this postal area code is smaller (a few doors) compared with the postal area code without the letters (a few blocks). A limitation here is the number of observations in the dataset. Using a full postal area code gives less observations in one postal area code.

After adding the postal area codes to the main dataset, the geographical dataset of the Netherlands was used to add the province and municipality this house belongs to. The geographical dataset includes all the 4-digit postal area codes of the Netherlands along with the province and municipality it belongs to. The geographical dataset was used as a lookup table to add the province and the municipality to the main dataset based on the 4-digit postal area codes.

The shapefiles contain the shape of the 4-digit postal area codes as well as the geographical references for these areas. This data also contains the 4-digit postal area codes and the name of the province and municipality. This dataset was merged with the dataset that contains all the data for the houses on funda and postal area codes.

3.2 Data manipulation and cleaning

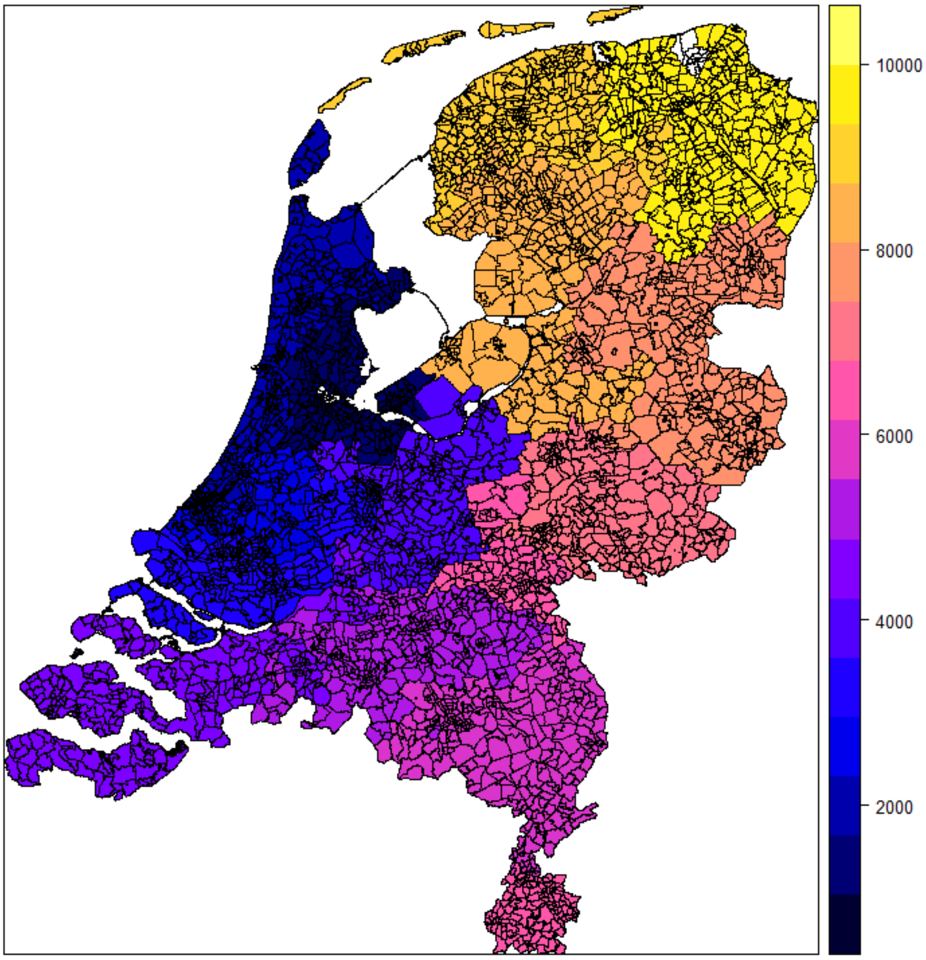
For the data to be useful in this research, multiple manipulations were made. The postal area group was split into groups for the first number. I.e., postal area code 3077 was put into group 3XXX and postal area code 2615 was put into group 2XXX. This was done due to having too many categorical variables when using the full 4-digit postal area code. For an interpretation of the postal area code, figure 3.1 shows a heat map of the postal area codes of the Netherlands.

The house type was split into 2 variables, because this variable has 2 specifications. First, the type of house, for example a family home, a bungalow or a mansion. The second variable of house type says something about the houses next to the house of interest, for example if the house is a semi-detached house or a detached house.

The variable roof was split into multiple different variables, consisting of the roof type and the type of cover that was on the roof. There were multiple possibilities for the type of cover on a roof, and this was into 8 binary variables.

The variable rooms was split into the number of bedrooms and the number of rooms. The variable toilet was split into the number of bathrooms and the number of separate toilets. The variable floors was split into the number of floors and 3 dummy variables that tells us whether the house has a basement, attic and/or a loft.

Figure 3.1 Postal area code plot (4-digits) of the Netherlands



The position was also separated into 16 dummy variables. For example, the position of a certain house was formulated as: “in the center of the city and next to a quiet road”. The dummy variables that came from this string of text were “in the center of the city” and “next to a quiet road”. The variable garden was split into 4 different categorical variables, because some houses in the dataset had multiple gardens around the house.

The final dataset has a total of 4,272 observations, with in total 59 variables. The variables that are included in this final dataset are as follows: the city, the price of the house, the lot size, the living space, the build year, the build type (newly build or existing), the energy label, the house type, type of roof, number of rooms, number of toilets/bathrooms, how many floors the property has, the energy label of the house, the position of the house, the type of garden, the postal area code in 4 digits, the estimated neighborhood price per square meters, the number of floors, if there is a basement, attic, and loft, the type of garden and the names of the province, municipality the price per square meter, the longitude, the latitude and the postal area group.

All missing values were omitted from the database. Only complete cases were used. The addresses that were not matched with a postal area code with the Google Maps API were removed from the dataset.

This research only includes prices of houses and no other properties such as offices, garages/parking spaces and storage boxes. Even though these properties are also sold on funda.nl, these properties are not in the scope of this research.

3.3 Descriptive variable statistics

Table 3.1 shows the description of the variables in the dataset.

Table 3.1: Descriptive variable statistics

VARIABLE	TYPE	DESCRIPTION
ADDRESS	Character	States the address of the house
CITY	Factor with 1041 levels	States the city/village the house is located at
PRICE	Numeric	States the listing price in euros
LOT.SIZE.M2.	Numeric	Lot size in square meters
LIVING.SPACE.SIZE..M2.	Numeric	Living space size in square meters
BUILD.YEAR	Integer	Year the property was built in
BUILD.TYPE	Factor with 2 levels	Type of built, new building or an existing building
ENERGY.LABEL	Factor with 12 levels	The grade of energy label the property is assigned to with A++++ being the best
ESTIMATED.NEIGHBOURHOOD.PRICE.PER.M2	Numeric	Estimated neighborhood price per square meter
PC	Factor with 1929 levels	4-digit postal area code
HOUSE.TYPE.1	Factor with 8 levels	First degree of the type of house
HOUSE.TYPE.2	Factor with 9 levels	Second degree of the type of house
HOUSE.TYPE.3	Factor with 10 levels	Third degree of the type of house
IN.WOONWIJK	Integer	Binary variable that states if the house resides in a residential area

AAN.BOSRAND	Integer	Binary variable that states if the house resides next to a forest
AAN.DRUKKE.WEG	Integer	Binary variable that states if the house resides next to a busy road
AAN.PARK	Integer	Binary variable that states if the house resides next to a park
AAN.RUSTIGE.WEG	Integer	Binary variable that states if the house resides next to a quiet road
AAN.VAARWATER	Integer	Binary variable that states if the house resides next to a waterway
AAN.WATER	Integer	Binary variable that states if the house resides next to water
BEDRIJVENTERREIN	Integer	Binary variable that states if the house resides on a business park
BESCHUTTE.LIGGING	Integer	Binary variable that states if the house resides in a sheltered area
BUITEN.BEBOUWDE.KOM	Integer	Binary variable that states if the house resides outside the build-up area
IN.BOSRIJKE.OMGEVING	Integer	Binary variable that states if the house resides in a forest
IN.CENTRUM	Integer	Binary variable that states if the house resides in the city center
LANDELIJK.GELEGEN	Integer	Binary variable that states if the house resides in the rural area
OPEN.LIGGING	Integer	Binary variable that states if the house has an open view
VRIJ.UITZICHT	Integer	Binary variable that states if the house has a clear view
ZEEZICHT	Integer	Binary variable that states if the house has a coastal view
NO.OF.ROOMS	Integer	Number of rooms
NO.OF.BEDROOMS	Integer	Number of bedrooms
NO.OF.BATHROOMS	Integer	Number of bathrooms
NO.OF.SEP.TOILETS	Integer	Number of separate toilet
ROOF.TYPE	Factor with 8 levels	Type of roof on the building
NO.OF.FLOORS	Integer	Number of floors

ZOLDER	Integer	Binary variable that states if the house has an attic
KELDER	Integer	Binary variable that states if the house has basement
VLIERING	Integer	Binary variable that states if the house has a loft
PROVINCIE.NAME	Factor with 13 levels	Name of the province the house resides in
GEMEENTE.NAME	Factor with 334 levels	Name of the municipality the house resides in
ACHTERTUIN	Integer	Binary variable that states if the house has a backyard
VOORTUIN	Integer	Binary variable that states if the house has a front yard
ZIJTUIN	Integer	Binary variable that states if the house has a side yard
PLAATS	Integer	Binary variable that states if the house has a courtyard
ZONNETERRAS	Integer	Binary variable that states if the house has a sun terrace
TUIN.RONDOM	Integer	Binary variable that states if the house has a garden all around
PATIO.ATRIUM	Integer	Binary variable that states if the house has a patio
ROOF.COVER.PANNEN	Integer	Binary variable that states if the roof is covered with roof tiles
ROOF.COVER.ASBEST	Integer	Binary variable that states if the roof is covered with asbestos
ROOF.COVER.BITUMINEUZE.DAKBEDEKKING	Integer	Binary variable that states if the roof is covered with bitumen
ROOF.COVER.KUNSTSTOF	Integer	Binary variable that states if the roof is covered with plastic
ROOF.COVER.LEISTEEN	Integer	Binary variable that states if the roof is covered with slate
ROOF.COVER.METAAL	Integer	Binary variable that states if the roof is covered with metal

ROOF.COVER.OVERIG	Integer	Binary variable that states if the roof is covered none of the mentioned roof covers
ROOF.COVER.RIET	Integer	Binary variable that states if the roof is covered with reed
PRICE.PER.M2.LOT.SIZE	Numeric	Price of the house divided by the variable LOT.SIZE.M2.
PRICE.PER.M2.LIVING.SPACE	Numeric	Price of the house divided by the variable LIVING.SPACE.SIZE..M2.
LON	Numeric	Longitude degree of the house
LAT	Numeric	Latitude degree of the house

3.4 Summary statistics

Table A.1 in appendix A shows all the variables that were used in the models for this research. Table A.2 to A.4 in appendix A gives the statistics of all the variables that are included for this research. In this part, the summary statistics will discuss most variables. We see that the minimum price of the houses in the dataset is equal to € 149,000, the median price is € 460,000, the average price is € 558,399, and the maximum price is € 4,700,000. Figure 3.2 shows the histogram of the price. From this, we see that there are many outliers in the dataset for the price.

We see that the mean lot size equals 256 m^2 while the mean living space size equals 146.1 m^2. The most frequent living space size in the dataset equals 110 m^2 that was found 83 times across the dataset. Figure 3.3 shows the histogram of the living size in square meters across the dataset.

Combining these variables into a scatterplot, we see a somewhat positive linear relationship between the price and the living space size in square meters. The scatterplot can be found in figure 3.4. The estimated neighborhood price is set at € 3,124 per m^2. The average year of a house is built is 1969 and most houses are graded with energy label C.

Location wise, around 74% of the houses are in a residential area, while roughly 2% of the houses are next to a forest. 3% is near a busy road but 6% is next to a quiet road, 4% is next to a park, 3% is next to a waterway and 7% is next to water. Around 12% are in a sheltered location or in the city center. 3% is outside the built-up area, 8% is in a forest area and 6% is in the rural area. 5% is in an open area and 2% have a clear view.

Figure 3.2 Histogram of the price

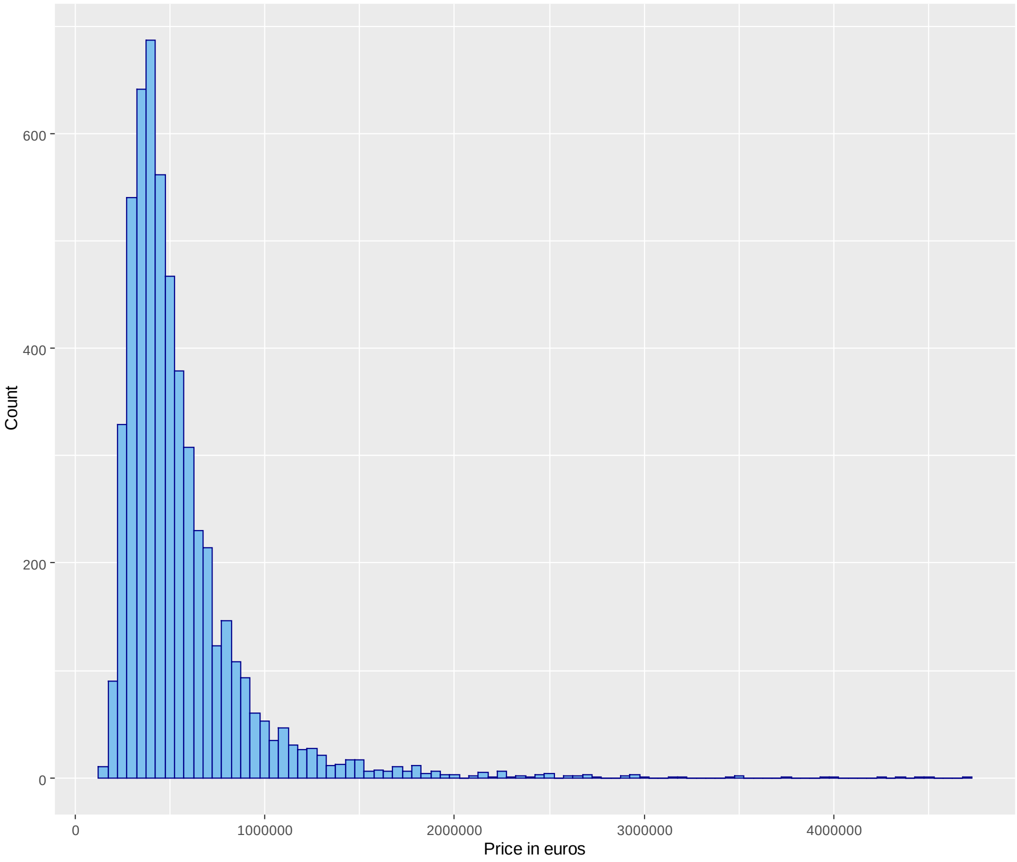


Figure 3.3 Histogram of the living space size in square meters across the dataset

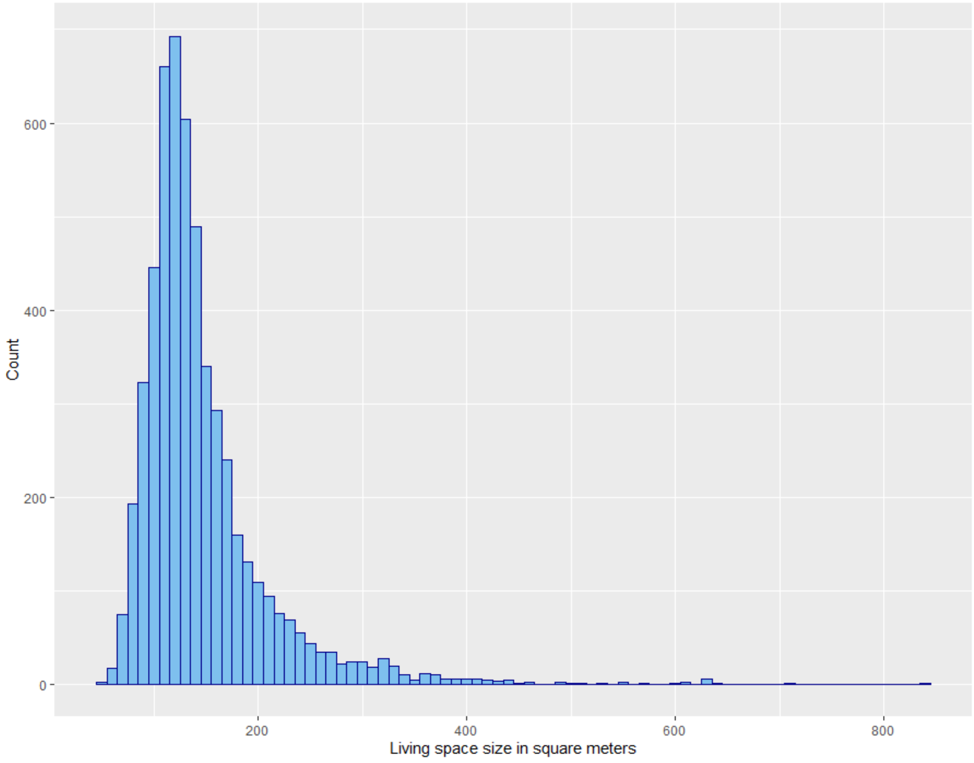
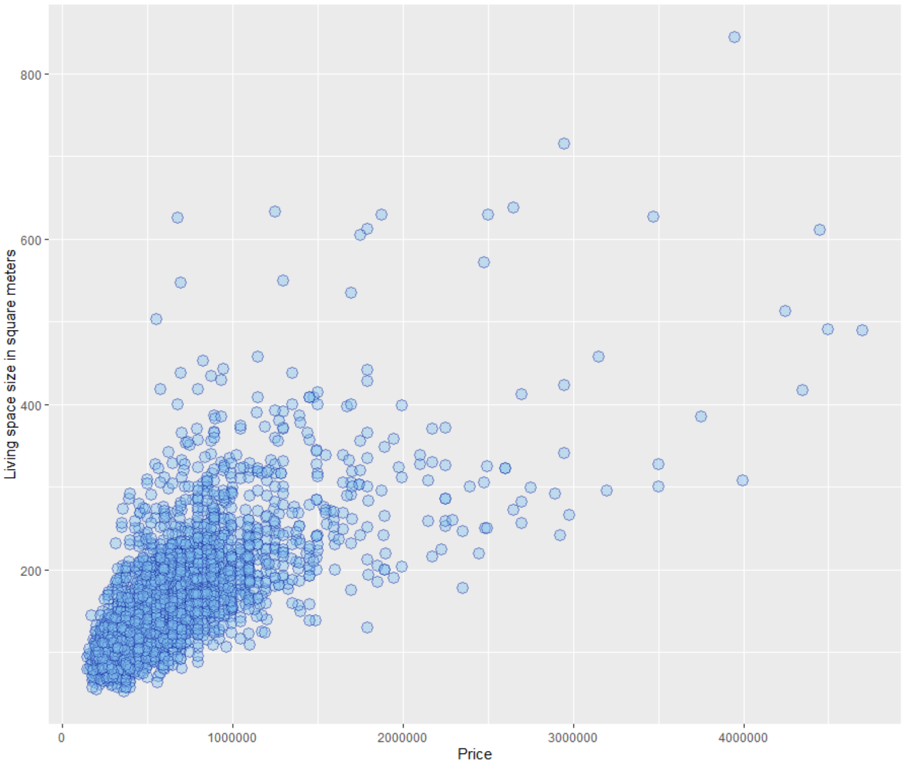


Figure 3.4 Scatterplot of the price and living space size in square meters



3.5 Correlation between variables

In this part, the correlation between the variables in the dataset will be discussed. Due to the size of the dataset, the correlation matrix was restricted in groups of variables. The most important correlation matrix can be found in figure 3.5. The bigger correlation matrices can be found in appendix B under figures B.1 to B.4.

First, the correlation matrix considering price, energy label, size, rooms and build type is shown in figure B.1 in appendix B. From this figure, it can be concluded that the price of a house is highly positive correlated with the living space in square meters. Thus, the bigger the living space size, the higher the price will be. Since the number of bathrooms is also positively correlated with the price, it can be concluded that the number of bathrooms increases the price of the houses. Build type is negatively correlated since the build type is either a new house or an existing house.

In B.2 in appendix B the correlation between the price and the house type variables are given. This matrix shows that the price is negatively correlated with single-family homes, meaning that the price decreases when the house is a single-family home, but the price is positively correlated with the house type being a detached house or a villa. Also, there is a negative correlation between the house being a single-family home and a mansion or a villa. Finally, a positive correlation is found between the house type bungalow and a semi-bungalow.

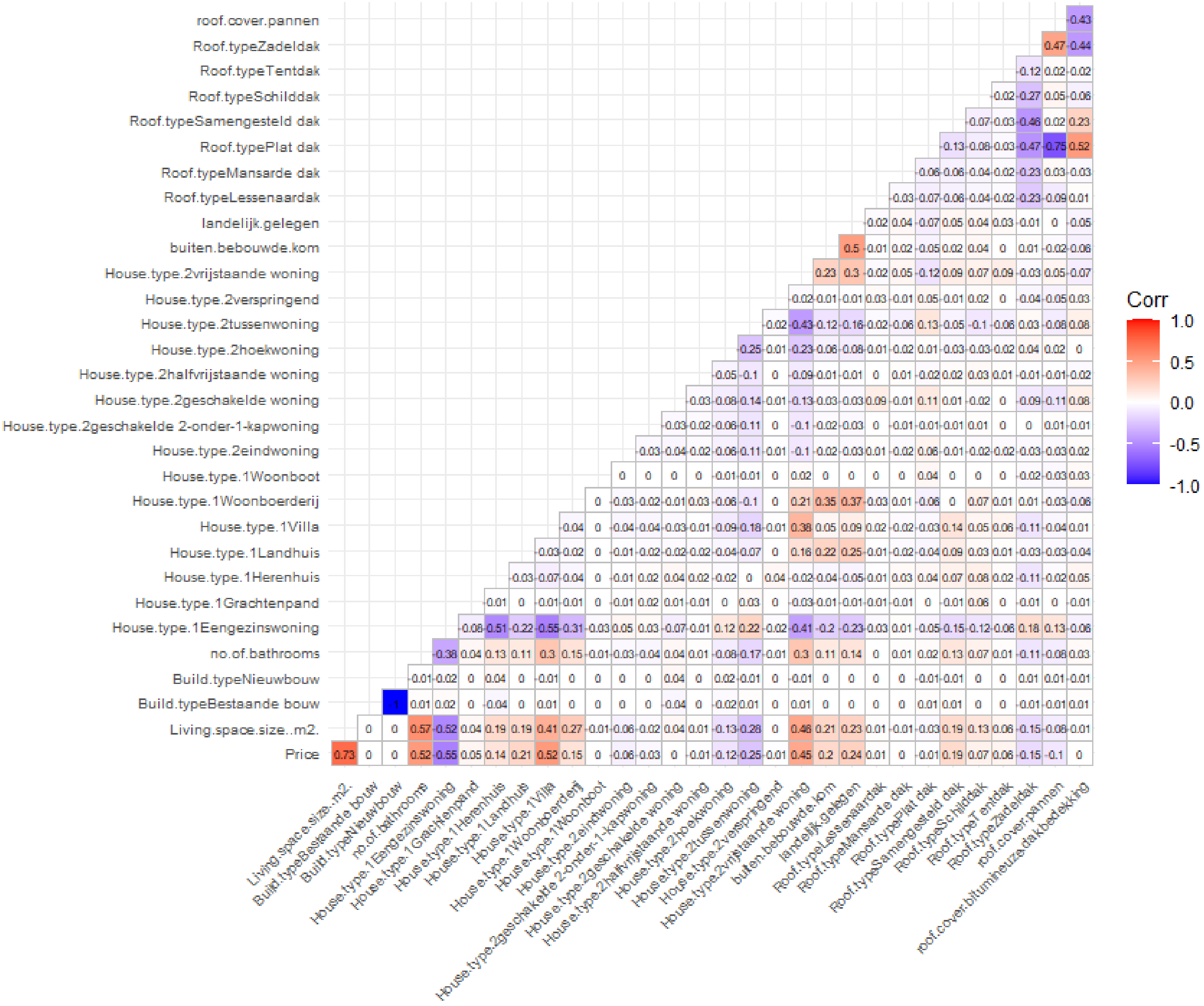
Figure B.3 in appendix B gives the correlation matrix of the house area variables. These variables tell where the property resides, for example in the city center, in a rural area, next to a busy road, et cetera.

There is no high correlation between the price and any of the other variables. However, there is a high correlation between the houses in a rural area and the houses outside the built-up area. This might be because houses in the rural area are mostly outside the built-up area.

Figure B.4 in appendix B shows more about the roof of the house. There is not a type of roof or roof cover that is highly correlated with the price of a house. However, there is a high negative correlation between a flat roof and roof tiles to cover a roof. In most cases, a flat roof does not use tiles to cover a roof. A flat roof is, however, more than often covered by bitumen. This can also be concluded from figure B.4 in appendix B.

Thus, from the correlation matrices in figure B.1 to B.4 in appendix B, prices are likely to be influenced by the living space, the number of bathrooms, single-family homes, detached house, and the house being a villa. Also, there is some multicollinearity due to correlation between variables.

Figure 3.5 Correlation matrix on price (summarized)



4 Methodology

In this chapter we will discuss what kind of machine learning methods will be used. The methods will be reviewed in this chapter and evaluated in the next chapter.

From the literature, we find that a simple linear regression is used as a baseline for the performance when comparing multiple methods. Therefore, we will first discuss the methodology of the simple linear regression. Then we will discuss the decision tree and the random forest and their variable importance. After that, the Support Vector Machine will be discussed. As mentioned in the literature, any other form of regression would be a better predictor compared to a linear regression. That is why multiple different methods will be used next to the linear regression.

