The University of Sheffield

School of Information, Journalism and Communication

MSc Data Science

Predicting House Prices
Final Report Data Mining & AI

Course Code: IJC428
Course: Data Mining and AI
Student Registration Number: 250208702
Word Count: 2724

Structured Abstract

Background. Accurate residential property valuation is a longstanding regression problem in real estate analytics. House prices reflect a complex interplay of structural, locational, and qualitative attributes, making them an instructive benchmark for predictive modelling.

Objective. This report compares the predictive performance and interpretability of Multiple Linear Regression and Random Forest for predicting log-transformed sale prices in the Ames, Iowa housing dataset (n = 1,300).

Methods. Data were partitioned into training (80%, n = 1,040) and held-out test (20%, n = 260) sets using a fixed random seed (42). Two preprocessing pipelines were implemented in KNIME 5.3.3, each tailored to the assumptions of its model. Hyperparameters were tuned via 10-fold cross-validation on the training set, then performance was evaluated on the held-out test set using R^2, RMSE, MAE, and mean signed difference.

Results. Multiple Linear Regression achieved test $R^2 = 0.863$ and RMSE = 0.139 (log units). The optimal Random Forest configuration (500 trees, unlimited depth) achieved $R^2 = 0.862$ and RMSE = 0.139. Both models generalised well, with no evidence of overfitting.

Conclusion. Linear regression and random forest produced essentially equivalent test-set performance despite LR's slight cross-validation advantage, illustrating that algorithm complexity does not always translate to predictive advantage on small-to-medium tabular datasets.

1. Introduction

Predicting residential property prices is a longstanding regression task in real estate analytics. The complexity arises from the joint influence of structural, locational, and qualitative attributes, often in non-linear and interacting ways. The Ames, Iowa housing dataset, compiled by De Cock (2011) and modified for this coursework, provides a rich tabular benchmark for evaluating predictive modelling techniques on this task.

The dataset comprises 1,300 residential property records with 30 predictor attributes including categorical features (such as Neighborhood, Building Type, and Garage Type), discrete ordinal features (Overall Quality and Overall Condition, both rated 1-10), and continuous numeric features (Lot Area, Above-Grade Living Area). The target variable is SalePrice, expressed in US dollars and ranging from $34,900 to $755,000 with a strongly right-skewed distribution.

Two supervised regression algorithms are compared: Multiple Linear Regression (MLR), a parametric method offering interpretable coefficients; and Random Forest (RF), a non-parametric ensemble method capable of capturing non-linear relationships and feature interactions. These methods were chosen to provide a clear methodological contrast, parametric versus non-parametric, linear versus non-linear, enabling a meaningful comparison of both predictive performance and explanatory transparency.

Section 2 reviews the theoretical foundations of the chosen methods; subsequent sections cover preprocessing, experimental design, results, and reflections.

2. Data Mining Theory

2.1 Multiple Linear Regression

Multiple Linear Regression (MLR) models the conditional expectation of a continuous target as a linear combination of predictors (Osborne, 2017; Hastie et al., 2009):

$$\hat{y}_i = \beta_0 + \beta_1 x\_i1 + \beta_2 x\_i2 + \ldots + \beta_p x\_ip + \varepsilon_i$$

where $\hat{y}_i$ is the predicted value for observation i, β_0 is the intercept, β_1 to β_p are coefficients for the p predictors x_i1 to x_ip, and ε_i is the residual error. Coefficients are estimated by Ordinary Least Squares (OLS), which has the closed-form solution $\hat{\beta} = (X^TX)^{-1}X^Ty$. MLR has no conventional hyperparameters; modelling decisions concern intercept inclusion, missing-value treatment, and feature engineering.

MLR offers direct interpretability, each coefficient expresses the average change in log-price per unit change in the predictor (Osborne, 2017), and is computationally efficient, well-suited to relationships that are approximately linear after appropriate target transformation (Li, 2022). Its disadvantages include strong distributional assumptions (linearity, homoscedasticity, normality of residuals), sensitivity to outliers and multicollinearity, and inability to capture non-linear interactions without explicit feature engineering (Hastie et al., 2009).

2.2 Random Forest Regression

Random Forest (RF), introduced by Breiman (2001), is an ensemble method that aggregates predictions from many decision trees, each trained on a bootstrap sample of the training data with feature sub-sampling at every split. For regression, the ensemble prediction is the arithmetic mean of individual tree predictions:

$$\hat{y}\_RF(x) = (1/B) \cdot \Sigma\_b \, T\_b(x)$$

where B is the total number of trees and T_b(x) is the prediction of the b-th tree. Each tree is grown on a bootstrap sample, with mtry features sampled randomly at each node, and splits selected by the variance reduction criterion. The bagging-plus-feature-sampling combination reduces predictive variance substantially without

inflating bias (Hastie et al., 2009). Principal hyperparameters are the number of trees B, the maximum tree depth or minimum node size, the number of features sampled per split (mtry), and the splitting criterion.

