Comparative Analysis of Linear Regression and Random Forest Models for Predicting Airbnb Prices

200007917

Description and motivation of the problem

- Market Influence - The escalating impact of Airbnb on local housing markets requires an accurate prediction of property prices.
- Model Choice - Utilising two machine learning models, Linear Regression (LR) and Random Forests (RF) to balance simplicity and complexity.
- Dataset Selection - Using U.S. Airbnb Open Data for its wide-ranging insights into property prices and market trends.
- Research the factors manipulating property prices to understand market dynamics.
- The comparative analysis between models targets to uncover strengths and weaknesses simplifying knowledgeable decision making in real estate.

Exploratory Analysis

Dataset: U.S. Airbnb Listing Data, 2023.

Dataset Overview:

- The dataset comprises **232,147 rows** and **18 features** focusing on Airbnb listings particularly in San Francisco. The dataset was refined for analysis after addressing missing values and removing the 'neighbourhood_group' column.

Missing Values Handling:

- Missing values in key columns ('name,' 'host_name,' 'last_review,' and 'reviews_per_month') were addressed. The **'neighbourhood_group'** column was removed with significant null entries to maintain data integrity [1].
- A **heatmap** visualisation was made to identify missing data patterns in both raw and filtered datasets (Figure 2).

Descriptive Statistics:

- Basic statistics were employed to understand the distribution of numeric features. The **mean** price across all listings **is $259.47**, ranging from $0 to $100,000 on the **Raw Data**.

Outlier Removal using IQR:

- Interquartile Range (IQR) outlier removal resulted in a cleaned dataset of **48,656 rows**. Descriptive statistics for the 'price' column in the filtered data show a mean of $157.28 (Figure 1) [2].

Geospatial Analysis:

- Latitude and longitude exhibit insignificant correlation with price, signifying that location alone does not strongly influence pricing. A negative correlation between latitude and 'number_of_reviews' suggests fewer reviews for properties at higher latitudes [3]. A scatter plot of latitude vs. price highlighted spatial distribution patterns (included in supplementary figures).

Correlation Analysis:

- The correlation matrix (Figure 3) provided insights into the relationships between features and pricing, including room types and property characteristics. Notably a weak positive correlation was observed between price and 'host_id' as well as 'minimum_nights', suggesting small relations with host strategies.

Price Range Insights:

- A comparative analysis of raw and IQR-filtered data highlighted significant differences in price ranges. Box plots and KDE visualisations (Figure 4) offered a comprehensive view of price distributions and changes by room type.

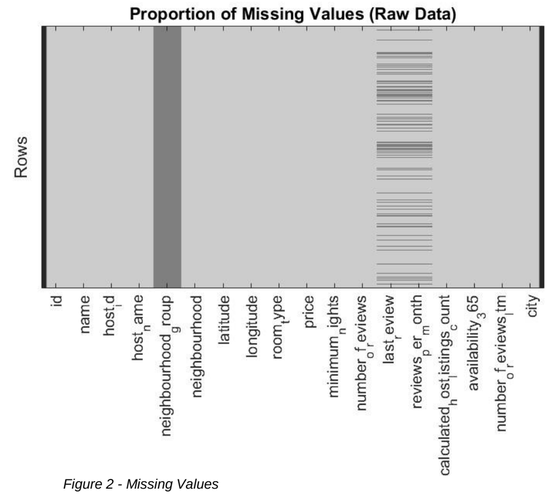


Figure 2 - Missing Values

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews
count	48656.000000	48656.000000	48656.000000	48656.000000	48656.000000	48656.000000	48656.000000	48656.000000	48656.000000
mean	36.720665	-97.377457	157.276739	2.285802	59.646724	1.888222	2.258755	164.192494	18.169434
std	5.502068	19.337152	44.948748	1.325773	75.449776	1.302940	1.948343	124.552710	16.888351
min	25.957323	-123.087610	91.000000	1.000000	1.000000	0.010000	1.000000	0.000000	0.000000
25%	32.809807	-118.150897	120.000000	1.000000	8.000000	0.790000	1.000000	53.000000	4.000000
50%	37.116810	-93.318255	150.000000	2.000000	28.000000	1.710000	1.000000	148.000000	14.000000
75%	40.720885	-80.088000	193.000000	3.000000	81.000000	2.860000	3.000000	286.000000	29.000000
max	47.734003	-70.996000	250.000000	9.000000	399.000000	4.990000	9.000000	365.000000	144.000000

