From *Funda*-mental Data to Smart Investments: How Machine Learning is Revolutionizing House Price Estimations in the Netherlands

Dimitrios Paschalidis
Professional Certificate in Data Science | IBM
M.Sc. Accounting & Finance | Erasmus School of Economics
B.Sc. Banking & Finance | University of Piraeus

November 30th, 2023

ABSTRACT

The housing market in the Netherlands has seen a significant increase in property prices over the last decade, with the House Price Index (HPI) increasing by 85.6% since 2015. This study examines the potential of Machine Learning in estimating house prices based on various attributes of a property, leveraging data from one of the most visited housing sites in the Netherlands, *funda.nl*. The study uses five algorithms - linear, ridge, lasso, random forest, and XGBoost regressions - to predict house prices. The Linear Regression model, trained on 124,012 observations, is found to be the most effective, with an expected deviation from the "true price" of ±EUR 36,000 or ±8.9%. The findings of this study have been applied to the development of PriceWise, a tool that helps users estimate a property's rightful price based on size, number of bedrooms, house type, location, energy label, insulation, and other house attributes.

Keywords: *Funda.nl; Dutch Housing Market; Machine Learning; House Price Prediction; PriceWise*

1. Introduction

The last decade was marked by the all-time low interest rates which led property prices soaring and to an unprecedented housing market in the Netherlands. According to the Dutch statistics institution, since 2015, the House Price Index (HPI) has increased by 77.8% (as of Q3 2023).[1] During the same time span, the EU19 HPI increased only by 42.3%.[2] The Dutch household rents have also followed this upward trend but lesser in magnitude (18.8%).[3] Additionally, it is common for Dutch banks to grant mortgages of 100% LTV for owner-occupied houses, whereas in most EU countries this percentage ranges from 70% to 90%. As a consequence, many residents seek house ownership because not only they might pay less in monthly mortgage installments than in rent payments, but they also "build" their equity in the house with 100% debt financing. But do buyers know whether they pay the "right" price for a house, especially in a bull market? This study examines the potential of machine learning in estimating house prices, by leveraging 132,025 transactions that have been advertised on *funda.nl* – one of the most visited housing sites in the Netherlands, with 4.8 million unique users per month.[4]

When someone wants to sell or buy a house, they usually compare it with known transactions from their friendly environment or network, search for similar properties on housing sites, or leave it up to the real estate agent to find the best price in the current market. Comparable transactions valuation is not only common in finance and M&A – it resembles the housing market, as there is no spot market to determine prices and the goods are not homogenous – but also in our everyday lives. Consider, for example, the time you wanted to sell a used item of yours. You probably went through eBay or Facebook's market place, had a look at similar (if not the same) products, and tried to place your product on the relative valuation scale. This involves more art than science, since the method can be proved quite subjective. What house sellers often want is just to "get a feeling" about how much their property is worth and if that feeling (i.e., the price) is satisfactory, "we have a deal"! However, this approach might not always be the optimal, since [a] the comparable transactions valuation can be easily influenced by non-fundamental factors (such as seasonal trends or market shortages), [b] data on transactions is limited for a retail seller, and [c] it is hard to define the "comparables". More to the science side, one can hire a real estate appraiser – in the Netherlands the cost ranges from €500 to €1,000 – who is a trained and experienced professional in the real estate market. On the other side of the spectrum, buyers usually weight the mortgage payments against their monthly liquidity needs to come up with the optimal loan amount (i.e., price) and then search for houses based on this price cap. For those who seek an own-occupied house, if the property is good enough, some overbidding is common to win the deal. In this case, the satisfaction of a "dream house" might overcome any suboptimal cost. For investing, however, overpaying can highly undermine ROI.

The purpose of this paper is to exploit publicly available data from *funda.nl* and specify an effective machine learning model that estimates residential property prices based on their fundamental characteristics (e.g., square meters, number of bedrooms, area, insulation, etc.). This approach dominates the comparable transactions by mitigating the subjectivity problem, addressing the lack of data for comparable real estate, and "decoding" other confounding factors that are not visible outside the statistics realm. The offspring of this study does not merely remain on script but the application of the model is directly accessible on the App Store and Google Play, under the name PriceWise. The rest of the paper is organized as follows: section 2 presents an overview of the data collection, the variables, and covers the data pre-processing and sample selection, section 3 goes over the descriptive statistics and correlations of the sample, section 4 describes the modeling methods and presents the evaluation results, section 5 concludes.

2. Data Collection & Pre-Processing

I built a web-scraper that browsed through *funda.nl* for approximately 15 consecutive days and fetched 228,642 raw observations for home properties around the Netherlands – 2,453 different municipalities, cities, towns, and villages sold between 2017 and 2023. I extracted 27 attributes, among them being the: city, area of the city, sale date, square meters, number of bedrooms, house type, year of construction, building-related outdoor space, plot size, number of floors, energy label, kind(s) of insulation, kind(s) of heating, kind(s) of garage(s), kind(s) of parking space(s), storage, contribution to the homeowner

[1] Centraal Bureau voor de Statistiek, Kadaster (https://www.cbs.nl/en-gb/figures/detail/83906eng)
[2] Eurostat (https://ec.europa.eu/eurostat/databrowser/view/prc_hpi_q/default/table?lang=en)
[3] Statista (https://www.statista.com/statistics/577189/housing-rent-increase-in-the-netherlands/)
[4] About funda *Accessed on November 30, 2023*

association (VvE – if applicable), and price in euros. Some landing pages' HTML layout showed inconsistencies, and with the web-scraper performing on fixed parameters, it fetched some garbage data and therefore the dataset had to be cleansed; either with manual inspection of the features' unique values or using regex – a library that specializes in parsing text. This data cleansing process reduced the sample size by 23,208 observations (10.15%). Further, I require every municipality, city, town, or village (from now on, the term municipality includes all of these) and sub-area to have at least 10 observations. This reduces the noise from geographical locations where real estate is traded less frequently. At this point of the sample selection process, the majority of sold properties lies between 2021 and 2023, with only 29 transactions during 2017-2020. Thus, years 2017, 2018, 2019, and 2020 are dropped, as the number of transactions is not only minor but also deemed "old" and may not reflect the current market prices.