RF automatically captures non-linear relationships and feature interactions without manual feature engineering (Antad et al., 2024), handles mixed numeric and categorical features natively, is robust to outliers and irrelevant features, and offers internal feature-importance measures (Palczewska et al., 2013). Disadvantages include reduced interpretability, regression-to-mean behaviour that under-predicts extremes (Breiman, 2001), higher computational cost, and a split-frequency importance measure biased toward high-cardinality categoricals (Hastie et al., 2009). Recent empirical work also shows that well-engineered linear models can match tree-based methods on small-to-medium tabular datasets (Grinsztajn et al., 2022).

2.3 Performance Evaluation

Model performance was assessed using four complementary metrics: the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean signed difference. RMSE penalises large errors more heavily and is the conventional choice when fitting by OLS, whilst MAE provides a robust measure of typical error magnitude, together these give a fuller picture of error structure than either alone (Chai & Draxler, 2014; Willmott & Matsuura, 2005). R^2 measures the proportion of variance explained, and mean signed difference detects systematic bias.

3. Data Exploration and Preparation

Initial exploratory analysis using the KNIME Statistics node revealed that only three predictors contained missing values: Alley (1,223 missing observations; 94.1%), LotFrontage (229; 17.6%), and GarageType (72; 5.5%). Inspection of feature semantics indicated that Alley and GarageType encode meaningful absence rather than unknown information, values are missing where the property has no alley access or no garage. These cases were imputed with the constant value "None". For

LotFrontage, missingness appeared structural, and median imputation was applied, a robust strategy under skewed distributions (Little & Rubin, 2020).

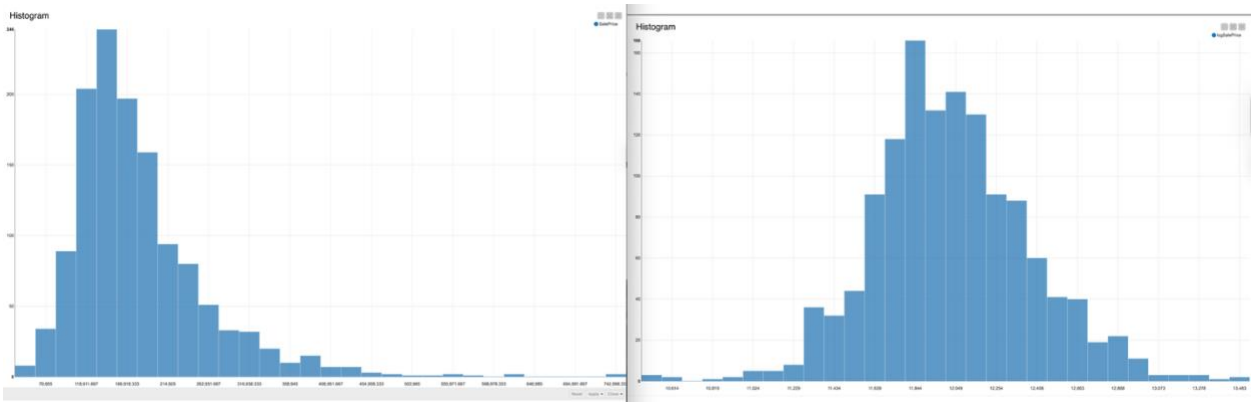

Figure 1. Distribution of SalePrice. (a) raw values, exhibiting strong right-skewness; (b) log-transformed values, approximately normally distributed.

The target SalePrice exhibited strong right-skewness (skewness = 1.97, kurtosis = 6.99), with a long tail of high-priced properties (Figure 1a). A natural logarithm transformation produced an approximately normal distribution (skewness = 0.15; Figure 1b), supporting the linear regression assumption of approximately normal residuals (Osborne, 2017). The log-transformed target (logSalePrice) was used as the response variable for both models, enabling direct comparison of error metrics on the same scale.

Pearson correlations identified OverallQual (r = 0.79), GrLivArea (r = 0.71), GarageCars (r = 0.64), GarageArea (r = 0.62), and TotalBsmtSF (r = 0.60) as the strongest linear predictors of SalePrice (Figure 2). Inspection also revealed pairwise multicollinearity: GarageCars/GarageArea (r = 0.88), GrLivArea/TotRmsAbvGrd (r = 0.83), and TotalBsmtSF/1stFlrSF (r = 0.82). Such pairs inflate coefficient variance in linear regression but have negligible effect on tree-based methods (Hastie et al., 2009).

