



Predicting Residential Property Sale Prices

Module: IJC428

Word Count: 2616

Registration Number: 250214208

Structured Abstract

Background: Accurate house price prediction is a critical challenge in real estate analytics, with significant implications for buyers, sellers and policymakers. This study addresses the regression problem of predicting residential property sale prices using supervised machine learning techniques applied to a dataset of 1,300 houses sold in Ames, Iowa.

Methods: Two supervised learning algorithms were implemented and compared: Multiple Linear Regression (MLR) and Random Forest (RF) regression. Data preprocessing included removal of high-missing-value features, median and most-frequent imputation for remaining missing values, one-hot encoding of categorical variables, min-max normalisation and manual feature selection to address multicollinearity. Both models were evaluated using 10-fold cross-validation on the training set and validated on a held-out test set comprising 20% of the data. The complete workflow was implemented in KNIME Analytics Platform 5.3.3.

Results: Random Forest outperformed Multiple Linear Regression across all evaluation metrics. On the test set, Random Forest achieved $R^2=0.884$, MAE=\$18,383 and RMSE=\$27,881, compared to $R^2=0.820$, MAE=\$23,390 and RMSE=\$34,685 for Linear Regression. Both models generalised well with no evidence of overfitting.

Conclusion: Random Forest demonstrated superior predictive accuracy for house price estimation, particularly benefiting from its ability to capture nonlinear relationships. Both models struggled with high-value outlier properties above \$400,000, suggesting this as an area for future improvement.

Introduction

House price prediction represents one of the most practically significant regression problems in data mining, with direct implications for homeowners, real estate agents, mortgage lenders and policymakers (Park & Bae, 2015). Accurate valuation models enable informed decision-making across the real estate market, yet predicting individual property prices remains challenging due to the complex and nonlinear interactions between numerous property attributes (Pai & Wang, 2020).

This report addresses the prediction of residential property sale prices, the variable `SalePrice`, using a dataset of 1,300 houses sold in Ames, Iowa. The dataset comprises 31 attributes encompassing physical characteristics such as above-ground living area, overall quality rating and garage capacity, alongside categorical features including neighbourhood classification and building type. `SalePrice` ranges from \$34,900 to \$755,000, with a mean of approximately \$181,000 and a right-skewed distribution indicating the presence of high-value outlier properties.

Two supervised machine learning methods were selected: Multiple Linear Regression as an interpretable baseline model (Tranmer et al., 2008), and Random Forest as a more sophisticated ensemble method capable of capturing nonlinear relationships (Breiman, 2001). Both methods have been successfully applied to house price prediction in prior literature (Adetunji et al., 2022; Truong et al., 2020). Model performance was evaluated using R^2 , MAE and RMSE, with 10-fold cross-validation employed to ensure robust model assessment.

Data Mining Theory

Multiple Linear Regression

Multiple Linear Regression (MLR) is a supervised learning technique that models the linear relationship between a continuous response variable and two or more explanatory variables (Tranmer et al., 2008). The general form is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

where β_0 is the intercept, $\beta_1 \dots \beta_p$ are regression coefficients representing the change in y for a one-unit increase in each predictor, and ε is the error term. Coefficients are estimated using Ordinary Least Squares, minimising the sum of squared residuals (Tranmer et al., 2008).

MLR is appropriate for house price prediction as SalePrice is a continuous variable with measurable linear relationships to attributes such as living area and quality rating. The interpretability of regression coefficients provides direct insight into feature contributions, and regression-based approaches represent one of the two principal research traditions in housing price modelling (Park & Bae, 2015). However, MLR relies on several assumptions, including linearity, normality of residuals, homoscedasticity and the absence of multicollinearity. Multicollinearity, occurring when explanatory variables are highly intercorrelated, inflates coefficient variance, widens confidence intervals and produces unreliable estimates (Kim, 2019). Categorical variables must also be converted to $n-1$ dummy variables per feature to avoid the dummy variable trap (Tranmer et al., 2008). A further limitation is MLR's sensitivity to outliers and its inability to model nonlinear relationships or feature interactions, potentially limiting accuracy in complex property markets (Smith et al., 2013).

