Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning for Property Valuation During a Housing Boom: Predictive Evaluation, Extrapolation Bias and Explainability in the Netherlands

Wouter Westerhof
University of Amsterdam
Rafiazka Millanida Hilman
r.m.hilman@uva.nl

University of Amsterdam https://orcid.org/0009-0003-9252-7765

Research Article

Keywords: machine learning, property valuation, explainable AI, model evaluation, urban analytics, Netherlands

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11846787/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Machine Learning for Property Valuation During a Housing Boom: Predictive Evaluation, Extrapolation Bias and Explainability in the Netherlands

Abstract

Machine learning is increasingly used to predict urban quantities such as property values, yet models are often evaluated with random splits and without simple benchmarks, which can hide how they behave when the predicted quantity trends over time. This study evaluates tree-based models for predicting official Dutch property valuations (WOZ values) during the 2014–2024 housing boom, combining a nationwide panel of 4,714 neighborhoods with 79,216 address-level valuations in Utrecht. Random forest and XGBoost are compared with naive persistence, drift and linear benchmarks under temporal hold-out, expanding-window and random validation, and interpreted with Shapley additive explanations (SHAP). XGBoost explains 93% of the variance in 2024 neighborhood values, against 86% for random forest and 76% for linear regression, yet carrying the previous valuation forward is more accurate than any attribute-based model. Over the boom, both ensembles underpredict by up to 13%, and their bias tracks market growth almost perfectly ($r = -0.95$). We trace this to the inability of trees to extrapolate the market level, which SHAP exposes as a frozen contribution of the year variable; lagged valuations reduce the bias but overshoot when growth slows. At the property level, valuation history provides most of the accuracy ($R^2 = 0.98$). Income, energy consumption, population density and location dominate the predictions, and errors are largest in non-urban and low-value neighborhoods, whose values are overpredicted. The results show that evaluation design changes what machine learning appears to achieve, and that forecasting trending urban quantities requires modeling the trend explicitly. The predictions are communicated through an interactive dashboard for planners, municipalities and residents.

Keywords— machine learning, property valuation, explainable AI, model evaluation, urban analytics, Netherlands

1 Introduction

Machine learning (ML) is increasingly used to predict and monitor urban and regional phenomena, from traffic and energy demand to rents and property values, and its outputs increasingly inform planning and policy. How such models are evaluated determines what their users can expect from them. Random train–test splits, high R^2 values and the absence of simple benchmarks can make a model look accurate while hiding how it behaves when the predicted quantity trends beyond the range of the training data, which is precisely the situation in which forecasts matter

most. This study examines these issues for one of the most policy-relevant urban quantities, the value of housing, during the Dutch housing boom of 2014–2024.

House prices in the Netherlands have risen almost continuously since the end of the global financial crisis, driven by a structural shortage of housing (Boelhouwer and van der Heijden, 2022; Boelhouwer, 2020). Population growth accelerated from 0.42% in 2014 to 1.25% in 2022 (Centraal Bureau voor de Statistiek, 2025b), and by 2024 the market tightness indicator of the Dutch association of real estate agents stood at 2.1, far below the value of around seven associated with a balanced market (NVM, 2024; Boelhouwer and van der Heijden, 2022). In the neighborhood data used here, the mean official valuation rose by 68% between 2014 and 2024. Because housing wealth and access have become a driver of social inequality (Boelhouwer, 2020), knowing where values rise, how well such changes can be anticipated, and which neighborhood characteristics accompany them matters for spatial planning, local taxation and housing policy.

A natural measure for such an analysis is the WOZ value (*Waardering Onroerende Zaken*), the official market value that each municipality assigns annually to every dwelling. WOZ values are the basis for local property taxation, enter income taxation and the points system for regulated rents, and are publicly available for every residential address. Unlike transaction prices, which are observed only for dwellings sold in a given year, WOZ values cover the entire housing stock and can be aggregated consistently to neighborhoods; they also underpin the official Dutch house price index (de Vries et al., 2009). As products of mass appraisal, however, they are smoothed and lag the market (Quan and Quigley, 1991; Geltner, 1993).

Machine learning (ML) has become a common technique for predicting property prices. Tree-based ensembles such as random forest and gradient boosting repeatedly outperform linear hedonic models because they capture non-linear relationships and interactions (Antipov and Pokryshevskaya, 2012; Ho et al., 2021; Soltani et al., 2022; Mathotaarachchi et al., 2024). Three features of this literature limit what it can tell planners and housing analysts. First, most studies use cross-sectional transaction data from a single city and evaluate models on random hold-out samples, which says little about predicting a future period (Bergmeir and Benítez, 2012; Roberts et al., 2017). Second, models are rarely compared with naive benchmarks, such as carrying the previous value forward, although such benchmarks are standard in forecasting (Hyndman and Koehler, 2006). Third, the inability of tree ensembles to extrapolate beyond the training data (Hastie et al., 2009) has received little empirical attention in housing markets, even though it matters most when prices rise fastest.

This study addresses these gaps at two spatial scales, asking (1) how accurately random forest and XGBoost predict WOZ values for neighborhoods and properties compared with naive and linear benchmarks, (2) how their accuracy changes over a housing boom, and why, (3) which neighborhood characteristics drive the predictions and where errors concentrate, and (4) what the models imply for the short-term development of values and how the results can be communicated to non-specialist users. We build a panel of 4,714 Dutch neighborhoods for 2014–2024 from the key figures of Statistics Netherlands (*Centraal Bureau voor de Statistiek*, CBS) and a dataset of 79,216 residential properties in Utrecht from the public WOZ register.

The core contribution is methodological. We show that the way tree-based models are evaluated changes what they appear to achieve: under out-of-time validation and against naive benchmarks, their accuracy for appraisal-based values is far more modest than random splits suggest, and in a rising market they underpredict systematically, by an amount almost perfectly explained by market growth. We diagnose the mechanism with three complementary checks, including SHAP explanations (Lundberg and Lee, 2017; Lundberg et al., 2020) that expose a frozen contribution of the year variable, and derive practices for evaluating ML models of trending

urban quantities. Empirically, the paper provides a nationwide neighborhood-level analysis of WOZ values over the full post-crisis boom, complemented by address-level evidence for one of the tightest Dutch housing markets, and identifies the neighborhood correlates of value and where predictions are least reliable. Practically, it delivers an interactive dashboard that maps predicted values together with their main determinants. Section 2 reviews the literature, Sections 3 and 4 describe the data and methods, Section 5 reports the results, and Sections 6 and 7 discuss and conclude.

2 Literature review

2.1 Determinants of house prices

Hedonic price theory treats a dwelling as a bundle of characteristics whose implicit prices are revealed by market transactions (Rosen, 1974). Reviews identify three groups of determinants: structural attributes such as floor area, age and condition; locational attributes such as accessibility to employment, services and transport; and neighborhood attributes such as income, safety, schools and green space (Sirmans et al., 2005; Soltani et al., 2022). In the Netherlands, the residential environment explains a substantial part of the spatial variation in house prices beyond dwelling characteristics (Visser et al., 2008), and mixed land use in the surroundings is capitalized into property values (Koster and Rouwendal, 2012). Neighborhood income is both a cause and a consequence of value, because households sort into neighborhoods according to their willingness and ability to pay for local amenities (Bayer et al., 2007). At the market level, prices respond to macroeconomic conditions such as income, interest rates and liquidity (Agnello and Schuknecht, 2011; Cohen and Karpavičiūtė, 2017), and price changes are positively autocorrelated in the short run (Case and Shiller, 1989), which makes past values informative but implies momentum that can overshoot at turning points.

2.2 Appraisal-based values

WOZ values are appraisals rather than prices. Appraisers combine recent transaction evidence with previous valuations, which is rational when individual transactions are noisy but produces values that are smoother and lag behind true market values (Quan and Quigley, 1991; Geltner, 1993). Mass appraisal amplifies this persistence because the same models and comparable are applied to the whole stock year after year. For prediction, smoothing implies that a property's previous valuation is an exceptionally strong predictor of its next one, so that naive persistence is a demanding benchmark. For interpretation, it implies that models of WOZ values learn the behavior of the valuation system as much as the behavior of the market. At the neighborhood level, averaging over many dwellings removes idiosyncratic appraisal noise and leaves the systematic component, which is what makes aggregated WOZ values a suitable basis for national comparisons (de Vries et al., 2009).

2.3 Machine learning for property valuation

Automated valuation increasingly relies on ML. Random forest (Breiman, 2001) averages many regression trees grown on bootstrap samples with random subsets of predictors, which reduces

variance and makes the method robust to outliers and multicollinearity. Gradient boosting (Friedman, 2001) instead builds trees sequentially, each fitting the residuals of the current ensemble, and XGBoost (Chen and Guestrin, 2016) adds regularization and second-order optimization to this scheme. Both families have been applied to mass appraisal of apartments (Antipov and Pokryshevskaya, 2012) and compared with hedonic regression and its updating methods (Mayer et al., 2019).