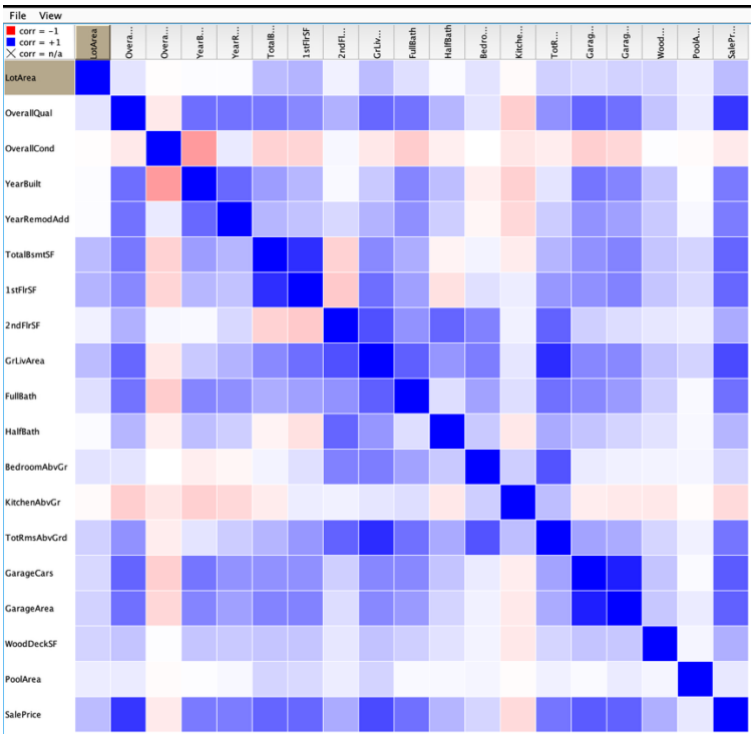

Figure 2. Pearson correlation heatmap of numeric features.

Outlier inspection (Figure 3) revealed several outlying observations in LotArea (38; 2.9%) and GrLivArea (18; 1.4%), including two extreme cases (GrLivArea > 4,000 sq ft with SalePrice < $200,000) widely discussed in the Ames literature (De Cock, 2011). These were retained as legitimate sales rather than data errors.

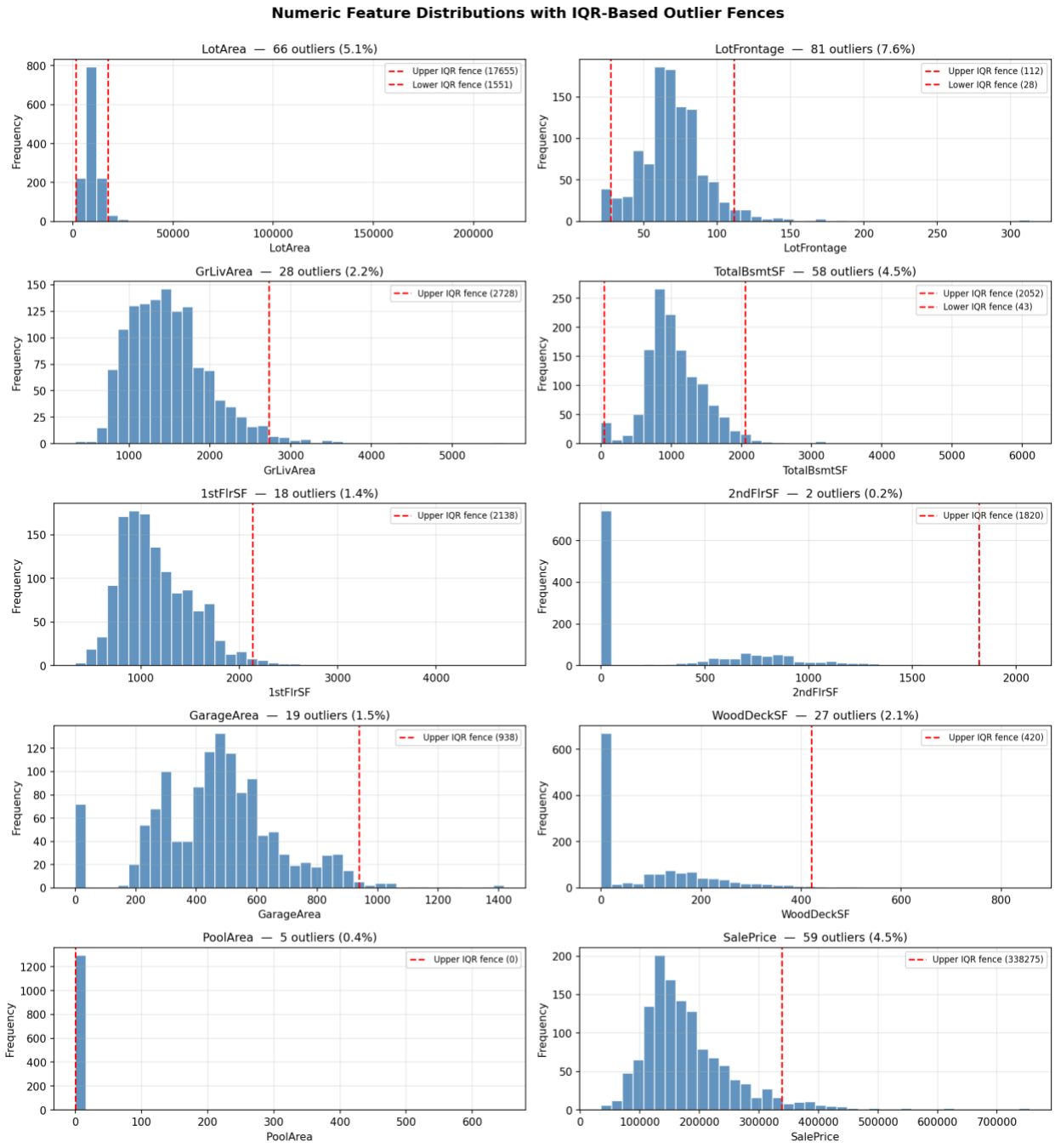

Figure 3. Outlier inspection histograms for key numeric features with IQR-based fences.