Random Forest Regression

Random Forest is an ensemble learning method introduced by Breiman (2001), constructing multiple decision trees and aggregating their predictions. In regression tasks, the forest prediction is the average of individual tree outputs:

$$\hat{y} = (1/k) \sum h_k(x)$$

where $h_k(x)$ is the prediction of the k th tree. Each tree is grown on a bootstrap sample of training data, and at each node only a randomly selected subset of features is considered for splitting. This combination of bagging and random feature selection reduces inter-tree correlation, which Breiman (2001) demonstrated is a key determinant of forest accuracy. A mathematically important property is that Random Forest generalisation error converges as trees increase, proving it does not overfit as more trees are added (Breiman, 2001).

Random Forest is well suited to house price prediction for several reasons. It naturally handles nonlinear relationships, such as the diminishing returns of additional bedrooms in luxury properties, which MLR cannot capture (Smith et al., 2013). It implicitly models feature interactions without requiring explicit specification, and is robust to outliers without the distributional

assumptions imposed by MLR (Smith et al., 2013). Adetunji et al. (2022) demonstrated competitive Random Forest accuracy on a housing dataset, further supporting its suitability. Despite these advantages, the ensemble nature reduces interpretability compared to MLR, and Denil et al. (2014) note that despite widespread practical success, a theoretical gap between Random Forest models and their practical implementation persists. Key parameters include the number of trees, set to 100 in this study following Breiman's (2001) recommendation, and the number of features considered at each split, set to the KNIME default of the square root of total features. The Tree Ensemble Statistics node confirmed 100 trees were built with an average depth of 21.04, reflecting the complexity of nonlinear interactions captured.

Evaluation Metrics

Three metrics were employed to evaluate and compare model performance. The coefficient of determination (R^2) measures the proportion of variance in SalePrice explained by the model, with Adjusted R^2 additionally penalising redundant predictors (Tranmer et al., 2008). Root Mean Squared Error (RMSE) computes the square root of average squared prediction errors, penalising large errors more heavily due to the squaring operation, making it appropriate when large financial misvaluations are particularly undesirable (Chai & Draxler, 2014). Mean Absolute Error (MAE) measures average absolute prediction error in dollars, providing a more robust and interpretable measure less sensitive to outliers (Hodson, 2022). Chai & Draxler (2014) recommend using both RMSE and MAE together, as each captures different aspects of model error, a practice adopted in this study.

Data Exploration and Preparation

Dataset Description and Exploratory Analysis

The dataset comprises 1,300 residential property records with 31 attributes, spanning numeric and categorical feature types. Table 1 summarises key feature characteristics identified during exploratory analysis.

Feature	Type	Min	Max	Missing
SalePrice	Numeric	\$34,900	\$755,000	0%
GrLivArea	Numeric	334 sq ft	5,642 sq ft	0%
OverallQual	Numeric	1	10	0%
LotFrontage	Numeric	21 ft	313 ft	17.6%
GarageType	Categorical	-	-	5.5%
Alley	Categorical	-	-	94.1%

Table 1: Summary of selected feature characteristics

Exploratory data analysis was conducted on the full dataset prior to splitting, using statistical summaries, histograms, scatter plots and box plots in KNIME, consistent with the approach recommended by Han et al. (2012).

