AIRBNB PRICE PREDICTION USING MACHINE LEARNING

KIU MAY YEE

MASTER OF COMPUTER SCIENCE

MULTIMEDIA UNIVERSITY

August 2023

AIRBNB PRICE PREDICTION USING MACHINE LEARNING

BY

KIU MAY YEE

Msc.Computer Science, Multimedia University, Malaysia

THESIS SUBMITTED IN FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE

(by ODL)

in the

Faculty of Computer Science

MULTIMEDIA UNIVERSITY

MALAYSIA

August 2023

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institution of learning.

____

Kiu May Yee

ABSTRACT

Airbnb as a term for Air bed and breakfast, started the business back in year 2008, in San Francisco. With the rapid growth of the sharing economy, Airbnb has become a prominent platform for short term rental. Bangkok, as one of the most popular tourist destinations around the world, Airbnb business has experienced a substantial growth as well. Numbers of listings increased from 7000 listings (before year 2019) to more than 17,000 listings in year 2023. In this final year project, it aims to develop a machine learning model for predicting rental prices for Airbnb listings in Bangkok. Airbnb business has faced big challenges during novel coronavirus (COVID-19) pandemic in year 2020. Since COVID-19 is now become endemic, Airbnb has been given a strong demand that people want to travel and working from home in another home. To be able to accurately predict rental prices, this can give a proper guidance for both hosts and customers in making informed decisions. However, due to the diverse range of factors that influencing rental prices, the task is complex. Due to there is no available Airbnb listings data for Malaysia, yet, Bangkok and Malaysia share the similar holiday season, and weather. Therefore, price prediction model that is identified from Bangkok dataset will be applied to improve the dynamic pricing in Malaysia.

In this study, latest Bangkok Airbnb listings data (up to March 2023) was captured from InsideAirbnb.com, which contain a comprehensive set of features related to Airbnb rental price. Datasets originally consists of 75 fields, including a column of rental price. In this study, only 27 fields are kept for price prediction model training. The objective of the study is to identify the important features that will be able to help host to improve the listing and thus increase the profit. By training validate a dataset with several machine learning models and fivefold cross validation technique, performance of machine learning models is evaluated. Feature selection technique, Lasso regression is also implemented in linear regression and neural network model. This is to select important features only and exclude the features with 0 coefficient. This helps to improve fitness of prediction model. Performance of all the implemented machine learning models were compared and discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Airbnb business started in 2008, in San Francisco, after the United state housing bubble. In order to raise money for room rental, the two founders, Brian Chesky and Joe Gebbia decided to rent out air mattresses in their apartments to attendees of a conference as most of the hotels had been fully booked. Not only professional companies can provide lodging, but individuals are able to offer the unused spaces for short term rentals on Airbnb. It is a digital marketplace which connects property owners who have extra spaces in the house or apartments to the people who need short-term accommodation.

Unlike the traditional business to customer (B2C) business model, Airbnb is run as a travel accommodation platform, is more than a travel company and social networking company, it is a technology company. It provides a platform for house owners as a host to offer their property spaces to travellers for a short term or long-term stay. In this case, hosts and travellers are the both customers to the platform, and Airbnb runs in a disruptive innovation based, customer to customer business model. It provides a more personal and unique experience to travellers, with the feeling of living locally.

In today's world, people have increased the flexibility about when and where they want to travel than they ever have before. Due to the covid pandemic, more companies encourage their employees to work remotely, thus people are now spreading out to thousands of cities and towns. When the COVID-19 pandemic, in the year Q1 '2020, countries around the world closed their borders and restricted travel, no doubt the tourism industry was heavily impacted. Booking on flights and travel accommodation were reduced to a minimum level. Now Covid-19 is identified as endemic, people are more comfortable to travel in the year 2023, and the tourism industries are slowly rebounding. Things start to change for Airbnb too, people who want to work from home in another home and travel again after Covid-19 pandemic gives a strong demand on Airbnb short term rental.

Price determination on Airbnb rental property is an arduous task. A reasonable rental price can attract more customers to a property and provide maximum benefits to property owners. Technologies such as big data analysis and machine learning are implemented to provide recommendations to customers based on room type preference and pricing guidance, and fraud detection for property owners. By building a price prediction model for Airbnb, it is possible to provide property owners with a more accurate price setting based on the factors such as property location, property type, amenities included in the price and external factors such as holiday season, and supply in the market. This helps to improve a fair marketplace, and improves the occupancy rate and helps customers to find the best deal on Airbnb based on the specified budgets and preferences.

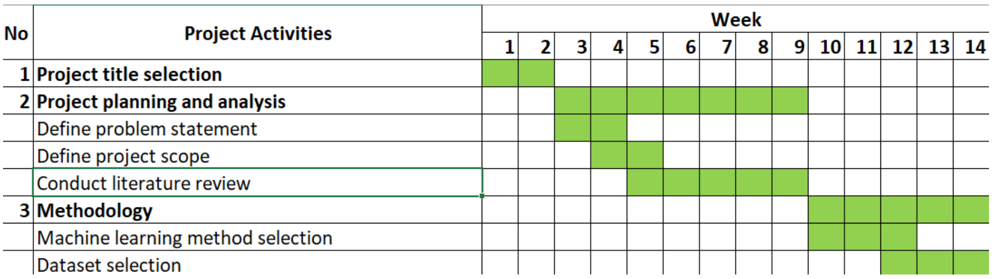
Three objectives are identified in this project.

1. to identify the set of important features for Airbnb pricing prediction.
2. to build a price prediction model using machine learning
3. to evaluate the performance of the predictive model.

Not all the features included in the dataset will have a strong correlation with the Airbnb price determination. Thus, in order for property owner and customer can make an informed decision in determining the price, it is important to identify the important features that are significant in Airbnb price prediction model.

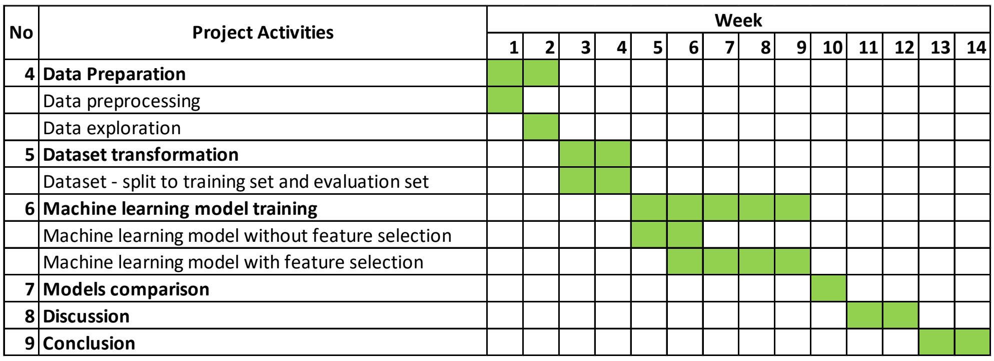
Scope of the project is identified to focus on Thailand Airbnb dataset. This is due to there is no available public dataset for Malaysia Airbnb listings. Thailand is one of the closest countries to Malaysia and both of the countries have the similar weather season and holiday patterns. Therefore, it is taken as a reference, which will be able to predict Airbnb price in Malaysia. Table 1 and table 2 show the Gantt chart for the Final year project 1 and project 2 for this project.

Table 1 Gantt chart for Final Year Project 1



No	Project Activities	Week													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	**Project title selection**	🟩	🟩												
2	**Project planning and analysis**			🟩	🟩	🟩	🟩	🟩	🟩	🟩					
	Define problem statement			🟩	🟩										
	Define project scope				🟩	🟩									
	Conduct literature review					🟩	🟩	🟩	🟩	🟩					
3	**Methodology**										🟩	🟩	🟩	🟩	🟩
	Machine learning method selection										🟩	🟩	🟩		
	Dataset selection												🟩	🟩	🟩

Table 2 Gantt chart for Final Year Project 2



No	Project Activities	Week													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
4	**Data Preparation**	🟩	🟩												
	Data preprocessing	🟩													
	Data exploration		🟩												
5	**Dataset transformation**			🟩	🟩										
	Dataset - split to training set and evaluation set			🟩	🟩										
6	**Machine learning model training**					🟩	🟩	🟩	🟩	🟩					
	Machine learning model without feature selection					🟩	🟩								
	Machine learning model with feature selection						🟩	🟩	🟩	🟩					
7	**Models comparison**										🟩				
8	**Discussion**											🟩	🟩		
9	**Conclusion**													🟩	🟩

CHAPTER 2: LITERATURE REVIEW

2.1 Price Prediction Model

With the various public datasets that are provided by Airbnb according to region, with archived data containing last 12 months listings data, many studies can be performed in the sharing economy industry. Price prediction for Airbnb by using machine learning techniques is one of the popular topics among researchers. From the research studies, various machine learning techniques have been adopted to predict prices for Airbnb short term rentals. Supervised techniques such as linear regression, support vector regression (SVR), random forest regression, artificial neural network (ANN), gradient boosting machine (GBM) and extreme gradient boosting (XGBoost) are widely studied.

Linear regression gives a linear model which finds the linear relationship between dependent and independent variables. Linear regression techniques are used in research study by (Y. Yang et al., 2016) to study the relationship between market accessibility and hotel prices in the Caribbean. A price prediction model was constructed based on a three-level mixed effect linear regression model and concluded the dependent variable -hotel prices are determined by the significant independent variables- level of market accessibility, services provided by hotel and online quality signalling factors.

SVR is another popular technique, which uses the support vector machine (SVM) to predict continuous variables. SVR fits the best line within a threshold value, by allowing tolerance in the model. SVR is recommended in study of real estate price prediction models by (Yu & Jiafu Wu, 2016) and Airbnb price prediction study by (Rezazadeh Kalehbasti et al., 2021)

Random forest regression uses ensemble learning methods by combining decision tree methods with regression and hence make more accurate predictions. (Wang & Nicolau, 2017) developed a model by using random forest regression and principal component analysis to predict Airbnb listings price in New York. In this study, the model was able to achieve an average absolute error of $20.49, which is lower than other baseline models.

GBM is another ensemble method which combines a decision tree. Shortcomings of the weak learners are identified through gradients in the loss function which indicates the fitness of the model's coefficients to the underlying data. In studies conducted by (Ye et al., 2018) on a pricing system consisting of three components is deployed at Airbnb to assist property owners to set an optimal price. A binary classification model is used to predict booking probability of each listing. Whereas a prediction model is used to obtain optimal price for each listing with a custom loss function. (Ahuja et al., 2021) concluded that by combining multiple features with the use of method Light GBM to predict Airbnb rental prices can help to improve accuracy of price prediction. Features identified in the studies including property characteristics, location and reviews by customers are impacting Airbnb rental prices.

XGBoost is another supervised machine learning method similar to GBM, with advanced regularisation to further improve generalisation capabilities. With XGBoost, a strong model is constructed from multiple weak models. XGBoost trains faster and can be performed parallelly across clusters for datasets containing mixtures of categorical and numerical values. (Trang et al., 2021) combined XGBoost with a clustering technique k-means to predict an optimal price suggestion for property listing on Airbnb. With the integration of clustering techniques and XGBoost, price prediction efficiency is highly improved in R2 and MSE value.

