University of Milan-Bicocca
CdLM Data Science

14 January 2026
Machine Learning Project

House price prediction with regression models

Lorenzo Lecci
matr. 881473
l.lecci@campus.unimib.it

Diana Carbone Radchenko
matr. 857792
d.carboneradchenko@campus.unimib.it

Martina Tropeano
matr. 945224
m.tropeano1@campus.unimib.it

Abstract

Real estate valuation has traditionally relied heavily on subjective expertise, a reliance that often leads to pricing inconsistencies and variance. To address this challenge, this report details the development of a data-driven price estimation tool utilizing the King County House Sales (USA) dataset. The study pursues the dual objective of constructing a predictive model and interpreting the primary determinants of market value. The workflow, implemented in KNIME, includes features advanced preprocessing and spatial feature engineering, specifically the application of K-Means clustering ($K = 15$) to synthesize high-cardinality ZIP codes into homogeneous neighborhoods. A comparative analysis was conducted between a parametric Linear Regression baseline and non-parametric ensemble methods (Random Forest and XGBoost) highlighting distinct preprocessing strategies ranging from logarithmic transformations to the utilization of raw spatial coordinates. Quantitative results designate the XGBoost algorithm, trained on the full feature set, as the best model: it achieved a Test R^2 of 0.901, successfully breaching the 0.90 threshold. Feature importance analysis across all architectures reveals that Structural Magnitude, Construction Quality and Precise Location constitute the three fundamental pillars defining the non-linear function of real estate prices in King County.

Keywords: *Machine Learning, Real Estate Valuation, K-Means Clustering, Random Forest, XGBoost Tree Enseble, Spatial Feature Engineering, KNIME.*

1 Introduction

The real estate market is driven by complex structural and locational factors, however traditional valuation remains susceptible to human bias and subjectivity. This research aims to support a hypothetical real estate agency in developing a data-driven price estimation tool to standardize assessments and minimize portfolio variance. The study pursues two primary objectives: developing a robust machine learning model to predict market value based on objective data and conducting a feature importance analysis to identify the key drivers of property value. A major challenge lies in effectively encoding location, particularly for the parametric baseline. Standard ZIP codes were deemed unsuitable for Linear Regression, as One-Hot Encoding the 70 unique identifiers would have introduced excessive dimensionality. Similarly, raw latitude and longitude coordinates were excluded from this model, as preliminary analysis revealed no linear relationship with the target variable. To address these limitations, we implemented K-Means clustering ($K = 15$) to model local sub-markets. This step was essential to provide the linear model with a usable spatial structure. The analytical workflow was implemented in KNIME, prioritizing the use of native nodes to maintain a consistent low-code architecture. However, to address specific limitations in KNIME's standard regression reporting and plotting capabilities, Python View and Python

Script nodes were strategically integrated to enable advanced visualization and granular statistical diagnostics. This framework compares a baseline Multiple Linear Regression (reliant on spatial clusters and log-transformations) against two non-parametric ensemble methods: Random Forest and XGBoost. In distinct contrast, these tree-based algorithms were trained on minimally preprocessed data, leveraging raw coordinates to capture non-linearities. Finally, Backward Feature Elimination was applied across all models to optimize the balance between complexity and predictive accuracy.

The remainder of this report is structured as follows: **Section 2** explores the dataset and **Section 3** describes the preprocessing steps and the spatial clustering methodology. The subsequent sections detail the implementation and optimization of the specific models: **Section 4** focuses on Linear Regression, **Section 5** on Random Forest and **Section 6** on XGBoost Tree Ensemble. **Section 7** synthesizes the findings regarding Feature Importance across all models and finally **Section 8** draws conclusions and outlines future work regarding stratified market segmentation.

2 Data Exploration

The empirical analysis presented in this study is grounded in the ”House Sales in King County, USA” dataset, retrieved from the Kaggle repository, which encompasses housing transactions recorded between May 2014 and May 2015. The dataset comprises 21.613 observations and 21 features: the target variable is *price*, while *id* serves as a unique identifier for each transaction. To facilitate a structured evaluation, the independent variables were conceptually categorized into three domains:

- **Structural Features:** includes physical dimensions (*bedrooms*, *bathrooms*, *floors*), area measurements (*sqft_living*, *sqft_lot*, *sqft_above*, *sqft_basement*), quality indices (*grade*, *condition*) and neighborhood context metrics (*sqft_living15*, *sqft_lot15*).
- **Spatial Features:** defines location via coordinates (*lat*, *long*), *zipcode* and positional amenities such as *waterfront* (binary indicator) and *view* (visual quality index).
- **Temporal Features:** records the asset's timeline, specifically the transaction *date*, original construction year (*yr_built*) and last renovation year (*yr_renovated*).

Central to this investigation is the target variable **price**. An examination of its probability density function, as illustrated in Figure 1, reveals that the distribution is heavily right-skewed. The majority of transactions are concentrated in the lower-to-mid price range, with a long tail extending towards high-value luxury properties.

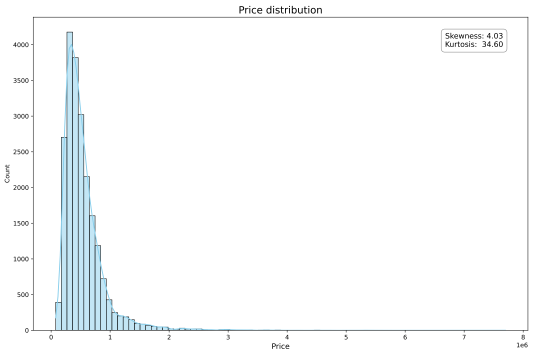


Figure 1: Histogram of price distribution

This marked deviation from normality presents a significant challenge for linear parametric models, which typically rely on assumptions of residual normality and homoscedasticity. Consequently, this observation motivates the application of a logarithmic transformation[1] during the preprocessing stage to approximate a Gaussian distribution and stabilize variance.