Histogram analysis revealed SalePrice is strongly right-skewed with a skewness coefficient of 1.96, with the majority of properties concentrated between \$100,000 and \$200,000 and a long tail extending to \$755,000. This is characteristic of residential property markets where luxury properties constitute a small but disproportionate segment (Pai & Wang, 2020). The right-skewed distribution is an important consideration for Linear Regression, as it may violate the assumption of normally distributed residuals (Tranmer et al., 2008). Box plot analysis confirmed numerous high-value outliers in SalePrice, with several properties exceeding \$400,000. Scatter plot analysis of GrLivArea against SalePrice revealed a clear positive linear relationship, supporting its inclusion as a key predictor, with two notable outliers representing large properties with disproportionately low prices.

Missing Value Treatment

Three attributes required attention. The Alley column, with 94.1% missing values, was removed entirely as it contained insufficient information for prediction. The Id column, a meaningless row index, was also removed. LotFrontage, with 17.6% missing values, was imputed using the median, as the right-skewed distribution makes mean imputation susceptible to upward bias (Emmanuel et al., 2021). GarageType, with 5.5% missing values, was imputed using the most frequent value, appropriate for categorical variables. These decisions reflect established missing data handling practice, where imputation method should reflect the statistical properties of the variable (Emmanuel et al., 2021).

Feature Selection, Encoding and Normalisation

Highly correlated feature pairs were identified and addressed to reduce multicollinearity, which produces unreliable coefficient estimates in regression (Kim, 2019). TotRmsAbvGrd, GarageArea, 1stFlrSF, PoolArea and WoodDeckSF were removed, producing a stable 15-feature model with negligible loss of predictive information, consistent with Kim's (2019) finding that excluding multicollinear variables stabilises regression without meaningful R^2 reduction.

Nine categorical columns were one-hot encoded using the One to Many node, with one dummy column removed per variable to avoid the dummy variable trap (Tranmer et al., 2008). Min-max normalisation was applied to all numeric features, fitted exclusively on the training set and applied to the test set via the Normalizer Apply node, preventing data leakage. Ahsan et al. (2021) demonstrated that appropriate feature scaling is critical for coefficient-based algorithms such as Linear Regression.

Experimental Setup

Data Partitioning

The dataset was partitioned into 80% training (1,040 records) and 20% test (260 records) using the Partitioning node in KNIME, with random sampling and a fixed seed of 12345 to ensure reproducibility. The test set was held aside and remained completely unseen throughout all preprocessing and model training stages, ensuring an unbiased final evaluation. All preprocessing steps were applied separately to each model's training pipeline, preventing data leakage, which Cawley & Talbot (2010) identify as a significant source of optimistic bias in performance evaluation.

Cross-Validation Strategy

10-fold cross-validation was applied to the training set for both models independently using the X-Partitioner and X-Aggregator nodes in KNIME. The training data was partitioned into 10 equal subsets, with each fold serving as the validation set once while the remaining nine folds were used for training. The cross-validation performance estimate is the average across all 10 folds (Kohavi, 1995).

The choice of 10 folds is well supported in the literature. Kohavi (1995) demonstrated through large-scale experimentation that 10-fold cross-validation provides the best balance between bias and variance for model selection, outperforming leave-one-out cross-validation, which exhibits high variance, and lower fold counts, which introduce pessimistic bias. Nti et al. (2021) confirmed that $k=10$ is the most widely adopted value among practitioners, producing reliable generalisation estimates. Cross-validation was conducted exclusively on the training set, ensuring the test set remained unseen during model development, consistent with the requirement that model selection must be treated as an integral part of the fitting procedure and performed independently of test data (Cawley & Talbot, 2010). A fixed random seed of 12345 was applied consistently across both the Partitioning and X-Partitioner nodes, ensuring that fold assignments were reproducible and results could be independently verified.

Linear Regression Configuration

The Linear Regression Learner node was configured with SalePrice as the target and 15 numeric features as predictors. Categorical dummy columns were excluded due to severe multicollinearity, which produced coefficients of the order 10^{18} in initial experiments, consistent with the dummy variable trap described by Tranmer et al. (2008). Linear Regression has no hyperparameters requiring tuning.

