Predicting Airbnb Rental Prices Across Regions Using Machine Learning

Team 3 (Green)

Caleb O'Neel, Nikhil Bhargava, Junbo Guan, Rhayoung Park, Xiaohan Yang

I. Abstract

Our team created a random forest regression model to predict Airbnb pricing using listing features and geographic information. The data used to train our model comes from six major cities within the United States, but through a variety of normalization techniques and controlling for cost of living, our model is able to successfully generalize to cities in western Europe far better than baseline predictions. This diverges from models previously created in the space which are constrained to their city or geographic region their training data comes from. Our model's ability to translate universally makes it more accessible.

II. Introduction

Airbnbs are a fantastic way to find a place to stay when visiting a city that offers a more unique experience than staying in a hotel. They are not without their risks, however. Since the owner of the Airbnb controls the pricing, pictures and information you see about a rental property- they can create a somewhat false image of how nice their rental is. This leaves customers at risk of overpaying based upon limited information. On the flip side, it can be difficult for owners to assess the market by looking at other Airbnb's in their location and creating accurate comparisons to give an idea of how much they should list their property for.

Our model attempts to dispel some of the ambiguity around appropriate Airbnb pricing by evaluating the features, location, and amenities of an Airbnb listing and determining what the average price you would expect to pay for such a property. With this information Airbnb customers can assess if they are paying a reasonable price, and Airbnb hosts can price their listings at a reasonable rate that will help them maximize their revenue and occupancy rate. Additionally, unlike other research that has been previously done around Airbnb price prediction, the aim of our study is to generalize a model beyond just the city or geographic location it was trained on. More specifically, we aim to predict Airbnb property prices per night in western Europe using a model exclusively trained on data from the United States.

III. Background

The problem of evaluating the pricing of a rental property has existed far longer than the introduction of Airbnb. Rental real estate is a massive industry with billions of dollars exchanging hands on an annual basis (Collins et al., 2021).

Our paper leveraged the ideas of several previous research projects done on similar topics. One of the concepts we took was the idea that housing data requires a significant amount of normalization when comparing across markets (Yu et al., 2016). This helps control for some of the external factors affecting prices that are unique to each city. Another idea that ultimately paid dividends for us was the concept of factoring in proximity to the center city (Ma et al., 2018).

Compared to general real estate posts, Airbnb posts include things such as ratings, user reviews, and host longevity. Airbnb rentals are also much shorter than the annual rentals most research in this space focuses on (Bivens et al., 2019). This makes working with the data slightly different as there is a larger social aspect to these decisions. For example, researchers at Stanford used sentiment analysis on reviews to predict Airbnb prices in New York City (Kalehbasti et al., 2019). This is also not the only unique aspect of working with Airbnb data. In the research done by Wang et al., traditionally predictive factors such as ratings and brand affiliation that are good indicators of popularity in the hotels do not seem effective in predicting Airbnb data (2017). Instead, they found host attributes, features and amenities, and property attributes hold the predictive signal key to unlocking an Airbnb's pricing.

Ultimately, we incorporated many ideas from the research papers, into our model design process. However, our research did differ from the aforementioned works in several ways. Most papers focused on specific cities or geographic areas to train and test their models. One focus of our research was to see how our predictions could translate to Western European cities after training exclusively on major cities within the United States, as there are differences in Airbnb etiquette between the two continents, (McMahan et al., 2017). One key difference we found in our study is that the distance to the city center was much more important in American cities than European. We hypothesize that this could be due to the fact that European cities have significantly better public transportation infrastructure giving tourists easier access to downtown areas (Cascajo et al., 2014). The vast majority of previous research on the topic is specific to one city or region, but we believe our model is better architected to generalize across regions.

IV. Data

A. Data Sources

For this study, four datasets were acquired, cleaned, and merged together to create comprehensive train and test datasets. The source, name, and description, of each data source can be found in Table 1 below.

Table 1: Data source, name, and description of each dataset used in the paper.

ID	Data Source	Name	Description
1	Kaggle	Airbnb Data in Major US Cities	This dataset contained the price and characteristics of various Airbnb listings in six major US cities -- Boston, Chicago, Los Angeles, New York, San Francisco, and Washington D.C. The initial dataset contained information on over 74,000 distinct Airbnb properties and 28 features from 2018. It was ultimately split into training, validation, and test sets.
2	Inside Airbnb	Airbnb Data in Global Cities	Inside Airbnb contained the price and characteristics of listings in major cities worldwide in 2021. To test for model generalizability abroad, we selected three prominent cities in Western Europe -- Madrid, London, and Paris. This dataset was concatenated together to form an International dataset. Each dataset was used for model testing.
3	Simple Maps	City Data	The simple maps data was joined on the Airbnb data to obtain population and city center coordinate data. This data was used in the feature engineering process.
4	World Data	Cost of Living Index	The Cost of Living Index (CLI) dataset contains metrics used to adjust the cost of products globally with respect to the US. This data was used to adjust our international predictions.

A sample set of features can be seen in Table 1 below and a comprehensive data dictionary can be found in Appendix I.

Table 2: Sample features used in the final training and test sets.

ID	Property Type	City - NYC	Room Type	Bath rooms	Number of reviews	# of Airbnbs within 1km	...	Distance to city center	Name Length	Bed to Bathroom ratio	Amenities – Beach essentials	Amenities – Pets Allowed
1	Apart -ment	1	Entire Home	2	208	1189	...	4.5	7	1.4	1	0
2	House	0	Private Room	1	87	754	...	9.6	5	1.9	0	1
3	Boat	0	Entire Home	0	46	38	...	20.8	8	1.2	0	0

B. Exploratory Data Analysis (EDA)

Our training data contained Airbnb properties from six major cities in the United States. A heatmap of the number of listings by city can be seen in Figure 1. Larger circles correspond to more properties contained in the dataset. A large majority of the dataset was from New York and Los Angeles, with a few thousand listings from each of the other cities. The proportion and number of properties by city in the dataset can be found in Table 2. These numbers are significant because average and median rental prices varied across these cities, which impacts how the model generalizes across cities and to our European test set. Figure 2 shows that San Francisco has the highest average list price. Somewhat surprisingly, New York had the second-lowest average listing price in the dataset.

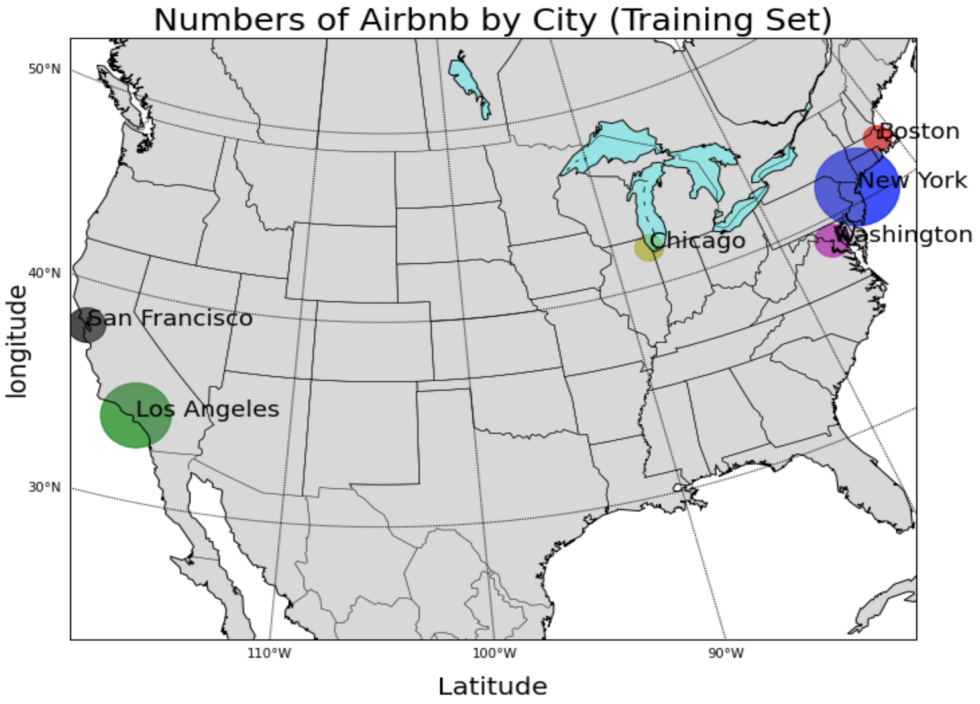


Figure 1: Heatmap of Airbnb listings across US cities.

