Hawaii Airbnb Price Prediction

Xingrou Mei
Dept. of Computer & Information Science
Fordham University
New York
xmei4@fordham.edu

***Abstract*— The pandemic has affected the world economy sufficiently, especially in tourism. Tourism is one of the largest economic sources for Hawaii. Now that the pandemic is somewhat stable, it will be interesting to see how tourism is doing in Hawaii currently. This paper will develop a model for price prediction based on the features that hosts offer in Airbnb listings. The goal is to help Airbnb hosts to set their prices more effectively and understand the pricing market in Airbnb better with different aspects. The algorithms that this paper will use are Linear regression, Random Forest, and Neural Network. The metrics to evaluate the performance of the models are MAE, RMSE, and R-squared scores. Based on the result, it states that the Random Forest has the best result.**

Keywords—Airbnb, Price Prediction, Neural Network, Random Forest, Hawaii

I. Introduction

Many scholars use machine learning and deep learning for price prediction. Airbnb is also another popular topic besides house price prediction. Airbnb is known to provide short-term rental with people's personal properties in the beginning. They also start to expand their business, such as offering long-term stays, tourism activities, and luxury homes.

Airbnb, AirBedandBreakfast, is an online platform that offers homestays for vacation rentals and activities worldwide. However, all listings on the Airbnb website are not owned by Airbnb, and it is personal property from the "hosts," anyone who owns a property and wants to rent it out for the short-term or long term. Airbnb is established in 2008, based in San Francisco. There are over 5.6 million active listings worldwide with at least 100,000 cities. Airbnb went public in December 2020. Its direct economic impact in 2018 was based on host income and guest spending. Guests spent about 35 billion at local cafes and restaurants, and its total revenue was over 117 billion in economic impact. The United States has the highest impact, which is about 33.8 billion. Covid impacts the entire tourism industry, including Airbnb. In March 2020, Airbnb had a 30% loss revenue due to covid. However, hosts earned about 1 billion during the pandemic [1].

To conclude, this paper aims to develop a price prediction based on the characteristics of its property using the Machine Learning and Deep Learning models. It includes Linear Regression, Random Forest, and Neural Network. The metrics to determine the performance are Mean Absolute Error, Root Mean Squared Error and R-squared score.

II. Background

Hawaii is the 50th U.S. state. It is a ground of volcanic islands in the central Pacific Ocean. One fun fact, Hawaii is the only state in the US with two official languages, English and Hawaiian [2]. During the pandemic, Hawaii is strictly monitored. Before October 15, 2020, 14-day of quarantine was mandatory for all persons entering the State, including traveling between any of the islands in the State. Beginning October 15, pre-travel COVID-19 testing of a negative result option can be an alternative to the mandatory 14-day quarantine. The testing must be a State-approved testing facility and within 72 hours from the final leg of departure. In December, mandatory self-quarantine was reduced to 10 days. Starting April 2021, travelers that are fully vaccinated can be an exception to the self-quarantine in addition to a negative test result [3].

III. Experiments

A. Dataset

Data source: [4] http://insideairbnb.com/get-the-data.html

The dataset for this paper is multiple datasets in months combined. I merged all the datasets between May 2020 to April 2021 to create a large dataset for Hawaii Airbnb listings.

Before preprocessing:
Total instances: 267,673 | Total features: 108
After preprocessing:
Total instances: 263, 134 | Total features: 78

The dataset is split into two different ranges, low-price range from $11 to $999 per night, and high-price range, from $1,000 to $25,000 per night. After filtering the dataset into two, there are 252,524 rows in the low-price range dataset. There are 10,610 rows in the high price range dataset.

All features after data cleaning includes,

[id, host id, host is super host, host has profile pic, host identity verified, latitude, longitude, accommodates, bedrooms, beds, price, minimum nights, maximum nights, has availability, availability 30, availability 60, availability 90, availability 365, number of reviews, review scores rating, review scores cleanliness, review scores check in, review scores communication, review scores location, instant bookable, calculated host listings count, calculated host listings count entire homes, calculated host listings count private rooms, calculated host listings count shared rooms, security deposit, cleaning fee, bathroom, year, month, host year, host month,

total year, email, phone, has self checkin, has kitchen, has full kitchen, has kitchenette, has family/kidfriendly, has long term stays allowed, has free parking on premises, has pets allowed, has bedlinens, has coffeemaker, has cooking basics, has tv, has washer, has dryer, has dishwasher, has pool, has private pool, has shared pool, has hottub, has shared hottub, has private hottub, has air conditioning, has internet, has wifi, has carbon monoxide detector, has elevator, has safe, has balcony, has private entrance, has security system, has bbq grill, has gym, day of week, host response time, neighbourhood cleansed, neighbourhood group cleansed, property type, room type, listing date]. **Full list of features in descriptions in appendix*.