Random Forest Configuration

The Random Forest Learner (Regression) node was configured with SalePrice as the target using the same 15 features. Three tree configurations were evaluated to identify the optimal setting, as shown in Table 2:

Number of Trees	CV R²	CV RMSE	Test R²	Test RMSE
50	0.856	\$30,168	0.886	\$27,572
100	0.856	\$30,160	0.884	\$27,881
200	0.858	\$30,042	0.883	\$27,969

Table 2: Random Forest performance across different numbers of trees

Differences across configurations are negligible, with R² varying by less than 0.003, demonstrating the stability of Random Forest with respect to tree count. This is consistent with Breiman's (2001) proof that generalisation error converges as trees increase. The configuration of 100 trees was selected as the final model, following Breiman's (2001) original recommendation that this provides reliable out-of-bag estimates while remaining computationally efficient. The consistency between cross-validation and test set results for both models confirmed no overfitting occurred (Cawley & Talbot, 2010).

Results and Discussion

Cross-Validation and Test Set Results

Table 3 presents cross-validation and test set performance for both models.

Metric	LR (CV)	LR (Test)	RF (CV)	RF (Test)
R ²	0.736	0.820	0.856	0.884
Adjusted R ²	0.736	0.809	0.854	0.877
MAE	\$23,460	\$23,390	\$18,407	\$18,383
RMSE	\$40,880	\$34,685	\$30,160	\$27,881
MAPE	14.0%	12.9%	11.1%	9.6%

Table 3: Model performance comparison

Random Forest outperformed Linear Regression across all metrics. On the test set, Random Forest achieved $R^2=0.884$ compared to 0.820 for Linear Regression, explaining 88.4% and 82.0% of SalePrice variance respectively. Both models exceeded the commonly accepted threshold of $R^2=0.70$ for acceptable regression performance (Pai & Wang, 2020). In terms of absolute prediction error, Random Forest predictions were on average \$5,007 closer to actual prices than Linear Regression as measured by MAE, with RMSE lower by \$6,804, indicating fewer large prediction errors. Given that large errors represent significant financial misvaluation, this difference is practically meaningful (Chai & Draxler, 2014).

Importantly, test set performance exceeded cross-validation performance for both models, with Linear Regression improving from $R^2=0.736$ to 0.820 and Random Forest from 0.856 to 0.884. This consistency provides strong evidence that neither model overfitted to the training data, validating the methodological approach of strict training and test set separation (Cawley & Talbot, 2010). The superior performance of Random Forest can be attributed to the nonlinear nature of house price determination, where the relationship between OverallQual and SalePrice accelerates at higher quality ratings rather than following the constant linear gradient assumed by MLR. Furthermore, MLR's exclusion of categorical features such as Neighborhood, necessitated by multicollinearity constraints, likely contributed to its lower R^2 , as location is widely recognised as a primary price determinant in residential markets (Park & Bae, 2015). This suggests the performance gap between models reflects not only fundamental algorithmic differences but also the limitations imposed by MLR's distributional assumptions on the preprocessing pipeline, a constraint that Random Forest does not share (Smith et al., 2013).

Feature Importance and Model Insights

Analysis of Linear Regression coefficients identified GrLivArea (coefficient: +\$340,319), OverallQual (+\$160,846) and GarageCars (+\$60,399) as the strongest positive predictors,

consistent with established understanding of residential property valuation (Adetunji et al., 2022). Statistically significant predictors at the 5% level included GrLivArea, OverallQual, LotArea, YearBuilt, GarageCars, TotalBsmtSF, BedroomAbvGr and KitchenAbvGr. Notably, BedroomAbvGr (coefficient: -\$54,113) and KitchenAbvGr (-\$58,155) showed negative coefficients, suggesting that holding living area constant, additional bedrooms or kitchens are associated with lower prices, possibly reflecting smaller room sizes. Several predictors including LotFrontage, YearRemodAdd, FullBath and HalfBath were not statistically significant, indicating limited independent predictive value once other features are controlled for (Tranmer et al., 2008). This pattern of insignificance may reflect correlations with included predictors rather than a genuine absence of relationship with SalePrice, a well-documented phenomenon in multiple regression analysis.