Table 3: Number and proportion of Airbnb listings in each city for the US dataset.

City	Number of Airbnbs	Proportion (%)
New York City	32,349	43.65
Los Angeles	22,453	30.3
San Francisco	6,434	8.68
Washington D.C.	5,688	7.67
Chicago	3,719	5.02
Boston	3,468	4.68
Total	**74,111**	**100**

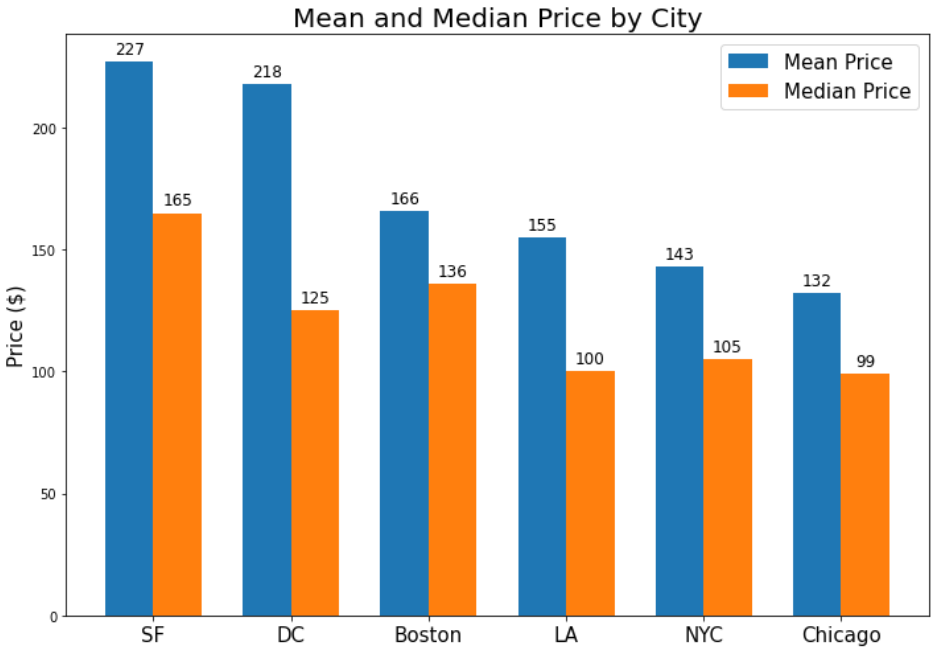


Figure 2: Mean and median listing prices across American cities.

Additionally, we visualized the distribution of listing prices based on property type. As seen in Figure 3, we found those entire apartments had significantly higher prices than shared apartments and rooms.

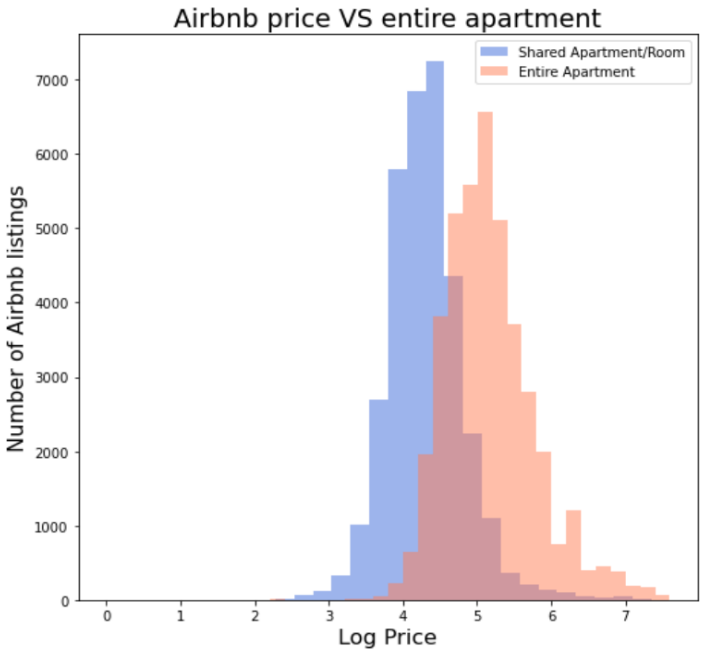


Figure 3 : Distribution of US Airbnb prices across property types.

Since a key component of our study was generalizing our model to other countries, it was important to understand how they differed from the US. Figure 4 depicts the distributions of prices of Airbnb Listings across different regions. Despite European cities having a few Airbnbs that were highly-priced, the American ones had a higher median overall price.

Figure 4: Distribution of Airbnb listing prices across regions.

C. Feature Engineering

Based on studies done by Zhang et al. (2017) and Ma et al. (2018) on geographical features influencing Airbnb prices, we took a similar approach to feature engineering. In particular, Zhang et al. determined how distances from the nearest highway and the Nashville Convention Center impacted prices of Nashville Airbnbs (2017). These studies inspired us to derive a similar but scalable approach to this by calculating the distance of Airbnbs to their respective city's city center. The potential impact this feature could provide can be seen below in Figure 5. The closer an Airbnb is to D.C.'s downtown, the higher the average price of an Airbnb is.

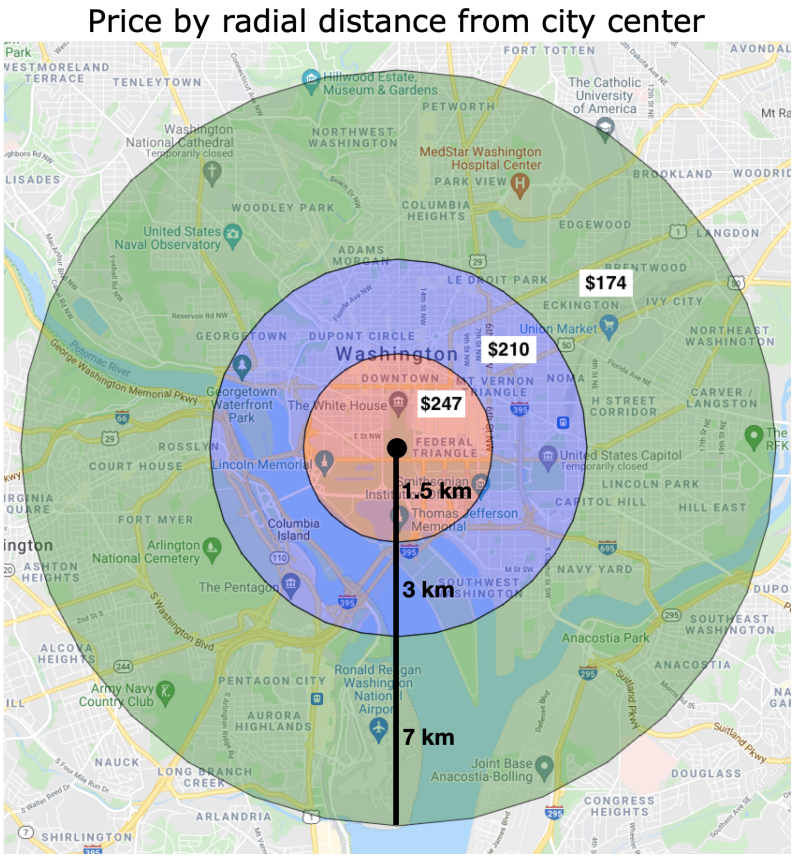


Figure 5: Average Airbnb Listing Prices in Relation to Washington D.C.'s City Center.