Comparative studies converge on the superiority of tree ensembles. In Bengaluru, XGBoost achieved the highest R^2 among linear, regularized, support vector and boosting regressions (Manasa et al., 2020). For housing in Ames, Iowa, gradient boosting and random forest outperformed seven other regressors (Kangane et al., 2021), and in Hong Kong random forest and gradient boosting machines produced lower errors than support vector machines (Ho et al., 2021). Using UK Land Registry transactions, Mathotaarachchi et al. (2024) found random forest and gradient boosting models to be the most accurate, and Soltani et al. (2022) showed that gradient-boosted trees outperform linear models once spatio-temporal dependence is incorporated. Feature selection before training, for example with recursive feature elimination, can further reduce prediction error on real estate datasets (Chanasit et al., 2021).

Prediction and explanation are, however, different tasks (Mullainathan and Spiess, 2017). Importance scores of trained ensembles rank the predictors a model relies on (Alsahaf et al., 2022), but impurity-based scores are biased toward continuous and correlated predictors (Strobl et al., 2007). Shapley additive explanations (SHAP) attribute each prediction additively to the predictors and can be computed exactly for tree ensembles, which yields consistent global importance measures and the direction of each predictor's association with the prediction (Lundberg and Lee, 2017; Lundberg et al., 2020). Neither approach identifies causal effects, because correlated predictors share credit and the models are fitted to observational data.

Beyond that, evaluation design matters. For temporally ordered data, random cross-validation allows the model learn from observations that are close in time to the test observations, which inflates accuracy. Therefore, out-of-time evaluation is the appropriate test for any forecasting use (Bergmeir and Benítez, 2012; Roberts et al., 2017). Accuracy should also be judged against simple benchmarks and with scale-free measures that reveal bias as well as dispersion (Hyndman and Koehler, 2006).

2.4 Hypotheses

H1 (model choice). Gradient boosting outperforms random forest and linear regression for neighborhood values, and the advantage is larger in the heterogeneous national panel than in the property data, where a single predictor carries most information.

H2 (persistence). Because appraised values are smoothed, naive persistence is a demanding benchmark, and at the property level most of the accuracy of ML models derives from valuation history.

H3 (extrapolation). Because tree ensembles predict averages of training observations, attribute-based predictions for a future year are anchored at the market level of the last training years, so they underpredict in proportion to market growth. Lagged values reduce this bias in persistent trends but carry momentum past turning points.

H4 (determinants). Neighborhood socioeconomic composition, especially income, and proxies for dwelling size and type are the main drivers of predicted neighborhood values.

3 Data

Two complementary datasets are constructed (Table 1): a national neighborhood panel for 2014, the first year after the bust, to 2024, and a property dataset for the municipality of Utrecht. Variable definitions are listed in Supplementary Material Tables S1 and S2. The study uses only publicly available administrative data on neighborhoods and residential properties and involves no human participants; ethical approval was therefore not required.

Table 1: **Overview of the two datasets**.

	Neighborhood panel (CBS)	**Property data (Utrecht)**
Source	CBS *Kerncijfers wijken en buurten*	*WOZ-waardeloket* and BAG
Unit of observation	Neighborhood (*buurt*) per year	Residential address
Coverage	The Netherlands, 2014–2024	Municipality of Utrecht, valuations 2014–2023
Observations	4,714 neighborhoods in 203 municipalities × 11 years = 51,854	79,216 properties
Predictors	68: population, housing, energy, income, social security, businesses, vehicles, facilities, area, urbanity, year and encoded location codes	17: year built, garden indicator, floor area, encoded postcode, street and neighborhood, coordinates and WOZ values 2014–2022
Target	Mean WOZ value (€1,000)	WOZ value 2023 (€)
Validation	Temporal: train 2014–2023, test 2024; expanding window 2015–2024	Random: 80% train, 20% test

3.1 Neighborhood panel

The annual *Kerncijfers wijken en buurten* releases of CBS provide public statistics for all municipalities, districts and neighborhoods of the Netherlands (Centraal Bureau voor de Statistiek, 2025a), covering population, housing, energy, income, social security, businesses, vehicles, facilities, land area, postal codes and urbanity, together with the mean WOZ value of the residential stock. Combining eleven releases raises two difficulties. First, boundaries change: between 2014 and 2024 the number of municipalities fell from 403 to 342 through mergers, while the numbers of districts and neighborhoods increased (Supplementary Material Figure S1). Using the CBS change indicator, units whose code and delineation are unchanged are linked backward year by year from the 2024 release, and only neighborhood-level units are kept. Second, variable definitions change, so only variables present in every release are retained.

Neighborhoods are kept if they are observed in all eleven years, have fewer than six missing WOZ values and have on average at least 100 inhabitants. Energy consumption by dwelling type,

which is largely missing (Supplementary Material Figure S2), is dropped in favor of neighborhood totals, except for detached houses, which signal high-value dwellings. The resulting panel contains 51,854 neighborhood-years for 4,714 neighborhoods in 203 municipalities and 71 variables, of whose values 8.5% are missing, including 3.6% of the WOZ values. Table 2 summarizes the main variables. The mean neighborhood WOZ value rose by 68%, from €253,100 in 2014 to €425,400 in 2024, and its dispersion widened, while average income per resident increased by 44% and household energy consumption fell.

Table 2: **Descriptive statistics of selected neighborhood variables, 2014 and 2024**.

		2014		2024	
Variable	**Unit**	**Mean**	**SD**	**Mean**	**SD**
Mean WOZ value	€ 1,000	253.1	115.5	425.4	169.7
Average income per resident	€ 1,000	23.8	5.5	34.3	8.1
Average income per income recipient	€ 1,000	30.4	7.5	40.8	9.9
Households in top 20% of income	%	24.0	12.1	23.6	11.7
Electricity consumption	kWh per dwelling	3,552	862	2,923	726
Natural gas consumption	m^3 per dwelling	1,513	518	1,112	411
Population density	inhabitants/km^2	3,128	3,289	3,270	3,395
Address density	addresses/km^2	1,101	1,145	1,227	1,259
Owner-occupied dwellings	%	67.2	21.0	68.7	21.1
Housing-association dwellings	%	19.8	21.0	18.9	20.0
Inhabitants	persons	1,585	1,859	1,674	1,935

Note: N = 4,714 neighborhoods. Values after imputation (Section 4.1).

3.2 Property-level valuations

WOZ values of residential properties are publicly accessible through the *WOZ-waardeloket* (WOZ-waardeloket, 2025), which also reports year of construction, lot size and floor area from the Basic Registration of Addresses and Buildings (BAG). Starting from all 175,299 BAG addresses in the city of Utrecht, each address is geocoded with the national location service (PDOK) and its valuation history for reference years 2014–2023 is retrieved. Because the interface allows at most 10,000 requests per day, data collection is restricted to Utrecht, one of the tightest housing markets in the country. Of the 97,281 addresses with a residential WOZ record, 79,615 are complete, and removing the highest 0.5% of 2023 valuations leaves 79,216. Lot size is recoded into an indicator of whether a dwelling has its own plot, which is absent for 63.4% of properties, mostly apartments. The mean valuation more than doubled, from €190,400 in 2014 to €424,500 in 2022, before declining slightly to €416,300 in 2023 in line with the national correction.

4 Methodology

The analysis follows the pipeline in Figure 1: the datasets are prepared and imputed, ensembles and benchmarks are trained and evaluated under different validation designs, the predictions are explained, and the best model is used for forecasting and mapping.

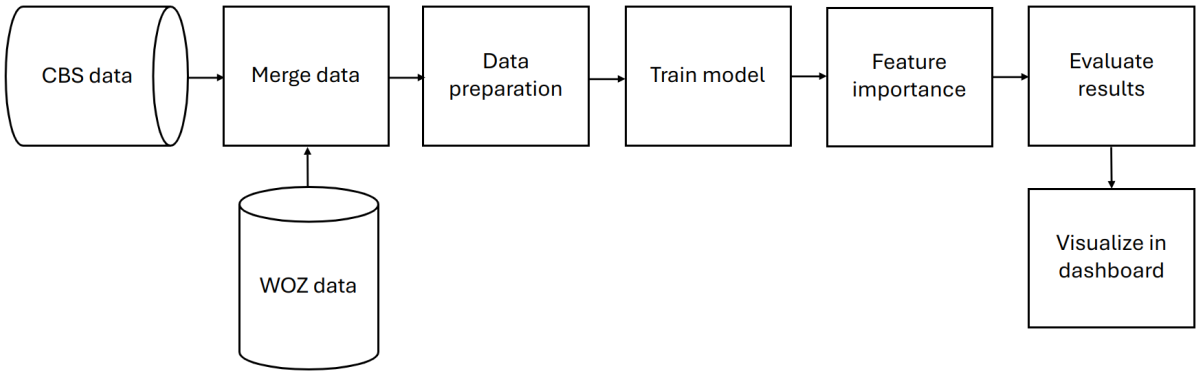

Figure 1: **Analysis pipeline**. Neighborhood statistics (CBS) and property valuations (WOZ) are merged and prepared, models are trained, feature importance is derived, results are evaluated, and predictions are mapped in an interactive dashboard.