The methods that will be discussed in this chapter will be compared in terms of predictive performance and will enlighten how these methods differ from each other. Also, ways of finding important characteristics of housing prices will be established in this chapter.

4.1 Multiple Linear Regression

A linear regression is a supervised machine learning method. This model estimates the relationship between the independent variables and one dependent variable by using a straight line (Molnar, 2023). The mathematical model can be written as follows:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon$$

Where Y is the dependent variable, β_i are the coefficients of regression, X is the independent variable and ϵ is the error term. In this research, we are interested in reducing the closeness from our predicted value with our model to the actual observation represented as ϵ in the formula above. Therefore, the least squares criterion is important when evaluating our results. The residual sum of squares can be written as follows:

$$\mathrm{RSS} = \sum_{\mathrm{i}=1}^{n} \epsilon_i^2$$

A Multiple Linear Regression includes more than one independent variable. In order to estimate a correct linear regression, assumptions of the relationships in the dataset must be met. These assumptions include linearity, normality, homoscedasticity, independence, fixed features, and absence of multicollinearity (Molnar, 2023). The linear regression for this research will have the following form:

$$Price_i = \beta_0 + \sum_{i=1}^{j} \beta_i X_i$$

Where:

$Price_i$ is the predicted price of house *i* and represents the weighted sum of all the features of the input variables*;*

β_0 is the coefficient of the intercept of the regression;

β_i is the coefficient of the variable X_i;

X_i is the value of the variable X_i.

All the variables included in the linear regression can be found in appendix A under table A.1.

The dataset includes numerical, categorical, and binary variables. Interpretation of these variables in a linear regression varies. For numerical variables, an increase of variable X_i by one unit increases the predicted outcome $Price_i$ with $\beta_i X_i$, ceteris paribus. In the case of a binary variable (where the binary variable is either set to 0 or 1), when the value of the variable X_i is set to 1, the predicted outcome $Price_i$ increases with β_i, ceteris paribus. Finally, a categorical variable includes L categories and in the case of a categorical variable, an increase of variable X_i from the reference category to the other category increases the predicted outcome $Price_i$ with $\beta_i X_i$, ceteris paribus (Molnar, 2023).

An advantage of the linear regression is that the model is easy to understand in how predictions are produced due to modeling the predictions as a weighted sum. Also, this model is highly used in many studies and places for predictive modeling and interference, and thus there is a high level of collective experience and expertise. Finding the optimal coefficients is likely from the mathematics behind a linear regression due to its relative simplicity of estimating the coefficients (Molnar, 2023).

However, linear regression models only represent a linear relationship, and a nonlinear relationship has to be set manually as an input feature. Linear models also do not have a great predictive performance because the model oversimplifies real life scenarios (Molnar, 2023).

Evaluating the Multiple Linear Regression will be done by comparing the RSME and R^2 with the other models in this research. This tells us how accurate the models are in predicting housing prices. Evaluating important characteristics of housing prices can be done by examining the linear regression model. Important to understand is what variables have a significant impact on the housing prices and in which way.

4.2 Support Vector Machine

A more advanced type of regression is the Support Vector Machine (hereafter: SVM). As mentioned in the literature section, this model might be useful in finding a good prediction model for housing prices. This study focuses on a regression problem and that is why a SVM as a regression model is used.

Compared to a linear regression, SVM allows to determine how much error is acceptable in our model and will find an appropriate hyperplane to fit the data. SVM wants to minimize the coefficients, where the error term is a part of the constraint. The absolute error is lower than or equal to ϵ. ϵ is the maximum error that is allowed for the SVM model. Since ϵ does not cover all observations, errors larger than ϵ should be taken into account. Therefore, the slack variable and hyperparameter ξ denotes the deviation from the margin for any value that falls outside of ϵ. This creates the following minimize function:

$$MIN\frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{n}|\xi_i|$$

The constraint is given as follows:

$$|y_i - w_i x_i| \leq \epsilon + |\xi_i|$$

Where y_i equals the value of the observation i and w_i is the coefficient of variable x_i. C is the variable and hyperparameter that is tuned for the tolerance towards the points outside of ϵ. An increase in C leads to an increase in tolerance for observations outside of ϵ. If $C = 0$, the model will be simplified and there would be no tolerance for observations outside of ϵ.

Finding the optimal value for C will be done by a grid search. After this, the mean absolute error will be compared to find the optimal C. Comparing the performance of the SVM model with other models will be done by comparing the RSME and the R^2.

4.3 Decision Tree

Taking into account that variables have a relationship, and the outcome of the prediction is not linear, the decision tree method is useful in this case. Therefore, in this study the decision tree will also be used. Decision trees can be used for both regression and classification. Decision tree models split the data several times due to feature cutoff values. These cutoff values split the dataset into different subsets, with each instance belonging to one subset. The final subset is a terminal or leaf node, while the intermediate subsets are called split or intermediate nodes. The outcome of each leaf node is the average outcome of the training data that was used in this part of the tree (Molnar, 2023). In this research, the CART algorithm is used for tree induction.

The following formula describes the relationship between the predicted outcome y and features x:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

Each prediction falls into one specific leaf node. The leaf node in this formula is denoted as R_m and thus the final subset. $\mathrm{I}_{\{x \in \mathrm{R_m}\}}$ is the identity function that return 1 if x is in the subset R_m and 0 if this is not the case. $\hat{y} = \mathrm{c_m}$ if an observation falls into leaf node R_m and c_m is the average value of all training observations in subset R_m.

The CART algorithm uses the Gini index to estimate variable importance and creates the subsets for the nodes. This algorithm takes a variable and establishes the cut-off point that minimizes the variance of y for a regression task. The variance is then minimized to a point where the training observations in the nodes have a similar value for the new prediction. From this it is evident that the best cut-off point is where two nodes are as different as possible with respect to the target outcome (Molnar, 2023). Then, the algorithm chooses the variable for splitting that will yield the best partition in terms of variance and adds it to the tree. The algorithm repeats a search-and-split process in both new nodes until a stop criterion is reached (Molnar, 2023).

Advantages of decision trees are that the tree structure captures interaction between variables. Interpretation is easier compared to a linear regression due to the data ending up in different groups versus points on a multidimensional hyperplane. The tree structure, with its nodes and edges, also provides a natural visualization. Finally, trees create good explanation of the data (Molnar, 2023).

Disadvantages of decision trees include the coping with linear relationships between variables due to the nature of splitting from a decision tree and creating a step function. There is also the lack of smoothness; small changes have a big impact on the outcome. Decision trees are also seen as unstable. Small changes in the dataset might lead to a different tree. Finally, interpretation is clear as long as the tree is short. The number of terminal nodes increases quickly with depth (Molnar, 2023).

The decision tree will also be evaluated by comparing the RSME and R^2. Evaluating the characteristics of that make housing prices high or low will be done by variable importance.

4.4 Random Forest

An extended version of the decision tree is the Random Forest model. A Random Forest is a supervised machine learning algorithm. A big advantage of this model is that both regression and classification problems can be solved. A Random Forest builds a number of decision trees on bootstrapped training samples. The individual decision trees have low bias but a high variance. Combining all the decision trees into a Random Forest decreases the variance. Thus, a Random Forest makes numerous decision trees and combines them together for a better and more accurate prediction because of this low bias and decreased variance. For splitting the trees, a Random Forest searches for the best variable amongst a random subset of variables. This results in a better model compared to traditional decision trees. That is why in a Random Forest only a random subset of the features is taken into account when splitting a node (James, Witten, Hastie, & Tibshirani, 2021).

Another advantage of the Random Forest model is that the algorithm itself has hyperparameters that produce good prediction results. Important parameters are the minimum numbers of nodes, the number of trees, the type of splitting rule, and the number of variables available for splitting at each node ('mtry').

When creating a Random Forest, the amount of decision trees and ‘mtry’ are both parameters that need to be optimized. Mtry is an integer that indicates how many variables are randomly sampled at each split. When using a higher mtry this will most likely lead to a model that becomes better at predicting the training data and a bit worse when predicting new testing data.

A Random Forest on the other hand has much higher predictive power compared to a regression model. That is also the biggest advantage of this method (James et al. 2021). This research focuses on creating the best prediction method for housing prices. Keeping this mind, it needs to be said that this method also has its drawbacks. It will take for example more time to compute these results and the results will also be less interpretable compared to using a multiple linear regression.

4.5 Variable importance – global interpretation method

In this research we also focus on important characteristics of a house for the asking. Therefore, to get a better understanding of the Support Vector Machine, Decision Tree model, and the Random Forest model, this study will also investigate the variable importance of these machine learning methods. Variable importance is a global interpretation method. With the variable importance, the significance of each variable in the dataset with respect to its effect on the generated model is checked. In a variable importance table, the predictors are ranked based on the contribution of the predictors that they make to the Decision Tree model and the Random Forest model. This creates a better understanding of these models. Advantages of this interpretation are that it contains a highly compressed global insight in the model's behavior, it does not require retraining of the model and it takes all interactions into consideration (Molnar, 2023).

The variable importance is established from an algorithm. The algorithm for the permutation variable importance has three steps. The input includes the trained model f̂, variable matrix X, target vector y and the error measure $L(y, \hat{f})$. First, the original model error $e_{orig} = L\left(y, \hat{f}(X)\right)$ is estimated. Then, for each feature $j \in \{1, \dots, p\}$, generate the feature matrix X_{perm} by permuting variable j in the dataset X, estimate error $e_{perm} = L\left(Y, \hat{f}(X_{perm})\right)$ based on the prediction of the permuted data and calculate the permutation feature importance as quotient $FI_j = e_{perm}/e_{orig}$ or by difference $FI_j = e_{perm} - e_{orig}$. Third, sort FI in descending order (Molnar, 2023). The first variable shows then the biggest importance, the second variable shows the second biggest importance and so on. A disadvantage of the permutation variable importance is multicollinearity. When there is a correlation between variables, permutation variable importance can perform poorly. Also, the scores are relative, it only shows the relative predictive power of the variables. At last, there is no statistical interference, because there is no insight in the nature of the relationship (Molnar, 2023).

4.6 Partial Dependence Plot – global interpretation method

Diving deeper into the variables for black box models, we will use partial dependence plots to show the relationship between the independent variable and the dependent variable. A partial dependence plot (in short: PDP) can show if this relationship is linear, complex or monotonic (Molnar, 2023). The partial dependence function for regression is written as follows:

$$\widehat{f_S}(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)dP(X_C)$$

Where the x_S are the variables for which the partial dependence function is plotted and X_C are the other variables that are used in $\hat{f}$ (Molnar, 2023). The partial function, which gives the average marginal effect of variable S on the independent variable is written as follows:

$$\widehat{f_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

In the formula above, $x_C^{(i)}$ represents the actual variable values from the dataset for the variables in which we are not interested, and n is the number of instances in the dataset (Molnar, 2023).

Advantages of PDP that the calculation of these plots is intuitive since a specific variable value represents the average prediction if we force all observations to assume that variable value. Also, when the variable in the PDP is not correlated with other variables, we have a clear interpretation of the variable its partial dependence. Furthermore, PDPs are ease to implement and the calculation for these plots has a causal interpretation (Molnar, 2023).

Disadvantages are the number of features in a PDP, since a PDP is a 3D representation. Also, the assumption of independence is an issue with PDP as well as heterogenous effects because these might be hidden (Molnar, 2023).

4.7 SLX model

In this research, we are also interested in the local effect on the asking price of a house. For this, we need to investigate the spatial spillover effects of houses. Therefore, the spatial lag of X model (also knows as the SLX model) is used in this research. Spatial spillovers are the impact of changes to independent variables in a particular unit on the dependent variable values in other units $j \neq i$ (Vega & Elhorst, 2015). Within the data, there is a possibility of $N(N-1)$ relations but only N data observations are available. A spatial weight matrix (mathematically written as W) reduces the number of parameters to be estimated from the total number of possible relationships to a number that corresponds with the SLX model (Vega & Elhorst, 2015). The SLX model is written as follows:

$$\mathrm{Y} = \alpha\iota_N \ + \ \beta X + WX\theta + \ \epsilon$$

Where:

Y represents a vector of one observation on the dependent variable for every unit in the dataset;

X denotes a vector $(N \ x \, K)$ that contains the value of the independent variable that is associated with the parameter vector $(K \ x \ 1)\beta$;

ϵ is a vector of independent and identical distributed disturbance terms with zero mean and variance;

W represents the weight matrix;

WX represents the exogenous spatial lags among the independent variables;

θ estimates the spillover effect of the variables, and;

β estimates the direct effect of variable (Vega & Elhorst, 2015) .

The advantage of the SLX model over other spatial econometric models is this model can be used to see whether asking prices are endogenous. Also, the specific functional form of an SLX model is not assumed when building this model. This model also provides more flexibility if it is not clear how sensitive interaction is to distance (Vega & Elhorst, 2015).

5 Results

This chapter is divided into the prediction models and the spatial dependence model. First, the predictive models will be discussed as well as the results these models reported. As mentioned before, the models will be evaluated on their RMSE and the R^2. Second, the spatial dependence model will be discussed where we will see the impact of neighboring houses on the house prices.

Note that the test dataset shrunk due to models that were not able to predict on variables that were included in the training dataset and not in the test dataset. Therefore, only complete cases were used. Outliers for the price were deleted. This was done to create models that are more accurate in terms of RMSE. Including the most expensive models in the houses, would generate a higher RMSE. The top 5 percent of the houses with the highest price were removed from the dataset. This resulted in a price-cutoff at a price of € 1,146,000. This means that the models constructed below are not able to predict houses with an extremely high price.

Imputation was not used due to creating a biased model and possibly violating the independence assumption. Also, imputed data does not have an error term included. Since this research tries to estimate a model that can predict house prices and the evaluation criteria is based on the variance and the error term, imputation would lead to models that would be biased.

For the predicted models, for each type of model, 2 models were constructed. The dependent variable for the models were "Price" and "Price per square meters". The variables that were used in the models can be found in appendix A under table A.1.

5.1 Multiple Linear Regression Model

The first 2 models that were constructed were a multiple linear regression. The full models can be found in appendix C, under table C.1. In table 5.1 and 5.2 we see the significant coefficients of the variables for respectively the model with the dependent variable being "Price" and "Price per square meter" with the standardized beta coefficients. These are the significant coefficients from table C.1 in appendix C.

5.1.1 Price model

Since there are 42 significant variables in the price model, we will discuss the variables with the top 3 standardized beta coefficients. This can also be seen in table 5.1. First, we will investigate the model with dependent variable being "Price". We see that the living space size in square meters has the highest standardized beta coefficient. This indicates that this variable has the greatest effect on the dependent variable "Price". If the living space size increases with 1 square meter, this will result in an increase in the price of the house with € 1,605, ceteris paribus. House type 2 being an in-between house instead of a 2-under-1 roof house (the reference category) decreases the price of the house with € 81,501, ceteris paribus. At last, if the longitude degree of a house increases with 1 degree, meaning that if the house resides more to the eastern part of the Netherlands, then this results into a price decrease of € 40,310, ceteris paribus. Other variables that have the most effect on the dependent variable are a garden around the house, house type 1 being a single family home instead of a bungalow (the reference category), postal area code that starts with a 2 or 7 instead of a 1 (the reference category), house type 2 being a corner house or a detached house compared to a 2-under-1 roof house (the reference category), and the house having energy label C instead of A (the reference category).

The RMSE for this model on the train set is 99,440 and for the test set is 100,667. The R^2 of this model on the training set is 0.7291 and the R^2 on the test set is 0.7217. This indicates that the multiple linear model explains 73% of the total variance in the train set and 72% of the variance in the test set.

Table 5.1 The significant coefficients of the linear regression model with price as dependent variable (sorted by the standard coefficient in absolute value)

Variable	Coefficient		Standardized coefficient	Standard Error
Living.space.size..m2.	1,605	***	0.3799	70
House.type.2tussenwoning	-81,501	***	-0.2021	6,233
lon	-40,310	***	-0.1614	9,175
Tuin.rondom	82,660	***	0.1433	10,760
House.type.1Eengezinswoning	-75,409	***	-0.1374	16,908
PCGROUP2XXX	74,695	***	0.1261	11,762
House.type.2hoekwoning	-62,394	***	-0.1107	6,985
PCGROUP7XXX	-59,883	***	-0.1022	17,502
Energy.labelC	-43,187	***	-0.1019	5,879
House.type.2vrijstaande woning	43,995	***	0.0970	7,766
no.of.bathrooms	37,641	***	0.0811	5,345
Energy.labelG	-74,985	***	-0.0773	11,589
PCGROUP9XXX	-63,634	**	-0.0673	22,090
Energy.labelF	-57,513	***	-0.0662	10,508
Energy.labelD	-38,223	***	-0.0630	7,840
PCGROUP6XXX	-40,293	*	-0.0609	17,926
House.type.2eindwoning	-64,880	***	-0.0599	11,628
no.of.sep.toilets	27,397	***	0.0536	5,307
Energy.labelB	-26,410	***	-0.0531	6,022
no.of.floors	16,221	**	0.0486	5,027
Zijtuin	20,595	***	0.0437	5,541
aan.vaarwater	54,878	***	0.0433	13,555
no.of.rooms	6,225	*	0.0412	2,719
Build.year	-207	**	-0.0397	78
Energy.labelE	-29,419	**	-0.0390	9,137
roof.cover.riet	82,126	**	0.0388	25,614
House.type.1Villa	37,780	*	0.0385	19,255
House.type.1Woonboerderij	-58,709	*	-0.0367	26,259

no.of.bedrooms	-6,864	*	-0.0351	3,235
roof.cover.metaal	-112,912	***	-0.0338	34,122
Estimated.neighbourhood.price.per.m2	2	**	0.0338	1
Lot.size..m2.	33	**	0.0326	13
in.bosrijke.omgeving	25,376	**	0.0324	8,320
Zonneterras	28,382	**	0.0306	9,436
Energy.labelA+++	111,509	**	0.0279	41,206
buiten.bebouwde.kom	33,757	*	0.0251	16,214
vrij.uitzicht	12,280	*	0.0247	5,347
kelder	20,214	*	0.0243	8,988
in.centrum	-14,434	*	-0.0232	6,636
House.type.2gesch. 2-onder-1-kapwoning	-28,816	*	-0.0226	13,121
open.ligging	-21,935	*	-0.0223	10,229
aan.bosrand	28,505	*	0.0211	13,947

Note: *p<0.1; **p<0.05; ***p<0.01

5.1.2 Price per square meter model

The "Price per square meter model" predict the price per square meters with the same independent variables as the "Price" model. Since this model has 38 significant variables (this can also be found in table 5.2), we only discuss the variables with the top 3 standardized beta coefficients.

Again, we see that the living space size has the most effect on our dependent variable "Price per square meter". If the living space size of the house increases with 1 square meter, the price per square meter decreases with € 9, ceteris paribus. If house type 2 is an in-between house instead of a 2-under-1-roof house (the reference category), we see that this decreases the price per square meter with € 559, ceteris paribus. We also see that the longitude degree has a big effect on the price per square meter according to the standardized coefficient. If the longitude degree increases with 1, meaning that the houses that reside in the more eastern part of the Netherlands, the price per square meter decreases with € 263, ceteris paribus. Other variables that have a big effect on the dependent variable are house type 1 being a single family home instead of a bungalow (the reference category), the postal area code that starts with a 2 or 7 instead of a 1 (the reference category), a garden around the house, house type 2 being a corner house or a detached house instead of a 2-under-1-roof house (the reference category) and energy label C instead of A (the reference category).

The RMSE for the price per square meter model on the train set is 689 and for the test set is 708. The R^2 of this model on the training set is 0.5238 and the R^2 on the test set is 0.4931. This indicates that the multiple linear model explains 52% of the total variance in the train set and 49% of the variance in the test set.

Table 5.2 The significant coefficients of the linear regression model with price per square meter as dependent variable (sorted by the standard coefficient in absolute value)

Variable	Coefficient		Standardized coefficient	Standard Error
Living.space.size..m2.	-9.41	***	-0.4261	0.48
House.type.2tussenwoning	-558.69	***	-0.2651	43.18
lon	-262.50	***	-0.2011	63.56
House.type.1Eengezinswoning	-484.81	***	-0.1691	117.13
PCGROUP7XXX	-502.84	***	-0.1642	121.24
PCGROUP2XXX	494.77	***	0.1598	81.48
Tuin.rondom	472.46	***	0.1567	74.54
House.type.2hoekwoning	-423.50	***	-0.1437	48.39
House.type.2vrijstaande woning	306.25	***	0.1292	53.80
Energy.labelC	-278.72	***	-0.1258	40.73
no.of.bathrooms	290.41	***	0.1197	37.03
PCGROUP6XXX	-382.52	**	-0.1107	124.18
PCGROUP9XXX	-487.47	**	-0.0987	153.03
PCGROUP4XXX	-285.89	*	-0.0868	128.19
House.type.2eindwoning	-474.37	***	-0.0839	80.55
Energy.labelD	-250.48	***	-0.0790	54.31
Build.year	-2.14	***	-0.0784	0.54
Energy.labelF	-348.94	***	-0.0768	72.79
Energy.labelB	-195.07	***	-0.0751	41.72
Energy.labelG	-378.63	***	-0.0747	80.28
Zijtuin	170.49	***	0.0693	38.38
PCGROUP8XXX	-266.13	*	-0.0665	105.91
no.of.bedrooms	-64.90	**	-0.0636	22.41
no.of.sep.toilets	156.91	***	0.0588	36.76
in.bosrijke.omgeving	237.54	***	0.0581	57.63
Estimated.neighbourhood.price.per.m2	0.02	***	0.0556	0.00
landelijk.gelegen	244.75	**	0.0499	81.06
aan.vaarwater	318.10	***	0.0481	93.90
House.type.1Woonboerderij	-378.69	*	-0.0453	181.91
Zonneterras	201.72	**	0.0416	65.36
Energy.labelE	-159.77	*	-0.0405	63.29
buiten.bebouwde.kom	269.90	*	0.0385	112.32
vrij.uitzicht	97.98	**	0.0378	37.04
roof.cover.riet	397.48	*	0.0359	177.43
open.ligging	-170.96	*	-0.0333	70.86

Lot.size..m2.	0.17	*	0.0328	0.09
Energy.labelA+++	656.01	*	0.0315	285.45
House.type.3drive-in woning	-290.43	*	-0.0268	144.17

Note: *p<0.1; **p<0.05; ***p<0.01

Comparing tables 5.1 and 5.2, we see some resemblance in the significant variables for both models. For instance, in both models we see a positive effect for when the house has an energy label of A+++ instead of A (the reference category). If the postal area code of the house starts with a 2 instead of 1 (the reference category), this also has a positive effect for both models. If there is either a garden around the house, a sun terrace, or a side garden, this also influences the dependent variable positively in both models. Other instances when variables have a positive effect on the dependent variable is when the roof cover is made from reed, when the house is next to a waterway, in a forestry area, outside the build up area or has an open view, when house type 2 is a detached house instead of a 2-under-1-roof house (the reference category), if the number of bathrooms or separate toilets increases or when the house has a sun terrace. Finally, both the lot size in square meters and the estimated neighborhood price per square meter increases, the dependent variable in both models also increases.

We see a negative effect on the dependent variable in both models if the build year increases by 1. Also, energy labels B to G instead of A (the reference category) all have a negative effect on both dependent variables in both models. The same applies if house type 1 is a single-family house or a farmhouse instead of a bungalow (the reference category) or when house type 2 is an end house, a corner house, an in-between house, or a detached house instead of a 2-under-1-roof house (the reference category). An increase in the longitude degree also has a negative impact on both dependent variables, as well as an increase in the number of bedrooms or when a house has an open view. Finally, we see that for both models the dependent variable decreases when the postal area code of the house starts with a 6, 7 or 9 instead of 1 (the reference category).

An increase in the living space size in square meters has a positive effect on the dependent variable “Price per square meter” but it has a negative effect on the dependent variable “Price”.

5.1.3 Model diagnostics

Figures 5.1 and 5.2 show the regression diagnostics plots for respectively the model with “Price” as independent variable and “Price per square meter” as independent variable. In the residuals vs fitted part (top left), we can indicate if there is a linear relationship between the residuals and the fitted values when there is a horizontal line with no distinct pattern. For both models, we can conclude that there is no distinct pattern, which indicates that there is a linear relationship between the independent and dependent variables.

In the Scale-Location part (bottom left) we see the homogeneity of the variance plot. This plot tells us if the residuals are evenly spread across the ranges of independent variables. From this plot we can conclude that the variances of the residuals increase with the value of the fitted values. This suggests non-constant variances in the residuals errors. Put simply, we can speak of heteroscedasticity for both models.

The normality assumption can be checked by the top right part of the figures 5.1 and 5.2, with the Normal Q-Q plot. In both cases, we see that the normality assumption is violated, since a normal distribution in the residuals would be in a straight line in both these plots. This is not the case for both models.

Looking at outliers, we see that there are some outliers in the bottom right plot of the figures 5.1 and 5.2. An outlier is considered when the standardized residuals are larger than 3 (James et al. 2021). Therefore, it can be concluded that there are many outliers. Also, a data point has a high leverage is it has extreme values for the independent variables. This is the case for a a few datapoints for both models. Omitting the top 5 percent of houses with the highest price did not remove all outliers.