Based on an IEEE article that mentioned how Airbnb "constructs" neighborhoods through previous rental prices, we used another geographic approach to pick up latent density signals (2015). By calculating the number of Airbnbs within a one kilometer radius of each listing, we believe we could identify potential hotspots within a city. Figure 6 shows a quick example of how this feature works on five randomly sampled Airbnbs within New York City. Many Airbnbs within the city have a large number of surrounding Airbnbs, despite being far from the city center (yellow star). For modeling purposes, this feature was normalized by the total number of listings within a city to account for larger cities.

Figure 6: Number of Airbnbs Within a 1km Radius of 5 Randomly Sampled Airbnbs in NYC.

Another feature of interest, amenities, was extracted from free text input directly by hosts. Using RegEx, we identified, cleaned, and standardized over 130 unique amenities. We then manually grouped similar amenities, and accounted for rare amenities by either removing them, or adding them to a larger group to prevent the model from overfitting on amenities potentially only found in the United States.

V. Methods

A. Process

As shown in Figure 7, we took an eight-step approach to create a generalized solution to predicting international Airbnb prices. The first five steps have been covered in the prior data section.

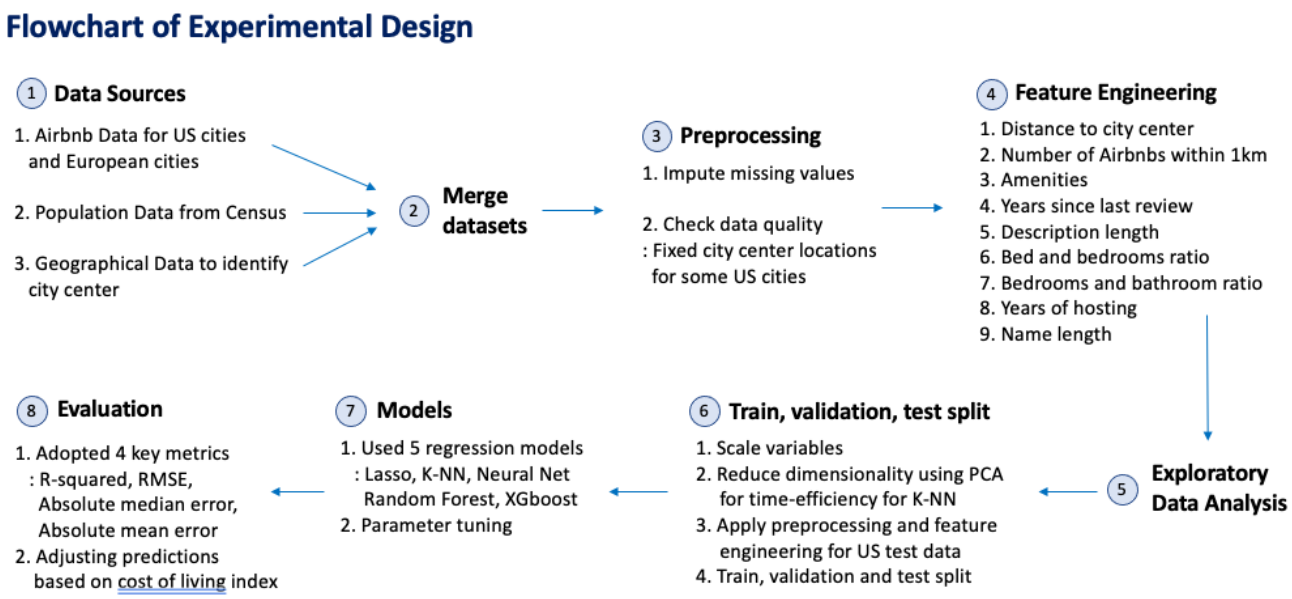


Figure 7: Flowchart of Experimental Design.

Once the training data was prepared for modeling, we set aside 20% to create a validation set to finetune our models' hyperparameters. For this study, five different test sets were used -- US, International, Madrid, London, Paris. The US test set contained about 14,000 observations with the same six cities found in the training set. The European test sets consisted of around 70,000 observations from London and Paris, respectively, and about 25,000 observations from Madrid. The International test set was created by concatenating the observations from all three European cities. Each test set was updated by imputing null values with the mean of the training set for their respective columns, and the same features were engineered to make the test set mimic the train set.

B. Models

Five models were trained to determine the effectiveness of different modeling approaches on predicting American Airbnb listing prices and then generalizing those predictions internationally. The optimal model selected for this problem was based on metrics defined in the Evaluation section below.

i. Linear (Lasso) Regression
Linear regression was selected as a potentially effective model to generalize predictions to European listings. By using Lasso Regression, we intended to reduce the amount of multicollinearity and noise within the features of our training set to help improve the model's overall performance on the test sets (Tersakyan, 2019).

ii. K-Nearest Neighbors Regressor

Thuraiya et al. showed that KNN is a good baseline model for predicting housing prices (2020). Considering the research by Yu et al., where algorithms with PCA performed better than the ones without, and also the issues with training time, we performed Principal Component Analysis (PCA) on the dataset before fitting the model (2016). Based on the cumulative variance explained and validation accuracy, the optimal number of components was 15 and the optimal K-value chosen was 7.

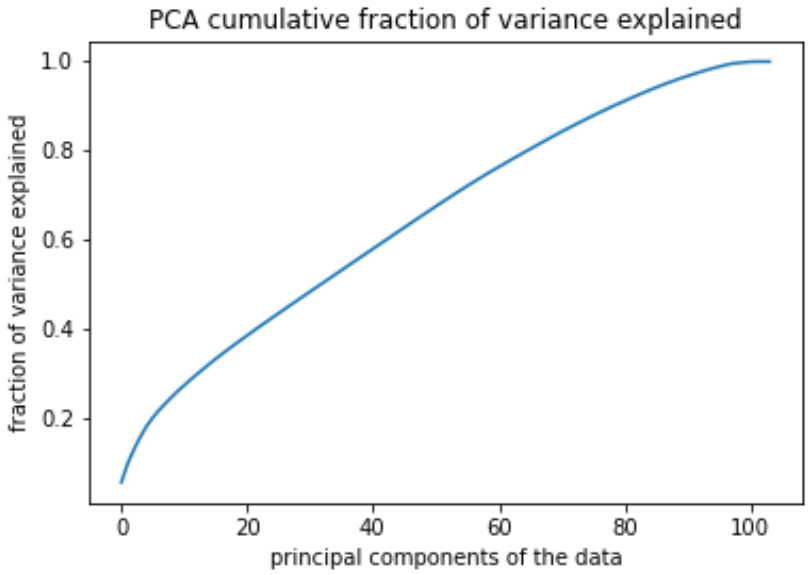


Figure 8: Cumulative Fraction of Variance Explained by PCA by N Components.

iii. Random Forest

Based on the results of the KNN regressor, we selected another non-parametric model to pick up on the flexible relations found within the training dataset. The optimal model was selected through a Random Search over a few hyperparameters such as the number of trees, the maximum depth of the tree, and a few others. Additionally, the random forest was bootstrapped to reduce overfitting while training.

iv. XGboost Regressor

From the study conducted by Kokasih et. al., their implementation of XGboost was found to perform extremely well at predicting rental prices (2020). Additionally, the success of another tree-based in Random Forests, encouraged us to try more sophisticated tree-based models like XGboost.

v. Multi-Layer Perceptron (MLP) Regressor

An MLP regressor was chosen to see how well a black box model could perform compared to our other models. Based on a previous study by Kalehbasti et al. (2019), neural networks were able to fit similar training datasets well with a high R^2. Using Random Search, the optimal network architecture contained 2 hidden layers with 200 neurons in each layer, with relu activation functions.