Scatter Plot Analysis and Limitations

Figure 1 presents the scatter plot of predicted versus actual SalePrice for Linear Regression on the test set. A clear positive diagonal trend is evident, confirming the model captures the general price relationship. However, one property received a negative predicted price of approximately -\$20,000, visible at the bottom left of Figure 1, highlighting a fundamental limitation of Linear Regression, which is unconstrained in its output range and can produce physically impossible predictions (Smith et al., 2013). Additionally, high-value properties above \$400,000 are consistently underestimated, with predicted values falling below the diagonal, reflecting the influence of the dense cluster of moderately priced properties on coefficient estimation.

Scatter Plot

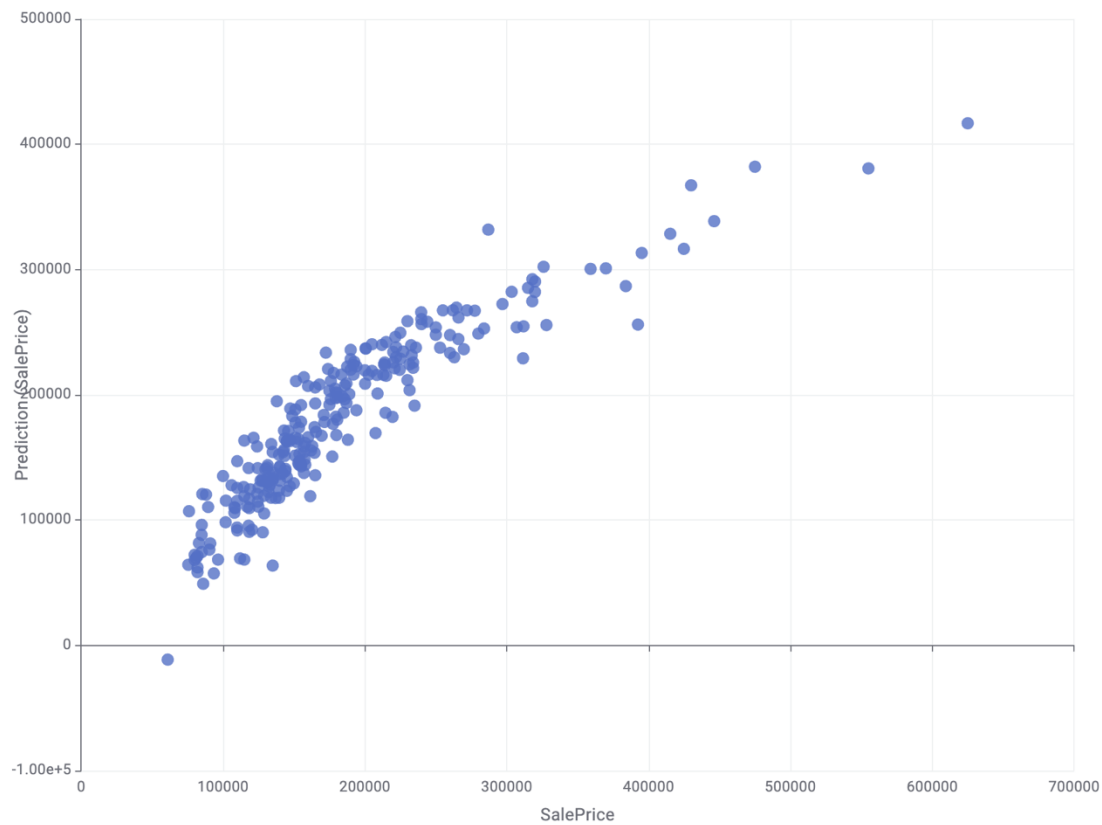


Figure 1: Predicted versus actual SalePrice for Multiple Linear Regression on the test set ($n=260$).

