Predicting Airbnb Homestay Prices Using Machine Learning

Jiaqin Wu

1 Executive Summary

This paper explores potential factors influencing Airbnb homestay prices, including institutional characteristics and demographic information, such as regional GDP per capita and regional house prices per squared meter. By understanding these factors, policymakers can develop targeted policies to maximize the benefits of Airbnb homestays for both hosts and guests while promoting sustainability and equity in the tourism industry. Additionally, policymakers can address any negative consequences, such as housing shortages or rising rents, by gaining insights into the potential impact of Airbnb on the local housing market.

To make predictions on Airbnb homestay prices, I utilized four machine learning techniques and tested the models on homestay data from Kaggle, which included information from 10 European cities. In addition, I incorporated demographic information from the IMF and Numbeo websites to enhance prediction accuracy. Although my model achieves a prediction accuracy of approximately 60%, there is still room for improvement through further feature engineering and exploration of alternative modeling techniques.

2 Background

Airbnb has become one of the most popular online marketplaces, providing people with a diverse range of options for short-term homestays and experiences. As travelers continue to seek out alternatives to traditional hotels, Airbnb homestays have emerged as a popular and cost-effective option (Dudás, Kovalcsik, Vida, Boros, & Nagy, 2020). Therefore, researching what variables influence Airbnb homestay prices and constructing a reliable machine-learning model to predict their prices is crucial.

From the market's perspective, identifying the variables that affect Airbnb homestay prices can offer valuable insights for investors, policymakers, and entrepreneurs to make informed decisions about their business strategies (Reinhold & Dolnicar, 2021). The research can contribute to our understanding of the economic impact of the sharing economy in the form of homestays (Toader, Negruşa, Bode, & Rus, 2022). Researchers can estimate the economic benefits and costs of the platform for different regions and sectors, providing valuable information for policymakers and investors.

From a stakeholder's perspective, Airbnb hosts and property managers can benefit from research on the variables that influence homestay prices by developing effective pricing strategies. By understanding the factors that affect Airbnb prices, hosts can adjust their pricing based on market demand and supply, competition, and other factors, leading to increased revenue and profitability (Toader et al., 2022). The research can also shed light on the factors that influence consumer behavior when selecting Airbnb accommodations, which can help hosts and property managers enhance the guest experience and meet the needs and preferences of their target audience (Hati, Balqiah, Hananto, & Yuliati, 2021).

From a policymaker's perspective, they can design regulations that balance the needs of different stakeholders, such as hosts, guests, and local communities by understanding the factors that affect homestay prices (Bivens, n.d.). Accurate predictive models can help policymakers adjust their strategies aimed at housing prices, which can be highly relevant to people's quality of life.

The development of accurate predictive models for studying Airbnb homestay prices is crucial, especially in today's sharing economy where demand for affordable and comfortable accommodation is increasing (Michael, 2015). In this paper, we analyze Airbnb homestay information and prices from various European countries to construct machine-learning models. If policymakers have access to these models, they can adjust their strategies aimed at housing prices and regulations to enhance tourism, promote sustainable economic growth, and improve residents' quality of life (Hati et al., 2021). Additionally, policymakers can use the model to assess the economic impact of the sharing economy on local communities and the effects of different regulatory approaches on the industry (Toader et al., 2022). Ultimately,

the development of a reliable prediction model for Airbnb homestay prices can benefit all stakeholders involved in the sharing economy, including hosts, guests, investors, and policymakers (Hall et al., 2022).

3 Data

This paper examines how various factors affect homestay prices, which are determined by pricing strategies for optimal profitability. The dataset used in this study is from Kaggle (The Devastator, 2023) and contains information on the 10 most popular European cities. Each listing is evaluated for different attributes such as room types, cleanliness, satisfaction ratings, bedrooms, distance from the city center, and more to capture a comprehensive understanding of Airbnb prices on both weekdays and weekends. To construct our machine-learning models, we added potentially region-related economic indicators to this dataset, including the average price per square meter to buy apartments and GDP per capita in each country. The housing price per square meter data is from NUMBEO (*Numbeo*, 2023), and the GDP per capita data is from IMF (Fund, 2023).

To prepare the dataset for analysis, we merged 20 separate CSV files into one and added two extra columns named 'week_time' and 'city.' The final dataset contains 23 columns, which we divided into two groups: continuous and categorical variables. The descriptive analyses of the continuous variables in the data set are shown in Table 1. We can see our target variable 'realSum' is a continuous variable with a mean value of 279.88. The descriptive analyses of the part of categorical variables in the data set are shown in Table 2. Since some data are discrete, we regarded them as categorical variables here. The *observations* column shows the distribution of each variable.

