Unveiling Location-Specific Price Drivers: A Two-Stage Cluster Analysis for Interpretable House Price Predictions

Research Paper

Paul Gümmer[1,2], Julian Rosenberger[3], Mathias Kraus[3], Patrick Zschech[2,4], and Nico Hambauer[3]

[1] University of Vienna, Vienna, Austria
paulg46@univie.ac.at
[2] Leipzig University, Leipzig, Germany
[3] University of Regensburg, Regensburg, Germany
{julian.rosenberger, mathias.kraus, nico.hambauer}@ur.de
[4] TU Dresden, Dresden, Germany
patrick.zschech@tu-dresden.de

Abstract. House price valuation remains challenging due to localized market variations. Existing approaches often rely on black-box machine learning models, which lack interpretability, or simplistic methods like linear regression (LR), which fail to capture market heterogeneity. To address this, we propose a machine learning approach that applies two-stage clustering, first grouping properties based on minimal location-based features before incorporating additional features. Each cluster is then modeled using either LR or a generalized additive model (GAM), balancing predictive performance with interpretability. Constructing and evaluating our models on 43 309 German house property listings from 2023, we achieve a 36 % improvement for the GAM and 58 % for LR in mean absolute error compared to models without clustering. Additionally, graphical analyses unveil pattern shifts between clusters. These findings emphasize the importance of cluster-specific insights, enhancing interpretability and offering practical value for buyers, sellers, and real estate analysts seeking more reliable property valuations.

Keywords: House Pricing, Cluster Analysis, Interpretable Machine Learning, Location-Specific Predictions

1 Introduction

The German housing market faces unprecedented challenges in price prediction. House price predictions remain challenging due to numerous factors including market heterogeneity, location characteristics, and diverging feature effects. While machine learning promises more accurate valuations, many current approaches either rely on simple linear regression (LR) or on overly complex *black-box* algorithms that offer little interpretability, making their decision logic opaque (Lorenz et al. 2023, Janiesch et al. 2021). Modeling heterogeneous markets with either overly complex models or with

20th International Conference on Wirtschaftsinformatik
September 2025, Münster, Germany

simple linear models presents a dilemma. Practitioners must choose between highly accurate, but less explainable predictions from complex models, or more transparent but less accurate estimates from simple linear models (Hong et al. 2020, Kruschel et al. 2025).

Traditionally, machine learning approaches apply models across an entire dataset, treating it as a single problem, respectively assuming homogeneous feature effects across samples (Hong et al. 2020). We therefore refer to them as global machine learning models. On the contrary, house price prediction challenges arise from heterogeneous data representing diverse local markets. Therefore, developing methodologies that accommodate this market granularity is essential for robust predictive modeling.

Recent advances in machine learning demonstrate that clustering techniques can improve prediction performance by accounting for market heterogeneity and individuality (Azimlu et al. 2021, Hambauer et al. 2025). These approaches overcome the limitations of global models that employ one-size-fits-all methodologies. However, existing house pricing research typically implements clustering merely as a preprocessing step, subsequently compromising interpretability by employing complex models in a secondary phase (Mayer et al. 2019). What remains unexplored is an integrated approach that leverages clustering to enhance both prediction quality and model interpretability simultaneously, thus resolving the apparent trade-off between transparent and accurate models. To address this gap, we formulate the following research question:

Research Question: *"Can a clustering approach improve the predictive performance of interpretable models for house price prediction?"*

Our paper addresses this question by proposing a fine-granular interpretable cluster analysis that decomposes the complex pricing problem into smaller, manageable subproblems. Our two-stage clustering approach first groups properties by key location features, then applies refined clustering with additional property characteristics. For each cluster, we train LR and Explainable Boosting Machine (EBM) models. EBM is a generalized additive model (GAM) from the InterpretML package (Nori et al. 2019), which maintains interpretability while capturing non-linear feature effects through intuitive shape plots (Kruschel et al. 2025, Lou et al. 2013). Unlike linear approaches, EBM reveals complex data patterns, enabling detailed feature effect visualization as demonstrated in Section 5.3. We evaluate our approach on $43\,309$ German house property listings from 2023. The main contributions of this paper are:

1. A two-stage clustering approach that uses location-based and property-based features in each clustering stage respectively.
2. A comparative performance assessment of two interpretable models at cluster-level: LR and EBM.
3. Empirical evidence showing substantial performance improvements of $58\,\%$ for LR and $36\,\%$ for EBM when using the proposed two-stage clustering approach compared to global models.
4. A visual assessment and interpretation of model outputs using the EBM, which reveals cluster-specific price effects across different location clusters.

