Comparative Analysis of Linear Regression and Random Forest for Predicting House Prices

Prepared by 2441638

Abstract

This project evaluates the performance of two machine learning regression models, Linear Regression and Random Forest, in predicting housing prices based on real estate dataset. Linear Regression is an interpretable method, chosen for its simplicity and effectiveness in modelling linear relationships. Random Forest, an ensemble method was chosen for its robustness in handling complex and non-linear relationships. Both models were trained on the dataset and their performances were assessed using Mean Squared Error (MSE) and R^2. Random Forest outperformed linear regression, achieving lower MSE and higher R^2. This showed its ability to capture complex patterns in the data. This report provides a detailed analysis of the models, their advantages and disadvantages, and an evaluation on their performance. Strategies for improving both models are proposed, ensuring better results in future applications. The project demonstrates the value aligning model selection with dataset characteristics to achieve the best results has.

Keywords: Linear Regression, Random Forest, Machine Learning, Regression Models

Contents

1. Introduction

Machine Learning (ML) models are used in predictive tasks across different industries. The purpose of this project is to predict housing prices based on features such as the number of bedrooms, square footage and other property characteristics. Accurate predictions in this domain is necessary for decision-making in real estate markets, as they guide investors, sellers, and buyers.

Linear Regression (LR) and Random Forest (RF) were compared. The main goal was to evaluate the performance of these methods, identify their strength and limitations, and propose methods to improve their effectiveness. This report examines the background of both models and gives a detailed performance evaluation using metrics such as MSE and R^2. The mechanisms of each algorithm and their suitability for this specific regression task will be explained and a systematic approach to building, evaluating, and refining these models was followed.

2. Linear Regression

LR is a statistical model used to predict a continuous target variable based on one or more input features. The model assumes a linear relationship between the independent variables and the dependent variable. LR minimises the sum of squared residuals to find the best-fitting hyperplane, making it computationally efficient and interpretable (Montgomery et al., 2021). It works well for datasets where relationships between variables are linear.

2.1. Advantages of Linear Regression

The primary advantage of LR is its simplicity. The model is easy to understand, implement, and interpret, making it an excellent choice for tasks where model transparency is important (Montgomery et al., 2021). Each coefficient represents the expected change in the target variable for a one-unit change in the corresponding feature, holding all other features constant. This interpretability allows practitioners to derive actionable insights from the model's outputs (Anandhi and Nathiya, 2023). LR is computationally efficient, even for moderately large datasets. Its training process involves solving a system of linear equations, which is computationally inexpensive compared to iterative algorithms used in more complex models. Additionally, LR performs well when the assumptions of linearity, independence, and homoscedasticity are met. Regularisation techniques, such as L1 (Lasso) and L2 (Ridge), can further improve its robustness by reducing overfitting and addressing multicollinearity among features(Anandhi and Nathiya, 2023).

2.2. Limitations of Linear Regression

LR's reliance on a linear assumption is its most significant limitation. It struggles to model complex, non-linear relationships, leading to suboptimal performance on datasets with such characteristics (Montgomery et al., 2021). The model is also highly sensitive to outliers, as large deviations can disproportionately influence the fitted line, resulting in biased predictions. This sensitivity necessitates careful data preprocessing and outlier detection. Another limitation is the requirement for feature scaling. Differences in the magnitude of feature values can lead to misleading coefficient estimates, making standardisation or normalisation a prerequisite (Montgomery et al., 2021). LR also assumes that the input features are independent of each other. Multicollinearity, where features are highly correlated, can destabilise the model's coefficients, reducing the reliability of its predictions. Moreover, LR assumes that the variance of residuals is constant across all levels of the independent variables (homoscedasticity). Violations of this assumption can lead to inefficient estimates and biased standard errors (Anandhi and Nathiya, 2023).

3. Random Forest

RF, a supervised ML algorithm that works by constructing a multitude of decision trees during training and outputting the average of their predictions for regression tasks. It is an ensemble method. Meaning, it combines the predictions of several models to create a more accurate and stable result. Each tree in the RF is trained on a bootstrap sample, a random subset of the training data, and at each split, the model considers only a random subset of features. This randomness helps the model reduce variance and avoid overfitting (Rigatti, 2017).

The key mechanism behind RF is aggregation. By averaging the predictions of multiple trees, RF smoothens out individual errors, resulting in a model that is both accurate and resilient to overfitting.