Figure 2 presents the equivalent scatter plot for Random Forest. The overall fit is visibly tighter around the diagonal compared to Figure 1, consistent with the superior R^2 of 0.884. Crucially, no negative predictions were produced, as Random Forest outputs are bounded by the range of training values (Breiman, 2001). One notable outlier is visible in the upper right of Figure 2, where a property with an actual price of approximately \$620,000 received a predicted value of approximately \$570,000, suggesting the model partially captured but underestimated this luxury property. Both models share a common limitation in predicting high-value properties above \$400,000, reflecting regression toward the mean for outlying observations in ensemble tree methods (Louppe, 2014).

Scatter Plot

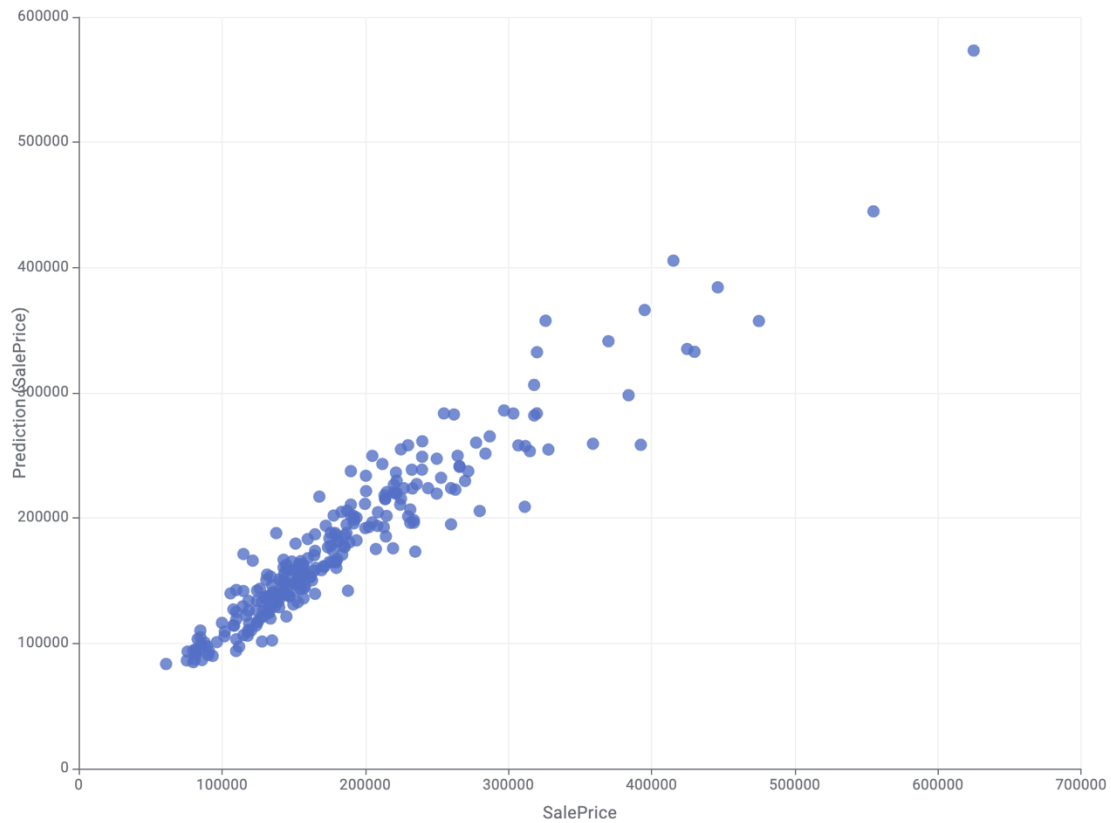


Figure 2: Predicted versus actual SalePrice for Random Forest regression on the test set (n=260).