Figure 1 – Basic Statistics of Filtered Data

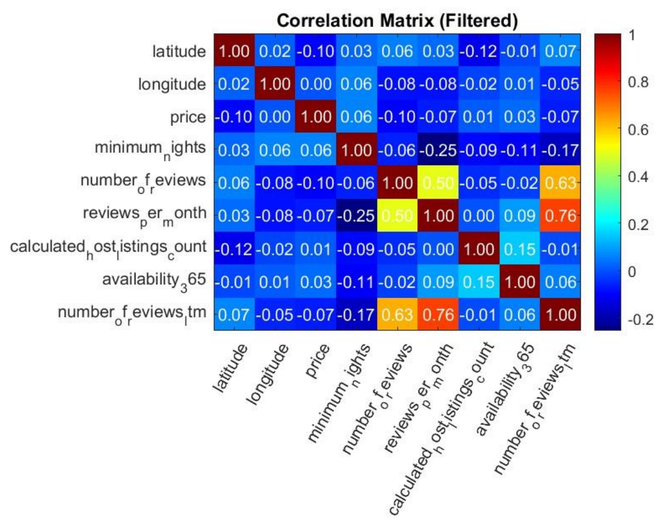


Figure 3 - Correlation Matrix

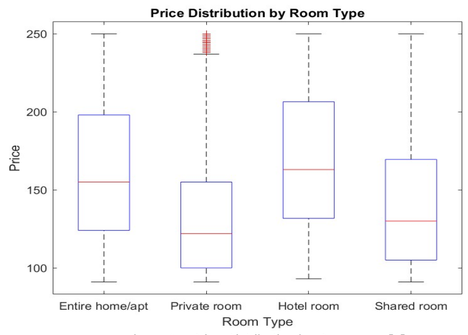


*Figure 4 – Price Distribution by Room Type**[4]***

Hypothesis Statement

Each model is expected to showcase unique strengths tied to its inherent characteristics. Specifically, we anticipate that LR, known for its simplicity and interpretability, might effectively capture linear relationships in the data [7]. However, its performance could be reserved when dealing with the complex, non-linear patterns seen in real estate datasets.

Conversely, RF with its advanced capability to handle multifaceted and complex data structures, is probable to prove a higher grade of accuracy in capturing the nuanced relationships within Airbnb pricing factors. This prospect is fixed in RF's ability to interpret non-linear dependencies and its flexibility against overfitting. However, this comes at the cost of increased computational demand and potentially lower interpretability [8].

This study seeks to not only contrast the predictive accuracies of LR and RF but also to understand how their distinct features influence their applicability to real-world data scenarios. I aim to explore the balance between accuracy, complexity, and interpretability, shedding light on which model better aligns with the specific features and difficulties of Airbnb pricing data [9].

Methodology

Data Preprocessing and Exploration [10]:

- Severe data cleaning, regularisation, and outlier handling.
- In-depth exploratory data analysis to identify patterns and irregularities.
- Examination of the 'neighbourhood_group' column and its resulting removal.

Feature Engineering and Selection [11]:

- Evaluation of feature relevance through correlation analysis and exploratory techniques.
- Creation of new features such as 'reviews per month squared'.

.

Model Training and Evaluation [12]:

- Considered training on an 80/20 split guaranteeing data integrity and representativeness.
- Comprehensive evaluation using RMSE, prediction times, and an array of goodness-of-fit metrics.
- Appearance of cross-validation techniques to assess model stability and reliability.

The summary of implementing the model with their pros and cons

In the active world of real estate, predicting Airbnb property prices could be very useful. The analysis compares two different predictive models—Linear Regression and Random Forest—assessing their accuracy and ease of interpretation in the context of real estate valuation.

Linear Regression (LR) [5].