(Kalehbasti et al., 2019) performed a study on Airbnb price prediction in Istanbul and found that using XGBoost method, price can be predicted accurately in the Istanbul area as compared to Linear regression method. Researchers found that the prediction model was not generalizable to be used in predicting Airbnb rental in Amsterdam and Rome.

ANN is the heart of deep learning, which is inspired by the human brain. It is a series of algorithms which is used to recognise relationships in the dataset. Artificial neural network consists of a minimum of three layers, an input node layer, one or more hidden layers and an output layer. Figure 1 shows the typical neural network structure. Every node is connected to each other to feed data generated from linear regression to the activation functions. In general, ANN can be categorised into three categories, feedforward neural networks, convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Feedforward neural network, also called deep neural network (DNN), is the foundation for natural language processing (NLP), CNNs and RNNs. CNNs are widely applied in image and pattern processing data, and computer vision related problems. RNNs work with sequential data, and are widely applied to speech recognition, machine translation and language modelling and generating text.

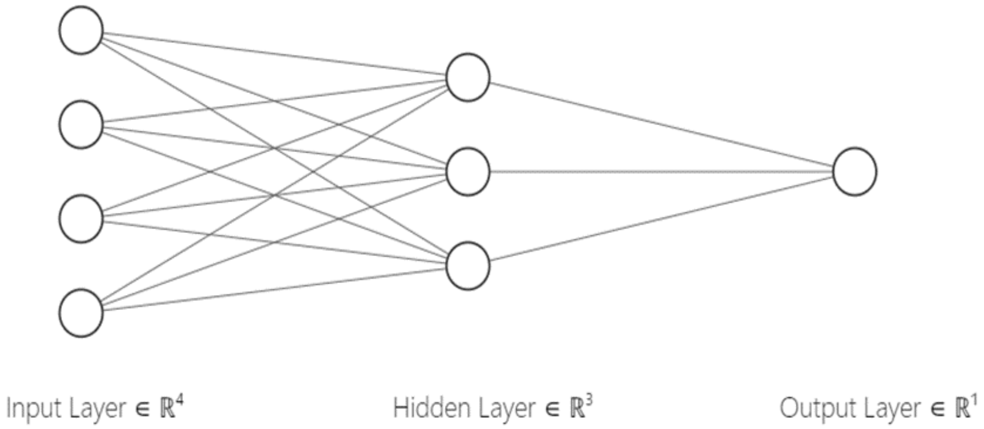


Figure 1 Artificial Neural Network Diagram

(Peng et al., 2020) uses NLP to extract textual comments reviewed by customers, and developed Airbnb price prediction model 4 types of machine learning models, Linear regression, SVR, XGBoost and DNN. Pricing models built with multi-modality data increase prediction efficiency. (Lewis, 2019) compared the Airbnb price prediction model which is built by XGBoost and DNN respectively by using Airbnb London listings. (Thakur et al., 2022) built a DNN with L1 Regularization to predict Airbnb price in Rio de Janeiro and suggested the property price can be set higher during warm season and room type of having at least a private bedroom with two separate bed and a bathroom is popular as most of the customers travel with family.

Bagging, an ensemble learning technique, which is also known as bootstrap aggregating. It is used in machine learning to improve the model predictive accuracy and stability of model. By creating multiple versions of a base model, using is a decision tree, using different subsets of the training data, the final prediction is determined by averaging the predictions from all the versions of regression models. Empirical and theoretical findings suggest that bagging can substantially improve the performance of a weak but unstable algorithm by moving them closer to the optimal results. (Otero Gomez et al., 2020) has compared bagging regression and LightBGM to forecast housing price in Colombia and obtained a RMSE value at 0.24. (Pandey et al., 2019) has also implemented bagging regression in forecast future gold and diamond price. From the study, bagging regressor model is shown to have produced model with highest accuracy at 0.9658.

2.2 Feature selection

In machine learning, feature selection is used to select the critical input variables by using only relevant data and eliminating noise in the dataset. By selecting most relevant and most critical features to the price of listings, it helps to increase the prediction power of the of the selected algorithms. There are various types of features, for example number of bedrooms, beds, bathrooms, location of property, customer reviews, and etc. Some of the features can be more important than the other features, and shows a greater correlation between the features and the listing price. Sub setting the redundant features can reduce the generalisation of the model, which is described as over-fitting, the price prediction model can be more robust and more interpretable.

(S. Yang, 2021) predicts Airbnb price in Beijing by using XGBoost method, and NN with feature importance selection. Feature selection is performed at data pre-processing stage by removing the not useful data column in the dataset. Subsequently, data cleaning is performed to convert the column into useful information, for example on date column and fees columns are filtered and categorized to provide better category.

Various studies have been conducted on the suitable techniques that can be used to select the most significant features in predicting listing price. Supervised machine learning techniques can be applied for feature selection in Airbnb price prediction models. supervised machine learning algorithms use labelled training data to learn a mapping between input features and output labels. Algorithms such as random forest and GBM have built-in methods to select critical features based on the relevance to the output label. Besides that, external techniques are also applied to select critical features for the price prediction model. Lasso regression, a widely used technique, is effective when there are many features that may be irrelevant or redundant, as it can select a subset of the critical features while shrinking the coefficient of the others to zero. It is adopted in a study conducted by (Masrom et al., 2022) to predict the Airbnb price listing. Researcher (Liu, 2021) uses Lasso regression by implementing L1 penalty and ridge regression as L2 penalty in feature identification. It is found that room type is the entire room/apartment and contains private rooms is the most popular preference in the New York area. XGBoost model is selected as the most efficient model with R2 score of 0.6321, and identifies statistically significant features in the price prediction model.

Another popular feature selection technique, Correlation based feature selection (CFS) is a filter technique, which identifies features by sub-setting the features with the highest correlation to the output by calculating the correlation between each input variable to the output label. A feature subset is considered as good if it can predict categories in the instance space that have not been predicted by other features. It is recommended to be used with datasets where the relationship between input variables and output label is not complex or in linear relationship.

(Kirkos, 2021) applied CFS to select more significant features that influenced the Airbnb listing price. Study concluded that the host plays a central role in price prediction. A property owner who quickly responds to a customer's request and a listing with more details on property were selected as significant features that influence the Airbnb listing price.

2.3 Summary

In summary, there are various studies implement machine learning methods in predicting Airbnb price in various locations. However, some of the features are redundant and not relevant for price prediction. These features are either redundant or irrelevant features which do not contribute to the Airbnb price prediction. Building a model with features that are redundant and irrelevant could impact the performance of model by introducing noise and reduce the model's generalisability.

To address this problem, this research will apply feature selection techniques to identify the most informative and relevant features that are strongly correlated to predict Airbnb price. By selecting a subset of strongly relevant features with the Airbnb price, then these features can be used to train machine learning model, including linear regression, SVM, or NN. With the application of feature selection, the machine learning model's performance can be enhanced, and generalizability can be improved.

CHAPTER 3 METHODOLOGY

3.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM is a de facto standard released in 2000 and a process model for data mining projects. CRISP-DM is an industry independent process model, from understanding customer's requirements until software deployment. The initial purpose of this methodology is to develop modelling techniques for industry. Based on the user guide of CRISP-DM, there are six steps in the process model, shown in Figure 2. Each step consists of second level steps in order to apply into all data mining applications.

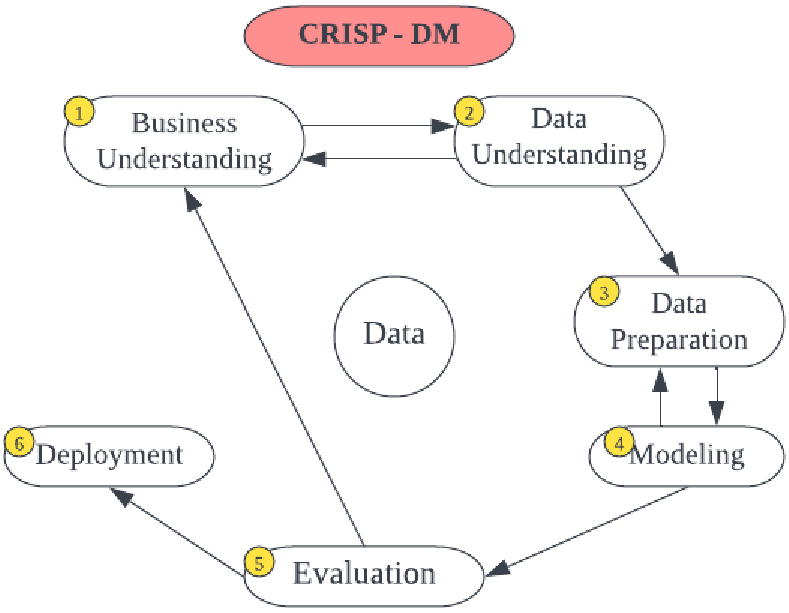


Figure 2 CRISP-DM Process Model Workflow

Business understanding is the foundation to initiate a project. This phase focuses on understanding the business objectives, the goal that the customer wants to achieve. Then define business success criteria. Sub-tasks of phase 1:

1. Assess situation: identify resources availability, business requirements of a project, perform risk assessment, identify contingency planning, and conduct cost-benefit analysis.

2. determine the data mining goals: define success criteria for technical data mining perspective.
3. Produce product plan: use tools and technologies that are suitable for the identified objectives, with the available resources. Define detailed plans for every project phase.

In the business understanding phase, problem statement of this project, project objectives and project scope are defined in Chapter 1. Literature review was conducted to identify research gaps, which are presented in Chapter 2.

Data understanding is the process of collecting data from data sources, exploring the relevant data and analysing the datasets that can help to accomplish the project goals. Sub-tasks of phase

1. Collect initial data: obtain and analyse data from data sources. In this study, datasets are acquired from Inside Airbnb.com.
2. Describe data: verify the data format, number of records, and field information.
3. Explore data: query and analyse the datasets. Explore the relevant data from Airbnb to gain insights into the features that are most relevant to predict rental prices.
4. Verify data quality: documents any quality issue in the data.

In this second phase, the Thailand Airbnb listing dataset is extracted from Inside Airbnb. Data exploration is performed to verify the quality is meeting the project goals. Further explanation is to be discussed in Chapter 3.

Data preparation is the critical part in the data mining project. Final datasets for the price prediction model are prepared. Cleaning, transforming the dataset into a format suitable for machine learning and selecting features from the data to prepare it for machine learning. Sub-tasks of phase 3:

1. Select data: identify useful datasets and include the reason for data inclusion/exclusion.
2. Clean data: correct, impute or remove erroneous values in the dataset. Example of data with missing values, encoding categorical variables, and normalising numerical feature
3. Construct data: derive new meaningful attributes from obtained datasets.
4. Integrate data: create new datasets by combining original data and the new attributes constructed in the previous step.
5. Format data: ensure the data has correct format in order to perform calculation and data visualization.