Subsequent multivariate analysis focused on identifying primary value drivers and potential redundancies. A Pearson correlation matrix (Figure A.1) indicates that *sqft_living* and construction *grade* exhibit the strongest positive correlation with the target variable. The analysis also highlights significant multicollinearity among predictor variables: specifically, *sqft_living* is highly correlated with *sqft_above* and *sqft_living15*. It was observed also that *sqft_living* is the sum of *sqft_above* and *sqft_basement*.

Finally, the geographical distribution of prices was visualized to assess the impact of location (Figure 2). The data confirms that high-value properties are spatially clustered, particularly in northern sectors and waterfront areas. Nevertheless, this spatial dependency is non-linear and irregular. For a linear regression model, raw coordinates (*lat*, *long*) fail to capture neighborhood boundaries effectively, while administrative *zipcodes* introduce excessive cardinality without strictly aligning with economic zones. This limitation underscores the necessity for advanced spatial feature engineering, providing the rationale for the K-Means clustering methodology detailed in the following section.

[1] In particular $log(x + 1)$ transformation was applied to all skewed features to handle zero values.

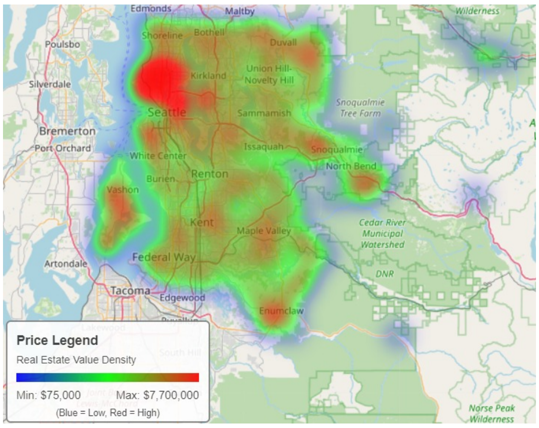


Figure 2: Geographic Heatmap

3 Data Preprocessing and Clustering

The data preparation phase was designed with a dual pipeline approach to accommodate the distinct requirements of the algorithms employed. While the non-parametric models (Random Forest and XGBoost) are intrinsically robust to skewed distributions and capable of handling raw spatial data, the Linear Regression model requires strict compliance to statistical assumptions. Consequently, the transformations detailed in this section were implemented to optimize the Linear Regression model, whereas the tree-based ensembles were trained on a minimally processed version of the dataset.

Data Cleaning and Distributional Transformations

The integrity of the dataset was first assessed through a systematic inspection, from which emerged the absence of missing values. Regarding outlier management, a strategic decision was made to deviate from standard noise reduction techniques: while clear structural errors were removed (e.g. an anomalous 33-bedroom property), high-leverage observations representing luxury estates were consciously retained. This approach ensures the model remains generalizable across the entire market spectrum, including the high-end segment.

To enhance informational content, specific feature engineering steps were executed. The attributes *yr_built* and *yr_renovated* were synthesized into a new $House\ Age = Year - yr\_built$ metric and a binary *is_renovated* indicator captured the property's effective lifespan and modernization status. At the same time, temporal granularity was refined by extracting numerical *Year* and *Month (Number)* values, while non-informative identifiers such as the transaction *ID* were discarded.

A necessary step involved addressing the distribution of the data: as highlighted in the exploratory analysis (Figure 1), key variables exhibited high skewness. To satisfy the assumption of residual normality and homoscedasticity of the linear model, logarithmic transformations were applied to the target variable (*price*) and to skewed structural features, including *sqft_living*, *sqft_basement sqft_lot*, *sqft_above* and *sqft_living15*. Post-transformation statistical checks confirmed that the skewness of these variables was successfully reduced to near-zero values.

Finally, the categorical variable *zipcode* posed a dimensionality challenge for the linear model due to its high cardinality (70 unique values). Since converting it to numerical values would imply a false ordinal relationship and one-hot encoding would excessively increase dimensionality, the *zipcode* feature was removed for the Linear Regression baseline. It was effectively replaced by the spatial clustering derived in the following subsection.

Spatial Feature Engineering via K-Means Clustering

Standard coordinate data (Latitude and Longitude) presents a challenge for linear regression, which assumes linear relationships between predictors and the target. However, real estate prices follow nonlinear spatial patterns, fluctuating irregularly across neighborhoods. To address this, a K-Means clustering algorithm was implemented to partition the county into K distinct economic sub-markets.

Since K-Means relies on Euclidean distance, it is highly sensitive to feature scaling. Therefore, prior to clustering, the input variables (*lat*, *long*) underwent Z-score normalization to ensure that both dimensions contributed equally to the distance calculations. Then the determination of the optimal number of clusters, K, followed a multi-stage validation process testing values from 5 to 20. First, the Silhouette analysis was conducted.

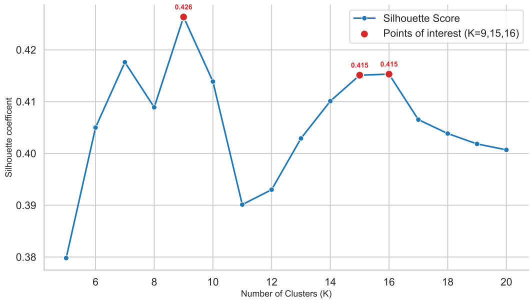


Figure 3: Silhouette Plot

As shown in Figure 3, while a global maximum was observed at $K = 9$ (≈ 0.426), the values $K = 15$ and $K = 16$ emerged as local maxima with comparable efficient (≈ 0.415). Although $K = 7$ presented a higher coefficient than the latter candidates, it was discarded as insufficient to capture the fine-grained variations of such a vast geographic area. To resolve the selection, an internal validation of the three candidates was performed by calculating Cohesion and Separation metrics.