Figure 5.1 Multiple linear regression diagnostics with dependent variable "Price"

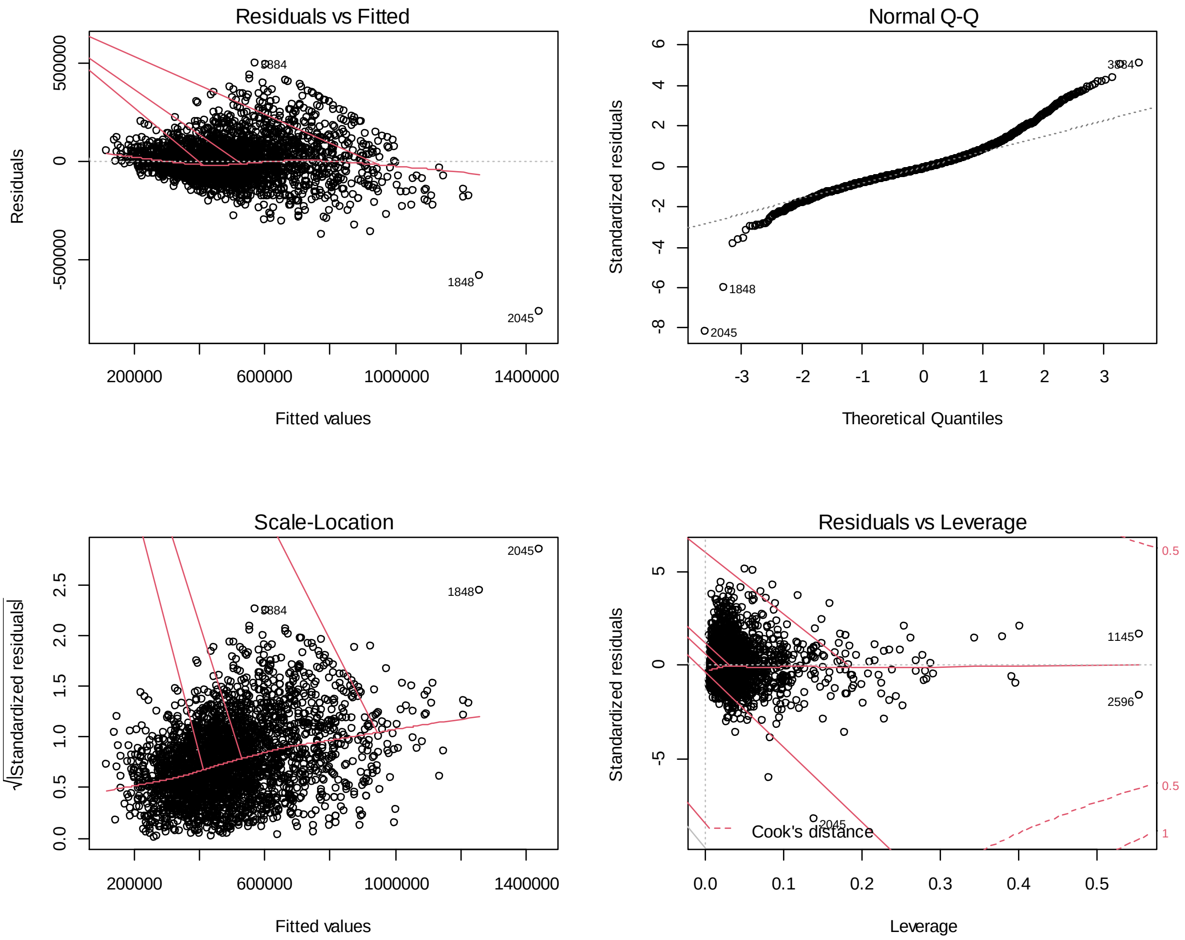
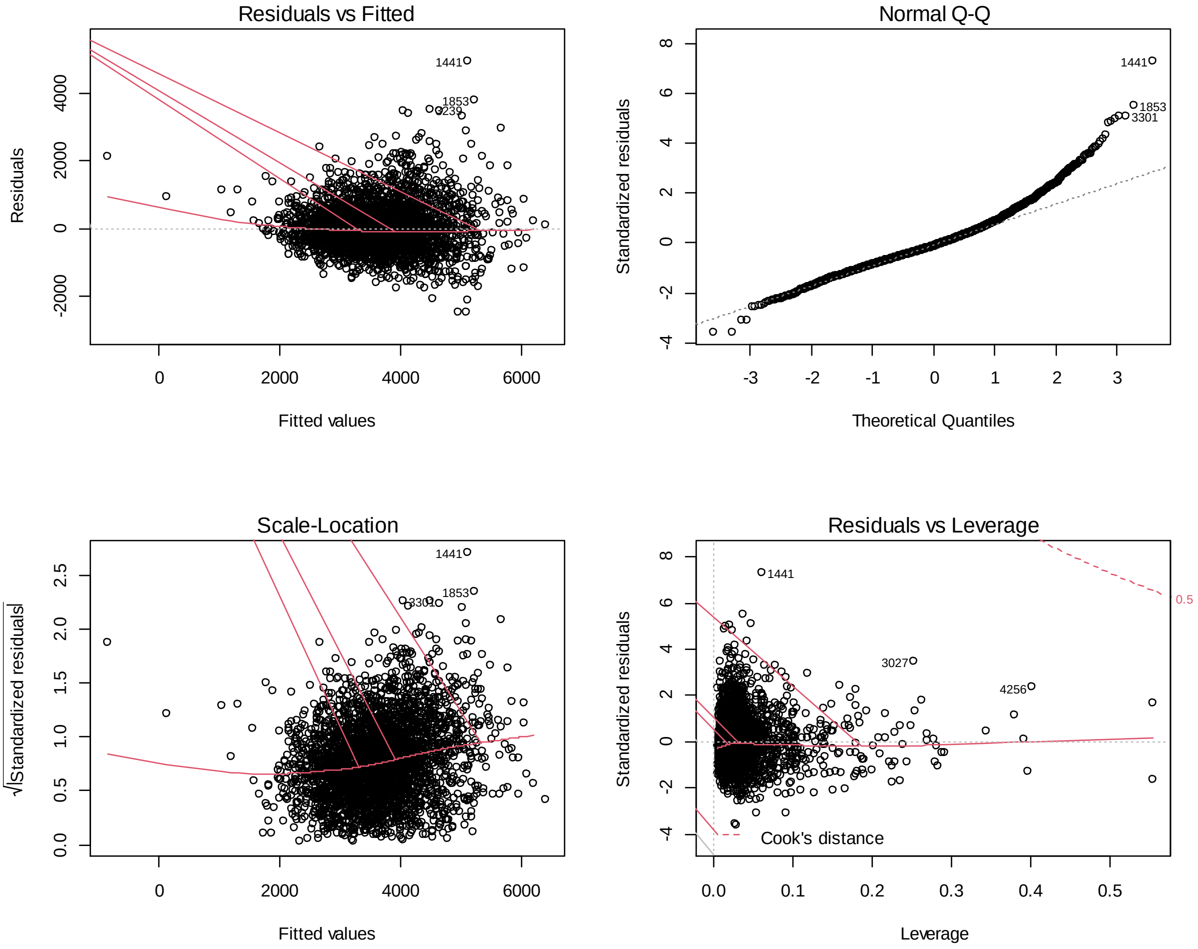


Figure 5.2 Multiple linear regression diagnostics with dependent variable "Price per square meter"



To solve the issues with the assumptions, it would be wise to balance the data. However, in the real world, house prices are not equally spread. The higher the price, the lower the amount of houses that are for sale. The tradeoff with balanced data then would be that the models would be biased. Using the logarithmic value of the dependent variable did not yield a better Normal Q-Q plot (top right part of figure 5.1 and 5.1). After transforming the dependent variables into logarithmic values, the normality assumption was still violated. Transformation does not solve the normality problem in this case.

5.2 Support Vector Model

For the SVM models, a grid search was performed. The variables used for both models can be found in appendix A under table A.1. 2 models were constructed, with the variable “Price” and “Price per square meter” as dependent variable. Then, the best value for C was found with the grid search. As mentioned before, C is the variable and hyperparameter that is tuned for the tolerance towards the points outside of ϵ. Also, the linear kernel was used for these models and the models were cross validated 10 times.

5.2.1 Price model

For the model with “Price” as dependent variable, we found that a C value of 3.5897 resulted in the lowest RMSE and highest R^2. This can also be found in table D.1 in appendix D. This model has a RMSE of 110,591 on the train set and a RMSE 100,001 on the test set. The R^2 of this model on the train set is 0.6676 and on the test set is 0.7276. This indicates that the SVM model explains 67% of the total variance in the train set and 73% of the variance in the test set.

Investigating the variable importance plot of figure 5.3, we see that the most important variable for the SVM price model is the living space per square meter is, followed by the categorical variable house type 1 and the lot size in square meters. Other important variables are the categorical variable house type 2, the number of bathrooms, the number of rooms, a garden around the house, the build year, the number of bedrooms, and a backyard.

5.2.2 Price per square meter model

For the model with “Price per square meter” as dependent variable, we found that a C value of 4 resulted in the lowest RMSE. This can also be found in table D.2 in appendix D. This model has a RMSE of 731 on the train set and a RMSE 707 on the test set. The R^2 of this model on the train set is 0.4749 and on the test set is 0,5002. This indicates that the SVM model explains 47% of the total variance in the train set and about 50% in the test set.

Investigating the variable importance plot of figure 5.4, we see that the most important variable for the SVM price per square meter model the categorical variable province name, followed by the categorical variable postal area code group and the longitude degree. Other important variables are the latitude degree, the build year, the living space per square meter, the categorical variables house type 1 and 2, the lot size in square meter and the number of rooms.

Combining figures 5.3 and 5.4, we see that both models have some resemblance considering the importance of the variables in the models. We see that living space and lot size in square meter, the build year, categorical variables house type 1 and 2 and the number of rooms.

Figure 5.3 Variable importance plot of the support vector machine with dependent variable "Price"

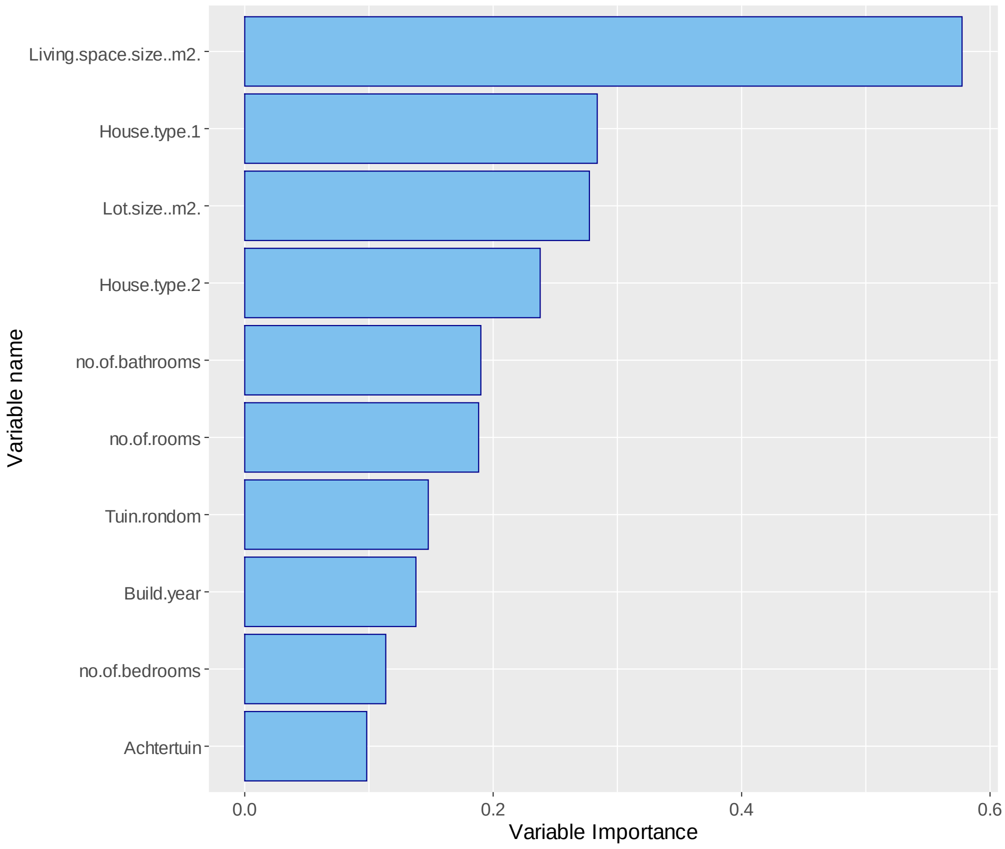
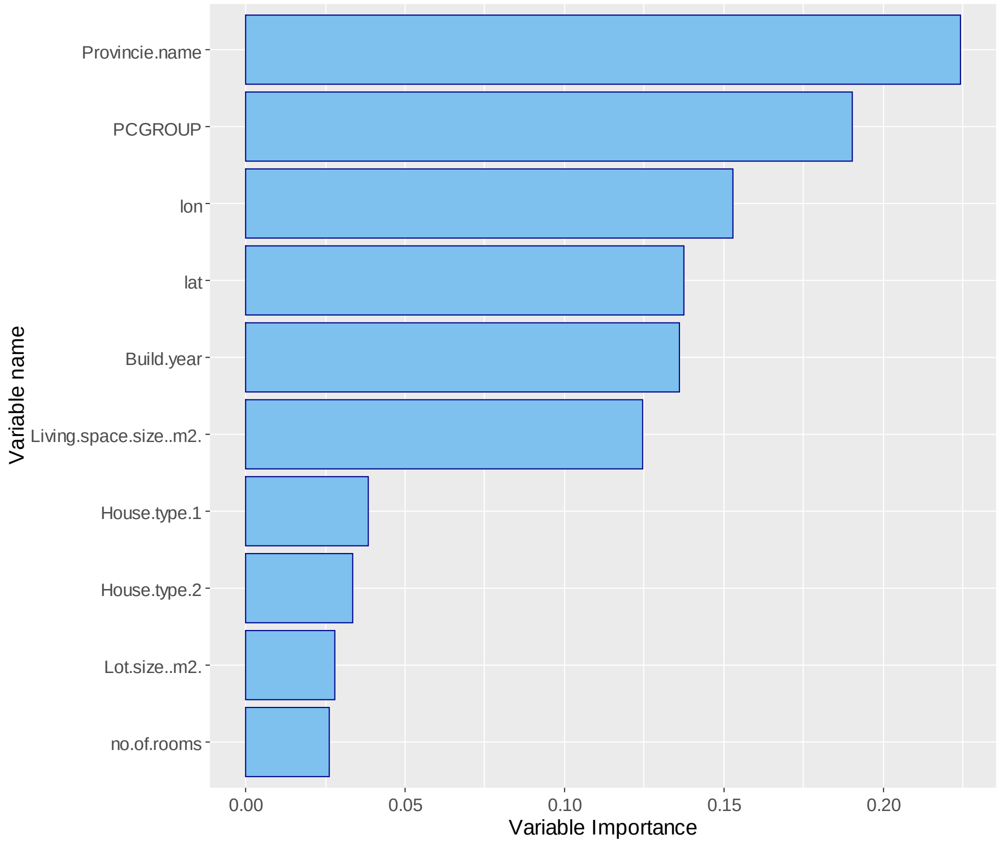


Figure 5.4 Variable importance plot of the support vector machine with dependent variable "Price per square meter"



5.3 Decision Tree

Also, for the decision tree models, a grid search was performed. This grid search was performed across different values for the complexity parameter (cp) and 10 times cross validated.

5.3.1 Price model

The grid search table for the model with dependent variable "Price" can be found in table D.3 in appendix D. Then, the model with the lowest RMSE was selected. The final model with dependent variable “Price” has a cp value of 0.008830 with the lowest RMSE of 124,960 and a R^2 of 0.5734.

The decision tree model that was used with dependent variable “Price” has a RMSE of 125.531 and a R^2 of 0.5676 on the test set, and therefore, this model explains 59% of the variance in the test set.

Taking a look at the decision tree for the model with “Price” as dependent variable in figure 5.7, we see that this decision tree is primarily based on the living space size in square meters (in nodes 1, 2 and 13), house type 1 being a single-family home (in node 2) and the longitude degree (in nodes 4, 5 and 6). The latitude degree, if there is a garden around the house and the lot size in square meters are also shown in nodes 9, 10 and 11. The endnotes show us the final predicted values of the house prices with the endnote of a house price of € 431.095 as the value that returns the most in the training data.

According to the variable importance plot, given in figure 5.5, the most important variable for the decision tree with “Price” as dependent variable is the living space size in square meter. Other important variables for this model are house type 2 being a detached house, house type 1 being a single-family house, a garden around the house, the longitude degree, the number of bathrooms, the lot size in square meters, the latitude degree, the province in which the house resides is Noord-Holland and having a backyard.

5.3.2 Price per square meter model

The grid search table for the model with dependent variable "Price per square meter" can be found in table D.4 in appendix D. The final model with dependent variable “Price per square meter” has a cp value of 0.0150527 with the lowest RMSE of 836 and a R^2 of 0.3021. These values for the RMSE and R^2 all correspond to the training set.

The decision tree model that was used with dependent variable “Price per square meter” has a RMSE of 846 and a R^2 of 0.2796 on the test set. In this case, the model explains 28% of the variance in the test set.

Considering the decision tree for the model with “Price per square meter” as dependent variable in figure 5.8, we see that this decision tree is primarily based on the longitude degree (in node 1), the latitude degree (in nodes 3, 4 and 15) and a garden around the house (in node 2). Also important are the living space in square meters (in node 5), house type 2 being a detached house (in node 6), the build year (in node 7) and the lot size in square meters (in node 14). The endnotes show us the final predicted values of the price per square meter with the endnote with a price of € 3.278 per square meter as the value that returns the most in the training data.

From the variable importance plot in figure 5.6, we see that the most important variable for the decision tree with “Price per square meter” as dependent variable is the lot size in square meters, followed by the latitude degree and the longitude degree. Also important are house type 2 being a detached house or an in-between house, the build year, house type 1 being a single-family home, the province in which the house resides is Noord-Holland or Gelderland and a garden around the house.

Comparing figures 5.5. and 5.6, we see that for both models the variables house type 1 being a single-family house, a garden around the house, the longitude and latitude degree, the lot size in square meters, and province in which the house resides is Noord-Holland all are important in both models.

Figure 5.5 Variable importance plot of the decision tree with dependent variable "Price"

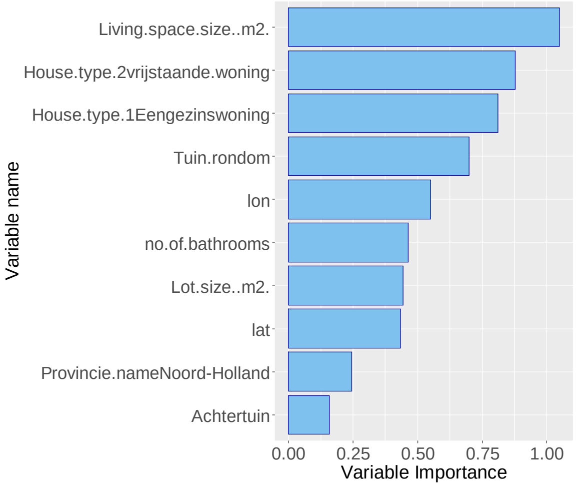


Figure 5.6 Variable importance plot of the decision tree with dependent variable "Price per square meter"

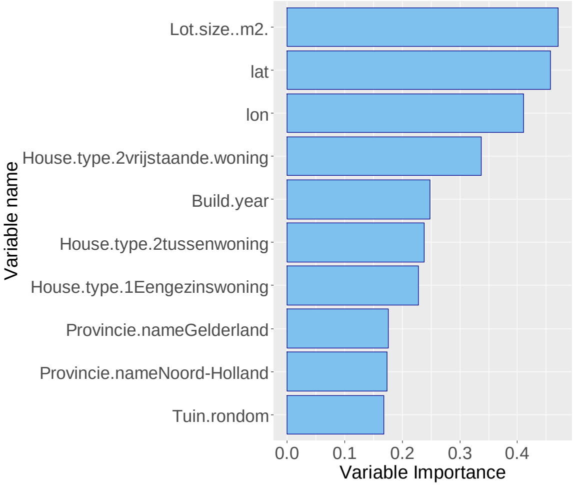


Figure 5.7 Decision tree plot with dependent variable "Price"

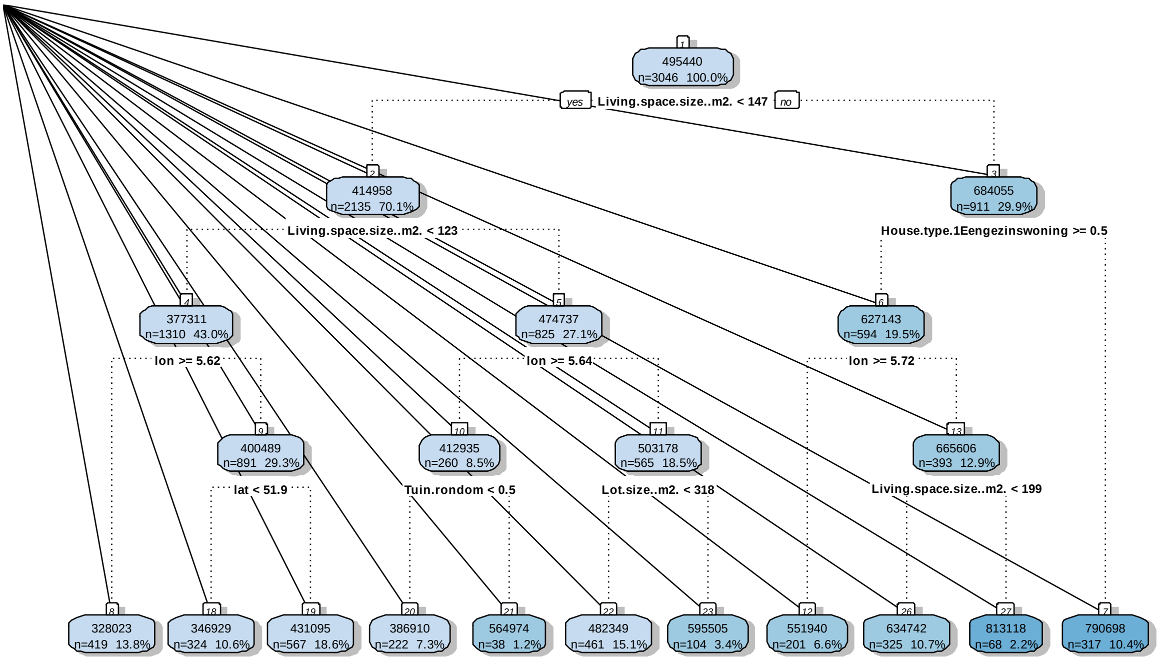
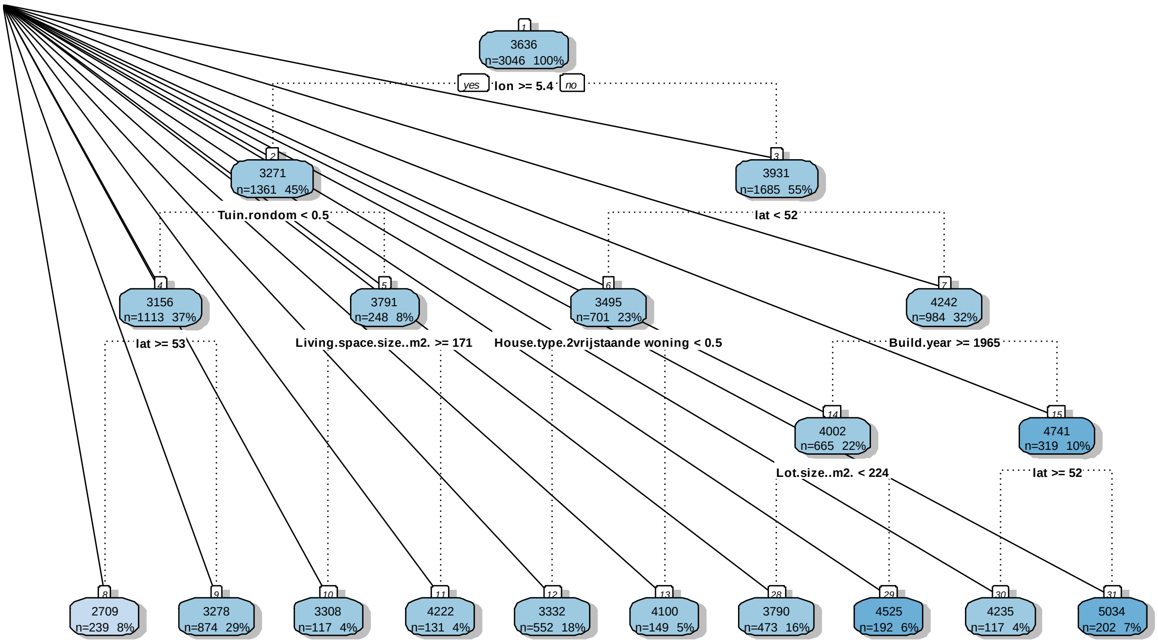


Figure 5.8 Decision tree plot with dependent variable "Price per square meter"



5.4 Random Forest Model

For the Random Forest model, another grid search was performed. The grid search included a mtry of 2, 54 and 106, the splitrules of variance and extratrees and a minimum node size of 5, 10 and 15 and the number of trees set to 500, 1,000 and 1,500. The results are shows in appendix D under the tables D.5 and D.6.

5.4.1 Price model

From the grid search, the final Random Forest model was estimated for the model with "Price" as dependent variable with a mtry of 54, splitrule set to variance, the minimum node size of 5 and the number of trees set to 1,500. This model had the lowest RSME compared to the other models in the grid search. All these results were cross validated 10 times.

The RSME from the trained model has a value of 91,010 and the R^2 that was found for this model is 0.7769. These values correspond to the training data. For the test data, this model has a RSME of 91,839 and a R^2 of 0.7703. The model explains about 78% of the total variance on the train set and about 77% on the test set.

Diving deeper into the Random Forest model, we see from the variable importance plot in figure 5.9 that the living space size in square meters is the most important variable for this random forest model. On the second place the house type 1 being a single-family home has an important role for the random forest model. Next, also important are the variables the longitude degree, the latitude degree, the lot size per square meter, house type 2 being a detached house, the build year, the number of bathrooms, the longitude, the estimated neighborhood price per square meter and the name of the province being Noord-Holland.