4.1 Handling missing values

Because most algorithms require complete data, and because deleting incomplete records would remove many neighborhoods, the 8.5% missing values in the panel are imputed with a rule-based procedure that exploits the time series of each neighborhood (Supplementary Material Figure S3). Each variable is grouped by neighborhood, and the method is chosen according to the number of missing years, the variance of the observed values and the strength of their linear trend.

For a neighborhood with between one and six missing years, the sample variance of the observed values is computed first. If it does not exceed one, values are nearly constant and missing years are filled with the mean of the observed years (variance below 0.01) or carried forward from the previous year. Otherwise, a linear regression of the observed values on year is fitted. If its coefficient of determination satisfies $R^2 \geq 0.7$, isolated missing years are filled by linear interpolation and runs of consecutive missing years by stepwise linear extrapolation from the neighboring observed values. If $R^2 < 0.7$, a natural cubic spline is used, which fits piecewise third-order polynomials between observed years and avoids the oscillations of a single high-order polynomial. Values that remain missing, mainly in neighborhoods with seven or more missing years in a variable, are completed with multivariate imputation by chained equations, in which each incomplete variable is iteratively regressed on all others (van Buuren and Groothuis-Oudshoorn, 2011; Pauzi et al., 2021). This step uses the iterative imputer of scikit-learn (Pedregosa et al., 2011) on standardized values of all 69 numerical variables (at most 50 iterations, tolerance 10^{-4}) and produces a single completed dataset. Extrapolation occasionally produces implausible values as the ten imputed neighborhood WOZ values, mostly in 2014–2016, are zero or negative. They are retained in the main models and excluded only where a logarithmic transformation requires positive values (Section 4.3).

4.2 Prediction models

Let y_i denote the WOZ value of observation i and $\mathbf{x}_i$ its vector of predictors. Both models approximate y_i by an ensemble of regression trees, each of which partitions the predictor space into regions R_j with constant predictions.

Random forest (Breiman, 2001) grows M trees T_m independently, each on a bootstrap sample

of the training data. At every split only a random subset of predictors is considered, and the split is chosen to minimize the squared error within the resulting nodes. The prediction is the average over trees:

$$\hat{y}^{\mathrm{RF}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} T_m(\mathbf{x}). \tag{1}$$

Averaging decorrelated trees reduces variance but does not correct systematic errors of individual trees.

Gradient boosting (Friedman, 2001) builds trees sequentially. Starting from a constant F_0, at iteration m a tree is fitted to the negative gradient of the loss function L evaluated at the current model,

$$r_{im} = - \left[\frac{\partial L\big(y_i, F(\mathbf{x}_i)\big)}{\partial F(\mathbf{x}_i)} \right]_{F=F_{m-1}}, \tag{2}$$

and the model is updated as

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \sum_{j=1}^{J_m} \gamma_{jm}\, \mathbf{1}(\mathbf{x} \in R_{jm}), \tag{3}$$

where γ_{jm} is the optimal value in leaf R_{jm} and ν the learning rate. For the squared error loss, r_{im} is the ordinary residual, so that each tree corrects the errors left by its predecessors. XGBoost (Chen and Guestrin, 2016) extends this scheme with L1 and L2 penalties on the leaf weights and a second-order approximation of the loss, which improves generalization and speed.

A property of both models is central to the interpretation of the temporal results. Every leaf value is an average of training targets, so that any prediction lies within the convex hull of the training values. More importantly for a panel, a split on `year` can only use thresholds observed in training; a test year beyond the training period is therefore treated exactly like the last training year. In the absence of lagged values, the predicted change between years must come entirely from changes in the other predictors.

For the neighborhood panel, a five-fold cross-validated grid search over 27 XGBoost configurations (200, 300 or 500 trees; learning rate 0.01, 0.05 or 0.1; maximum depth 3, 5 or 7) favored a learning rate of 0.1 and 500 trees. The evaluated XGBoost model uses 700 trees, a learning rate of 0.1 and a maximum depth of 5. The random forest uses 200 trees with a random subset of $\sqrt{p}$ predictors at each split, a maximum depth of 40, and at least five observations to split a node and two per leaf, following a randomized search. The same configurations are used for both datasets (Supplementary Material Table S3). Models are implemented in Python with scikit-learn (Pedregosa et al., 2011) and the XGBoost library (Chen and Guestrin, 2016). Categorical identifiers, namely municipality and region code in the panel and street and neighborhood in the Utrecht data, are label-encoded as integers.

4.3 Benchmarks and robustness checks

To judge whether the ensembles add value, they are compared with three benchmarks. *Naive persistence* predicts the current value by the previous year's value of the same unit. *Naive drift* multiplies the previous value by the average growth factor of the last observed year (2022–2023 for the neighborhoods; 2022 to 2023 in the Utrecht training sample). *Linear regression* (ordinary least squares on standardized predictors) represents a hedonic-type model on the same predictors

as the ensembles. Persistence and drift require the unit's own valuation history; the ensembles and the linear model in the main specification do not.

Three robustness checks are run with XGBoost on the neighborhood panel. The first removes the encoded location identifiers (municipality, region code and most common postal code), which may let the model memorize neighborhoods. The second models the logarithm of the WOZ value, which stabilizes the variance and turns additive into multiplicative effects; it is estimated on the training observations with positive values. The third adds the WOZ values lagged by one, two and three years and their rolling three-year mean, which turns the attribute-based model into a history-based one. Lags undefined in the first years of the panel are left missing, which XGBoost handles natively. For Utrecht, the ensembles are additionally trained on dwelling characteristics and location only, without valuation history, to separate the contribution of history from that of attributes.

4.4 Feature selection and explanation

Three complementary approaches identify the most influential predictors. Before training, recursive feature elimination (RFE) (Guyon et al., 2002) with a random forest regressor repeatedly fits the model, ranks the predictors by importance and removes the least important ones, five at a time; the number of retained predictors is chosen by three-fold cross-validation on the training data. For the neighborhood panel this retains 18 of the 68 predictors (Supplementary Material Table S4). For the Utrecht data, cross-validation retains all 17 predictors, so the reduced set consists of the ten highest-ranked ones. Each model is then trained both on the full and on the reduced feature set. After training, the built-in importance scores of both models are extracted (gain for XGBoost, total reduction in squared error for random forest), both for a single model trained on the full panel and summed over ten models trained on expanding windows.

Because impurity-based scores are biased toward continuous and correlated predictors (Strobl et al., 2007) and do not indicate direction, the main interpretation relies on SHAP values (Lundberg and Lee, 2017). For a prediction $\hat{y}(\mathbf{x}_i)$, SHAP decomposes the deviation from the average training prediction ϕ_0 into contributions ϕ_{ik} of the p predictors,

$$\hat{y}(\mathbf{x}_i) = \phi_0 + \sum_{k=1}^{p} \phi_{ik}, \tag{4}$$

where ϕ_{ik} is the Shapley value of predictor k, its average marginal contribution over all orderings of the predictors. For tree ensembles the values are computed exactly (Lundberg et al., 2020); here they are obtained with the XGBoost implementation for the 4,714 neighborhoods in the 2024 test set. Global importance is measured by the mean absolute SHAP value of each predictor, and direction by the correlation between a predictor's value and its SHAP value across neighborhoods.

4.5 Validation design and accuracy measures

The neighborhood panel is split by time. Models are trained on 2014–2023 and tested on 2024, so that predictions are always made for a year that was not seen during training. The current WOZ value is excluded from the predictors. To trace accuracy over the boom, an expanding-window evaluation repeats the procedure for each year from 2015 to 2024, with models trained on all preceding years. For the Utrecht data, the target is the 2023 valuation and the predictors include the valuations of 2014–2022; the properties are split randomly into 80% for training and 20% for

testing with a fixed random seed. Because all predictors refer to earlier years, this split does not leak information from the target year, but it does let the model learn the 2022–2023 relationship from neighboring properties, which makes it a cross-sectional rather than a forecasting test.