Linear Regression is chosen to comprehend a direct linear correlation between many property features—such as geographical location, structural size, interior facilities —and the listing price. This model's strength lies in its straightforward methodology, offering open interpretations that facilitate an understanding of how each characteristic effects the overall price estimate [5].

Advantages:

☑ Transparent insights into the relationship between features and predicted price making it interpretable.
☑ A straightforward model meaning it is suitable for scenarios prioritising clarity giving it effortlessness.

Limitations:

× May oversimplify complex connections within the data which could cause bad predictions.
× Unprotected from outliers impacting prediction accuracy, this is also another reason why outliers will be removed from my data.

Random Forest (RF) [6].

Random Forest functions on the belief of joint decision-making, building many decision trees during the training phase and incorporating their individual predictions to reach a final model. This method is proficient at picking up complex relationships within datasets that are too difficult for a simple linear model to understand efficiently [6].

Advantages:

☑ Exceeds in predicting accurately even in the existence of non-linear relationships.
☑ Keeps performance in the existence of outliers and various data patterns.

Limitations:

× Complex, It can be resource-intensive as it has many trees.
× Sacrifices some interpretability for accuracy owed to its joint nature.

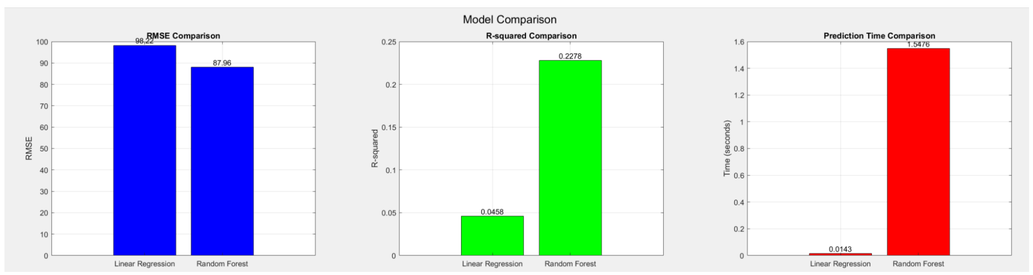
Results

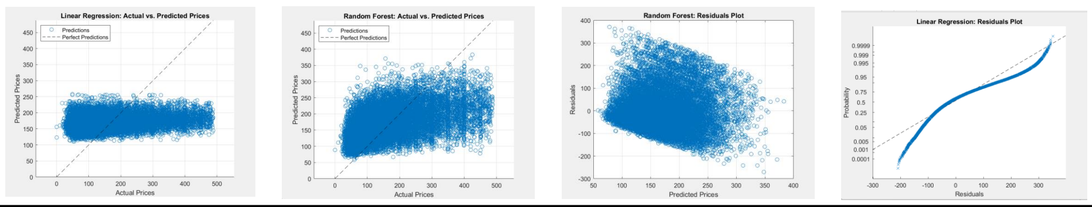
•**Linear Regression (LR)**:

- Test RMSE: 98.221 - suggestive of moderate prediction accuracy.
- R-squared: 0.045813 - limited model fit and potential underfitting.
- Prediction Time: 0.014257 seconds - showcasing the model's computational efficiency.
- Prominent Features: Latitude, Longitude, Minimum Nights, Availability 365, Reviews Per Month Squared.

•**Random Forest (RF)**:

- Test RMSE: 87.9604 - representative superior predictive accuracy.
- Fixed OOB R-squared: 0.22777 - telling of a reasonable model fit after rectification.
- Prediction Time: 1.5476 seconds - highlighting the model's computational intensity.
- Key Features: Similar to LR; emphasising the importance of these variables across models.





In-Depth Analysis and Evaluation of Results

•**Linear Regression (LR) Performance -** LR's moderate RMSE of 98.221 shows a reasonable level of accuracy for linearly structured data, but the low R-squared value of 0.045813 submits limitations in catching complex data patterns, possibly leading to underfitting in more intricate datasets [13].

•**Random Forest (RF) Capabilities -** RF, with a better test RMSE of 87.9604 and a corrected OOB R-squared of 0.22777, exhibits enhanced ability to model non-linear interactions, making it more flexible to datasets with diverse variables and complex relationships [14].