4. Experimental Setup

The complete workflow was implemented in KNIME Analytics Platform 5.3.3 (Figure 4). The dataset was partitioned into training (80%, n = 1,040) and held-out test (20%, n = 260) sets using random sampling with a fixed seed of 42, ensuring

reproducibility and strict separation between model development and final evaluation (Hastie et al., 2009).

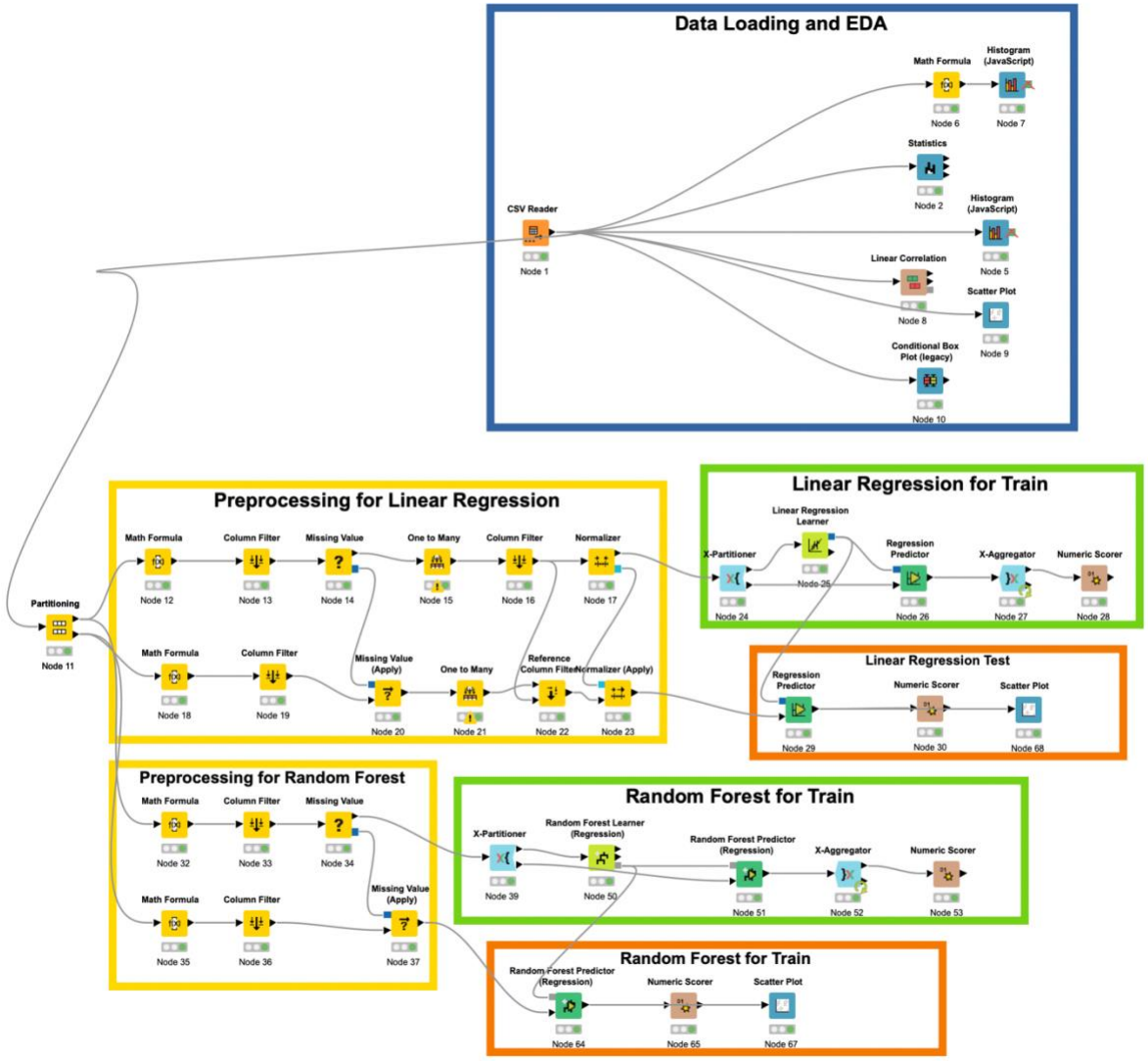

Figure 4. Complete KNIME workflow, comprising exploratory analysis (top), Linear Regression pipeline (upper-middle), and Random Forest pipeline (lower).

4.1 Feature selection and column-level justification

The dataset contained 30 predictors plus the target SalePrice. The Id column was excluded as a non-informative row identifier. The remaining 29 attributes were retained as candidate features for both models, with method-specific preprocessing applied as follows.

