A Comparative Analysis of Multilinear Regression and the Random Forest Regression using the house pricing dataset.

Report by: Clifford Emmanuel Akai-Nettey

Introduction

This task involves building a regression model based on two different algorithms ie: Multi-Linear Regression (MLR) and Random Forest Regression (RFR). This report includes a comparative analysis of the two algorithms and proposes approaches to improve them. The dataset chosen for this study is the King County housing price data (named kc_house_data.csv) provided.

Accuracy comparison

The table below compares the accuracy of the models built with the two algorithms based on the following metrics:

- Root Mean Squared Error (RMSE)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R squared score (R^2)

The models were trained by varying the random states for the train test split 10 times.

Random state	RMSE (MLR)	RMSE (RFR)	MSE (MLR)	MSE (RFR)	MAE (MLR)	MAE (RFR)	R- squared (MLR)	R- squared (RFR)
0	186153.51	129775.67	34653130000.00	16841720000.00	108701.96	72307.38	0.71	0.86
1	222399.007	178286.44	49461320000.00	31786050000.00	115985.14	81009.89	0.71	0.82
2	201910.87	148452.69	40768000000.00	22038200000.00	113823.96	78126.74	0.71	0.84
3	197914.95	135532.54	39170330000.00	18369070000.00	112075.83	75918.13	0.70	0.86
4	202926.29	145521.93	41179080000.00	21176630000.00	115431.54	77289.68	0.68	0.84
5	206074.77	149481.27	42466810000.00	22344650000.00	115806.63	79433.46	0.69	0.84
6	204149.07	141313.44	41676840000.00	19969490000.00	114906.90	76566.99	0.69	0.85
7	205646.93	148696.43	42290660000.00	22110630000.00	117376.27	77434.36	0.68	0.83
8	200325.48	151949.95	40130300000.00	23088790000.00	113208.54	75885.83	0.72	0.84
9	208387.45	143573.00	43425330000.00	20613210000.00	115994.90	77246.87	0.69	0.85
10	203985.05	155058.02	41609900000.00	24042990000.00	112846.90	75728.24	0.72	0.84
Average	203624.85	147967.40	41530153974.81	22034675015.61	114196.23	76995.24	0.70	0.84

Observations and deductions

- Random Forest Outperforms Multi-Linear Regression:
    - The Random Forest model achieves significantly lower MSE (22 billion vs. 41 billion) and MAE (76,995 vs. 114,196). These metrics indicate that the Random Forest predictions are closer to the true target values compared to the Multi-Linear Regression model.
    - The R^2 score is higher for the Random Forest model (0.8416 vs. 0.6993), suggesting that it explains more variance in the target variable (house prices).
- Error Spread:

  The RMSE of Random Forest is also smaller (147,967 vs. 203,624), confirming that the overall errors are lower compared to Multi-Linear Regression.

Why Does Random Forest Perform Better?

- Non-Linearity Handling:
    - Multi-Linear Regression assumes a linear relationship between features and the target variable. However, housing prices often have complex, non-linear relationships (e.g., the effect of sqft_living or grade on price might not be strictly linear).
    - Random Forest, being a tree-based model, can capture these non-linear interactions effectively.
- Feature Interactions:
    - Random Forest can automatically account for feature interactions (e.g., how grade and sqft_living combine to influence price) without needing explicit feature engineering. Multi-Linear Regression would require manually adding interaction terms.
- Robustness to Outliers:
    - Housing datasets often contain outliers (e.g., extremely expensive houses). Linear regression is sensitive to such outliers, as they can disproportionately influence the coefficients. Random Forest, being based on decision trees, is more robust to outliers.
- Flexibility with Complex Data:
    - Random Forest can handle complex, high-dimensional data better because it partitions the data recursively to find patterns. Linear models struggle in such scenarios unless the relationships are explicitly specified.

Conclusion

The Random Forest model is the better choice for predicting house prices based on these metrics because it:

- Captures non-linear relationships.
- Handles outliers and feature interactions better.
- Provides lower error metrics and higher explanatory power (R^2).

However, Random Forest models are more computationally expensive and less interpretable than Multi-Linear Regression. If interpretability is a key requirement (e.g., understanding which features most influence prices), the Multi-Linear Regression can be preferred despite its lower performance.