The extracted dataset includes ten numeric variables, out of which six are retained: price, square meters, number of bedrooms, building-related outdoor space, external storage size, and VvE. To elaborate on the dropped numeric features: the number of rooms is excluded while the number of bedrooms is retained because the former shows high correlation with the latter, while the number of bedrooms offers higher R^2, and larger and more statistically significant coefficient in a univariate OLS. The number of floors and the étage are also dropped because the former is conflated (not always clear whether it refers to the number of floors of a single house or the total number of floors in a condominium) and the latter has too many missing values – a preliminary inspection of this variable showed that missing values do not always correspond to ground floor (i.e., étage = 0). Lastly, plot size is also removed due to its high number of missing values.

Handling outliers is the most enduring and pervasive methodological challenge. According to Aguinis et al. (2013), there are 14 mutually exclusive outlier definitions, 39 outlier identification techniques, and 20 different ways of handling outliers. One of the most common techniques of handling outliers, especially with large samples, is to adjust the observed values to fall within plausible limits and remove any values outside of those limits from the dataset (i.e., truncation). Due to the objective of this study, – to provide a predictive model to the average house-buyer – I apply subjective judgment on the limits, accompanied by evidence from standard preliminary analysis plots (boxplots, histograms, and Q-Q plot). For the variables except price, I also plot the average price per grouped value of each variable. This way, it becomes clearer at which x-axis points the X-Y relations become distorted. Table 1 provides the details of the truncations.

TABLE 1: TRUNCATIONS

Variable	**Range**		**Observation loss**	
	Original	*Truncated*	*Absolute*	*Percentiles*
1. Price (€)	[60,000 – 8,750,000]	[100,000 – 1,000,000]	5,215	0.04% – 97.92%
2. Square meters (m^2)	[10 – 950]	[30 – 250]	1,420	0.15% – 99.38%
3. Number of bedrooms	[1 – 10]	[1 – 6]	1,260	0.00% – 99.80%
4. Building-related outdoor space (m^2)	[0 – 1,750]	[0 – 70]	511	0.00% – 99.72%
5. External storage (m^2)	[0 – 982]	[0 – 60]	1,444	0.00% – 99.20%
6. VvE (€)	[0 – 1,309]	[0 – 300]	1,598	0.00% – 99.11%

Note: Percentiles indicate the inclusive cutoff range points of the remaining distribution at each truncation step.

Price (prc)

The original price distribution is highly skewed to the right. Houses (or apartments) worth above EUR 1 million are not in the list of an every-day buyer and therefore the upper limit is set to this amount. On the other hand, houses worth less than EUR 100,000 usually need some kind of repairment or bare other hidden costs such as poor insulation. Despite the upper cap of EUR 1 million, a substantial amount of price outliers remains and therefore I proceed with the log-transformation of this variable.

Square meters (sqm)

Very small apartments usually come with shared toiles and interestingly enough, the smallest houses in the pre-truncated sample have relatively high VvE costs.[5] Conversely, approximately after 250m^2, the variance of the mean price increases, therefore

[5] Square meters ≤ 20: average VvE = €94 p.m. | Square meters > 20 & ≤ 40: average VvE = €74 p.m. | Square meters > 40: average VvE = €35 p.m.

indicating a distortion in the relation. In this regard, "normal" homes are deemed above 30m^2 and below 250 m^2 – also log-transformed.

Number of bedrooms (bed)
Most of the houses (or apartments) in the sample have 2 to 4 bedrooms. The inspection of the mean price across grouped observations (on number of bedrooms) revealed that it is increasing until (and including) the group with 7 bedrooms, decreases by 11% for the group with 8, before rebouncing again. This inconsistency may indicate a distortion between the relation of the two variables and thus, the threshold of 6 is chosen as the "healthy" one.

Building-related outdoor space (out)
Outdoor space is presumably related only with houses; not apartments. Therefore, this variable shows 46% zeros, while the rest 54% show a value between 1 and 1,750 (pre-truncation). The inspection of the mean price across grouped observations (grouped per 1 sq.m.) shows a linear trend with relatively stable variance up until 70 m^2, thus this threshold is chosen for the upper bound. Due to outliers, I log-transformed this variable as well; the log distribution excluding zeros resembles a normal. To preserve those observations with none outdoor space (0), I add the unity before the log-transformation.

External storage (ext)
External storage refers to space that is available for storing things that would not fit within in-house storerooms (e.g., tools, equipment, and building material), or garage space. One would expect that the greater the external storage area, the higher the price of the property, but this assumption is true up until certain square meters. An average household has 1-2 cars and limited stuff that needs to be stored outside the main building. The data show that 60m^2 is the "sweet point" after which noise is introduced into a property's price. As with the building-related outdoor space, zeros are dominating this variable (30%) and outliers are present even after truncation, I apply log-transformation after adding the unity.

VvE (vve)
When buying a property in the Netherlands in a building with other apartments, one will automatically become member of the *Vereniging van Eigenaren* ("Home-Owners Association"). The associations are responsible for common parts of the building, such as halls, roof, and the lift. For the maintenance of these parts, the VvE may provide for monthly contributions from the home owners. In the original sample, these contributions are as high as €1,309. Usually, more luxurious – and therefore expensive properties – come with a higher need of common spaces maintenance, leading to higher VvE. Even though the majority of the properties (77%) in the sample has no VvE contributions, closer data inspection shows that this is true up to roughly €300. This variable is not log-transformed, because the distribution, excluding zeros, shows no outliers.