Table 1: Descriptive Statistics for Continuous Variables (N=51707)

Variable Label	Description	Mean	Med.	Std.	Min.	Max.
realSum	The total price of the Airbnb listing	279.88	211.343	327.948	34.779	18545.45
dist	The distance from the city center	3.191	2.614	2.394	0.015	25.285
metro_dist	The distance from the nearest metro station	0.682	0.413	0.858	0.002	14.274
attr_index_norm	Normalized attraction index (0-100)	13.423	11.468	9.81	0.926	100
rest_index_norm	Normalized restaurant index (0-100)	22.786	17.542	17.804	0.593	100

Before running the models, I explored the relationships between several features and the target variable 'Real_sum'. I expected some associations to exist in my data, and the visualizations tested my assumptions.

Firstly, I created a scatter plot to explore the relationship between the distance to the city center and the homestay price (shown in Figure 1). The figure demonstrates that the homestay prices decrease as the distance to the city center increases. Moreover, the homestays in London and Berlin are located farther away from the city center than the homestays in other cities. Similarly, I created a scatter plot to visualize the relationship between the distance to the nearest metro station and the homestay price (shown in Figure 2). The findings indicate that the relationship between the distance to the metro station and the homestay price is almost the same as the relationship between the distance to the city center and the homestay price.

To gain further insights into the relationship between cities and homestay prices, we created two visualizations, Figure 3 and Figure 4. Figure 3 shows a linear relationship between GDP per capita and mean homestay price, where the higher the GDP per capita, the higher the mean homestay price. However, Rome, Berlin, and Vienna are three exceptions, as they have relatively higher GDP per capita but lower homestay prices. In Figure 4, we observe a partly linear relationship between house price per m2 and mean homestay price, where the higher the house price per m2, the higher the mean homestay price. However, Lisbon, Rome, Vienna, and Berlin are four exceptions, as they have relatively higher house prices per m2 but lower mean homestay prices.

In addition, I also investigated the impact of week time on mean homestay prices. As depicted in Figure 5, the majority of homestay prices on weekends were found to be higher than those on weekdays, except for Budapest and Barcelona. Notably, in Athens, the homestay prices on weekends were significantly higher than those on weekdays.

The institutional characteristics of homestays in each city can also contribute to the price difference. To compare these characteristics, I analyzed the superhost proportion, cleanliness rate, and historical guest satisfaction of homestays in each city through visualizations in bar plots. The plots are shown in

Table 2: Descriptive Statistics for Part of Categorical Variables (N=51707)

Variable Label	Description	Observations
room_type	The type of room being offered	Entire home/apt: 32648
		Private room: 18693
		Shared room: 366
room_shared	Whether the room is shared or not	False: 51341
		True: 366
room_private	Whether the room is private or not	False: 33014
		True: 18693
person_capacity	The maximum number of people that can stay in the room	2 rooms: 24333
		3 rooms: 6165
		4 rooms: 14000
		5 rooms: 2935
		6 rooms: 4274
host_is_superhost	Whether the host is a superhost or not	False: 38475
		True: 13232
multi	Whether the listing is for multiple rooms or not	False: 36642
		True: 15065
biz	Whether the listing is for business purposes or not	False: 33599
		True: 18108
cleanliness_rating	The cleanliness rating of the listing	2: 143
		3: 10
		4: 143
		5: 86
		6: 501
		7: 947
		8: 4352
		9: 15458
		10: 30067
week_time	Whether the listing is for weekday or weekend	weekends: 26207
		weekdays:25500

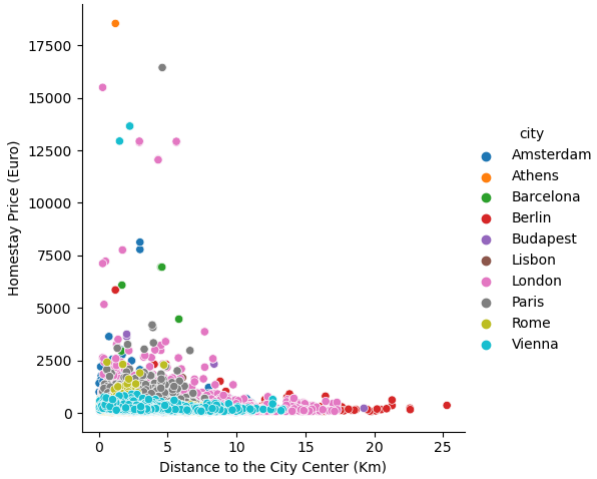

Figure 1: Relationship between the distance to the city center and homestay price

Figure 6, 7, and 8. Based on these plots, we can observe that Athens has the highest scores in all these variables, which may increase the homestay price to some extent. On the other hand, Paris has the lowest scores in these variables, which may slightly decrease the original homestay price. Overall, most of the homestays in these cities have relatively high cleanliness rates and guest satisfaction rates.