Our findings have important implications for both practice and research: practitioners gain more reliable and interpretable valuations, while researchers benefit from a new

methodological approach that bridges the gap between prediction performance and interpretability in real estate valuation. The rest of this paper is structured as follows: Section 2 outlines related work that use clustered and locally weighted approaches. Section 3 reports our experimental setup, whereas Section 4 explains our two-staged clustered analysis. Section 5 reports the results and Section 6 discusses our findings and outlines implications for researchers and practitioners. Finally, Section 7 summarizes our paper with concluding remarks.

2 Related Work

Recent years have seen a shift towards machine learning in real estate valuation (Park & Bae 2015, Ho et al. 2021). Traditional valuation methods like the comparison approach, income approach, and cost approach are often limited, especially when considering complex relationships and spatial effects in housing markets. Additionally, these methods struggle to capture non-linear feature effects and interactions between features. Such types of models, which estimate property values based on a few constituent features, are referred to as hedonic pricing models. They often suffer from functional form misspecification when the assumed relationship between features and price does not capture the true underlying dynamics (Root et al. 2023).

Earlier studies in house price predictions demonstrated that neural networks achieved higher performance than traditional regression models (Din et al. 2001, Peterson & Flanagan 2009). However, they also tend to require a larger amount of samples in the dataset and imply a more intricate model parametrization. While neural networks effectively addressed form misspecifications of hedonic pricing models, they also introduced challenges, such as overfitting and limited interpretability in identifying key price determinants (Hui & Cheung 2009). As an alternative, geographically weighted regression (GWR) was introduced to enhance traditional regression models by allowing location-specific coefficients. GWR proved particularly valuable for mass property valuation by balancing performance with interpretability (McCluskey et al. 2012). Further research compared GWR with GAMs and ordinary least squares (OLS) regression using German rental listings from 2013 to 2015 (Cajias & Ertl 2018). Their findings indicated that simpler OLS models yielded superior results, as GWR weights remained constant over time, potentially resulting in suboptimal predictions in dynamic markets, which highlights the importance of model simplicity.

Further developments showed that random forest (RF) models outperformed LR in predicting apartment prices in South Korea (Hong et al. 2020). Similarly, an evaluation of six different valuation methods using over $123\,000$ single-family home transactions in Switzerland identified Gradient Boosting as the most effective approach (Mayer et al. 2019). These studies highlighted the importance of regularization in tree-based models to avoid prediction bias and overfitting in real estate price predictions (Mullainathan & Spiess 2017). More recently, deep learning models have gained traction in property price estimation. A comparative study on deep learning applications demonstrated that transformer models and RF outperformed traditional regression-based approaches in predictive performance, while long short-term memory networks and gated recurrent units performed poorly, suggesting that not all deep learning methods are well-suited

for real estate price forecasting (Shi et al. 2023). This research highlights the potential of feature selection and dimensionality reduction in improving model performance, emphasizing the necessity of balancing complexity with interpretability.

A significant advancement came with the introduction of market segmentation before applying regression models (Azimlu et al. 2021). By first clustering properties using k-means and then applying regression models to each group, this approach achieved improved predictive performance, especially for datasets with high variance. This suggested that different market segments might require different prediction models.

Recent research has also focused on identifying key price-influencing factors. Processing location data plays a crucial role (Alfaro-Navarro et al. 2020) and environment factors such as proximity to schools, public transportation, and shopping facilities are highly predictive (Wang et al. 2021). Additionally interpretable machine learning methods demonstrated that property age and size were major factors in pricing, with feature combinations showing more impact than individual characteristics (Lorenz et al. 2023).

Next to traditional real estate features, location data, and environmental factors, alternative data sources have been explored to enhance real estate price prediction. Sun et al. (2014) proposed an integrated approach combining online news articles and web search behavior to forecast real estate price fluctuations. The authors find that incorporating search engine query data alongside sentiment analysis from news articles improved predictive performance, suggesting that consumer behavior and market sentiment are important factors in real estate valuation.

Despite these advancements, a gap remains between interpretable but less accurate linear models and complex but opaque machine learning approaches like RF. While the approach of clustering properties before applying interpretable machine learning models shows promise, this combination remains underexplored (Hong et al. 2020). It is still unclear which features substantially influence price in each cluster and whether they differ across market segments. Combining clustering with interpretable models could lead to not only more accurate predictions but also more actionable insights into specific features affecting property prices in various market segments.