Boxplots, histograms, and Q-Q plots before and after truncation and log-transformations are available in Appendix A.

After the truncations, I re-apply the minimum of 10 observations per municipality and sub-area, keep observations with *year of construction* no later than 2025,[6] and drop missing values. At this point, the sample consists of 132,025 observations on which the initial models' trainings are commenced. Table 2 presents the condensed sample selection steps, along with the absolute and relative sample reductions. For the detailed sample selection table, see Appendix B.

TABLE 2: SAMPLE SELECTION

	Number of observations	Δ	%Δ
Raw data	228,642	-	-
Data cleansing	205,434	-23,208	-10.15%
Data filtering*	176,053	-29,381	-14.30%
Year of construction ≤ 2025	176,046	-7	0.00%
Drop missing values	132,025	-44,021	-25.01%

**Judgmental truncation of the numerical variables and deselection of observations that do not pass certain categorical thresholds.*

[6] Properties can be sold before they have been constructed.

3. Descriptive Statistics & Correlations

Table 3 presents the descriptive statistics for the numeric features; before log-transformations. By observing the ranges (min & max), it is clear that the variables remain bounded at the given judgmental thresholds. The average property in the sample is sold for slightly below EUR 400,000, has an area of 111 m^2 and 3 bedrooms. A typical house is sold for EUR 369,500, while at least 75% of the properties are sold for EUR 475,000 or less, and have an area of 131m^2 or less. The *building-related outdoor space* and *external storage* mean (6.4 and 7.8) and median (2 and 6) differ significantly, underlining the need of the natural logarithm transformation. Less than 25% have a *building-related outdoor space* greater than 9m^2 and *external storage* greater than 10m^2. At least 75% of *VvE* show zero values – nevertheless, this variable is kept as is, since the distribution of the non-zero values lacks outliers and resembles the normal distribution. For the categorical variables, see Appendix C.

TABLE 3: DESCRIPTIVE STATISTICS

Variable	N	Mean	St. Dev.	Min	25th pctl	Median	75th pctl	Max
prc	132,025	399,443	156,000	100,000	289,000	369,500	475,000	1,000,000
sqm	132,025	111	34	30	86	110	131	250
bed	132,025	3.2	1.1	1	2	3	4	6
out	132,025	6.4	9.6	0	0	2	9	70
ext	132,025	7.8	8.8	0	0	6	10	60
vve	132,025	32.1	65.8	0	0	0	0	300

Table shows number of observations (N), mean (Mean), standard deviation (St. Dev.), minimum (Min), 25th percentile (25th pctl), median (Median), 75th percentile (75th pctl), and maximum (Max) of the numeric variables used in this study. Transaction price (*prc*) and home-owners association monthly contribution (*vve*) are in euros; square meters (*sqm*), building-related outdoor space (*out*), and external storage (*ext*) are in m^2; number of bedrooms (*bed*) is a natural number. The log-transformed variables are converted back into the original measures for better data intuition.

Table 4 reports Pearson correlations between the numeric features. Most of the linear relations are intuitive, while the signs and the magnitudes are as expected. *Price* is significantly correlated with every other variable – substantially stronger with *square meters* – and the signs are as mostly expected. *Square meters* shows also highly significant correlations with the rest of the features, but the intuition behind the negative relation with *VvE* is not clear *prima facie*; probably the underlying rationale is that bigger properties are houses (not apartments), therefore no common spaces with other properties and no VvE contributions. |*Number of bedrooms* shows a negative correlation with the *building-related outdoor space* and a positive with e*xternal storage*. The negative linear relation with *VvE* is not obvious at first glance, but the same rationale applies, as with the *square meters*. *Building-related* o*utdoor space* shows a negligible correlation with *external storage*, while *external storage* presents a weak negative correlation with *VvE*.

TABLE 4: CORRELATION MATRIX

	prc	*sqm*	*bed*	*out*	*ext*	*vve*
prc	1					
sqm	0.610***	1				
bed	0.387***	0.735***	1			
out	0.100***	0.038***	-0.096***	1		
ext	-0.045***	0.039***	0.095***	-0.070***	1	
vve	-0.153***	-0.395***	-0.479***	0.255***	-0.015***	1

*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

Table reports Pearson correlations between the numeric variables used in the study. *prc* is the natural logarithm of the transaction price in euros; *sqm* is the is the natural logarithm of the area of a property in m^2; *bed* is the number of bedrooms; *out* is the is the natural logarithm +1 of the building-related outdoor space in m^2; *ext* is the is the natural logarithm +1 of the area of the external storage in m^2; and *vve* is the home-owners association (*Vereniging van Eigenaren*) monthly contribution in euros.

4. Model Training & Evaluation

Five different models were chosen to predict house prices, namely: linear, ridge, lasso, random forest, and XGBoost regression. For each model, a 10-fold cross validation was applied, meaning that the data-set is split into 9:1 train-test parts and the evaluation metrics of each fold is retained. After the completion of all 10 train-test iterations, the averages and standard deviations of the evaluation metrics are taken.

Linear Regression is a statistical technique used to establish a relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fit line that describes the relationship between the variables by minimizing the sum of squared errors between the predicted values and the actual values of the dependent variable. The line is described by an equation in the form of $y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$, where y is the dependent variable, x_1, x_2, ..., x_n are the independent variables, and b_0, b_1, b_2, ..., b_n are the intercept and slopes of the regression line, respectively. Once the coefficients have been estimated, the model can be used to make predictions on new data.

Ridge Regression is a type of regularized linear regression model that is used to prevent overfitting in a regression analysis. It works by adding a penalty term to the ordinary least squares (OLS) objective function, which shrinks the coefficients towards zero and reduces their variance. This penalty term is proportional to the square of the magnitude of the coefficients and is controlled by a hyperparameter called the regularization parameter or lambda (λ). The effectiveness of ridge regression depends on the appropriate choice of the regularization parameter λ. Lambda values from 0.001 to 100 have been used in order to “fine tune” this model.