Now, we will investigate the partial dependence plot in figure 5.11. The top left part of figure 5.11 shows the partial dependence of the living space size in square meter. We see that the price increases if the living space increases and stays relatively stable after 200 square meters. The distribution of the observations, however, are concentrated around the 0 to 200 square meters mark, which indicates that the PD estimates are less reliable in the less concentrated areas. On the bottom left part of figure 5.11, we see the PDP for the build year. The price stays relatively constant from 1500 to approximately 1800 and then the price decreases until around 1980, from where the prices rise again. The distribution of the observations is mostly around 1900 to 2000, so the PD estimates are the most reliable in this area. Next, we have the top middle part of figure 5.11 which shows the PDP for the lot size in square meters. The price decreases from 0 to 150 square meters and then increases after around 150 square meters. PD estimates are the most reliable between 0 and 600 square meters. In the bottom middle plot of figure 5.11, we see the longitude degree. We see an increase in price till around 4,8 degrees and from there the price stays relatively constant till it decreases around 5,3 degrees. So, the more east the house resides, the more the price of the house decreases. The distribution of the observations is spread over the whole axis, which indicate that the PD estimates are reliable. The top right part of figure 5.11 shows the PDP for the variable estimated neighborhood price per square meter. We see a general decrease in price till around an estimated neighborhood price of 2,000 euros per square meters. Hereafter, the price increases till around an estimated neighborhood price of 2,000 euros per square meters. From there the variable stays relatively constant. The distribution of the observations is spread between the values of 0 and 6,000, indicating reliable PD estimates within an estimated neighborhood price per square meter of 0 to 6,000 euros. Finally, we have the bottom right plot of figure 5.11 which shows the latitude degree. We see that the latitude degree increases till 52,3 and from there decreases. This indicates that if the house resides in the northern and southern part of the Netherlands, the price then is lower compared to the middle part of the Netherlands. The observations are mainly distributed around the middle part of the Netherlands (51,5 to 53 degrees), which indicates that the PD estimates are not very reliable in the most northern and southern part of the Netherlands.

5.4.2 Price per square meter model

From the grid search, the final Random Forest model was estimated for the model with "Price per square meter" as dependent variable with a mtry of 54, splitrule set to variance, the minimum node size of 5 and the number of trees set to 1,500. This model had the lowest RSME compared to the other models in the grid search. All these results were cross validated 10 times.

The RSME from the trained model has a value of 615 and the R^2 that was found for this model is 0.6320. These values correspond to the training data. For the test data, this model has a RSME of 619 and a R^2 of 0.6178. The model explains about 63% of the total variance on the train and test set and about 62% on the test set.

Diving deeper into the Random Forest model, we see from the variable importance plot in figure 5.10 that the latitude degree the most important variable is for the random forest model. On the second place the longitude degree has an important role for the random forest model. Next, also important are the variables build year, the lot size and living space size per square meter, the province being Noord-Holland, the estimated neighborhood price per square meter, house type 2 being a detached house, the number of rooms and if there is a garden around the house.

Next, we will investigate the partial dependence plot in figure 5.12. The top left part of figure 5.12 shows the partial dependence of the living space size in square meter. We see that the price per square meter decreases if the living space increases. The distribution of the observations, however, are concentrated around the 0 to 200 square meters mark, which indicates that the PD estimates are less reliable in the less concentrated areas. On the bottom left part of figure 5.12, we see the PDP for the build year. The price per square meter stays relatively constant from 1500 to approximately 1800 and then the price per square meter decreases until around 1980, from where the price per square meter rise again. The distribution of the observations is mostly around 1900 to 2000, so the PD estimates are the most reliable in this area. Next, we have the top middle part of figure 5.12 which shows the PDP for the lot size in square meters. The price per square meter decreases from 0 to 150 square meters and then increases after around 150 square meters. PD estimates are the most reliable between 0 and 600 square meters. In the bottom middle plot of figure 5.12, we see the longitude degree. We see an increase in price per square meter till around 4,8 degrees and from there the price per square meter stays relatively constant till it decreases around 5,3 degrees. So, the more east the house resides, the more the price per square meter of the house decreases. The distribution of the observations is spread over the whole axis, which indicate that the PD estimates are reliable. The top right part of figure 5.12 shows the PDP for the variable estimated neighborhood price per square meter. We see a general decrease in price till around an estimated neighborhood price of 2,000 euros per square meters. Hereafter, the price increases till around an estimated neighborhood price of 2,000 euros per square meters, which has a positive effect on the dependent variable "price per square meters". From there the variable stays relatively constant. The distribution of the observations is spread between the values of 0 and 6,000, indicating reliable PD estimates within an estimated neighborhood price per square meter of 0 to 6,000 euros. Finally, we have the bottom right plot of figure 5.12 which shows the latitude degree. We see that the price per square meter increases till 52,3 and from there decreases. This indicates that if the house resides in the northern and southern part of the Netherlands, the price per square meter then is lower compared to the middle part of the Netherlands. The observations are mainly distributed around the middle part of the Netherlands (51,5 to 53 degrees), which indicates that the PD estimates are not very reliable in the most northern and southern part of the Netherlands.

Comparing figures 5.9 and 5.10, we see that some variables return in both plots. For both models the living space size, lot size and estimated neighborhood price in square meters are important variables. Also, the province name in which the house resides in being Noord-Holland, house type 1 being a single-family home, the longitude and latitude degree and the build year are important variables for both models. Comparing figures 5.11 and 5.12, we see that all variables move in the same direction except for the living space size in square meters. For the random forest price model, we see that if the living space size increases, the price increases, whereas for the random forest price per square meter model the price per square meter decreases when the square meters increases.

Figure 5.9 Variable importance plot of the Random Forest model with dependent variable "Price"

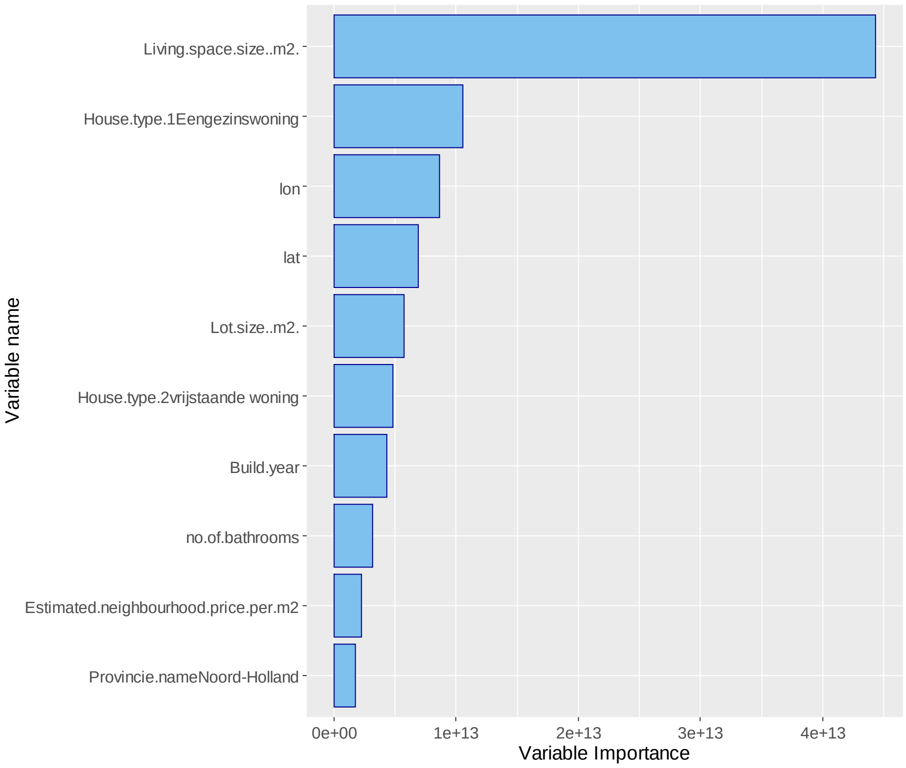


Figure 5.10 Variable importance plot of the Random Forest model with dependent variable "Price per square meter"

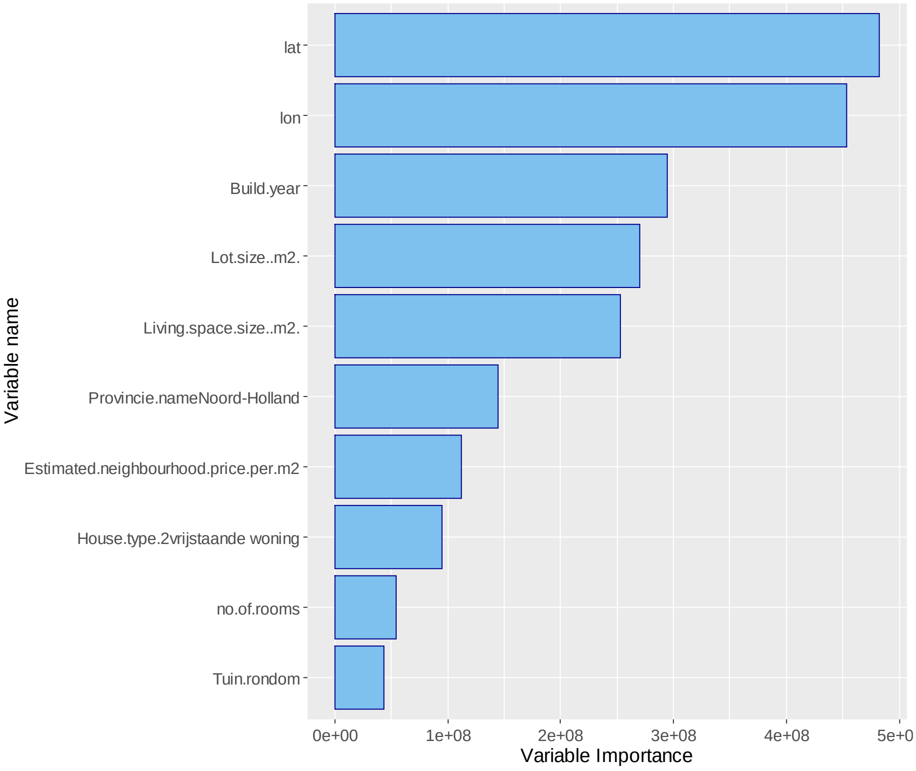


Figure 5.11 PDP for dependent variable "Price" and independent variables "Living space size in square meters", "Lot size in square meters", "Estimated neighborhood price per square meter", "Build Year", "Longitude degree", "Latitude degree"

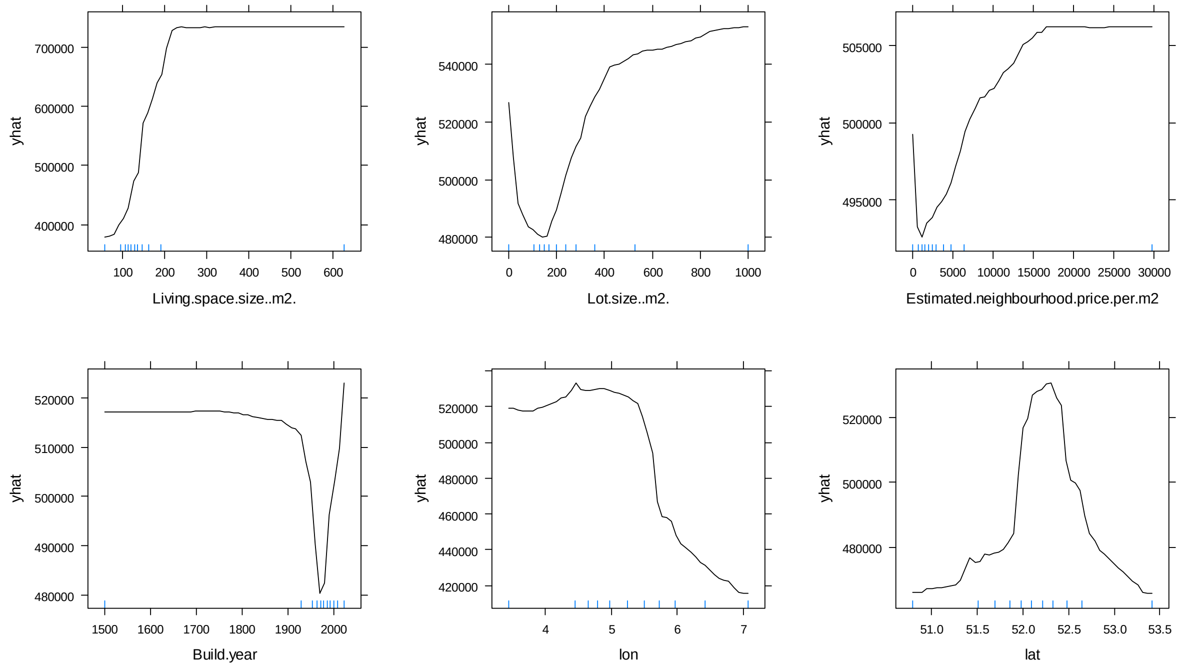
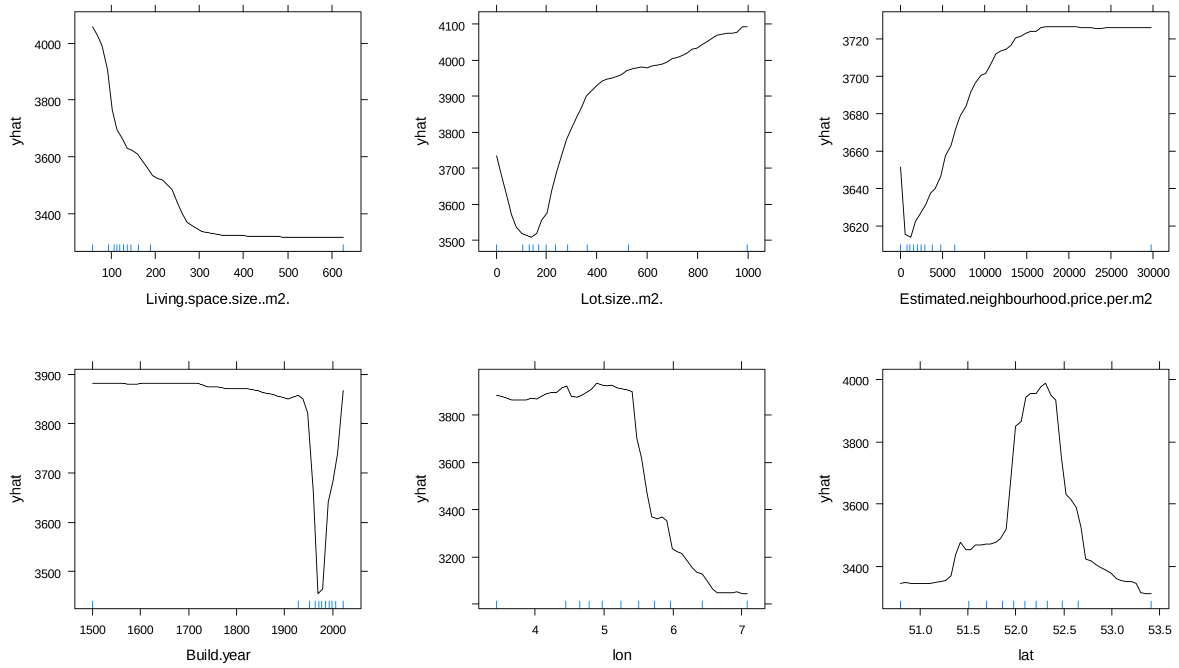


Figure 5.12 PDP for dependent variable "Price per square meter" and independent variables "Living space size in square meters", "Lot size in square meters", " Estimated neighborhood price per square meter ", "Build Year", "Longitude degree", "Latitude degree"



5.5 SLX Model

Finally, we will investigate if neighboring houses have an effect on a specific observation. This will be done with the SLX model. Compared to the models in chapter 5.1 to 5.4, the models in this part of the results use different variables compared to the other models. Variables such as the name of the province, postal area code group, longitude and latitude degree are left out because the SLX models uses spatial data to perform its operations. The variables that are left out are variables that tell us more about the geographical location of the house, and since the spatial dataset already takes care of this, the left-out variables are obsolete. The full SLX models as well as the multiple linear regression models that are used to compare with the SLX model can be found in appendix E under table E.1. For the SLX model, we mainly investigate the indirect effect of the lagged variables. The output of the impact of full model can be found in appendix E under tables E.2 and E.3. For these models, we reject the null hypothesis at a 5% significance level.

5.5.1 Price model

Before estimating a SLX model, we will first investigate whether there is any form of spatial dependence in the dataset. This is done by a Moran I test, which tells us if there is a spatial correlation between the residuals. In order to perform the Moran I test, a linear model needs to be constructed to test if there is any spatial dependence in the data. Therefore, a linear model was created with the same variable as the SLX model with dependent variable "Price". The Moran I test showed a P-value smaller than 1% which means that we reject the null hypothesis under the 5% significance level and accept the alternative hypothesis, which states that there is a spatial correlation in the residuals for dependent variable "Price". Therefore, looking into the SLX models is significant.

Looking at the fit on our data, we see that the RMSE of the linear model[1] is 121,190, while the SLX model returns a RMSE of 102,662. The R^2 of the linear model[2] is 0.5973, while the R^2 of the SLX model is 0.7111. The linear model[3] explains about 60% of the total variance, while the SLX model explains about 71% of the total variance. From this, we can conclude that, for modelling house prices, it is wise to include spatial data compared to using models with no spatial dependence included.

Interpretation of the impact from the SLX model is focused around a specific observation. For interpretation purposes, we call this specific observation "our house".

The SLX model with dependent variable "Price" estimated that the increase of the living space size in square meters from the neighboring houses has a negative effect on the price of our house. This means, that if the living space size of the neighboring houses increases while our living space size stays the same, this then decreases the price of our house. Neighboring houses with energy label D or E

[1] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

[2] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

[3] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

instead of A (the reference category) have a positive effect on our house. We also see, if the number of bathrooms, separate toilets or floors increases for neighboring houses while this variable stays the same for our house, then this influences our house price positively. The same applies for a front garden, side garden or if the neighboring house has an attic. Neighboring houses with house type 1 being a single-family house instead of being a bungalow (the reference category) has a negative effect on the price of our house. So, if our house is a bungalow, and our neighbors' house is a single-family house, this then decreases the price of our house. House type 2 being a detached house for a neighboring house instead of a 2-under-1-roof house (the reference category) has a negative effect on our house. House type 3 being a patio house has a positive effect on the price of our house. If the roof cover is made from bitumen, roof tiles or other materials that were not specified, we see that this has a negative effect on our house price as well. We see that, if the neighboring houses are located in a forestry area or next to a waterway, this then positively influences the price of our house. Neighboring houses next to a busy road have a negative effect on our houses. This can be seen as if the house of the neighbors is next to a busy road, this then decreases the price of our house, without our house being next to a busy road.

5.5.2 Price per square meter model

For the dependent variable "Price per square meter", we again perform the Moran I test, which tells us if there is a spatial correlation between the residuals. In order to perform the Moran I test, a linear model needs to be constructed to test if there is any spatial dependence in the data. Again, a linear model was created with the same variable as the SLX model with dependent variable "Price per square meter". The Moran I test showed a P-value smaller than 1% which means that we reject the null hypothesis under the 5% significance level and accept the alternative hypothesis, which states that there is a spatial correlation in the residuals for dependent variable "Price per square meter". Therefore, looking into the SLX models is significant.

Again, a look at the fit on our data, but now for the dependent variable "Price per square meter" we see that the RMSE of the linear model[4] is 868, while the SLX model returns a RMSE of 729. The R^2 of the linear model[5] is 0.2406, while the R^2 of the SLX model is 0.4651. The linear model[6] explains about 24% of the total variance, while the SLX model explains about 47% of the total variance. From this, we can conclude that, for modelling the price per square meter of a house, it is wise to include spatial data.

Again, interpretation of the impact from the SLX model is focused around a specific observation. For interpretation purposes, we call this specific observation "our house".

[4] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

[5] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

[6] This model contains the same variables as used in the SLX model and thus does not include any geographical variables. This is a different MLR model than the model discussed in chapter 5.1.

The SLX model with dependent variable “Price per square meter” shows that the lot size in square meters from neighboring houses has a negative effect on our house price per square meter. Relatively bad energy labels for neighboring houses, such as label D, E and G instead of A (the reference category), influence the price per square meter of our house positively. The number of bathrooms or separate toilets of neighboring houses influences the price per square meter of our house positively. This also applies for side gardens or attics from neighboring houses. Neighboring houses having a loft, or a back yard have a negative effect on our house. So, if the house of the neighbors has a loft or a back yard while our house does not, this then decreases the price per square meter of our house. Neighboring houses with house type 1 being a single-family house instead of a bungalow (the reference category) have a negative effect on the price per square meter of our house. House type 2 being a detached house, or a staggered house instead of a 2-under-1-roof house all have a negative effect on the price per square meter of our house. If the neighboring houses resides next to a busy road, then this influences the price per square meter of our house negatively. This is not the case for neighboring houses that resides next to a waterway or in a forestry location since this impacts our price per square meter positively. If the roof cover is made from bitumen, or other materials that were not specified for neighboring houses, while our roof cover is not made from one of these categories, we then see that this influences our price per square meter negatively.

Comparing the SLX models for both dependent variables, we see that some variables significantly impact both models. We see that the lot size in square meters for both models have a negative effect on the dependent variables as well as house type 1 being a single family house instead of a bungalow (the reference category), house type 2 being a detached house instead of a 2-under-1-roof house(the reference category), neighboring houses residing next to a busy road and neighboring houses with a roof cover made from bitumen or other materials that were not specified.

Variables that had a positive effect on both the dependent variables are if the neighboring houses have energy label D or E instead of A (the reference category) or if the neighboring houses resides next to a waterway or in a forestry area. If the number of bathrooms or toilets increases for neighboring houses, this then also increases the price and the price per square meter of our house. Finally, if the house has an attic or a side garden, this also increases the price and price per square meter of our house.

5.6 Summary of the models

Table 5.3 and 5.4 shows the RMSE and the R^2 of all the models that were used for predicting in this research. The RMSE and the R^2 are both given for the model on the train data as well as the test data.

From table 5.3, we can conclude that the Random Forest model performs better than the other models. This model has the smallest value for the RMSE and the highest value for the R^2. The same applies for the Random Forest in case of the training data. The decision tree performed the worst on both the train and test data. The decision tree model has the highest value for the RMSE and had the lowest value for the R^2. The support vector machine and the multiple linear regression scored for both the RMSE and the R^2 in between the best and worst model.

Table 5.3 Summary of the predictive models with "Price" as dependent variable

	Training		**Predicting**	
Model	RMSE	Rsquared	RMSE	Rsquared
Multiple Linear Regression	99,440	0.7291	100,667	0.7217
Support Vector Machine	110,591	0.6676	100,001	0.7276
Decision Tree	124,960	0.5734	125.531	0.5676
Random Forest	91,010	0.7769	91,839	0.7703

Table 5.4 Summary of the predictive models with "Price per square meter" as dependent variable

	Training		**Predicting**	
Model	RMSE	Rsquared	RMSE	Rsquared
Multiple Linear Regression	689	0.5238	708	0.4931
Support Vector Machine	731	0.4749	707	0.5002
Decision Tree	836	0.3021	846	0.2796
Random Forest	615	0.6320	619	0.6178

For the models with dependent variable "Price per square meter" (table 5.4) we see similar results as the models with dependent variable "Price". The Random Forest model performs better than the other models. Again, this model has the smallest value for the RMSE and the highest value for the R^2. The same applies for the Random Forest in case of the training data. The decision tree performed the worst on both the train and test data. The decision tree model has the highest value for the RMSE and had the lowest value for the R^2. The support vector machine and the multiple linear regression scored for both the RMSE and the R^2 in between the best and worst model, where the multiple linear regression outperformed the support vector machine on the training data and the support vector machine outperformed the multiple linear regression on the test data.

Investigating the SLX model, we see that the SLX model performs better than a multiple linear regression[7] from table 5.5. This is the case for both dependent variables. In all cases, the RMSE and R^2 had better scores under the SLX model. As the Moran I test suggested, there is some spatial dependence in the data and from using the spatial data we see indeed that this improves the model fit.

[7] The MLR models used in this table do not use any geographical variables and only use the same dependent variables as the SLX model. These MLR models are different than the models from table 5.3 and 5.4.

Investigating the important characteristics of the house, we investigate the variable importance plot of the decision tree, support vector machine, random forest as well as the coefficients of the multiple linear model. Also, we investigate the PDP for the random forest models.

Taking a look at the important variables for the asking price, we see that the living space size and the lot size in square meters seems to be important variables, since these return in multiple models. The same applies to house type 1 being a single-family house instead of a bungalow, house type 2 being a detached house instead of a 2-under-1-roof house, longitude and latitude degree, the build year, a garden around the house, the province in which the house resides in is Noord-Holland and the estimated neighborhood price per square meter.

Table 5.5 Summary of the SLX and MLR model on dependent variables "Price" and "Price per square meter"

	Dependent variable			
	Price		*Price per square meter*	
Model	RMSE	Rsquared	RMSE	Rsquared
Multiple Linear Regression[8]	121,190	0.5973	868	0.2406
SLX	102,662	0.7111	729	0.4651

Looking at the important variables for the asking price per square meter, we can conclude that we see that the living space size and the lot size in square meters seems to be important variables, since these also return in multiple models. We also see that the build year is an important characteristic, as well as the longitude and latitude degree. House type 1 being a single-family house instead of a bungalow, house type 2 being a detached house instead of a 2-under-1-roof house, longitude and latitude degree, the build year, a garden around the house, the province in which the house resides in is Noord-Holland and the estimated neighborhood price per square meter.

