Model Card

Model name

Airbnb Price Prediction

Model date and version

Built December 2024. No numbered versioning. The model was updated on January 2025.
All parties registered as a user of this model will be informed by email if we release an update.

Overview Model type

The model is a geospatial analysis framework integrating OLS, SFE, GWR, and MGWR to predict Airbnb prices by capturing spatial and non-spatial determinants. OLS (with and without geographic features) establishes a baseline, while SFE accounts for neighborhood-level variations. GWR enhances accuracy by modeling spatial heterogeneity, and MGWR refines this further by incorporating multi-scale effects. Trained on 7,757 records from Inside Airbnb (http://insideairbnb.com).Collected in November 2024, the dataset represents the most recent information available at the time. The model analyzes location, neighborhood characteristics, and proximity to attractions, providing valuable insights for hosts, policymakers, and urban planners in Texas City's Airbnb market.

Questions or comments

Please send any questions to:

gyawalisanjiv@gmail.com

Primary intended users

The urban planners and policymakers to identify priorities in city development strategies and zoning regulations.
The property managers and Airbnb hosts may prioritize property maintenance and guest experience based on market demand.
The researchers in urban economics and geospatial analytics to analyze urban trends and spatial data for informed decision-making.

Primary intended uses

To identify spatial and neighborhood-level factors influencing Airbnb pricing and use that information to optimize rental strategies.
To support policy decisions on urban zoning, housing affordability, and tourism management through data-driven insights.
To provide property owners with insights into pricing strategies based on geographic and property features to maximize revenue potential.

Out of scope uses

Real-time pricing predictions are intended for reference purposes due to the limitations of static datasets. They should not be relied upon for precise or dynamic pricing decisions.

Direct pricing recommendations should not be used without further validation, as they may not reflect real-time market conditions.

Generalizing insights to cities with significantly different urban dynamics or tourism profiles may lead to inaccurate conclusions and should be approached with caution.

Limitations

The pricing model is based on data collected in November 2024 and may not reflect changes in market trends post-data collection.

The dataset is skewed towards affluent and centrally located neighborhoods, which may introduce bias in the results.

The model is specifically tailored to Austin, Texas, and may not generalize well to other cities without re-training.

High-priced outliers were excluded, which may limit insights into the luxury property segment.

The following factors can affect classifications:

Proximity to attractions (pricing accuracy may vary based on distance to key locations).

Room type and property features (e.g., bedrooms, bathrooms) (certain property types may be underrepresented in the dataset).

Localized neighborhood characteristics (some neighborhoods may have limited data, affecting prediction reliability).

Metrics

The model is optimized for spatial interpretability over pure predictive accuracy, meaning that it prioritizes understanding the geographic variation in relationships rather than just minimizing error.

The model accounts for spatial heterogeneity, ensuring that local variations in relationships are captured. However, this also means that results may not generalize well to areas outside the study region without recalibration

Multicollinearity among geographically weighted variables is addressed, but due to the localized nature of the analysis, some spatially correlated predictors may still introduce bias in certain regions.

Training and evaluation data

The model was trained using Airbnb listings data from Austin, Texas. The dataset includes information on property characteristics, pricing, and geographic coordinates. It was trained on listings with complete data for key explanatory variables, including the number of accommodations, bedrooms, beds. Missing values in these variables were removed prior to training.

Key predictors included accommodations, number of bedrooms, number of beds, and review scores. Categorical variables such as room type were one-hot encoded.

The model incorporated geographical data (spatial coordinates: longitude, latitude) of listings to account for spatial heterogeneity.

The optimal bandwidth selection was determined using cross-validation with the Golden Section Search (GSS) algorithm, implemented via the mgwr.sel_bw module.

The process involved iteratively testing different bandwidth sizes to find the one that minimized the Akaike Information Criterion corrected (AICc), ensuring an optimal balance between local and global model complexity. The selected bandwidth of 118 represents the spatial scale within which relationships between variables remain statistically significant, allowing the model to capture localized price variations while preventing overfitting.

The model prioritizes spatial interpretability over pure predictive accuracy, making it useful for understanding geographic variations in price determinants rather than solely for price prediction.

Quantitative analysis

The GWR model exhibited the highest explanatory power among the tested models, achieving an R^2 of 0.667 and an Adjusted R^2 of 0.626.

The Residual Sum of Squares (RSS) for GWR was 1031.753, considerably lower than 1790.88 in OLS without geographic features and 1562.68 in OLS with geographic features, confirming better error minimization and model efficiency in capturing spatial heterogeneity.

The Akaike Information Criterion corrected (AICc) score for GWR was 8045.236, significantly lower than AICc scores for OLS models (10,650.71 for OLS without geographic features and 9621.215 for OLS with geographic features).

The Bayesian Information Criterion (BIC) for GWR was 13,889.61, demonstrating improved model performance.

The GWR model proved to be the most robust, outperforming traditional OLS and fixed-effects models by incorporating spatial variations in Airbnb pricing determinants. The reduction in RSS, AICc, and improved R^2 values confirm its ability to model localized price variations, making it highly suitable for spatial economic analysis and urban policy planning.

While MGWR offers a theoretically more flexible model, in this case, GWR provides a better balance between accuracy, interpretability, and generalizability. The higher R^2, lower AICc, and better log-likelihood values in GWR confirm that it is the optimal model for capturing spatial variations in Airbnb pricing while remaining computationally efficient and practically useful.

Ethical considerations

The dataset used for this model was fully anonymized, ensuring compliance with GDPR and CCPA regulations. No personally identifiable information was included in the training and testing data.

The model does not incorporate sensitive attributes such as ethnicity, gender, or religious affiliation, and their potential impact on the results was not investigated.

Efforts were made to mitigate geographic and socio-economic biases in the dataset by ensuring a balanced representation of different neighborhood characteristics. However, some disparities in listing distribution may still exist.

The findings from this model should not be used to justify policies that could exacerbate housing affordability challenges or encourage gentrification.

The model's predictions should be interpreted with caution in policy decisions.

Further validation is required to determine whether the model generalizes well beyond Austin, Texas, especially in cities with different urban structures and housing policies.

Feedback

All users are encouraged to share concerns or comments about the performance, accuracy, or ethical implications of this model by reaching out via the provided feedback channel.

http://bit.ly/3WY2T7M

Any feedback regarding potential biases, geographic limitations, or policy impacts will be carefully reviewed. You should receive a response within 48 hours, detailing how your comments will be processed and whether they will be escalated to relevant teams for further investigation. Your input is valuable in ensuring the responsible and effective use of spatial modeling techniques.

Additional notes and any other relevant factors

None