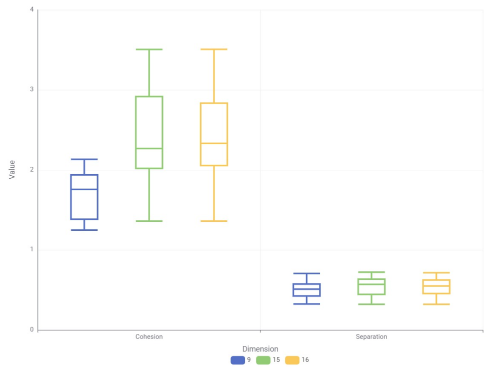


Figure 4: Cohesion and Separation Plot

The analysis (Figure 4) led to the exclusion of $K = 9$ due to suboptimal cluster compactness. With $K = 15$ and $K = 16$ showing similar internal performance, a final relative validation was conducted using the Calinski-Harabasz, Davies-Bouldin and Dunn indices (shown in figure A.2). The results favored $K = 15$, particularly as the Dunn Index showed a sharp decline for values greater than $K = 16$. So each property was assigned a *Cluster* from 0 to 14 (Figure A.3).

Then the variable *Cluster* was transformed into binary dummy variables (One-Hot Encoding) to prevent the model from inferring an incorrect ordinal hierarchy among neighborhoods and the original raw coordinates (*lat*, *long*) were removed in order to prevent multicollinearity.

4 Linear Regression

This section details the development of the linear regression model. The focus is placed on the selection of predictors to ensure model parsimony and the diagnostic evaluation of the model.

Feature Selection and Multicollinearity Resolution

To construct a robust and interpretable model, a Backward Feature Elimination strategy was applied to the Training Set (70% of the data). This iterative process was wrapped within a 5-Fold Cross-Validation loop to ensure that the selection was not biased by a specific data partition. The optimization objective was the maximization of the coefficient of determination (R^2). The automatic selection process demonstrated that reducing the dimensionality from the full feature set down to 20 variables resulted in a negligible degradation of predictive power, with the R^2 decreasing only marginally from 0.838 to 0.8298, while significantly reducing model complexity.

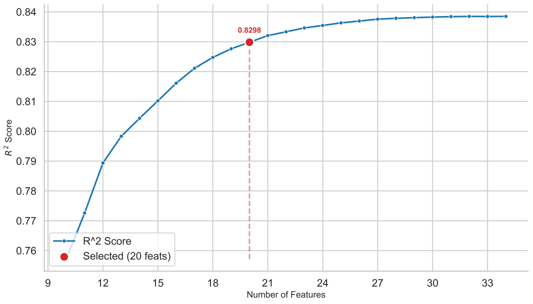


Figure 5: Feature Selection R2 Curve

As illustrated in the figure 5, the curve exhibits a point of diminishing returns at 20 features: while adding more variables continues to increase R^2, the rate of improvement slows down significantly. However, subsequent manual inspection of the Variance Inflation Factor (VIF) revealed a critical issue of multicollinearity between the log-transformed variables *sqft_living_log* and *sqft_above_log*, with VIF values exceeding the threshold of 5. Mathematically, the living area is the sum of the area above ground and the basement: thus, retaining both introduces redundancy. To resolve this, an A/B test was conducted. The removal of *sqft_above_log* led to superior model stability (reducing all VIF values below 3) and a final R^2 of 0.825, outperforming the alternative configuration. Consequently, the final model was built on an optimized set of 19 features.

Model Training and Diagnostics

Prior to training, the 19 selected features underwent Z-Score normalization to standardize their scales. Significantly, the normalization parameters (mean and standard deviation) were computed exclusively on the Training Set and subsequently applied to the Test Set. This strict separation was enforced to prevent data leakage, ensuring that the model's performance metrics reflect a true out-of-sample evaluation. The linear regressor was then trained on the processed data.

The model demonstrated high consistency between the training and testing phases. On the Training Set, the model achieved an $R^2 \approx 0.825$ with a Mean Absolute Error (MAE) of 0.316. Similarly, on the Test Set, the model resulted in an $R^2 \approx 0.826$ and an MAE of 0.315. The virtual identity of these metrics indicates that the model has generalized well to unseen data, with no evidence of overfitting.

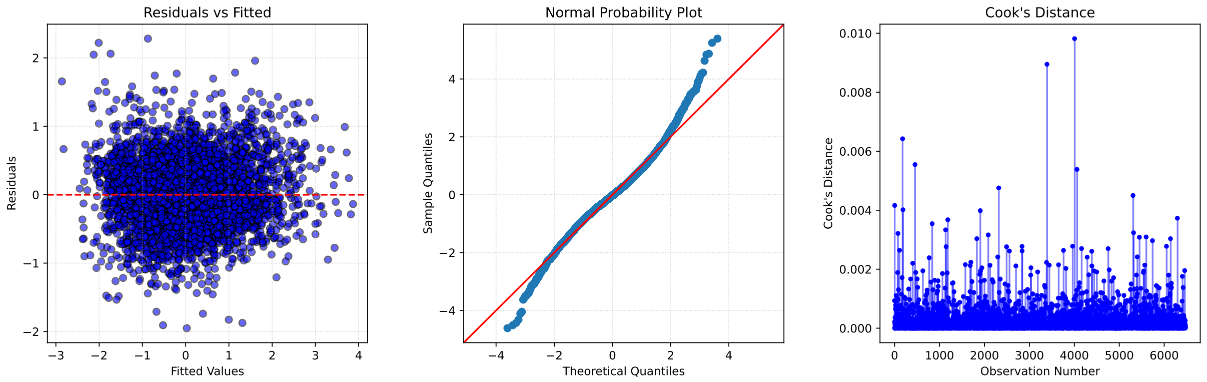


Figure 6: Diagnostic Plots

Index	Train	Test
R^2	0.825	0.826
Adj. R^2	0.825	0.826
MAE (log)	0.317	0.315
RMSE (log)	0.419	0.424

