Investing with Airbnb:

Predicting Price and Availability of Listings

Dauren Bizhanov
dauren.bizhanov@duke.edu

Cindy Chiu
yuting.chiu@duke.edu

Sydney Donati-Leach
sydney.donatileach@duke.edu

Aarushi Verma
aarushi.verma@duke.edu

A final project submitted in partial fulfillment of the requirements for the course

IDS 705- Principles of Machine Learning

14 April 2022

Abstract

Home-sharing services like Airbnb can become a viable source of passive income if the host chooses the correct price point and the listing is booked regularly. Previous studies focused on predicting listing prices and few focused on availability of a listing. Availability can be an important factor in predicting revenue as the listing will be profitless without any bookings. This research builds a model that is able to predict the price of a listing as well as anticipated availability based on Airbnb data. Investors could do a simple calculation of the annual occupancy rate multiplied by the nightly price to obtain the annual revenue of a listing. In this study, we used XGBoost to train on Hawaii data. We also perform cross-domain testing in Broward County, Florida and Crete, Greece. The price model performed well in both US locations while the performance of the availability model was limited.

Introduction and Motivation

Real estate is a billion-dollar industry in which many are trying their hand (Collins, 2022). Tourism is a trillion-dollar industry that by next year is expected to reach pre-pandemic levels (Binggeli et al., 2020). One of the best ways to capitalize on both of these massive markets is with a home-sharing site like Airbnb. Listing your space on Airbnb can be more profitable than simply renting out your property since you can charge more per night than you can for monthly rent (Lemke, 2022). Once a property is up and running, if the host chooses the correct price point and the listing is booked regularly, it can become a viable source of passive income.[1] The trick is figuring out the correct price and how to maximize occupancy.

This can be a very daunting task as there are many variables that impact these figures. Airbnb collects enough information on their listings that could help someone with their decision; however, it would be too much to sort through manually. Therefore, the purpose of this report is to build a model that is able to predict the price of a listing as well as anticipated availability based on all the features available to us in Airbnb data. In order to calculate how much annual revenue an investor could gain from a listing, we then do a simple calculation of the annual occupancy rate multiplied by the nightly price.

We also want to explore if our model is generalizable and can be used to predict price and availability in areas where it has not seen the data. Therefore, we will be training our model on Airbnb data from the entire state of Hawaii and testing it on data from Broward County, Florida (Miami area) and Crete, Greece. These two testing locations are similar to Hawaii in that they are both large tourist destinations and thus have many listings. They also have similar weather and should experience the same kinds of trends in seasonality as Hawaii.

Background

Pricing is widely acknowledged to be one of the most significant factors impacting success in the hospitality industry (Hung et al.) The availability of Airbnb open data has allowed for multiple projects in this space and makes Airbnb price prediction a popular research topic.

In 2017, Wang et al. worked on Airbnb datasets from 33 cities and identified 25 price determinants from accommodation rental offers using ordinary least squares and quantile regression analysis. By categorizing these determinants, they found that property-related attributes and host-related attributes most affect Airbnb pricing. Kalehbasti et al. used various models coupled with sentiment analysis of host reviews to predict Airbnb prices in the city of New York. In their methodology, they incorporated how the consumer's sentiment affects the price. Similarly, in 2015, Li et al. applied Linear Regression to clusters obtained from the Multi-Scale

[1] "Report: New Airbnb Hosts Have Earned $1 Billion during the Pandemic." *Airbnb Newsroom*, 16 Mar. 2021, https://news.airbnb.com/report-new-airbnb-hosts-have-earned-1-billion-during-the-pandemic/.

Affinity Propagation clustering method. They studied the New York market to provide evidence that hosts with multiple properties earn more than hosts with single properties. Maseiro et al. used a quantile regression model to evaluate the relation between travel traits, with holiday homes and hotel prices. They incorporated distances between a property and landmarks in the city to create the clusters. Yang et al. and Lee et al. in their research studied the influence of the location of the property and how distance from the city center affected the price. Consistent with other research, they also found that a shorter distance would result in a higher price. In our analysis, we differ from these by providing prospective hosts an estimate of their yearly revenue to be earned from an Airbnb listing. We also attempt to generalize our model by testing it in the European market.

In regards to the real estate side of this analysis, Airdna.co is a company created to help those interested in investing in home-sharing properties. They have combined data from Airbnb, VRBO, HomeAway, and Zillow to make millions of addresses available to its users through a private API. A user can pay for this API and then simply search an address or a neighborhood to generate a variety of reports. These insights include "property valuation data, Airbnb calendars, detailed listing insights, market research including occupancy rate in any city, and dynamic pricing with historical and forward-looking pacing"[2]. Airdna has built some machine learning models to offer these kinds of insights. More specifically, they are predicting the price and the occupancy rate, or inversely the availability, of a given property. From these predictions, a user can determine if listing a property or investing in a new one would be worth their time or money. However, there are limitations to what Airdna.co provides. For example, their model to predict the estimated daily rate does not consider amenities such as a pool, wifi, or air conditioning. This is something we are interested in improving upon in our model.

Data

We sourced our data from the website Inside Airbnb[3]. This website scrapes and compiles publicly available information from Airbnb. For our analysis, we shortlisted three locations to build and test our model (Table #1). The data contained the following two tables:

- Listings: Each row is a listing available on Airbnb's site for a specific location. The columns describe different characteristics of each listing including listing price. We have 74 different features such as room type, ratings, number of bathrooms and bedrooms among others (Appendix Figure i).
- Calendar: Each row represents a listing on a specific date and its price and availability status on that date. Since the data is scraped for 365 days, each listing will have 365 rows (Figure #1).

Summary Statistic	Hawaii, United States	Broward County, United States	Crete, Greece
Total unique listings	24,294	12,531	20,137
Minimum listing price	$10.00	$10.00	$9.00
Maximum listing price	$25,000.00	$10,929.00	$21,000.00
Median listing price	$245.00	$201.00	$80.00
Train/Test	Train	Test	Test

Table #1. *Summary Statistics of Source Data From Inside Airbnb. Prices for listings in Crete, Greece were converted from local currency to US dollars.*

[2] Vacation Rental Data to Set You Apart. Insights to Keep You Ahead. https://www.airdna.co/.
[3] "Get the Data." Inside Airbnb, http://insideairbnb.com/get-the-data/

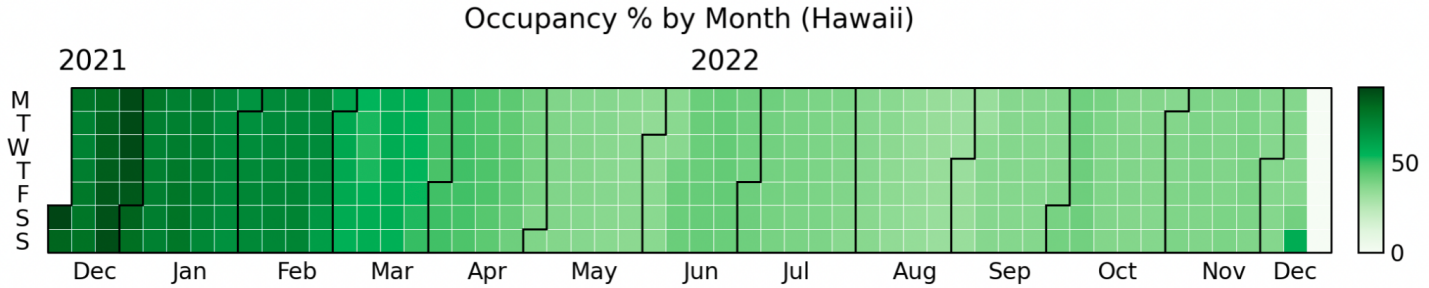

Figure #1. *Percent Occupancy Rate for Listings in Hawaii Dec 2021- Dec 2022: The occupancy rate is high during the start of the calendar data because this was closest to the time the data was scraped. Since we did not have access to past data, our analysis is limited to the future expected occupancy rates.*

Methods and Experiments

Our end goal was to calculate revenue of a listing for a hypothetical real estate investor, so we separated the project into two different aspects: predicting price and predicting availability. Revenue can then be calculated by multiplying the price of the listing with the percentage of occupancy.