For the Multiple Linear Regression pipeline, three further numeric columns, GarageArea, TotRmsAbvGrd, and 1stFlrSF, were removed for exhibiting a Pearson correlation above 0.80 with another retained predictor (GarageCars, GrLivArea, and TotalBsmtSF respectively); such collinearity inflates OLS coefficient variance without improving predictive accuracy (Hastie et al., 2009). All remaining categorical attributes were one-hot encoded, with one reference dummy dropped per categorical to avoid the dummy variable trap. The surviving numeric features were Z-score standardised so that coefficients are directly comparable across features.

For the Random Forest pipeline, all 29 candidate predictors were retained without further filtering. Tree-based methods make splits based on the order of values rather than their magnitude, so standardisation is unnecessary; they handle categorical features natively without one-hot encoding; they are robust to multicollinearity because the random feature sub-sampling at each node spreads predictive credit across correlated predictors; and they tolerate skewed distributions (Breiman, 2001). The RF preprocessing therefore consisted only of missing-value imputation and the log-transformation of SalePrice, the latter applied to keep the prediction scale identical to MLR for fair comparison of error metrics.

4.2 Cross-validation, hyperparameter tuning, and final evaluation

To prevent leakage of test-set information into preprocessing parameters, stateful transformations (median imputation, Z-score normalisation) were fitted on the training set only and applied to the test set via the Apply variants of the Missing Value and Normalizer nodes (Apicella et al., 2025; Cawley & Talbot, 2010).

Hyperparameter selection and validation were conducted exclusively on the training partition using 10-fold cross-validation via the X-Partitioner and X-Aggregator nodes (Kohavi, 1995), with the same random seed across all variants for fair comparison. For Multiple Linear Regression, two variants were tested: (A) the primary specification with multicollinear features removed; and (B) a comparison variant retaining them. Both achieved CV $R^2 = 0.833$ with RMSE of 0.165 and 0.166 respectively, confirming that multicollinearity removal preserved accuracy whilst

improving coefficient interpretability. For Random Forest, a grid search was conducted over the number of trees ∈ {100, 300, 500} and maximum depth ∈ {10, 20, unlimited}, implemented as five parallel cross-validation chains with identical 10-fold partitioning and a fixed model-level random seed.

After hyperparameter selection, each final model was applied to the held-out test set and scored using the Numeric Scorer node. Interpretability was assessed via standardised MLR coefficients (directly comparable due to Z-scoring) and RF split-frequency statistics (Palczewska et al., 2013).

5. Results and Discussion

Tables 1 and 2 present the cross-validation results for each method's variants. Table 3 summarises the final test-set performance of the chosen models.

Table 1. Linear Regression cross-validation variant comparison.

Variant	Features	R^2	RMSE	MAE
A (primary)	Multicollinear removed	0.833	0.165	0.105
B	All retained	0.833	0.166	0.105

Table 2. Random Forest hyperparameter grid search (10-fold CV).

Variant	Trees	Max depth	R^2	RMSE	MAE
RF-1	100	Unlimited	0.814	0.174	0.117
RF-2	300	Unlimited	0.819	0.172	0.115
RF-3 (winner)	500	Unlimited	0.820	0.171	0.114
RF-4	300	10	0.629	0.246	0.174
RF-5	300	20	0.803	0.180	0.120

Table 3. Final model performance on training (CV) and held-out test sets.

Model	CV R^2	CV RMSE	Test R^2	Test RMSE
Linear Regression	0.833	0.165	0.863	0.139
Random Forest (500, unlim.)	0.820	0.171	0.862	0.139

5.1 Linear regression performance and interpretation

Multiple Linear Regression achieved cross-validation $R^2 = 0.833$ (RMSE = 0.165) and test-set $R^2 = 0.863$ (RMSE = 0.139). The slight improvement from CV to test indicates a well-calibrated model with no evidence of overfitting. The mean signed difference of -0.010 confirms minimal systematic bias, predictions are, on average, accurate in direction as well as magnitude.

This level of explained variance is attributable to three factors. First, the log-transformation of the right-skewed target aligned the response with the OLS assumption of approximately normally distributed residuals (Osborne, 2017), removing the heteroscedasticity that would have biased estimates on the raw scale. Second, removal of the three most multicollinear features stabilised the coefficient estimates, since OLS can produce wildly inflated standard errors when predictors are near-linearly dependent (Hastie et al., 2009). Third, the high-leverage categorical features, particularly Neighborhood and MSZoning, which capture substantial location-based price variation, were properly encoded as dummy variables and contributed strongly to fit.

5.2 Random forest performance and the depth/tree-count effects

Among the five Random Forest variants, 500 trees at unlimited depth gave the best CV performance ($R^2 = 0.820$, RMSE = 0.171). Applied to the test set, this configuration achieved $R^2 = 0.862$ and RMSE = 0.139. As with LR, the test-set performance slightly exceeded CV, indicating that RF also generalised without overfitting.

Two patterns in the grid-search results merit explanation. The diminishing-returns pattern from 100 to 500 trees (RMSE: 0.174 → 0.172 → 0.171) reflects the convergence behaviour predicted by Breiman (2001): adding more trees reduces ensemble variance, but with rapidly decreasing marginal benefit once a sufficient number of decorrelated trees has been averaged. By 500 trees the variance reduction has largely saturated, so further additions would yield negligible further accuracy at increasing computational cost.