Table 1: Model Performance Comparison

The diagnostic evaluation initiates with an inspection of the Actual vs. Predicted scatter plot. As illustrated in the figure 7, the observed data points cluster tightly along the identity line. This strong linear alignment confirms the efficacy of the logarithmic transformation applied to the target variable, demonstrating that the model effectively captures the underlying market trends without exhibiting significant systematic bias or curvature.

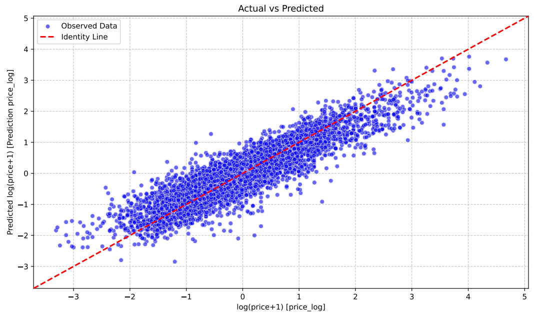


Figure 7: Actual vs Residuals

The validity of the model assumptions was verified through visual inspection of the diagnostic plots presented in Figure 6. The Residuals vs Fitted plot displays a random dispersion of points around the horizontal zero axis, devoid of distinct non-linear patterns or funnel shapes. This uniform spread supports the assumption of homoscedasticity, indicating that the variance of the error terms remains constant across the range of predicted values.

At the same time, the Normal Probability Plot confirms that the residuals closely follow a Gaussian distribution. The data points align tightly with the theoretical reference line, particularly in the central quantiles. Although a slight deviation is observable at the extreme tails (due to the presence of luxury or distressed properties), the overall distribution satisfies the normality assumption required for reliable statistical inference.

Finally, the Cook's Distance analysis demonstrates model robustness: the calculated values are uniformly low, indicating that no single observation exerts a disproportionate influence on the regression coefficients, effectively ruling out the presence of damaging leverage points.

To conclude the diagnostic assessment, the Durbin-Watson statistic was calculated at 1.9927: since this value is virtually indistinguishable from 2, it indicates a complete absence of autocorrelation among residuals, thereby satisfying the assumption of independence of error terms.

Cross-Validation and Robustness Check

To further validate the stability of the model, a 10-Fold Cross-Validation was performed on the Training Set. In each iteration, the data was re-partitioned and the Z-Score normalization was reapplied dynamically within the fold to ensure a rigorous evaluation.

The results demonstrated remarkable stability, with the R^2 scores across the 10 folds exhibiting very low variance, oscillating tightly between 0.81 and 0.83. This consistency confirms that the model's predictive capability is structural and robust, providing a reliable baseline for price estimation against which non-parametric models can be compared.

5 Random Forest

The investigation progressed to non-parametric ensemble methods, focusing on the Random Forest algorithm. A key advantage of this approach is its inherent ability to model complex and non-linear relationships with minimal data preparation. Consequently, the rigorous preprocessing steps applied in the previous section (such as normalization, logarithmic transformations and K-Means clustering) were deemed unnecessary. Instead, the algorithm leveraged the raw dataset, allowing for the direct inclusion of the *zipcode* variable and preserving original spatial information. Within this framework, the analysis compared two distinct configurations: a **Full Feature model** utilizing the complete set of predictors and a second version subjected to **Feature Selection**, designed to isolate the most impactful variables and reduce noise.

Feature Selection Strategy

To optimize model complexity and computational efficiency, a Backward Feature Elimination strategy was employed on the Training Set (70%), validated through a 3-Fold Cross-Validation loop. The objective was to identify the minimal subset of predictors that maximized the coefficient of determination (R^2).

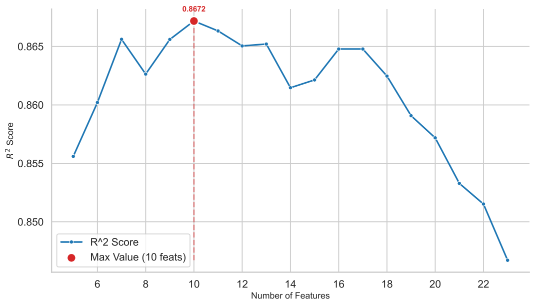


Figure 8: Feature Selection R^2 Curve

The optimization trajectory, visualized in the figure 8, reveals that the initial model, inclusive of all candidate features, exhibited sub-optimal performance due to the noise introduced by irrelevant variables. As features were iteratively removed, the R^2 score increased, reaching a Global Maximum ($R^2 \approx 0.867$) at a subset of exactly 10 features. Reducing the dimensionality beyond this point resulted in a sharp performance drop, indicating the loss of vital information. Consequently, the feature set was frozen at these 10 optimal variables for the subsequent modeling phase.

Random Forest with Feature Selection (Optimized)

The Optimized Model was trained using the 10 features identified in the previous step. The performance metrics indicate an improvement in predictive capability compared to the linear baseline.