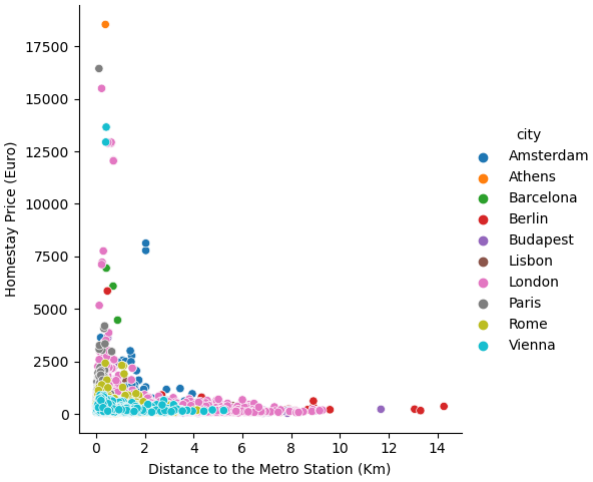

Figure 2: Relationship between the distance to the metro station and homestay price

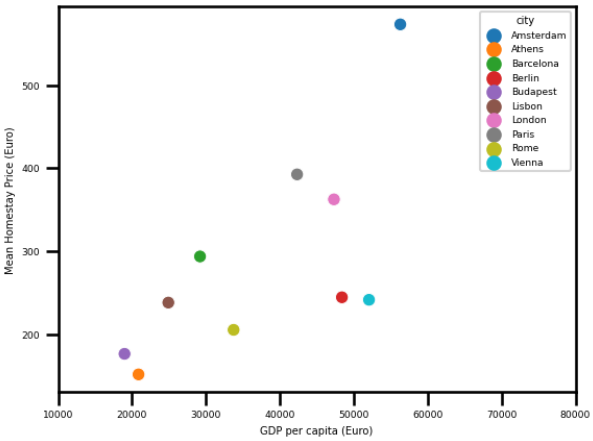

Figure 3: Relationship between GDP per capita and mean homestay price

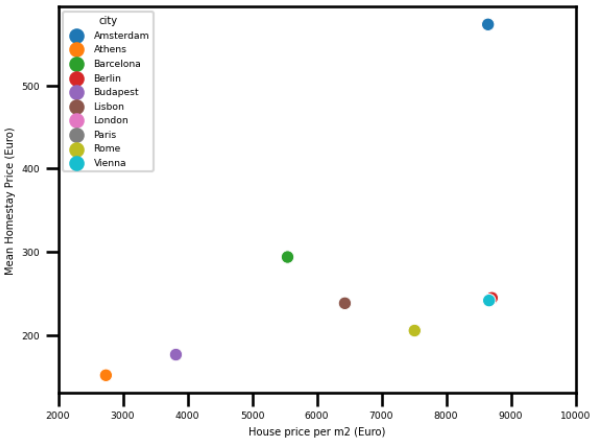

Figure 4: Relationship between house price per m2 and mean homestay price

4 Methodology

My project consisted of both prediction and inference analysis in the context of Airbnb homestay prices. The goal was to utilize machine learning models to predict homestay prices and to identify the characteristics that had an association with high or low homestay prices. The target array for all the mod-