Accuracy is measured by the coefficient of determination, the root mean squared error, the mean absolute error and two scale-free measures, the mean absolute percentage error (MAPE) and the mean percentage error (MPE), which reveals systematic over- or underprediction (Hyndman and Koehler, 2006):

$$\begin{aligned} R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \mathrm{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_i (y_i - \hat{y}_i)^2}, \quad \mathrm{MAE} = \tfrac{1}{n}\textstyle\sum_i |y_i - \hat{y}_i|, \\ \mathrm{MAPE} &= \tfrac{100}{n}\textstyle\sum_i \left|\frac{\hat{y}_i - y_i}{y_i}\right|, \qquad \mathrm{MPE} = \tfrac{100}{n}\textstyle\sum_i \frac{\hat{y}_i - y_i}{y_i}. \end{aligned} \tag{5}$$

A negative MPE indicates underprediction. Errors are also broken down by degree of urbanity and by quintile of the actual value.

All results reported below were obtained by re-running the analysis code with XGBoost 3.4.1 and scikit-learn 1.9.1. The mean squared errors of the four main neighborhood models differ by less than 1.5% from the original runs, and the property-level results are identical to three decimals in R^2 (Supplementary Material Table S6).

4.6 Diagnosing extrapolation bias

To test whether the temporal errors reflect the inability of trees to extrapolate (H3), three diagnostics are computed from the expanding-window results. First, the mean percentage error of each test year, MPE_t, is correlated with the growth rate of the mean WOZ value, $g_t = \bar{y}_t / \bar{y}_{t-1} - 1$; a strong negative correlation indicates that the models lag the market level. Second, the share of test values that exceed the highest value seen in training is computed for each year, to distinguish extreme target values from a shift in the overall market level. Third, the SHAP contribution of the `year` variable is inspected. Because a tree treats any test year as the last training year, this contribution is frozen at its training value, and its size shows how much of the predicted level the model attributes to market-wide appreciation rather than to neighborhood characteristics. The same diagnostics are applied to the model with lagged values to assess whether history-based predictors remove the bias or merely shift it.

4.7 Forecasting

For forecasting, XGBoost is retrained on all available data. In the neighborhood panel, the model (500 trees, learning rate 0.1, maximum depth 5) uses a reduced set of 40 neighborhood predictors together with the WOZ values lagged by one to three years and the rolling mean of the three preceding years, so that no information from the target year enters the predictors. Lags that are undefined in the first years of the panel are imputed by chained equations. To forecast 2025, a new observation is added for each neighborhood, its non-lag predictors are carried forward from 2024, and its lags are updated with the latest observed values; the procedure is repeated for 2026 using the predicted 2025 values. For Utrecht, an XGBoost model (200 trees, learning rate 0.1, maximum depth 7, the best configuration of a separate grid search) is trained to predict the 2023 valuation from the earlier ones and then applied after shifting the valuation history forward by one year, so that the 2023 valuation takes the role of the most recent predictor. Iterating this step yields forecasts for 2024 to 2026.

5 Results

5.1 Predictive performance and benchmarks

Table 3 reports the accuracy of all models on the 2024 neighborhood test set. Among the attribute-based models, XGBoost is the most accurate, with $R^2 = 0.931$, an RMSE of €44,600 and a MAPE of 7.4%; three in four neighborhoods (75.3%) are predicted within 10% of their actual value, and 95.2% within 20%. Random forest reaches $R^2 = 0.856$ (MAPE 9.9%) and linear regression only $R^2 = 0.757$ (MAPE 12.9%). The ensembles thus capture substantial non-linearity and interaction that a linear specification on the same predictors misses, and boosting captures more of it than averaging, in line with H1. Feature selection has opposite effects on the two ensembles. It lowers the accuracy of XGBoost slightly but improves that of random forest, which benefits from the removal of redundant predictors that would otherwise be drawn into its random subsets, whereas XGBoost already discounts uninformative predictors through its gain-based splitting (Alsahaf et al., 2022).

Table 3: **Accuracy of benchmarks, ensembles and robustness checks for neighborhood WOZ values, 2024 hold-out year**.

Model	R^2	RMSE	MAE	MAPE (%)	MPE (%)
Benchmarks					
Naive persistence (previous year)	0.982	22.5	16.5	3.9	−3.2
Naive drift (previous year × national growth)	0.894	55.3	47.4	11.1	11.0
Linear regression	0.757	83.6	55.2	12.9	−2.1
Attribute-based ensembles					
Random forest	0.856	64.4	43.4	9.9	−3.0
Random forest (RFE)	0.873	60.5	40.1	9.5	−0.6
XGBoost	0.931	44.6	30.7	7.4	1.4
XGBoost (RFE)	0.925	46.5	32.0	7.7	1.9
Robustness checks (XGBoost)					
Without location identifiers	0.890	56.3	39.4	9.6	2.6
Logarithmic target	0.931	44.5	29.5	6.7	−0.6
With lagged WOZ values	0.916	49.1	43.2	10.7	10.6

Note: RMSE and MAE in € 1,000; mean actual value €425,400; N = 4,714 neighborhoods. Models are trained on 2014–2023. RFE = recursive feature elimination (18 predictors). Location identifiers are the encoded municipality, region code and most common postal code. MPE < 0 indicates underprediction.

The benchmarks put these results in perspective. Carrying each neighborhood's 2023 value forward is more accurate than any attribute-based model ($R^2 = 0.982$, MAPE 3.9%), although it underpredicts by 3.2% on average because it ignores market growth. This supports H2 and is expected for smoothed appraisal data, in which the valuation system itself relies heavily on previous values. The comparison should, however, be read with the different information sets in mind. Persistence requires the unit's own valuation history and has nothing to say about why values differ; the attribute-based models predict a neighborhood's value from its current characteristics alone, which is the relevant task when no valuation history exists, for example for new or redrawn neighborhoods, and when the aim is to understand value differences. The drift benchmark, which extrapolates the 14.7% national increase of 2023 to all neighborhoods, overpredicts by 11.0% because growth slowed to 3.3% in 2024, illustrating the risk of extrapolating momentum at a turning point.

The robustness checks clarify what drives XGBoost's accuracy. Removing the location

identifiers lowers R^2 from 0.931 to 0.890 and raises MAPE from 7.4% to 9.6%, so that part of the accuracy derives from coarse location rather than from neighborhood characteristics, but most of it does not. Modeling the logarithm of the value leaves R^2 unchanged but lowers MAPE to 6.7% and almost removes the average bias, as expected when errors are proportional to the level of value. Adding lagged values does not improve the 2024 prediction since the model inherits the momentum of the 2023 increase and overpredicts by 10.6%, a result examined further in the next subsection.

At the property level (Table 4), random forest and XGBoost perform almost identically ($R^2 = 0.983$, RMSE about €22,000, or 5% of the mean 2023 valuation) and halve the MAPE of the persistence benchmark (2.3–2.8% against 5.5%). Linear regression on the same predictors ($R^2 = 0.970$, MAPE 4.7%) is better than persistence but clearly worse than the ensembles. Without the valuation history, however, the ensembles reach only $R^2 = 0.913$–0.926 and a MAPE of 4.6–6.0%, close to the persistence benchmark. Restricting the models to the ten predictors ranked highest by RFE lowers R^2 slightly, to 0.981 for XGBoost and 0.982 for random forest. Most of the property-level accuracy therefore stems from valuation history, consistent with H2, while the ensembles add value by learning how the history of each property should be updated given its characteristics and location. The smaller gap between the two algorithms than in the neighborhood panel is consistent with H1 when one predictor carries most information, the choice of algorithm matters little.

Table 4: **Accuracy of benchmarks and ensembles for Utrecht property values, 2023 (random 20% test sample)**.

Model	R^2	RMSE (€)	MAE (€)	MAPE (%)	MPE (%)
Naive persistence (2022 value)	0.960	34,155	23,662	5.5	2.3
Naive drift (2022 value × city growth)	0.964	32,590	22,031	5.1	0.4
Linear regression	0.970	29,646	20,052	4.7	0.6
Random forest	0.983	22,054	11,047	2.3	0.3
XGBoost	0.983	22,023	12,678	2.8	0.2
Without valuation history (characteristics and location only)					
Random forest	0.926	46,510	20,191	4.6	1.0
XGBoost	0.913	50,445	25,393	6.0	1.2

Note: $N = 15{,}844$ test properties; mean 2023 valuation €416,300. The city growth factor (0.981) is the ratio of mean 2023 to mean 2022 valuations in the training sample. Characteristics and location are year built, garden indicator, floor area, encoded postcode, street and neighborhood, and coordinates.