C. Adjusting Predictions

While performing EDA, we discovered that the distribution of Airbnbs prices were different across countries and especially when compared to the US. According to Handbury, one plausible reason for this discrepancy could be due to cost of living differences between regions (2019). Thus, we scaled our predictions by the CLI for each country relative to the US. For the international data set predictions, we took the weighted average between the CLI and the number of Airbnbs in each country to scale our predictions.

The most significant example of how this impacted our predictions can be found in predictions for Madrid. The CLI of Spain is 75.8, meaning products in Spain are 24.2% cheaper than they are in the US. Therefore, we scaled our price predictions by that same factor. The magnitude of this adjustment was visualized in Figure 9.

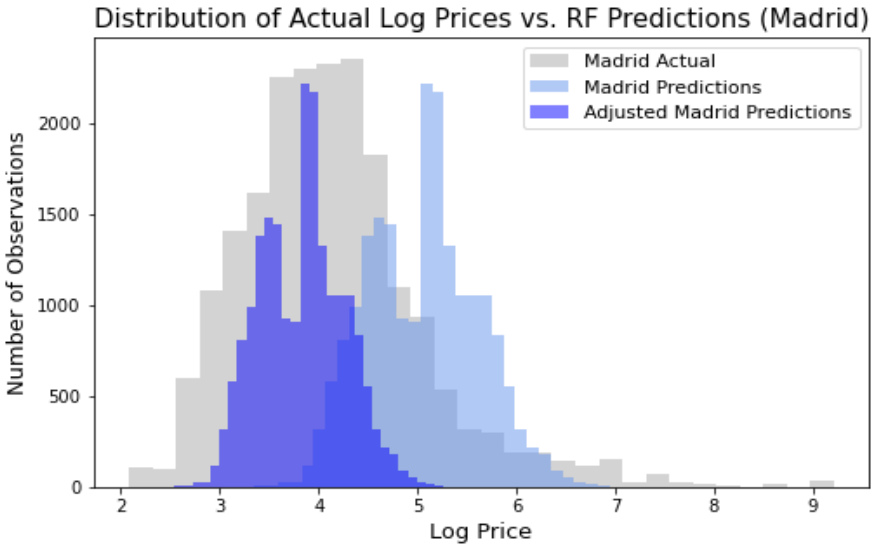


Figure 9: Distribution of Random Forest predictions for Madrid before and after being adjusted by CLI.

D. Evaluation Metrics

The most important metrics used to compare model accuracies were the R^2 and RMSE, particularly on the International and American test sets. Additionally, the mean and median absolute error were secondary metrics calculated to better interpret our results. These values correspond with the mean and median number of dollars our predictions were from actuality.

A baseline for each test set was derived by calculating R^2, RMSE, Mean and Median Absolute Error using the mean price of the training set to predict the price of every Airbnb in the respective test set.

VI. Results

The R^2, RMSE, Mean Absolute Error, and Median Absolute Error for each model and the baseline metrics can be found in Table 1 below. While every model outperformed the baseline R^2 scores, the same was not true for the RMSE. Both Linear (Lasso) Regression and the MLPRegressor struggled to translate American listing prices to their international counterparts and performed significantly worse than the baseline. Based on the distribution of predictions in Figure 10, this is likely caused by both models making larger price predictions, while others were more conservative. These errors likely skewed the RMSE values to be much larger than the baseline and other models.

Table 4: Accuracy Metrics for Each Model.

Model	Region (Test)	R^2	RMSE	M[ean]AE	M[edian]AE
Baseline* (Mean Guess)	US*	0	168.2	97.35	75.57
	International*	-0.32	305.43	106.78	91.57
	Madrid	-0.49	457.83	146.1	110.57
	London	-0.34	317.32	106.88	95.57
	Paris	-0.27	219.67	94.41	85.57
Linear (Lasso) Regression	US*	0.58	132.79	59.5	28.61
	International*	-0.23	594236.89	2531.26	58.12
	Madrid	0.01	7780.96	889.05	73.24
	London	-0.14	176437.65	739.63	55.66
	Paris	-0.63	918898.77	5164.26	59.8
KNN Regressor	US*	0.56	130.14	60.99	29.16
	International*	0.24	300.74	61	24.11
	Madrid	-0.03	463.11	91.56	19.08
	London	0.34	309.48	59.01	27.48
	Paris	0.07	214.02	54.32	24.08
Random Forest Regressor	US*	0.68	117.08	51.68	23.55
	International*	0.34	259.93	59.25	28.29
	Madrid	0.16	460.53	85.45	16.04
	London	0.46	304.68	55.43	27.33
	Paris	0.18	208.31	54.03	32.1
XGboost Regressor	US*	0.62	129.64	56.48	24.41
	International*	0.29	299.93	57.83	20.25
	Madrid	-0.23	464.03	93.9	22.24
	London	0.5	307.03	51.34	19.58
	Paris	0.01	213.59	53	22.08
MLP Regressor	US*	0.54	132.08	62.34	31.7
	International*	0.02	27791.23	230.78	46.63
	Madrid	-0.16	458.76	116.53	53.8
	London	0.18	7581.9	101.26	42.45
	Paris	-0.41	24427.67	298.77	61.59

*US and International test datasets are an aggregation of multiple cities.
*Baseline is not a model, just the mean guess.

Surprisingly, the KNN and XGboost regressors had lower median absolute errors in the International test set than the US. One possible explanation is there are only half the number of cities in the international test set than there are in the American one, and therefore still more price fluctuation.

Ultimately, the Random Forest performed the best on both the American and International test sets after having the largest R^2 and smallest RMSE. XGboost and KNN, both did well but performed slightly worse, particularly on the Madrid and Paris test sets. The Random Forest Regressor was able to predict the price of an Airbnb in the US within $52 compared to $59 internationally. Additionally, the median prediction was within about $24 in the US compared to $28 internationally. Compared to the results produced by Kokasih et. al., our models performed relatively well compared to the mean absolute error (2020). Although they had slightly lower errors, we suspect that this is due to their study being focused on Singapore.

The distribution of each model's predictions, along with the associated residual distributions can be found for the US in Figures 10 and for the European cities in Figure 11 below. Confirming what was found in Table 1 above, Figure 10 depicts how well the predicted price distributions overlap the actual Airbnb listings prices in the US. Each model, on both test sets, also had a bimodal residual distribution centered around zero. The closer these bimodal peaks are to zero, the better the model, such as the distribution for the Random Forest.