B. Data Preprocessing

*1) **Missing values***

List of all features with missing values in percentage:

security_deposit	78.101265
cleaning_fee	76.166442
bathrooms	74.292887
bathrooms_text	25.765766
review_scores_checkin	25.433645
review_scores_location	25.433645
review_scores_communication	25.421316
review_scores_cleanliness	25.416833
review_scores_rating	25.352576
bedrooms	10.382071
host_response_time	10.303990
host_acceptance_rate	5.221296
host_location	0.387413
host_identity_verified	0.266370
host_since	0.266370
host_has_profile_pic	0.266370
host_is_superhost	0.266370
beds	0.198376

*2) **Remove unwanted symbols***

There are numerical features with unwanted symbols and characters, and this makes them an object type. They must be removed and change to a float if there are decimals. For instance, features with a price have a dollar sign, $, and a comma. This includes price, cleaning_fee, and security_deposit. For property_type, there are characters that are not necessary, and they are all removed.

There are two columns related to the bathroom, 'bathrooms' and 'bathrooms_text.' The reason is that the datasets in 2020 only have data in 'bathrooms,' and the datasets in 2021 only have data in 'bathrooms_text.' Therefore, these features are combined into one column after removing unwanted characters. In addition, there are also decimal numbers instead of whole numbers. In the United States, bathrooms have different types: full bathroom, master, three-quarters, half, and quarter bathrooms. The full bathroom has a toilet, sink, shower, and bathtub/shower, and this will make it a 1. Half-bathroom represents a 0.5, three-quarter represents a 0.25, etc [4].

*3) **Fill Missing Values***

There are some missing values after the bathroom columns are combined, and the null values are removed. Features with less than 0.5 missing percentage, all the rows are dropped. These features are host_location, host_identity_verified, host_since, host_has_profile_pic, host_is_superhost, and beds. Features that can be filled with the mean values are below.

cleaning_fee
security_deposit
review_scores_checkin
review_scores_location
review_score_communication
review_scores_cleanliness
review_scores_rating

All null values for 'host_response_time' are filled with unknown. This feature is not numeric. Instead of filling the data with something unsure or removing a significant number of rows, 26688, it is better to fill it with unknown.

*4) **Binary Features***

All binary features with the letter 'F' or 'T' will change to 0 or 1, respectively. These features are 'instant_bookable', 'has_availability', 'host_has_profile_pic', 'host_is_superhost', and 'host_identity_verified.'

Features with a list in each row are extracted and changed to a binary feature as well. For instance, 'amenities' and 'host_verifications.' Each word in the lists is generated to a binary feature and this creates all the features with a 'has' in the front, such as 'has_pool' and 'has_kitchen.'

*5) **Features with Time***

'last_scaped' is the listing date of each listing. This feature is changed to the name 'listing_date.' The year and month from 'listing_date' and 'host_since' are extracted and created into a new column. This purpose is to see if there is any relationship with the price by month. Moreover, this dataset also added a new column with the number of years that the host has been an Airbnb host. This is calculated using today's year minus the year of 'host_since.'

*6) **Drop unwanted price***

The range for the price per night in column 'price' is from 0 to 94016. It does not make sense if the price is 0, and all the price with a 0 is removed. The highest price is researched manually, and this price is found to be the wrong price tag. Therefore, listings with this 94k per night are dropped. To check every listing with a high price randomly is not efficient. However, it is difficult to check which listing is wrong and which is not, especially when there are over 200k of data. All the incorrect price tags are spotted manually and with some simple common sense. For example, less than two accommodate, price per night is 5k to 10k, and property type is a private room. This clearly does not make sense. Upon researching on the Airbnb website, this listing is a 2-star hotel

with 2 double beds. Although this can be a monthly price, it is not confirmed, and it is better to drop it.

*7) **Label Encode and Scale***

There are some categorical variables, they need to be encoded to integer for better performance. After filling null values, dropping unwanted characters and data, creating new features, and encoding all categorical features. It is essential to normalize before working on the models. In this project, MinMaxScaler() is used to transform all features in the same scale from range 0 to 1.

C. Data Visualizations

Correlation with Price

Top Features:

bathroom	0.574410
bedrooms	0.473651
accommodates	0.434409
beds	0.384838
security_deposit	0.263124
property_type_en	0.246956
cleaning_fee	0.226904
has_balcony	0.185745
has_hottub	0.124763
has_securitysystem	0.105047
has_safe	0.103736
has_dishwasher	0.101900
host_has_profile_pic	-0.127941
number_of_reviews	-0.143531