3 Experimental Setup

In our experiments, data is gathered, examined, and preprocessed. After that, the data is grouped into clusters. During the data analysis phase, different clustering methods and algorithms are employed to categorize house property listings into groups that are more uniform, based on price and features. Then, the outcomes of various methods and clusters are compared to identify the characteristics of each cluster.

Subsequently, the cluster-level models undergo a training phase respectively. We employ both a LR model and EBM interchangeably. This allows us to compare these interpretable approaches in our experimental results. Since the EBM is a GAM and therefore captures feature-wise non-linear effects, it is presumed to excel in identifying more nuanced patterns for each cluster, thus enabling us to examine in more depth how these features drive the house price. Based on performance measures we comparatively assess whether and to what extent the clustering methods enhance the predictive performance

of real estate price predictions. Moreover, we combine cluster-level models in shared visualizations to compare the patterns between location clusters using the EBM.

3.1 Data Acquisition

This study uses a comprehensive dataset from the RWI-Leibniz Institute for Economic Research, entitled RED Campus dataset in its fifth version, provided by Immobilien-Scout24 for non-commercial use. The dataset includes location information mainly in the form of postal codes of the properties, detailed property characteristics, and price data. The dataset has been acquired, validated, with some preprocessing steps by the Research Data Center (FDZ) Ruhr (Schaffner & Thiel 2024).

The dataset in its fifth version is divided into two separate files: the Cross-Section dataset, which contains real estate listings from 2023 across Germany, and the Panel Campus dataset, which includes listings from the 15 largest cities over time between 2008-2023. The Cross-Section dataset differentiates between listings for house sales, apartment sales, and apartment rentals. We focus on house sale listings from the Cross-Section dataset, thereby analyzing a substantial part of the recent german house market.

3.2 Data Preprocessing

First, data cleaning was performed to avoid overfitting and skewing the analysis. This included removing samples that did not have the *price* target and removing duplicate samples that were likely re-published on the platform. To account for slight variations in price and area when re-published, we removed duplicates differing by up to $5\,\%$ in these features. We removed properties with a plot size less than 30 square meters. Moreover, samples with unrealistically high feature values were also removed manually. Further than that, no winsorization or outlier removal was necessary nor performed. The dataset resulting from acquisition and preprocessing had $43\,309$ entries.

Second, to improve the representation of geographical information, latitude and longitude features were derived from the postal codes. While postal codes provide location data, they are not always geographically sequential. By converting postal codes into latitude and longitude coordinates, we obtained continuous numerical features that more accurately reflect the properties' actual locations and distances between them.

Third, irrelevant features and those with more than $90\,\%$ missing values were removed (Schaffner & Thiel 2024). Missing values in key features were handled as follows: 1) *monthly rental income* values were set to zero, assuming the property was not rented in case no income was reported. 2) for the feature *last renovation year*, missing values were replaced by the property's construction year, indicating no renovations were performed since construction. 3) missing counts of bathrooms and bedrooms were estimated using LR based on the total number of rooms, reflecting their proportional relationship.

Fourth, feature selection was performed to reduce dimensionality and enhance model performance: An initial selection removed features irrelevant to price prediction, such as metadata about the listing (e.g., number of views). Features with high missing values that could not be imputed were excluded (Schaffner & Thiel 2024). The number of location features was reduced, retaining postal code, latitude and longitude as the primary

identifiers. The EBM model was leveraged to identify and retain the most important features influencing the property price (Lou et al. 2013). We resulted in a final set of 18 features. A list of these features will be made publicly in our online appendix.[1]

Lastly, categorical features were encoded to be suitable for our models. Binary features were encoded as 0 (No) and 1 (Yes). Ordinal features with a natural order (e.g., energy efficiency class, property condition) were label-encoded with integers reflecting their order (Müller & Guido 2016). Nominal features without a natural order (e.g., heating type, property type) were one-hot encoded to avoid introducing ordinal relationships where none exist (Müller & Guido 2016).

3.3 Evaluation and Validation

Model performance and clustering effectiveness were assessed using mean absolute error (MAE) and root mean square error (RMSE). MAE measures the average magnitude of errors in predictions, providing an intuitive understanding of predictive performance. RMSE highlights larger errors due to the squaring of residuals, offering insight into the models prediction outliers (Chai & Draxler 2014). K-fold cross-validation was employed to ensure the robustness of the models and prevent biases due to data splitting. Feature effects were visualized using the EBM model, as it reveals feature-wise non-linear feature effects that enable us to derive actionable insights. This enhances the results section by complementing model performance with model explanations in Section 5.3.