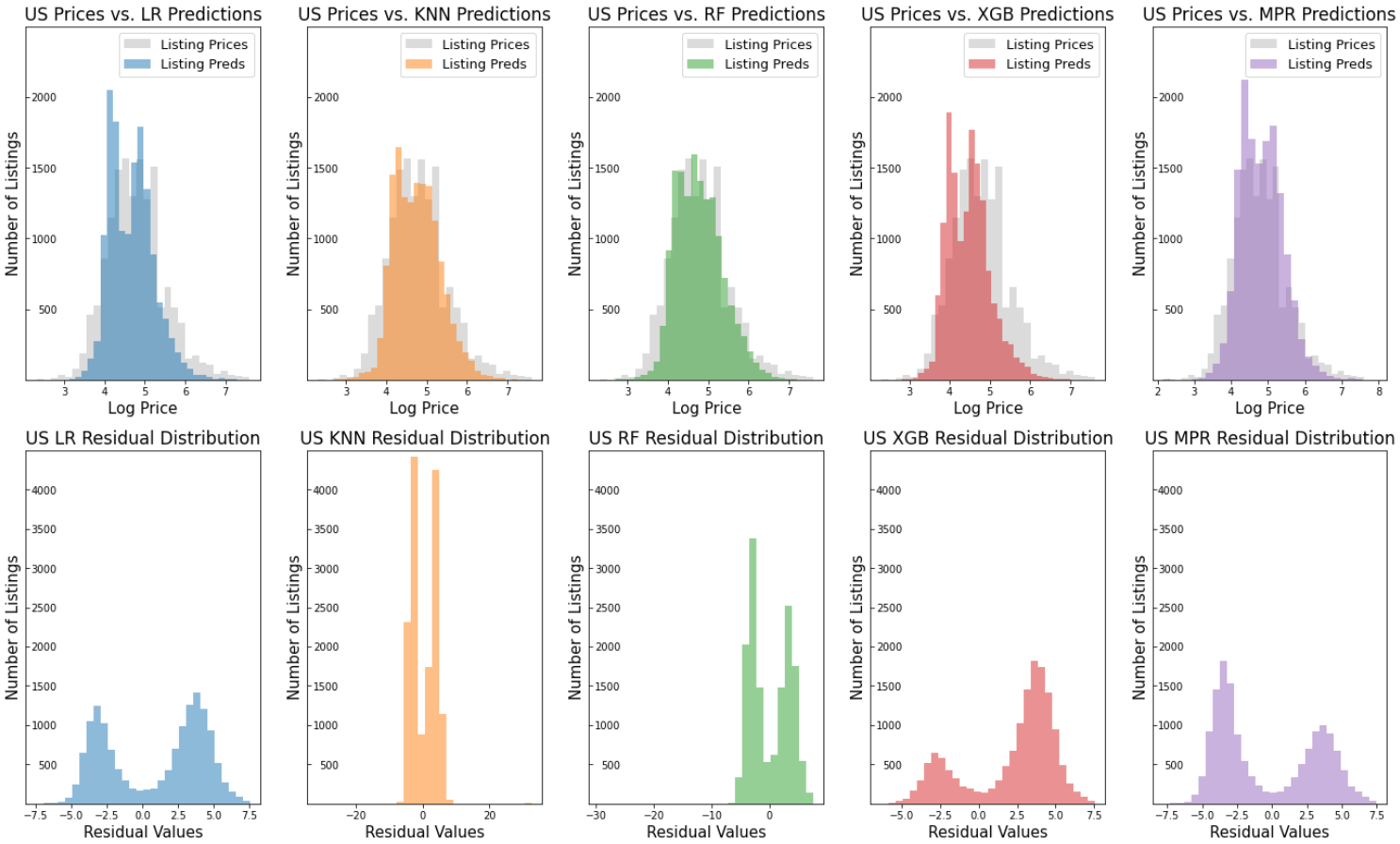


Figure 10: Distribution of price predictions versus actual (top) and distribution of residuals (bottom) for each Model on the US test set.

As expected, the prediction distributions of the KNN, Random Forest, and XGboost models were very similar on the international dataset. One slight difference that may have given the Random Forest a performance boost was its ability to predict a wider range of prices close to the median, as evident by its distribution curve in Figure 11.

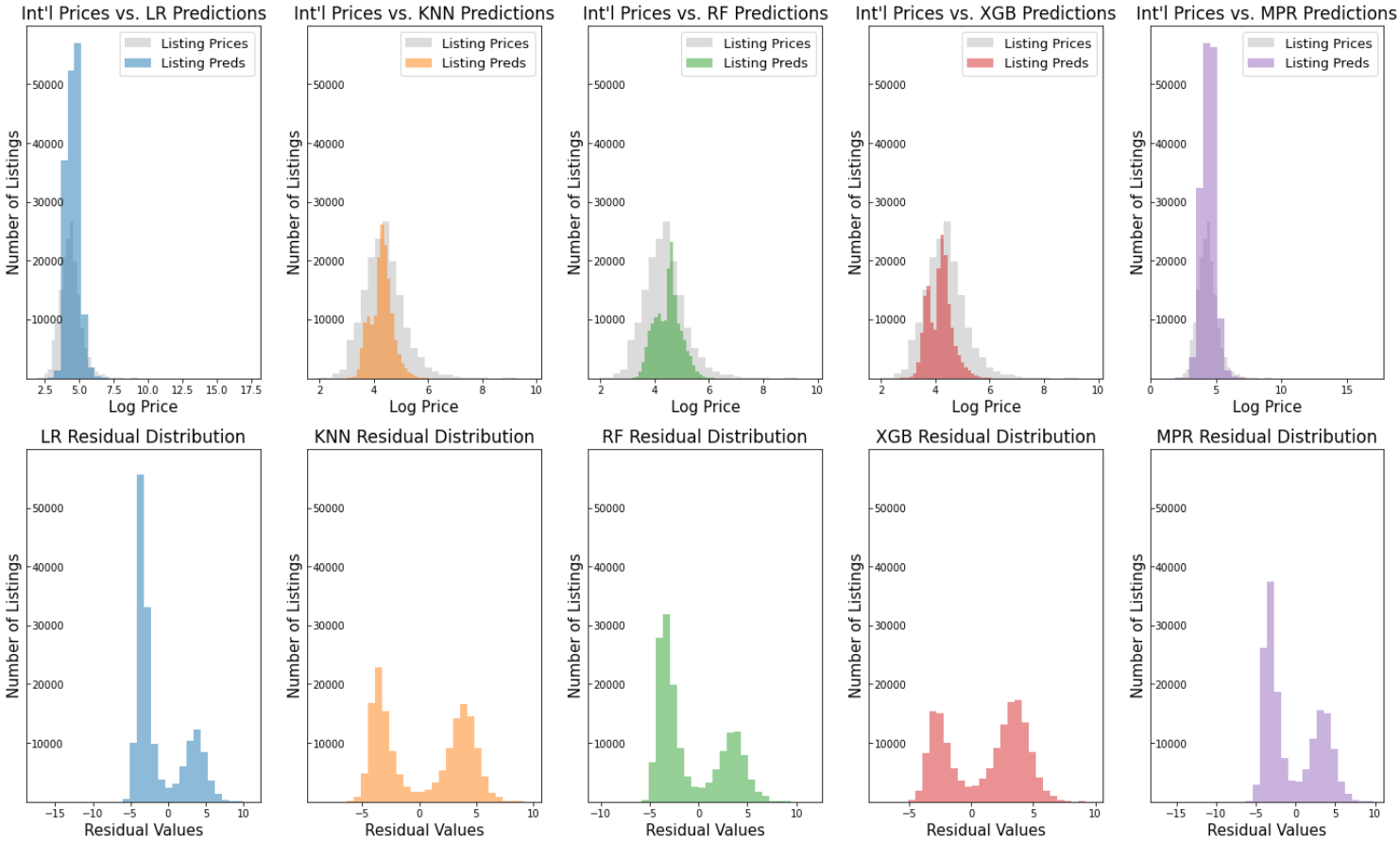


Figure 11: Distribution of price predictions versus actual (top) and distribution of residuals (Bottom) for each model on the International test set.

To get a better sense of what was important for predicting prices for the Random Forest model on the training set, the normalized Gini-Importance was calculated for each feature. The larger the value, the higher the importance. As can be seen in Figure 12, the single most important feature in predicting prices was whether or not an Airbnb was listed as an entire home or apartment. The geographic features engineered were two of the five most important features in predicting price, confirming the validity of the concept.