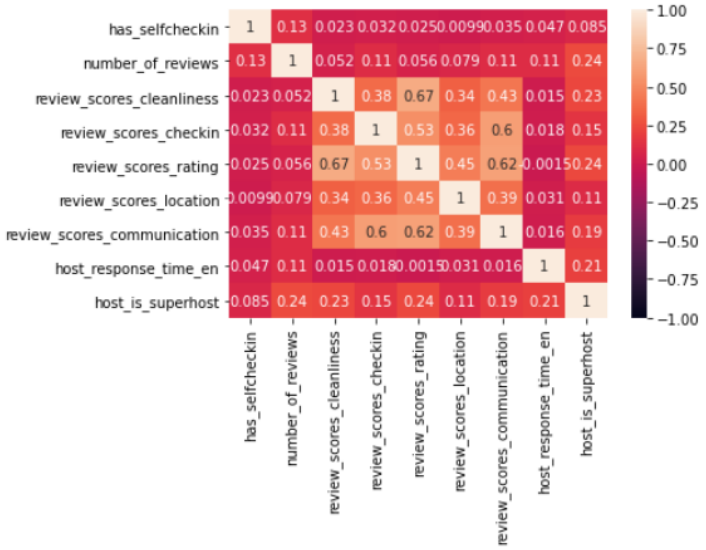
Features that are surprisingly does not have much correlation with price in this dataset:

has_privatepool	0.075319
neighbourhood_group_cleansed_en	0.075296
has_pool	0.070309
has_fullkitchen	0.062795
has_internet	0.060992
review_scores_rating	0.032067
has_wifi	0.017878
has_privateentrance	-0.024253
host_is_superhost	-0.036570
has_elevator	-0.040819
has_longtermstaysallowed	-0.068568
neighbourhood_cleansed_en	-0.085734

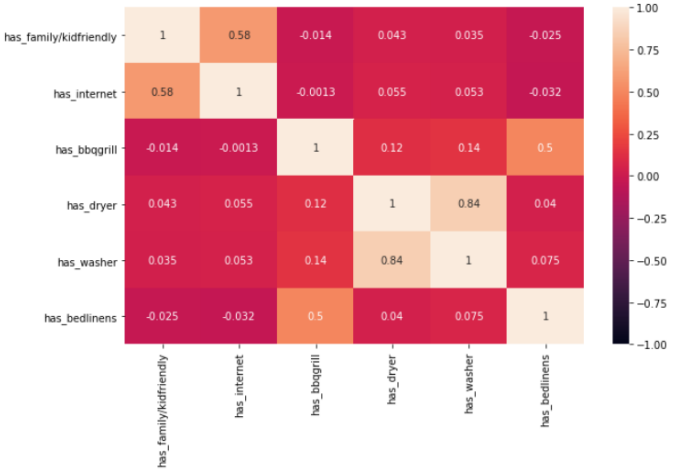
According to **Map 1**, The review score rating overall is 67% correlated with review scores in cleanliness, 62% correlated with review scores in communication, and 53% correlated with review scores in check-in. This determines that if a host has a better review score in cleanliness, communication, or check-in, then this impacts the overall rating by more than 50%. Host response time is only 1.6% correlated with review scores in communication. Therefore, the response time might not matter as much, but maybe the quality of communication matters the most.

Assuming self-check-in should correlate with review scores in check-in, but in fact, it is only 3.2% correlated. Both host response time and self-check-in correlate with the number of reviews, 11%, and 13%, respectively. For the amenities, **Map 2,** if a listing is family and kids friendly, most likely there is internet with a 58%. There is a 50% correlation between bbq grill and bed linen. With the highest correlation of 84% among all amenities, and they are dryers and washers. This is very straightforward, nothing surprising.

Map 1, correlation between reviews scores



Map 2, correlation with amenities



This paper uses data visualization to find patterns between features before working with the algorithms. Tableau is the tool that produces most of the visualizations for this paper.

It is essential to know when the price is high. Based on **Fig 1,** it shows that Wednesday is the most expensive day of the week. Tuesday and weekend are also high compared to other days. For the monthly chart, **Fig. 2**, the most expensive month is around December and July 2020. Since the state announced that pre-covid testing with a negative result is an alternative to the mandatory quarantine around October, the price also increased. Maybe there are more travelers around that time due to less restriction, and the host increased the price up a little each month. Especially around November and December are holiday

seasons. The price trend by month for all islands is similar. However, when it comes to the average price per night, Maui is more expensive, and Honolulu is less costly.

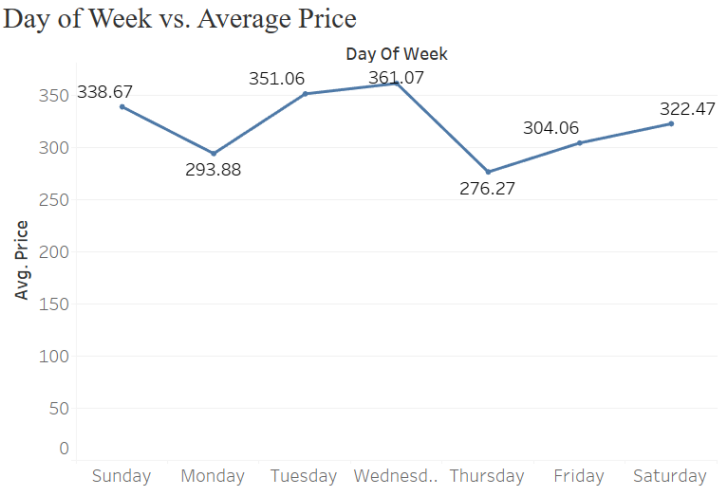


Fig 1, Price by day of the week.