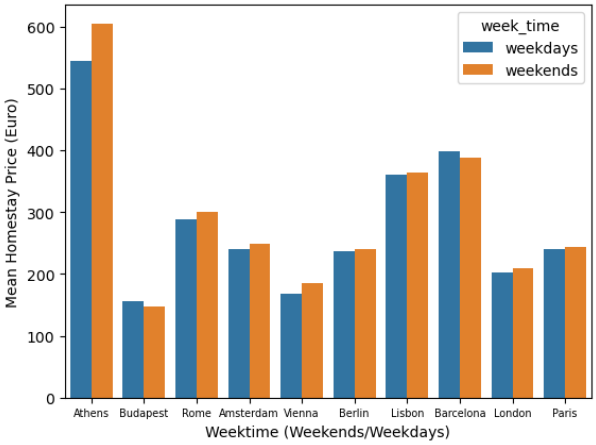

Figure 5: Relationship between week time and mean homestay price

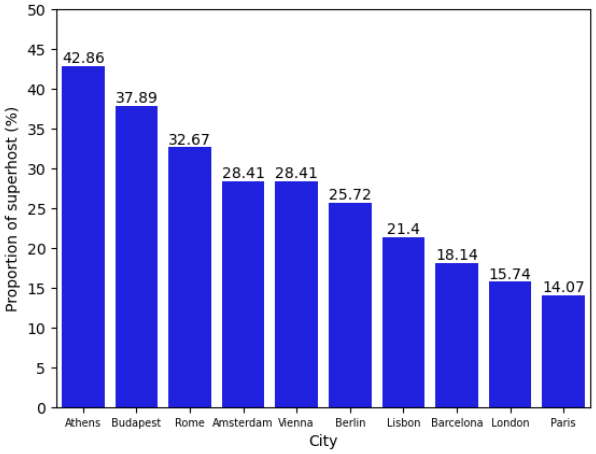

Figure 6: Superhost proportion of homestay in each city

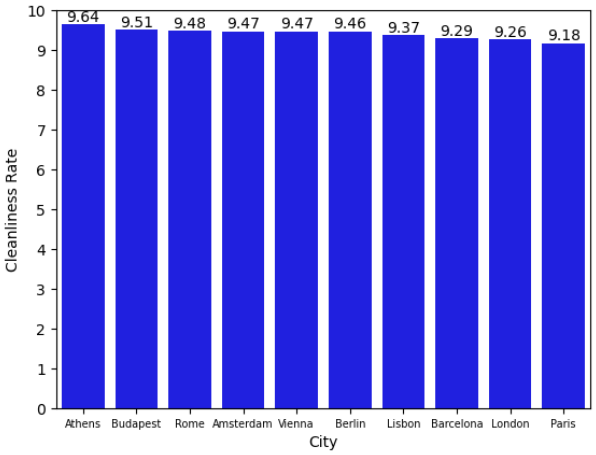

Figure 7: Mean cleanness rate of homestay in each city

els was the homestay prices ('realSum'), and the feature variables included the maximum number of guests ('person_capacity'), ownership of the listing by hosts ('multi', 'biz'), cleanliness rating ('cleanliness_rating'), guest satisfaction rate ('guest_satisfaction_overall'), number of bedrooms ('bedrooms'), distance from the city center ('dist'), distance from the metro station ('metro_dist'), attraction index ('attr_index_norm'), restaurant index ('rest_index_norm'), housing price per squared meter ('price_m2'),

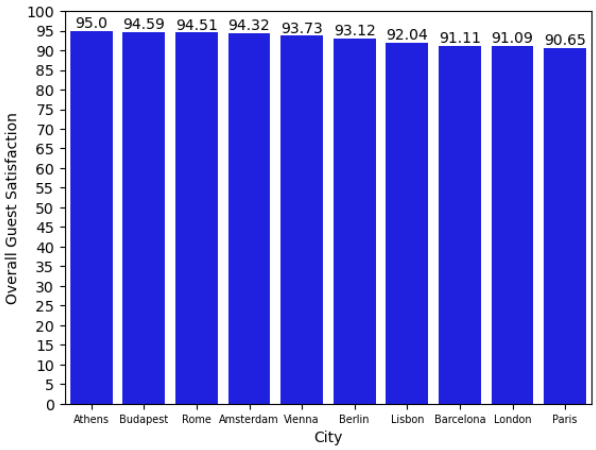

Figure 8: Mean guest satisfaction rate of homestay in each city

GDP per capita ('GDP_capita'), room type ('room_type_Private room', 'room_type_Shared room'), type of the host ('host_is_superhost_True'), and week time ('week_time_weekends').

4.1 Prediction

In order to predict more accurate homestay prices, I chose both parametric and non-parametric models, including linear regression models (Ridge and Lasso), Decision Tree model, Random Forest model, and XGBoost model, and assessed which had the lowest error. The desired final output was a continuous attribute representing the predicted homestay price based on a given set of features. The intuition behind each model's algorithm follows:

- Linear Regression Model: Linear regression is a statistical method used to predict a numeric value by fitting a linear equation to observed data. The equation minimizes the squared differences between predicted and actual values, and takes the form of Y = b0 + b1X1 + b2X2 + ... + bnXn, where Y is the predicted value, X1 through Xn are input variables, and b0 through bn are the coefficients. Lasso and Ridge algorithms are linear regression models that help prevent overfitting by adding a penalty term to the equation based on the input variable coefficients. This encourages the model to select a smaller subset of important variables, making it more useful for models with many variables.

  One example of the application of linear regression is a study published in the Journal of Medical (Baltimore), which used linear regression to analyze the long-term survival time and associated factors among patients with HIV/AIDS who received ART in Liangshan Prefecture for the first time (Zhang et al., 2016).