Comparing important characteristics for both the price models and the price per square meter models, we see that overall, all the variables that are important across the models with dependent variable "Price" and "Price per square meter" are also important across all the multiple linear regression models, support vector machine models, decision tree models and the random forest models.

Finally, investigating partial dependence plots, we see that the variables for the lot size in square meter, the estimated neighborhood price, the build year, longitude and latitude degree move in the same direction for the random forest price model and random forest price per square meter model. However, we see opposite effects for the living space size in square meters.

From the SLX model, we see that neighboring houses do have an effect on our house for the price and price per square meter. The following variables both have an indirect negative effect on our dependent variables: Lot size in square meters, house type 1 being a single-family house, house type 2 being a

[8] The MLR models used in this table do not use any geographical variables and only use the same dependent variables as the SLX model. These MLR models are different than the models from table 5.3 and 5.4.

detached house, neighboring houses that reside next to busy road, the roof cover made from bitumen or another material that was not specified. Variables from neighboring houses that both have a positive effect on the price or price per square meter of our house include energy label D or E, the house resides next to a waterway or in a forestry area, an increase in the number of bathrooms or separate toilets and the house having an attic or side garden.

6 Conclusion

In this chapter, we will discuss our findings and answer our main- and subquestions. At last, a recommendation for a follow-up research will be provided.

6.1 What are important characteristics of a house for the asking price?

From the different models we see that a lot of variables are important in predicting the house prices. For each machine learning method, 2 models were constructed with different dependent variables. First, we predicted the price based on the variables in our dataset, and then we did the same with price per square meter as dependent variable. We see that the lot size in square meters and the living space size in square meters are important variables for predicting house prices across different models. The build year seems to be rather important.

The multiple linear regression models tell us that the energy label is an important characteristic since the energy labels have a significant effect on the house prices and the price per square meter, ceteris paribus. The importance of the energy label variable is much lower for the other models if we consider the variable importance plots with the top 10 most important variables from figures 5.3, 5.4, 5.5, 5.6 5.9 and 5.10.

Generally speaking, the models show us that the house type (1 and 2) are important characteristics of a house. Specifically speaking about the models, we see that house type 1 being a single-family home or a farmhouse, is important for the multiple linear regression models. House type 1 being a single-family home is important for the multiple linear regressions, the decision tree and the random forest models. For house type 2, we see that an end house, a corner house, an in-between house and a detached house are important for the multiple linear regressions. House type 2 being a detached house seems important across the linear regression as well as the random forest. Having a garden all around the house is also important for multiple models and thus is an important characteristic for a house and its price.

Another important characteristic is where the house resides. We see from the multiple linear regression models that houses in a forestry area, outside the build-up area, next to a waterway or having an open view have an impact on the price and thus are important characteristics of a house. Also, the longitude degree and latitude degree are important across all the 4 types of models. The name of the province seems to be important across the support vector machine, decision tree and random forest. The estimated neighborhood price per square meter seems to be rather important for the multiple linear regression models and the random forest models.

The roof cover seems important for a house too, considering that the multiple linear regression models tell us that the roof cover being made from reed or metal have a significant effect on the house prices and price per square meter.

In general, considering all the predictive models, we found that important characteristics of a house are the living space and lot size per square meter, the number of rooms, bathrooms and separate toilets, the build year, the location of the house (postal area code group, longitude degree, latitude degree, and where the house resides (e.g. in the city center)), the energy label, the type of house and the roof cover of a house.

6.2 How does the neighborhood influence house prices?

The SLX models showed us that the houses in the neighborhood have an effect on the house price itself. The SLX models had as dependent variable the price and the price per square meter. It has shown that price and price per square meter of our house decreases when the lot size in square meters increases for neighboring houses. This is before any correction on our house. In case the neighbors' lot size of their house increases, while our house its lot size stays the same, then this decreases the price of our house compared to the neighbor. The same applies for neighboring houses that are a single-family house or a detached house. If a neighboring house is one of these types of houses, while our house is a bungalow (the reference category), this then influences our price and price per square meter negatively. If the neighboring house resides next to a busy road or has a roof cover made from bitumen or another material that was not specified, this also decreases the price and price per square meter of our house.

On the other hand, the price and price per square meter of our house increases when for neighboring houses have energy label D or E instead of A (the reference level). The same applies for the number of bathrooms and separate toilets. Furthermore, if the neighboring houses are located in a forestry area or next to a waterway, this then positively influences the price and price per square meter of our house. The same applies for neighboring houses when they have a side garden or an attic.

Neighboring houses influence the house price both negatively and positively. We see that the lot size per square meter, energy label, the location, the number of separate toilets and bathrooms, the energy label and the type of house from the neighboring houses have an effect on our house and thus the houses in the neighborhood influences house prices.

6.3 How do different machine learning methods differ in predicting housing prices?

For this research, we conducted 2 linear regression models, 2 support vector machines, 2 decision tree models, 2 random forest models and 2 SLX models. The linear regression models estimate the relationship between the independent variables and one dependent variable by using a straight line or plane (Molnar, 2023). It then predicts the value of a house by estimating the coefficients and adding these to find the predicted price.

This is different compared to the support vector machine because the support vector machine allows us to determine how much error is acceptable in our model and will find an appropriate hyperplane to fit the data. The support vector regression is restricted on its absolute error is lower than or equal to the maximum error that is allowed for the support vector machine. A support vector machine uses a kernel function that matches the test data to the house from the training data to predict the house price from the test data Both the multiple linear regression model and the support vector machine model are oriented around solving a linear relationship between variables.

Not all relationships are linear, that is why this research also considered the decision tree and the random forest. Decision tree models split the data several times due to feature cutoff values, leading to the dataset being split in various subsets. The final subset is a terminal or leaf node, and this is the average outcome of the training data that was used in this part of the tree (Molnar, 2023). This is also how a decision tree predicts a value for the test data. The decision tree, after being built, tries to fit the house onto the tree to find out at which node it splits and in which leaf node it ends up, and then estimates the price of that house.

An extension of the decision tree model is the random forest model. This model builds a number of decision trees on bootstrapped training samples, leading to a forest of trees, which reduces bias and variance. A prediction is made from the mean outcome of a various number of trees and then estimates the price of a house. From tables 5.3 and 5.4, we indeed see that the variance is better explained by our random forest model than our decision tree model.

At last, the SLX model is a derivative of the multiple linear regression model, but this model takes spatial lag into consideration. The SLX model thus is able to conduct computations based on the geographical location of the houses to identify neighboring effects.

6.4 How do models of house prediction perform compared to each other?

From our results, we can conclude that the best model for this research is the random forest model. This model seems to predict the house prices better than the other models, considering the RMSE and the R^2.

Comparing the SLX model with a linear model[9], we see that the SLX model has a better fit with the data in terms of RMSE and the R^2. This also showed us that there is spatial dependence in house prices for the Netherlands.

6.5 What drives the asking price of the Dutch housing market and how can we predict the housing prices?

What drives the asking price of the Dutch housing market and how can we predict the housing prices? From this research we can conclude that several important characteristics from the house itself as well as its geographical location drive the asking price of the Dutch housing market. We concluded i.e., that the number of rooms, the type of house, the type of roof cover and where the house resides all impact the asking price of a house. We also concluded that neighboring houses influence the asking price of a house. Characteristics of neighboring houses do impact the asking price of houses. The best way to make prediction for the asking price of houses on the Dutch housing market can be done by a random forest model. Also, for the houses in the Netherlands, we concluded that neighboring houses influence each other and therefore, there is spatial dependence in the price of houses in the Netherlands.

[9] This linear model includes the same variables as the SLX model. This MLR is also described in table 5.5.

6.6 Limitations and recommendations for a follow-up research

This research only conducted 4,057 houses being for sale while there were around 80,000 for sale on the same website. So, this part of the dataset might be biased towards a variable compared to all the 80.000 houses that were for sale in the same time period. Other houses that were for sale but not listed on funda are also a limitation on this research, since these houses might show us that the houses that were only listed on funda are biased towards a specific variable.

Considering the location of the houses, this dataset did not consist of an observation from every street or full postal area code. So, our estimated model might be far off in predicting houses in a small town that is not known by our model. Cities and municipalities were also not used in the models due to having a relatively small dataset.

For a follow up research, it is recommended to investigate a bigger dataset that has listings for different sources, as well as more houses with higher prices to reduce outliers. This means, however, that a lot of computational power is required to estimate and cross validate the models. For example, when testing my codes on the data, it took over 36 hours to conduct a random forest model with the dataset that was used with the cities and municipalities included (with an AMD Ryzen 9 5800X CPU). Therefore, it is advised to use a machine that is well capable of handling datasets that are i.e., 5 times bigger and estimating large models for these datasets.

Also, investigating into a neural network might lead to better predictions on prices, and therefore a follow-up research including a neural network could be insightful. Combining spatial dependence models with for example a random forest might also be interesting for prediction purposes.

Another limitation was predicting on class variables that were in the test dataset but not in the training dataset. This makes the models unable to predict prices of the houses. This is a model limitation since the R packages in RStudio do not handle predictions on not available variables well. This is also a data limitation, for not having well balanced data of all the variables for the house prices. On the other hand, a recommendation would be to compare different estimated models with and without imputation when a bigger and more complete dataset is not possible.

Appendix A Summary Statistics

Table A.1 shows all the variables that are used in the models for this research. The summary statistics can be found in the second, third and fourth table. Variables with a asterisk (*) are variables that were only used as dependent variables. The variables for the SLX model that were used, were also used for the multiple linear regression that the SLX model was compared to.

Table A.1 Variable list used in the models

Variable Name	MLR	SVM	DT	RF	SLX
Price*	X	X	X	X	X
Lot.size..m2.	X	X	X	X	X
Living.space.size..m2.	X	X	X	X	X
Build.year	X	X	X	X	X
Build.type	X	X	X	X	X
Energy.label	X	X	X	X	X
Estimated.neighbourhood.price.per.m2	X	X	X	X	
PCGROUP	X	X	X	X	
House.type.1	X	X	X	X	X
House.type.2	X	X	X	X	X
House.type.3	X	X	X	X	X
in.woonwijk	X	X	X	X	X
aan.bosrand	X	X	X	X	X
aan.drukke.weg	X	X	X	X	X
aan.park	X	X	X	X	X
aan.rustige.weg	X	X	X	X	X
aan.vaarwater	X	X	X	X	X
aan.water	X	X	X	X	X
bedrijventerrein	X	X	X	X	X
beschutte.ligging	X	X	X	X	X
buiten.bebouwde.kom	X	X	X	X	X
in.bosrijke.omgeving	X	X	X	X	X
in.centrum	X	X	X	X	X
landelijk.gelegen	X	X	X	X	X
open.ligging	X	X	X	X	X
vrij.uitzicht	X	X	X	X	X
zeezicht	X	X	X	X	X
no.of.rooms	X	X	X	X	X
no.of.bedrooms	X	X	X	X	X
no.of.bathrooms	X	X	X	X	X
no.of.sep.toilets	X	X	X	X	X
no.of.floors	X	X	X	X	X
zolder	X	X	X	X	X
kelder	X	X	X	X	X
vliering	X	X	X	X	X

Provincie.name	X	X	X	X	
Achtertuin	X	X	X	X	X
Voortuin	X	X	X	X	X
Zijtuin	X	X	X	X	X
Plaats	X	X	X	X	X
Zonneterras	X	X	X	X	X
Tuin.rondom	X	X	X	X	X
patio.atrium	X	X	X	X	X
Roof.type	X	X	X	X	X
roof.cover.pannen	X	X	X	X	X
roof.cover.asbest	X	X	X	X	X
roof.cover.bitumineuze.dakbedekking	X	X	X	X	X
roof.cover.kunststof	X	X	X	X	X
roof.cover.leisteen	X	X	X	X	X
roof.cover.metaal	X	X	X	X	X
roof.cover.overig	X	X	X	X	X
roof.cover.riet	X	X	X	X	X
Price.per.m2.living.space*	X	X	X	X	X
lat	X	X	X	X	
lon	X	X	X	X	

Table A.2 Summary statistics of the variables in the dataset (1)

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	NA's
Price	149000	365000	465000	558399	645000	4700000	7
Lot.size..m2.	1	133	197	256	322	998	0
Living.space.size..m2.	53.0	110.0	130.0	146.1	162.0	844.0	0
Build.year	1500	1955	1976	1969	1995	2022	102
Estimated.neighbourhood.price.per.m2	15	1280	2345	3124	4088	29775	361
in.woonwijk	0	0	1	0.7448	1	1	0
aan.bosrand	0	0	0	0.02424	0	1	0
aan.drukke.weg	0	0	0	0.02609	0	1	0
aan.park	0	0	0	0.03812	0	1	0
aan.rustige.weg	0	0	0	0.5761	0	1	0
aan.vaarwater	0	0	0	0.02813	0	1	0
aan.water	0	0	0	0.06773	0	1	0
bedrijventerrein	0	0	0	0.00185	0	1	0
beschutte.ligging	0	0	0	0.1186	0	1	0
buiten.bebouwde.kom	0	0	0	0.02942	0	1	0
in.bosrijke.omgeving	0	0	0	0.07846	0	1	0
in.centrum	0	0	0	0.1181	0	1	0
landelijk.gelegen	0	0	0	0.05996	0	1	0
open.ligging	0	0	0	0.04885	0	1	0
vrij.uitzicht	0	0	0	0.2019	0	1	0
zeezicht	0	0	0	0.0005551	0	1	0
no.of.rooms	2	5	5	5.414	6	18	0
no.of.bedrooms	1	3	4	3.849	4	16	8
no.of.bathrooms	1	1	1	1.215	1	7	0
no.of.sep.toilets	1	1	1	1.175	1	4	364
no.of.floors	1	2	3	2.581	3	6	276
zolder	0	0	0	0.1982	0	1	0
kelder	0	0	0	0.1157	0	1	0
vliering	0	0	0	0.09567	0	1	0
Achtertuin	0	1	1	0.8181	1	1	0
Voortuin	0	0	1	0.6778	1	1	0
Zijtuin	0	0	0	0.2085	0	1	0
Plaats	0	0	0	0.004441	0	1	0
Zonneterras	0	0	0	0.04885	0	1	0
Tuin.rondom	0	0	0	0.163	0	1	0
patio.atrium	0	0	0	0.01147	0	1	0
roof.cover.pannen	0	1	1	0.8157	1	1	0

roof.cover.asbest	0	0	0	0.001665	0	1	0
roof.cover.bitumineuze.dakbedekking	0	0	0	0.1817	0	1	0
roof.cover.kunststof	0	0	0	0.004441	0	1	0
roof.cover.leisteen	0	0	0	0.004811	0	1	0
roof.cover.metaal	0	0	0	0.003516	0	1	0
roof.cover.overig	0	0	0	0.007587	0	1	0
roof.cover.riet	0	0	0	0.01647	0	1	0
Price.per.m2.living.space	1078	2993	3564	3782	4310	13808	7
Price.per.m2.lot.size	277	1456	2152	49442	3237	3312192	7

Table A.3 Summary statistics of the variables in the dataset (2)

Variable Name			
City	Eindhoven : 100	Apeldoorn : 80	Almere : 77
Build.type	Bestaande bouw:5402	Nieuwbouw : 2	
Energy.label	C :1508	A :1204	B : 905
PC	4841 : 22	3881: 21	1506: 19
House.type.1	Eengezinswoning:4366	Villa : 361	Herenhuis : 316
House.type.2	tussenwoning :1715	vrijstaande woning :1522	2-onder-1-kapwoning: 907
House.type.3	:5167	semi-bungalow : 69	drive-in woning : 40
Roof.type	Zadeldak :3386	Plat dak : 633	Samengesteld dak: 616
Provincie.name	Noord-Holland:921	Gelderland :897	Noord-Brabant:893
Gemeente.name	Zaanstad : 119	Eindhoven : 100	Apeldoorn : 95

City	Amersfoort: 66	Enschede : 64	Zaandam : 62	(Other) :4955
Build.type				
Energy.label	D : 621	E : 410	F : 293	(Other): 463
PC	1851: 18	3772: 18	3207: 17	(Other):5289
House.type.1	Bungalow : 164	Woonboerderij : 126	Landhuis : 61	(Other) : 10
House.type.2	hoekwoning : 643	geschakelde woning : 230	eindwoning : 145	(Other) : 242
House.type.3	split-level woning: 38	dijkwoning : 34	hofjeswoning : 15	(Other) : 41
Roof.type	Schilddak : 221	Dwarskap : 169	Lessenaardak : 169	(Other) : 210
Provincie.name	Zuid-Holland :870	Utrecht :517	Overijssel :367	(Other) :939
Gemeente.name	Almere : 71	Amersfoort: 71	Enschede : 64	(Other) :4884

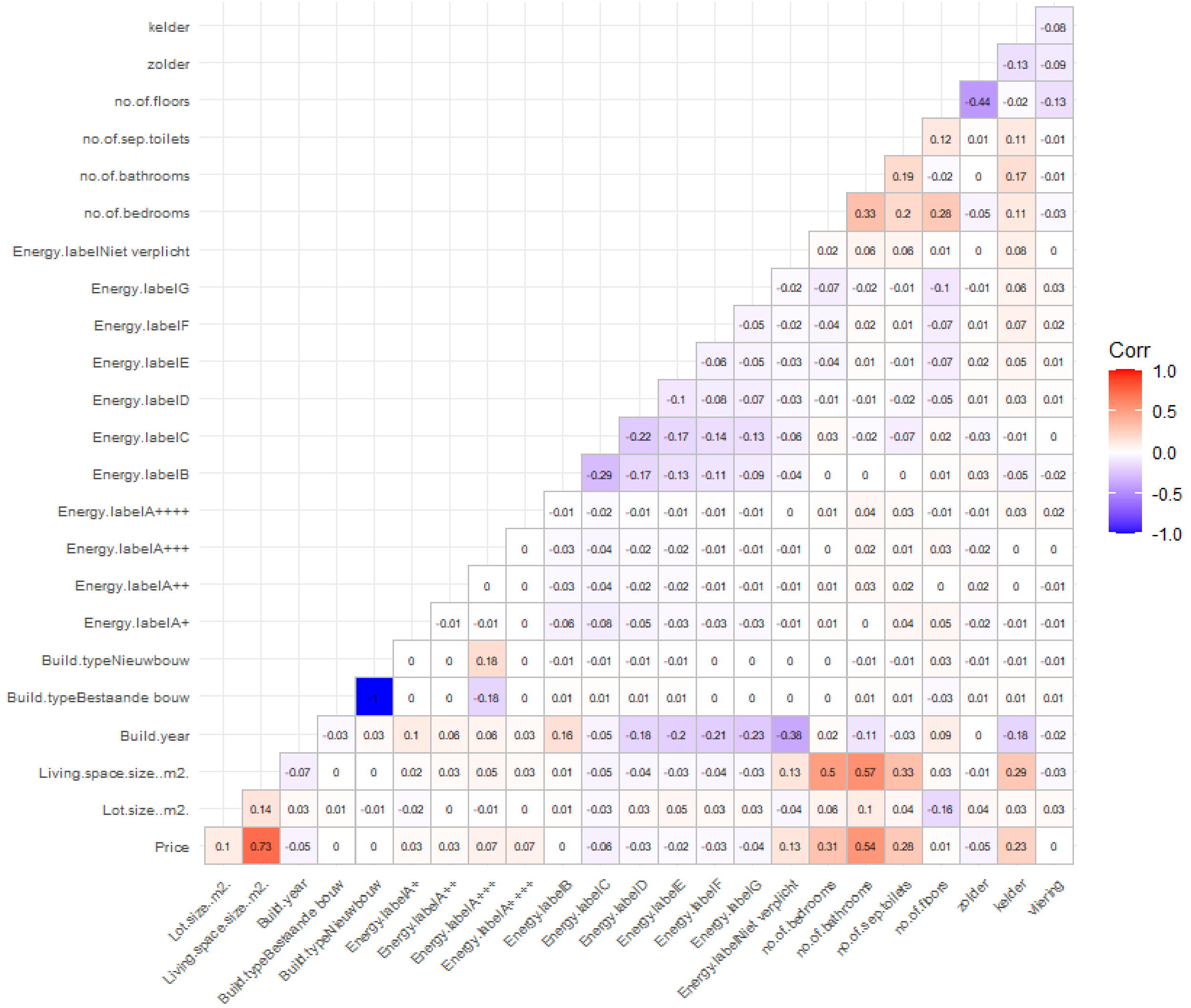
Table A.4 Summary statistics of the variables in the dataset (3)

Variable Name			
Address	Length:5404	Class :character	Mode :character

Appendix B Correlation Coefficient

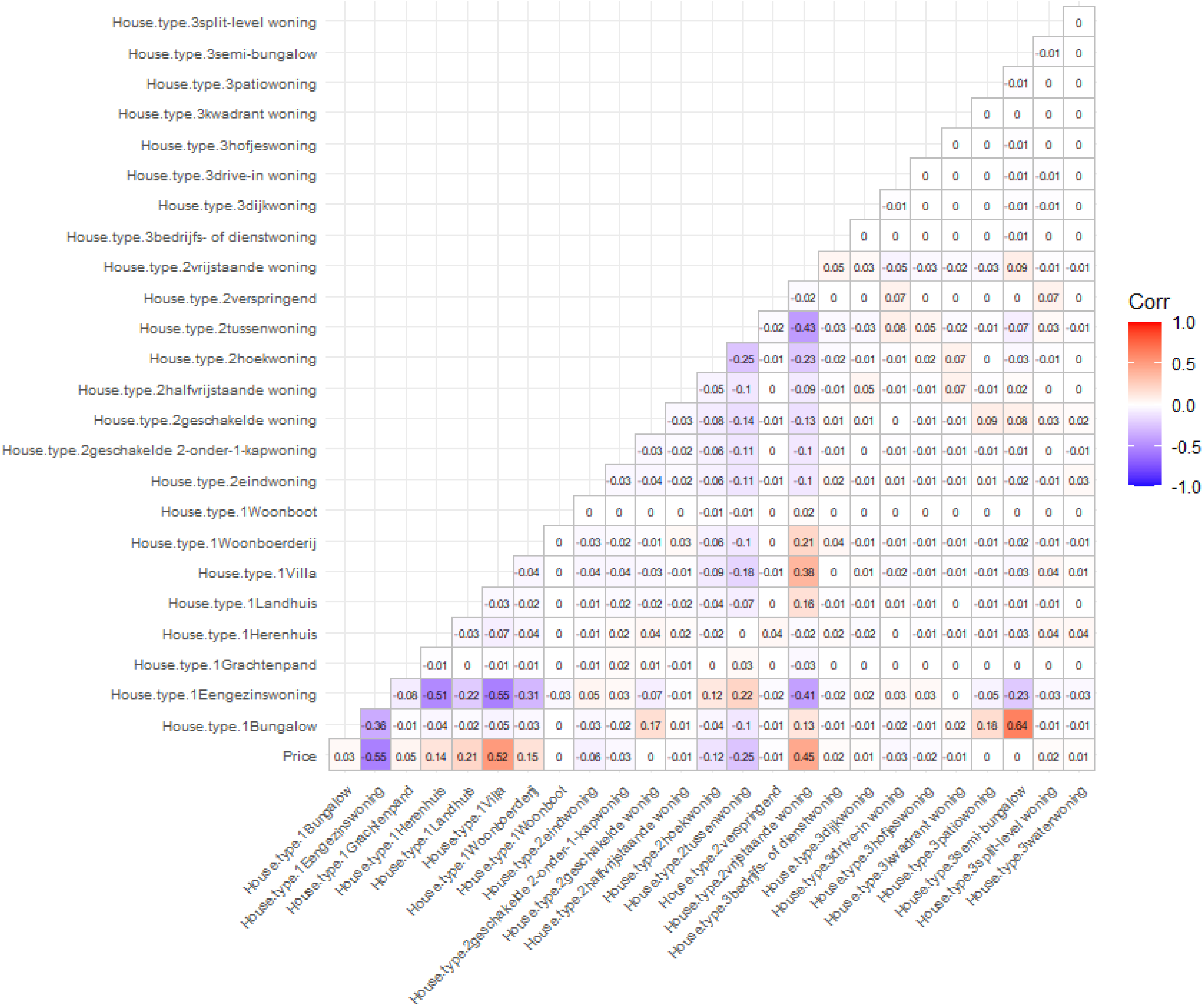
Figure B.1 to B.4 shows the correlation coefficients as discussed in chapter 3.