5.2 Performance over the housing boom

The single 2024 test year conceals strong variation over time. Table 5 and Figure 2 report the expanding-window results. XGBoost is more accurate than random forest in every year, with R^2 between 0.80 and 0.96 against 0.70 and 0.93. Both models, however, systematically underpredict from 2018 onward, by up to 13% in 2023, and their bias moves almost one for one with market growth. The correlation between the annual growth of the mean WOZ value and the MPE is -0.95 for XGBoost and -0.96 for random forest. The bias is smallest in years of weak growth (2016, 2017 and 2024) and largest in the years of strongest growth (2019, 2022 and 2023). In 2024, after the strong increase of 2023 entered the training data and growth slowed, the bias of XGBoost even turns slightly positive.

This pattern confirms H3 and has a simple mechanical explanation. Only 0.04% or fewer

Table 5: **Expanding-window evaluation of neighborhood WOZ predictions, 2015–2024.**

	Actual		R^2			MPE (%)		
Year	**Mean**	**Growth (%)**	**XGB**	**RF**	**XGB-lag**	**XGB**	**RF**	**XGB-lag**
2015	248.4	−1.9	0.920	0.876	–	5.5	7.7	–
2016	251.6	1.3	0.961	0.925	0.978	0.7	3.3	−2.0
2017	258.7	2.8	0.959	0.932	0.980	−1.4	0.6	−1.7
2018	270.4	4.5	0.946	0.910	0.974	−3.5	−3.5	−2.1
2019	291.7	7.9	0.906	0.835	0.949	−6.3	−8.1	−3.3
2020	311.0	6.6	0.906	0.802	0.978	−5.6	−8.1	−0.3
2021	330.2	6.2	0.958	0.890	0.988	−0.6	−5.2	0.4
2022	359.1	8.7	0.881	0.788	0.979	−8.6	−9.3	−2.9
2023	411.9	14.7	0.804	0.701	0.958	−13.2	−13.1	−5.9
2024	425.4	3.3	0.931	0.856	0.916	1.4	−3.0	10.6

Note: Each year is predicted by models trained on all preceding years. Mean actual WOZ value in €1,000; growth relative to the previous year. XGB-lag = XGBoost with WOZ values lagged by one to three years and their rolling three-year mean; it is not estimated for 2015, when only one training year is available.

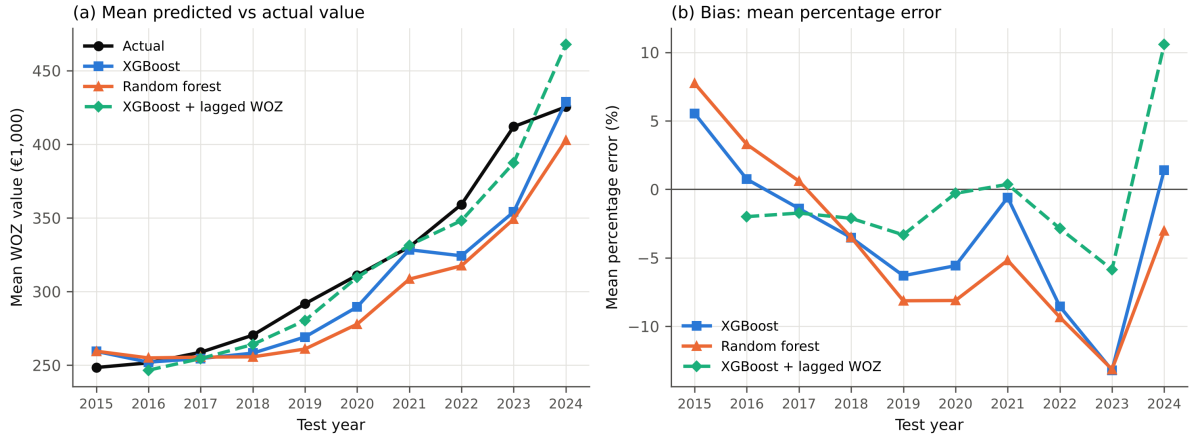

Figure 2: **Expanding-window evaluation, 2015–2024.** (a) Mean actual and predicted neighborhood WOZ values; (b) mean percentage error. Each year is predicted by models trained on all preceding years. Attribute-based models underpredict in years of rapid appreciation; the model with lagged values tracks the trend more closely but overshoots when growth slows in 2024.

of the test values in any year exceed the highest value seen in training, so the bias does not arise from extreme values beyond the training range. It arises from the market level. Because a tree treats any test year as the last training year (Section 4.2), the attribute-based models predict the value that a neighborhood with the observed characteristics had in the last training year. Neighborhood characteristics, such as income, change only gradually, and so the predicted change between years falls short of the actual appreciation whenever the market rises faster than these characteristics. Random forest, which averages more strongly, lags more.

Adding lagged values largely removes this bias as long as the trend persists. The lagged model attains R^2 between 0.95 and 0.99 between 2016 and 2023, and its underprediction in 2023 (5.9%) is less than half that of the attribute-based models. In 2024, however, it overpredicts by 10.6% and becomes less accurate than the attribute-based XGBoost, because it has learned

from the boom years that a strong increase is followed by another one. The two model types thus fail in opposite ways. Attribute-based models lag accelerating markets, while history-based models overshoot at turning points, in line with the short-run momentum documented for housing markets (Case and Shiller, 1989).

5.3 Where predictions err

Figure 3 and Table 6 break down the 2024 errors of XGBoost. Accuracy is similar across the four most urban classes (MAPE 6.1–7.0%) but lower in non-urban neighborhoods (8.5%), which also have the highest mean valuation in the panel and are overpredicted by 2.5% on average. Rural neighborhoods are more heterogeneous in dwelling type and plot size, which the neighborhood averages capture only imperfectly. By value, the model is least accurate and most biased for the lowest quintile (MAPE 8.6%, MPE 4.2%) and slightly underpredicts the highest quintile (−0.4%). This compression toward the mean is typical of ensemble predictions, which shrink extreme cases toward the average of similar training observations, and it implies that attribute-based models flatter the value of the cheapest neighborhoods.

Table 6: **Prediction error of XGBoost by degree of urbanity and value quintile, 2024**.

Segment	N	**Mean / range of actual value**	**MAPE (%)**	**MPE (%)**
Degree of urbanity				
1 Very strongly urban	576	346.8	6.9	1.0
2 Strongly urban	982	359.1	6.9	1.1
3 Moderately urban	709	395.3	6.1	0.1
4 Slightly urban	674	455.1	7.0	0.7
5 Non-urban	1,773	488.3	8.5	2.5
Quintile of actual value				
Q1 (lowest)	945	63–293	8.6	4.2
Q2	941	294–363	6.8	0.9
Q3	952	364–436	7.1	1.3
Q4	936	437–534	6.9	1.0
Q5 (highest)	940	535–2,113	7.5	−0.4

Note: Values in €1,000: mean for urbanity classes, range for quintiles. Degree of urbanity follows the CBS classification based on address density (`ste_mvs`).

5.4 Determinants of predicted values

Supplementary Material Figure S9 shows the predictors retained by RFE for the neighborhood panel. Average income per resident is by far the most important predictor, with roughly twice the importance of the second, average income per income recipient. Average household electricity consumption ranks third, followed by year, the most common postal code, the share of high-income households, population density, address density and gas consumption of owner-occupied dwellings. The built-in importance scores of the trained models, both for a single model and summed over the expanding windows, rank the same predictors highest (Supplementary Material Figure S5).