- Decision Trees: decision trees model is a type of non-parametric machine learning algorithm that predicts a numeric value by recursively splitting the data based on the values of input variables. Each split is chosen to maximize the reduction in the variance of the target variable. The final predicted value is the mean of the target variable for the leaf node to which the input data point belongs.

  An example of the application of decision trees is a study published at the 2021 5th International Conference, which used decision trees to predict solar energy production based on various features (Rahul, Gupta, Bansal, & Roy, 2021).

- Random Forest: random forest model is an ensemble method that combines multiple decision trees to improve prediction accuracy. It works by randomly selecting subsets of the input data and input variables and building decision trees on each subset. The final predicted value is the average of the predicted values of all the trees.

  An example of the application of random forest is a study published in the Journal of AGU Advances, which used the random forest to predict global marine sediment density based on various features (Graw, Wood, & Phrampus, 2021).

- XGBoost: XGBoost model is an ensemble method that combines the predictions of many individual models, typically decision trees, to improve overall prediction accuracy. The intuition behind the XGBoost model is to iteratively add new decision trees to the model, with each tree attempting to correct the errors of the previous trees. The XGBoost model also incorporates a regularization term in the objective function to prevent overfitting.

  An example of the application of XGBoost is a study published in the Journal of Interactive Learning Environments, which used XGBoost to predict the students' academic performances based on various features such as gender, age, previous academic performance, attendance, etc. (Asselman, Khaldi, & Aammou, 2021).

4.2 Inference

In addition to predicting homestay prices, I also aimed to conduct an inference analysis to identify the variables that had the highest impact on homestay prices. While we had some prior common sense about the variables that could potentially influence homestay prices, using machine learning models helped to verify our assumptions and provide more accurate information. By identifying the most significant variables, policymakers could use this information to make informed decisions and adjust policies related to housing. This approach can lead to better policies that are more aligned with the actual drivers of homestay prices, ultimately benefiting both homeowners and travelers.

5 Findings

I first divided the data set into the training set and the test set, in which the training set accounted for 80%. Since the target variable I predicted is continuous, I chose 5 indicators including MAE, MSE, RMSE, R2 score, and RMSE(Cross-Validation) to measure the model performance. I also adjusted the parameters of the models to achieve their best prediction performance and compared these six models' best performances. Part of the performance results of machine-learning models is below in Table 3.

Table 3: Performance Results of Prediction Models

Model	Training RMSE	Training R2 Score	Test RMSE	Test R2 Score	RMSE (Cross-Validation)
XGBoost	197.29	0.645	198.30	0.601	284.55
Random Forest	256.24	0.402	250.27	0.364	286.47
Decision Tree	271.77	0.327	244.06	0.396	294.98
Ridge	292.79	0.219	275.19	0.232	321.30
Lasso	292.79	0.219	275.19	0.232	325.87
LinearRegression	292.79	0.219	275.19	0.232	325.87

The linear regression model had the lowest accuracy, indicating that it could only predict around 21.9% of homestay prices. In contrast, the Lasso and Ridge models performed slightly better, as they regularized some of the variables in the model and reduced some of the coefficients. The RMSE and RMSE (Cross-Validation) values for all three models are very similar, suggesting that the linear regression model is not suitable for handling complex datasets. Furthermore, all three models were underfitted because the R2 score of the training dataset was lower than that of the test dataset, which further supports the idea that linear regression models are not ideal for predicting homestay prices.

The decision tree model achieved the highest R2 score of 0.402, indicating that nearly forty percent of homestay prices can be accurately predicted. However, the validation curve displayed in Figure 9 reveals that the decision tree model performs poorly. Until the max depth reaches 5, the testing R2 score remains higher than the training R2 score, suggesting that the decision tree model is underfitted. After the max depth exceeds 5, the decision tree model becomes overfitted, with the training R2 score significantly higher than the testing R2 score. We selected a max depth of 5 for the model, which is still underfitted, but it performs best at this level. The R2 score, RMSE, and RMSE (Cross-Validation) values are all better than those of the linear regression model.