Table B.1 Correlation coefficients on year, energy label, size, rooms and build type



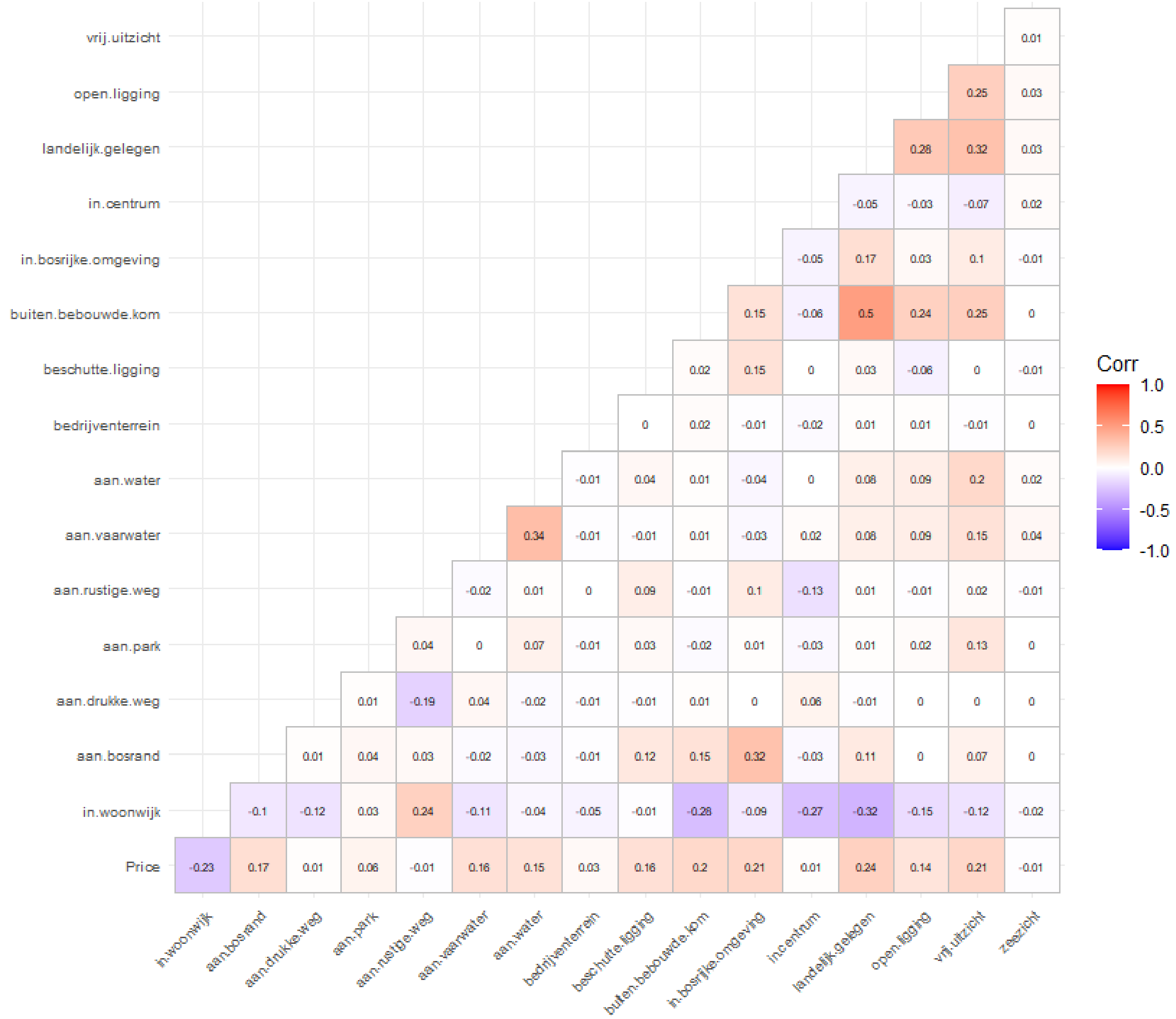
	Lot.size..m2.	Living.space.size..m2.	Build.year	Build.typeBestaande bouw	Build.typeNieuwbouw	Energy.labelA+	Energy.labelA++	Energy.labelA+++	Energy.labelA++++	Energy.labelB	Energy.labelC	Energy.labelD	Energy.labelE	Energy.labelF	Energy.labelG	Energy.labelNiet verplicht	no.of.bedrooms	no.of.bathrooms	no.of.sep.toilets	no.of.floors	zolder	kelder	vliering
kelder																							-0.08
zolder																						-0.13	-0.09
no.of.floors																					-0.44	-0.02	-0.13
no.of.sep.toilets																				0.12	0.01	0.11	-0.01
no.of.bathrooms																			0.19	-0.02	0	0.17	-0.01
no.of.bedrooms																		0.33	0.2	0.28	-0.05	0.11	-0.03
Energy.labelNiet verplicht																	0.02	0.06	0.06	0.01	0	0.08	0
Energy.labelG																-0.02	-0.07	-0.02	-0.01	-0.1	-0.01	0.06	0.03
Energy.labelF															-0.05	-0.02	-0.04	0.02	0.01	-0.07	0.01	0.07	0.02
Energy.labelE														-0.06	-0.05	-0.03	-0.04	0.01	-0.01	-0.07	0.02	0.05	0.01
Energy.labelD													-0.1	-0.08	-0.07	-0.03	-0.01	-0.01	-0.02	-0.05	0.01	0.03	0.01
Energy.labelC												-0.22	-0.17	-0.14	-0.13	-0.06	0.03	-0.02	-0.07	0.02	-0.03	-0.01	0
Energy.labelB											-0.29	-0.17	-0.13	-0.11	-0.09	-0.04	0	0	0	0.01	0.03	-0.05	-0.02
Energy.labelA++++										-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0	0.01	0.04	0.03	-0.01	-0.01	0.03	0.02
Energy.labelA+++									0	-0.03	-0.04	-0.02	-0.02	-0.01	-0.01	-0.01	0	0.02	0.01	0.03	-0.02	0	0
Energy.labelA++								0	0	-0.03	-0.04	-0.02	-0.02	-0.01	-0.01	-0.01	0.01	0.03	0.02	0	0.02	0	-0.01
Energy.labelA+							-0.01	-0.01	0	-0.06	-0.08	-0.05	-0.03	-0.03	-0.03	-0.01	0.01	0	0.04	0.05	-0.02	-0.01	-0.01
Build.typeNieuwbouw						0	0	0.18	0	-0.01	-0.01	-0.01	-0.01	0	0	0	0	-0.01	-0.01	0.03	-0.01	-0.01	-0.01
Build.typeBestaande bouw					-1	0	0	-0.18	0	0.01	0.01	0.01	0.01	0	0	0	0	0.01	0.01	-0.03	0.01	0.01	0.01
Build.year				-0.03	0.03	0.1	0.06	0.06	0.03	0.16	-0.05	-0.18	-0.2	-0.21	-0.23	-0.38	0.02	-0.11	-0.03	0.09	0	-0.18	-0.02
Living.space.size..m2.			-0.07	0	0	0.02	0.03	0.05	0.03	0.01	-0.05	-0.04	-0.03	-0.04	-0.03	0.13	0.5	0.57	0.33	0.03	-0.01	0.29	-0.03
Lot.size..m2.		0.14	0.03	0.01	-0.01	-0.02	0	-0.01	0	0.01	-0.03	0.03	0.05	0.03	0.03	-0.04	0.06	0.1	0.04	-0.16	0.04	0.03	0.03
Price	0.1	0.73	-0.05	0	0	0.03	0.03	0.07	0.07	0	-0.06	-0.03	-0.02	-0.03	-0.04	0.13	0.31	0.54	0.28	0.01	-0.05	0.23	0

Table B.2 Correlation coefficients of the house type variables



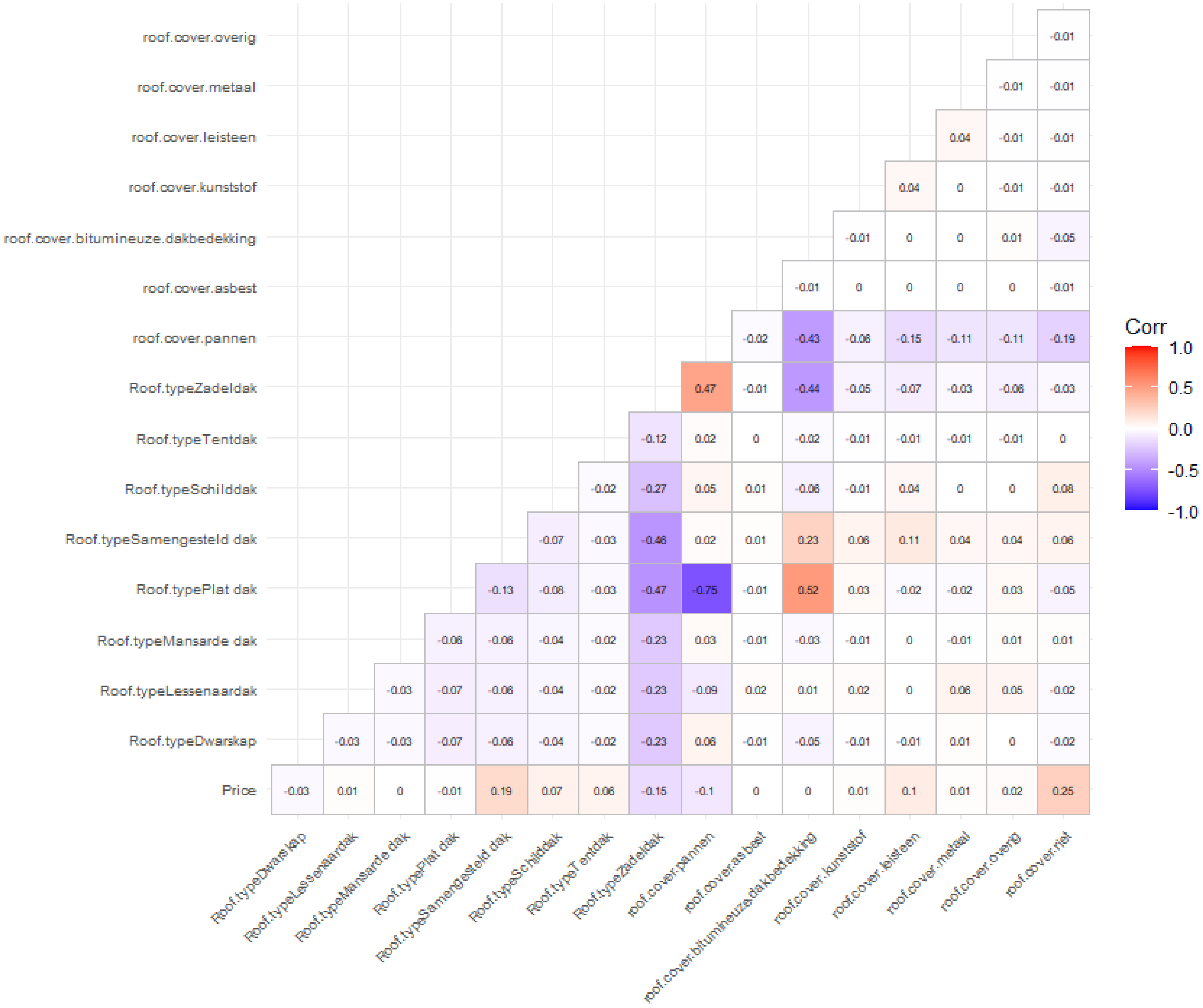
	House.type.1Bungalow	House.type.1Eengezinswoning	House.type.1Grachtenpand	House.type.1Herenhuis	House.type.1Landhuis	House.type.1Villa	House.type.1Woonboerderij	House.type.1Woonboot	House.type.2eindwoning	House.type.2geschakelde 2-onder-1-kapwoning	House.type.2geschakelde woning	House.type.2halfvrijstaande woning	House.type.2hoekwoning	House.type.2tussenwoning	House.type.2verspringend	House.type.2vrijstaande woning	House.type.3bedrijfs- of dienstwoning	House.type.3dijkwoning	House.type.3drive-in woning	House.type.3hofjeswoning	House.type.3kwadrant woning	House.type.3patiowoning	House.type.3semi-bungalow	House.type.3split-level woning	House.type.3waterwoning
House.type.3split-level woning																									0
House.type.3semi-bungalow																								-0.01	0
House.type.3patiowoning																							-0.01	0	0
House.type.3kwadrant woning																						0	0	0	0
House.type.3hofjeswoning																					0	0	-0.01	0	0
House.type.3drive-in woning																				0	0	0	-0.01	-0.01	0
House.type.3dijkwoning																			-0.01	0	0	0	-0.01	-0.01	0
House.type.3bedrijfs- of dienstwoning																		0	0	0	0	0	-0.01	0	0
House.type.2vrijstaande woning																	0.05	0.03	-0.05	-0.03	-0.02	-0.03	0.09	-0.01	-0.01
House.type.2verspringend																-0.02	0	0	0.07	0	0	0	0	0.07	0
House.type.2tussenwoning															-0.02	-0.43	-0.03	-0.03	0.08	0.05	-0.02	-0.01	-0.07	0.03	-0.01
House.type.2hoekwoning														-0.25	-0.01	-0.23	-0.02	-0.01	-0.01	0.02	0.07	0	-0.03	-0.01	0
House.type.2halfvrijstaande woning													-0.05	-0.1	0	-0.09	-0.01	0.05	-0.01	-0.01	0.07	-0.01	0.02	0	0
House.type.2geschakelde woning												-0.03	-0.08	-0.14	-0.01	-0.13	0.01	0.01	0	-0.01	-0.01	0.09	0.08	0.03	0.02
House.type.2geschakelde 2-onder-1-kapwoning											-0.03	-0.02	-0.06	-0.11	0	-0.1	-0.01	0	-0.01	-0.01	-0.01	-0.01	-0.01	0	-0.01
House.type.2eindwoning										-0.03	-0.04	-0.02	-0.06	-0.11	-0.01	-0.1	0.02	-0.01	0.01	-0.01	-0.01	0.01	-0.02	-0.01	0.03
House.type.1Woonboot									0	0	0	0	-0.01	-0.01	0	0.02	0	0	0	0	0	0	0	0	0
House.type.1Woonboerderij								0	-0.03	-0.02	-0.01	0.03	-0.06	-0.1	0	0.21	0.04	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01
House.type.1Villa							-0.04	0	-0.04	-0.04	-0.03	-0.01	-0.09	-0.18	-0.01	0.38	0	0.01	-0.02	-0.01	-0.01	-0.01	-0.03	0.04	0.01
House.type.1Landhuis						-0.03	-0.02	0	-0.01	-0.02	-0.02	-0.02	-0.04	-0.07	0	0.16	-0.01	-0.01	0.01	-0.01	0	-0.01	-0.01	-0.01	0
House.type.1Herenhuis					-0.03	-0.07	-0.04	0	-0.01	0.02	0.04	0.02	-0.02	0	0.04	-0.02	0.02	-0.02	0	-0.01	-0.01	-0.01	-0.03	0.04	0.04
House.type.1Grachtenpand				-0.01	0	-0.01	-0.01	0	-0.01	0.02	0.01	-0.01	0	0.03	0	-0.03	0	0	0	0	0	0	0	0	0
House.type.1Eengezinswoning			-0.08	-0.51	-0.22	-0.55	-0.31	-0.03	0.05	0.03	-0.07	-0.01	0.12	0.22	-0.02	-0.41	-0.02	0.02	0.03	0.03	0	-0.05	-0.23	-0.03	-0.03
House.type.1Bungalow		-0.36	-0.01	-0.04	-0.02	-0.05	-0.03	0	-0.03	-0.02	0.17	0.01	-0.04	-0.1	-0.01	0.13	-0.01	-0.01	-0.02	-0.01	0.02	0.18	0.64	-0.01	-0.01
Price	0.03	-0.55	0.05	0.14	0.21	0.52	0.15	0	-0.06	-0.03	0	-0.01	-0.12	-0.25	-0.01	0.45	0.02	0.01	-0.03	-0.02	-0.01	0	0	0.02	0.01

Table B.3 Correlation coefficients of the house type variables



	in.woonwijk	aan.bosrand	aan.drukke.weg	aan.park	aan.rustige.weg	aan.vaarwater	aan.water	bedrijventerrein	beschutte.ligging	buiten.bebouwde.kom	in.bosrijke.omgeving	in.centrum	landelijk.gelegen	open.ligging	vrij.uitzicht	zeezicht
vrij.uitzicht																0.01
open.ligging															0.25	0.03
landelijk.gelegen														0.28	0.32	0.03
in.centrum													-0.05	-0.03	-0.07	0.02
in.bosrijke.omgeving												-0.05	0.17	0.03	0.1	-0.01
buiten.bebouwde.kom											0.15	-0.06	0.5	0.24	0.25	0
beschutte.ligging										0.02	0.15	0	0.03	-0.06	0	-0.01
bedrijventerrein									0	0.02	-0.01	-0.02	0.01	0.01	-0.01	0
aan.water								-0.01	0.04	0.01	-0.04	0	0.08	0.09	0.2	0.02
aan.vaarwater							0.34	-0.01	-0.01	0.01	-0.03	0.02	0.08	0.09	0.15	0.04
aan.rustige.weg						-0.02	0.01	0	0.09	-0.01	0.1	-0.13	0.01	-0.01	0.02	-0.01
aan.park					0.04	0	0.07	-0.01	0.03	-0.02	0.01	-0.03	0.01	0.02	0.13	0
aan.drukke.weg				0.01	-0.19	0.04	-0.02	-0.01	-0.01	0.01	0	0.06	-0.01	0	0	0
aan.bosrand			0.01	0.04	0.03	-0.02	-0.03	-0.01	0.12	0.15	0.32	-0.03	0.11	0	0.07	0
in.woonwijk		-0.1	-0.12	0.03	0.24	-0.11	-0.04	-0.05	-0.01	-0.28	-0.09	-0.27	-0.32	-0.15	-0.12	-0.02
Price	-0.23	0.17	0.01	0.06	-0.01	0.16	0.15	0.03	0.16	0.2	0.21	0.01	0.24	0.14	0.21	-0.01

Table B 4 Correlation coefficients of the roof variables



	Roof.typeDwarskap	Roof.typeLessenaardak	Roof.typeMansarde dak	Roof.typePlat dak	Roof.typeSamengesteld dak	Roof.typeSchilddak	Roof.typeTentdak	Roof.typeZadeldak	roof.cover.pannen	roof.cover.asbest	roof.cover.bitumineuze.dakbedekking	roof.cover.kunststof	roof.cover.leisteen	roof.cover.metaal	roof.cover.overig	roof.cover.riet
roof.cover.overig																-0.01
roof.cover.metaal															-0.01	-0.01
roof.cover.leisteen														0.04	-0.01	-0.01
roof.cover.kunststof													0.04	0	-0.01	-0.01
roof.cover.bitumineuze.dakbedekking												-0.01	0	0	0.01	-0.05
roof.cover.asbest											-0.01	0	0	0	0	-0.01
roof.cover.pannen										-0.02	-0.43	-0.06	-0.15	-0.11	-0.11	-0.19
Roof.typeZadeldak									0.47	-0.01	-0.44	-0.05	-0.07	-0.03	-0.06	-0.03
Roof.typeTentdak								-0.12	0.02	0	-0.02	-0.01	-0.01	-0.01	-0.01	0
Roof.typeSchilddak							-0.02	-0.27	0.05	0.01	-0.06	-0.01	0.04	0	0	0.08
Roof.typeSamengesteld dak						-0.07	-0.03	-0.46	0.02	0.01	0.23	0.06	0.11	0.04	0.04	0.06
Roof.typePlat dak					-0.13	-0.08	-0.03	-0.47	-0.75	-0.01	0.52	0.03	-0.02	-0.02	0.03	-0.05
Roof.typeMansarde dak				-0.06	-0.06	-0.04	-0.02	-0.23	0.03	-0.01	-0.03	-0.01	0	-0.01	0.01	0.01
Roof.typeLessenaardak			-0.03	-0.07	-0.06	-0.04	-0.02	-0.23	-0.09	0.02	0.01	0.02	0	0.06	0.05	-0.02
Roof.typeDwarskap		-0.03	-0.03	-0.07	-0.06	-0.04	-0.02	-0.23	0.06	-0.01	-0.05	-0.01	-0.01	0.01	0	-0.02
Price	-0.03	0.01	0	-0.01	0.19	0.07	0.06	-0.15	-0.1	0	0	0.01	0.1	0.01	0.02	0.25

Appendix C Model Summary Multiple Linear Regression

This appendix shows the full models that were used when predicting the house prices.

Table C.1 Summary of the multiple linear regression models with dependent variable price and price per square meter

	Dependent variable	
	Price	Price.per.square.meter
	(1)	(2)
Lot.size..m2.	33.083***	0.174**
	(12.719)	(0.088)
Living.space.size..m2.	1,605.255***	-9.409***
	(69.872)	(0.484)
Build.year	-207.450***	-2.139***
	(77.560)	(0.537)
Build.typeNieuwbouw	-33,520.330	-44.036
	(75,383.290)	(522.204)
Energy.labelA+	13,704.000	105.419
	(15,801.770)	(109.464)
Energy.labelA++	26,219.470	284.072
	(30,078.570)	(208.364)
Energy.labelA+++	111,508.900***	656.007**
	(41,206.300)	(285.449)
Energy.labelB	-26,409.680***	-195.072***
	(6,021.882)	(41.715)
Energy.labelC	-43,187.020***	-278.724***
	(5,879.068)	(40.726)
Energy.labelD	-38,222.830***	-250.480***
	(7,839.512)	(54.307)
Energy.labelE	-29,419.320***	-159.769**
	(9,136.914)	(63.294)
Energy.labelF	-57,513.070***	-348.944***
	(10,507.540)	(72.789)
Energy.labelG	-74,985.030***	-378.627***
	(11,589.450)	(80.284)

Energy.labelNiet verplicht	21,163.880	116.022
	(27,819.950)	(192.718)
Estimated.neighbourhood.price.per.m2	2.192***	0.019***
	(0.672)	(0.005)
PCGROUP2XXX	74,695.230***	494.766***
	(11,762.440)	(81.482)
PCGROUP3XXX	7,873.289	1.937
	(13,690.590)	(94.839)
PCGROUP4XXX	-27,910.600	-285.893**
	(18,504.630)	(128.187)
PCGROUP5XXX	-6,562.351	-144.320
	(20,261.350)	(140.357)
PCGROUP6XXX	-40,293.340**	-382.518***
	(17,925.910)	(124.178)
PCGROUP7XXX	-59,882.630***	-502.838***
	(17,502.230)	(121.243)
PCGROUP8XXX	-28,498.140*	-266.127**
	(15,288.120)	(105.906)
PCGROUP9XXX	-63,633.930***	-487.467***
	(22,090.130)	(153.025)
House.type.1Eengezinswoning	-75,408.980***	-484.814***
	(16,908.490)	(117.130)
House.type.1Grachtenpand	9,606.305	-39.739
	(63,627.060)	(440.765)
House.type.1Herenhuis	16,222.490	28.614
	(19,368.590)	(134.172)
House.type.1Landhuis	8,825.212	-170.074
	(31,177.250)	(215.975)
House.type.1Villa	37,779.890**	-43.981
	(19,254.880)	(133.385)

House.type.1Woonboerderij	-58,708.640**	-378.688**
	(26,259.290)	(181.906)
House.type.2eindwoning	-64,880.120***	-474.374***
	(11,628.370)	(80.553)
House.type.2geschakelde 2-onder-1-kapwoning	-28,816.410**	-178.153*
	(13,120.520)	(90.890)
House.type.2geschakelde woning	-16,403.270	-90.463
	(10,273.960)	(71.171)
House.type.2halfvrijstaande woning	26,188.540*	189.113*
	(14,595.920)	(101.111)
House.type.2hoekwoning	-62,393.840***	-423.497***
	(6,985.398)	(48.390)
House.type.2tussenwoning	-81,500.700***	-558.688***
	(6,233.100)	(43.179)
House.type.2verspringend	-100,459.200*	-712.818*
	(53,037.940)	(367.411)
House.type.2vrijstaande woning	43,995.460***	306.246***
	(7,765.988)	(53.797)
House.type.3bedrijfs- of dienstwoning	73,601.450*	507.777*
	(41,600.440)	(288.179)
House.type.3dijkwoning	-27,053.180	-359.161*
	(27,054.940)	(187.418)
House.type.3drive-in woning	-30,725.810	-290.426**
	(20,812.510)	(144.175)
House.type.3hofjeswoning	48,027.290	438.354
	(42,323.990)	(293.192)
House.type.3kwadrant woning	-16,735.520	-6.754
	(47,094.610)	(326.239)
House.type.3patiowoning	45,930.940	214.637
	(39,618.300)	(274.448)
House.type.3semi-bungalow	-45,687.200*	-306.985*

	(23,332.760)	(161.633)
House.type.3split-level woning	-5,131.277	-30.621
	(22,937.790)	(158.897)
House.type.3waterwoning	-77,767.140	-514.395
	(51,966.110)	(359.986)
in.woonwijk	3,391.686	1.877
	(5,616.802)	(38.909)
aan.bosrand	28,504.730**	78.098
	(13,946.930)	(96.615)
aan.drukke.weg	-18,645.680	-153.262*
	(12,314.380)	(85.306)
aan.park	3,830.999	53.001
	(10,399.620)	(72.041)
aan.rustige.weg	746.897	10.821
	(4,024.688)	(27.880)
aan.vaarwater	54,878.470***	318.099***
	(13,554.740)	(93.898)
aan.water	2,062.277	-24.738
	(8,715.242)	(60.373)
bedrijventerrein	-36,264.300	-241.320
	(41,454.530)	(287.169)
beschutte.ligging	3,124.483	8.558
	(6,242.253)	(43.242)
buiten.bebouwde.kom	33,757.160**	269.903**
	(16,213.600)	(112.317)
in.bosrijke.omgeving	25,376.020***	237.541***
	(8,319.839)	(57.634)
in.centrum	-14,434.260**	-33.983
	(6,635.711)	(45.968)
landelijk.gelegen	22,263.320*	244.749***
	(11,701.420)	(81.059)

open.ligging	-21,934.970**	-170.965**
	(10,229.250)	(70.861)
vrij.uitzicht	12,279.730**	97.982***
	(5,346.526)	(37.037)
zeezicht	102,880.600	828.814
	(103,288.900)	(715.515)
no.of.rooms	6,224.789**	22.746
	(2,718.777)	(18.834)
no.of.bedrooms	-6,863.600**	-64.896***
	(3,235.447)	(22.413)
no.of.bathrooms	37,640.740***	290.409***
	(5,345.154)	(37.028)
no.of.sep.toilets	27,396.570***	156.914***
	(5,306.900)	(36.763)
Roof.typeLessenaardak	214.354	-57.282
	(15,256.650)	(105.688)
Roof.typeMansarde dak	-6,585.105	-57.858
	(15,940.060)	(110.422)
Roof.typePlat dak	6,105.408	13.378
	(15,048.910)	(104.249)
Roof.typeSamengesteld dak	5,591.423	-2.207
	(12,703.980)	(88.004)
Roof.typeSchilddak	894.180	44.899
	(14,851.480)	(102.881)
Roof.typeTentdak	-12,227.340	-131.044
	(23,961.380)	(165.988)
Roof.typeZadeldak	-5,143.142	-35.144
	(11,096.300)	(76.868)
no.of.floors	16,220.940***	30.413
	(5,026.618)	(34.821)

zolder	-2,978.922	-61.022
	(6,286.151)	(43.546)
kelder	20,214.120**	121.069*
	(8,988.169)	(62.264)
vliering	11,442.750	49.472
	(7,136.475)	(49.437)
Provincie.nameDrenthe	24,880.190	688.339
	(102,803.300)	(712.151)
Provincie.nameFlevoland	-2,296.099	513.669
	(104,880.600)	(726.541)
Provincie.nameFryslÃ¢n	-49,438.700	151.716
	(103,416.800)	(716.400)
Provincie.nameGelderland	48,723.190	927.237
	(104,513.800)	(724.000)
Provincie.nameGroningen	-11,525.190	442.699
	(102,953.000)	(713.188)
Provincie.nameLimburg	-81,708.940	39.256
	(107,552.700)	(745.051)
Provincie.nameNoord-Brabant	-5,541.548	525.773
	(107,143.800)	(742.219)
Provincie.nameNoord-Holland	85,492.220	1,229.757*
	(105,689.800)	(732.147)
Provincie.nameOverijssel	40,227.540	832.724
	(103,971.900)	(720.246)
Provincie.nameUtrecht	92,146.090	1,298.536*
	(105,336.600)	(729.700)
Provincie.nameZeeland	-125,897.500	-289.101
	(109,594.900)	(759.199)
Provincie.nameZuid-Holland	-21,232.570	449.582
	(106,545.000)	(738.071)
Achtertuin	14,218.390	59.501