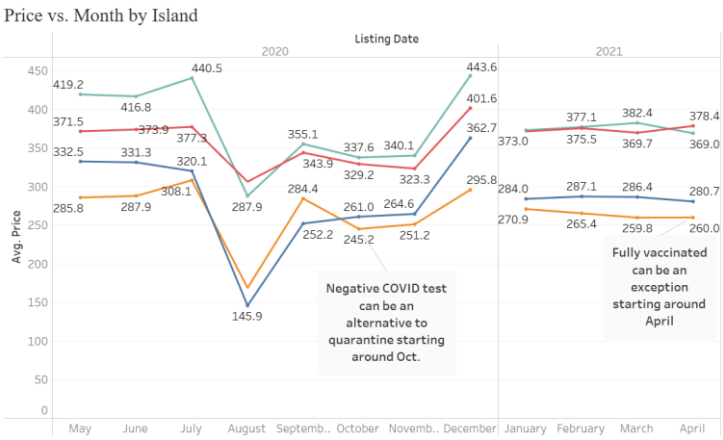


Fig 2, Price by month

As mentioned previously, Maui is more costly, on average, and the listing count on this island also has the most counts. Kauai does not have as much as the other island, but the average price per night is the second highest. **Fig. 3** is based on the entire dataset from range $11 to $25,000. Due to the wide range of prices, the price per night by average cannot give a better perspective on the high price range. In addition, **Fig 2**, shows that Honolulu is the cheapest island, but the average price in the high price range is higher than other islands in **Fig 4**. This only indicates that there are more expensive listings in Honolulu, but overall is still not the most expensive island.

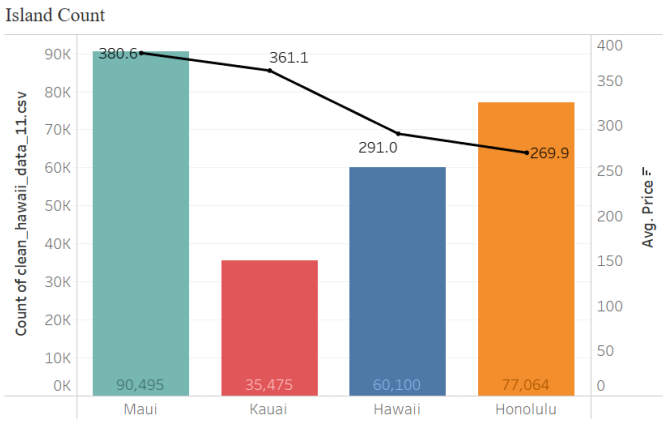


Fig. 3 Island count

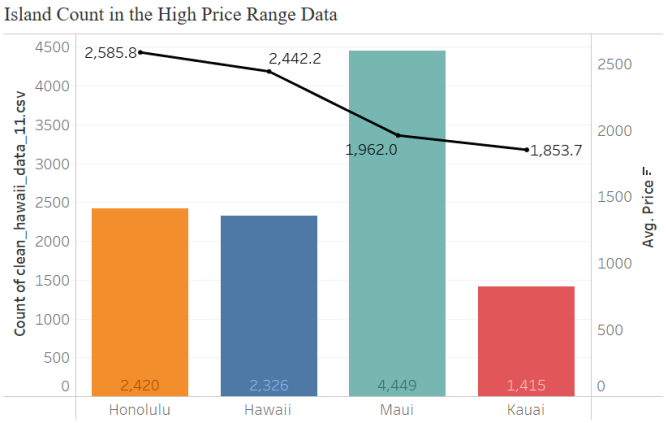


Fig. 4 Island count (high)

D. Algorithms

1) Linear Regression

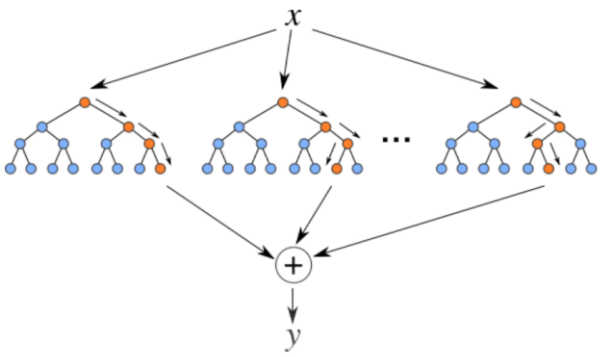
Linear Regression is basic and simple. It uses the relationship between two variables by fitting a linear equation. In this project, linear regression is a baseline model.