The SHAP decomposition of the XGBoost predictions for 2024 (Table 7 and Figure 4) adds the direction and size of each association. The largest single contribution comes from `year`,

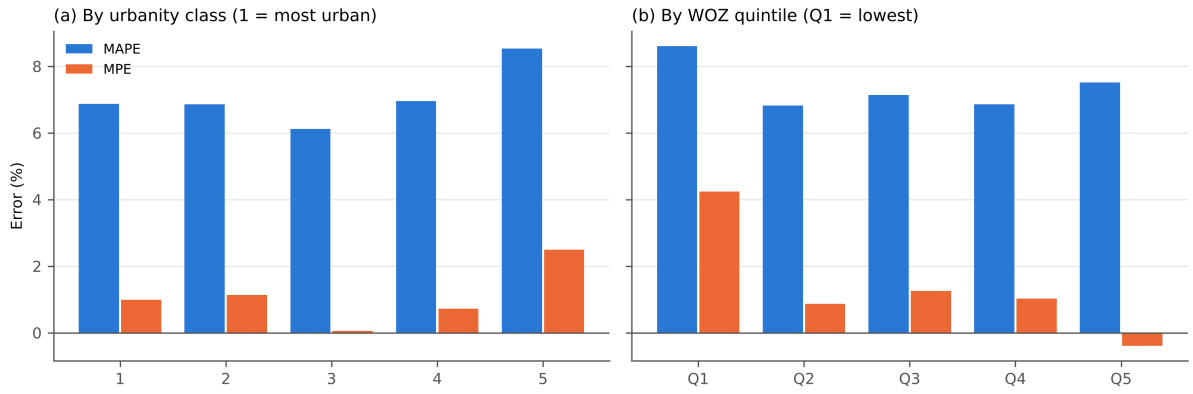

Figure 3: **Mean absolute percentage error (MAPE) and mean percentage error (MPE) of XGBoost for neighborhoods in 2024, by (a) degree of urbanity and (b) quintile of the actual WOZ value**.

which accounts for 26% of the total mean absolute SHAP value and raises the 2024 predictions by about €93,000 on average relative to the average training prediction. This is the model's representation of the market-wide appreciation since 2014, frozen at its 2023 level, and thus the quantity behind the temporal bias discussed above. Among neighborhood characteristics, the two income measures contribute most (together 18%), followed by electricity consumption, the most common postal code, the share of high-income households and population density. Income, electricity and gas consumption and the share of high-income households are positively associated with predicted value, with correlations between predictor and SHAP value of 0.76 to 0.96. Population density and the number of dwellings are negatively associated. Denser neighborhoods contain more and smaller apartments, so their mean value per dwelling is lower even where prices per square meter are high. The negative association with the most common postal code reflects the national gradient from the high-value western Randstad (postal codes starting with 1 to 3) to the east and south.

These results support H4. Predicted neighborhood values are driven primarily by income composition, which reflects the sorting of high-income households into high-value neighborhoods (Bayer et al., 2007), and by energy consumption, which most likely acts as a proxy for dwelling size and type, as larger and detached dwellings consume more electricity and gas. Two cautions apply. First, the share of persons in the lowest 40% of the income distribution is positively associated with the prediction once the other income variables are accounted for, which is unlikely to be a substantive effect and illustrates how SHAP values divide credit among strongly correlated predictors. Second, the encoded identifiers (postal code, region code and municipality) together account for about 10% of the attribution, and their removal lowers accuracy (Table 3), indicating that the models partly learn location rather than characteristics.

At the property level, the previous valuation dominates all other predictors. In the RFE ranking for Utrecht, the 2022 valuation obtains an importance of 0.963, followed by floor area (0.004) and the valuations of 2021 (0.003), 2020 (0.003) and 2018 (0.003) (Supplementary Material Table S5). The influence of past valuations decreases with the time gap, which indicates strong short-run persistence of property values, in line with appraisal smoothing (Quan and Quigley, 1991) and with the serial correlation documented for housing markets (Case and Shiller, 1989).

Table 7: **Most important predictors of neighborhood WOZ values: SHAP values of XGBoost, 2024 test set.**

Predictor	Description	Mean $\|\phi\|$	Share (%)	Direction
`year`	Year (market level)	93.2	26.1	–
`g_ink_po`	Average income per income recipient	32.8	9.2	+0.96
`g_ink_pi`	Average income per resident	32.2	9.0	+0.94
`g_ele`	Average electricity consumption per dwelling	22.7	6.4	+0.89
`pst_mvp`	Most common postal code	22.4	6.3	−0.80
`p_hh_hi`	Households in the top 20% of income	14.3	4.0	+0.96
`bev_dich`	Population density	13.9	3.9	−0.65
`gwb_code`	Region code (encoded)	8.7	2.5	−0.49
`p_ink_li`	Persons in the lowest 40% of income	8.7	2.4	+0.89
`a_woning`	Number of dwellings	8.0	2.2	−0.55
`a_soz_ow`	Recipients of state old-age pension	6.2	1.7	+0.56
`p_ink_hi`	Persons in the highest 20% of income	5.6	1.6	+0.89
`g_gas_ko`	Average gas consumption, owner-occupied dwellings	5.4	1.5	+0.76

Note: Mean $|\phi|$ is the mean absolute SHAP value in €1,000 over the 4,714 neighborhoods; share is its percentage of the sum over all 68 predictors. Direction is the correlation between the predictor's value and its SHAP value; for `year`, which is constant in the test set, it is undefined.

5.5 Forecasts and dashboard

Values in Utrecht started well below the national neighborhood mean in 2014 but rose faster, overtaking it from 2018 (Figure 5). The forecast mean neighborhood value rises from €425,400 in 2024 to €441,200 in 2025 and €457,000 in 2026, roughly 3.6–3.7% per year, and the Utrecht mean from €416,300 in 2023 to €456,000 in 2026. Both imply a continued but more moderate increase than during 2019–2023. In light of the expanding-window results, these figures are trend-continuation scenarios that would overstate growth if the market turned.

The predictions are mapped in an interactive dashboard, the *Residential Valuation Observatory*, built with Streamlit (Streamlit Inc., 2025) and Plotly (Plotly Technologies Inc., 2025). A national map shows the predicted percentage change in WOZ value of each neighborhood, with the most important predictors, such as income per resident, electricity consumption and population density, displayed on hovering (Figure 6). A second map shows predicted values of individual Utrecht properties projected on their neighborhoods (Supplementary Material Figure S8). The dashboard lets users relate the spatial distribution of expensive and affordable areas to the characteristics of their neighborhoods.

6 Discussion

6.1 Interpretation of the findings

XGBoost is the most accurate attribute-based model for neighborhood values in every evaluation year, which confirms for a national panel and an out-of-time test what comparative studies

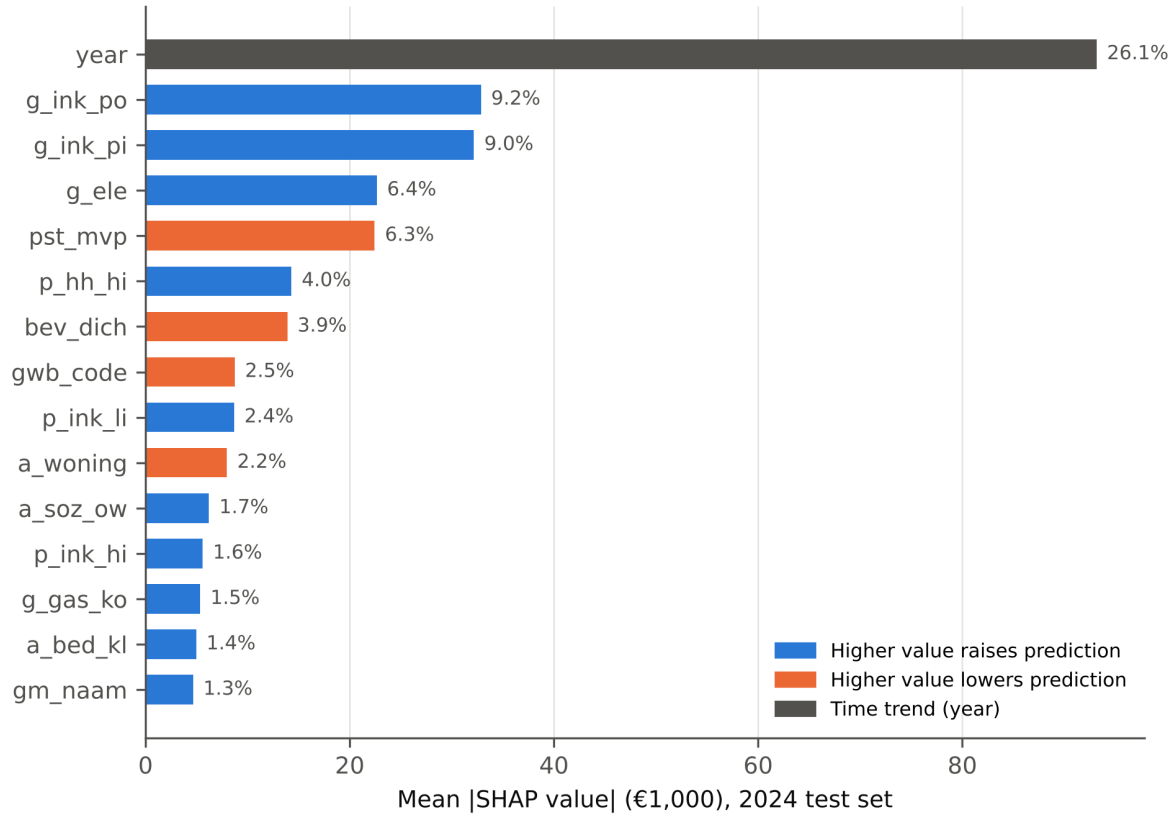

Figure 4: **Mean absolute SHAP values of the 15 most important predictors of XGBoost for neighborhood WOZ values in 2024, with their share of the total**. Color indicates whether higher values of the predictor raise or lower the prediction. Variable names are defined in Supplementary Material Table S1.

have reported for cross-sectional transactions (Manasa et al., 2020; Kangane et al., 2021; Mathotaarachchi et al., 2024). The benchmark comparisons, however, qualify the high accuracy figures common in this literature. Naive persistence beats all attribute-based neighborhood models in 2024 and achieves $R^2 = 0.96$ for properties without any model, so high R^2 values for appraisal-based targets say less than they appear to, and studies using previous values or random splits of panel data should report persistence as a minimum benchmark (Hyndman and Koehler, 2006). The ensembles nevertheless add value: for properties they halve the percentage error of persistence, and for neighborhoods they predict values from current characteristics alone, which persistence cannot do.