3.1. Advantages of Random Forest

A key strength of RF is its ability to handle non-linear relationships and interactions between features. This makes it particularly effective for datasets where relationships between variables are complex and difficult to model using simpler algorithms. RF also has built-in mechanisms for reducing overfitting. By averaging predictions across multiple trees, it stabilises outputs and ensures better generalisation to unseen data. Additionally, RF does not require feature scaling or normalisation, as it splits data based on thresholds rather than relying on magnitude compar-

Random State	Linear Regression	Random Forest
Random_state=0	MSE: 0.000143, R^2: 0.753	MSE: 0.000074, R^2: 0.873
Random_state=3	MSE: 0.000148, R^2: 0.757	MSE: 0.000076, R^2: 0.875
Random_state=7	MSE: 0.000152, R^2: 0.751	MSE: 0.000075, R^2: 0.877
Random_state=10	MSE: 0.000146, R^2: 0.744	MSE: 0.000076, R^2: 0.867
Random_state=25	MSE: 0.000154, R^2: 0.745	MSE: 0.000081, R^2: 0.866
Random_state=42	MSE: 0.000148, R^2: 0.751	MSE: 0.000074, R^2: 0.875
Random_state=123	MSE: 0.000145, R^2: 0.750	MSE: 0.000075, R^2: 0.872
Random_state=2023	MSE: 0.000156, R^2: 0.748	MSE: 0.000082, R^2: 0.867
Random_state=88	MSE: 0.000153, R^2: 0.749	MSE: 0.000078, R^2: 0.872
Random_state=55	MSE: 0.000151, R^2: 0.742	MSE: 0.000080, R^2: 0.863
Average Score	**MSE:0.00015**, **R^2:0.749**	**MSE:0.0000772**, **R^2:0.871**

Table 1: Average MSE & R^2 of both Models (decimal point format)

isons. This makes it easy to apply without extensive preprocessing (Ao et al., 2019). Another advantage is the model's ability to assess feature importance. RF provides metrics that indicate the contribution of each feature to the overall prediction. These insights can be valuable for feature selection and for understanding the underlying structure of the data. Furthermore, RF performs well on high-dimensional datasets, where it effectively handles large numbers of features without significant performance degradation (Ao et al., 2019).

3.2. Limitations of Random Forest

RF is computationally intensive, requiring significant resources to train and store multiple decision trees, especially for large datasets (Ao et al., 2019). This can make it impractical for real-time applications or environments with limited computational power. RF is also less interpretable compared to simpler models like LR. While feature importance metrics provide some insight, the internal workings of the model, spread across numerous trees, are challenging to explain comprehensively. Another drawback is its sensitivity to hyperparameter settings. Parameters such as the number of trees, maximum depth, and minimum samples per leaf require careful tuning to achieve optimal performance. Poorly chosen hyperparameters can lead to overfitting or underfitting. Additionally, while RF is robust against overfitting compared to individual decision trees, it may still struggle with datasets containing high levels of noise or irrelevant features (Ao et al., 2019).

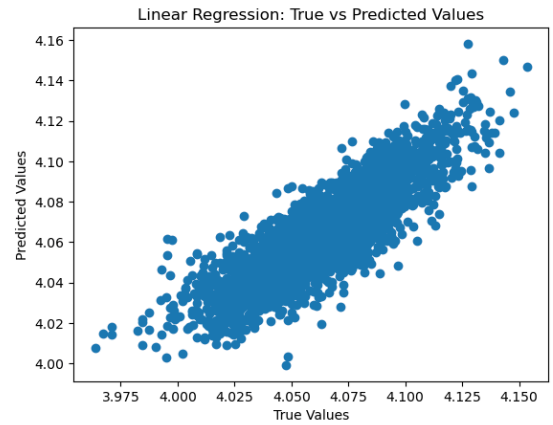
4. Evaluation of Model Performance

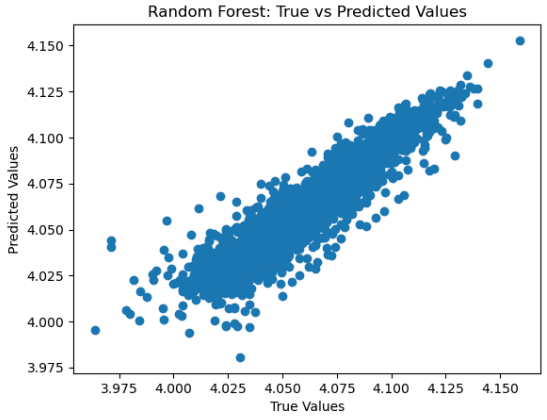
RF and LR were compared on this dataset to evaluate their strengths, limitations and overall performance. The evaluation of the models involved running the models ten times with different random states and measuring their performance using Mean Squared Error (MSE) and R^2 metrics. RF emerged as the better performing model, as it suited the datasets characteristics more. It achieved a lower MSE and high R^2 values, consistently outperforming LR across all iterations 1. This shows that RF was able to minimise prediction errors and explain the variance in the target variable better than LR. In one iteration, RF achieved an MSE of 7.40e-05 and an R^2 of 0.873, while LR recorded an MSE of 1.46e-04 and an R^2 of 0.751. These results highlight RF's capacity to capture the non-linear relationships present in the dataset. Its ensemble approach allows it to learn intricate patterns, even in the presence of noise. The code implementation further supports this, as feature importance metrics derived from the RF model identified significant predictors such as sqft_living and grade, which had strong correlations with the target variable. Conversely, LR's performance was constrained by its linear assumption. While it provided a good baseline model, its higher MSE and lower R^2 values suggest that it could not adequately capture the complexity of the dataset. This was particularly evident in the residual analysis, where the residuals showed patterns indicative of non-linear relationships that LR failed to model. The simplicity of the model and the reliance on scaled features were evident in the code outputs, where the lack of interaction terms or polynomial features limited its ability to generalise. When comparing both models, it becomes clear that the choice of algorithm significantly influences performance. RF's ability to model complex interactions contrasts sharply with LR's reliance on a single global hyperplane. However, RF's computational cost and lack of interpretability stand out as potential trade-offs. LR, despite its limitations, offered interpretability and simplicity, making it a valuable model for understanding the general trends in the dataset. Additionally, incorporating feature selection techniques can reduce computational complexity and enhance interpretability. For LR, introducing polynomial or interaction terms can help capture non-linear re-