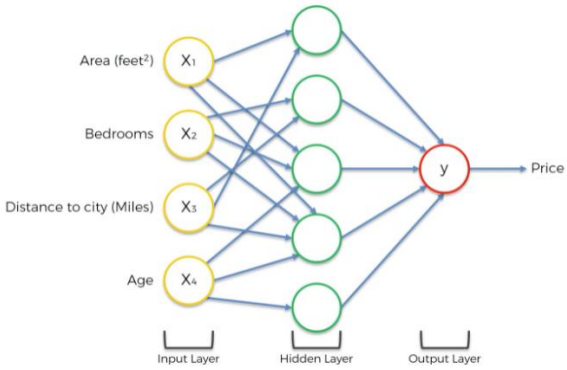
$$[\ Y(pred) = b0 + b1*x\ ]$$

2) Random Forest

Random Forest is a supervised learning algorithm, it can be used for regression and classification problems. It is an ensemble method on decision trees, meaning the decision is made with majority voting. When splitting an internal node, the best split is chosen randomly from a subset of features.

x = dataset
+ = averaged all predictions from all the trees
y = final decision



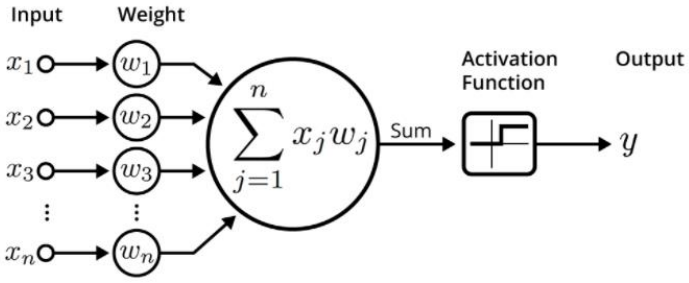


Feature importance

The theory behind the Random Forest model makes it easy to evaluate the importance of the feature. This project is using the python package Scikit-learn. The concept for Sklearn on feature importance is based on how much each feature contributes to decreasing the weighted impurity. Since this is based on a Random Forest model, it is averaging the impurity [5].

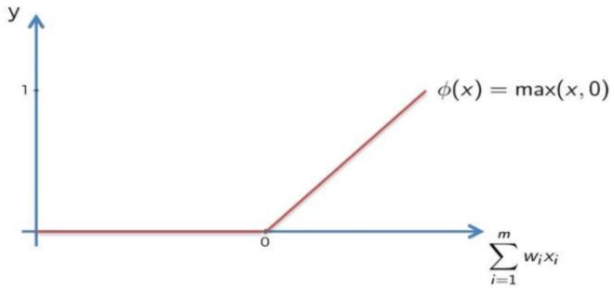
3) Neural Network

The neural network is a deep learning technique, and it is designed to mimic the human brain. The human brain consists of neurons that transmit, and process information received from the five senses. This is where the idea of the neural network comes from.The neural network is composed of layers of nodes, and each node is like a neuron in the brain. The first layer is the input layer, then the hidden layers, and then the output layer. Neurons or nodes receive input signals. The input is from the raw dataset or the previous layer. Once the signals are received, it calculates the sum of the multiplication with the corresponding weight, and then it passes through the activation function. Activation function can be, ReLU, Rectifier function, or Sigmoid functions. These two functions are commonly used in the Neural Network. Lastly, the output of the activation function is passed into the next layer, if there are more, or else that will be the output [6][7].



ReLU Activation Function

If the input value is negative, then the function will return 0. If it's a positive value, the function will output its input value.



Early Stopping

This technique is from Keras. This callback allows model to stop training process once it is not showing any improvements. This is a good method to avoid overfitting the dataset.

E. Metrics

Below are the metrics to evaluate the performance of the models.

1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}|P_i - \hat{P}_i|$$

2) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2}$$

3) R-Squared Score (R2 Score)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Mean Absolute Error measures the absolute average of the difference between the predicted value and actual value. MAE is more robust to outliers and does not penalize the errors as much as MSE and RMSE. This indicates that it is not suitable for datasets where outliers need to pay more attention.

Root Mean Square Error is the square root of the average of squared differences between the predicted value and actual value. It measures how far are the data points are to the regression line on average. [8] RMSE is a known metric for regression tasks. The value is squared before averaging, which gives a high penalty for large errors. It works better with outliers.

R2 score is to compare the current model with a constant baseline and identify how much the model is better. The constant baseline is taking the mean of the data and then draw a line of the mean. R2 is scale-free, which implies that it does not matter whether the values are too different in range, and this is because R2 will always be less than or equal to 1 [8].

IV. Results

This project implements three regression models, Linear Regression, Random Forest, and Neural Network. The dataset is split into two ranges before running on the entire dataset: low price range and high price range.

count	252524.000000	count	10610.000000	count	263134.000000
mean	246.532719	mean	2195.112260	mean	325.102683
std	176.003802	std	1941.635199	std	573.281960
min	11.000000	min	1000.000000	min	11.000000
25%	126.000000	25%	1199.000000	25%	129.000000
50%	195.000000	50%	1555.000000	50%	200.000000
75%	304.000000	75%	2450.000000	75%	332.000000
max	999.000000	max	25000.000000	max	25000.000000

*Fig 5. Left(low-price), Middle(high-price), Right(all price)

All the algorithms are performed by using Python and Sklearn packages. There are three metrics to evaluate the models, MAE, RMSE, and R-Squared. The testing data is 30% of the dataset, which the training data is 70%.