In this project, Thailand Airbnb dataset will be cleaned to remove the not useful data, transformed into appropriate format and construct new attributes. This ensures all the data columns have the correct format in order to perform calculation and price prediction.

Modelling is a phase to choose an appropriate machine learning algorithm and fitting it to the data obtained from the previous phase. This phase can be an iterative step in order to obtain the best model. Sub-tasks of phase 4:

1. Identify modelling techniques. The appropriate machine learning algorithm depends on the business problem statement and the available datasets.
2. Building test case.
3. Build model.
4. Assess model. Evaluate and compare multiple models against the evaluation criteria, such as pre-defined success criteria, domain knowledge and the test design.

In this project, various machine learning algorithms will be modelled and compared for Airbnb price prediction model

Evaluation is the next phase to assess the model in terms of meeting the business requirements. Sub-tasks in phase 5:

1. Evaluate results based on identified business success criteria. Fine tune the model if necessary.
2. Review work progress, and summarize observations.
3. Determine next steps based on the summary whether it should be proceeded to deployment, or re-iterate on the phase 4 modelling or initiate a new project.

Deployment: deployed the trained and validated model to predict the rental prices of new Airbnb listings. In this Airbnb price prediction model project, a prototype system will be the deliverable, this deployment phase is beyond the scope of this project.

3.2 Feature Analysis and Selection

Feature analysis also known as feature exploration. It is a crucial step in machine learning. Through the analysis, we can gain insights into how each feature relates to the target variable. Ultimately to guide feature selection, and data preprocessing decisions.

In this study, individual feature is first analysed to examine summary statistics. Histogram and box plots are used to examine distributions of the features. Besides that, heat map analysis and SelectKBest are used to investigate correlation between each feature and target variable. Heat map as a graphical representation of the correlation matrix between numerical variable in the dataset, with the correlation values, range from -1 to 1. Positive value indicates a positive correlation, negative values indicate negative correlation (as feature's value increase, the target variable will decrease), and value close to 0 indicate weak or no correlation.

SelectKBest as a feature selection technique is a technique to rank input features based on the relationship with the target variable. Relationship of input features with the target variable is shown by a statistical score for each feature. In this study, scoring function of ‘mutual_info_regression’ is used, as it can captures the relationship between input features and target variable.

3.3 Cross Validation

5-fold cross validation technique is used in this project to assess the performance of machine learning model, and to estimate how well it can perform on unseen data. In 5-fold cross validation, dataset is split to five subsets of data. The cross-validation process involves training and evaluating the model five times, by using different subsets as the training-testing set in each iteration. During each iteration, four of the subsets are used as the training set, and the remaining one set is used as the testing set.

Machine learning models that are applied in this project include linear regression, neural network, bagging regression, gradient boosting regression, XGBoost regression, random forest regression and decision tree regression.

3.3.1 Linear Regression

In this price prediction model, there is more than one independent variable are used, therefore multiple linear regression which can consider multiple independent variables is considered. Multiple linear regression expands upon the principles of simple linear regression to encompass multiple independent variables. The goal is to find the best fitting linear relationship between independent variables to the target variables. In mathematical terms, the formula for multiple linear regression is as follows:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn + \varepsilon \tag{1}$$

where y is the target variable, price in this study,

β0 is the intercept of the model,

β1, β2, … βn are the coefficients for the respective independent variables *x1, x2,…xn*,

x1, x2,…xn are the independent variables that influence the target variable.

ε represents the error term or residuals, which captures the difference between the predicted price value and the actual observed value.

The equation describes a hyperplane in an *n*-dimensional space, where each input features contributes to the result by certain scoring which is determined by its coefficient. The model learns these coefficients from the training data, and allows it to make predictions on new and unseen data.

By finding the coefficients for independent variables that minimize the sum of squared differences between the predicted values and the actual values. This process involves of fitting the model to training dataset by estimating the coefficients that gives the best fitting line.

In this study, a module ‘sklearn.linear_model’ within ‘scikit learn’ library in Python is used. It includes classes and functions for performing linear regression, lasso regression, logistic regression, and other linear model variations. To use the linear regression from scikit learn, it will need to import LinearRegression from sklearn.linear_model. After import module, linear regression model can be created by specifying parameters value. Default parameter is used in this study.

Below is the coding that is used to create Linear Regression model in Jupyter notebook:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression (fit_intercept =True, copy_X =True, n_jobs
=None, positive =False).
```

3.3.2 Neural Network Model

Neural network is a machine learning model which is inspired by the structure and functioning of the human brain. It consists of various layers of interconnected nodes (neurons) that process and transform data. In this study, where price prediction is a regression task, the goal is to create a neural network model that can predict continuous numerical value – Airbnb price based on multiple input features from the training dataset.

As the architecture of a neural network includes an input layer, one or more hidden layers, and an output layer. The number of independent variables that is used in the training is considered as the number of neurons in the input layer. For the output layer consists of a single neuron, which represents the predicted Airbnb price. There can be multiple hidden layers, with varying number of neurons. Deeper networks might capture more complex patterns, which can also lead to overfitting if not properly regularized. Activation functions can be applied to the output of each neuron to introduce nonlinearity and capture complex relationships in the data. For output layer in regression, there is no activation function is often used.

In this study, Multi-Layer Perceptron (MLP) Regressor, which is a type of neural network designed for regression task in used. In Python environment, a class 'MLPRegressor' from scikit learns 'neural_network' module is imported. This class allows us to create and train a multiple layer's perceptron model that is specifically regression tasks. Below is the code which is used in the study to generate a neural network model. In this study, default hidden layer sizes are followed, which the training model consist of single hidden layer with 100 neurons, single layer perceptron and one output layer.

```
from sklearn.neural_network import MLPRegressor
model=MLPRegressor(random_state=1, max_iter=500)
```

In this study, parameter used for MLP Regressor is set as default, and the max iteration is set at 500, parameters can be adjusted to create a network with different architectures, depending on the problem complexity and requirements. Default hidden layer size is (100,1), which is 100 neurons in single layer. Choice of hidden layer sizes can significantly impact the model's performance.

3.3.3 Bagging Regression Model

Bagging also known as bootstrap aggregating, is an ensemble learning technique. Bagging starts by creating multiple bootstrap samples from the original training dataset. Each bootstrap sample is obtained by randomly selecting data points from the training data with replacement. These bootstrapped samples as used to train a separate instance of base model. After the training, each base model able to predict the new target variable for the new data points. The final ensemble prediction is finalized by averaging the combined prediction values from all the available base models. Formula of bagging regression task is represented as follows:

$$Ensemble\ Prediction = \frac{PredictionModel_{1}(x) + PredictionModel_{2}(x) + \cdots + PredictionModel_{\boldsymbol{B}}(x)}{\boldsymbol{B}}$$

where *Prediction_Model*$_i$*(x)* represents the prediction of the *i*-th base model for the input X, and **B** is the number of base models in the ensemble.

To apply bagging regression model in Python, class ‘BaggingRegressor’ from ‘sklearn.ensemble’ can be used. This class provides an implementation of the bagging ensemble technique that specifically designed for regression tasks. Default base estimator for bagging regressor is decision tree regressor.

```
from sklearn.svm import SVR
from sklearn.ensemble import BaggingRegressor
model = BaggingRegressor(estimator=SVR(),n_estimators=10,
random_state=42).fit(X, y)
```

In this study, selected base regressor is SVR, and the default number of base model is used, which is set at 10. 10 base models are created, and final ensemble prediction for Airbnb listing price is the average value of the individual prediction value on the Airbnb listing price.

3.3.4 Gradient Boosting Regression Model

Gradient boosting regression model is a machine learning technique for regression tasks. This is an ensemble learning method that is derived from a decision tree. Decision trees uses a hierarchical structure, beginning at the tree’s root, branching based on conditions, and heading towards the leaves, which is the predicted target variable value. Gradient boosting regression model is used to mitigate the risk of overfitting, it combines multiple weak learners to create a strong predictive model.

The fundamental concept of gradient boosting regression model is a process of iterative refinement, where a sequence of base models is meticulously constructed to rectify the weakness and inaccuracies in the previous base models. Consequently, in each iteration, the ensemble can refine the predictive power and able to adapt to complex patterns present within the dataset. The formula for the prediction in Gradient Boosting Regression can be represented as follows:

$$y_{pred} = \sum learning\ rate\ x\ base\ model_i(x) \tag{3}$$

where y_{pred} is the final predicted value, base model$_i$(x) is the prediction of the i-th base model for the input x. and learning rate is a hyperparameter that controls the contribution of each base model's prediction.

To apply gradient boosting regression model in Python environment, class GradientBoostingRegressor from 'sklearn.ensemble' module can be used. In this study, all the parameters in the regressor are set as default.

```
from sklearn.ensemble import GradientBoostingRegressor
model=GradientBoostingRegressor(random_state=0)
#to identify influential features
importance = model.feature_importances_
```

Besides that, for this gradient boosting regression model, it has an implicit feature selection attribute, where we can use this to identify the importance scores for each independent variable, and understand which features contribute the most to the model's predictive performance and overall behaviour. Feature analysis can be performed by study and visualize the scoring in the model.

3.3.5 Decision Tree Regression Model

A decision tree regression model is a predictive modelling technique that uses tree-like structure. In this case study, Airbnb price prediction is done based on a set of decision rules inferred from the training data. The process starts with the entire dataset, and the tree is built through recursively splitting the data into subsets based on the values of each independent variables. At each node, the algorithm evaluates different independent variables and their potential thresholds to find the best split that can reduce the variance in the target variable values. The process continues until a stopping criterion such as maximum depth or minimum number of samples per leaf is reached. Finally, the predicted value is the average of the target variable values in each leaf node. Formula of the decision tree regression model can be represented as follows:

$$y_{pred} = \sum average(y\ in\ Leaf_i) \tag{3}$$

where y_{pred} is the predicted target variable value, $Leaf_i$ represents each terminal node in the tree. In Python code, class DecisionTreeRegressor from sklearn.tree can be used to train the prediction model.

Similar to gradient boosting regression model, decision tree regression model has an implicit feature importance attribute, that can be used to study the influential features in the data modelling. Feature analysis can be performed to identify the importance features considered in the training model.

```
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor(random_state=42)
#to identify influential features
importance = model.feature_importances_
```

3.3.6 Random Forest Regression Model

Random forest regression model is an ensemble learning technique that combines the prediction power of multiple decision trees to create a robust and accurate predictive model. It can address the limitations of individual decision trees by reducing the overfitting and increase the prediction accuracy. The process starts by creating multiple bootstrap samples from the original training dataset. Besides, random forest also selects a subset of features for each tree. This diversifies among the trees and helps capture different patterns in the data. Next, a decision tree is constructed for each bootstrap sample and feature subset. The final prediction for the target variable is obtained by averaging predictions outcome from all the independent decision tree. This ensemble approach reduces variances and improves prediction accuracy.

Formula of random forest regression model can be represented as follows:

$$y_{pred} = \frac{1}{N} \times \sum predicted\ value\ of\ tree_i(x) \tag{5}$$

where y_{pred} is the predicted target variable value, N is the number of trees in the random forest. In Python environment, class RandomForestRegressor from sklearn.ensemble can be used .In this study, no changes on the default parameter in the class.