Index	Train	Test
R^2	0.982	0.884
Adj. R^2	0.982	0.884
MAE	26,655	69,869
RMSE	48,940	128,351

Table 2: Performance Metrics for the Optimized RF

As summarized in Table 2, the model achieved a near-perfect fit on the Training Set ($R^2 \approx 0.982$), suggesting that the ensemble successfully captured almost all variance within the training data. While this significant gap between training and test scores (≈ 0.10) flags a potential tendency towards overfitting, the generalization capability remains robust. Of note, the Test Set R^2 of 0.884 represents a substantial improvement over the Linear Regression model (0.826), demonstrating the superior capacity of non-linear modeling to capture the complexities of the King County housing market.

Random Forest without Feature Selection (Baseline)

To validate the efficacy of the feature selection process, a comparative "Full-Feature" baseline model was trained including all available predictors. This unoptimized configuration yielded a Test R^2 of approximately 0.868, which is notably lower than the 0.884 achieved by the optimized model.

Index	Train	Test
R^2	0.979	0.868
Adj. R^2	0.979	0.868
MAE	27,589	72,675
RMSE	53,184	136,595

Table 3: Performance Metrics for the full RF

This degradation in performance can be attributed to the Curse of Dimensionality. The inclusion of variables with low importance (such as *Month* or *sqft_basement*) effectively diluted the predictive signal with noise, leading to a model that generalized less effectively to unseen data. This comparison empirically validates the necessity of the features selection step.

Cross-Validation and Robustness

Finally, to ensure that the reported predictive superiority of the optimized model was not an artifact of a specific random split, a 10-Fold Cross-Validation was conducted. The results confirmed the structural stability of the model: the training scores remained invariant across all folds ($R^2 \approx 0.982$), indicating consistent learning behaviour. Regarding generalization, the test scores averaged approximately 0.874, oscillating within a range of 0.853 to 0.895.

It is pertinent to note that even in the worst-case iteration (0.853), the Random Forest outperformed the best configuration of the Linear Regression model. This confirms that the Random Forest's architectural advantages are robust and statistically significant.

6 XGBoost Tree Ensemble

The final modeling phase involved the implementation of Extreme Gradient Boosting algorithm (XGBoost). In contrast to the Random Forest algorithm, which utilizes a bagging approach to construct independent decision trees in parallel, XGBoost operates on a boosting framework where trees are built sequentially. Each subsequent tree is explicitly designed to correct the residual errors generated by the preceding ensemble members. This fundamental algorithmic distinction heavily influenced the behaviour of the feature selection process and the final model architecture.

Feature Selection Strategy

To ensure methodological comparability with previous sections, the Backward Feature Elimination strategy was applied to the Training Set under a 3-Fold Cross-Validation framework. The objective remained the maximization of the R^2 score.

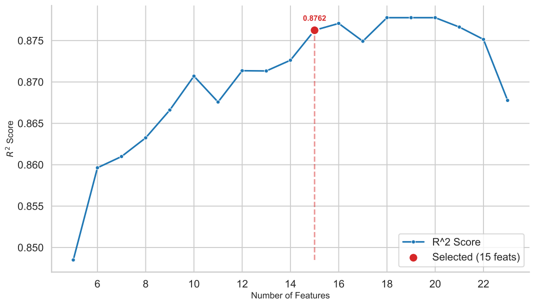


Figure 9: Feature Selection R^2 Curve

The optimization curve, illustrated in Figure 9, demonstrates a clear upward trend as informative variables are added. It is worth noting that while the absolute mathematical maximum is observed slightly beyond the selected threshold (peaking around 18 features), the subset of 15 features ($R^2 \approx 0.8762$) was identified as the optimal strategic compromise. In this point the model achieves near-peak performance before the marginal gains of adding further complexity become negligible or potentially detrimental due to overfitting. This dimensionality is significantly higher than the 10 features required by the Random Forest, suggesting that the Gradient Boosting mechanism effectively leverages secondary predictors. As evidenced by the steep drop in the curve to the left of the selected point, reducing the set below 15 features causes a loss of information.

XGBoost with Feature Selection (Optimized)

The Optimized Model was subsequently trained on the identified set of 15 features. The performance metrics indicate a superior predictive capacity compared to previous iterations.

Index	Train	Test
R^2	0.975	0.895
Adj. R^2	0.975	0.895
MAE	40,867	67,864
RMSE	57,463	121,863

Table 4: Performance Metrics for the Optimized XGBoost

As shown in Table 4, the model achieves a Test R^2 of 0.895. This result surpasses both the Linear Regression (0.82) and Random Forest (0.88) benchmarks, approaching the 0.90 threshold. Furthermore, this configuration recorded the lowest Root Mean Squared Error (RMSE) observed up to this stage of the analysis, indicating a high degree of precision.

XGBoost without Feature Selection (Full Model)

In a deviation from previous findings, a baseline model was trained on the full dataset to evaluate the algorithm's intrinsic robustness to high-dimensional data. Conversely to the Random Forest results, this "Full Model" outperformed the feature-optimized version.

Index	Train	Test
R^2	0.979	0.901
Adj. R^2	0.979	0.901
MAE	37,700	66,316
RMSE	52,733	118,572

Table 5: Performance Metrics for the full XGBoost

As detailed in Table 5, the Full Model achieved a Test R^2 of 0.901, successfully breaking the 0.90 accuracy barrier. This result contrasts sharply with the Random Forest analysis, where the inclusion of excess features degraded performance. It appears that XGBoost effectively utilized contextual variables (such as *sqft_living15*) that were filtered out during the selection process. This demonstrates the algorithm's superior noise robustness and its capacity to "digest" and extract value from complex high-dimensional information.

Cross-Validation and Final Verdict

To validate this hierarchy and ensure the results were not specific to the static train-test split, a 10-Fold Cross-Validation was conducted. The analysis confirmed the superiority of the Full Model, which yielded a mean Test R^2 of approximately 0.872, compared to 0.863 for the Optimized version.