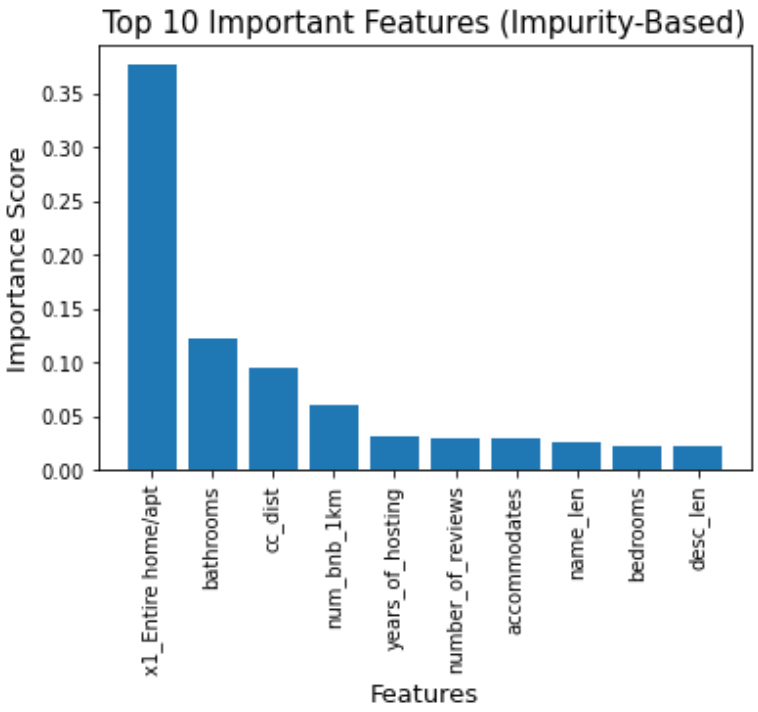


Figure 12: Top 10 important features for predicting listing prices based on Gini Importance (Random Forest). The feature, 'entire home/apt' refers to the property type, 'bathroom' is number of bathrooms, 'cc_dist' is the distance to the city center, 'num_bnb_1km' is the number of Airbnbs within 1km radius, 'years_of_hosting' indicates the years that the host has been hosting, 'name_len' refers to the length of Airbnb name, and 'desc_len' indicates length of the Airbnb description.

VII. Conclusion

This study developed a modeling and feature engineering approach that not only performs well at predicting Airbnb rental prices within the United States but also globally. Based on property characteristics and geographic features, we were able to train multiple models that generalized beyond the geographic location it was trained upon. Based on the predictive power, we concluded that property type, the number of bathrooms, and the distance an Airbnb is to the city center is critical in predicting the price of a listing.

The results from the Random Forest Regressor showed that the model was able to outperform other techniques such as Lasso Regression, KNNs, XGboost, and MLP Regressors based on key metrics such as R^2 and RMSE. The median prediction of this model was within $24 of the actual Airbnb price in the US compared to $28 internationally. Considering the nightly cost of anywhere from a couple of bucks to thousands of dollars, this range is acceptable.

One limitation our study did not account for is the impact Covid may have had on rental prices internationally. Although much of the international test data was acquired in 2021, our training data was acquired in 2018. According to Forbes, 29% of American hosts listed their properties at reduced prices, especially for those considered essential workers. The pandemic could have affected our study overall.

Some future work can be done to improve upon our results. First, this project was focused primarily on large, metropolitan areas. Obtaining training samples from smaller cities and testing the performance would be critical to implementing a generalized model on a global scale. Second, we believe that city demographics and tourism data have a huge impact on Airbnb demand and prices. Additional datasets related to these concepts may allow for greater prediction accuracy.

References

1. Bivens, Josh. “The Economic Costs and Benefits of Airbnb: No Reason for Local Policymakers to Let Airbnb Bypass Tax or Regulatory Obligations.” Economic Policy Institute, 30 Jan. 2019, www.epi.org/publication/the-economic-costs-and-benefits-of-Airbnb-no-reason-for-local-policymakers-to-let-Airbnb-bypass-tax-or-regulatory-obligations.
2. Cascajo, Rocio & Alonso, Andrea & Monzón, Andrés. (2015). Comparative analysis of passenger transport sustainability in European cities. Ecological Indicators. 48. 578-592. 10.1016/j.ecolind.2014.09.022.
3. Collins, Gord. “The US Rental Property Market Outlook.” ManageCasa, 31 Mar. 2021, managecasa.com/articles/us-rental-property-market.
4. D. Wang and J. L. Nicolau, “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com,” International Journal of Hospitality Management, vol. 62, pp. 120–131, 2017.
5. Dogru, Tarik, and Osman Pekin. What Do Guests Value Most in Airbnb Accommodations? An Application of the Hedonic Pricing Approach, 7 June 2017, www.bu.edu/bhr/2017/06/07/airbnb-guest-pricing-value/.
6. H. Yu and J. Wu, “Real estate price prediction with regression and classification,” CS229 (Machine Learning), 2016.
7. Handbury, J. (2019). Are poor Cities cheap for Everyone? Non-homotheticity and the cost of living Across U.S. Cities. *SSRN Electronic Journal*. doi:10.2139/ssrn.3497831
8. Hill, Dan. The Secret of Airbnb’s Pricing Algorithm. IEE Spectrum, 2 Aug. 2015, spectrum.ieee.org/computing/software/the-secret-of-Airbnbs-pricing-algorithm.
9. Kalehbasti, Pouya Rezazadeh, Liubov Nikolenko, and Hoormazd Rezaei. "Airbnb price prediction using machine learning and sentiment analysis." arXiv preprint arXiv:1907.12665 (2019).
10. Kokasih, Marco Febriadi, and Adi Suryaputra Paramita. "Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm." IJIIS: International Journal of Informatics and Information Systems 3.2 (2020): 54-59.
11. Kuvalekar, Alisha and Manchewar, Shivani and Mahadik, Sidhika and Jawale, Shila, House Price Forecasting Using Machine Learning (April 8, 2020). Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020, Available at SSRN: https://ssrn.com/abstract=3565512
12. Lane, Lea. “How Bad Are Covid-19 Pandemic Effects On Airbnb Guests, Hosts?” Forbes, Forbes Magazine,09 June, 2020, www.forbes.com/sites/lealane/2020/06/09/how-bad-are-covid-19-pandemic-effects-on-airbnb-guests-hosts/?sh=154e07b97432
13. McMahan, Dana. “Here’s What You Need to Know About European Airbnbs.” Kitchn, 9 Oct. 2017, www.thekitchn.com/10-things-to-know-about-european-Airbnbs-249707.
14. Mohd, T., Jamil, N. S., Johari, N., Abdullah, L., & Masrom, S. (2020). An overview of real estate modelling techniques for house price prediction. *Charting a Sustainable Future of ASEAN in Business and Social Sciences, 1*, 321-338. doi:10.1007/978-981-15-3859-9_28
15. Perez-Sanchez, V. Raul, et al. "The what, where, and why of Airbnb price determinants." Sustainability 10.12 (2018): 4596.
16. Tersakyan, Alen. “Airbnb Price Prediction Using Linear Regression (Scikit-Learn and StatsModels).”Medium,4Dec.2019,towardsdatascience.com/Airbnb-price-prediction-using-linear-regression-scikit-learn-and-statsmodels-6e1fc2bd51a6.
17. Tang, Emily, and Kunal Sangani. "Neighborhood and price prediction for San Francisco Airbnb listings." Departments of Computer science, Psychology, economics–Stanford University (2015).

18. Y. Ma, Z. Zhang, A. Ihler, and B. Pan, “Estimating warehouse rental price using machine learning techniques.,” International Journal of Computers, Communications & Control, vol. 13, no. 2, 2018.
19. Zhang, Zhihua, et al. "Key factors affecting the price of Airbnb listings: A geographically weighted approach." Sustainability 9.9 (2017): 1635.
20. Zhu, Ang, Rong Li, and Zehao Xie. "Machine Learning Prediction of New York Airbnb Prices." 2020 Third International Conference on Artificial Intelligence for Industries (AI4I). IEEE, 2020.