A. Result for LinearRegression

Before working on the complex models, Linear Regression is used as a guideline or baseline to see where the starting point is. First starts with the low-price range with Linear Regression. Based on **Table 1**, MAE is about $88 difference between the predicted value and the actual value. Now looking at the RMSE, the error is about $125. Since the 'outliers' in the entire dataset, and therefore, using the RMSE is a better metric to evaluate the model. Comparing the RMSE with the average, minimum, and maximum price will be easier to understand. **Fig. 5** shows that the mean price for the low-price range is about $246, and the RMSE is about $125 for both train and test sets. It shows that the model performs poorly. It is because linear regression is simple. Next, **Table 2**, the high price range is about $1,480 on the testing set. This error seems like a large number, but the mean price for this dataset is $2,195. If the price is higher, the RMSE is usually higher as well. It is still not a good result. Finally, **Table 3**, RMSE on the testing set for all the prices is $417, and the mean price is $325. Overall, linear regression does not perform that well with this dataset.

B. Result for Random Forest

For Random Forest in **Table 1**, RMSE is about $46, compared to the mean price of $246, it is a relatively good performance. The R-squared score is 0.93, which indicates that this model is 93% better than the constant baseline with the mean of the data. For the high price range, **Table 2**, RMSE is about $563. Besides looking at the mean price, compared to the RMSE in the linear regression, it is about $1000 less. It is clearly stating that Random Forest is a better model compared to the baseline. When Random Forest applies to the entire dataset, **Table 3**, RMSE is about $158. Overall, Random Forest performs well on all the datasets.

Features are essential when it comes to price prediction. Below is showing the top 5 features for low, high, and all the prices. Based on the result in **Fig 6-8**, the bathroom is the most important feature to price for all datasets. For price in the low range, accommodates is second, then latitude and longitude.

For the price in the high range, besides the bathroom, the second most important is a host with a profile picture, then its security deposit. The average price for the security deposit on the higher end is about $470 and on the low-price range is about $246. The maximum price on the high price range is $20,000, and the low-price range is $5000. Therefore, the security deposit is more important in the high price range than the low-price range. Besides the bathroom, latitude is also relevant in high, low, and all-price range. For the entire dataset, the second most important feature is the number of reviews. Property type also matters in the all-price range. It states that, in general, property type matters for all the listings in this dataset. On the low-price range, without some expensive listings, neighborhbor_group is relevant to price. Some island is more expensive than other on average, such as Maui.

*Fig 6, Top 5 features on low price

importance	feature
0.280896	bathroom
0.099093	accommodates
0.091736	latitude
0.077863	longitude
0.043106	neighbourhood_group_cleansed_en
0.040024	number_of_reviews

*Fig 7, Top 5 features on high price

importance	feature
0.241859	bathroom
0.055043	host_has_profile_pic
0.046226	security_deposit
0.042308	latitude
0.039996	minimum_nights
0.035639	id

*Fig 8, Top 5 features on the entire dataset

importance	feature
0.396502	bathroom
0.036873	number_of_reviews
0.035432	latitude
0.034055	longitude
0.032602	property_type_en
0.032080	host_id

C. *Result for Neural Network*

First, this model is used in 2 hidden layers, 77 nodes, 2 hidden layers, 1 output layer, the activation function is ReLU, 32 batch size, 100 epochs, and loss curve using Mean Squared Error. The result for the training set on the MSE is 1709729, RMSE at 1307, and R2 score at 0.55. This model is still decreasing on error. Later, the model trains with 200 epochs, and it looks like it still needs more training with an R2 score of 0.69. However, when the model is changed to 3 hidden layers, the R2 score is at 0.83 after 100 epochs. Using 3 hidden layers seems to get a better result with less training time. To avoid overfitting the dataset, this project is using a technique called EarlyStopping() from Tensorflow Keras with patience of 10. This means when the loss error stops to decrease and after 10 more epochs and still increasing, the model will stop. When this model is applied to the high price range, it stops at 143 epochs with an R2 score at 0.88 and 0.79 on the train and test set, **Table 1**. For low price range data, it stops at 95 epochs with R2 scores at 0.85 and 0.80 on train and test, **Table 2**. For the entire dataset, it stops at 89 epochs with an R2 score at 0.91 on the train set, and 0.84 on the test set, **Table 3**.
Below are all loss curves for the model with all the price range datasets.