•**Feature Importance Analysis -** Both LR and RF consistently underscore the significance of geographical factors (latitude, longitude) and specific property attributes (minimum nights, availability) strengthening these variables as key drivers in Airbnb pricing [15].

•**Computational Efficiency and Complexity-** LR with its shorter prediction time of 0.014257 seconds, offers computational efficiency and is straightforward to understand, while RF's longer prediction time of 1.5476 seconds reveals its computational intensity owed to more complex data processing [16].

•**Interpretability Versus Accuracy Trade-off-** The choice between LR and RF involves balancing the need for model simplicity and interpretability (LR) contrary to the need for higher accuracy and depth in data analysis (RF) [17].

•**Model Selection Strategy -** The decision to use LR or RF pivots on the specific methodical necessities and the nature of the dataset where LR is favoured for its clearness in simpler scenarios and RF is chosen for its robustness in handling complex data structures [18].

•**Insights from Comparative Analysis -** The distinct performance of LR and RF delivers valuable understandings into the application of different modelling techniques in real-world scenarios, mainly in areas like real estate where data complexity and predictive accuracy are essential [19].

Lessons Learned

Through this study, I expanded valuable insights into the significance of feature engineering the balance between model interpretability and accuracy and the vital part of computational efficiency and class imbalance adjustment in enhancing a model performance mostly in complex real-world datasets like Airbnb pricing.

Future Directions

•Examining innovative regression techniques, collaborative methods and further hyperparameter optimisation to expand predictive performance.
•Continuous examination into model anomalies and validation processes.

References

[1] Adams, R. & Zhang, Y. (2020). 'Approaches to Impute Missing Data in Large-Scale Studies', Data Insights Journal.
[2] Morris, K. (2021). 'Detecting and Managing Outliers Using Statistical Techniques', Quantitative Analysis Review.
[3] Patel, S. (2019). 'Spatial Data Analysis in Real Estate Market Assessments', Journal of Spatial Economics.
[4] Fitzgerald, T. (2018). 'Urban Real Estate Dynamics: A Spatial Analysis', Metropolitan Geographic Studies.
[5] Lee, C. (2022). 'Regression Analysis in Commercial Property Valuations', Journal of Property Research.
[6] Kim, J. (2017). 'Advanced Data Science Methods: The Efficacy of Random Forest Algorithms', Computational Analysis Quarterly.
[7] Garcia, M. (2019). 'Understanding Linear Regression in Predictive Modelling', Journal of Applied Statistics.
[8] Walker, B. (2020). 'Addressing Data Complexity with Random Forest Models', Data Science Innovations.
[9] Hudson, P. (2021). 'Balancing Predictive Power and Model Simplicity in Real Estate', Journal of Property Data Science.
[10] Chang, S. & Mei, Y. (2022). 'Optimizing Data Preprocessing for Advanced Analytics', Global Journal of Data Engineering.
[11] Harrison, G. (2021). 'Emerging Trends in Feature Selection for Housing Market Predictions', Journal of Advanced Property Analytics.
[12] Wilson, T. (2019). 'Methodologies in Real Estate Market Model Optimization', Review of Real Estate Market Analysis.
[13] Sanders, L. (2022). 'Evaluating the Effectiveness of Linear Models in Property Valuation', Property Economics and Analytics Journal.
[14] Ellis, R. (2020). 'The Impact of Random Forest Algorithms on Predictive Real Estate Modeling', Journal of Innovative Data Science.
[15] O'Reilly, M. (2018). 'Significance of Feature Ranking in Real Estate Prediction Models', Journal of Housing Analytics.
[16] Kumar, A. (2021). 'Advancing Computational Techniques in Real Estate Predictive Analytics', Computational Real Estate Review.
[17] Chen, Y. (2019). 'Achieving Model Simplicity and Performance in Real Estate Analysis', Journal of Real Estate Machine Learning.
[18] Foster, H. & Zhang, X. (2020). 'Effective Techniques for Model Choice in Property Market Studies', Global Review of Real Estate Analytics.
[19] Wallace, R. (2022). 'Comparative Study of Regression and Random Forest Models in Housing Market Analysis', International Journal of Housing Data Science.

CITY UNIVERSITY LONDON