To predict the price of a listing given all the attributes captured in our data, we wanted to build a supervised regression model because the outcome variable is continuous. To achieve this, we leveraged both the calendar data and listing data. We incorporated each listing's average monthly rate from the calendar data to capture seasonality since the price for an Airbnb listing tends to fluctuate throughout the year (Svetec et al., 2022). Due to the skewness of the price variable we log transformed it (Figure #2). We performed feature engineering on almost all of the columns to eliminate null values and create robust information for our models (Appendix Figure i). The evaluation metrics of our price prediction model were both RMSE and MAE. We also generated a baseline model that made its predictions from the listing's average monthly rate. We compared the RMSE and MAE on the validation set to pick the optimal model.

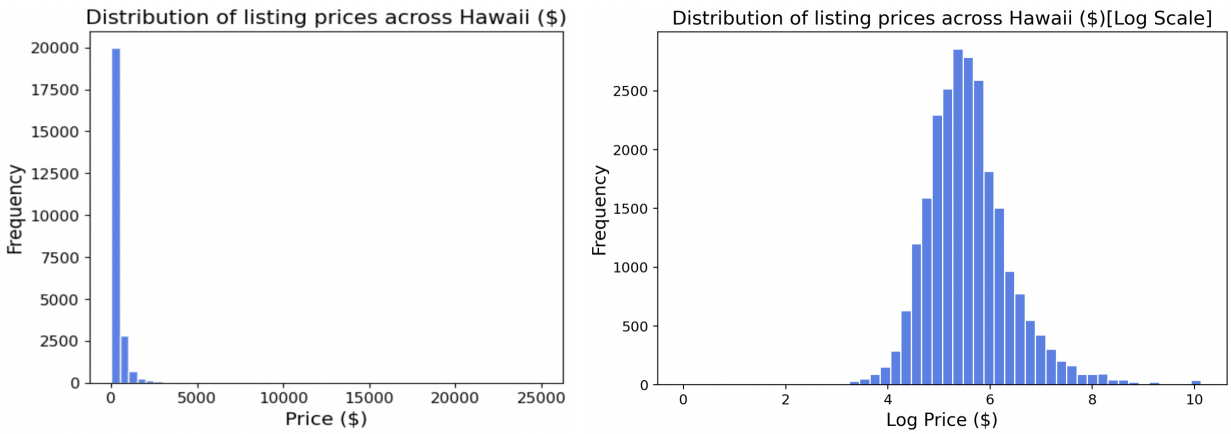

Figure #2. *Distribution of Listing Price per Night for Hawaii: The outcome variable, price, is skewed (left), but after converting price to log scale, it becomes normally distributed (right).*

To predict the percent yearly occupancy, or availability of a listing, we categorized the outcome into three different buckets (0-29.9%, 30-69.9% and 70-100%) and performed supervised classification. This binning strategy balanced the number of records within each category. The evaluation metrics of our availability prediction model were precision, recall, and F1-scores. We used confusion matrices to understand in which class the model performed the worst. We also generated a baseline model which made random predictions for each listing.

After preprocessing all of our Hawaii training data, we separated it into 20% validation and 80% training. We created models using XGBoost (Chen et al.), LGBM (Ke et al.), and Random Forest (Ho). We made the decision to use tree-based models because they are known to be superior in predicting outcomes for tabular data (Shwartz-Ziv et al., 2021). To test whether our model is generalizable across different locations, we performed cross-domain testing on two other locations. A more detailed overview of our experiment can be seen in Figure #3.

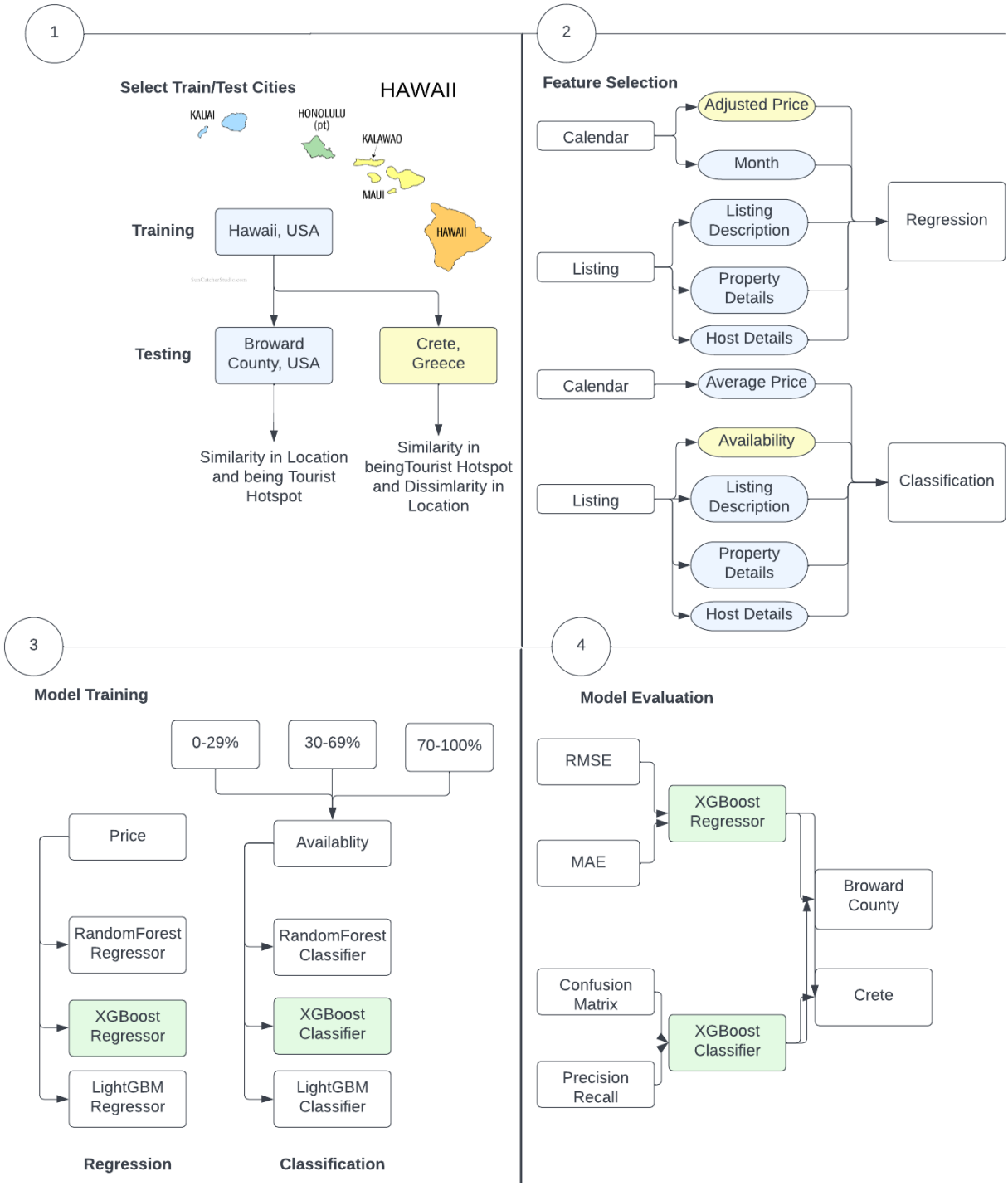

Figure #3. *Flowchart of Experimental Design: This begins with the selection of our training and testing locations (1), continuing with feature selection which included feature engineering of the two different outcome variables: listing price and availability (2). Different models were explored when training on the Hawaii data (3), and evaluated with RMSE and MAE for regression, and confusion matrices and Precision/Recall for classification (4). The XGBoost model achieved the best performance in both cases and was used to test in Broward County and Crete, Greece.*

Results

Price Model

For our price prediction model, Table #2 provides the results on the validation set. Even though we tried to optimize the hyperparameters of our models using Bayesian optimization, we were not able achieve better performance. Therefore, we chose XGBoost with default hyperparameters as our optimal model.