```
from sklearn import ensemble
model=ensemble.RandomForestRegressor(random_state= 42,n_estimators=100)
#to identify influential features
importance = model.feature_importances_
```

3.3.7 XGBoost Regression Model

XGBoost is built upon the principles of gradient boosting and introduces several enhancements to improve both accuracy and efficiency. XGBoost follows the gradient boosting framework, where an ensemble of weak base models is iteratively constructed to correct the errors of the previous base models. Objective function of XGBoost combines loss function that quantifying the error and L1 and L2 regularization terms that prevent overfitting. The formula of XGBoost regression model can be expressed as follows:

$$y_{pred} = \sum learning\ rate\ \times base\ model_i(x) \tag{6}$$

where y_{pred} is prediction for target variable value, base model$_i$(x) is the prediction of the i-th base model for the input 'x' and learning rate is a hyperparameter that controls the contribution of each base model's prediction. The predicted value is the sum of the predictions from each base model, weighted by learning rate.

In Python environment, class XGBRegressor from xgboost module is used to train a Airbnb price prediction model. By default, L1 and L2 regularization term is set to 0, there is no fine tune being performed in this case. Similarly, there is a built-in feature selection function, which is stored in attribute feature_importances, and it can be used to analyse the feature scoring to understand the impact of each individual independent variable in the prediction model.