Lasso Regression works by adding a penalty term to the standard linear regression objective function that shrinks the coefficients of the less important features to zero. In other words, it "selects" the most important features and reduces the impact of the others. This is achieved by adding a L1 regularization term to the objective function of the standard linear regression model. This regularization term penalizes the sum of the absolute values of the coefficients of the model, which results in some coefficients becoming exactly zero. This makes lasso regression useful for feature selection, as it effectively removes the least important features from the model.

Random Forest Regression is an algorithm that uses an ensemble of decision trees for regression tasks. In random forest regression, a large number of decision trees are created, each one being trained on a random subset of the training data, and using a random subset of the features. Each tree in the forest makes a prediction for the target variable, and the final prediction is obtained by averaging or taking the median of the predictions made by all the trees in the forest. The main advantage of random forest regression is its ability to handle both linear and nonlinear relationships between the features and the target variable, its ability to handle large datasets with many features, and its robustness to outliers and missing data. A hyper-parameter tuning was performed for the number of variables to be randomly sampled at each split and the number of trees contained in the ensemble.

XGBoost Regression (Extreme Gradient Boosting Regression) is a type of ensemble learning method that combines multiple decision trees to make predictions. The algorithm builds a series of decision trees, where each subsequent tree is trained to correct the errors of the previous tree. The final prediction is then made by combining the predictions of all trees. The XGBoost algorithm uses a gradient boosting framework, which means that it optimizes an objective function by iteratively adding decision trees. The objective function measures the difference between the predicted and actual values and seeks to minimize it. The gradient descent algorithm is used to find the best values for the model parameters. One of the key advantages of XGBoost is that it is highly scalable and can handle large datasets with a large number of features. It also allows for feature importance ranking, which can help identify the most important variables in a given dataset. A hyper-parameter tuning was performed for the maximum depth of a tree and the step size of each boosting step (eta).

The models are evaluated with a series of metrics, namely: Adjusted R-squared (Adj. R^2), Root Mean Standard Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Relative Squared Error (RSE).

Adjusted R^2 (or coefficient of determination) evaluates the proportion of variance in the dependent variable that can be explained by the independent variables. In other words, it represents how well the data fits the regression model, also known as goodness of fit. Adjusted R^2 values range between 0 and 1. The calculation is detailed in Equation (1).

$$Adj.R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \quad (1)$$

RMSE measures the average difference between the predicted values and the actual values in a dataset, and it is calculated by taking the square root of the mean of the squared differences between the predicted and actual values. RMSE penalizes larger errors more than smaller ones, due to the squaring operation. This means that if a model has a few very large errors, it will have a higher RMSE compared to a model with many small errors, even if the overall error rate is the same. The calculation is detailed in Equation (2).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

MAE measures the average absolute difference between the predicted values and the actual values. It has several advantages over other regression metrics, such as mean squared error (MSE), because it is less sensitive to outliers. Since MAE takes the absolute values of the errors, it gives equal weight to all errors, whether they are positive or negative. This makes it a more robust measure of error than MSE, which penalizes large errors more heavily. MAE is also easy to interpret because it represents the average magnitude of the errors in the units of the target variable. The calculation is detailed in Equation (3).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (3)$$

MAPE measures the average relative deviation of the actual values from the predicted values. Its value ranges from 0 to infinity. The lower the MAPE value, the better the model's performance. A value of 0 indicates a perfect forecast, while a value of infinity indicates that the actual value is 0, which is not possible. MAPE is useful for comparing the performance of different forecasting models, as it provides a standardized measure of error. However, MAPE can be sensitive to extreme values, and it cannot be used when the actual values are zero. Also, it assumes that all observations have equal importance, which may not always be the case. The calculation is detailed in Equation (4).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

RSE is a normalized version of the Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values of the target variable. RSE measures the proportion of the variance of the target variable that is explained by the regression model. A low RSE indicates that the model is a good fit for the data and can explain a significant amount of the variation in the target variable. A high RSE, on the other hand, indicates that the model is not a good fit and is not able to explain much of the variation in the target variable. It is important to note that RSE can take on values greater than one if the MSE is larger than the variance of the target variable. The calculation is detailed in Equation (5).

$$RSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (5)$$

Where:

- R^2 is the sample R-squared [$= 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$];
- N is the total sample size;
- p is the number of independent variables;
- y_i is the actual value of the i_{th} observation;
- $\hat{y}_i$ is the predicted value of the i_{th} observation; and
- $\bar{y}$ is the sample mean

Table 5 - Panel A reports the average metrics per 10-fold cross validation per model; with their standard deviations. The Linear Regression and XGBoost Regression models illustrate the best metrics, with a slight advantage upon the linear regression. To further mitigate the effect of outliers in the data, I use Cook's distance on the Linear Regression and assess the influence that data points have on all regression coefficients as a whole (Aguinis et al., 2013). In a regression analysis, Cook's distance measures how much the estimated regression coefficients change when a single observation is omitted from the dataset. Specifically, it is the ratio of the change in the estimated regression coefficients to their standard errors, when the observation is included and when it is excluded. This way, the data points with Cook's D above the threshold $\{4/N - p - 1\}$[7] are removed from the sample and the Linear and XGBoost models are retrained and re-evaluated. The reduced sample size comes down to 124,012.[8] Table 5 - Panel B presents the final model evaluation results. For the variable importance, see Appendix E.