Model	RMSE		MAE	
	Log Scale ($)	Actual Scale ($)	Log Scale ($)	Actual Scale ($)
Naive prediction	0.84	1299.93	0.63	321.42
XGBoost with default hyperparameters	0.43	1122.06	0.30	186.29
LGBM with default hyperparameters	0.45	1133.88	0.32	196.78
Random Forest with default hyperparameters	0.46	1136.00	0.30	186.67
LGBM with optimized hyperparameters	0.44	1125.32	0.31	190.18

Table #2. *RMSE and MAE of the Price: XGBoost with default parameters has the smallest RMSE and MAE. Our model was able to predict the price with a mean average error of $186.29.*

To evaluate exactly where the XGBoost model performs well and where it does not perform well, we can look at a few different plots of our outcome variable. If we first look at a distribution of the model's residual values, we can immediately see a normal curve (Figure # 4). However, we had a bit of a right tail on the distribution of the log price which should make us question if we can predict as well for all listings. To dive deeper into that question, we can look at a plot of each listing's predicted price versus its actual price (Figure #5). This shows us that our model struggled to predict extremely high prices.

This limitation is due to the small number of samples with such prices in our training data. In addition, our feature engineering may have influenced the quality of predictions. For example, we binned discrete features into groups to reduce the cardinality of the feature space (Appendix Figure i). The rationale behind this decision was to achieve more robust results on the test data and not have to remove any listings that had null values. However, this meant we would be keeping outliers in our data; therefore, any predictions higher than $8000 should be treated skeptically.

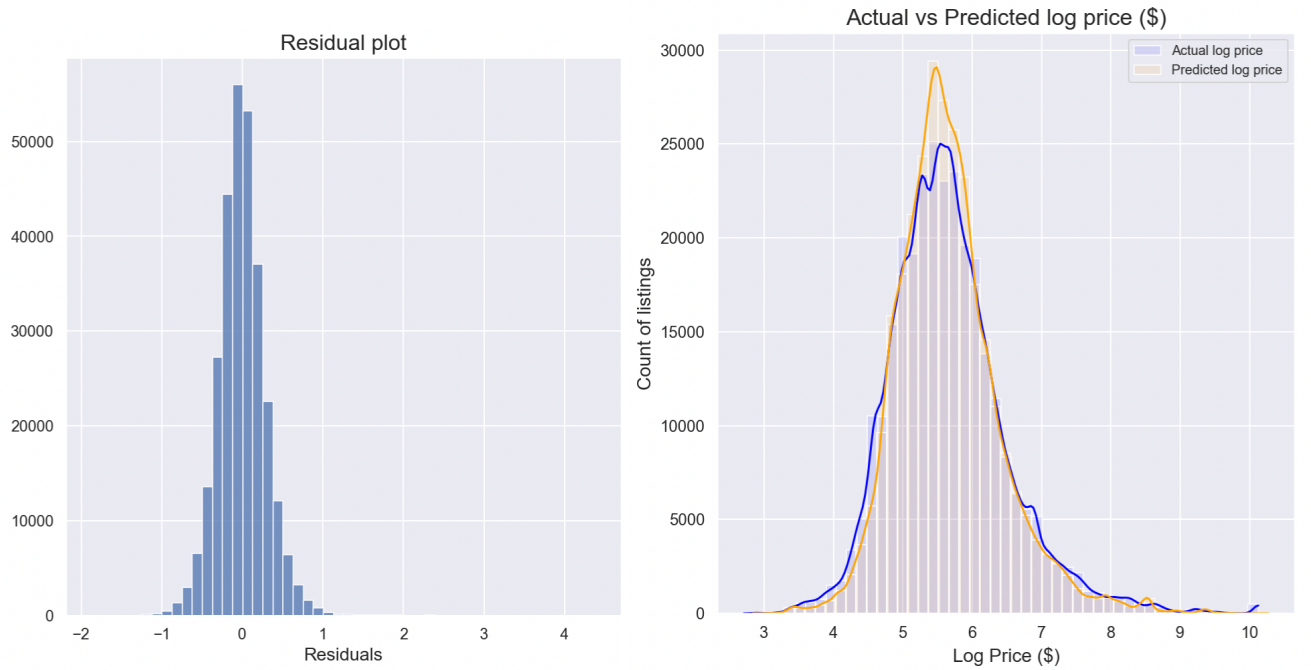

Figure #4. *Predicted and Actual Price Distributions: The residual plot (left) shows that our predictions were centered around zero, meaning they did not deviate much from the actual values. The plot on the right shows the distribution of the predicted log price (orange) overlays the actual log price (purple) well, except for a right tail which indicates that we did not have as much data for higher priced listings.*

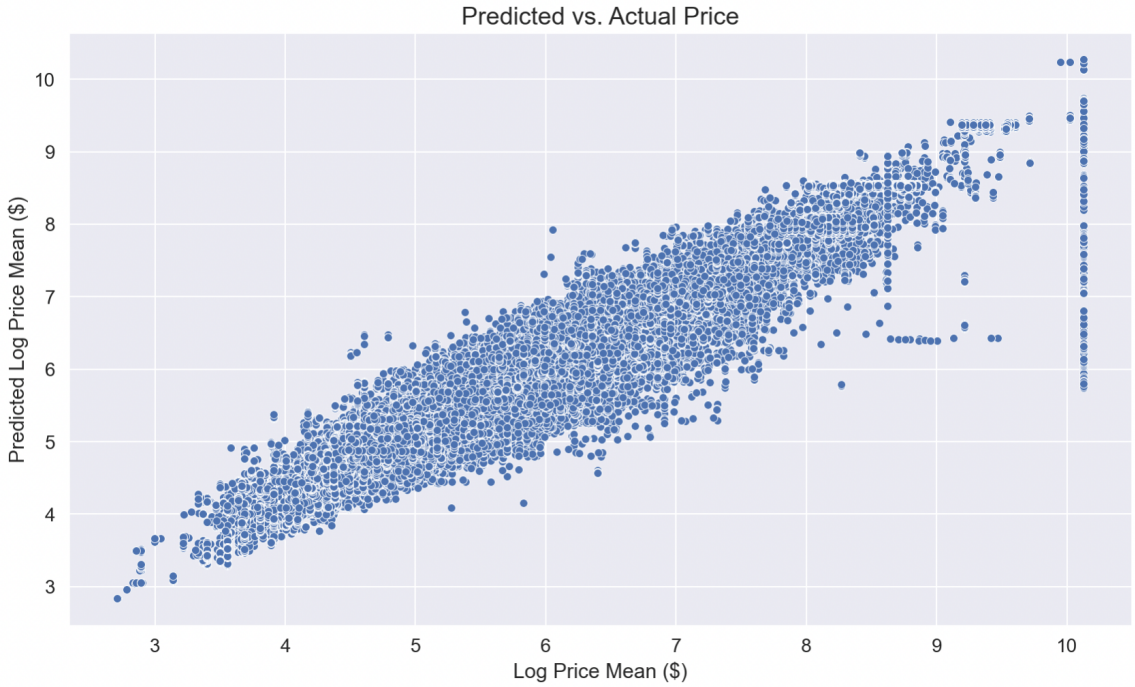

Figure #5. *Predicted vs Actual Price: Each point represents a listing. Ideally all of the points should have the same value on the x-axis as it does on the y-axis. This signifies that the predicted value of the price is the same as the actual price. The most deviation from the actual log price can be seen in the upper right hand corner of the plot, where price is the highest.*

Another area where our XGBoost price prediction model did not perform as well was when trying to capture the seasonality of the listing prices. There is a clear decline in prices in May and September because these months are an off season for travel and tourism (Masson, 2020). Consequently, we see a peak in price in the winter months (December, January, February) as these are warm destinations that attract tourists from the northern hemisphere. Our model tried to account for this seasonality by having a feature that determined the price that occurred for each listing each month. Unfortunately, it was not able to capture all the variation that actually occurs across months.