Consequently, the Full XGBoost configuration is declared the best model of this study. It offers the highest predictive power ($R^2 > 0.90$ on the static split) and the best stability across validations, proving that for the algorithm within this specific real estate context, data richness outweighs parsimony.

7 Feature Importance Analysis

This section explores the interpretability of the predictive models by analyzing their internal decision-making logic. Specifically, it examines and compares the feature importance hierarchies derived from Linear Regression, Random Forest and XGBoost to identify the distinct variables that drive the most significant influence on housing prices within the King County market.

Linear Regression

For the parametric model, feature importance was evaluated using the **Absolute Value** of the linear regression coefficients. Since the input variables were Z-Score normalized, the magnitude of these coefficients provides a direct comparison of their influence on the predicted price.

The analysis identifies Geography as the primary driver of value, surprisingly surpassing structural attributes. As illustrated in the plot (Figure 10, the specific spatial zone *cluster_0* emerges as the single most significant predictor (Coefficient ≈ 0.37), followed by *sqft_living_log* (≈ 0.33) and the secondary zone *cluster_5* (≈ 0.32). This hierarchy validates the effectiveness of the Spatial Clustering strategy detailed in Section 3: knowing the precise neighborhood zone proves to be more critical for the linear model than the physical size of the property.

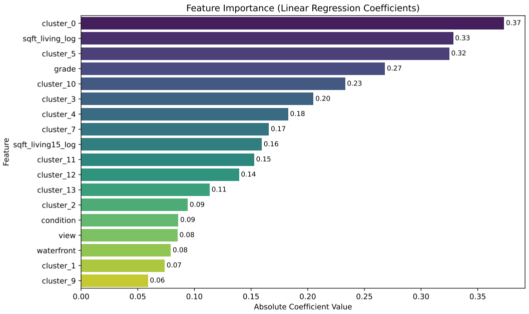


Figure 10: Linear Regression Feature Importance

While *grade* remains a strong structural predictor (≈ 0.27), variables typically associated with high value, such as *waterfront* and *view*, appear with relatively low importance coefficients (< 0.10). This is not an indication of irrelevance, but rather a result of multicollinearity with the spatial clusters. The value contribution of a "view" is likely absorbed by the high-value spatial clusters (e.g., a cluster covering a luxury coastal area already captures the location premium), leaving the binary variables with less residual variance to explain.

Random Forest

For the Random Forest analysis, importance was measured by the **Total Number of Splits** that is the frequency with which a feature is selected to partition the data across the ensemble of trees. A comparative analysis between the "Full" model and the "Optimized" model (post-feature selection) highlights the impact of noise on model focus.

In the full model, the predictive power is diluted. The primary feature, *sqft_living*, accounts for approximately 249 splits. The model allocates a significant number of decision nodes to noisy or low-relevance features such as *Month*, *floors* and *sqft_basement*, effectively "distracting" the ensemble.

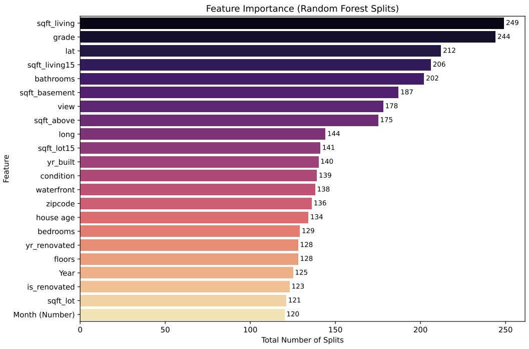


Figure 11: RF Full Feature Importance

On the other hand, the optimized model demonstrates a phenomenon of *Signal Concentration.* Upon removing the noise, the usage of key features increases significantly. The attribute *sqft_living* jumps to 394 splits (a 58% increase) and the construction *grade* rises to 381 splits. Regarding location, *lat* (Latitude) establishes itself as the third pillar (312 splits). Unlike the Linear model, which relied on pre-computed clusters, the Random Forest natively learns the North-South value gradient through coordinate splits. Furthermore, the removal of noise allows rare but critical features to surface. Notably, *waterfront*, which was buried at 13th place in the Full Model, rises to 5th place (269 splits) in the Optimized version. This confirms that the feature selection is useful for the interpretability of critical market drivers.

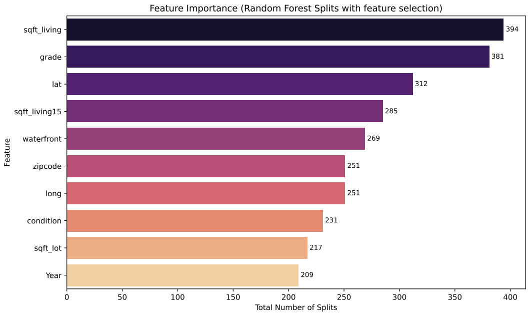


Figure 12: RF Optimized Feature Importance

XGBoost

Finally, the XGBoost models were analyzed using the **Total Gain** metric that quantifies the average reduction in prediction error (loss) contributed by a feature.

In the optimized configuration, XGBoost presents a slight inversion of the hierarchy observed in other models: *grade* emerges as the number one predictor (1.20×10^{15} gain), followed by *sqft_living*. This suggests that for the Gradient Boosting algorithm, the qualitative Grade of construction contains higher information density than raw dimension.