The outcome variable is the natural logarithm of the price in euros – due to outliers. However, metrics denominated in log terms are not much intuitive and hard to interpret. For example, the Mean Absolute Error (MAE) in log terms is 0.0889; a unit-free number. The rough interpretation would be that it shows the typical size of percentage error on the original scale. Note that it is almost identical to the Mean Absolute Percentage Error (MAPE), thus not conveying any new information. Therefore, to make the metric figures more intuitive, I turn the estimated log prices and the actual log prices into their original (euro) scale, before proceeding with the metrics' calculation.

Overall, the features in the Linear Regression explain more than 91% of the price variance. In absolute terms, the model is expected to "miss", on average, the actual price by EUR 36,000, or 8.9%. For example, if a property's observed price is EUR 400,000, the model is expected to predict a price between EUR 436,000 and EUR 364,000. Furthermore, the inspection of the error distribution shows that the error in log-terms follows a normal distribution ~**N(0, 0.1)** – Figure 1. Anti-log transforming the estimated and observed price values before calculating the errors gives the distribution in Figure 2, and further taking the absolute error values, the distribution becomes as in Figure 3 (this figure actually shows the MAE distribution). The median absolute error is EUR 26,482 (50% of the errors are below this number), and there is a 5% probability for the model to make an absolute error of more than EUR 103,310 and a 1% probability to make an absolute error of more than EUR 163,684.

TABLE 5: EVALUATION METRICS

Panel A: Cook's D Not Applied

	Adj. R^2		RMSE		MAE		MAPE		RSE	
Model	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Linear	0.882	0.004	59,107	429	41,480	163	0.103	0.0003	0.144	0.002
Ridge	0.531	0.007	107,815	817	78,798	654	0.202	0.0018	0.479	0.003
Lasso	0.520	0.007	107,910	816	78,802	662	0.202	0.0018	0.480	0.004
Rand Forest	0.552	0.006	91,180	714	64,997	511	0.166	0.0014	0.342	0.004
XGBoost	0.873	0.005	60,812	636	41,634	244	0.103	0.0003	0.152	0.002

Panel B: Cook's D Applied

	Adj. R^2		RMSE		MAE		MAPE		RSE	
Model	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Linear	0.910	0.007	49,761	400	35,992	273	0.089	0.0007	0.107	0.003
XGBoost	0.895	0.006	53,139	729	37,261	397	0.092	0.0008	0.122	0.004

Note: Models in Panel A are trained on 132,025 observations, while models in Panel B are trained on 124,012 observations.

[7] N is the sample size; p is the number of predictors

[8] For the descriptive statistics of the reduced sample, see Appendix D.

Figure 1:
Error in log terms

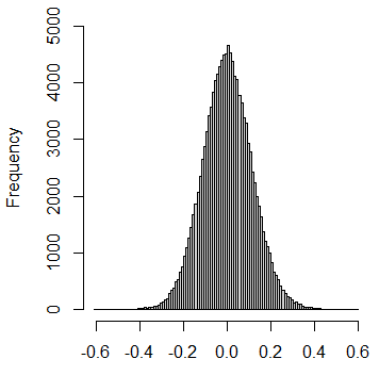


Figure 2:
Error in original scale

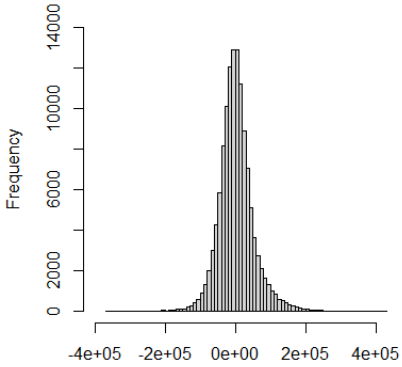
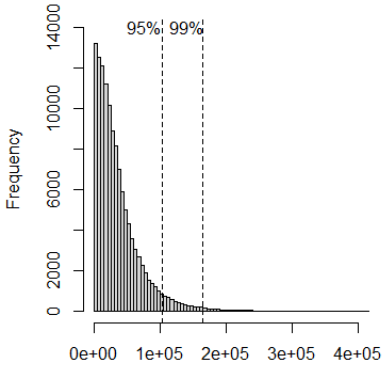


Figure 3:
Abs. error in original scale



5. Conclusion

In this study, five machine learning algorithms are used to estimate property prices based on attributes scraped from *funda.nl*. The models employed were the linear regression, ridge regression, lasso regression, random forest regression, and XGBoost regression. Linear regression, which was trained on a dataset of 124,012 observations, was found to be the most effective algorithm in estimating the price of a property, with an expected error of EUR ±36,000, or ±8.9%. The median absolute error is EUR 26,482, while there is a 5% probability for the absolute error to exceed EUR 103,310 and a 1% to exceed EUR 163,684. This specific model's parameters are used on the back-end of the PriceWise app (available on the App Store and Google Play), to help users estimate a property's "rightful" price based on size, house type, location, and many other house attributes.

With the vast amounts of data available online, accurately predicting the prices of properties has never been more attainable. Machine learning algorithms can analyze massive amounts of data in seconds, identify patterns, and make predictions with impressive accuracy. This technological advancement has enabled prospective home-owners and investors to make better-informed decisions, ultimately leading to smarter investments and more profitable outcomes. All in all, the integration of machine learning in the real estate industry has opened up new possibilities for investors, agents, and buyers alike. With the power of Machine Learning at our fingertips, the ability to make precise predictions about the cost of houses has become easier than ever, which can assist in making more intelligent choices when buying or investing in real estate.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301.