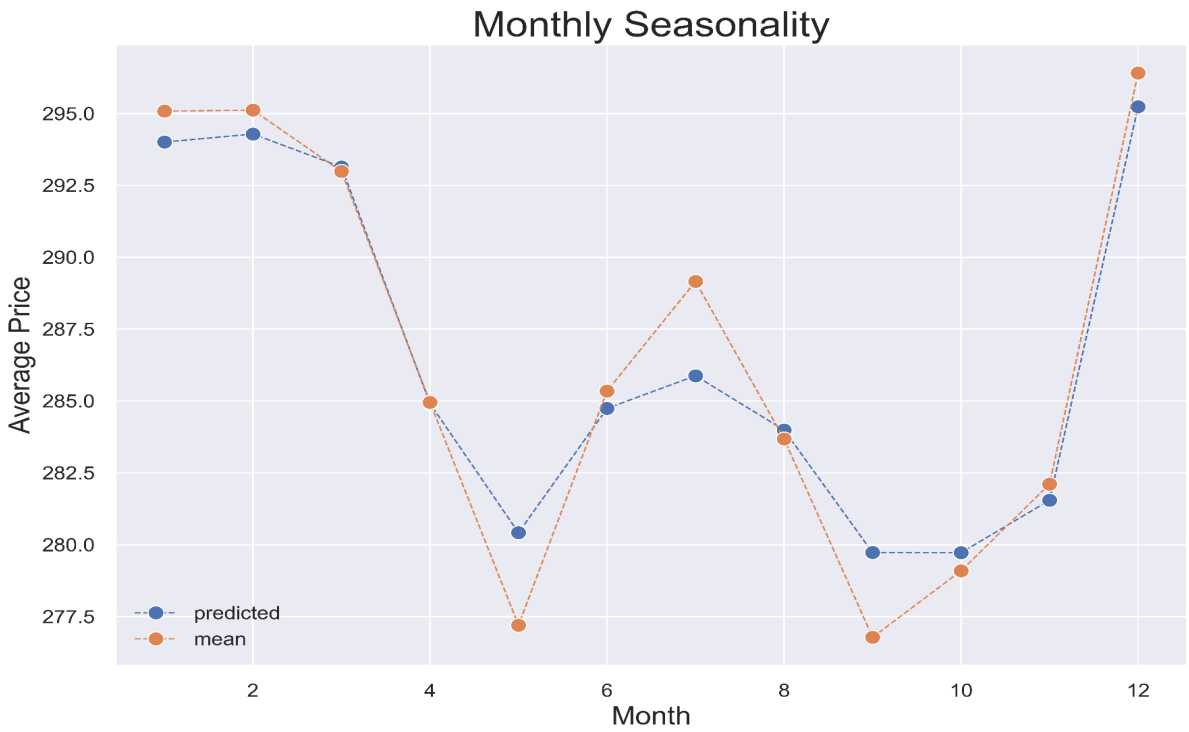

Figure #6. *Monthly average and predicted prices. This figure shows the average price between the actual data and the predicted price. Our model captured the monthly trend but had less variation across each month. Therefore, we overpredicted in May and September and underpredicted in June.*

After taking all of this into consideration, we can move forward with testing the performance of the XGBoost model in Broward County, FL and Crete, Greece. The results of the price prediction model on the test data is shown in Table #3. In comparison to Hawaii, MAE degraded by 76.7% in Broward and by 286.6% in Crete. The significant drop in performance in Crete is not surprising as its price distribution was very different from the training data. Crete had a median listing price of only $80 whereas Hawaii was around $245.

City	**RMSE**		**MAE**	
	Log Scale ($)	Actual Scale ($)	Log Scale ($)	Actual Scale ($)
Broward	0.66	414.50	0.53	167.94
Crete	1.38	1162.66	1.24	406.35

Table #3. *Model Performance of XGBoost model on test datasets. Broward County, FL performs better than Crete, Greece with a mean average error of $167.94.*

Another strategy we employed to understand why our model did not perform as well in Crete was to look at the features the model found important. With an XGBoost model, this can be done with the Gain metric.

This metric gives insight into a feature's relative contribution calculated by taking each feature's contribution for each tree in the model (Chen et al.). A higher value of this metric when compared to another feature implies it is more important for generating a prediction. We found that the price model puts high importance on features that describe the listing's physical space (Figure #7). European spaces are fundamentally different and more compact than spaces in the United States (McMaken, 2019); therefore, this could also be why our model does not generalize well in Crete.

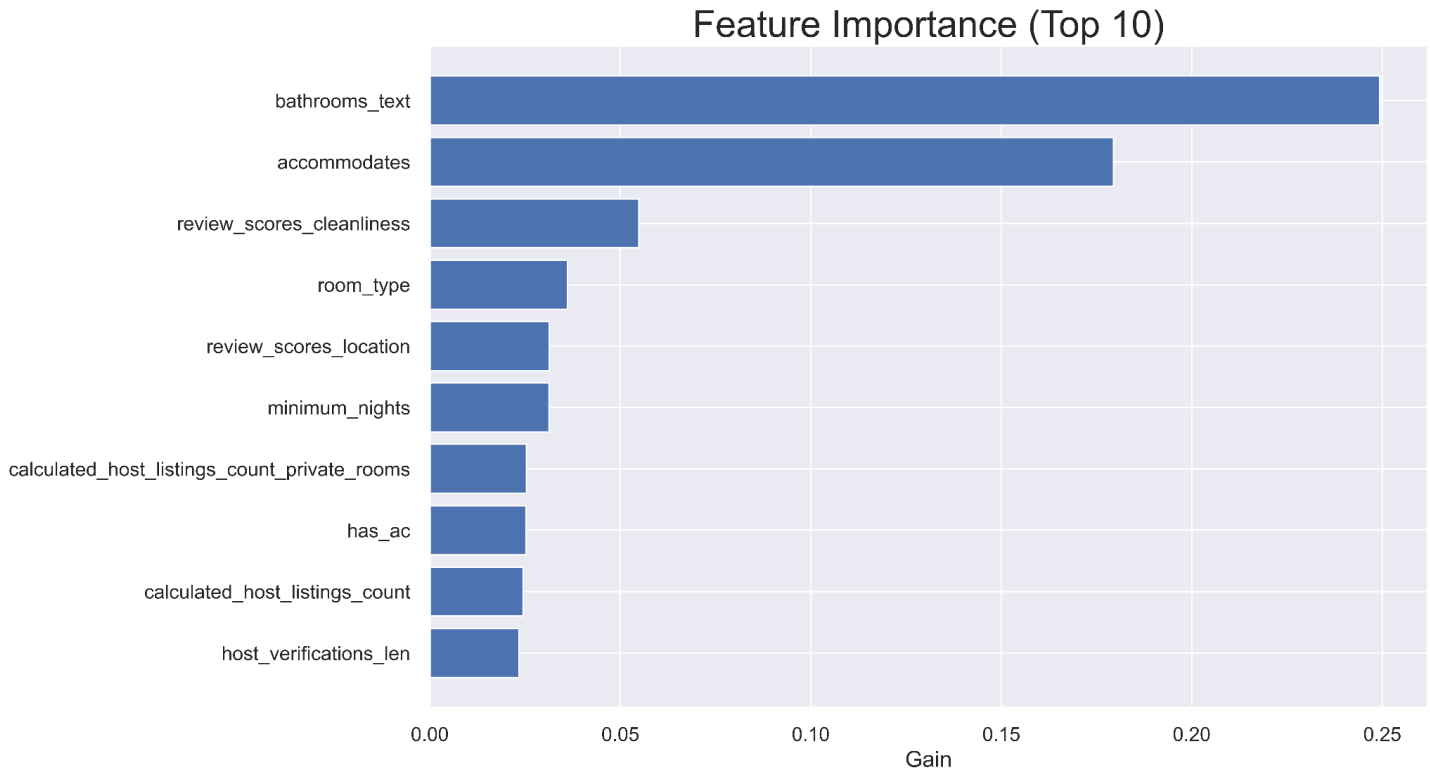

Figure #7*. Gain based feature importance of the price prediction model: The top features are highly dependent on the listing's space, such as number of bathrooms and number of people the space can accommodate.*

Availability Model

To continue with our goal of calculating revenue, we also needed to predict a listing's availability. While the listing's availability rate is a continuous variable in the data, we decided to create three separate classes for it. Our reasoning behind this was motivated by looking at the problem from a business perspective. The availability of a listing may vary significantly if the guest decides to alter their booking by a day or two. Additionally, only 20% of the listings had bookings two months in advance, but normally listings reach an average occupancy of 80% only ten days in advance (May, 2016). Therefore, an availability range would be safer for an investor's decision-making process.