4 Method

The novelty of our method lies in the application of a granular two-stage clustering approach, followed by the use of interpretable models at the cluster-level. The first clustering stage uses a small feature set, while the second clustering stage uses the entire feature set. In the first stage, mainly latitude, and longitude are used to group geographically proximate houses with similar price levels combined with relying on information from the *price* target. In the second stage, a more nuanced clustering approach with a broader feature set is employed. In the final prediction stage, our method proposes to make use of simple models, namely LR and EBM, to make house price predictions more interpretable. We make our implementation freely available for reproduction and application in other contexts.[2]

For clustering, we use regression trees, k-means, and k-nearest neighbors (KNN), which are among the most frequently applied clustering approaches (Hambauer et al. 2025). Regression trees split the feature space via thresholding and assign samples to their cluster via leaf-modeling (De Caigny et al. 2018). K-means and KNN identify k centroids, by searching for dense data regions in the feature space (Friedman et al. 1975).

For the first stage we chose k-means as it revealed the best results for $k = 2$. For KNN and regression trees we would only indirectly set a maximum cluster size through

[1] https://osf.io/jh2kb/?view

[2] https://gitlab.com/house-price-predicition/cluster-analysis-for-interpretable-house-price-predictions

other hyperparameters. Therefore, KNN and regression trees produced a larger number of clusters, which made us disregard them for the first stage. We deliberately opted for only two clusters in this initial clustering stage, as each additional cluster would have increased the number of sub-clusters by a factor of eight, which would reduce interpretability and human comprehensibility.

In the second stage, we compared k-means with a regression tree. Guided by the regression tree, we fixed the tree depth at three, which yields eight terminal leaves. A depth of two produced clusters that were almost indistinguishable, whereas a depth of four generated 16 leaves, many of them contained too few observations to build stable cluster-specific models. To ensure a fair comparison, we therefore also set $k = 8$ for the k-means variant. We used the Elbow method to validate that, for the set of features, 8 clusters are feasible (see online appendix).[3] The plot shows diminishing returns starting around 6-8 clusters, with 8 clusters providing a good balance between model complexity and within-cluster variance reduction. Our guiding principle throughout was to limit the number of clusters to a scale that remains interpretable for human analysts while still capturing the key heterogeneity in the data.

Based on their interpretability, two models were selected: A simple LR using an L1 penalty and a slightly more complex yet interpretable non-linear GAM, specifically the EBM (Lou et al. 2013). LR performs feature selection by penalizing the absolute size of the coefficients using an L1 penalty, effectively reducing less important feature weights to zero and preventing overfitting (Tibshirani 1996). The EBM is an advanced non-linear model that combines predictive power with inherently interpretability to capture complex relationships without becoming a *black-box* and demonstrated superior performance over traditional models for various problems (Kruschel et al. 2025).

5 Results

In this section, we systematically present our results. First, we compare the predictive performance of a global model without clustering against three different clustering approaches. These three approaches include a simple approach using location-based k-means clustering, and two variations of repeated two-stage clustering with variants of both k-means and regression trees.

5.1 Assessment of Predictive Performance

Table 1 illustrates the comparative assessment of predictive performance for both LR and the EBM. All performance measures, respectively errors are in Euros (€). In all four approaches, the EBM achieves better performance than LR. Nevertheless, LR as part of various clustered approaches shows similar performance compared to the EBM while still being intrinsically easier to interpret, due to the small set of coefficients. Moreover, both prediction models perform substantially better once the data is clustered.

Location-based clustering brings the largest relative performance difference looking at LR without clustering (170 745 MAE) vs. location clustering (127 272 MAE) and

[3] `https://osf.io/jh2kb/?view`

Table 1. Assessment of predictive performance for clustering approaches under mean absolute error (MAE) and root mean squared error (RMSE) in Euros. Best approach is underlined.