Acknowledgements

The idea of this working paper came after reading the: Xu, K., & Nguyen, H. (2022). Predicting housing prices and analyzing real estate market in the Chicago suburbs using Machine Learning. *arXiv preprint arXiv:2210.06261*

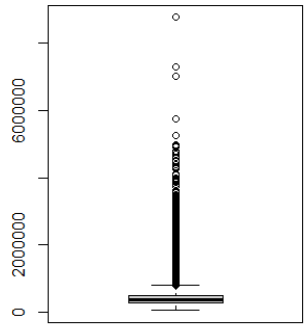
Author's note

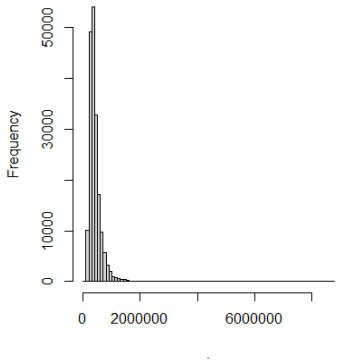
PriceWise holds the model parameters on the Google Cloud. Every month, the sample is expanded, the model is retrained, and the parameters and evaluation metrics are updated. If the reader wants to be using the latest version, please consider updating the PriceWise app in regular intervals.

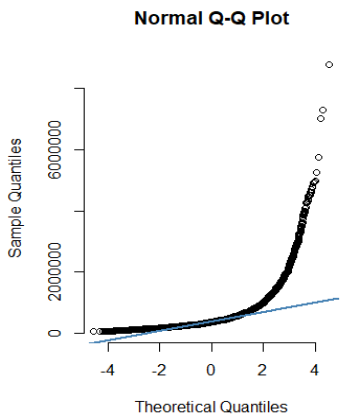
Appendix A

-Price-

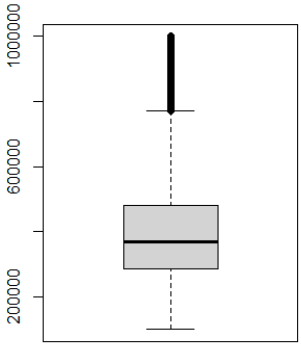
Original range

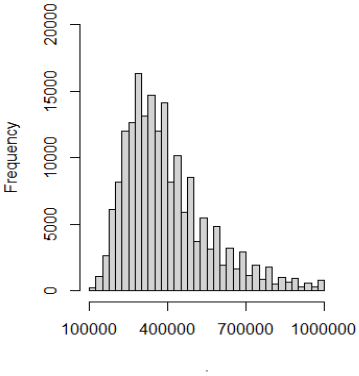


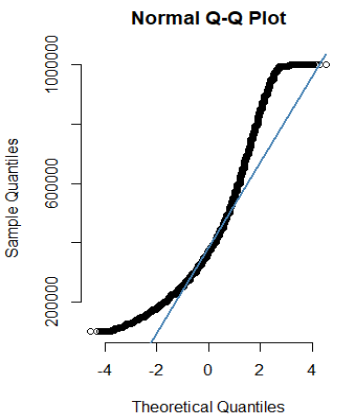




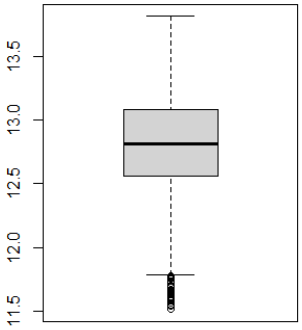
Truncated range

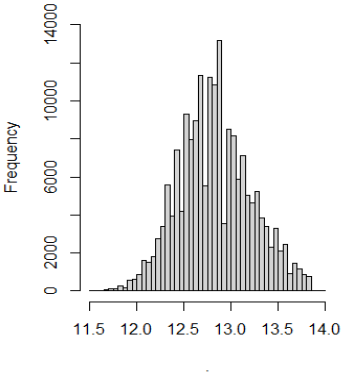


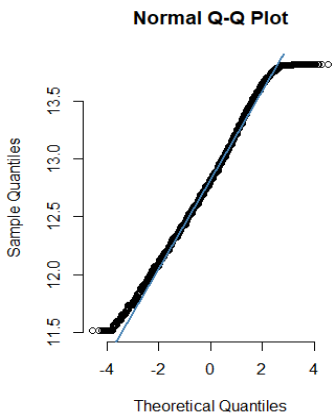




Log-transformed

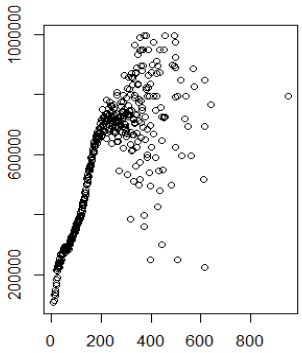


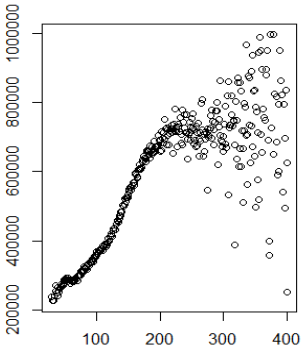


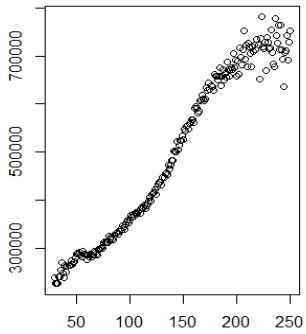


-Square Meters-

Grouped by square meters and summarized on average price



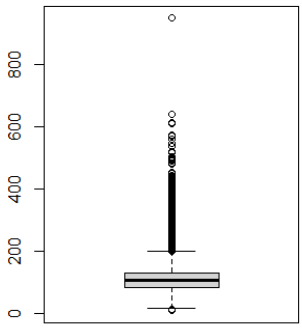


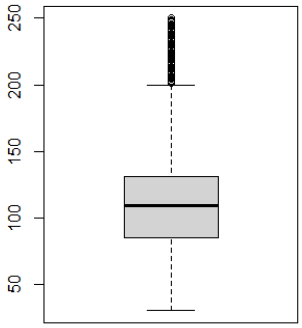


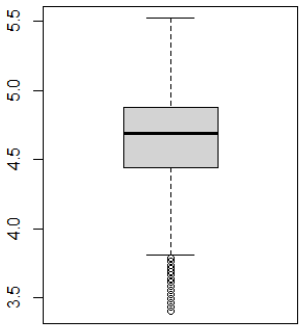
Original range

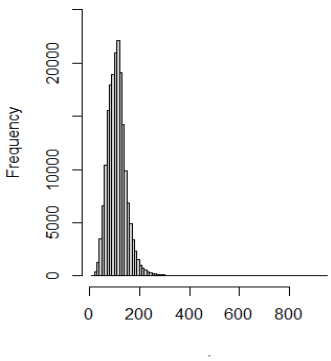
Truncated range

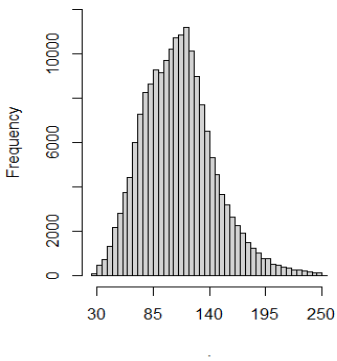
Log-transformed

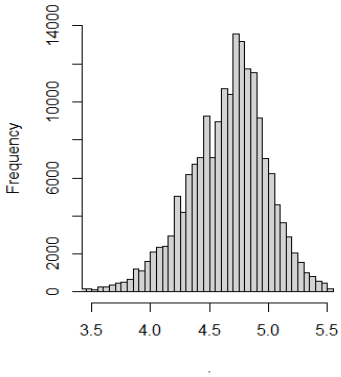


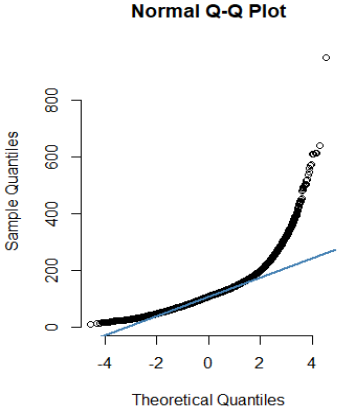


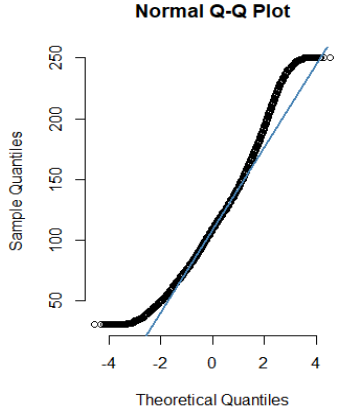


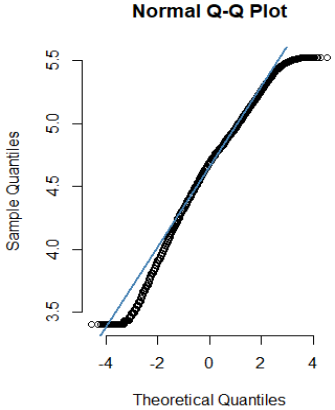






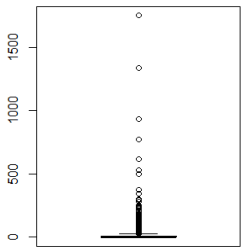




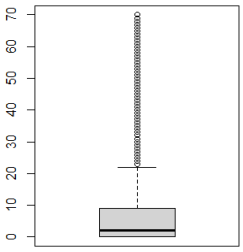


-Building-related outdoor space-

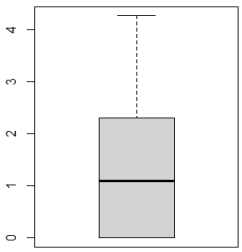
Original range



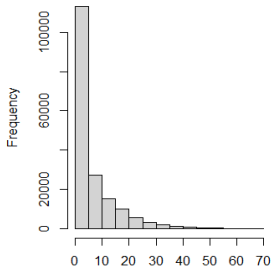
Truncated range



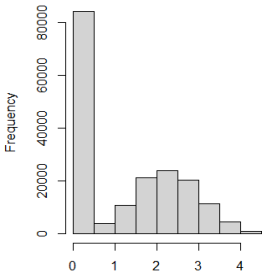
Log-transformed



Truncated range

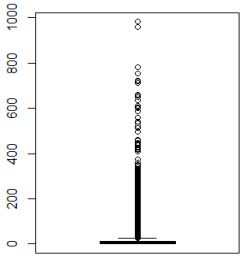


Log-transformed

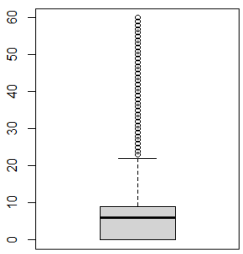


-External storage-

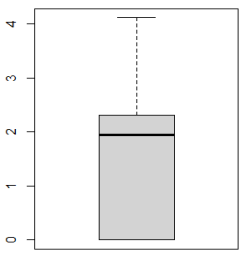