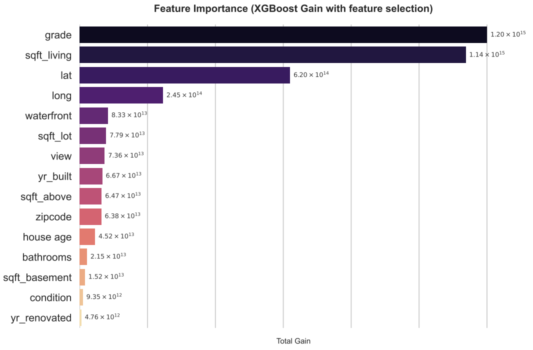


Figure 13: XGBoost Optimized Feature Importance

The analysis of the Full Model reveals why it outperformed the optimized version: the full model ranks *sqft_living15* (the average size of the nearest 15 neighbors) as the 5th most important feature. In the feature selection phase, *sqft_living15* was removed as ”redundant” due to its correlation with *sqft_living*. However, XGBoost effectively utilized this variable to establish Neighborhood Context (e.g., distinguishing a large house in a small-house neighborhood from a large house in a large-house neighborhood). The loss of this contextual nuance in the optimized dataset explains the slight drop in performance, demonstrating that Gradient Boosting benefits from a richer feature set.

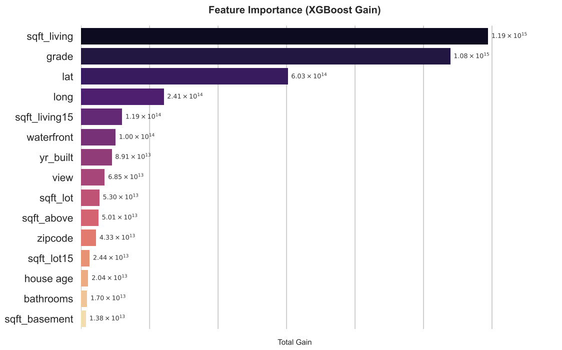


Figure 14: XGBoost Full Feature Importance

In conclusion, while the three modeling approaches exhibit distinct preferences in how they ingest data, a unified narrative regarding market value emerges. The price of real estate in King County is a non-linear function defined by three fundamental pillars: Structural Magnitude (represented by *sqft_living*), Construction Quality (represented by *grade*) and Precise Location (captured either through *Cluster* in linear models or raw *Latitude* in tree-based ensembles).

8 Conclusions and Future Work

This study successfully implemented an end-to-end machine learning pipeline for the objective valuation of real estate in King County, transitioning from parametric baselines to advanced gradient boosting architectures. The comparative analysis establishes the **XGBoost (Full Feature)** configuration as the superior architecture. With a Test R^2 of **0.901** and a Mean Absolute Error of **66,316**, it is the only model to breach the $R^2 \approx 0.90$ threshold, significantly outperforming both the Linear Regression baseline ($R^2 \approx 0.826$) and the Optimized Random Forest ($R^2 \approx 0.884$). Essentially, the gradient boosting algorithm demonstrated a superior capacity to ingest high-dimensional data, effectively leveraging secondary contextual variables (such as *sqft_living15*) to refine residuals and capture com-

Model Architecture	Configuration	Test R^2	Test MAE
Linear Regression	Baseline (Log-Target)	0.826	0.315*
Random Forest	Optimized (10 Features)	0.884	69,869
XGBoost	**Full (No FS)**	**0.901**	**66,316**

*Note: Linear Regression MAE is reported in log-scale units.

Table 6: Final Model Comparison: Performance Summary

plex neighborhood dynamics that simpler models discarded as noise.

These quantitative findings corroborate the study's primary goals. The ability to explain over 90% of price variance confirms that a data-driven price estimation model can be established, effectively minimizing subjective bias in valuation. At the same time, the investigation into value determinants revealed that market value is fundamentally driven by **Precise Location** (via Spatial Clusters or Latitude/Longitude), **Structural Quality** (Grade) and **Physical Magnitude** (Sqft Living). However, the specific success of the Full XGBoost model highlights that **Neighborhood Context** is the critical factor distinguishing high-accuracy prediction from standard estimation.

Despite the global model's robustness, the application of a uniform decision logic across a heterogeneous market remains a limitation. Future development should therefore focus on **Stratified Segmentation Modeling**, partitioning the dataset into distinct price tiers (e.g., Economy, Mid-Range, Luxury). It is hypothesized that feature importance is dynamic: aesthetic variables such as *view* and *waterfront* likely dominate the high-end segment, whereas functional attributes like *condition* and *house_age* are expected to drive the economy sector. Training specific sub-models for each tier allows for the capture of these granular nuances, offering a pathway to further reduce residual variance in extreme market segments.