	(9,586.806)	(66.411)
Voortuin	8,334.375	48.833
	(5,478.967)	(37.955)
Zijtuin	20,595.060***	170.495***
	(5,540.669)	(38.382)
Plaats	-55,389.860*	-53.364
	(29,426.420)	(203.846)
Zonneterras	28,381.920***	201.716***
	(9,435.610)	(65.363)
Tuin.rondom	82,660.260***	472.459***
	(10,760.120)	(74.539)
patio.atrium	16,886.350	138.996
	(21,027.000)	(145.661)
roof.cover.pannen	-13,795.540	-84.214
	(8,816.780)	(61.077)
roof.cover.asbest	-45,566.360	-281.711
	(47,458.150)	(328.758)
roof.cover.bitumineuze.dakbedekking	-9,852.327	-54.047
	(6,228.153)	(43.144)
roof.cover.kunststof	-27,081.780	-54.943
	(28,291.320)	(195.983)
roof.cover.leisteen	-6,641.259	37.749
	(35,717.040)	(247.423)
roof.cover.metaal	-112,911.900***	-207.434
	(34,122.070)	(236.374)
roof.cover.overig	-35,219.100	-243.839
	(25,789.460)	(178.652)
roof.cover.riet	82,125.630***	397.482**
	(25,613.540)	(177.433)
lon	-40,309.720***	-262.501***
	(9,174.936)	(63.558)

lat	-25,151.310	-189.425*
	(16,596.150)	(114.967)
Constant	2,149,090.000**	19,888.670***
	(908,221.200)	(6,291.534)
Observations	3,046	3,046
R2	0.729	0.524
Adjusted R2	0.719	0.507
Residual Std. Error (df = 2939)	101,233.900	701.279
F Statistic (df = 106; 2939)	74.609***	30.503***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Appendix D Grid Search Tables

Table D.1 Grid search table for the SVM with dependent variable "Price"

cp	RMSE	Rsquared	MAE
0.1026	111,035	0.6650	77,175
0.2051	110,741	0.6662	77,013
0.3077	110,701	0.6665	77,010
0.4103	110,687	0.6666	76,997
0.5128	110,659	0.6668	76,985
0.6154	110,665	0.6668	77,000
0.7179	110,669	0.6668	76,992
0.8205	110,668	0.6668	76,990
0.9231	110,657	0.6669	76,983
1.0256	110,667	0.6669	76,979
1.1282	110,676	0.6668	76,992
1.2308	110,664	0.6669	76,987
1.3333	110,663	0.6669	76,988
1.4359	110,671	0.6669	76,995
1.5385	110,649	0.6671	76,974
1.6410	110,660	0.6670	76,984
1.7436	110,630	0.6671	76,977
1.8462	110,645	0.6671	76,975
1.9487	110,641	0.6671	76,974
2.0513	110,654	0.6671	76,987
2.1538	110,632	0.6671	76,982
2.2564	110,644	0.6672	76,981
2.3590	110,633	0.6672	76,972
2.4615	110,639	0.6672	76,979
2.5641	110,643	0.6672	76,981
2.6667	110,632	0.6673	76,972
2.7692	110,636	0.6673	76,971
2.8718	110,628	0.6673	76,965
2.9744	110,618	0.6674	76,962
3.0769	110,634	0.6673	76,968
3.1795	110,622	0.6674	76,960
3.2821	110,603	0.6675	76,947
3.3846	110,610	0.6675	76,952
3.4872	110,624	0.6674	76,961
3.5897	110,591	0.6676	76,932
3.6923	110,610	0.6675	76,958
3.7949	110,598	0.6676	76,938
3.8974	110,614	0.6675	76,955
4.0000	110,600	0.6676	76,946

Table D.2 Grid search table for the SVM with dependent variable "Price per square meter"

cp	RMSE	Rsquared	MAE
0.1026	751	0.4426	541
0.2051	748	0.4492	540
0.3077	748	0.4528	538
0.4103	745	0.4545	538
0.5128	744	0.4576	537
0.6154	743	0.4596	536
0.7179	742	0.4611	535
0.8205	741	0.4629	534
0.9231	740	0.4630	534
1.0256	740	0.4645	533
1.1282	739	0.4659	532
1.2308	738	0.4665	532
1.3333	738	0.4683	531
1.4359	737	0.4682	531
1.5385	737	0.4685	531
1.6410	737	0.4689	531
1.7436	736	0.4700	530
1.8462	736	0.4688	531
1.9487	737	0.4710	530
2.0513	736	0.4723	529
2.1538	735	0.4724	529
2.2564	735	0.4721	529
2.3590	735	0.4733	529
2.4615	734	0.4711	530
2.5641	734	0.4721	529
2.6667	734	0.4734	528
2.7692	741	0.4661	533
2.8718	734	0.4730	528
2.9744	733	0.4726	528
3.0769	734	0.4739	527
3.1795	736	0.4701	529
3.2821	733	0.4735	527
3.3846	734	0.4710	528
3.4872	735	0.4740	528
3.5897	734	0.4711	528
3.6923	731	0.4752	527
3.7949	732	0.4744	527
3.8974	734	0.4731	527
4.0000	731	0.4749	526

Table D.3 Grid search for the decision tree with dependent variable Price

cp	RMSE	Rsquared	MAE
0.008830	124,960	0.5734	94,211
0.009254	125,443	0.5700	94,637
0.009774	125,951	0.5665	95,142
0.013044	129,556	0.5410	98,524
0.013138	129,856	0.5389	98,841
0.013462	130,852	0.5319	99,777
0.015775	133,498	0.5129	102,268
0.043221	137,757	0.4807	106,566
0.049736	143,839	0.4342	111,677
0.415932	173,304	0.3830	135,696

Table D.4 Grid search for the decision tree with dependent variable Price per square meter

cp	RMSE	Rsquared	MAE
0.0150527	836	0.3021	630
0.0155844	839	0.2970	632
0.0169746	842	0.2913	635
0.0199884	852	0.2738	643
0.0228127	863	0.2552	652
0.0243246	866	0.2490	655
0.0269416	876	0.2324	664
0.0387914	892	0.2033	679
0.0752393	925	0.1436	705
0.1079225	977	0.0790	744

Table D.5 Grid search of the Random Forest model with dependent variable "Price"

num.trees	mtry	splitrule	min.node.size	RMSE	Rsquared	MAE
500	2	variance	5	228311	0,68759	128775
500	2	variance	10	229403	0,68490	129524
500	2	variance	15	229955	0,68200	129628
500	2	extratrees	5	241451	0,64058	138932
500	2	extratrees	10	242545	0,63546	139741
500	2	extratrees	15	242995	0,63620	139835
500	51	variance	5	169640	0,76900	88237
500	51	variance	10	169775	0,76877	88532
500	51	variance	15	170953	0,76595	89135
500	51	extratrees	5	181815	0,73759	95316
500	51	extratrees	10	183598	0,73278	95948
500	51	extratrees	15	183722	0,73294	96555
500	100	variance	5	171668	0,75733	90535
500	100	variance	10	173649	0,75208	91351
500	100	variance	15	174619	0,74987	91622
500	100	extratrees	5	180568	0,73870	93786
500	100	extratrees	10	181528	0,73611	94475
500	100	extratrees	15	182470	0,73414	94911
1000	2	variance	5	228359	0,68777	128797
1000	2	variance	10	229011	0,68682	129176
1000	2	variance	15	230033	0,68258	129709
1000	2	extratrees	5	241787	0,63896	139122
1000	2	extratrees	10	242558	0,63671	139574
1000	2	extratrees	15	243027	0,63583	139803
1000	51	variance	5	169687	0,76918	88206
1000	51	variance	10	170049	0,76787	88611
1000	51	variance	15	171056	0,76571	89132

1000	51	extratrees	5	181399	0,73892	95183
1000	51	extratrees	10	183185	0,73388	95841
1000	51	extratrees	15	183662	0,73305	96436
1000	100	variance	5	172035	0,75618	90482
1000	100	variance	10	173361	0,75281	91196
1000	100	variance	15	174264	0,75065	91622
1000	100	extratrees	5	180847	0,73797	93747
1000	100	extratrees	10	181285	0,73695	94434
1000	100	extratrees	15	182408	0,73413	94973
1500	2	variance	5	228411	0,68802	128820
1500	2	variance	10	228954	0,68686	129179
1500	2	variance	15	229882	0,68390	129684
1500	2	extratrees	5	241598	0,63984	139075
1500	2	extratrees	10	242417	0,63821	139474
1500	2	extratrees	15	243076	0,63593	139814
1500	51	variance	5	169544	0,76973	88146
1500	51	variance	10	170035	0,76822	88578
1500	51	variance	15	171073	0,76562	89142
1500	51	extratrees	5	181436	0,73862	95143
1500	51	extratrees	10	182745	0,73525	95638
1500	51	extratrees	15	183720	0,73284	96514
1500	100	variance	5	172057	0,75617	90567
1500	100	variance	10	173043	0,75382	91029
1500	100	variance	15	174197	0,75084	91675
1500	100	extratrees	5	180950	0,73776	93749
1500	100	extratrees	10	181181	0,73729	94370
1500	100	extratrees	15	182364	0,73432	94976

Table D.6 Grid search of the Random Forest model with dependent variable "Price per square meter"

num.trees	mtry	splitrule	min.node.size	RMSE	Rsquared	MAE
500	2	variance	5	991,585	0,49040	694,375
500	2	variance	10	993,087	0,48950	695,461
500	2	variance	15	994,673	0,48763	696,188
500	2	extratrees	5	1028,907	0,44213	724,000
500	2	extratrees	10	1030,560	0,44078	724,930
500	2	extratrees	15	1031,666	0,43942	725,626
500	51	variance	5	759,803	0,62950	505,126
500	51	variance	10	762,868	0,62663	507,814
500	51	variance	15	764,915	0,62509	510,468
500	51	extratrees	5	827,199	0,55747	559,760
500	51	extratrees	10	828,591	0,55628	562,253
500	51	extratrees	15	831,353	0,55378	565,120
500	100	variance	5	763,017	0,62372	507,953
500	100	variance	10	765,225	0,62167	510,153
500	100	variance	15	768,074	0,61888	513,275

500	100	extratrees	5	820,708	0,56377	549,792
500	100	extratrees	10	821,688	0,56287	551,553
500	100	extratrees	15	823,364	0,56145	553,907
1000	2	variance	5	991,374	0,49162	694,187
1000	2	variance	10	992,997	0,49027	695,444
1000	2	variance	15	994,809	0,48860	696,274
1000	2	extratrees	5	1028,627	0,44366	723,957
1000	2	extratrees	10	1030,277	0,44200	724,767
1000	2	extratrees	15	1031,567	0,44079	725,522
1000	51	variance	5	759,659	0,62969	505,025
1000	51	variance	10	761,851	0,62782	507,334
1000	51	variance	15	764,881	0,62516	510,288
1000	51	extratrees	5	826,791	0,55797	559,513
1000	51	extratrees	10	828,696	0,55621	562,092
1000	51	extratrees	15	830,922	0,55432	564,821
1000	100	variance	5	762,701	0,62409	507,705
1000	100	variance	10	764,913	0,62203	509,812
1000	100	variance	15	767,711	0,61925	512,793
1000	100	extratrees	5	820,253	0,56432	549,476
1000	100	extratrees	10	821,442	0,56320	551,342
1000	100	extratrees	15	823,197	0,56163	553,755
1500	2	variance	5	991,342	0,49184	694,159
1500	2	variance	10	993,042	0,49043	695,451
1500	2	variance	15	994,580	0,48913	696,221
1500	2	extratrees	5	1028,564	0,44401	723,875
1500	2	extratrees	10	1030,137	0,44236	724,700
1500	2	extratrees	15	1031,732	0,44094	725,668
1500	51	variance	5	759,459	0,62995	504,841
1500	51	variance	10	761,917	0,62775	507,347
1500	51	variance	15	764,621	0,62543	510,175
1500	51	extratrees	5	826,641	0,55817	559,466
1500	51	extratrees	10	828,430	0,55659	561,962
1500	51	extratrees	15	830,699	0,55454	564,716
1500	100	variance	5	762,528	0,62425	507,509
1500	100	variance	10	764,713	0,62221	509,709
1500	100	variance	15	767,551	0,61943	512,669
1500	100	extratrees	5	819,794	0,56477	549,327
1500	100	extratrees	10	821,242	0,56339	551,240
1500	100	extratrees	15	823,124	0,56168	553,745

Appendix E Model Summary SLX

Table E.1 Model summary of the multiple linear regression models and the SLX models on dependent variables "Price and "Price per square meter"

	Model:			
	MLR	**MLR**	**SLX**	**SLX**
	Dependent variable:			
	Price	*Price per square meter*	*Price*	*Price per square meter*
	(1)	*(2)*	*(3)*	*(4)*
Lot.size..m2.	-1.241	-0.069	36.325***	0.212***
	(13.202)	(0.095)	(11.514)	(0.082)
Living.space.size..m2.	1,593.978***	-9.727***	1,613.058***	-9.562***
	(72.903)	(0.523)	(63.658)	(0.452)
Build.year	-222.186***	-2.294***	-11.078	-0.706
	(79.172)	(0.568)	(69.770)	(0.496)
Build.typeNieuwbouw	-85,357.040	-488.152	-102,298.700	-621.184
	(89,421.340)	(641.137)	(77,201.520)	(548.410)
Energy.labelA+	10,352.640	62.721	2,900.614	34.625
	(16,776.510)	(120.285)	(14,539.330)	(103.282)
Energy.labelA++	34,556.280	305.419	30,858.110	261.775
	(31,173.060)	(223.506)	(27,039.490)	(192.078)
Energy.labelA+++	73,920.960*	442.924	92,959.310***	539.584**
	(40,766.950)	(292.293)	(35,208.950)	(250.111)
Energy.labelB	-25,661.770***	-180.375***	-24,188.120***	-171.901***
	(6,312.062)	(45.256)	(5,473.402)	(38.881)
Energy.labelC	-45,576.740***	-300.192***	-43,316.620***	-283.138***
	(6,154.390)	(44.126)	(5,376.417)	(38.192)
Energy.labelD	-38,967.360***	-248.680***	-40,613.930***	-254.109***
	(8,167.968)	(58.563)	(7,101.959)	(50.450)
Energy.labelE	-27,690.030***	-149.539**	-37,852.900***	-220.400***
	(9,611.600)	(68.914)	(8,375.601)	(59.497)
Energy.labelF	-56,565.810***	-351.480***	-64,282.380***	-402.361***
	(10,860.200)	(77.866)	(9,443.868)	(67.086)

Energy.labelG	-76,897.620***	-397.856***	-79,744.250***	-416.829***
	(11,975.660)	(85.864)	(10,398.210)	(73.865)
Energy.labelNiet verplicht	21,208.900	107.262	34,835.090	192.849
	(27,951.040)	(200.405)	(24,236.420)	(172.166)
House.type.1Eengezinswoning	-85,490.250***	-559.321***	-49,271.980***	-307.739***
	(17,506.200)	(125.517)	(15,179.150)	(107.827)
House.type.1Grachtenpand	-114,297.400*	-918.038**	-36,219.640	-364.569
	(60,489.270)	(433.698)	(52,541.780)	(373.237)
House.type.1Herenhuis	22,220.250	72.431	32,758.010*	153.792
	(20,082.510)	(143.988)	(17,429.970)	(123.816)
House.type.1Landhuis	12,493.610	-141.946	37,076.050	60.544
	(32,313.780)	(231.685)	(28,357.430)	(201.440)
House.type.1Villa	60,101.200***	164.352	67,704.190***	214.827*
	(20,054.260)	(143.786)	(17,390.630)	(123.536)
House.type.1Woonboerderij	-62,693.640**	-402.121**	-2,883.257	-7.935
	(26,742.100)	(191.737)	(23,458.760)	(166.642)
House.type.2eindwoning	-34,067.840***	-248.545***	-63,431.720***	-472.201***
	(12,234.890)	(87.722)	(10,673.790)	(75.823)
House.type.2geschakelde 2-onder-1-kapwoning	-12,874.530	-69.787	-19,241.560*	-101.474
	(13,206.670)	(94.690)	(11,469.960)	(81.478)
House.type.2geschakelde woning	-12,160.610	-75.834	-17,289.890*	-107.778
	(10,791.950)	(77.377)	(9,346.902)	(66.397)
House.type.2halfvrijstaande woning	10,654.790	77.677	12,929.120	97.677
	(14,933.910)	(107.074)	(12,958.590)	(92.053)
House.type.2hoekwoning	-31,854.950***	-182.153***	-57,361.790***	-367.869***
	(7,292.999)	(52.290)	(6,362.335)	(45.196)
House.type.2tussenwoning	-44,892.140***	-280.632***	-76,957.540***	-518.169***
	(6,400.223)	(45.889)	(5,657.183)	(40.186)
House.type.2verspringend	-120,612.300**	-818.571**	-125,569.100**	-842.028**
	(57,082.130)	(409.270)	(49,301.940)	(350.222)

House.type.2vrijstaande woning	40,493.450***	262.061***	50,484.070***	328.417***
	(8,081.670)	(57.944)	(7,035.680)	(49.979)
House.type.3bedrijfs- of dienstwoning	66,246.280	497.824	74,243.590**	571.047**
	(43,040.120)	(308.591)	(37,419.330)	(265.813)
House.type.3dijkwoning	-7,056.261	-137.191	-801.726	-96.170
	(27,417.810)	(196.581)	(23,732.180)	(168.584)
House.type.3drive-in woning	-29,132.250	-256.628	-26,564.440	-227.047*
	(22,105.340)	(158.492)	(19,111.560)	(135.761)
House.type.3hofjeswoning	28,771.490	286.881	-16,924.620	-55.852
	(35,801.970)	(256.694)	(31,218.920)	(221.767)
House.type.3kwadrant woning	-37,385.620	-174.136	-5,325.089	45.326
	(51,275.720)	(367.639)	(44,318.790)	(314.824)
House.type.3patiowoning	93,627.790**	592.440**	71,163.440**	412.250*
	(38,513.880)	(276.138)	(33,473.570)	(237.783)
House.type.3semi-bungalow	-47,756.370**	-353.255**		
	(23,706.490)	(169.972)		
House.type.3split-level woning	1,007.014	3.554		
	(24,895.500)	(178.497)		
House.type.3semi.bungalow			-26,452.460	-200.589
			(20,558.940)	(146.043)
House.type.3split.level.woning			15,953.680	108.871
			(21,568.790)	(153.216)
House.type.3waterwoning	-105,758.700*	-736.518*	-78,886.860	-556.149
	(55,589.610)	(398.569)	(48,024.170)	(341.145)
in.woonwijk	8,865.334	61.490	-12,831.630**	-96.366***
	(5,806.714)	(41.633)	(5,094.000)	(36.186)
aan.bosrand	18,787.720	31.051	13,768.920	-4.362
	(15,000.800)	(107.553)	(12,991.730)	(92.288)
aan.drukke.weg	-35,721.960***	-260.934***	-35,774.820***	-262.626***
	(13,129.300)	(94.135)	(11,391.620)	(80.922)

aan.park	18,915.830*	158.833**	8,159.762	76.622
	(10,575.840)	(75.827)	(9,133.837)	(64.883)
aan.rustige.weg	-4,544.340	-31.196	2,267.413	14.322
	(4,171.545)	(29.909)	(3,632.064)	(25.801)
aan.vaarwater	56,882.430***	377.466***	58,522.560***	376.483***
	(13,811.940)	(99.029)	(12,263.550)	(87.116)
aan.water	35,475.380***	214.703***	25,868.080***	130.732**
	(8,957.305)	(64.222)	(7,808.427)	(55.468)
bedrijventerrein	45,631.210	307.175	-7,555.071	-101.501
	(42,833.740)	(307.111)	(36,990.830)	(262.769)
beschutte.ligging	20,948.080***	129.941***	11,082.880**	51.308
	(6,437.055)	(46.153)	(5,624.205)	(39.952)
buiten.bebouwde.kom	22,295.550	183.268	42,977.470***	349.017***
	(16,680.260)	(119.595)	(14,576.310)	(103.545)
in.bosrijke.omgeving	21,754.220**	200.611***	24,553.540***	228.021***
	(8,548.633)	(61.292)	(7,647.795)	(54.327)
in.centrum	-10,984.930	-3.905	-2,578.501	55.668
	(6,883.700)	(49.355)	(6,001.876)	(42.635)
landelijk.gelegen	23,270.010**	234.477***	31,073.460***	283.908***
	(11,672.050)	(83.687)	(10,153.410)	(72.126)
open.ligging	6,210.520	19.619	-1,141.509	-40.051
	(10,192.870)	(73.081)	(8,857.767)	(62.922)
vrij.uitzicht	9,758.861*	73.663*	8,731.271*	68.359**
	(5,466.392)	(39.193)	(4,738.135)	(33.658)
zeezicht	-9,559.835	-94.318	282,559.700**	1,542.185
	(123,341.100)	(884.336)	(135,274.700)	(960.940)
no.of.rooms	-5,002.580*	-41.867**	-943.900	-10.285
	(2,782.615)	(19.951)	(2,458.403)	(17.464)
no.of.bedrooms	4,023.161	1.412	-632.660	-33.299
	(3,284.691)	(23.551)	(2,875.956)	(20.430)
no.of.bathrooms	41,363.820***	319.200***	33,078.050***	258.936***

	(5,539.169)	(39.715)	(4,803.733)	(34.124)
no.of.sep.toilets	45,365.240***	279.617***	32,786.600***	185.349***
	(5,509.083)	(39.499)	(4,797.944)	(34.083)
no.of.floors	32,857.840***	160.530***	11,990.210***	3.938
	(5,257.172)	(37.693)	(4,585.305)	(32.572)
zolder	-9,018.368	-112.976**	-5,919.960	-89.402**
	(6,567.579)	(47.089)	(5,722.871)	(40.653)
kelder	-14,286.640	-103.069	1,476.154	4.612
	(8,853.846)	(63.481)	(7,813.932)	(55.507)
vliering	32,288.050***	211.334***	18,474.120***	109.529**
	(7,446.942)	(53.393)	(6,464.272)	(45.920)
Achtertuin	-2,485.583	-87.283	-8,599.452	-127.508**
	(10,122.190)	(72.575)	(8,802.893)	(62.532)
Voortuin	9,012.563	68.995*	16,925.020***	130.088***
	(5,697.737)	(40.852)	(4,999.591)	(35.515)
Zijtuin	30,964.520***	243.352***	21,132.640***	171.413***
	(5,760.374)	(41.301)	(5,004.209)	(35.548)
Plaats	-77,021.210**	-337.689	-37,285.330	-66.657
	(30,839.910)	(221.117)	(26,710.980)	(189.745)
Zonneterras	14,171.210	96.481	14,937.340*	104.384*
	(9,856.971)	(70.673)	(8,636.246)	(61.349)
Tuin.rondom	61,828.450***	313.250***	66,216.950***	350.379***
	(11,412.640)	(81.827)	(9,861.356)	(70.051)
patio.atrium	26,724.140	253.221*	7,419.560	113.972
	(20,952.380)	(150.225)	(18,223.320)	(129.452)
Roof.typeLessenaardak	1,389.212	-32.035	2,338.566	-9.249
	(15,856.880)	(113.691)	(13,820.810)	(98.178)
Roof.typeMansarde dak	19,138.970	154.747	-8,980.582	-43.286
	(16,500.660)	(118.307)	(14,320.640)	(101.728)
Roof.typePlat dak	6,561.727	24.999	-6,283.435	-56.749
	(15,521.570)	(111.287)	(13,462.490)	(95.632)