Original range



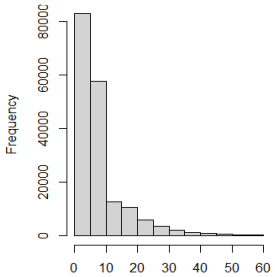
Truncated range



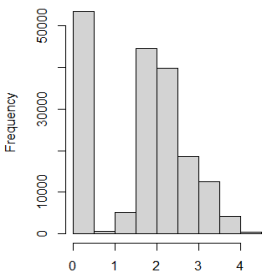
Log-transformed



Truncated range

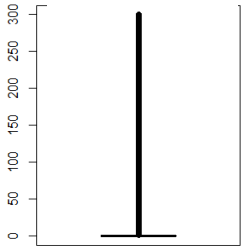


Log-transformed

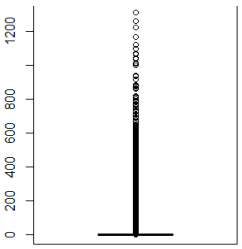


-VvE-

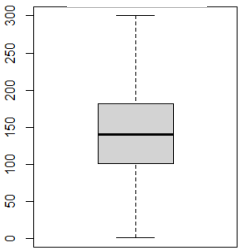
Original range



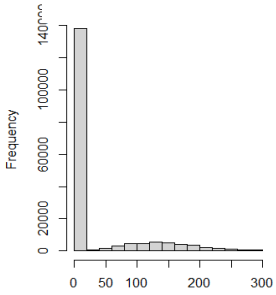
Truncated range



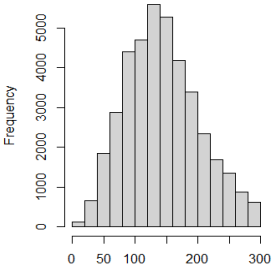
Non-zero



Truncated range



Non-zero



-Number of bedrooms-

Number of bedrooms	N	Average price (€)
1	13,142	289,781
2	39,696	361,689
3	55,822	385,207
4	54,949	429,169
5	15,429	518,872
6	2,270	590,575
7	293	629,595
8	48	573,281
9	16	670,937
10	1	750,000
n.a.	902	355,777

Appendix B

SAMPLE SELECTION – DETAILED TABLE

	Number of observations	Δ	%Δ
Raw data	228,642	-	-
Data cleansing	205,434	-23,208	-10.15%
Areas with at least 10 observations	189,463	-15,971	-7.77%
Municipalities with at least 10 observations	189,232	-231	-0.12%
Sales in years: 2021, 2022, 2023	189,203	-29	-0.02%
Price range: €100,000 to €1,000,000	183,988	-5,215	-2.76%
Square meters range: $30m^2$ to $250m^2$	182,568	-1,420	-0.77%
Number of bedrooms range: 1 to 6	181,308	-1,260	-0.69%
Outside related space range: $0m^2$ to $70m^2$	180,797	-511	-0.28%
External storage range: $0m^2$ to $60m^2$	179,353	-1,444	-0.80%
VvE range: €0 to €330	177,755	-1,598	-0.89%
Areas with at least 10 observations	176,082	-1,673	-0.94%
Municipalities with at least 10 observations	176,053	-29	-0.02%
Year of construction ≤ 2025	176,046	-7	0.00%
Drop missing values	132,025	-44,021	-25.01%
Drop outliers identified with Cook's distance	124,012	-8,013	-6.07%

This table shows the number of observations, the absolute and relative changes of the sample size at stages of data cleansing, filtering, truncations, and dropping missing values and outliers.

Appendix C

CATEGORICAL VARIABLES

Municipality			Area			House Type		
Amsterdam	6,586	5.0%	*Centrum*	1,124	0.9%	*Eengezinswoning*	87,573	66.3%
Rotterdam	4,720	3.6%	*Binnenstad*	454	0.3%	*Portiekflat*	11,634	8.8%
The Hague	3,579	2.7%	*Bomenbuurt*	341	0.3%	*Bovenwoning*	8,058	6.1%