```
from xgboost import XGBRegressor
model = XGBRegressor(objective='reg:squarederror')
#to identify influential features
importance = model.feature_importances_
```

3.4 Model Evaluation

In this project, various metrics will be used to evaluate the model that is trained. In particular, R^2, mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE).

3.4.1 Mean Squared Error (MSE)

A risk function corresponding to the expected value of the squared error loss. It measures how close a regression line is to a set of data points, by calculate the mean of error squared from data as it relates to a function. A larger MSE indicates that more data points are widely dispersed around the mean. A smaller MSE is preferred because it indicates that the data points are focus around the mean, has less errors and it is not skewed. Equation 1 shows the formula to evaluate MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{1}$$

where y_i is the observed values, and p_i is the predicted values.

3.4.2 Mean absolute error (MAE)

A measure that is used to evaluate the performance of regression models. It measures the mean absolute difference between an observed data sets and the predicted values. MAE is calculated with the formula as

$$MSE = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}|y_i - p_i| \tag{2}$$

where y_i is the observed values, and p_i is the predicted values

3.4.3 Root Mean Squared Error (RMSE)

RMSE is a weighted measure of model accuracy by calculate the average differences between a statistical model's predicted values and the observed dataset values. It is the standard deviation of the prediction errors. A value of 0 indicates less error produced in the model predictions. The formula to calculate RMSE is:

$$RMSE = \sqrt{\frac{\sum(y_i - p_i)^2}{N - P}} \tag{3}$$

where *N* is the number of observed dataset values, and *P* is the number of parameter estimates.

3.4.4 R-Squared (R2)

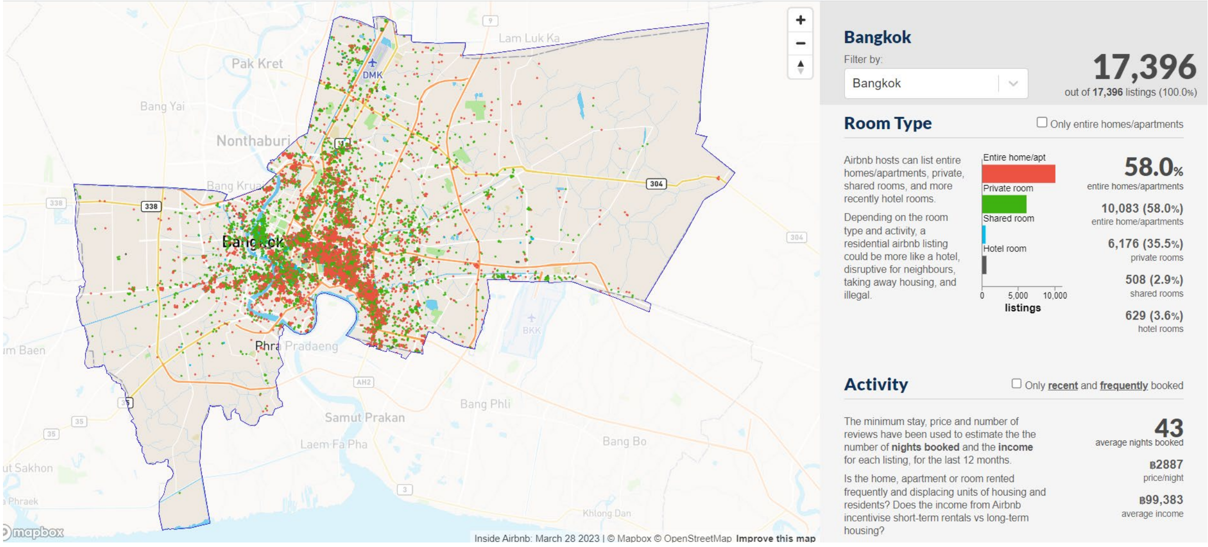
An R-Squared measure the goodness-of-fit of fitted regression line to the scatter of the actual data. It is referred as the coefficient of determination, and it has a value range between 0 and 1. 0 indicates the model does not explain the variability of the data around its mean, and 1 indicates that 100% of the data correspond to a model around its mean. In general, model with R2 less than 0.30 is considered a weak model, a model with R2 in between 0.30 to 0.50 is considered a moderate model. If a model that has R2 value more than 0.70, indicates that the independent variables give a strong effect in predicting the target variable. The formula is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - p_i)^2}{\sum_{i=1}^{n}\left(y_i - \frac{\sum y}{N}\right)^2} \tag{4}$$

where *N* is the number of observed dataset values, y_i is the observed values, and p_i is the predicted values.

CHAPTER 4: DATA PREPROCESSING AND ANALYSIS

Airbnb open dataset from Inside Airbnb that consists of general information of property listing and reviews from customers. In this study, the most recent dataset (March 2023) which listed the Airbnb property listing in Bangkok, Thailand is chosen. Figure 3 shows the cumulative listings in Bangkok as of March 2023 as 17,936 accommodations are listed on Airbnb.



Bangkok

Filter by: Bangkok

17,396

out of 17,396 listings (100.0%)

Room Type

□ Only entire homes/apartments

Airbnb hosts can list entire homes/apartments, private, shared rooms, and more recently hotel rooms.

Depending on the room type and activity, a residential airbnb listing could be more like a hotel, disruptive for neighbours, taking away housing, and illegal.

58.0% entire homes/apartments

10,083 (58.0%) entire home/apartments

6,176 (35.5%) private rooms

508 (2.9%) shared rooms

629 (3.6%) hotel rooms

Activity

□ Only recent and frequently booked

The minimum stay, price and number of reviews have been used to estimate the the number of **nights booked** and the **income** for each listing, for the last 12 months.

Is the home, apartment or room rented frequently and displacing units of housing and residents? Does the income from Airbnb incentivise short-term rentals vs long-term housing?

43 average nights booked

฿2887 price/night

฿99,383 average income

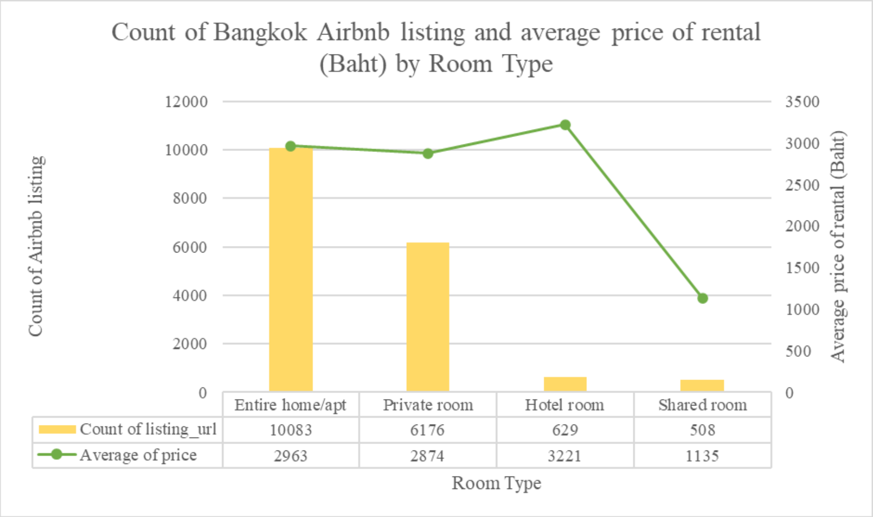
Inside Airbnb: March 28 2023 | © Mapbox © OpenStreetMap Improve this map

Figure 3 Airbnb listing in Bangkok, Thailand (captured from InsideAirbnb.com as of March 2023)

4.1 Data Exploration

In general, there are four types of room types available in Airbnb selection. With an entire room or apartment, which usually includes a bedroom, a bathroom, a kitchen and a separate entrance. It is the best choice for travellers who are looking for a home away from home. For private room type, travellers will be offered a private room for sleeping and may share some common space with others. This provides a local connection with another guest or host. For hotel rooms, it offers a level of service and hospitality associated with traditional hotels. For hotel room types, it typically includes a vibrant common area. On Airbnb, travellers can have a shared room with other guests or hosts if they do not mind sharing space with others. This room type is suitable for travellers who are looking for new friends and budget friendly stays. From the dataset, we can see there is 58% of room type is entire room/apartment, and it is mostly at the upper side of Chao Phraya River, where majority of the tourist hot spots can be found in this location. Figure 4 provides a summary chart depicting the count of Airbnb listings and the average rental prices based on different room types. Among the room types, hotel rooms have the highest rental prices, followed by the entire apartments/rooms and private rooms. On the other hand, shared rooms have the lowest average rental price, typically at 1,135 Baht. However, it is important to note that the availability of shared rooms is limited supply to Bangkok area compared to other room types. In contrast, entire apartments/rooms are the most commonly found accommodation option in the city.

Figure 5 provides a summary of the dataset, displaying the number of listings categorized by the year of the last review date. The data reveals that out of the total 17,936 Airbnb listings, a significant portion (at least 8,494 listings) remain active as of the years 2022 and 2023. Furthermore, in the figure summarized the rental price trends from early 2019 up to the most recent data available in March 2023 based on different room types.



	Entire home/apt	Private room	Hotel room	Shared room
Count of listing_url	10083	6176	629	508
Average of price	2963	2874	3221	1135

Figure 4 Count of Airbnb listing (Bangkok) and average rental (Baht) by Room Type

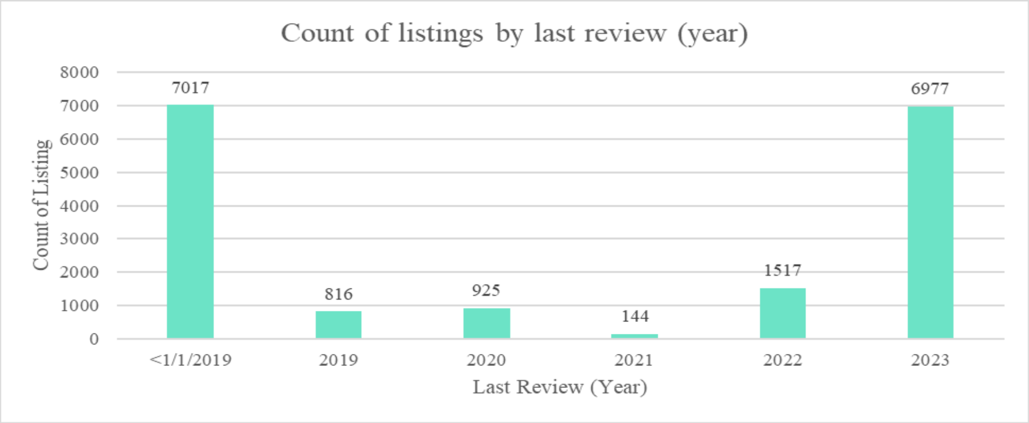
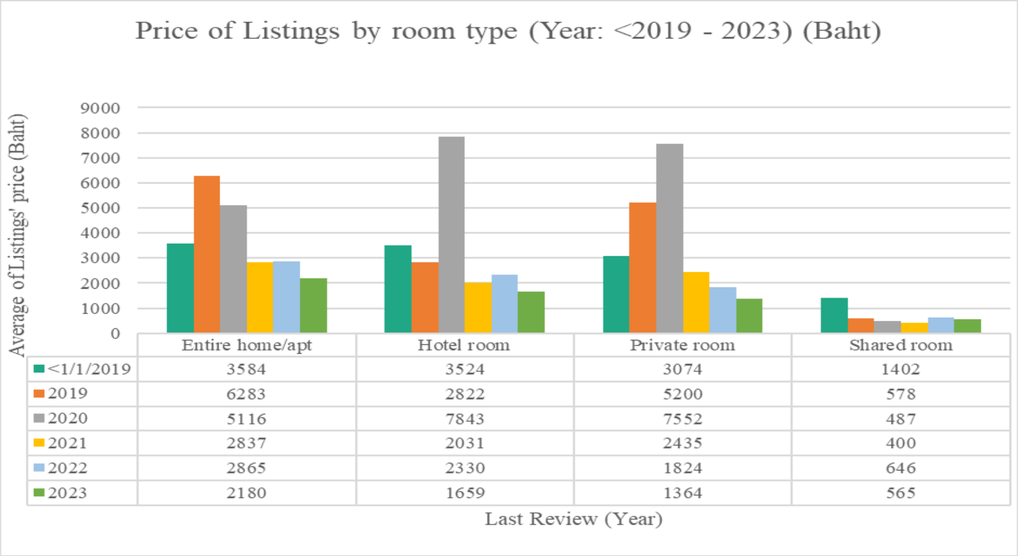


Figure 5 Count of listings by year of last review

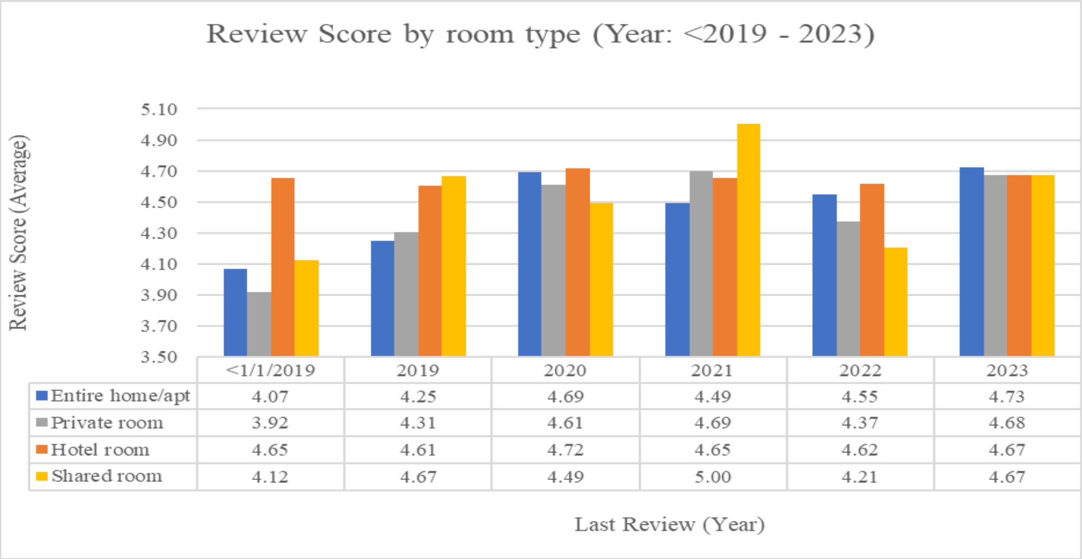
From Figure 6, there is a noticeable decline in the rental prices of listings after the year 2020. This drop was particularly significant for hotel rooms and private rooms, and lastly followed by entire homes/apartments. Furthermore, based on the most recent rental price data, it can be observed that entire homes/apartments have the highest average rental price, reaching an average of 2,180 Baht. On the other hand, shared rooms offer the most affordable rental option, with an average price of 565 Baht. These findings provide insights into the current pricing trends in the Airbnb market, highlighting the varying costs associated with different types of accommodations.