Model	Approach	Clustering Phases		Average MAE (in €)	Average RMSE (in €)
		1st Phase	2nd Phase		
EBM	No Clustering	-	-	110 387	165 509
	Location clustering	k-means	-	87 276	115 627
	Two-stage clustering	k-means	regression tree	88 696	115 291
	Two-stage clustering	k-means	k-means	<u>80 925</u>	<u>104 189</u>
LR	No Clustering	-	-	170 745	244 552
	Location Clustering	k-means	-	127 272	160 076
	Two-stage clustering	k-means	regression tree	122 295	151 510
	Two-stage clustering	k-means	k-means	107 996	134 297

EBM without clustering (110 387 MAE) vs. location clustering (87 276 MAE). For the second stage of clustering which goes beyond just location, by using various features for clustering, applying k-means twice in both clustering stages is the most effective for both models. This further improved the prediction performance comparing LR with only location clustering (127 272 MAE) vs. clustering twice (104 189 MAE) and EBM with only location clustering (87 276 MAE) vs. clustering twice (80 925 MAE). Overall the two-stage approach improves predictive performance by approximately 36%, looking at EBM without clustering (110 387 MAE) and two-stage clustering with k-means (num80 925 MAE) as well as approximately 58% looking at LR without clustering (170 745 MAE) vs. two-stage clustering with k-means (107 996 MAE).

5.2 Visualizations of Location-based Assessment

Location-based clustering proved to have a large impact on the predictive performance. In the previous analysis we relied on k-means, while alternative clustering approaches would also be possible. Therefore, we applied the location-based clustering approach using three different algorithms on a limited set of features and visualize the results in Figure 1 on a map of Germany.

The left part of Figure 1 visualizes the clustering result obtained by using k-means. A K-Nearest-Neighbor (KNN) regressor is used to predict property prices based on the location, which are then grouped into two clusters using binning (middle part of Figure 1). The right part of Figure 1 visualizes the clusters provided by a pruned regression tree with a maximum depth of fifteen aggregated by setting a threshold for the leaf nodes. The differences between the outcome of this analysis become especially obvious with the k-means algorithm, which distinctly identifies major metropolitan areas such as Hamburg, Berlin, Munich, and Frankfurt am Main. Based on the graphical representation, k-means provides the best separation between more expensive and less expensive properties when only two clusters are used.

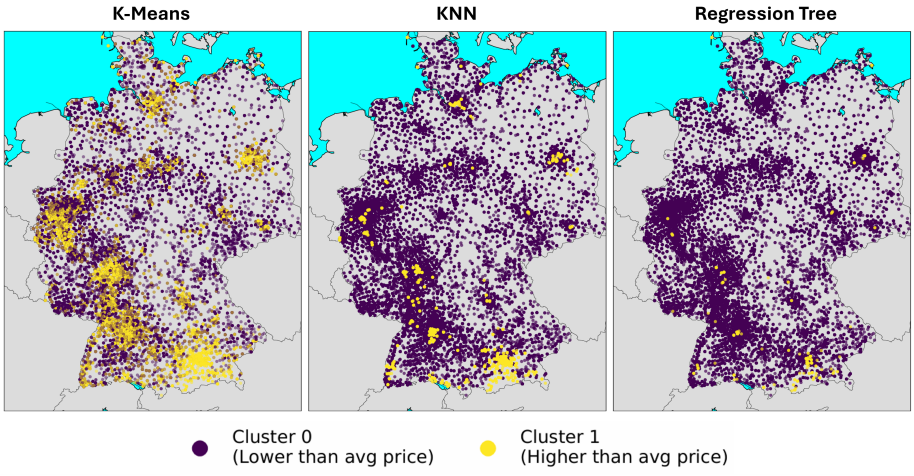

Figure 1. Results of the location clusters for different algorithms

5.3 Assessment of Cluster-specific Price Effects

The application of location-based clustering effectively reduced the dominant influence of location in the subsequent sub-clustering step, allowing other property characteristics to become more prominent in the modeling phase using LR and EBM. An analysis of the feature relationships learned by the EBM model reveals that the same feature can have different effects on price depending on the respective location cluster and sub-clusters, underscoring the heterogeneity of market segments.

Figure 2 illustrates the influence of *construction year* on property prices in two distinct sub-clusters. Notably, the different sub-clusters have distinct feature ranges due to the clustering. The line in purple representing sub-cluster 4 begins at a later *construction year* since that cluster only includes samples that have the feature *construction year* around that feature range. The cluster-specific EBM model learns a different relationship for properties in sub-cluster 4 compared to sub-cluster 5. For instance, at *construction year* 1950, there's approximately a $70\,000$€ difference in feature effect between the two sub-clusters, with sub-cluster 4 showing a negative effect on price while sub-cluster 5 demonstrates a positive effect in that area. Moreover, comparing sub-cluster 5 to sub-cluster 4 in general, we can see that in sub-cluster 5 older houses have a high negative influence of approximately $62\,500$ between 1720 until approximately 1850, after which the influence increases continuously up to the year 2000, reaching a positive feature effect by roughly 1930. In sub-cluster 4 however, houses built around 1900 start with having a positive influence, which decreases until about 1975 and then rises again. These patterns suggest a specific preference for historical, more luxurious properties in sub-cluster 4, while sub-cluster 5 could include more standardized houses.