We used the same set of models to predict a listing's availability as we did for price prediction: XGBoost, LGBM, and Random Forest. We decided to move forward with the XGBoost model because the evaluation metrics between it and the other models were very similar if not the same, and it required less computational power to run (Table #4).

Model	Precision	Recall	F1-Score
Random Guess	0.33	0.33	0.33
XGBoost Default	0.66	0.65	0.65
LGBM Default	0.66	0.65	0.65
Random Forest Default	0.67	0.66	0.66
Random Forest Optimized	0.68	0.65	0.65

Table #4. *Availability model performance on the validation set (weighted average). All models performed almost at the same level in terms of precision with slight dominance with the optimized Random Forest model. However, the training and prediction speed of the Random Forest model is much slower than XGBoost and LGBM classifiers. Therefore, we decided to use the XGBoost model and be consistent with our model choices. All models perform approximately 50% better than random guessing.*

Now that we have a general understanding of which model performs best, we can dive into how it performs within each class (Figure #8, Table #5). Our model predicts high occupancy, class 0, with the highest AUC of 0.85 and the highest precision of 0.74. We decided to focus on precision metrics in our problem because from an investor's perspective it is more important to be more confident in the predictions and it is less critical to miss some opportunities (recall). To be more specific, an investor would be comfortable settling with less opportunities that they know are more accurate. The class with the highest uncertainty is the medium occupancy group, class 1, which has an AUC of 0.79 and a precision of 0.59 on the validation data set.

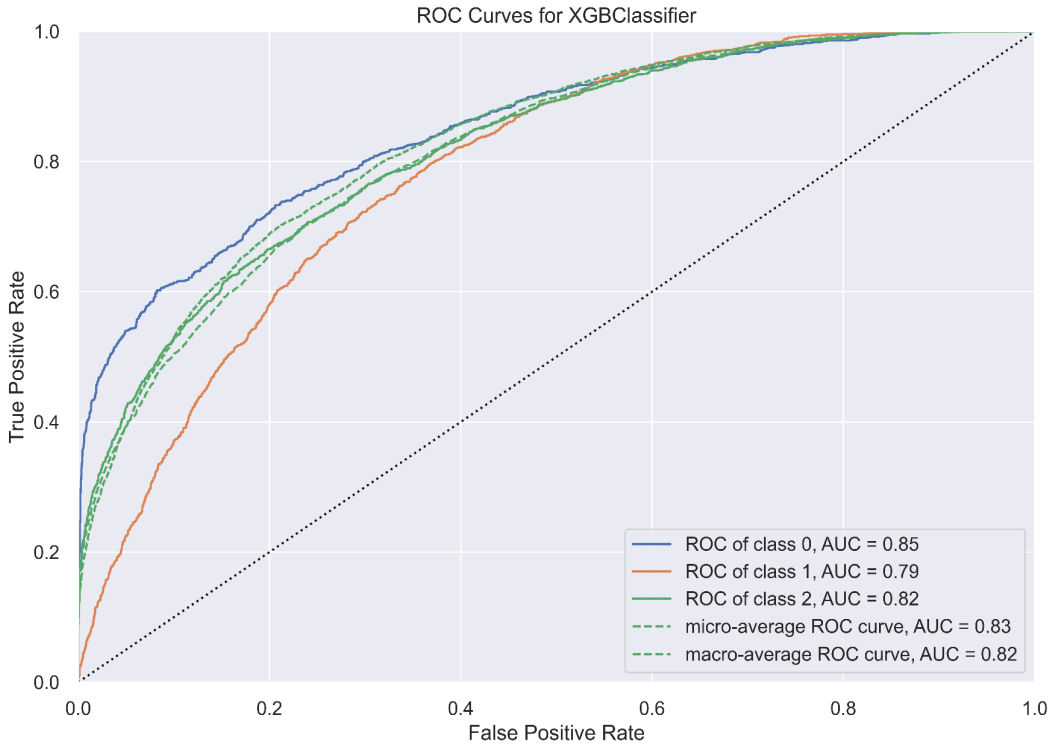

Figure #8. *XGBoost ROC Curves for the validation data set. The high occupancy rate category, class 0, performs the best out of all three classes. We have high AUC for each class and the micro and macro-averages (green dashed lines) measure the performance across all 3 classes and the average AUC is 0.82, which outperformed our baseline model (black dashed line).*

Class	Precision	Recall	F1-Score	Support (No. of listings)
0: High occupancy (70-100%)	0.74	0.58	0.65	1285
1:Medium occupancy (30-69.9%)	0.59	0.74	0.66	1904
2: Low occupancy (0-29.9%)	0.68	0.60	0.64	1676

***Table #5**. Performance metrics within each class on the validation set. Looking at the precision score, the high occupancy class performs the best at 0.74 whereas the medium occupancy class performs the worst at 0.59.*

We wanted to explore what caused the precision in class 1 to be lower than the other classes, so we examined the confusion matrix of the validation data. Figure #9 shows that many of the actual class 1 listings are predicted as class 0 or class 2. This is most likely due to how we put this originally continuous variable into three categories. Even though listings with 29% and 30% occupancy are relatively similar, listings with 29% occupancy will be categorized as class 2 and listings with 30% occupancy will be categorized as class 1. Our model is not able to easily identify the cutoff points between classes, and mispredicts one class to the other class that is right next to it.

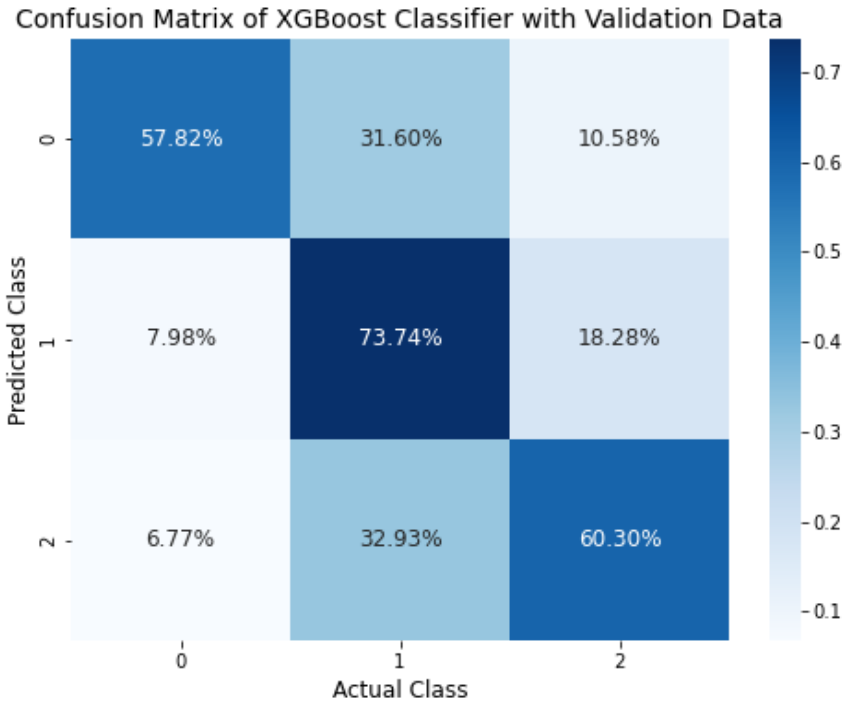

***Figure #9.** Confusion matrix of XGBoost Classifier with validation data. The model often mispredicts listings in class 1 as class 0 or class 2. The model is not able to distinguish the cutoff points between classes.*

Even though we know our model already does not perform well in all classes, we still went ahead and tested it on both Broward County and Crete to see if that pattern continues. Table #6 shows the performance metrics of XGBoost for the testing data. The precision of our testing data is worse in general. We examined the confusion matrices and further explored the performance of each class. It had the same issue as the validation data and more data is classified to the category right next to it (Figure #10).