The temporal analysis yields the main methodological insight. The underprediction of tree ensembles in a rising market is a systematic property, almost perfectly explained by market growth, that stems from the inability of trees to extrapolate the market level; SHAP makes it visible as a large but frozen contribution of the year variable. Lagged values replace it with the opposite error at turning points. Neither model type is therefore safe for forecasting in a changing market without an explicit treatment of the trend (Section 6.2).

Spatially, the determinants identified by RFE, impurity-based importance and SHAP agree: income composition, energy consumption, density and location are the main neighborhood correlates of WOZ values, in line with evidence that socioeconomic composition and the residential environment explain much of the spatial variation in Dutch house prices (Visser et al., 2008). Property values are thus embedded spatially, through neighborhood characteristics, and tempo-

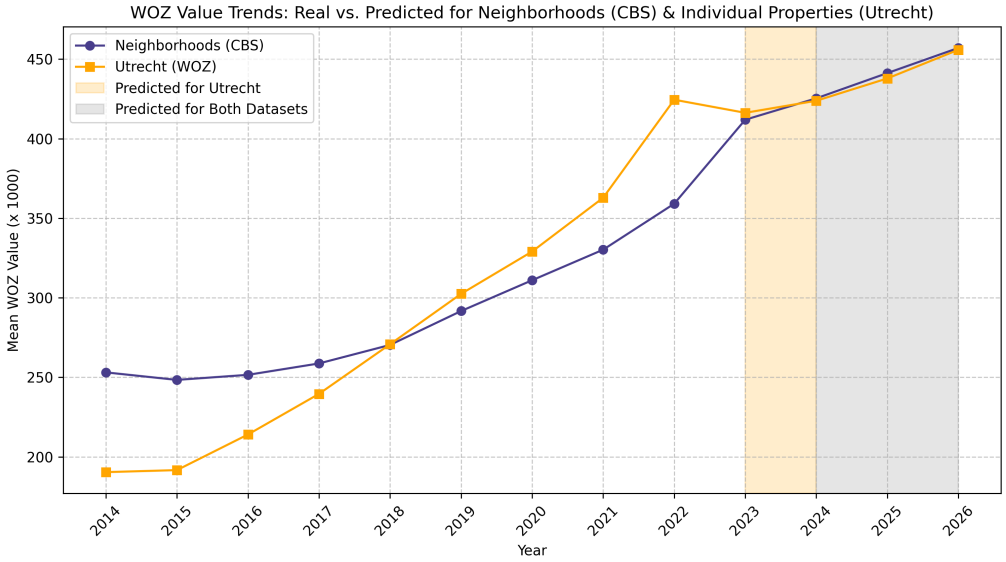

Figure 5: **Observed and forecast WOZ values**. Mean values (€1,000) for Dutch neighborhoods (CBS) and Utrecht properties (WOZ). Shaded areas mark years forecast by XGBoost for Utrecht only (orange) and for both datasets (gray).

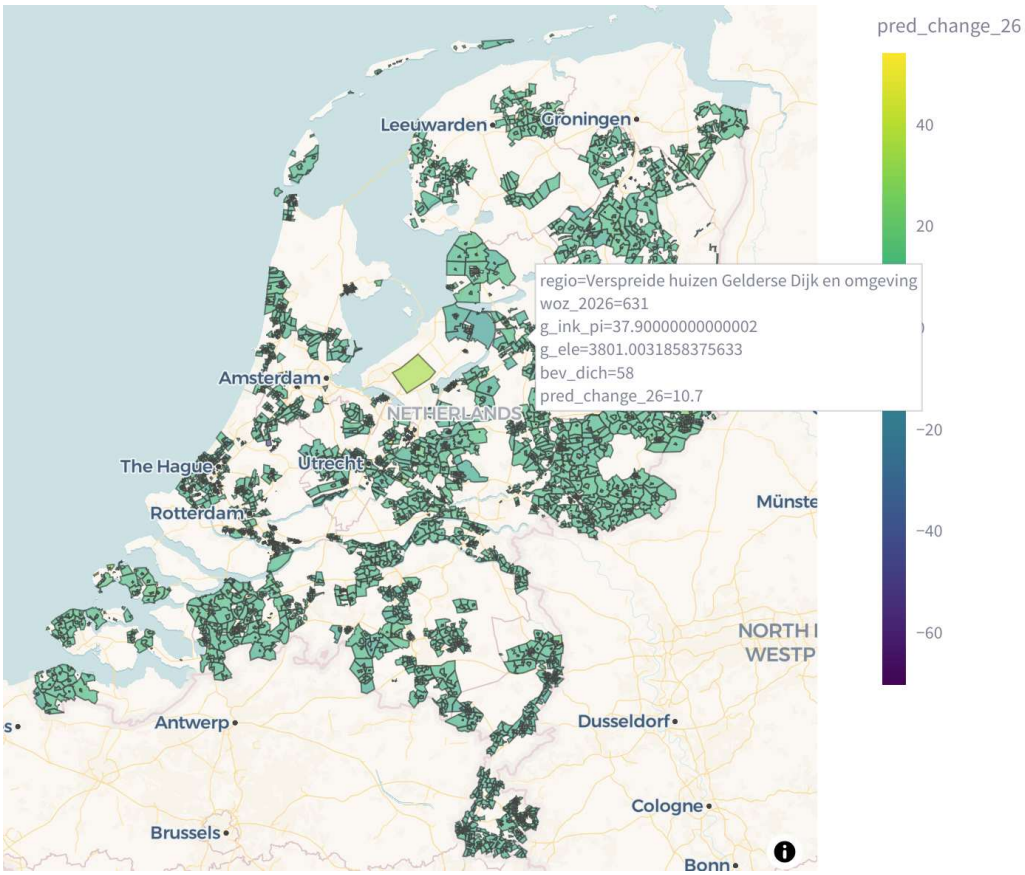

Figure 6: **Dashboard map of predicted changes in neighborhood WOZ values in the Netherlands**. Hovering over a neighborhood shows its predicted change and main predictors.

rally, through their own valuation history.

6.2 Implications for urban analytics

The findings generalize beyond housing. Many urban quantities that planners seek to predict, such as rents, land prices, population, traffic volumes or energy demand, trend over time, and tree ensembles are among the most widely used models for them. Whenever the target trends beyond the training range, such models lag the trend in proportion to its speed, and random validation does not reveal it. Five practices follow. First, models intended for forecasting should be evaluated out of time, ideally with expanding windows that expose how errors vary with market conditions (Bergmeir and Benítez, 2012; Roberts et al., 2017). Second, accuracy should be reported against naive benchmarks such as persistence and drift, together with a signed error measure such as MPE that reveals bias (Hyndman and Koehler, 2006). Third, the relation between errors and the growth rate of the target should be monitored as a routine diagnostic, and SHAP contributions of time variables offer a simple check of whether a model's level is frozen. Fourth, trending targets should be modeled explicitly, for example through growth rates, detrending with an index (de Vries et al., 2009), or hybrid models that combine trees with a time-series component. Fifth, spatial applications should report where predictions are least reliable, because compression toward the mean redistributes errors systematically across places.

6.3 Implications for policy

For municipalities, the compression of predictions toward the mean means that any use of attribute-based models in mass appraisal would overvalue the cheapest and non-urban neighborhoods and, because WOZ values determine local taxes and regulated rents, shift the tax burden toward lower-value areas unless corrected, a concern about vertical equity in property assessment. For planners and policymakers, the strong association between income composition and property values signals that rising values reinforce socio-spatial sorting, which is relevant for policies on affordable housing, mixed neighborhoods and social rental allocation (Boelhouwer, 2020). For residents, the dashboard offers a transparent overview of where valuations are expected to rise and which neighborhood characteristics accompany these increases.