References

[1] **L. Breiman**, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

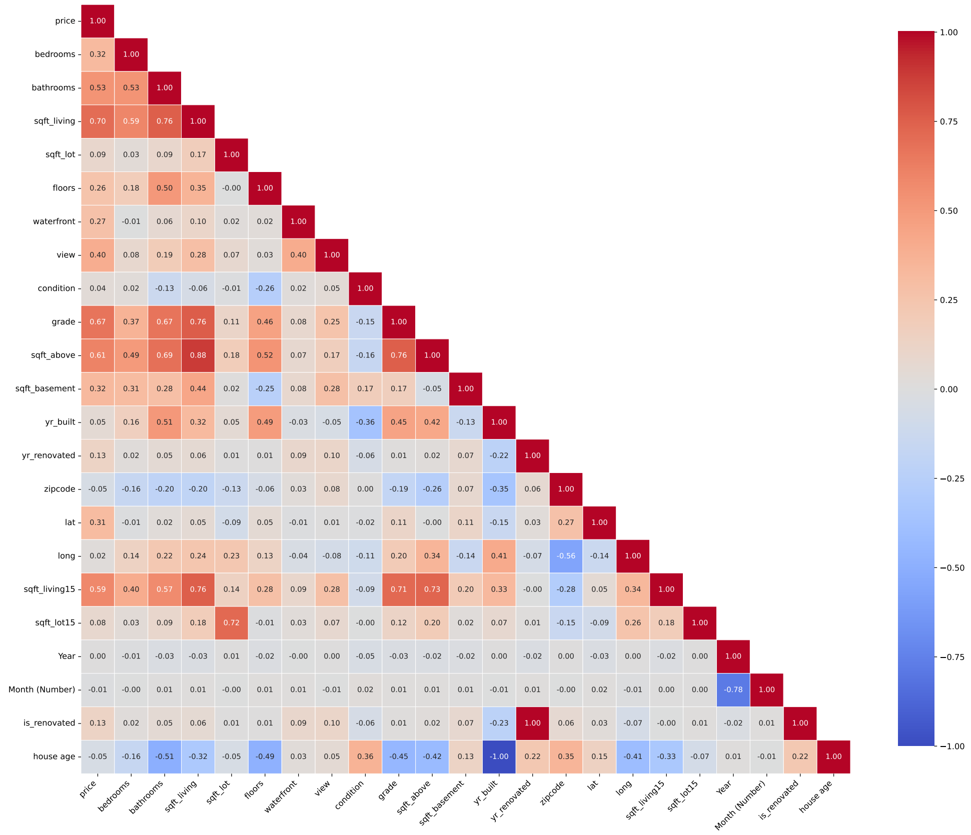
[2] **T. Chen and C. Guestrin**, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[3] **KNIME**, *KNIME Analytics Platform.* Zurich, Switzerland, 2025. Available at: `https://www.knime.com`.

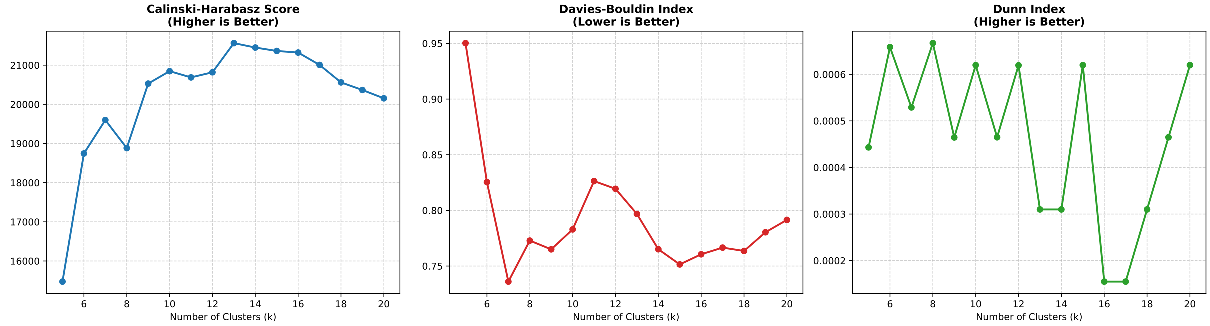
[4] **H. Foxem**, *House Sales in King County, USA*. Kaggle, 2016. Available at: `https://www.kaggle.com/harlfoxem/housesalesprediction`. (Accessed: 2026).

A Appendix: Additional Visualizations

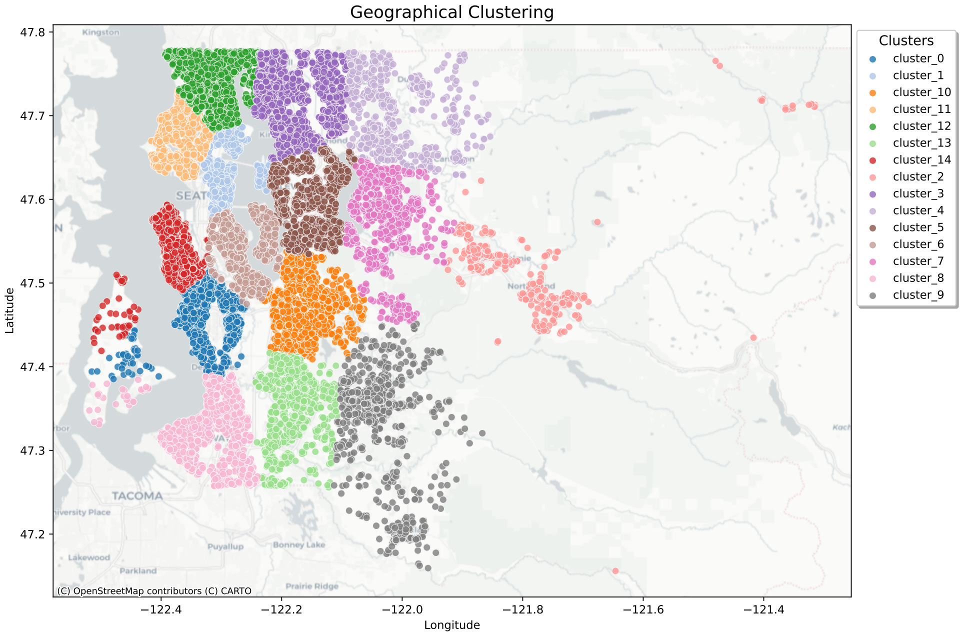
Correlation Matrix Heatmap



Calinski-Harabasz, Davies-Bouldin and Dunn indices



Map of the 15 Spatial Clusters


Geographical Clustering
Latitude
Longitude
47.8
47.7
47.6
47.5
47.4
47.3
47.2
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
Clusters
cluster_0
cluster_1
cluster_10
cluster_11
cluster_12
cluster_13
cluster_14
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
cluster_9
Kingston
SEATTLE
TACOMA
University Place
Lakewood
Parkland
Puyallup
Bonney Lake
Prairie Ridge
North Bend
(C) OpenStreetMap contributors (C) CARTO