Roles:

1. Caleb O'Neel: 1-E [Results], 2-D [Video lead]

   Responsible for compiling, writing, editing, and presenting the final presentation. Worked on text feature engineering, trained and tuned the linear regression (LASSO) model. In charge of collecting, compiling, and presenting performance metrics. Help to bring them together into a coherent discussion of the results, combine plots when appropriate for comparison. Provide a coherent assessment of the results of all models and visualize them.

2. Junbo Guan: 1-D [Methods], 2-A [Project coordinator]

   Responsible for articulating the approach the team takes to solve the problem or answer the question at the core of the project. Dealt with missing values imputation, duplicates detection and one hot encoding. Worked on PCA and the KNN model. Create a timeline, coordinate and convene weekly meetings on the project as well as facilitating an agenda for each meeting.

3. Nikhil Bhargava: 1-C [Data: Preprocessing and Exploration], 2-E [Github lead]

   Led exploratory data analysis of the dataset, merged datasets, data cleaning, preparation, and feature extraction & engineering. Made sure training dataset was prepared for modeling, as well as cleaned and formatted all international datasets. In charge of random forest modeling and visualizing model metrics and predictions for the models (and adjusting predictions). Also responsible for ensuring that the final project content is well organized, documented, and commented in Github and could be easily used by other data scientists.

4. Rhayoung Park: 1-A [Abstract, Introduction, and Conclusions], 2-B [Proposal and report lead]

   Lead the work on and contribute the most to the proposal and progress reports. Interpret the results and place them in the context of other work or potential impact on application domains. Responsible for the final review, quality assurance, and submission of the reports. Merge external datasets and word on data processing. Train and tune the XGBoost model.

5. Xiaohan Yang: 1-B [Background and References], 2-C [Report integrator lead]

   Dive deeply into related references and interesting projects that have been conducted. Review the content and make sure the pieces flow reasonably well and work with the authors to improve the content and help with rewriting. Deal with text feature processing, create a matching dictionary for amenities. Train and tune Neural Network Regressor.

Timeline of Activity:

Activity	Timeline
Brainstorming & Data Collection	
Meeting 1, Discuss the project topic; divide roles	Jan 27
Meeting 2, Decide project theme; find related datasets	Feb 2
Find and read related references	Feb 7
Exploratory Data Analysis	
EDA on the raw dataset, Finish final proposal	Feb 10
Project Proposal Due	Feb 10
Meeting 3, Discuss final proposal feedback; find European test sets	Feb 15
Data Pre-processing (deal with missing values, duplicates); more EDA and visualization	Feb 21
Meeting 4, Finish data processing on the US training data; complete progress report	Feb 23
Feature Engineering	
Compare the features between US training data and European testing data and discard some features only appear in one dataset	Feb 26
Meeting 5, Discuss feature engineering	Mar 2
Read references; merge external data; feature transformation(for both US and European data)	Mar 4
Meeting 6, Discuss text feature engineering	Mar 5
Finish text feature engineering: One hot encoding, Split train, validation, test data, upload data processing notebook to git repo	Mar 12
Modeling	
Meeting 7, Split modeling tasks	Mar 17
Progress Report Due	Mar 21
Train, validate, and tune the models	Mar 23
Meeting 8, Discuss the model performance; unify the evaluation metrics for all models	Mar 25
Upload evaluation metric for all 5 models and compare the result	Mar 28
Meeting 9, Discuss plausible reasons for this discrepancy in performance on US test data and European test data	Apr 1
Read more references, Find out several reasons and adjust the prediction	Apr 7

Result Discussion & Finalizing	
Meeting 10, Compare results to determine the final model	Apr 12
Upload the models to git repo and finalize git repo	Apr 15
Meeting 11, Prepare for final project presentation and final report	Apr 17
Finish the final project presentation, Upload the video to youtube, and win the judge's prize!!!	Apr 20
Final Presentation	Apr 21
Meeting 12, Finalize the final report	Apr 26
Final Report Due	Apr 26