Roof.typeSamengesteld dak	28,802.270**	174.737*	11,694.360	64.420
	(12,937.420)	(92.759)	(11,255.050)	(79.952)
Roof.typeSchilddak	1,774.232	50.267	8,628.728	105.461
	(15,215.350)	(109.092)	(13,209.080)	(93.832)
Roof.typeTentdak	25,467.420	180.360	13,397.530	86.779
	(24,317.110)	(174.350)	(21,007.750)	(149.231)
Roof.typeZadeldak	1,899.090	24.871	-1,088.462	14.145
	(11,233.060)	(80.539)	(9,789.680)	(69.542)
roof.cover.pannen	-29,441.290***	-199.558***	-18,204.810**	-111.289*
	(9,208.696)	(66.025)	(8,006.912)	(56.878)
roof.cover.asbest	-14,689.080	-70.829	-59,124.330	-383.390
	(56,737.840)	(406.801)	(48,900.600)	(347.371)
roof.cover.bitumineuze.dakbedekking	-18,111.030***	-108.376**	-13,949.520**	-78.866**
	(6,502.576)	(46.622)	(5,656.790)	(40.184)
roof.cover.kunststof	-47,034.000	-238.382	-44,197.980*	-209.705
	(29,718.610)	(213.078)	(25,751.150)	(182.926)
roof.cover.leisteen	-26,936.650	-213.408	-31,533.240	-243.449
	(37,215.380)	(266.828)	(32,139.830)	(228.309)
roof.cover.metaal	-106,638.300***	-254.971	-117,406.200***	-340.330
	(37,361.360)	(267.875)	(32,183.790)	(228.621)
roof.cover.overig	-54,242.850**	-405.430**	-55,787.570***	-405.797***
	(22,363.350)	(160.342)	(19,499.700)	(138.518)
roof.cover.riet	10,383.400	-32.123	27,503.390	133.948
	(25,124.020)	(180.135)	(22,326.240)	(158.597)
lag.Lot.size..m2.			-134.246***	-0.948***
			(27.241)	(0.194)
lag.Living.space.size..m2.			180.300	0.320
			(158.191)	(1.124)
lag.Build.year			-49.339	-0.393*
			(31.189)	(0.222)

lag.Build.typeNieuwbouw	32,697.670	807.270
	(201,027.200)	(1,428.021)
lag.Energy.labelA.	-80,911.450*	-881.072***
	(42,337.960)	(300.753)
lag.Energy.labelA..	28,078.020	352.758
	(82,230.280)	(584.132)
lag.Energy.labelA...	-190,758.500*	-1,557.486**
	(108,565.700)	(771.210)
lag.Energy.labelB	20,210.090	174.471*
	(13,926.030)	(98.925)
lag.Energy.labelC	25,706.030**	138.224
	(12,637.170)	(89.770)
lag.Energy.labelD	113,282.400***	849.092***
	(17,010.430)	(120.836)
lag.Energy.labelE	102,490.300***	673.538***
	(19,816.430)	(140.768)
lag.Energy.labelF	75,339.380***	477.643***
	(20,915.730)	(148.577)
lag.Energy.labelG	117,500.100***	937.616***
	(22,767.300)	(161.730)
lag.Energy.labelNiet.verplicht	76,694.650	568.946
	(50,056.930)	(355.585)
lag.House.type.1Eengezinswoning	-137,024.600***	-968.439***
	(39,559.070)	(281.013)
lag.House.type.1Grachtenpand	-858,325.700***	-7,375.158***
	(146,102.500)	(1,037.856)
lag.House.type.1Herenhuis	-68,946.980	-579.231*
	(44,071.730)	(313.069)
lag.House.type.1Landhuis	-202,682.300***	-1,223.436**
	(72,601.820)	(515.736)

lag.House.type.1Villa	-36,439.700	-192.173
	(45,776.950)	(325.182)
lag.House.type.1Woonboerderij	-188,765.900***	-1,375.039***
	(56,340.420)	(400.221)
lag.House.type.2eindwoning	45,703.250	331.559
	(30,239.260)	(214.808)
lag.House.type.2gesch.2.onder.1.kapwoning	13,756.380	36.994
	(28,909.470)	(205.362)
lag.House.type.2geschakelde.woning	-87,609.320***	-706.346***
	(26,714.280)	(189.768)
lag.House.type.2halfvrijstaande.woning	-83,450.460**	-539.403**
	(38,356.330)	(272.469)
lag.House.type.2hoekwoning	35,705.930**	315.011**
	(17,577.130)	(124.861)
lag.House.type.2tussenwoning	76,684.730***	555.188***
	(14,490.620)	(102.936)
lag.House.type.2verspringend	-349,374.300*	-3,210.443**
	(181,287.400)	(1,287.797)
lag.House.type.2vrijstaande.woning	-99,360.100***	-633.501***
	(17,603.220)	(125.047)
lag.House.type.3bedrijfs..of.dienstwoning	-72,049.710	-666.013
	(75,051.250)	(533.135)
lag.House.type.3dijkwoning	149,427.400**	716.526
	(74,372.180)	(528.312)
lag.House.type.3drive.in.woning	-93,684.470*	-874.211**
	(54,437.370)	(386.702)
lag.House.type.3hofjeswoning	14,352.070	110.139
	(69,792.600)	(495.780)
lag.House.type.3kwadrant.woning	-1,061.189	192.964
	(170,981.500)	(1,214.587)

lag.House.type.3patiowoning	401,645.500***	2,420.075***
	(99,413.930)	(706.199)
lag.House.type.3semi.bungalow	-149,573.300**	-1,209.815***
	(63,473.410)	(450.891)
lag.House.type.3split.level.woning	-44,972.930	-191.633
	(65,424.840)	(464.753)
lag.House.type.3waterwoning	-422,166.800**	-3,105.633**
	(183,106.100)	(1,300.716)
lag.in.woonwijk	36,725.950***	251.406***
	(12,740.260)	(90.502)
lag.aan.bosrand	-92,815.310***	-794.337***
	(35,346.650)	(251.089)
lag.aan.drukke.weg	-73,397.690***	-595.438***
	(28,192.370)	(200.268)
lag.aan.park	84,801.090***	532.539***
	(28,320.310)	(201.177)
lag.aan.rustige.weg	8,611.509	80.838
	(9,294.310)	(66.023)
lag.aan.vaarwater	-15,524.130	-71.864
	(29,425.610)	(209.028)
lag.aan.water	71,571.430***	500.726***
	(22,186.480)	(157.604)
lag.bedrijventerrein	-10,413.420	-173.812
	(81,906.010)	(581.829)
lag.beschutte.ligging	25,216.030*	226.324**
	(13,677.620)	(97.161)
lag.buiten.bebouwde.kom	32,128.490	131.220
	(32,499.680)	(230.865)
lag.in.bosrijke.omgeving	114,723.700***	832.107***
	(17,289.140)	(122.815)

lag.in.centrum	22,073.390	181.320*
	(14,764.750)	(104.883)
lag.landelijk.gelegen	-45,461.390*	-190.051
	(24,579.260)	(174.602)
lag.open.ligging	-27,283.180	-89.492
	(21,019.440)	(149.314)
lag.vrij.uitzicht	-6,671.789	14.174
	(12,949.890)	(91.991)
lag.zeezicht	87,433.880	-108.687
	(116,213.400)	(825.535)
lag.no.of.rooms	6,829.066	54.430
	(5,739.087)	(40.768)
lag.no.of.bedrooms	-21,001.690***	-123.454**
	(7,525.886)	(53.461)
lag.no.of.bathrooms	76,724.500***	510.679***
	(11,294.090)	(80.229)
lag.no.of.sep.toilets	80,721.050***	599.106***
	(12,845.750)	(91.251)
lag.no.of.floors	67,398.130***	610.322***
	(12,487.530)	(88.707)
lag.zolder	-50,744.210***	-291.137***
	(14,922.860)	(106.006)
lag.kelder	-87,936.290***	-644.199***
	(17,219.900)	(122.324)
lag.vliering	72,106.300***	510.667***
	(19,290.020)	(137.029)
lag.Achtertuin	-96,021.480***	-661.695***
	(22,221.260)	(157.851)
lag.Voortuin	40,533.570***	137.053
	(12,732.560)	(90.447)
lag.Zijtuin	63,173.570***	479.962***

	(13,170.080)	(93.555)
lag.Plaats	-251,169.100***	-2,342.249***
	(86,678.340)	(615.730)
lag.Zonneterras	-31,613.710	-186.039
	(20,703.960)	(147.073)
lag.Tuin.rondom	-40,736.390	-419.927**
	(24,932.210)	(177.109)
lag.patio.atrium	7,944.284	100.980
	(49,984.660)	(355.072)
lag.Roof.typeLessenaardak	63,937.390*	193.113
	(37,470.420)	(266.176)
lag.Roof.typeMansarde.dak	54,112.310	244.126
	(37,756.970)	(268.211)
lag.Roof.typePlat.dak	58,103.970	197.029
	(35,568.300)	(252.664)
lag.Roof.typeSamengesteld.dak	44,014.080	87.469
	(29,586.020)	(210.168)
lag.Roof.typeSchilddak	-34,354.860	-422.176
	(36,490.220)	(259.213)
lag.Roof.typeTentdak	149,358.700**	1,007.416**
	(58,576.430)	(416.105)
lag.Roof.typeZadeldak	29,945.950	76.830
	(25,410.430)	(180.506)
lag.roof.cover.pannen	-69,212.260***	-583.086***
	(20,775.720)	(147.583)
lag.roof.cover.asbest	130,656.200	1,084.474
	(184,190.600)	(1,308.419)
lag.roof.cover.bitumineuze.dakbedekking	-70,108.520***	-432.019***
	(15,629.530)	(111.026)
lag.roof.cover.kunststof	-135,970.100***	-1,144.322***

			(52,573.140)	(373.459)
lag.roof.cover.leisteen			-289,732.100**	-2,306.654***
			(115,929.800)	(823.521)
lag.roof.cover.metaal			-153,699.200	-1,691.915**
			(108,974.200)	(774.111)
lag.roof.cover.overig			-155,717.500***	-1,219.138***
			(47,601.960)	(338.146)
lag.roof.cover.riet			5,413.807	124.957
			(47,379.840)	(336.568)
Constant	641,722.900***	9,338.452***	207,579.200	6,115.280***
	(160,254.100)	(1,148.996)	(141,776.300)	(1,007.125)
Observations	4,057	4,057	4,057	4,057
R2	0.597	0.241	0.711	0.465
Adjusted R2	0.589	0.225	0.699	0.442
Residual Std. Error	122,464.50	878.051	104,842.20	744.759
	(df = 3973)	(df = 3973)	(df = 3890)	(df = 3890)
F Statistic	71.011***	15.164***	57.666***	20.372***
	(df = 83; 3973)	(df = 83; 3973)	(df = 166; 3890)	(df = 166; 3890)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table E.2 Impact measures of the SLX model with dependent variable "Price"

	Impact measures (SlX, estimable, n-k):					
	Direct		Indirect		Total	
Lot.size..m2.	36	**	-134	***	-98	***
Living.space.size..m2.	1,613	***	180		1,793	***
Build.year	-11		-49		-60	
Build.typeNieuwbouw	-102,299		32,698		-69,601	
Energy.labelA+	2,901		-80,911		-78,011	
Energy.labelA++	30,858		28,078		58,936	
Energy.labelA+++	92,959	**	-190,759		-97,799	
Energy.labelB	-24,188	***	20,210		-3,978	
Energy.labelC	-43,317	***	25,706	*	-17,611	
Energy.labelD	-40,614	***	113,282	***	72,668	***
Energy.labelE	-37,853	***	102,490	***	64,637	**
Energy.labelF	-64,282	***	75,339	***	11,057	
Energy.labelG	-79,744	***	117,500	***	37,756	
Energy.labelNiet verplicht	34,835		76,695		111,530	*
House.type.1Eengezinswoning	-49,272	**	-137,025	***	-186,297	***
House.type.1Grachtenpand	-36,220		-858,326	***	-894,545	***
House.type.1Herenhuis	32,758		-68,947		-36,189	
House.type.1Landhuis	37,076		-202,682	**	-165,606	*
House.type.1Villa	67,704	***	-36,440		31,264	
House.type.1Woonboerderij	-2,883		-188,766	***	-191,649	**
House.type.2eindwoning	-63,432	***	45,703		-17,728	
House.type.2gesch.2-onder-1-kapwoning	-19,242		13,756		-5,485	
House.type.2geschakelde woning	-17,290		-87,609	**	-104,899	***
House.type.2halfvrijstaande woning	12,929		-83,450	*	-70,521	
House.type.2hoekwoning	-57,362	***	35,706	*	-21,656	
House.type.2tussenwoning	-76,958	***	76,685	***	-273	
House.type.2verspringend	-125,569	*	-349,374		-474,943	*
House.type.2vrijstaande woning	50,484	***	-99,360	***	-48,876	**
House.type.3bedrijfs- of dienstwoning	74,244	*	-72,050		2,194	
House.type.3dijkwoning	-802		149,427	*	148,626	
House.type.3drive-in woning	-26,564		-93,684		-120,249	*
House.type.3hofjeswoning	-16,925		14,352		-2,573	
House.type.3kwadrant woning	-5,325		-1,061		-6,386	
House.type.3patiowoning	71,163	*	401,646	***	472,809	***
House.type.3semi-bungalow	-26,452		-149,573	*	-176,026	**
House.type.3split-level woning	15,954		-44,973		-29,019	
House.type.3waterwoning	-78,887		-422,167	*	-501,054	**
in.woonwijk	-12,832	*	36,726	**	23,894	
aan.bosrand	13,769		-92,815	**	-79,046	*
aan.drukke.weg	-35,775	**	-73,398	**	-109,173	***
aan.park	8,160		84,801	**	92,961	**

aan.rustige.weg	2,267		8,612		10,879	
aan.vaarwater	58,523	***	-15,524		42,998	
aan.water	25,868	***	71,571	**	97,440	***
bedrijventerrein	-7,555		-10,413		-17,968	
beschutte.ligging	11,083	*	25,216		36,299	*
buiten.bebouwde.kom	42,977	**	32,128		75,106	*
in.bosrijke.omgeving	24,554	**	114,724	***	139,277	***
in.centrum	-2,579		22,073		19,495	
landelijk.gelegen	31,073	**	-45,461		-14,388	
open.ligging	-1,142		-27,283		-28,425	
vrij.uitzicht	8,731		-6,672		2,059	
zeezicht	282,560	*	87,434		369,994	*
no.of.rooms	-944		6,829		5,885	
no.of.bedrooms	-633		-21,002	**	-21,634	**
no.of.bathrooms	33,078	***	76,724	***	109,803	***
no.of.sep.toilets	32,787	***	80,721	***	113,508	***
no.of.floors	11,990	**	67,398	***	79,388	***
zolder	-5,920		-50,744	***	-56,664	***
kelder	1,476		-87,936	***	-86,460	***
vliering	18,474	**	72,106	***	90,580	***
Achtertuin	-8,599		-96,021	***	-104,621	***
Voortuin	16,925	***	40,534	**	57,459	***
Zijtuin	21,133	***	63,174	***	84,306	***
Plaats	-37,285		-251,169	**	-288,454	**
Zonneterras	14,937		-31,614		-16,676	
Tuin.rondom	66,217	***	-40,736		25,481	
patio.atrium	7,420		7,944		15,364	
Roof.typeLessenaardak	2,339		63,937		66,276	
Roof.typeMansarde dak	-8,981		54,112		45,132	
Roof.typePlat dak	-6,283		58,104		51,821	
Roof.typeSamengesteld dak	11,694		44,014		55,708	
Roof.typeSchilddak	8,629		-34,355		-25,726	
Roof.typeTentdak	13,398		149,359	*	162,756	**
Roof.typeZadeldak	-1,088		29,946		28,857	
roof.cover.pannen	-18,205	*	-69,212	***	-87,417	***
roof.cover.asbest	-59,124		130,656		71,532	
roof.cover.bitumineuze.dakbedekking	-13,950	*	-70,109	***	-84,058	***
roof.cover.kunststof	-44,198		-135,970	**	-180,168	**
roof.cover.leisteen	-31,533		-289,732	*	-321,265	**
roof.cover.metaal	-117,406	***	-153,699		-271,105	*
roof.cover.overig	-55,788	**	-155,717	**	-211,505	***
roof.cover.riet	27,503		5,414		32,917	

Table E.3 Impact measures of the SLX model with dependent variable "Price per square meter"

	Impact measures (SlX, estimable, n-k):					
	Direct		Indirect		Total	
Lot.size..m2.	0.2118	**	-0.9477	***	-0.7359	***
Living.space.size..m2.	-9.5618	***	0.3197		-9.2421	***
Build.year	-0.7062		-0.3932		-1.0994	*
Build.typeNieuwbouw	-621.1838		807.2700		186.0863	
Energy.labelA+	34.6249		-881.0723	**	-846.4474	**
Energy.labelA++	261.7754		352.7583		614.5337	
Energy.labelA+++	539.5844	*	-1,557.4862	*	-1,017.9018	
Energy.labelB	-171.9007	***	174.4707		2.5701	
Energy.labelC	-283.1384	***	138.2244		-144.9140	
Energy.labelD	-254.1092	***	849.0920	***	594.9828	***
Energy.labelE	-220.3997	***	673.5379	***	453.1382	**
Energy.labelF	-402.3606	***	477.6431	**	75.2825	
Energy.labelG	-416.8289	***	937.6155	***	520.7866	**
Energy.labelNiet verplicht	192.8492		568.9461		761.7953	*
House.type.1Eengezinswoning	-307.7388	**	-968.4392	***	-1,276.1780	***
House.type.1Grachtenpand	-364.5694		-7,375.1584	***	-7,739.7279	***
House.type.1Herenhuis	153.7917		-579.2308		-425.4391	
House.type.1Landhuis	60.5438		-1,223.4361	*	-1,162.8922	*
House.type.1Villa	214.8265		-192.1728		22.6537	
House.type.1Woonboerderij	-7.9351		-1,375.0395	***	-1,382.9745	**
House.type.2eindwoning	-472.2009	***	331.5586		-140.6423	
House.type.2gesch. 2-onder-1-kapwoning	-101.4735		36.9935		-64.4800	
House.type.2geschakelde woning	-107.7778		-706.3456	***	-814.1234	***
House.type.2halfvrijstaande woning	97.6772		-539.4031	*	-441.7259	
House.type.2hoekwoning	-367.8690	***	315.0115	*	-52.8575	
House.type.2tussenwoning	-518.1694	***	555.1878	***	37.0183	
House.type.2verspringend	-842.0281	*	-3,210.4428	*	-4,052.4709	**
House.type.2vrijstaande woning	328.4167	***	-633.5011	***	-305.0844	*
House.type.3bedrijfs- of dienstwoning	571.0467	*	-666.0133		-94.9665	
House.type.3dijkwoning	-96.1701		716.5261		620.3560	
House.type.3drive-in woning	-227.0473		-874.2110	*	-1,101.2583	**
House.type.3hofjeswoning	-55.8521		110.1387		54.2866	
House.type.3kwadrant woning	45.3258		192.9644		238.2902	
House.type.3patiowoning	412.2499		2,420.0752	***	2,832.3251	***
House.type.3semi-bungalow	-200.5894		-1,209.8151	**	-1,410.4045	**
House.type.3split-level woning	108.8712		-191.6327		-82.7615	
House.type.3waterwoning	-556.1486		-3,105.6329	*	-3,661.7815	**
in.woonwijk	-96.3659	**	251.4059	**	155.0400	
aan.bosrand	-4.3620		-794.3371	**	-798.6991	**
aan.drukke.weg	-262.6258	**	-595.4383	**	-858.0641	***
aan.park	76.6219		532.5390	**	609.1609	**

aan.rustige.weg	14.3216		80.8383		95.1598	
aan.vaarwater	376.4831	***	-71.8643		304.6188	
aan.water	130.7322	*	500.7263	**	631.4585	***
bedrijventerrein	-101.5014		-173.8123		-275.3138	
beschutte.ligging	51.3080		226.3235	*	277.6316	**
buiten.bebouwde.kom	349.0174	***	131.2200		480.2374	
in.bosrijke.omgeving	228.0208	***	832.1069	***	1,060.1277	***
in.centrum	55.6680		181.3202		236.9881	*
landelijk.gelegen	283.9078	***	-190.0514		93.8563	
open.ligging	-40.0510		-89.4916		-129.5426	
vrij.uitzicht	68.3592	*	14.1743		82.5335	
zeezicht	1,542.1849		-108.6866		1,433.4982	
no.of.rooms	-10.2850		54.4299		44.1450	
no.of.bedrooms	-33.2993		-123.4541	*	-156.7533	**
no.of.bathrooms	258.9362	***	510.6794	***	769.6155	***
no.of.sep.toilets	185.3489	***	599.1057	***	784.4546	***
no.of.floors	3.9385		610.3223	***	614.2607	***
zolder	-89.4018	*	-291.1367	**	-380.5385	***
kelder	4.6117		-644.1992	***	-639.5876	***
vliering	109.5292	*	510.6669	***	620.1961	***
Achtertuin	-127.5079	*	-661.6953	***	-789.2032	***
Voortuin	130.0876	***	137.0528		267.1404	**
Zijtuin	171.4129	***	479.9623	***	651.3752	***
Plaats	-66.6573		-2,342.2487	***	-2,408.9059	***
Zonneterras	104.3839		-186.0389		-81.6550	
Tuin.rondom	350.3785	***	-419.9274	*	-69.5489	
patio.atrium	113.9724		100.9804		214.9528	
Roof.typeLessenaardak	-9.2493		193.1132		183.8639	
Roof.typeMansarde dak	-43.2865		244.1261		200.8396	
Roof.typePlat dak	-56.7486		197.0294		140.2808	
Roof.typeSamengesteld dak	64.4203		87.4685		151.8888	
Roof.typeSchilddak	105.4614		-422.1758		-316.7143	
Roof.typeTentdak	86.7788		1,007.4158	*	1,094.1946	*
Roof.typeZadeldak	14.1455		76.8300		90.9755	
roof.cover.pannen	-111.2885		-583.0858	***	-694.3744	***
roof.cover.asbest	-383.3898		1,084.4735		701.0837	
roof.cover.bitumineuze.dakbedekking	-78.8665	*	-432.0189	***	-510.8854	***
roof.cover.kunststof	-209.7047		-1,144.3221	**	-1,354.0269	**
roof.cover.leisteen	-243.4490		-2,306.6536	**	-2,550.1026	**
roof.cover.metaal	-340.3297		-1,691.9150	*	-2,032.2446	*
roof.cover.overig	-405.7974	**	-1,219.1380	***	-1,624.9354	***
roof.cover.riet	133.9482		124.9571		258.9053	

References

Anh Le My, P. (2016). *Comparison of Support Vector Regression and Neural Networks.* University of Minnesota Duluth.

Centraal Bureau voor de Statistiek. (2022, 3 11). *Centraal Bureau voor de Statistiek*. Retrieved from Wat dreef de stijgende prijzen van koopwoningen sinds medio 2013?: https://www.cbs.nl/nl-nl/nieuws/2022/10/wat-dreef-de-stijgende-prijzen-van-koopwoningen-sinds-medio-2013-

De Hypotheker. (2021). *De voor- en nadelen van een hogere WOZ-waarde*. Opgehaald van https://www.hypotheker.nl/actueel/nieuwsberichten/2021/de-voor-en-nadelen-van-een-hogere-woz-waarde/

De Nederlandsche Bank. (sd). *Actuele economische vraagstukken over de woningmarkt*. Opgehaald van https://www.dnb.nl/actuele-economische-vraagstukken/woningmarkt/

Dubin, R. A. (1998). Predicting House Prices Using Multiple Listings Data. *Journal of Real Estate Finance and Economics, 17*(1), 35-59.

Funda Index. (2023, Juni). Opgehaald van Funda: https://www.funda.nl/funda-index/juni-2023/

Hoe wordt de vraagprijs van mijn woning bepaald? (2022, November 30). Opgehaald van Funda: https://www.funda.nl/meer-weten/verkopen/hoe-wordt-de-vraagprijs-van-mijn-woning-bepaald/

Imran, Zaman, U., Waqar, M., & Zaman, A. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Soft Computing and Machine Intelligence Journal*, 11-23.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021, August 4). An Introduction to Statistical Learing with Applications in R.

Langenberg, H., & Jonkers, W. (2022). Achtergrond bij de huizenprijsstijgingen vanaf 2013. *De Nederlandse economie*.

Liu, X. (2013). Spatial and Temporal Dependence in House Price Prediction. *The Journal of Real Estate Finance and Economics, 47*, 341-369. doi:10.1007/s11146-011-9359-3

Manasa, J., Narahari, N. S., & Gupta, R. (2020). Machine Learning based Predicting House Prices. *IEEE*.

Molnar, C. (2023, March 2). Interpretable Machine Learning; A Guide for Making Black Box Models Explainable.

Ollongren, K. (2021, July 5). *Staat van de Woningmarkt 2021 [Letter of government]*. Opgehaald van https://open.overheid.nl/repository/ronl-60430922-d774-4f5b-9914-4f3597f8d4f7/1/pdf/aanbieding-rapport-staat-van-de-woningmarkt-2021.pdf

Park, B., & Bae, J. (2015). Using machine learning algorithms for housing price prediction:. *Elsevier*, 2928-2934.

Rijksoverheid. (sd). *Waardering onroerende zaken (WOZ)*. Opgehaald van https://www.rijksoverheid.nl/onderwerpen/waardering-onroerende-zaken-woz/vraag-en-antwoord/woz-waarde-bepalen#:~:text=Gemeenten%20bepalen%20de%20waarde%20van,Dit%20is%20de%20waardepeildatum

Risse, M., & Kern, M. (2016). Forecasting house-price growth in the Euro area with dynamic model averaging. *North American Journal of Economics and Finance*, 70-85.

Steven, B. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Research, 32*(2).

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2022). Housing Price Prediction via Improved Machine Learning Techniques. *Elsevier*, 433-442.

Vega, S. H., & Elhorst, J. P. (2015). The SLX Model. *Journal of Regional Science*(3), pp. 339-363. doi:10.1111/jors.12188

Zhang, Y., Zhang, D., & Miller, E. J. (2021). Spatial Autoregressive Analysis and Modeling of Housing Prices in City of Toronto. *American Society of Civil Engineers*. doi:10.1061/(ASCE)UP.1943-5444.0000651

Ziets, J., Zietz, E., & Simans, G. (2008). Determinants of House Prices: A Quantile Regression Approach. *The Journal of Real Estate Finance and Economics*. doi:10.1007/s11146-007-9053-7