Figure 3 depicts the effect of *living space* on house property prices. In sub-cluster 2 prices rise with increasing square meters up to a peak at approximately 175 square meters, remaining roughly constant afterwards. In contrast, in sub-cluster 5 the price

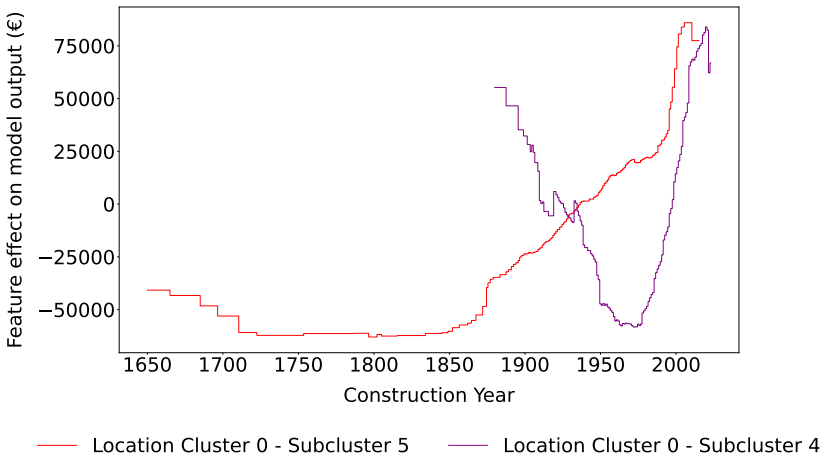

Figure 2. Feature effect of construction year on the price using the EBM model based on two distinct clusters respectively.

increases slowly, but almost linearly in relation to *living space*. This could indicate that in sub-cluster 2, demand for very large properties tapers off, whereas in sub-cluster 5, demand remains stable even for larger properties.

These varying feature-price relationships across sub-clusters highlight the advantages of our two-staged clustering approach with interpretable models at cluster-level. It allows for a more nuanced analysis of cluster-specific trends and leads to a marked improvement in predictive performance.

6 Discussion

6.1 Implications

This study makes several key contributions to the field of real estate price prediction by addressing the presumed trade-off between model interpretability and predictive performance through a novel two-stage clustering approach. Researchers and practitioners might benefit in various ways. While existing research often applies clustering as a preprocessing step without leveraging its interpretability benefits, our approach integrates clustering directly into the modeling process, enhancing both prediction performance and explainability.

First, we introduce a hierarchical clustering framework that groups properties based on location-specific features before refining clusters using property-specific attributes. This segmentation improves market differentiation and enhances predictive performance.

Second, we evaluate LR and EBM across these clusters, demonstrating that two-stage clustering improves predictive performance by up to 58% for LR and 36% for EBM compared to non-clustered approaches.

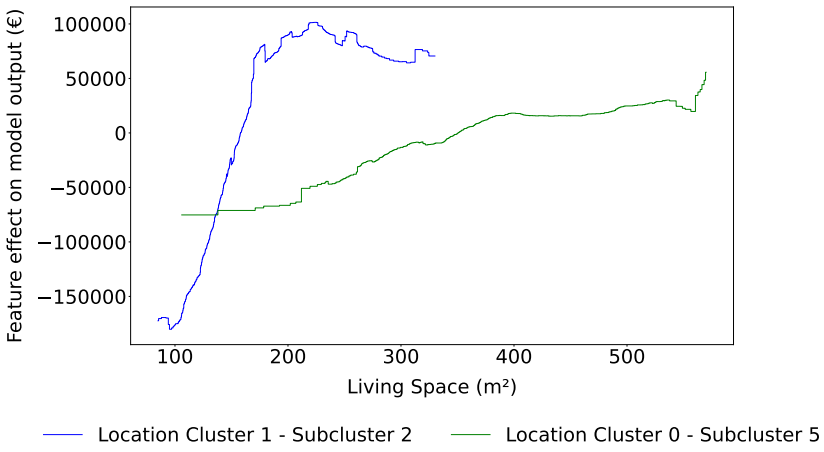

Figure 3. Feature effect of living space on the price using the EBM model based on two distinct clusters respectively.