City	Precision	Recall	F1-Score
Broward	0.44	0.42	0.42
Crete	0.43	0.41	0.42

Table #6. *Availability model performance on the testing set (weighted average). This shows our model is not generalizing well on locations other than Hawaii.*

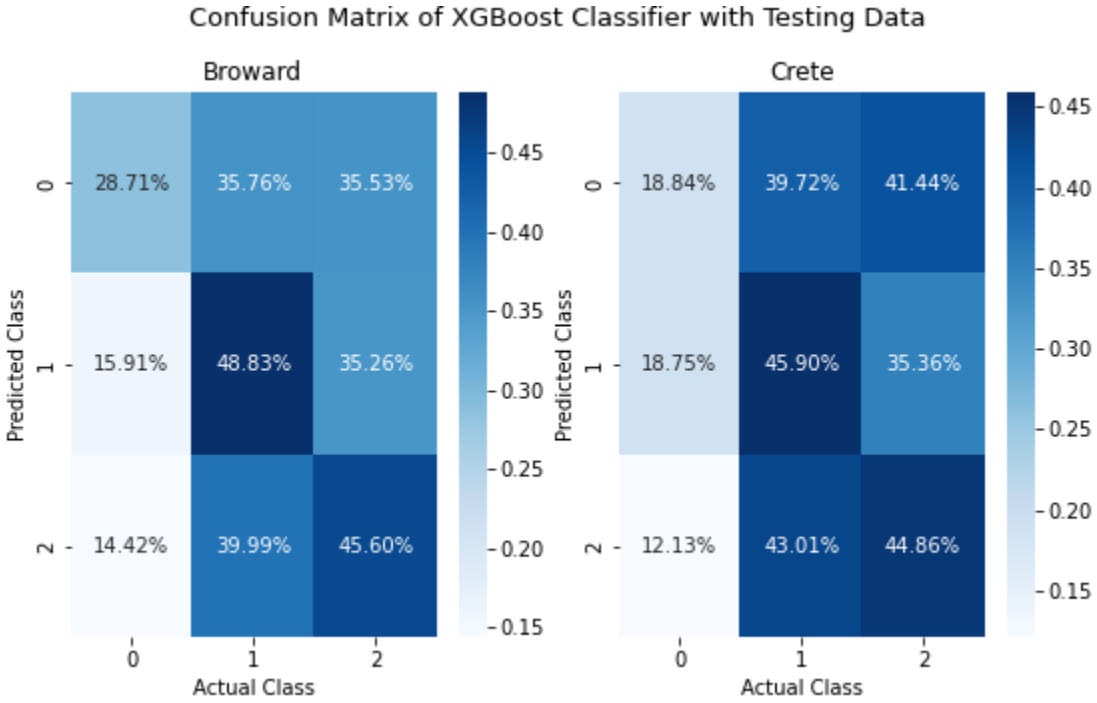

Figure #10. *The confusion matrix of test data within each class. This shows our data is not generalizing well on predicting the occupancy classes and tends to misclassify the listing to another class.*

Another method to understand why our model did not perform well in Hawaii, Broward or Crete is to look at the gain based feature importances of the classification model (Figure #11). Unlike the price regression model, the most important features for availability are host-based. These include whether the listing is instantly bookable and the number of listings the host owns. There are many listings where the host does not add this amount of detail, therefore these values would be null. There are also text-based features such as neighborhood description (Appendix Figure i #129-131) and listings descriptions (Appendix Figure i #124-128) that are relatively important in this model. This is cause for speculation as the description fields in the test sets contain a lot of vocabulary that possibly cannot be generalized to other listings or to other locations.

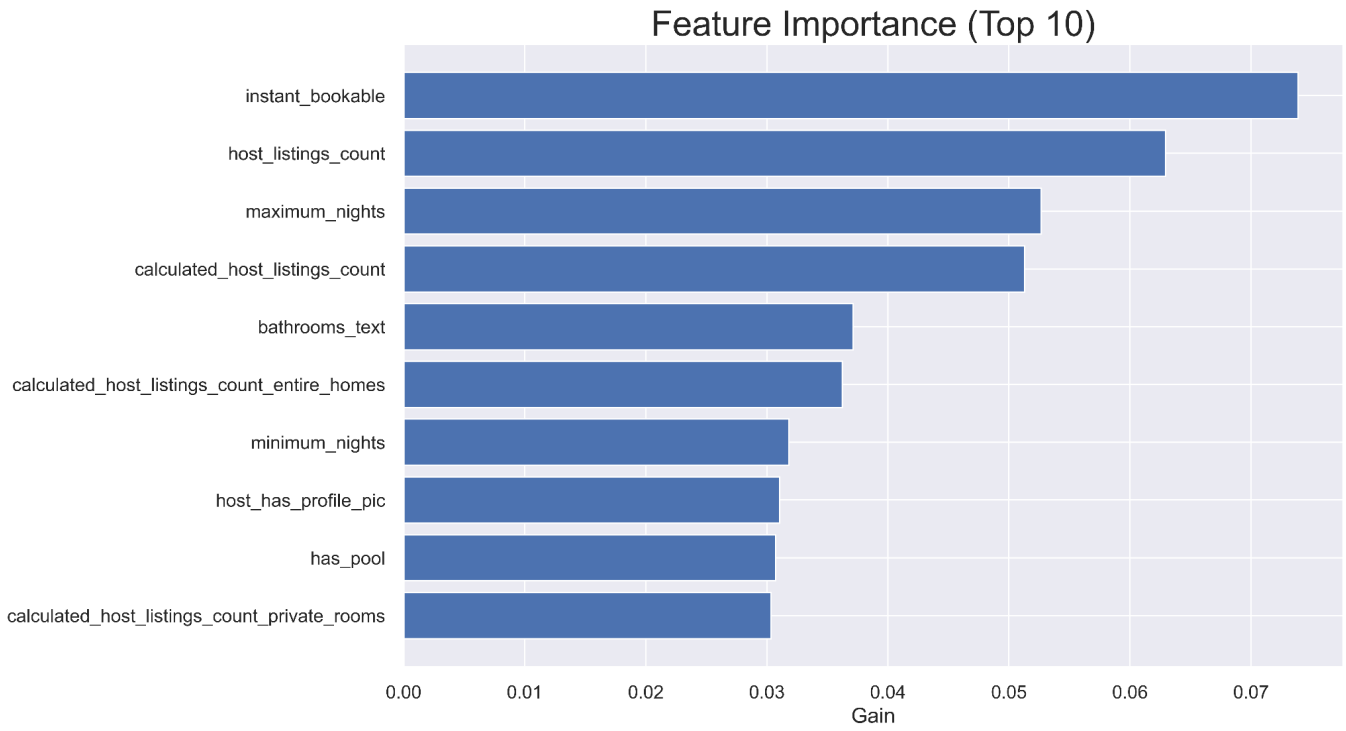

Figure #11. *Gain based feature importance of the availability prediction model. The top features are highly dependent on the host's behavior, such as whether the listing can be booked without communication and the number of Airbnb listings the host owns.*

Revenue

Our model to predict listing price and our model to predict a listing's occupancy rate are two important tools that a real estate investor could use to determine their potential revenue from a property. Since our predicted availability falls into three different categories with a lower and upper bound for each, the potential revenue would also be a range with a lower and upper bound. For example, an interested investor tells us their property has 2 bedrooms, less than 3 bathrooms, can accommodate 6 people and guests must stay a minimum of 3 nights. Our model would predict they can charge around $165 per night and would fall within 30-69.9% annual occupancy, so they could earn revenue between $1,690 and $12,389 per year. They can of course provide more specifications about their property which will allow us to nail down their price point and availability even more precisely.

Conclusions

Our price model using XGBoost Regressor performed well on predicting Airbnb listing prices in Hawaii and Broward County, USA. It is not generalizable outside the US as it performed much worse in Crete, Greece. Important features for predicting listing prices were space dependent such as how many guests can be accommodated and how many bathrooms there are in the property.

For predicting yearly availability, our model using XGBoost Classifier performed moderately well. There was a lot of variability within the medium occupancy class due to the lack of ability to identify listings close to the cutoff point. This also means the model was not able to perform well in Broward County or Crete. Therefore, our model can be used to predict revenue with lower and upper bounds only in Hawaii.

There are a few limitations in this analysis. First, the data only provided the insight for booking 365 days in advance, and did not include any historical information. Airbnb guests do not book their stays far in advance, so we would not be able to capture the true occupancy rate until the booking date has passed. Furthermore, there is no clear identifier in the data to distinguish whether a listing was booked or blocked by the host. Future work can incorporate methodologies from other studies such as the distance from the city center which can impact price. Also, access to data over multiple years could improve the price model to better capture the seasonality changes, and improve the availability model to achieve better accuracy.