Aspect	Linear Regression	Random Forest
Assumptions	Assumes a linear relationship between features and the target variable.	No assumptions about the relationship between features and the target variable. Handles non-linear patterns well.
Feature Scaling	Requires feature scaling (e.g., standardisation or normalisation) to ensure coefficients are comparable.	Does not require feature scaling; works with raw feature values.
Overfitting	Prone to underfitting or overfitting if regularisation (L1, L2) is not applied.	Less prone to overfitting due to averaging across multiple decision trees.
Interpretability	Highly interpretable; coefficients directly indicate the impact of features.	Low interpretability; the aggregated decision-making process across trees is difficult to explain.
Sensitivity to Outliers	Highly sensitive to outliers, which can disproportionately influence predictions.	Less sensitive to outliers; splits are based on feature thresholds.
Performance on Small Datasets	Performs well on small to medium-sized datasets if assumptions are met.	Performs adequately but may require careful tuning for small datasets.

Table 2: Comparison of Linear Regression and Random Forest Models

lationships; while addressing multicollinearity and outliers can stabilise coefficient estimates and improve performance. Both models would benefit from data preprocessing steps, such as scaling features, balancing classes, and augmenting the dataset with additional samples or synthetic data.





5. Suggestions for Improvement

To enhance the performance of RF, hyperparameter tuning should be prioritised. Adjusting parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), and minimum samples per leaf (min_samples_leaf) can improve the model's generalisation (Robnik-Šikonja, 2004). Additionally, feature selection could be used to reduce computational complexity and improve interpretability by focusing on the most influential predictors. Ensemble techniques, such as Gradient Boosting or XGBoost, could also be explored to refine the model further (Robnik-Šikonja, 2004).

For LR, introducing polynomial or interaction terms could help capture non-linear relationships within the dataset (Arpna and Dalal). For example, terms like sqft_living × grade could provide the model with more flexibility to approximate non-linear patterns. Addressing multicollinearity by using techniques like variance inflation factor analysis or ridge regression would stabilise coefficient estimates (Arpna and Dalal). Outlier detection and removal should also be performed, as outliers can disproportionately impact the model's performance. Data augmentation techniques, such as SMOTE, could be employed to enrich the training dataset, ensuring better generalisation across both models.

6. Conclusion

This study provides a comprehensive evaluation of RF and LR models for predicting house prices. The analysis highlights the strengths and weaknesses of both models, offering insights into their practical applications. RF's ability to model complex patterns and interactions allowed it to outperform LR in accuracy, as evidenced by lower MSE and higher R^2 values. This makes RF particularly suitable for tasks where predictive accuracy is paramount, such as real estate valuation or financial forecasting. LR, while less accurate, remains an essential tool due to its simplicity and interpretability. It is well-suited for applications where understanding the relationship between variables is crucial, such as identifying key drivers of house prices or evaluating policy impacts. The differences in model performance demonstrate the

importance of aligning model selection with the dataset's characteristics and the specific goals of the analysis. Future work should focus on applying targeted improvements to both models. For RF, optimising hyperparameters and integrating advanced ensemble techniques could further enhance its predictive power. For LR, incorporating non-linear features and addressing issues such as multicollinearity and outliers would improve its ability to generalise. Additionally, exploring hybrid approaches that combine the strengths of both models could provide a balanced solution, leveraging the accuracy of RF and the interpretability of LR. Ultimately, this study emphasises the need for a thoughtful approach to model selection and improvement, ensuring that the chosen methods align with the data and the problem's requirements. By addressing these considerations, both models can be refined to provide reliable and actionable insights in predictive analytics.

Bibliography

P Anandhi and E Nathiya. 2023. Application of linear regression with their advantages, disadvantages, assumption and limitations. *International Journal of Statistics and Applied Mathematics*, 8(6).

Yile Ao, Hongqi Li, Liping Zhu, Sikandar Ali, and Zhongguo Yang. 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering*, 174:776–789.

Nikhil Arpna and Surjeet Dalal. Optimizing linear regression model in water hardeness prediction for industry 4.0. *Trends in Mechatronics Systems: Industry 4.0 Perspectives*, page 73.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2021. *Introduction to Linear Regression Analysis*. Wiley.

Steven J Rigatti. 2017. Random forest. *Journal of Insurance Medicine*, 47(1):31–39.

Marko Robnik-Šikonja. 2004. Improving random forests. In *European conference on machine learning*, pages 359–370. Springer.