	Entire home/apt	Hotel room	Private room	Shared room
<1/1/2019	3584	3524	3074	1402
2019	6283	2822	5200	578
2020	5116	7843	7552	487
2021	2837	2031	2435	400
2022	2865	2330	1824	646
2023	2180	1659	1364	565

Figure 6 Price of listings by room type (Baht)

Figure 7 illustrates the trend of review scores for Airbnb listings from the year 2019 to 2023. The data reveals that all types of rooms, including the entire homes/apartments, private rooms and shared rooms, have experienced an improvement in their review scores. On average, the review scores have increased from 4.67 to 4.73, which is now comparable to the review score of hotel rooms. Interestingly, hotel rooms have maintained a consistent review score, averaging at 4.65 throughout the time period analysed. This finding suggests the overall quality and satisfaction level of Airbnb listings, regardless of room type, have shown improvement over the time, reaching similar levels of guest satisfaction as hotel rooms. This emphasizes the positive impact of review and guest feedback in shaping the quality and reputation of Airbnb accommodations.



	<1/1/2019	2019	2020	2021	2022	2023
■ Entire home/apt	4.07	4.25	4.69	4.49	4.55	4.73
■ Private room	3.92	4.31	4.61	4.69	4.37	4.68
■ Hotel room	4.65	4.61	4.72	4.65	4.62	4.67
■ Shared room	4.12	4.67	4.49	5.00	4.21	4.67

Figure 7 Review score by room type

Figure 8 presents the distribution of rental prices for Airbnb listings in Bangkok. The majority of rentals fall within the price range of 1,000 – 1,999 Baht, followed by listings with prices below 1,000 Baht. Out of the total count of 6,811 Airbnb listings within the 1,000-1,999 Baht range, 65% of them are categorized as entire homes/apartments, while 32% are classified as private rooms.

To address the potential outliers and normalize the rental price data, a log transformation has been applied. This transformation helps to mitigate the impact of extreme values and provides a more balanced representation of the rental prices. The results of this log transformation are shown in Figure 9, which demonstrates a more normalized distribution of the Airbnb rental prices in Bangkok.

By applying the log transformation, the data becomes more suitable for analysis, as it reduces the influence of outliers that might skew the results. This normalization process allows for a more accurate examination of the overall rental price distribution and helps to mitigate any potential bias introduced by extreme values.

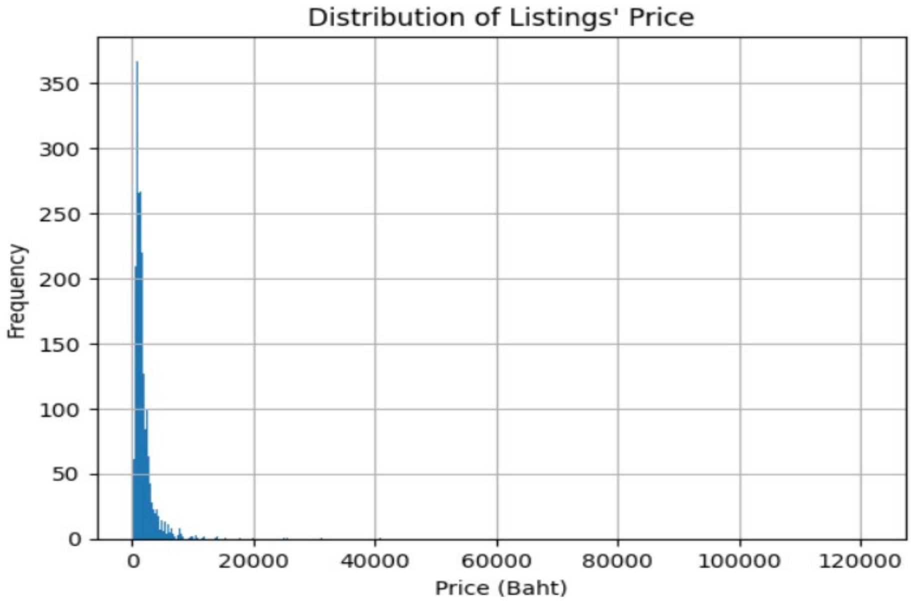


Figure 8 Price distribution of Bangkok Airbnb listings.

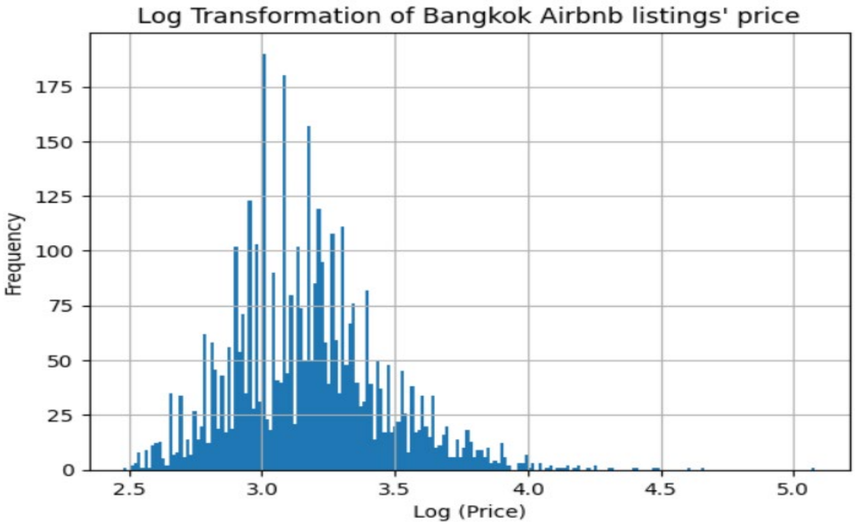


Figure 9 Log Transformation of Bangkok Airbnb listings' price

Among the total of 17,936 Airbnb listings in Bangkok, a significant proportion of hosts, specifically 10,354 listings, are giving respond within an hour, which is summarized in Figure 10. This accounts for approximately 60.2% of the total listings. The majority of these highly responsive hosts are from entire homes/apartments, followed by private rooms. Despite the prompt response times, it does not have a substantial impact on the review scores provided by guests. This finding as shown in Figure 11 suggests that while it is commendable for hosts to have quick response times, it does not necessarily translate into higher satisfaction levels or improved guest experiences. Other factors, such as the quality of accommodation, amenities provided, and overall hospitality, likely play a more significant role in determining the review scores.

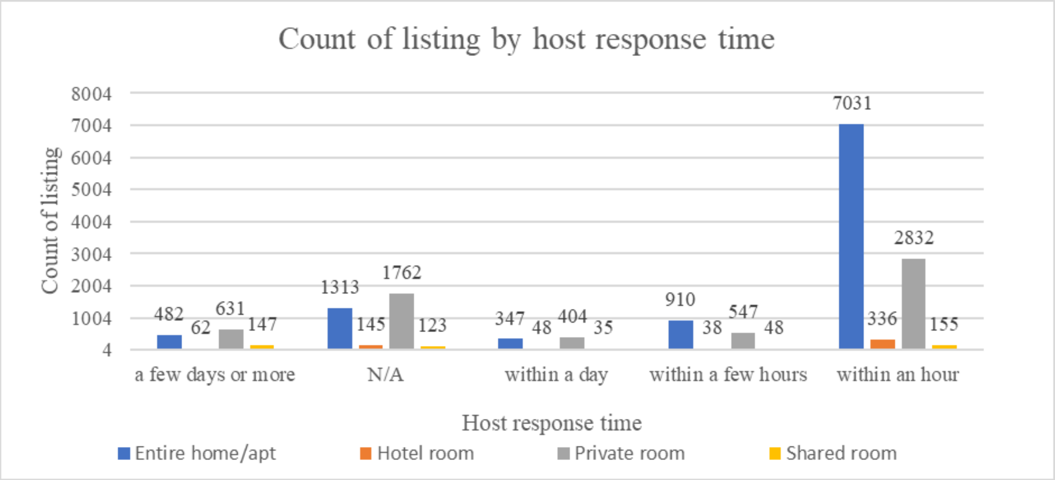


Figure 10 Host response time by room type.

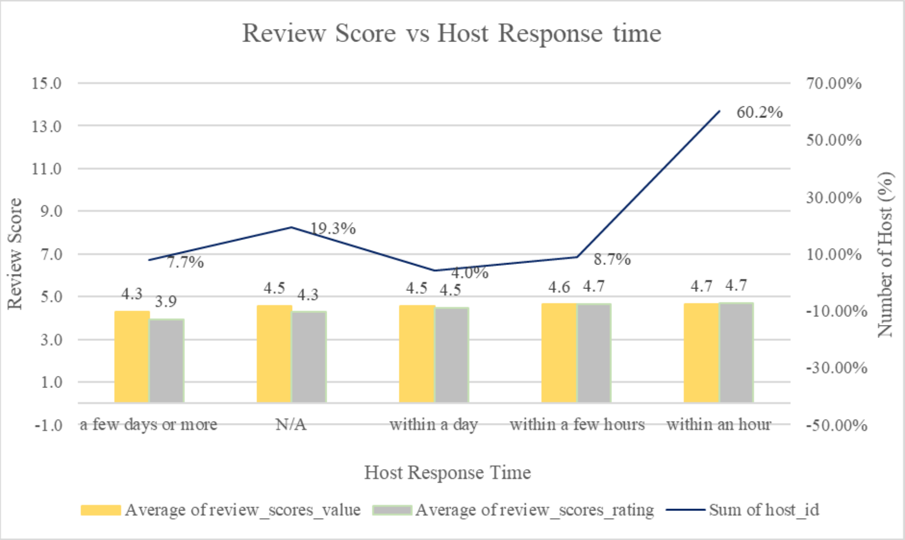


Figure 11 Relationship of review score to host response time.

Figure 12 provides a summary of the availability of Airbnb listings within a 365-day timeframe. The date is divided into different ranges of availability, specifically within 30 days, 60 days, 90 days or for the entire year. Out of the total 17,936 Airbnb listings in Bangkok, a substantial number of listings, specifically 14,245 listings have an availability of more than 90 days. This indicates that a significant portion of Airbnb hosts in Bangkok offer their listings for extended periods throughout the year. This highlights the availability and flexibility of Airbnb accommodations, with a large number of hosts willing to rent out their properties for extended durations. It suggests that travellers have a wide range of options to choose from when it comes to book accommodations for longer stays or extended periods in Bangkok.

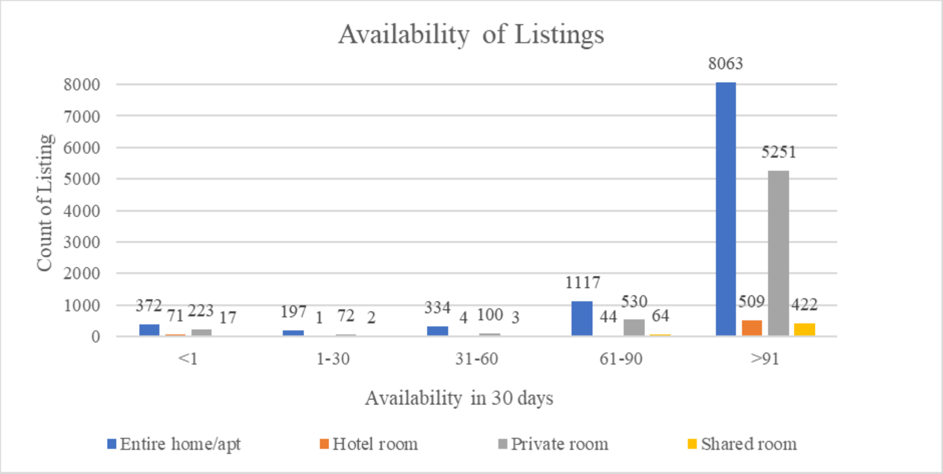


Figure 12 Availability of Thailand Airbnb listings in 365 days.

4.2 Data Pre-processing

In the Bangkok Airbnb dataset, a total of 75 fields/columns are available in the list. After reviewing the dataset, columns are first filtered and keep only the remaining 27 useful columns in the dataset for further exploration, due to the several reasons:

- data is duplicate in two columns: calendar updated, calendar last scrapped, last scraped, minimum_nights_avg_ntm, maximum_nights_avg_ntm, host identity verified, bathrooms, property type, etc.
- not relevant to observations: host verifications, host has profile picture, neighbourhood cleansed, host thumbnail url, host picture url, picture url, host about, host location, scrape id, source, etc.
- most of the column's values are missing: licensed, neighbourhood group cleansed, etc.

Next, data types of the column are converted to numbers (integer, date, currency). Empty entries were removed. Table 3 summarized the description of each column and categorized it into host's profile, host's responsiveness, property information, property availability and property reviews.

Table 3 Description of features in Bangkok Airbnb dataset.

No	Field	Type	Description
Host Profile			
1	host_identity_verified	Boolean	[t=true; f=false]
2	host_is_superhost	boolean	[t=true; f=false]
Host responsiveness			
3	host_response_time	Datetime	
4	host_acceptance_rate		That rate at which a host accepts booking requests.
5	host_response_rate		
Property information			
6	room_type	text	[Entire home/ apartment |Private room| Shared room| Hotel] All homes are grouped into the following three types of room: Entire place

No	Field	Type	Description
			Private room Shared room
7	accommodates	integer	The maximum capacity of the listing
8	bathrooms_text	string	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 are used.
9	bedrooms	integer	The number of bedrooms
10	beds	integer	The number of bed(s)
11	amenities	json	
12	price	currency	daily price in local currency
13	minimum_nights	integer	minimum number of night stay for the listing (calendar rules may be different)
14	instant_bookable	boolean	[t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.
Property Availability			
15	has_availability	Boolean	[t=true; f=false]
16	availability_30	integer	The availability of the listing 30 days in the future as determined by the calendar.
Property Reviews			
17	number_of_reviews	integer	The number of reviews the listing has
18	number_of_reviews_ltm	integer	The number of reviews the listing has (in the last 12 months)