More dramatically, restricting maximum depth to 10 substantially degraded performance (RMSE = 0.246, about 44% worse than the unlimited-depth baseline). This indicates that the true decision logic governing house prices on this dataset is deep, capturing localised concepts (e.g. specific combinations of neighbourhood, storey count, and garage type) requires more than ten sequential splits, and restricting depth forces the trees to average over too-broad regions of feature space.

5.3 Comparing linear regression and random forest

Linear regression slightly outperformed random forest in cross-validation (by 0.013 R^2 and 0.006 log-RMSE), but the two methods converged to essentially identical performance on the held-out test set (R^2 0.863 vs 0.862; RMSE 0.139 for both). Several factors explain this outcome. First, with only ~1,040 training observations the tree ensemble had limited data per split, particularly for deep levels of the trees. Second, after log-transformation the response appears to be approximately linear in the predictors, so little non-linear structure remains for the more flexible method to exploit, consistent with Grinsztajn et al.'s (2022) finding that well-engineered linear models match tree-based methods on small-to-medium tabular data. Third, the regression-to-mean behaviour of bagged ensembles (Breiman, 2001) compressed RF's predictions toward the centre of the price distribution, causing systematic under-prediction of the highest-priced houses visible in the test-set scatter (Figure 6).

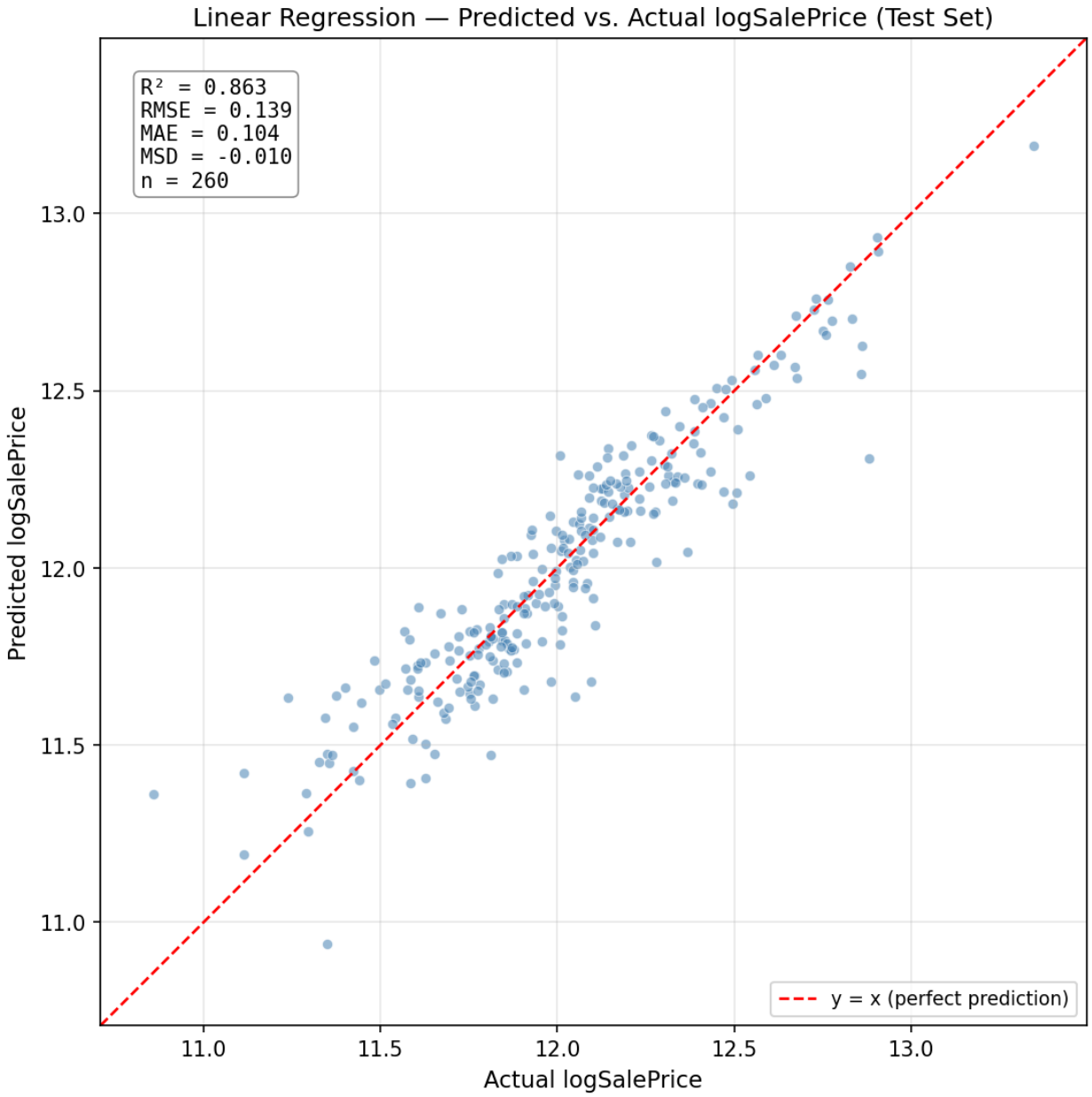

Figure 5. Predicted vs. actual logSalePrice on the test set, Linear Regression.