Roles

Aarushi- She was able to use her background in consulting to frame our story and put a positive spin on our model outcomes and limitations. She worked with Cindy on the EDA for the listing data and worked on feature engineering for variables in the listing data. She focused on telling the audience the significance of our project and sourcing past studies as references. She also created presentation slides and competed with Sydney in a "presentation-off".

Cindy- She primarily worked on the EDA for the listing data and calendar data. She also worked on feature engineering to aid Dauren in model building. Cindy has written the Methods and Experiments part of the report. She also created advanced plots to turn our findings into great visualizations which are easy for the audience to understand. She was also in charge of recording the video presentation. She took the leadership to chalk out the group's timeline to work on the project and ensure all tasks were completed as per the deadlines

Dauren- He was our main coder and resident expert in machine learning. He automated the scripts and generated reproducible code in github. He built out and fine tuned several models to predict both price and availability and evaluated them based on regression and classification standards. He documented the results section of the paper. He was the key driver for all our model experiments and explorations.

Sydney- Through EDA she discovered how prices are affected by seasonality, and also determined the difference between 'price' and 'adjusted_price' so the team could decide which would be used as the outcome of our predictions. She created feature engineering for the ratings by determining the percentiles that would separate each rating to make it a categorical variable. She also created our presentation slides, competed with Aarushi in a "presentation-off" and focused on telling the audience our motivation and results.

References

Binggeli, Urs, et al. “Covid-19 Tourism Spend Recovery in Numbers.” *McKinsey & Company*, McKinsey & Company, 5 Nov. 2020, https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/covid-19-tourism-spend-recovery-in-numbers.

Chen, Tianqi, and Carlos Guestrin. “XGBoost.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016

Collins, G. (2022, January 17). The US Rental Property Market Outlook. ManageCasa. Retrieved April 12, 2022, from https://managecasa.com/articles/us-rental-property-market/

Feldman, Jess, and Tim Lemke. “Pros and Cons of Airbnb as an Investment Strategy.” *The Balance*, The Balance, 6 Jan. 2022, https://www.thebalance.com/pros-and-cons-of-airbnb-as-an-investment-strategy-4776231.

Hung, Wei-Ting, et al. “Pricing Determinants in the Hotel Industry: Quantile Regression Analysis.” *International Journal of Hospitality Management*, vol. 29, no. 3, 2010, pp. 378–384.

Kam, Ho Tin. "Random decision forest." Proceedings of the 3rd international conference on document analysis and recognition. Vol. 1416. Montreal, Canada, August, 1995.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 3146-3154

Lee, Seul Ki, and SooCheong (Shawn) Jang. “Premium or Discount in Hotel Room Rates? the Dual Effects of a Central Downtown Location.” *Cornell Hospitality Quarterly*, vol. 53, no. 2, 2012, pp. 165–173.

Li, Yang, et al. “Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering.” 2016 35th Chinese Control Conference (CCC), 2016.

Masiero, Lorenzo, et al. “A Demand-Driven Analysis of Tourist Accommodation Price: A Quantile Regression of Room Bookings.” International Journal of Hospitality Management, vol. 50, 2015, pp. 1–8.

Masson, Thibault. “Real Airbnb Booking Data Show Which Travel Trends Should Do Well in Winter, Spring, and Summer 2021.: Rental Scale.” *Rental Scale-Up*, 8 Dec. 2020, https://www.rentalscaleup.com/real-airbnb-booking-data-show-which-travel-trends-should-do-well-in-winter-spring-and-summer-2021/.

May, Kevin. “Number-Crunching Reveals Best Time to Book on Airbnb.” *PhocusWire*, PhocusWire,6 Dec. 2016, https://www.phocuswire.com/Number-crunching-reveals-best-time-to-book-on-Airbnb.

McMaken, Ryan. “Americans Have Much More Living Space than Europeans: Ryan McMaken.” *Mises Institute*, 11 Mar. 2019, https://mises.org/power-market/americans-have-much-more-living-space-europeans.

Rezazadeh Kalehbasti, Pouya, et al. “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis.” International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2021, 2021, pp. 173–184.

Shwartz-Ziv, Ravid, and Amitai Armon. "Tabular data: Deep learning is not all you need." *Information Fusion* 81 (2022): 84-90.

Svetec, James, and Symon He. “Baseline Pricing for Your Airbnb.” Airbnb For Dummies, 17 Mar. 2022, https://www.dummies.com/article/home-auto-hobbies/travel/baseline-pricing-for-your-airbnb-271329/.

Wang, Dan, and Juan 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, 2017, pp. 120–131.

Yang, Yang, et al. “Hotel Location Evaluation: A Combination of Machine Learning Tools and Web Gis.” *International Journal of Hospitality Management*, vol. 47, 2015, pp. 14–24., https://doi.org/10.1016/j.ijhm.2015.02.008.

Appendix

Figure i. Data Dictionary

#	Column Name	Column Description	Data Set	Feature Type
1	id	Unique id for each property/listing	listings.csv	Original
2	listing_url	Airbnb url pertaining to the listing	listings.csv	Original
3	scrape_id	Inside Airbnb "Scrape" this was part of	listings.csv	Original
4	last_scraped	Date and time the listing was last scraped	listings.csv	Original
5	name	Name of the property/listing	listings.csv	Original
6	description	Description of the property/listing	listings.csv	Original
7	neighborhood_overview	Host's description of the neighborhood	listings.csv	Original
8	picture_url	URL to the Airbnb hosted regular sized image for the listing	listings.csv	Original
9	host_id	Airbnb's unique identifier for the host/user	listings.csv	Original
10	host_url	The Airbnb page for the host	listings.csv	Original
11	host_name	Name of the host. Usually just the first name(s).	listings.csv	Original
12	host_since	The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.	listings.csv	Original
13	host_location	The host's self reported location	listings.csv	Original
14	host_about	Description about the host	listings.csv	Original
15	host_response_time	time taken by host to respond	listings.csv	Original
16	host_response_rate	Rate of host's responses to messages (%)	listings.csv	Original
17	host_acceptance_rate	That rate at which a host accepts booking requests.	listings.csv	Original
18	host_is_superhost	Is the host a superhost	listings.csv	Original
19	host_thumbnail_url	AirBnB URL for host's thumbnail	listings.csv	Original
20	host_picture_url	AirBnB URL for host's picture	listings.csv	Original
21	host_neighbourhood		listings.csv	Original
22	host_listings_count	no. of host's listing within specified region (as per file name)*	listings.csv	Original
23	host_total_listings_count	total listing by the host on Airbnb*	listings.csv	Original
24	host_verifications	list of verifications performed by the host. (email, phone, reviews) etc.	listings.csv	Original
25	host_has_profile_pic	Whether the host has a profile pictures	listings.csv	Original
26	host_identity_verified	Whether host's identity has been verified	listings.csv	Original
27	neighbourhood	where the property is located	listings.csv	Original
28	neighbourhood_cleansed	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	listings.csv	Original

29	neighbourhood_group_cleansed	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	listings.csv	Original
30	latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	listings.csv	Original
31	longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	listings.csv	Original
32	property_type	Apartment, House, Boat, Yurt etc.	listings.csv	Original
33	room_type	Entire Home, Private Room, Shared Home etc.	listings.csv	Original
34	accommodates	The maximum capacity of the listing	listings.csv	Original
35	bathrooms	The number of bathrooms in the listing	listings.csv	Original
36	bathrooms_text	The number of bathrooms in the listing. On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, bathrooms is used.	listings.csv	Original
37	bedrooms	The number of bedrooms	listings.csv	Original
38	beds	The number of bed(s)	listings.csv	Original
39	amenities	list of amenities available at the property	listings.csv	Original
40	price	Daily price for listing in USD	listings.csv	Original
41	minimum_nights	minimum number of night stay for the listing	listings.csv	Original
42	maximum_nights	maximum number of night stay for the listing	listings.csv	Original
43	minimum_minimum_nights	the smallest minimum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
44	maximum_minimum_nights	the largest minimum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
45	minimum_maximum_nights	the smallest maximum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
46	maximum_maximum_nights	the largest maximum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
47	minimum_nights_avg_ntm	the average minimum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
48	maximum_nights_avg_ntm	the average maximum_night value from the calender (looking 365 nights in the future)	listings.csv	Original
49	calendar_updated		listings.csv	Original
50	has_availability	whether property is available	listings.csv	Original
51	availability_30	The availability of the listing 30 days in the future as determined by the calendar	listings.csv	Original
52	availability_60	The availability of the listing 60 days in the future as determined by the calendar	listings.csv	Original
53	availability_90	The availability of the listing 90 days in the future as determined by the calendar	listings.csv	Original
54	availability_365	The availability of the listing 365 days in the future as determined by the calendar	listings.csv	Original
55	calendar_last_scraped	Date and time the calendar was last scraped	listings.csv	Original