These findings are consistent with prior literature. Truong et al. (2020) similarly found Random Forest achieved the lowest error among several machine learning methods for housing data. The contrast with Smith et al. (2013), who found MLR outperforming Random Forest in neuroscience data, demonstrates that algorithm superiority is dataset dependent, with nonlinear property market dynamics favouring ensemble methods in the present study.

Conclusion and Reflections

Summary of Findings

This study applied Multiple Linear Regression and Random Forest regression to predict residential property sale prices in Ames, Iowa. Random Forest outperformed Linear Regression across all evaluation metrics, achieving $R^2=0.884$, $MAE=\$18,383$ and $RMSE=\$27,881$ on the test set, compared to $R^2=0.820$, $MAE=\$23,390$ and $RMSE=\$34,685$ for Linear Regression. Both models exceeded the commonly accepted $R^2=0.70$ threshold for acceptable regression performance (Pai & Wang, 2020), and the consistency between cross-validation and test results confirmed that neither model overfitted (Cawley & Talbot, 2010). Feature importance analysis identified GrLivArea, OverallQual and GarageCars as the strongest price predictors, consistent with established property valuation literature (Adetunji et al., 2022).

Reflections on Methods

Several methodological choices merit reflection. Excluding categorical features from Linear Regression, while necessary to address multicollinearity, removed potentially informative predictors such as Neighborhood, widely recognised as a primary price determinant (Park & Bae, 2015). With hindsight, dimensionality reduction techniques such as Principal Component Analysis could have retained categorical information while addressing multicollinearity. More sophisticated missing value imputation methods, such as k-nearest neighbour or Random Forest-based imputation, may have produced more accurate substituted values than simple median imputation (Emmanuel et al., 2021).

Potential Improvements

Both models consistently underestimated luxury properties above \$400,000. Log-transformation of SalePrice could address the right-skewed distribution and improve predictions in this range (Tranmer et al., 2008). Advanced ensemble methods such as Gradient Boosting or XGBoost, which Truong et al. (2020) demonstrated outperform standard Random Forest through stacked generalisation, represent a natural extension of this work. Incorporating additional features capturing macroeconomic variables such as interest rates and employment levels, identified by Pai & Wang (2020) as influential real estate factors, could also improve predictive accuracy. Despite these limitations, both models demonstrated strong and reliable performance, with Random Forest providing a robust approach to automated house price estimation.

References

- Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199, 806–813.
- Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3), 52.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11, 2079–2107.
- 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.
- Denil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. International conference on machine learning (pp. 665-673). PMLR.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, 8(1), 140.
- Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and. *Techniques*, Waltham: Morgan Kaufmann Publishers, 2012–2013.
- Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 2022, 1–10.
- Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, 72(6), 558–569.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* (Vol. 14, No. 2, pp. 1137-1145)
- Louppe, G. (2014). *Understanding random forests: From theory to practice*. Universite de Liege (Belgium).
- Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology and Computer Science*, 13(6), 61–71.
- Pai, P.-F., & Wang, W.-C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, 10(17), 5832.
- 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.
- Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, 220(1), 85–91.
- Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 5(5), 1–5.
- Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174, 433–442.
- Gilroy, K. E. (2023). The Large Competitors in the Fashion E-Retailing Marketplace: Has the Market Become Saturated?.

Dekimpe, M. G., & van Heerde, H. J. (2023). Retailing in times of soaring inflation: What we know, what we don't know, and a research agenda. *Journal of Retailing*, 99(3), 322-336.

Hjort, K., & Lantz, B. (2016). The impact of returns policies on profitability: A fashion e-commerce case. *Journal of Business Research*, 69(11), 4980-4985.

(2026, April 23). Asos half-year loss narrows as it keeps full-year guidance. *Alliance News UK*. <https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:6JDK-NVJ3-RWWC-402N-00000-00&context=1519360>