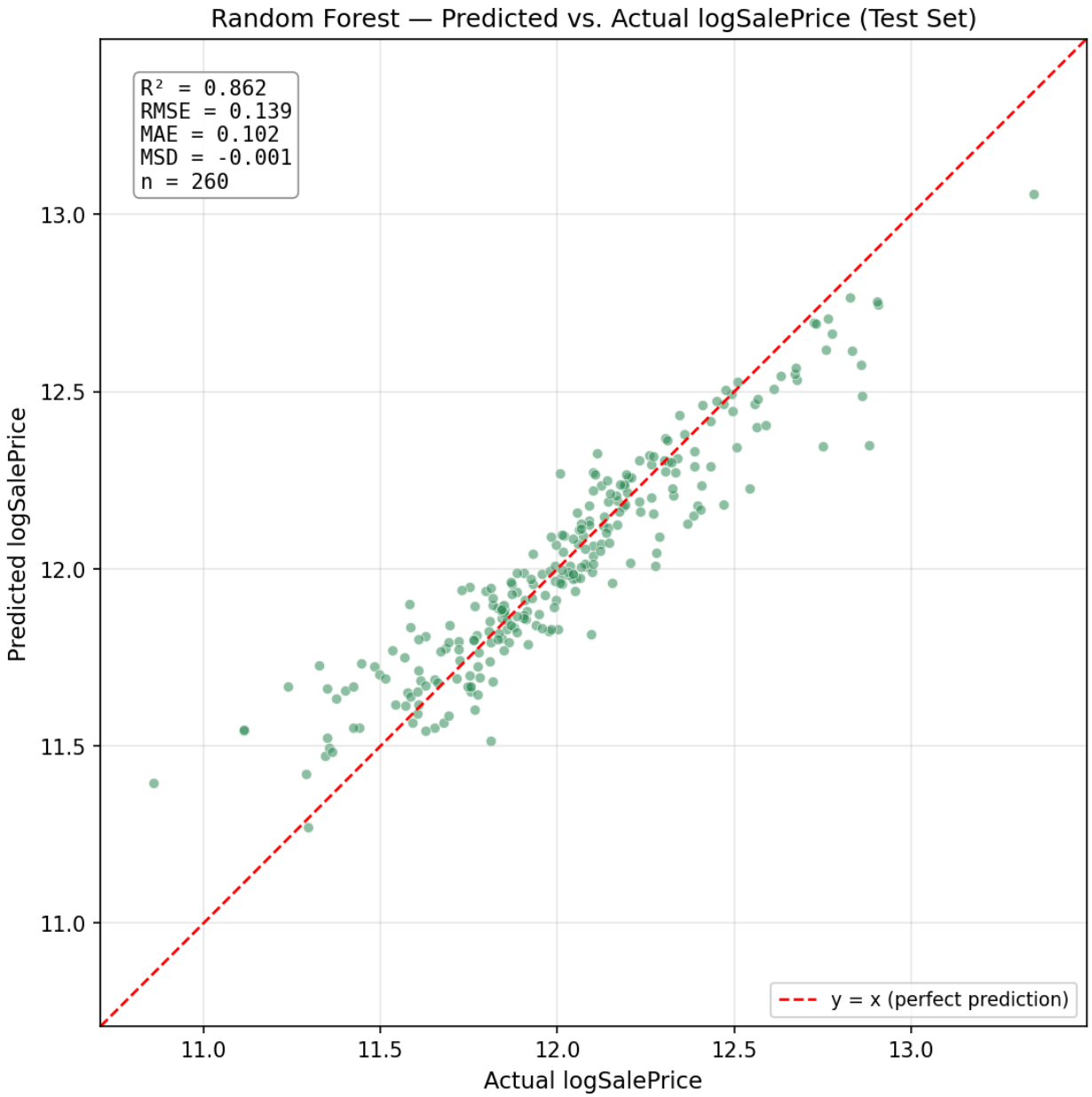

Figure 6. Predicted vs. actual logSalePrice on the test set, Random Forest.

5.4 Feature-importance comparison

The two methods agreed on the central importance of Neighborhood, Building Type, and Zoning, but differed in how they ranked continuous features. Linear regression identified GrLivArea (β = 0.118, standardised), OverallQual (β = 0.086), GarageCars (β = 0.060), and OverallCond (β = 0.041) as the strongest continuous predictors of log-price. The largest categorical effects were commercial zoning (C(all)_MSZoning, β = -0.431), premium neighbourhoods StoneBr (β = +0.255) and

NridgHt (β = +0.195), and very irregular lots (IR3_LotShape, β = -0.271). Random Forest split-frequency analysis, by contrast, prioritised high-cardinality categoricals, Neighborhood (315 splits), RoofStyle (267), HouseStyle (253), and MSZoning (229), with continuous predictors such as OverallQual (85) and GrLivArea (63) appearing further down. This divergence reflects the well-documented bias of the split-frequency measure toward variables with many possible split points (Palczewska et al., 2013; Hastie et al., 2009) and does not indicate that the continuous predictors are unimportant in absolute predictive terms.