56	number_of_reviews	The number of reviews the listing has	listings.csv	Original
57	number_of_reviews_ltm	The number of reviews the listing has (in the last 12 months)	listings.csv	Original
58	number_of_reviews_l30d	The number of reviews the listing has (in the last 30 days)	listings.csv	Original
59	first_review	The date of the first/oldest review	listings.csv	Original
60	last_review	The date of the last/newest review	listings.csv	Original
61	review_scores_rating	On a scale of 1-5 rating given	listings.csv	Original
62	review_scores_accuracy	On a scale of 1-5 review scores for accuracy	listings.csv	Original
63	review_scores_cleanliness	On a scale of 1-5 review scores for cleanliness	listings.csv	Original
64	review_scores_checkin	On a scale of 1-5 review scores for check in	listings.csv	Original
65	review_scores_communication	On a scale of 1-5 review scores for communication	listings.csv	Original
66	review_scores_location	On a scale of 1-5 review scores for location	listings.csv	Original
67	review_scores_value	On a scale of 1-5 review scores for value	listings.csv	Original
68	license	The licence/permit/registration number	listings.csv	Original
69	instant_bookable	Whether the guest can automatically book the listing without the host requiring to accept their booking request	listings.csv	Original
70	calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.	listings.csv	Original
71	calculated_host_listings_count_entire_homes	The number of Entire home/apt listings the host has in the current scrape, in the city/region geography	listings.csv	Original
72	calculated_host_listings_count_private_rooms	The number of Private room listings the host has in the current scrape, in the city/region geography	listings.csv	Original
73	calculated_host_listings_count_shared_rooms	The number of Shared room listings the host has in the current scrape, in the city/region geography	listings.csv	Original
74	reviews_per_month	The number of reviews the listing has over the lifetime of the listing	listings.csv	Original
75	listing_id	Unique id for each property/listing	calendar.csv	Original
76	date	Calendar date	calendar.csv	Original
77	available	Whether listing is available	calendar.csv	Original
78	price	Listing price is USD	calendar.csv	Original
79	adjusted price	Adjusted price based on offer made by host to guest	calendar.csv	Original
80	minimum_nights	Minimum nights the listing can be booked for	calendar.csv	Original
81	maximum_nights	Maximum nights the listing can be booked for	calendar.csv	Original
82	host_response_time	"1": within an hour, "2": within a few hours, "3": within a day, "4": a few days or more, "5": missing	revenue.csv	Engineered

83	room_type	"1": Entire home/apt,"2": Private room, "3": Hotel room, "4": Shared room, "5": missing	revenue.csv	Engineered
84	host_has_profile_pic	"1": yes, "0": no	revenue.csv	Engineered
85	host_identity_verified	"1": yes, "0": no	revenue.csv	Engineered
86	bathrooms_text	"1": 1, "2": 2, "3": 3, "4": more than 3, "5": "missing	revenue.csv	Engineered
87	bedrooms	"1": 1, "2": 2, "3": 3, "4": 4, "5": more than 4, 6: missing	revenue.csv	Engineered
88	beds	"1": 1, "2": 2, "3": 3, "4": 4, "5": more than 4, 6: missing	revenue.csv	Engineered
89	review_scores_rating	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
90	review_scores_accuracy	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
91	review_scores_cleanliness	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
92	review_scores_checkin	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
93	review_scores_communication	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
94	review_scores_location	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
95	review_scores_value	Quantile classification of rating scores : "1": [0-25), "2": [25-50), "3": [50-75), "4": [75-100], "missing": 5	revenue.csv	Engineered
96	instant_bookable	"1": yes, "0": no	revenue.csv	Engineered
97	has_wifi	"1": yes, "0": no	revenue.csv	Engineered
98	has_pool	"1": yes, "0": no	revenue.csv	Engineered
99	has_kitchen	"1": yes, "0": no	revenue.csv	Engineered
100	has_washer	"1": yes, "0": no	revenue.csv	Engineered
101	has_dryer	"1": yes, "0": no	revenue.csv	Engineered
102	has_ac	"1": yes, "0": no	revenue.csv	Engineered
103	has_self_checkin	"1": yes, "0": no	revenue.csv	Engineered
104	has_workspace	"1": yes, "0": no	revenue.csv	Engineered
105	has_pet_allowed	"1": yes, "0": no	revenue.csv	Engineered
106	has_free_parking	"1": yes, "0": no	revenue.csv	Engineered
107	id	Unique id for each property/listing	revenue.csv	Original
108	host_response_rate	Rate of host's responses to messages (%)	revenue.csv	Original

109	host_acceptance_rate	That rate at which a host accepts booking requests.	revenue.csv	Original
110	host_listings_count	no. of host's listing within specified region (as per file name)*	revenue.csv	Original
111	accommodates	The maximum capacity of the listing	revenue.csv	Original
112	minimum_nights	minimum number of night stay for the listing	revenue.csv	Original
113	maximum_nights	maximum number of night stay for the listing	revenue.csv	Original
114	number_of_reviews	The number of reviews the listing has	revenue.csv	Original
115	number_of_reviews_ltm	The number of reviews the listing has (in the last 12 months)	revenue.csv	Original
116	calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.	revenue.csv	Original
117	calculated_host_listings_count_entire_homes	The number of Entire home/apt listings the host has in the current scrape, in the city/region geography	revenue.csv	Original
118	calculated_host_listings_count_private_rooms	The number of Private room listings the host has in the current scrape, in the city/region geography	revenue.csv	Original
119	calculated_host_listings_count_shared_rooms	The number of Shared room listings the host has in the current scrape, in the city/region geography	revenue.csv	Original
120	reviews_per_month	The number of reviews the listing has over the lifetime of the listing	revenue.csv	Original
121	name_len	number of characters in listing name	revenue.csv	Engineered
122	neighborhood_overview_len	number of characters in the overview	revenue.csv	Engineered
123	host_verifications_len	number of verification methods (selfie, facebook, etc...)	revenue.csv	Engineered
124	desc_1	features extracted from the descriptions using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
125	desc_2	features extracted from the descriptions using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
126	desc_3	features extracted from the descriptions using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
127	desc_4	features extracted from the descriptions using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
128	desc_5	features extracted from the descriptions using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
129	n_1	features extracted from the neighborhood description using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
130	n_2	features extracted from the neighborhood description using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered

131	n_3	features extracted from the neighborhood description using Tf-Idf and SVD for dimensionality reduction	revenue.csv	Engineered
132	month	1-jan, 2-feb,...,12-dec	revenue.csv	Engineered
133	log_price_mean	mean price aggregated by month and listing_id from calendar data and log transformed	revenue.csv	Outcome
134	log_price_std	std aggregated by month and listing_id from calendar data and log transformed	revenue.csv	Outcome
135	predicted_price	log price predicted by xgboost regression model	revenue.csv	Outcome
136	target	based on availability_365 original column normalized to be in [0, 1] and categorized: "0": [0, 0.3), "1": [0.3, 0.7), "2": [0.7, 1]	revenue.csv	Outcome
137	availability_predicted	predicted availability by xgboost classification model	revenue.csv	Outcome
138	lower_t	lower threshold value for revenue calculation	revenue.csv	Outcome
139	upper_t	lower threshold value for revenue calculation	revenue.csv	Outcome
140	lower_revenue	lower revenue range based on predicted price and using boundaries from availability predicted column	revenue.csv	Outcome
141	upper_revenue	upper revenue range based on predicted price and using boundary from availability predicted column	revenue.csv	Outcome