19	reviews_per_month	numeric	The number of reviews the listing has over the lifetime of the listing
20	last_review	date	The date of the last/newest review
21	review_scores_rating	Numeric	
22	review_scores_accuracy	Numeric	
23	review_scores_cleanliness	Numeric	
24	review_scores_checkin	Numeric	
25	review_scores_communication	Numeric	
26	review_scores_location	Numeric	
27	review_scores_value	Numeric	

4.3 Analysis of Features Relationship

Correlation between independent variables and the target variables are explored to reveal hidden connections, and patterns that can inform feature selection and model training. These insights are critical for designing an effective predictive model and making informed decisions based on data driven evidence.

In this study, two techniques: heat map analysis and SelectKBest are used. Heat map analysis is a graphical representation data in a matrix format, where values are represented by colors. Figure 13 shows a heat map is drawn to study the correlation between independent variables to target. Dark colour indicates lower correlation between Airbnb price and the independent variables. The lighter the colour, indicates the feature is higher correlated to Airbnb price. It is found that feature such as bedrooms, accommodates, beds, and bathrooms_text have a strong correlation to Airbnb price.

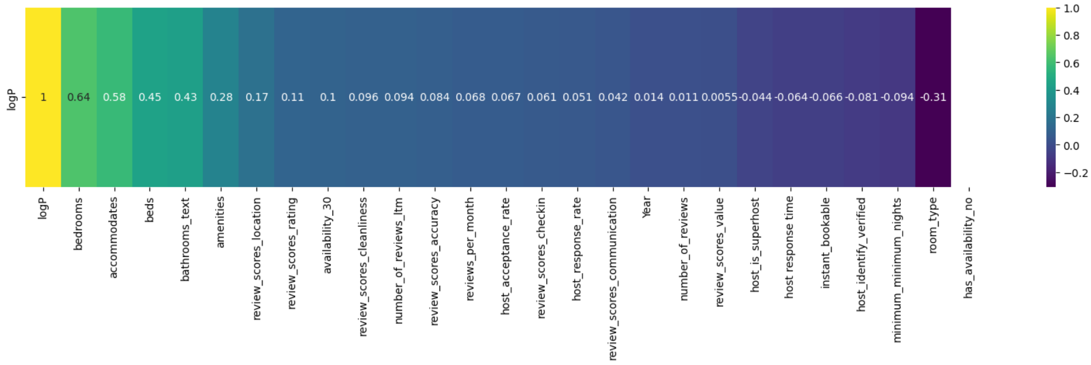


Figure 13 Heat Map

Figure 14 summarized the importance score of every independent variable that was computed by using SelectKBest method. By comparing these two feature analysis tools, heat map only shows limited insight into the feature correlation with the target variable, but it can help to understand which features have positively or negatively correlated to target variable.

SelectKBest is a method used for feature selection in machine learning, with a specific emphasis on statistical analysis and the ranking of features based on their connection to the target variable. This technique assigns scores to features, indicating the strength of their relationship with the target variable. The primary purpose of SelectKBest is to identify those features that possess the greatest ability to contribute to the predictive accuracy of the model. It's a valuable tool for enhancing the efficiency of the model by focusing on the most impactful features.

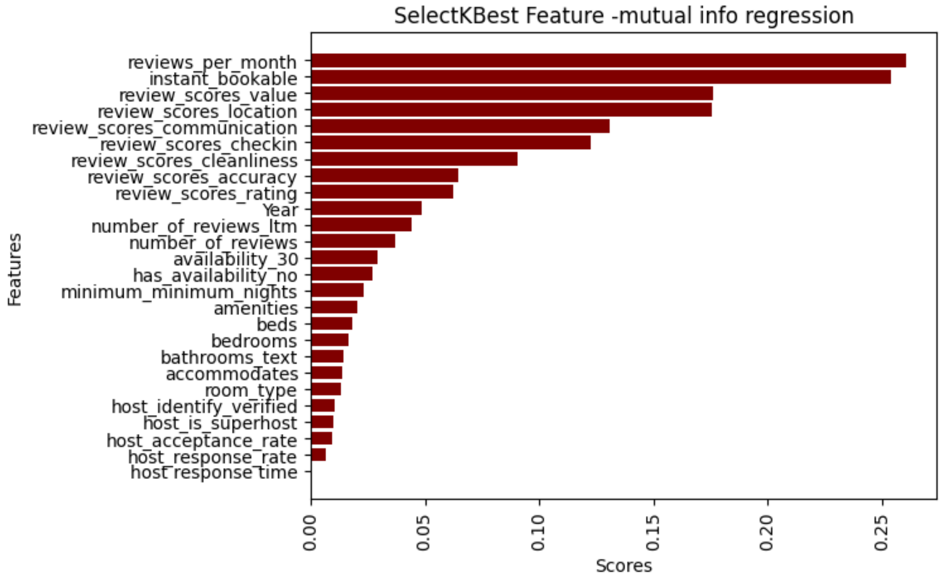


Figure 14 Score of features in SelectKBest

In SelectKBest method, it shows that the most important feature is reviews_per_month (average number of reviews in a month), followed by instant_bookable (is the Airbnb listing is instantly bookable), review_scores_value (review scoring of a listing), review_scores_location (review scoring given by a customer to the location of Airbnb listing), review_scores_communication (review scoring given by a customer to the communication between host and customer), review_scores_checkin (review scoring given in terms of checkin process), review_scores_cleanliness (review scoring given to the cleanliness of an Airbnb listing), and year (date of the last reviewed given to the Airbnb listing). The other features show lower scores in predicting Airbnb listings price.

CHAPTER 5: IMPLEMENTATION

Data is cleaned according to methods specified in Section 4.2, and feature of last_review is simplified as Year to indicate the last review year to the Airbnb listing. With the cleaned and normalized dataset, Airbnb price prediction models are constructed by implementing linear regression, neural network, bagging regressor, gradient boosting, decision tree regressor, random forest regression and XGBoost methods. Model performance is evaluated using the metrics described in Section 3.4.

In the next step, Lasso regression is implemented together with linear regression, neural network and bagging regressor, which these models do not have implicit feature selection during the model building process. For other machine learning methods: decision tree regression, gradient boosting regression, XGBoost regression and random forest regression have implicit feature selection function, which provides feature importance scores.

5.1 Lasso Regression

In Lasso regression, L1 penalty is applied to prediction model, and optimal lambda, λ is identified by using LassoCV for each model. For linear regression, λ = 0.0030 is selected, for neural network, alpha is set to 0.0030 and for bagging regression, optimal λ = 0.0030.

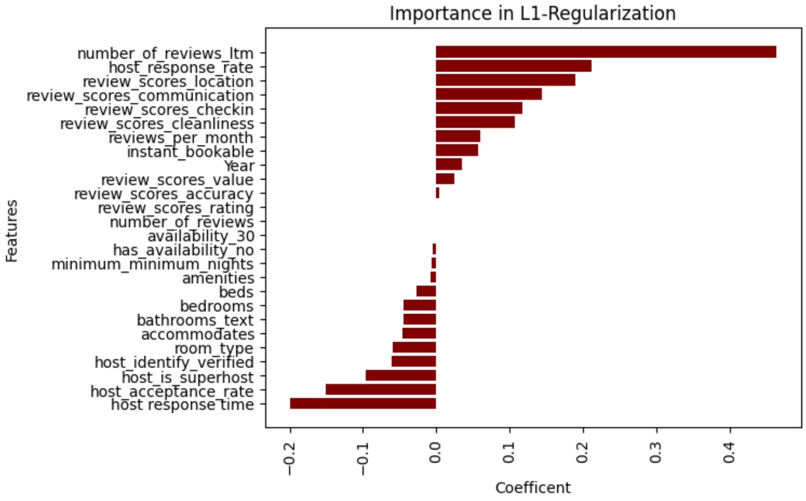


Figure 15 Coefficient of features in Linear Regression model.

Figure 15, shows the coefficient of features in lasso regression model order by the coefficient value. From the figure, it is found that significant features are, with the positive relationship number_of_reviews_ltm, host_response_rate, review_scores_location, review_scores_communication, and review_scores_checkin. Features with negative relationship with target variables are: host_response_time, host_acceptance_rate and host_is_superhost.

For ensemble models, gradient boosting regression, decision tree regression, random forest regression and XGBoost regresson, implicit feature selection is available. Feature importance ranking is more obviously seen by a positive number between 0 to 1. The larger the coefficient for the features, the feature is more accurate in predicting the Airbnb price.

Figure 16 – Figure 19 summarize the feature_importances_ of all the features for each model. By comparing all four model, it is found that importance features that are considered in the ensemble techniques are: number of reviews to the listings, host response rate and location of the listing.

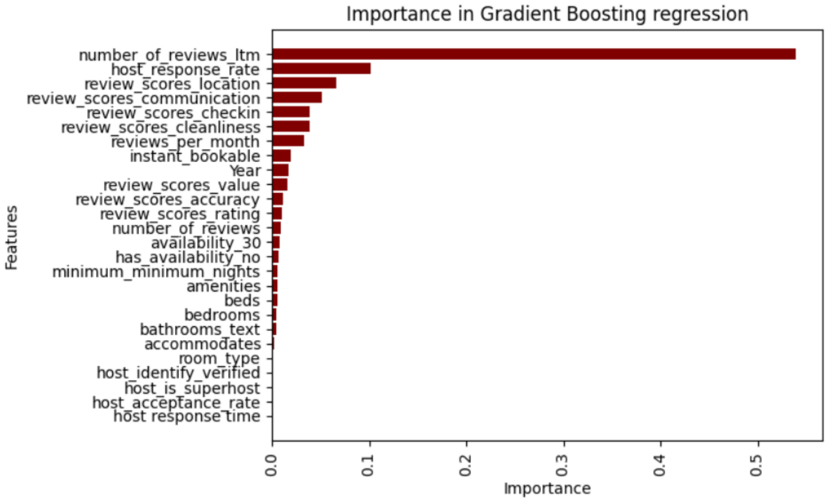


Figure 16 Feature importance in gradient boosting regression model.

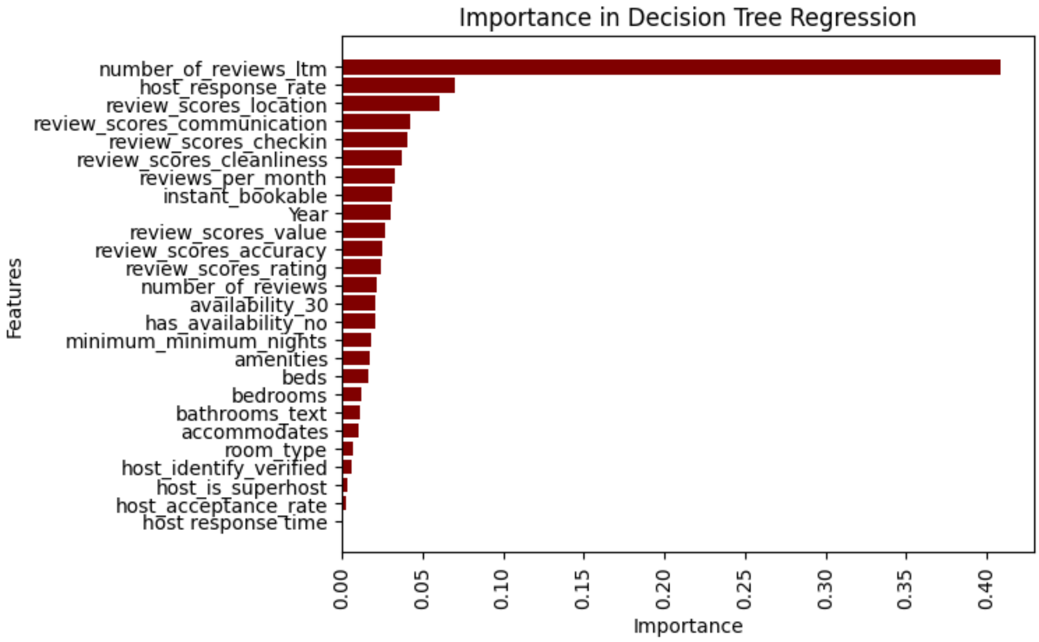


Figure 17 Feature importance in decision tree regression model.

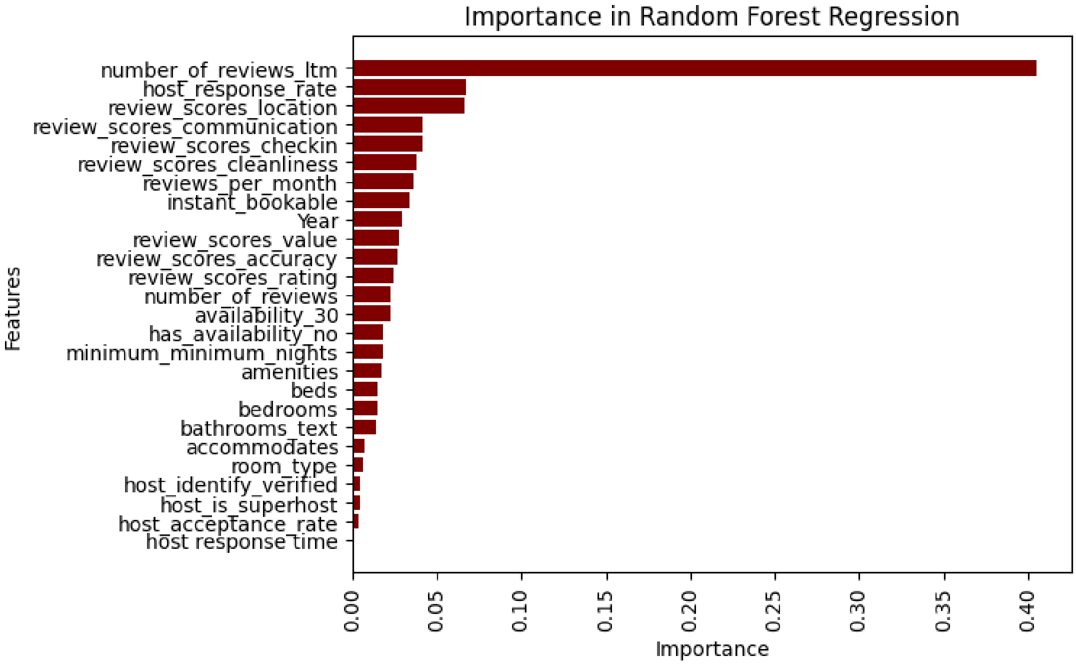


Figure 18 Feature importance in random forest regression model.

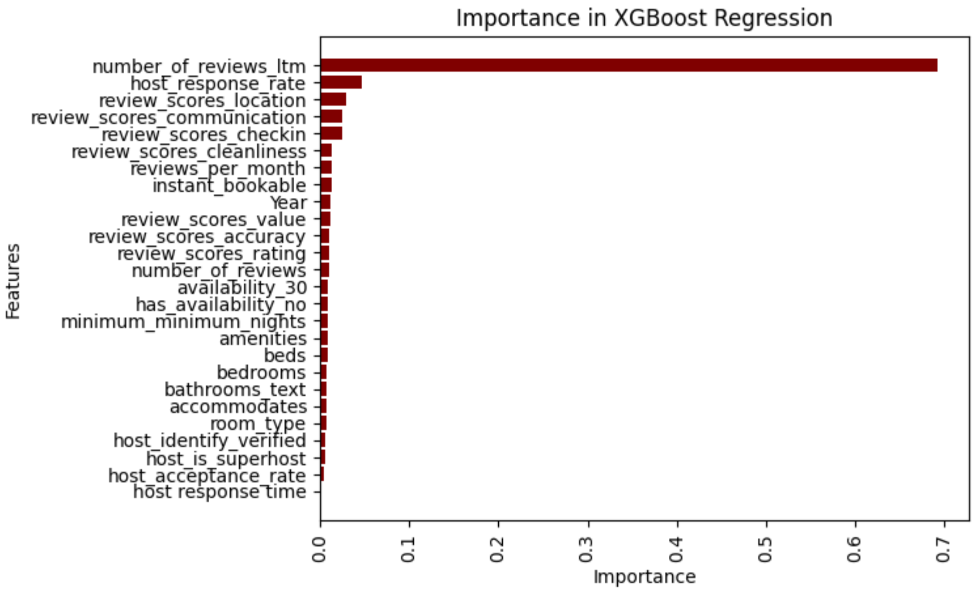


Figure 19 Feature importance in XGBoost regression model.

The feature importance ranking shows that there are variances when different models selecting the features. Different models assign different coefficient to the features in order to predict Airbnb price more accurately. Yet, the top pareto of the significant features are still similar among the models. From LASSO regression technique and the implicit feature selection technique, number of reviews, host response rate and the location of Airbnb listing are the top consideration when a customer choose an Airbnb listing.

CHAPTER 6: DISCUSSION

This project is implemented in Numpy and scikit-learn. 5-fold cross validation is used to split, train and validate the prediction model. 80% of the data is used for training, and the remaining 20% data is used for validation. Total 5 iterations were repeated in order to compute the fitness and prediction error of the machine learning model.

Summary of the results for each machine learning model are tabulated in Table 4 for comparing the metric performance for single method techniques model. This includes linear regression, neural network model, and bagging regression models.