6. Conclusion and Reflections

This report compared Multiple Linear Regression and Random Forest for predicting Ames, Iowa house sale prices. The principal finding is that, after appropriate preprocessing, log-transformation of the right-skewed target, median imputation, one-hot encoding with reference-dummy removal, multicollinearity reduction, and Z-score standardisation, Multiple Linear Regression and Random Forest produced essentially equivalent test-set performance (R^2 0.863 vs. 0.862; RMSE 0.139 for both), with LR holding only a small advantage in cross-validation (R^2 0.833 vs. 0.820). Both models generalised well from cross-validation to the held-out test set, evidence that the experimental design, strict train/test separation, hyperparameter tuning restricted to the training partition, fixed random seeds, successfully avoided overfitting and leakage.

Three substantive reflections emerge. First, the success of MLR confirms the value of careful preprocessing: the log-transformation alone reduced target skewness from 1.97 to 0.15 and aligned the response with the assumption of normally distributed residuals (Osborne, 2017). Second, the agreement between linear coefficients and random-forest split-frequency on the influence of Neighborhood and Zoning increases confidence that these are genuine signals. Third, the cardinality bias of split-frequency importance cautions against interpreting it as a direct measure of predictive contribution (Palczewska et al., 2013); MLR's coefficient-based interpretation offered a clearer signal regarding continuous predictors.

6.1 Methodological limitations

Several aspects of the experimental design warrant critical discussion. First, the final Random Forest model used for test prediction was the model produced by the last cross-validation fold rather than a fresh model refit on the full training set; while the performance difference is expected to be minor at this data volume, the methodologically correct procedure is to refit. Second, the grid search explored only five hyperparameter combinations; a finer grid, or principled approaches such as Bayesian optimisation, might identify configurations closer to the global optimum. Third, only a single random seed was used for the 80/20 partition: repeated partitions or nested cross-validation would tighten estimates of expected generalisation error and account for variance introduced by the single split (Cawley & Talbot, 2010). Fourth, feature engineering was limited, additional informative features could have been derived (for example, house age from YearBuilt, or interaction terms between Neighborhood and GrLivArea), and ordinal features such as OverallQual could have been encoded with monotonic transformations rather than treated as numeric. Finally, regularised linear regression (Ridge or LASSO) would have addressed multicollinearity automatically rather than through manual feature removal, and gradient-boosted ensembles such as XGBoost would constitute a stronger non-parametric comparator (Grinsztajn et al., 2022).

6.2 Data and ethical considerations

Predictive models for housing valuation carry non-trivial ethical implications. The strong predictive influence of Neighborhood, observed in both models, raises concerns familiar from the broader literature on algorithmic fairness: features that are predictive of price may also encode historical patterns of geographic and socioeconomic discrimination, including the documented legacy of redlining practices in the United States. A model assigning a 35% price discount to commercially-zoned properties or a 29% premium to a specific neighbourhood, if deployed in valuation or lending decisions, would reproduce and potentially amplify those historical patterns. Within the scope of this academic exercise the concerns are

theoretical, but in a real deployment mitigations would include excluding or carefully auditing geographic variables, applying fairness constraints during training, and validating across demographic subgroups.

Two further choices have ethical dimensions. Retaining rather than removing extreme outliers (the two large but inexpensive houses in Figure 3) preserves the empirical distribution but inflates measured error on those observations; a more rigorous treatment would investigate each case individually. Fixed random seeds and full documentation of the pipeline support reproducibility, a precondition for ethical scrutiny. Finally, generalisability is limited: the model is trained on Ames, Iowa data from a specific period and would require re-validation before application to other markets or time periods.

References

1. Antad, S., Bag, V., Waghmode, O., Wattamwar, S., Wagh, A., & Zite, A. (2024). Kisan Dhan - Crop Price Prediction Using Random Forest. Communications on Applied Nonlinear Analysis. https://doi.org/10.52783/cana.v31.762
2. Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *The Artificial Intelligence Review*, *58*(11), Article 339. https://doi.org/10.1007/s10462-025-11326-3
3. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.
5. 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. https://doi.org/10.5194/gmd-7-1247-2014
6. De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, *19*(3). https://doi.org/10.1080/10691898.2011.11889627

7. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* https://doi.org/10.48550/arxiv.2207.08815
8. Hastie, T., Friedman, J. H., & Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Second edition.). Springer. https://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=6314834
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 103). New York: springer.
10. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).
11. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.
12. Li, X. (2022). Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model. *Computational Intelligence and Neuroscience*, *2022*, Article 3590224. https://doi.org/10.1155/2022/3590224
13. Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (Third edition.). Wiley. https://onlinelibrary.wiley.com/book/10.1002/9781119482260
14. Osborne, J. W. (2017). *Regression & linear modeling : best practices and modern methods* (1st ed.). SAGE Publications, Inc.
15. Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013, August). Interpreting random forest models using a feature contribution method. In *2013 IEEE 14th international conference on Information Reuse & Integration (IRI)* (pp. 112-119). IEEE.
16. Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, *12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03
17. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079