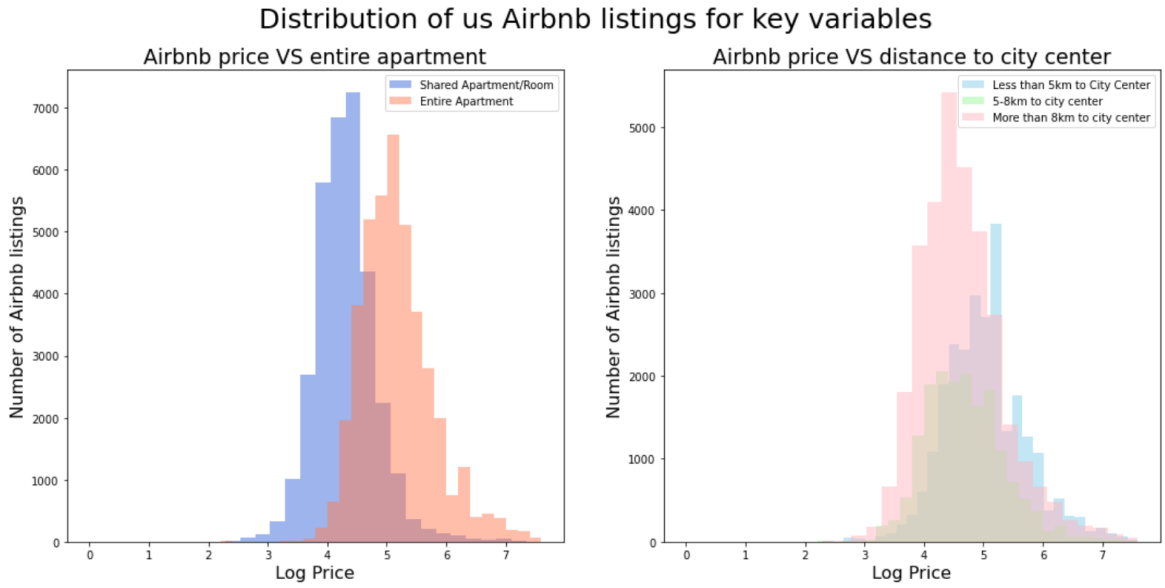
Appendix I. Data Dictionary

Features	Description	Dataset
id	Unique ID of Airbnb listings	Airbnb
price	Price of the listing (USD)	Airbnb
property_type	Apartment, House, Boat, Yurt etc	Airbnb
room_type	Entire Home, Private Room	Airbnb
amenities		Airbnb
accomodates	Number of people accommodated in this property	Airbnb
bathrooms	Number of bathrooms	Airbnb
city	Boston, LA, San Fransisco, Chicago, NYC, DC	Airbnb
description	Description of the property	Airbnb
first_review	The date when the first review was uploaded	Airbnb
host_has_profile_pic	Whether the host has a profile picture or not	Airbnb
host_identity_verified	Whether the host identity is verified or not	Airbnb
host_response_rate	Whether the host responses to the messages (%)	Airbnb
host_since	The date since the host has been hosting	Airbnb
instant_bookable	Whether the property is instantly bookable	Airbnb
last_review	The date that the most recent review was uploaded	Airbnb
latitude	The latitude of the property	Airbnb
longitude	The longitude of the property	Airbnb
cancellation_policy	How strict the cancellation policy is (strict / moderate / flexible)	Airbnb
name	Name of the property	Airbnb
neighborhood	Where the property is located	Airbnb
number_of_reviews	Number of reviews for the property	Airbnb
review_scores_rating	Ratings for the property	Airbnb
thumbnail_url	The url link for the thumbnail image	Airbnb
cleaning_fee	Whether the customer needs to pay cleaning fee (T/F)	Airbnb
zipcode	The zipcode for each property	Airbnb
bedrooms	Number of bedrooms	Airbnb
bed_type	Type of the bed	Airbnb
beds	Number of beds	Airbnb
city	City name	Simple Maps
lat	Latitude of the city center	Simple Maps
long	Longitude of the city center	Simple Maps
population	City population	Simple Maps
num_airbnb_x_mile	Number of airbnbs in 1km radius	Feature Engineering
dist_city_center	Distance of airbnb from city center	Feature Engineering
bed_bath_ratio	Ratio of bedrooms to bathrooms	Feature Engineering
description_length	Number of letters in description	Feature Engineering
name_length	Number of letters in bnb name	Feature Engineering
bed_bedrooms_ratio	Number of beds to bedroom ratio	Feature Engineering
amenities_{amenity_name}	Multiple boolean columns indicating if amenity exists or not	Feature Engineering
amenities_accessibility	Smooth pathway to front door, Disabled parking spot, Wide doorway, Wheelchair	Feature Engineering
amenities_24_hour_checkin	24 hour check-in	Feature Engineering
amenities_airconditioning	Air conditioning	Feature Engineering
amenities_bbq_grill	BBQ grill	Feature Engineering
amenities_baby_kid_friendly	Baby bath, Baby monitor, Changing table, Crib, High chair, Table corner guards	Feature Engineering
amenities_bathroom	Bath towel, Toilet paper, Bathtub, Hot tub, Hair dryer, Body soap, Shampoo	Feature Engineering
amenities_beach_essentials	Beach essentials	Feature Engineering
amenities_bedlinens	Extra pillows and blankets, Firm mattress	Feature Engineering
amenities_buzzer_wireless	Buzzer, Wireless	Feature Engineering
amenities_tv	Cable TV, TV	Feature Engineering
amenities_pets	Cats, Dogs, Other pets, Pets allowed, Pets live on this property	Feature Engineering
amenities_cleaning_before_checkout	Cleaning before checkout	Feature Engineering

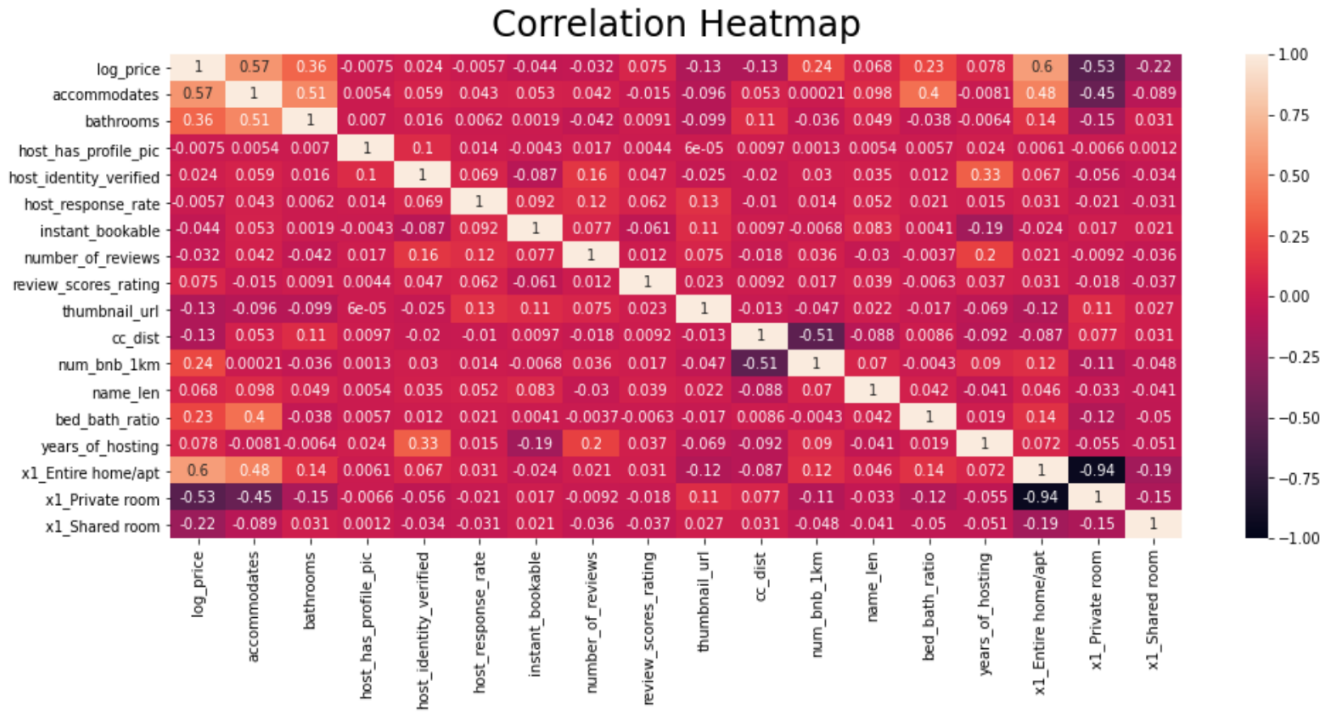
Features	Description	Dataset
amenities_doorman	Doorman	Feature Engineering
amenities_washer/dryer	Washer, Dryer	Feature Engineering
amenities_ev_charger	EV charger	Feature Engineering
amenities_doorman	Doorman	Feature Engineering
amenities_elevator	Elevator, Elevator in building	Feature Engineering
amenities_essentials	Essentials	Feature Engineering
amenities_free_parking	Free parking	Feature Engineering
amenities_game_console	Game console	Feature Engineering
amenities_garden_or_backyard	Garden or backyard	Feature Engineering
amenities_gym	Gym	Feature Engineering
amenities_heating	Heating	Feature Engineering
amenities_hostsgreetsyou	Host greets you	Feature Engineering
amenities_hotwater	Hot water	Feature Engineering
amenities_indoorfireplace	Indoor Fireplace	Feature Engineering
amenities_clothes	Iron, Hangers	Feature Engineering
amenities_kitchen	Refrigerator, Microwave, Dishwasher, Stove, Coffee maker, Oven, Cooking basics	Feature Engineering
amenities_lake	Lake	Feature Engineering
amenities_work	Work	Feature Engineering
amenities_lockonbedroomdoor	Lock on bedroom door	Feature Engineering
amenities_lockbox	Lock box	Feature Engineering
amenities_longtermstaysallowed	Long term stay allowed	Feature Engineering
amenities_waterfront	Water front	Feature Engineering
amenities_selfcheckin	Self check-in	Feature Engineering
amenities_single_level_home	Single-level home	Feature Engineering
amenities_smoking_allowed	Smoking allowed	Feature Engineering
amenities_safety	Smartlock, Smoke detector, Smartlock, Fire extinguisher, First-aid, Window guards	Feature Engineering
amenities_pool	Pool	Feature Engineering
amenities_wirelessinternet	Pocket wifi, Wireless internet, Ethernet, Internet	Feature Engineering
amenities_privacy	Private bathroom, Private entrance, Private living room, Room-darkening shades	Feature Engineering
amenities_paid_parking	Paid parking	Feature Engineering
amenities_luggage_dropoff_allowed	Luggage	Feature Engineering
amenities_other	Other	Feature Engineering
amenities_suitable_for_events	Suitable for events	Feature Engineering
amenities_breakfast	Breakfast	Feature Engineering
years_since_lastreview	Number of years since last review (i.e. small number = has recent reviews)	Feature Engineering
years_of_hosting	Number of years the host has been doing airbnb	Feature Engineering

Appendix II. Additional Visualizations & Exploratory Analysis

Distribution of Airbnb Listing Grouped by Key Variables:



Correlation Heatmap for Main Features:



Appendix III. Github Repo

[Github Repo Link]