The random forest model outperforms the decision tree model as it involves multiple trees. The validation curve shown in Figure 10 indicates that the model becomes underfitted before reaching a max depth of 2, while after a max depth of 6, it appears to be overfitted. Therefore, I chose a max depth of 5 to achieve the best performance. Although the random forest model's prediction accuracy of around 36.4% is slightly lower than that of the decision tree model, its RMSE after cross-validation is better, indicating

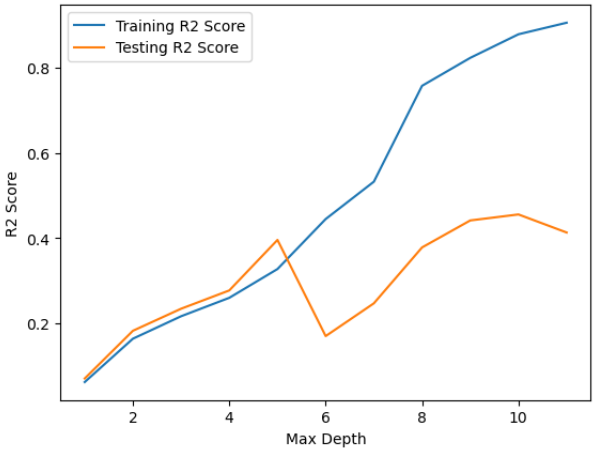

Figure 9: Validation Curve for Decision Tree Regressor

better overall performance. Moreover, the random forest model can handle complicated datasets well and select more important features, reducing the impact of noisy or irrelevant features. Therefore, it is a more reliable model for predicting homestay prices than the decision tree model.

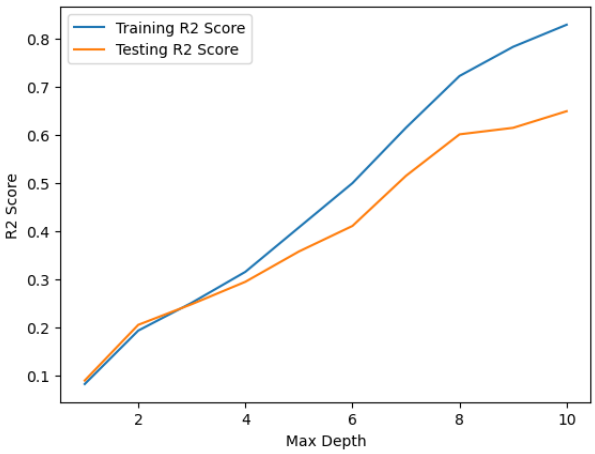

Figure 10: Validation Curve for Random Forest Regressor

The XGBoost model demonstrated the best performance among the six models evaluated. Figure 11 depicts the prediction accuracy of the XGBoost model with different estimators ranging from 100 to 1000. It is observed that the XGBoost model performs well when the estimators are between 100 to 400. However, when the estimators are above 400, the XGBoost model becomes overfitted. Based on this observation, the estimators were selected as 400 to train the final XGBoost model. The final XGBoost model demonstrated an accuracy of over 60%, which is significantly better than the other models. The reason for the superior performance of the XGBoost model could be attributed to its algorithm, which is a gradient-boosting algorithm. Gradient boosting is an ensemble method that combines multiple weak models, usually decision trees, to create a strong model. Additionally, the XGBoost algorithm adds new models to the ensemble, and each model attempts to correct the errors made by the previous models. This approach allows the XGBoost model to have superior performance compared to other models.

Although the homestay price prediction models did not perform well, the feature importance analysis can still provide some useful insights. Figure 12 displays the feature importance of the decision tree, random forest, and XGBoost models. Surprisingly, the decision tree model gave more weight to fewer features, while the XGBoost model distributed similar weights among more features. Notably, variables such as the number of bedrooms, the attraction index of the homestay, and the latitude and longitude had a significant impact on the homestay prices. Demographic information such as GDP per capita and housing prices per square meter were also found to be influential. On the other hand, variables such as week time, cleanliness rate, and host condition, although appearing to have an effect on the homestay, had

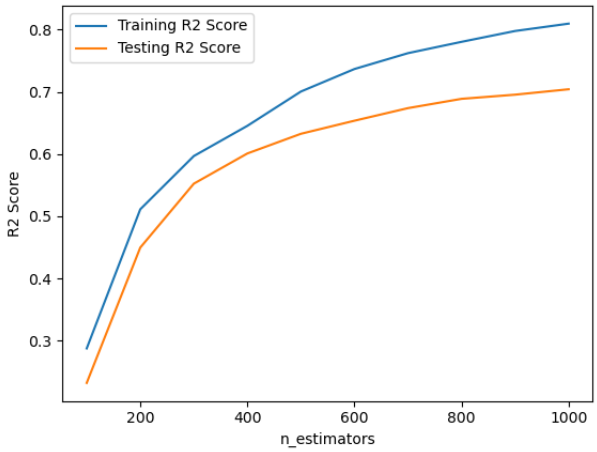

Figure 11: XGBoost Model Performance for Different Estimators

low importance scores in some of the models. Based on the feature importance analysis, the number of bedrooms and the attraction index of the homestay appear to have the greatest impact on the estimated prices of homestays.