Utrecht	3,336	2.5%	*Zeeheldenbuurt*	330	0.2%	*Galerijflat*	7,126	5.4%
Groningen	2,169	1.6%	*Bloemenbuurt*	313	0.2%	*Benedenwoning*	5,176	3.9%
Eindhoven	2,112	1.6%	*Staatsliedenbuurt*	302	0.2%	*Herenhuis*	3,210	2.4%
(Other)	109,523	83.0%	*(Other)*	129,161	97.8%	*(Other)*	9,248	7.0%

Construction Type			Energy Label			Storage		
Bestaande bouw	130,204	98.6%	*C*	36,433	27.6%	*Vrijstaande houten berging*	29,608	22.4%
Nieuwbouw	1,821	1.4%	*A*	32,753	24.8%	*No Storage*	29,010	22.0%
			B	20,933	15.9%	*Vrijstaande stenen berging*	27,833	21.1%
			D	16,135	12.2%	*Inpandig*	18,243	13.8%
			E	10,681	8.1%	*Aangebouwde stenen berging*	11,748	8.9%
			F	6,032	4.6%	*Box*	11,017	8.3%
			(Other)	9,058	6.9%	*(Other)*	4,566	3.5%

Note: An observation can take only one value of the above variables. The rest categorical variables, which multiple values can be assigned per property, are: insulation, heating, location, garage type, parking options, and additional house areas (attic and cellar).

Appendix D

DESCRIPTIVE STATISTICS – REDUCED SAMPLE

Variable	N	Mean	St. Dev.	Min	25th pctl	Median	75th pctl	Max
prc	124,012	397,4300	152,250	100,000	289,000	369,000	475,000	1,000,000
sqm	124,012	110	34	30	86	109	130	250
bed	124,012	3.2	1.1	1	2	3	4	6
out	124,012	6.3	9.5	0	0	2	9	70
ext	124,012	7.6	8.5	0	0	6	9	60
vve	124,012	32.8	66.2	0	0	0	0	300

Table shows number of observations (N), mean (Mean), standard deviation (St. Dev.), minimum (Min), 25th percentile (25th pctl), median (Median), 75th percentile (75th pctl), and maximum (Max) of the numeric variables used in this study. Transaction price (*price*) and home-owners association monthly contribution (*vve*) are in euros; square meters (*sqm*), building-related outdoor space (*out*), and external storage (*ext*) are in m^2; number of bedrooms (*bed*) is a natural number. This sample is reduced due to the elimination of outliers identified with Cook's distance (threshold: 4/N – p – 1, where N is sample size and p the number of predictors). The log-transformed variables are converted back into the original measures for better data intuition.

Appendix E

SHAP VALUES