Third, we provide cluster-specific insights into price drivers, revealing variations in the influence of factors such as living space and construction year. This enhances model transparency and supports better decision-making for real estate professionals by providing actionable insights.

Finally, our work contributes to the discussion on interpretable machine learning, bridging the gap between high-performance but black-box models and transparent but less predictive approaches (Kruschel et al. 2025). Our findings highlight the potential of segmentation-driven modeling for more accurate and actionable property valuations.

6.2 Limitations & Future Research

While our proposal of combining clustering techniques and predictive models for house prices yielded promising results, several inherent limitations affect the predictive quality and reliability.

A prevailing challenge is forecasting prices for luxury properties, which exhibit high heterogeneity and unique features often absent from structured datasets. Features such as aesthetic design elements, special features, or high-end renovations are not adequately captured in tabular data, being only discernible in textual descriptions or images. Additionally, emotional factors and personal buyer preferences play an important role in real estate (Ben-Shahar & Golan 2014), further complicating price predictions. Future research could explore multi-modal inputs incorporating text and image data, and engage with real estate professionals to better understand feature importance and validate the practical utility of model insights.

Moreover, we focused on interpretable models rather than black-box regressors such as RF or XGBoost, aiming to derive inherently interpretable explanations without relying on post-hoc explainers (Rudin 2019). Future work could investigate model-agnostic interpretability techniques like SHAP (Lundberg et al. 2020) or LIME (Ribeiro et al. n.d.)

when using black-box models, and conduct comprehensive comparisons with other state-of-the-art methods including advanced ensemble methods and deep learning approaches to better position our methodology within the broader predictive modeling landscape.

Another limitation is the handling of outliers and our location-driven clustering approach. Some properties received exceptionally high valuations despite features suggesting lower prices, indicating that critical attributes may not have been fully captured. Additionally, our clustering was primarily driven by location as the initial splitting criterion, and future research could investigate whether alternative features might be more suitable as the primary clustering dimension.

In addition, our dataset is based on online house property listings, which introduces potential biases as it remains unclear whether listed prices reflect actual transaction prices or include overpriced properties that remained unsold. Access to finalized transaction data would enhance prediction reliability.

Beyond real estate prediction, developing systematic methodologies for determining optimal clustering characteristics and model parameters across different domains represents an important avenue for future research.

7 Conclusion

In this study, we applied a two-stage clustering approach to house property listings across Germany for the year 2023. The first clustering stage grouped properties exclusively by location, opting for k-means. In the second stage, the clusters were further divided based on additional property features using k-means and regression trees.

Our findings demonstrate that clustering improves house price prediction while providing deeper insights into market segments across different locations and property types. The interpretability of clustered models revealed heterogeneous pricing mechanisms, showing that feature impacts on property prices vary across market segments. This segmentation enhances market transparency and benefits buyers, sellers, and real estate professionals by enabling more targeted valuation approaches.

A two-stage k-means clustering approach combined with EBM achieved the best predictive performance, improving MAE by up to 36% over non-clustered models. When paired with linear regression, the clustering methodology yielded even greater improvements of approximately 58%.

Despite clustering improvements, limitations remain, particularly reduced predictive performance for higher-priced properties due to the greater complexity and heterogeneity of luxury real estate. Future studies could integrate image-based features or textual descriptions to enhance predictions for these high-value properties.

Overall, our study demonstrates that a two-stage clustering approach significantly enhances real estate price prediction performance. By grouping properties into homogeneous categories, this methodology enables targeted feature analysis and yields more precise, transparent valuations. This approach provides valuable insights for real estate professionals, investors, and policymakers through a data-driven pricing framework. Our findings confirm geographical location as the most critical price determinant, while interpretable models ensure transparency and practical applicability, offering clear advantages over black-box approaches where explainability is essential.

References

Alfaro-Navarro, J.-L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M. & Larraz, B. (2020), 'A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems', *Complexity* **2020**(1), 5287263.

Azimlu, F., Rahnamayan, S. & Makrehchi, M. (2021), House price prediction using clustering and genetic programming along with conducting a comparative study, *in* 'Proceedings of the Genetic and Evolutionary Computation Conference Companion', pp. 1809–1816.

Ben-Shahar, D. & Golan, R. (2014), 'Real estate and personality', *Journal of Behavioral and Experimental Economics* **53**, 111–119.