6.4 Limitations and future research

Several limitations should be noted. First, 8.5% of the neighborhood data, including 3.6% of the WOZ values themselves, had to be imputed, and ten imputed WOZ values are implausibly non-positive. Because a single completed dataset is used rather than multiple imputations, imputation uncertainty is not propagated to the results (Rubin, 1987). Likewise, the hyperparameter grid search used random folds over all years, so the configuration was not selected strictly out of time. Second, neighborhoods whose boundaries changed were excluded; if boundary changes are more frequent in growing or merging areas, the panel may not be fully representative of all Dutch neighborhoods. Third, the SHAP and importance measures describe predictive associations, not causal effects, and divide credit among correlated predictors (Strobl et al., 2007; Lundberg et al., 2020); encoded region and postal codes capture location only coarsely and without spatial structure. Fourth, the property-level analysis is limited to Utrecht because of access restrictions and uses a random rather than a temporal split. Fifth, WOZ values are appraisals with a one-year valuation lag rather than transaction prices, so that the models learn the behavior of

the valuation system as much as that of the market (Geltner, 1993). Finally, forecasts assume that neighborhood characteristics remain at their latest values and inherit the momentum bias documented above.

Future research could model spatial dependence explicitly, for example with coordinates, spatial lags or spatial cross-validation (Roberts et al., 2017), test whether prediction errors cluster in space, link the property and neighborhood levels in a multilevel model, and extend the property data to other municipalities. Modeling growth rates or detrended values, adding macroeconomic variables, amenities, green space and energy labels, and comparing WOZ-based predictions with transaction prices would further test how far the findings generalize.

7 Conclusion

This study evaluated tree-based machine learning models for predicting official property valuations in the Netherlands during the post-2014 housing boom, using a national neighborhood panel and address-level data for Utrecht. XGBoost predicts neighborhood WOZ values more accurately than random forest and linear regression from neighborhood characteristics alone, with income, energy consumption, density and location dominating the predictions and errors concentrated in non-urban and low-value neighborhoods. Yet the evaluation design changes the picture: naive persistence is a demanding benchmark for appraisal-based values, property-level accuracy derives mainly from valuation history, and tree ensembles underpredict in proportion to market growth, while models with lagged values overshoot when growth slows. These results call for out-of-time evaluation, naive benchmarks and explicit modeling of trends whenever machine learning is used to predict trending urban quantities. Built entirely on public administrative data and communicated through an interactive map, the approach offers a transparent and reproducible basis for monitoring housing values and their neighborhood context.

CRediT authorship contribution statement

Wouter Westerhof: conceptualization, methodology, data collection, data curation, software, formal analysis, investigation, validation, visualization, writing – original draft. **Rafiazka Millanida Hilman:** conceptualization, methodology, resources, supervision, writing – review and editing.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

No funding was obtained for this research.

Data availability statement

The neighborhood statistics (*Kerncijfers wijken en buurten*) and WOZ valuations (*WOZ-waardeloket*) used in this study are publicly available from Statistics Netherlands and the WOZ-waardeloket. The code is available from the corresponding author upon request. The predictions can be explored in the *Residential Valuation Observatory* dashboard at `https://residential-valuation.rafiazkahilman.com`.

Declaration of generative AI use

The authors used Claude (Anthropic) to assist with language editing during the preparation of this manuscript. The authors reviewed and edited all content and take full responsibility for the research design, analysis, interpretation and final text of the manuscript.

References

L. Agnello and L. Schuknecht. Booms and busts in housing markets: Determinants and implications. *Journal of Housing Economics*, 20(3):171–190, 2011. doi: 10.1016/j.jhe.2011.04.001.

A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi. A framework for feature selection through boosting. *Expert Systems with Applications*, 187:115895, 2022. doi: 10.1016/j.eswa.2021.115895.

E. A. Antipov and E. B. Pokryshevskaya. Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2):1772–1778, 2012.

P. Bayer, F. Ferreira, and R. McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638, 2007. doi: 10.1086/522381.

C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

P. Boelhouwer. The housing market in The Netherlands as a driver for social inequalities: Proposals for reform. *International Journal of Housing Policy*, 20(3):447–456, 2020. doi: 10.1080/19491247.2019.1663056.

P. Boelhouwer and H. van der Heijden. De woningcrisis in Nederland vanuit een bestuurlijk perspectief: Achtergronden en oplossingen [The housing crisis in the Netherlands from an administrative perspective: Backgrounds and solutions]. *Bestuurskunde*, 31(1):19–33, 2022. doi: 10.5553/Bk/092733872022031001002.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

K. E. Case and R. J. Shiller. The efficiency of the market for single-family homes. *American Economic Review*, 79(1):125–137, 1989.

Centraal Bureau voor de Statistiek. Kerncijfers wijken en buurten 2004–2024 [Key figures for districts and neighborhoods 2004–2024] [dataset]. `https://www.cbs.nl/nl-nl/reeksen/publicatie/kerncijfers-wijken-en-buurten`, 2025a. Accessed February 21, 2025.

Centraal Bureau voor de Statistiek. Bevolkingsgroei [Population growth]. `https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/bevolkingsgroei/groei`, 2025b. Accessed February 21, 2025.

K. Chanasit, E. Chuangsuwanich, A. Suchato, and P. Punyabukkana. A real estate valuation model using boosted feature selection. *IEEE Access*, 9:86938–86953, 2021. doi: 10.1109/ACCESS.2021.3089198.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

V. Cohen and L. Karpavičiūtė. The analysis of the determinants of housing prices. *Independent Journal of Management & Production*, 8(1):49–63, 2017. doi: 10.14807/ijmp.v8i1.521.

P. de Vries, J. de Haan, E. van der Wal, and G. Mariën. A house price index based on the SPAR method. *Journal of Housing Economics*, 18(3):214–223, 2009. doi: 10.1016/j.jhe.2009.07.002.

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

D. Geltner. Estimating market values from appraised values without assuming an efficient market. *Journal of Real Estate Research*, 8(3):325–345, 1993.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

W. K. O. Ho, B.-S. Tang, and S. W. Wong. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1):48–70, 2021. doi: 10.1080/09599916.2020.1832558.

R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001.

P. Kangane, A. Mallya, A. Gawane, V. Joshi, and S. Gulve. Analysis of different regression models for real estate price prediction. *International Journal of Engineering Applied Sciences and Technology*, 5(11), 2021. doi: 10.33564/IJEAST.2021.v05i11.041.

H. R. A. Koster and J. Rouwendal. The impact of mixed land use on residential property values. *Journal of Regional Science*, 52(5):733–761, 2012.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

J. Manasa, R. Gupta, and N. S. Narahari. Machine learning based predicting house prices using regression techniques. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 624–630. IEEE, 2020. doi: 10.1109/ICIMIA48430.2020.9074952.

K. V. Mathotaarachchi, R. Hasan, and S. Mahmood. Advanced machine learning techniques for predictive modeling of property prices. *Information*, 15(6):295, 2024. doi: 10.3390/info15060295.

M. Mayer, S. C. Bourassa, M. Hoesli, and D. Scognamiglio. Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, 12(1):134–150, 2019.

S. Mullainathan and J. Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017. doi: 10.1257/jep.31.2.87.

NVM. Analyse woningmarkt 3e kwartaal 2024 [Housing market analysis, third quarter 2024]. Technical report, Nederlandse Coöperatieve Vereniging van Makelaars en Taxateurs in onroerende goederen NVM, 2024.

N. A. M. Pauzi, Y. B. Wah, S. M. Deni, S. K. N. A. Rahim, and Suhartono. Comparison of single and MICE imputation methods for missing values: A simulation study. *Pertanika Journal of Science and Technology*, 29(2), 2021. doi: 10.47836/pjst.29.2.15.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Plotly Technologies Inc. Plotly: Modern visualization for the web [Computer software]. `https://plotly.com/`, 2025. Accessed February 21, 2025.

D. C. Quan and J. M. Quigley. Price formation and the appraisal function in real estate markets. *Journal of Real Estate Finance and Economics*, 4(2):127–146, 1991.

D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

S. Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

G. S. Sirmans, D. A. Macpherson, and E. N. Zietz. The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1):1–44, 2005.

A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities*, 131:103941, 2022. doi: 10.1016/j.cities.2022.103941.

Streamlit Inc. Streamlit: The fastest way to build and share data apps [Computer software]. `https://streamlit.io/`, 2025. Accessed February 21, 2025.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

P. Visser, F. van Dam, and P. Hooimeijer. Residential environment and spatial variation in house prices in the Netherlands. *Tijdschrift voor Economische en Sociale Geografie*, 99(3):348–360, 2008.

WOZ-waardeloket. WOZ-waardeloket [WOZ value portal] [dataset]. `https://www.wozwaardeloket.nl/`, 2025. Accessed February 21, 2025.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Housingsupplementary.pdf