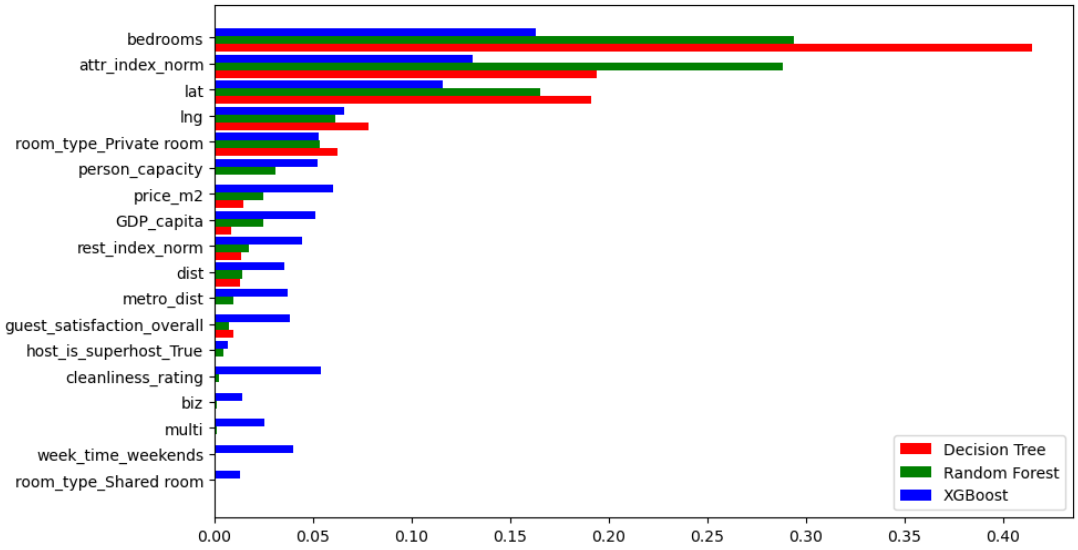

Figure 12: Comparison of Feature Importance of Different Models

While the XGBoost model showed the highest prediction accuracy of around 60% among the tested models, it still had limited precision in predicting homestay prices. Nevertheless, the feature importance analysis of the XGBoost model can provide useful insights for policymakers to understand which factors influence homestay prices, such as the number of bedrooms and attraction index. This information can be used to inform policy decisions related to the homestay industry.

6 Conclusion

While my current models have achieved a 60% prediction accuracy, it is clear that the complex nature of homestay pricing requires further refinement of the models. There are various factors that can impact homestay prices, beyond the homestay characteristics and geographic information included in our dataset. For instance, institutional variables like the homestay's neighborhood or the presence of specific amenities like bathtubs can affect pricing. Furthermore, demographic information like the impact of COVID-19, currency inflation rates, and local economic conditions may also be relevant.

To improve the accuracy of our models, it is crucial to include more comprehensive variables in our

dataset. By incorporating a broader range of institutional and demographic factors, we can better capture the nuances of homestay pricing and achieve more accurate predictions. This can be accomplished by leveraging various data sources and employing advanced analytical techniques, such as natural language processing and sentiment analysis. Ultimately, a more refined model can help homestay owners optimize their pricing strategies and provide guests with more transparent and competitive pricing options.

6.1 Limitation

The performance and effectiveness of our machine learning models are heavily dependent on the quality and relevance of the data used for training. While it's true that the absence of specific institutional variables can restrict the accuracy of our models, it's essential to acknowledge that demographic factors also play a significant role in determining the quality of our dataset.

Regarding institutional variables, the inclusion of additional characteristics can indeed lead to a higher homestay price. For instance, amenities such as bathtubs and swimming pools can significantly increase the value of a homestay. However, the collection of this type of information can be challenging due to the limited availability of homestays with these features, and the high costs associated with data collection.

On the other hand, demographic variables can also play a critical role in the effectiveness of our machine learning models. When considering demographic information, it's important to take into account more factors such as the income level of the homestay owner, the location of the property, and the crime rates around the homestay. Furthermore, it's essential to recognize that differences can exist even within the same city, such as varying housing prices and GDP levels in different regions. However, collecting the regional information could be very challenging, which also limits the accuracy of our prediction models.

6.2 Policy Implications

Despite the current limitations of my machine learning model, it can still provide valuable insights for policymakers to make informed decisions regarding homestay regulations, taxation policies, and urban planning.

With regard to homestay regulations and taxation policies, policymakers can leverage these models to set fair pricing practices and establish appropriate tax rates for homestay owners based on predicted prices. By doing so, policymakers can help maintain the competitiveness of the homestay market while ensuring that homestay owners are not unfairly burdened by taxes.