Cajias, M. & Ertl, S. (2018), 'Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions?', *Journal of Property Investment & Finance* **36**(1), 32–49.

Chai, T. & Draxler, R. R. (2014), 'Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding rmse in the literature', *Geoscientific model development* **7**(3), 1247–1250.

De Caigny, A., Coussement, K. & De Bock, K. W. (2018), 'A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees', *European Journal of Operational Research* **269**(2), 760–772.

Din, A., Hoesli, M. & Bender, A. (2001), 'Environmental variables and real estate prices', *Urban Studies* **38**(11), 1989–2000.

Friedman, J. H., Baskett, F. & Shustek, L. J. (1975), 'An algorithm for finding nearest neighbors', *IEEE Transactions on Computers* **100**(10), 1000–1006.

Hambauer, N., Stoecker, T., Zschech, P. & Kraus, M. (2025), Benchmarking Cluster-Then-Predict Models to Challenge Prevailing Global Machine Learning Models, *in* 'Proceedings of the 58th Hawaii International Conference on System Sciences'.

Ho, W. K., Tang, B.-S. & Wong, S. W. (2021), 'Predicting property prices with machine learning algorithms', *Journal of Property Research* **38**(1), 48–70.

Hong, J., Choi, H. & Kim, W.-s. (2020), 'A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea', *International Journal of Strategic Property Management* **24**(3), 140–152.

Hui, Y. & Cheung, S. (2009), Property price modelling: The regression model and the neural newtwork model, *in* 'ICEB 2009 Proceedings'.

Janiesch, C., Zschech, P. & Heinrich, K. (2021), 'Machine learning and deep learning', *Electronic Markets* **31**(3), 685–695.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M. & Zschech, P. (2025), 'Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models', *Business & Information Systems Engineering* pp. 1–25.

Lorenz, F., Willwersch, J., Cajias, M. & Fuerst, F. (2023), 'Interpretable machine learning for real estate market analysis', *Real estate economics* **51**(5), 1178–1208.

Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. (2013), Accurate intelligible models with pairwise interactions, *in* 'Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 623–631.

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

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

McCluskey, W., Davis, P., Haran, M., McCord, M. & McIlhatton, D. (2012), 'The potential of artificial neural networks in mass appraisal: the case revisited', *Journal of Financial Management of Property and Construction* **17**(3), 274–292.

Mullainathan, S. & Spiess, J. (2017), 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

Müller, A. C. & Guido, S. (2016), *Introduction to machine learning with Python: a guide for data scientists*, " O'Reilly Media, Inc.".

Nori, H., Jenkins, S., Koch, P. & Caruana, R. (2019), 'Interpretml: A unified framework for machine learning interpretability', *arXiv preprint arXiv:1909.09223* .

Park, B. & Bae, J. K. (2015), 'Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data', *Expert systems with applications* **42**(6), 2928–2934.

Peterson, S. & Flanagan, A. (2009), 'Neural network hedonic pricing models in mass real estate appraisal', *Journal of Real Estate Research* **31**(2), 147–164.

Ribeiro, M. T., Singh, S. & Guestrin, C. (n.d.), " why should i trust you?" explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 1135–1144.

Root, T. H., Strader, T. J. & Huang, Y.-H. J. (2023), 'A review of machine learning approaches for real estate valuation', *Journal of the Midwest Association for Information Systems* **2023**(2), 2.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature machine intelligence* **1**(5), 206–215.

Schaffner, S. & Thiel, P. (2024), Fdz data description: Real-estate data for germany campus files (rwi-geo-red panel and rwi-geo-red cross v5)-advertisements on the internet platform immobilienscout24 for teaching purposes, Technical report, RWI Datenbeschreibung.

Shi, D., Zhang, H., Guan, J., Zurada, J., Chen, Z. & Li, X. (2023), Deep learning in predicting real estate property prices: A comparative study, *in* 'Proceedings of the 56th Hawaii International Conference on System Sciences'.

Sun, D., Du, Y., Xu, W., Zuo, M. Y., Zhang, C. & Zhou, J. (2014), 'Combining online news articles and web search to predict the fluctuation of real estate market in big data context', *Pacific Asia Journal of the Association for Information Systems* **6**(4), 2.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Wang, P.-Y., Chen, C.-T., Su, J.-W., Wang, T.-Y. & Huang, S.-H. (2021), 'Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism', *IEEE access* **9**, 55244–55259.