Fig. 9, all price loss curve

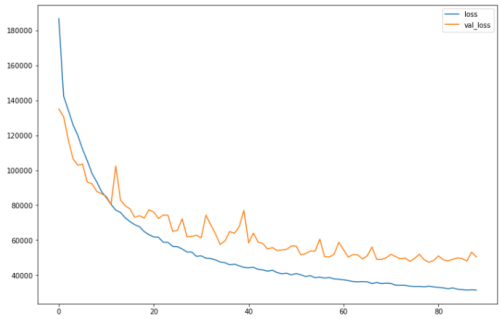


Fig. 10, high price loss curve

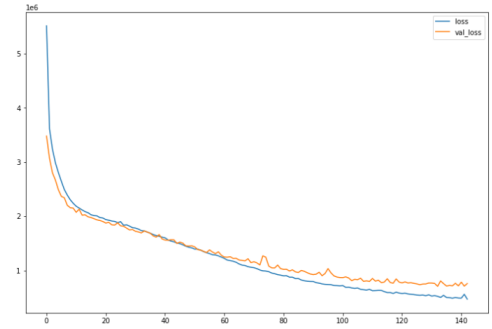
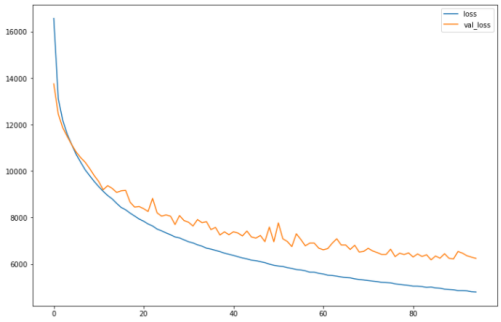


Fig. 11, low price loss curve



D. *Result, summary and metrics for all in charts*

Table 1	**Price Prediction on Low**			
	Algorithms	***MAE***	***RMSE***	***R2***
Train	Linear Regression	88.18	125.84	0.4903
Test	Linear Regression	88.19	125.76	0.4859
Train	Random Forest	9.04	12.21	0.9905
Test	Random Forest	24.23	46.28	0.9304
Train	Neural Network	47.27	68.18	0.8504
Test	Neural Network	52.71	79.00	0.7971

Table 2	**Price Prediction on High**			
	Algorithms	***MAE***	***RMSE***	***R2***
Train	Linear Regression	895.43	1501.12	0.4172
Test	Linear Regression	891.03	1484.73	0.3777
Train	Random Forest	121.98	259.52	0.9826
Test	Random Forest	298.15	563.32	0.9109
Train	Neural Network	488.04	696.28	0.8800
Test	Neural Network	577.22	871.82	0.7933

Table 3	**Price Prediction on all Price**			
	Algorithms	***MAE***	***RMSE***	***R2***
Train	Linear Regression	200.32	442.52	0.4209
Test	Linear Regression	199.15	417.20	0.4321
Train	Random Forest	15.07	60.37	0.9892
Test	Random Forest	40.30	158.64	0.9179
Train	Neural Network	89.87	178.87	0.9057
Test	Neural Network	96.76	224.51	0.8359

V. Related Works

There are many related papers on Airbnb price prediction. Previously, Luo, X. Zhou, and Y. Zhou [10] implemented machine learning approaches for price prediction for NYC, Paris, and Berlin Airbnb listings. It includes linear regression as a baseline, Random Forest, and Neural Network. The best approach is NN with r-squared values for various cities in the range of 0.716 to 0.773. In addition, Tang and Sangani [11] developed an SVM to predict Airbnb listings in San Francisco, its price, and its neighborhood. For the price prediction, they achieved a testing accuracy of 81.2 percent using all features. Besides Airbnb price prediction, Yu and Wu [12] implemented various regressions and classifications for real estate price prediction with given features. The models they used for both regression and classification including SVM and Random Forest. The best model is SVR with the gaussian kernel, RMSE of 0.5271, for regression and the best classification model is SVC with linear kernel, with an accuracy of 69 percent. Lastly, Ma, Zhang, Ihler, and Pan [13] estimated warehouse rental price with its features using algorithms including Linear regression, Regression Tree, and Random Forest. The result shows that price is closely related to its location in Beijing. In addition, the best model is Random Forest using RMSE and correlation coefficient as performance evaluation with correlation coefficient of 0.57. Liu [14] demonstrates price prediction using Linear Regression, Regression Tree and Gradient Boosting, Linear Model Ridge, Linear Lasso, SVM, Neural Network. Liu also adds customer reviews to train the models. Price prediction without using customer view produces the highest MAE, 0.609, with Linear Lasso model. Regression Tree with Gradient Boost output the least MAE, 0.287. Performances for all models are improved when putting customer review into count. Besides machine learning, using mathematical formulas can also determine which features are a good predictor for pricing. Wang D. and J. L. Nicolau [14] uses Airbnb datasets from 33 cities to identify price determinants using ordinary least squares and quantile regression analysis. They worked on 25 features in 5 categories, host attributes, site and property attributes, amenities and services, rental rules and number of online reviews and ratings. The ordinary least square shows that 24 out of 25 determinants are good predictors of price, and the Quantile regression analysis indicates that all the attributes are relevant.