For linear regression model, with Lasso regression that chose lambda, λ = 0.0030, R2 value is slightly improved from 0.5419 to 0.5529, MAE is reduced from 0.5146 to 0.5100, and RMSE is reduced from 0.6721 to 0.6687. However, for bagging regression, model performs slightly better without using lasso regression to identify the important features. For neural network model, with L1- like regularization is created by using the alpha parameter. In this case, alpha is set to 0.0030, and the model performance is improved, R2 value is improved to 0.4256 from 0.3986, MAE is improved from 0.5810 to 0.5693 and RMSE is improved from 0.7691 to 0.7529. In this case, linear regression model performs better compared to the other two models. 54.23% data was able to be explained through the prediction model.

Table 4 Comparison of model performance with and without feature selection.

Model	**Without L1-regularization**			**With L1-regularization**			
	MAE	**R2**	**RMSE**	**λ**	**MAE**	**R2**	**RMSE**
Linear Regression	0.5146	0.5419	0.6721	0.0030	0.5142	0.5423	0.6717
Bagging regression	0.4854	0.5665	0.6537	0.0030	0.5146	0.5418	0.6721
				alpha	**MAE**	**R2**	**RMSE**
Neural Network	0.5810	0.3986	0.7691	0.0030	0.5693	0.4256	0.7529

Table 5 summarized metric performance for model that is trained with ensemble models. Among the four models, gradient boosting has the best performance with MAE 0.4938, R2 score 0.5738 and RMSE 0.6481. This is followed by random forest regression, and XGBoost regression. In this case, 57.38% of the price data was able to be explained in gradient boosting model, and the RMSE is also lower by 0.0236 as compared to linear regression model.

In overall, adding penalty to tackle overfitting in neural network shows improvement in prediction accuracy. However, of all the prediction model, gradient boosting gives the highest accuracy, with lower overfitting to the dataset. Random forest regression gives comparable accuracy to gradient boosting regression model. Among all the models, gradient boosting performs the best with highest R2 and lowest errors.

Table 5 Model performance for ensemble techniques.

Model	**MAE**	**R2**	**RMSE**
Gradient Boosting	0.4938	0.5738	0.6481
Decision tree regressor	0.6977	0.1145	0.9324
Random Forest Regression	0.4890	0.5718	0.6490
XGBoost	0.5074	0.5458	0.6693

CHAPTER 7: CONCLUSION

The objective of the project is to identify Airbnb price prediction model and include feature selection technique for models which do not have implicit feature selection.

To achieve the objective, latest dataset as of March 2023 is obtained from InsideAirbnb.com, and data exploration is performed to have more understanding on the correlation between features and the target variable. In order to reduce the noise in the dataset, independent variables that are not relevant, independent variables that have too many missing data or with duplicated data between two columns and outliers in independent variables are filtered out. 5-fold cross validation is used to train, and validate the performance of machine learning models. Lasso regression is added to linear regression, neural network and bagging regression to add L1 penalty term based on absolute values of the coefficients. Features with zero coefficients are removed from the model, and thus reduce the overfitting on the model. For ensemble technique, implicit feature selection is available, features selection is automatically performed in the machine learning model.

In conclusion, gradient boosting regression model gives the best accuracy with MAE 0.4938, R2 score 0.5738 and RMSE 0.6481. From the feature analysis, important features include number of reviews in the last twelve months, location of the Airbnb listing, cleanliness of the listing, host response rate of an Airbnb listing and a superhost can significantly impact the price of an Airbnb listing.

In this research, the algorithms were trained using the default parameters without any performance tuning. Future work can be considered to perform hyper parameter tuning to fine tune the optimal parameters to be used in machine learning models. To further improve the model prediction accuracy, information such as latitude and longitude can be used to study the popular geographical location of an Airbnb listing, sentiment analysis can be integrated to process the text in column such as reviews, amenities, neighbourhood overview, etc. thus extract the textual data that express sentiment about the listing.

REFERENCE

Ahuja, A., Lahiri, A., & Das, A. (2021). Predicting Airbnb Rental Prices Using Multiple Feature Modalities. *ArXiv*, *abs/2112.06430*.

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *CoRR*, *abs/1907.12665*. http://arxiv.org/abs/1907.12665

Kirkos, E. (2021). Airbnb listings' performance: determinants and predictive models. *European Journal of Tourism Research*, *30*, 3012. https://doi.org/10.54055/ejtr.v30i.2142

Lewis, Laura. (2019). Predicting Airbnb prices with machine learning and deep learning. *Medium-Toward Data Science*.

Liu, Y. (2021). Airbnb Pricing Based on Statistical Machine Learning Models. *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, 175–185. https://doi.org/10.1109/CONF-SPML54095.2021.00042

Masrom, S., Baharun, N., Razi, N. F. M., Rahman, R. A., & Abd Rahman, A. S. (2022). Particle Swarm Optimization in Machine Learning Prediction of Airbnb Hospitality Price Prediction. *International Journal of Emerging Technology and Advanced Engineering*, *12*(1), 146–151. https://doi.org/10.46338/IJETAE0122_14

Otero Gomez, D., Manrique, M. A. C., Sierra, O. B., Laniado, H., Mateus C, R., & Millan, D. A. R. (2020). Housing-Price Prediction in Colombia using Machine Learning. *OSF Preprints*, *w85z2*.

Pandey, A. C., Misra, S., & Saxena, M. (2019). Gold and diamond price prediction using enhanced ensemble learning. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–4.

Peng, N., Li, K., & Qin, Y. (2020). Leveraging Multi-Modality Data to Airbnb Price Prediction. *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, 1066–1071. https://doi.org/10.1109/ICEMME51517.2020.00215

Rezazadeh Kalehbasti, P., Nikolenko, L., & Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 173–184). Springer International Publishing.

Thakur, N., Jain, R., Mahajan, A., & Islam, S. M. N. (2022). Deep Neural Network based Data Analysis and Price Prediction framework for Rio de Janeiro Airbnb. *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, 1–7. https://doi.org/10.1109/I2CT54291.2022.9824383

Trang, L. H., Huy, T. D., & Le, A. N. (2021). Clustering helps to improve price prediction in online booking systems. *International Journal of Web Information Systems*, *17*(1), 45–53. https://doi.org/10.1108/IJWIS-11-2020-0065

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. https://doi.org/https://doi.org/10.1016/j.ijhm.2016.12.007

Yang, S. (2021). Learning-based Airbnb Price Prediction Model. *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*, 283–288. https://doi.org/10.1109/ECIT52743.2021.00068

Yang, Y., Mueller, N. J., & Croes, R. R. (2016). Market accessibility and hotel prices in the Caribbean: The moderating effect of quality-signaling factors. *Tourism Management*, *56*, 40–51. https://doi.org/10.1016/J.TOURMAN.2016.03.021

Ye, P., Qian, J., Chen, J., Wu, C., Zhou, Y., De Mars, S., Yang, F., & Zhang, L. (2018). Customized Regression Model for Airbnb Dynamic Pricing. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 932–940. https://doi.org/10.1145/3219819.3219830

Yu, H., & Jiafu Wu. (2016). Real estate price prediction with regression and classification. *CS229 (Machine Learning) Final Project Reports*.