Moreover, these models can also be used to inform urban planning policies. By understanding the factors that influence homestay prices, policymakers can make informed decisions about zoning regulations and infrastructure planning that take into account the needs of both homestay owners and guests. For example, policymakers can determine where to allocate resources for infrastructure development based on the predicted demand for homestays in different areas. Additionally, these models can be used to analyze the potential impact of new developments on the homestay market and inform decisions about whether to incentivize or restrict such developments.

6.3 Future considerations

To further enhance the accuracy of my machine learning models for Airbnb homestay price prediction, there are several factors that should be taken into consideration in the future.

Firstly, I should explore more advanced approaches to feature engineering, such as deep learning or natural language processing, in order to extract more meaningful information from the available data. Secondly, experimenting with different types of models, such as neural networks or SVM models, can help me identify the most suitable model for this specific problem.

Moreover, it is important to stay up-to-date with any changes or updates to the Airbnb platform or homestay regulations, as these may impact the factors that influence homestay prices or the data that is available for analysis. This can be achieved through continuous monitoring of the platform and relevant industry news sources.

Finally, I should also consider incorporating more external data sources, such as housing prices in each region and distance to tourist attractions into my models, as these probably have an impact on homestay prices. This can help to refine and enrich my models and provide more accurate predictions for Airbnb homestay prices.

References

Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning xgboost algorithm. *Interactive Learning Environments*, *0*(0), 1-20. Retrieved from https://doi.org/10.1080/10494820.2021.1928235 DOI: 10.1080/10494820.2021.1928235

Bivens, J. (n.d.). *The economic costs and benefits of airbnb: No reason for local policymakers to let airbnb bypass tax or regulatory obligations.* https://www.epi.org/publication/the-economic-costs-and-benefits-of-airbnb-no-reason-for-local-policymakers-to-let-airbnb-bypass-tax-or-regulatory-obligations/. (Accessed: 2023-3-26)

Dudás, G., Kovalcsik, T., Vida, G., Boros, L., & Nagy, G. (2020, May). Price determinants of airbnb listing prices in lake balaton touristic region, hungary. *Eur. J. Tour. Res.*, *24*, 2410.

Fund, I. M. (2023). *Imf data mapper: Gross domestic product per capita, current prices (u.s. dollars) - european union, united kingdom.* https://www.imf.org/external/datamapper/NGDPDPC@WEO/UVK/EURO/EU.

Graw, J. H., Wood, W. T., & Phrampus, B. J. (2021, January). Predicting global marine sediment density using the random forest regressor machine learning algorithm. *J. Geophys. Res. Solid Earth*, *126*(1).

Hall, C. M., Prayag, G., Safonov, A., Coles, T., Gössling, S., & Koupaei, S. N. (2022). Airbnb and the sharing economy. *Current Issues in Tourism*, *25*(19), 3057-3067. Retrieved from https://doi.org/10.1080/13683500.2022.2122418 DOI: 10.1080/13683500.2022.2122418

Hati, S. R. H., Balqiah, T. E., Hananto, A., & Yuliati, E. (2021, October). A decade of systematic literature review on airbnb: the sharing economy from a multiple stakeholder perspective. *Heliyon*, *7*(10), e08222.

Michael. (2015, December). *Airbnb and the sharing economy: Creating value for everyone.* https://d3.harvard.edu/platform-rctom/submission/airbnb-and-the-sharing-economy-creating-value-for-everyone/. (Accessed: 2023-3-26)

Numbeo. (2023). https://www.numbeo.com/cost-of-living/region_prices_by_city?itemId=100&region=150. (Accessed: March 25, 2023)

Rahul, Gupta, A., Bansal, A., & Roy, K. (2021, May). Solar energy prediction using decision tree regressor. In *2021 5th international conference on intelligent computing and control systems (ICICCS).* IEEE.

Reinhold, S., & Dolnicar, S. (2021). *The evolution of airbnb's business model.* figshare.

The Devastator. (2023, February). *Airbnb prices in european cities.*

Toader, V., Negruşa, A. L., Bode, O. R., & Rus, R. V. (2022, December). Analysis of price determinants in the case of airbnb listings. *Econ. Res.-Ekon. Istraž.*, *35*(1), 2493–2509.

Zhang, G., Gong, Y., Wang, Q., Deng, L., Zhang, S., Liao, Q., ... Liu, Z. (2016, July). Outcomes and factors associated with survival of patients with HIV/AIDS initiating antiretroviral treatment in liangshan prefecture, southwest of china. *Medicine (Baltimore)*, *95*(27), e3969.