VI. Conclusion

The best algorithm is Random Forest based on the metrics. Testing set using the high price range data, RMSE is $563.32, MAE is $298.15 difference between predicted and actual value, and R-squared is 0.91. Using the low price range data, RMSE is $46.28, MAE is $24.23, and R-square is 0.93. It performs very well using Random Forest for the high and low price ranges. For the entire dataset, RMSE is $158.64, MAE is $40.30, and R-squared is 0.92. The price range is broad, and there are many listings over $1000 per night, and it is considered low with a $158 in RMSE. Neural Network performs not as well as the Random Forest on the testing set. However, the training set in Neural Network and the testing set in Random Forest are similar in RMSE, with an error rate of $158 in Random Forest and $178 in Neural Network.

For Feature importance, there is only one feature in this dataset with more than a 0.50 correlation coefficient with the price. The most important feature is the bathroom with a 0.5744 correlation coefficient, and then bedrooms, accommodations, and beds. In the Random Forest Feature Importance, the bathroom also evaluates as an important feature. Although the rest of the features does not correlate with the price that much, but the top features are the number of reviews, property type, longitude, and latitude. Besides the correlation with price, review scores are correlated with each other. For instance, a host with a higher review score in cleanliness, communication, and check-in will eventually have a high score rating overall. There is very little correlation between price and review rating score, about 3%.

In terms of location, on average, Maui has the highest price than other islands. Honolulu has more listings over $10,000 per night. The neighborhood in Honolulu, called Primary Urban Center, has the most count of listings.

In conclusion, if host have properties in Maui, they can set the price higher than the properties in other island, unless they have very expensive villa, then the location does not matter. The location does not matter that much for the luxury homes. Host needs to keep in mind that bathroom, bedrooms, beds, and accommodates can be a factor and price determinants. According to feature importance, the number of reviews is somewhat important, and "host is a super host" is 24% correlates with the number of reviews [**Map 1**]. This is stating that, host might need to take this as a consideration. A super host may not be important in this dataset but being one can be a good method to have more power in setting the price higher.

References

[1] The zebra. 2021. "Airbnb statistics and host insights [2021]". [online] Available at: <https://www.thezebra.com/resources/home/airbnb-statistics/#:~:text=Airbnb%20has%205.6%20million%20active,guests%20have%20stayed%20at%20Airbnbs.>

[2] HISTORY. 2019. "Hawaii". [online] Available at: <https://www.history.com/topics/us-states/hawaii>

[3] Ige, D., 2021. COVID-19 Emergency Proclamations. [online] Governor.hawaii.gov. Available at: <https://governor.hawaii.gov/category/covid-19/covid-19-emergency-proclamations/>

[4] L. Wallender, "5 Types of Bathrooms," The Spruce. [Online]. Available at: <https://www.thespruce.com/types-of-bathrooms-4800093>

[5] Lewinson, E., 2021. Explaining Feature Importance by example of a Random Forest. [online] Medium. Available at: <https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e>

[6] Panchal, Shubham, 2018. 'Artificial Neural Networks — Mapping the Human Brain'. [online] Medium. Available at: <https://medium.com/predict/artificial-neural-networks-mapping-the-human-brain-2e0bd4a93160>

[7] McCullum, Nick. 2020. 'Deep Learning Neural Networks Explained in Plain English'. [online] Free code camp. Available at <https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/>

[8] Jj, "MAE and RMSE - Which Metric is Better?," Medium, 23-Mar-2016. [Online]. Available at: <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>

[9] Mishra, Divyanshu. 2019. 'Regression: An Explanation of Regression Metrics And What Can Go Wrong'. [online] Towards data science. Available at: <https://towardsdatascience.com/regression-an-

explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914>

[10] Yuanhang Luo, Xuanyu Zhou, Yulian Zhou, "Predicting Airbnb Listing Price Across Different Cities" Available at: <http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647491.pdf>

[11] Emily Tang and Kunal Sangani, "Neighborhood and Price Prediction for San Francisco Airbnb Listings" Available at: <http://cs229.stanford.edu/proj2015/236_report.pdf>

[12] Hujia Yu and Jiafu Wu, "Real Estate Price Prediction with Regression and Classification" Available at: <http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf>

[13] Y. Ma, Z. Zhang, A. Ihler, B. Pan, "Estimating Warehouse Rental Price using Machine Learning Techniques" Available at: <https://www.researchgate.net/publication/324512689_Estimating_Warehouse_Rental_Price_using_Machine_Learning_Techniques>

[14] Shravan Kuchkula, "Use Data Science to find your next Airbnb getaway" Available at: https://towardsdatascience.com/use-data-science-to-find-your-next-airbnb-getaway-3cb9c8333ad1

[15] Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com. International Journal of Hospitality Management, 62, 120-131.