ORIE 5741 Final Report

Jiongjiang Duan(jd2253), Sirawich Tippawanich(st895), Jiaqi Zhang (jz2232)

December 4, 2021

1 Dataset Introduction

The data set used for this project was published on Kaggle by a user named Ahmad Bhat. The data set contains Airbnb (a popular home-sharing platform) listings in 10 major cities including information about hosts, pricing, location, room types and over 5 million historical reviews. Among 30 features, there are 11 nominal, 2 ordinal, 5 discrete and 12 continuous variables. There are a total of 180,000 unique hosts who were responsible for 280,000 unique listings in the data set.

2 Problem definition and business importance

Given the data set of listings and review scores, the following questions can be raised:

- Do track records of hosts influence their behavior in determining the listing prices? For review metrics to be good incentives for hosts to maintain their standards, a relationship should exist between higher scores and ability to ask for higher prices. Platform managers should assess the effectiveness of the incumbent review system and find a replacement if it is found to be ineffective.
- Does the effectiveness of the review system vary across major cities or neighborhoods within the city? How do they vary across different property tiers? Do owners of high-value properties wield more power in determining listing prices? Do higher number of claimed amenities allow hosts to price higher relative to their review performances?

Home-sharing platforms should act as moderators between hosts and guests in ensuring that there is a system in place to protect both parties in all transactions. The ability of guests to provide reviews is a powerful tool that can assure fairness between pricing and underlying service the guests receive. Rigorous analysis of the dynamic between reviews and market pricing of properties can help a platform identify anomalies and vulnerability in their policies and governance as a moderator.

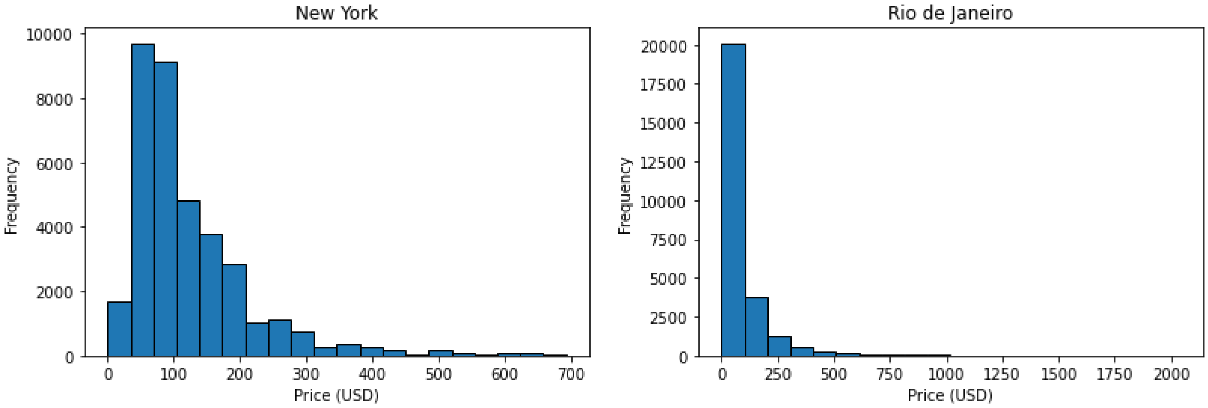
To ensure business longevity, platforms such as Airbnb must satisfy both hosts and guests. Guest dissatisfaction can arise from sources such as unfair pricing and inaccurate listing details, both of which the platform is ultimately responsible for regulating. Similarly, to ensure meritocratic-based competition between hosts, the review process should be effective in establishing a market equilibrium between quality and pricing. Insights developed from review data analysis can bring to light areas of improvement in the platforms' operations.

3 Data Exploration and Visualization

Histograms are a powerful tool in representing the distribution of a variable within a data set. Plots of Airbnb listing prices against their frequencies show that prices follow right-skewed distributions for all cities within the data set. Some cities feature price distributions where the modes are at the lowest price bin (Rio De Janeiro, Cape Town, Mexico City and Istanbul). Paris, New York, Bangkok , Sydney, Rome and Hong Kong feature distributions where the lowest price bin was not the mode. Histograms showing price distributions for New York and Rio de Janeiro (representing the two characteristics of distributions mentioned) are shown in Figure 1.

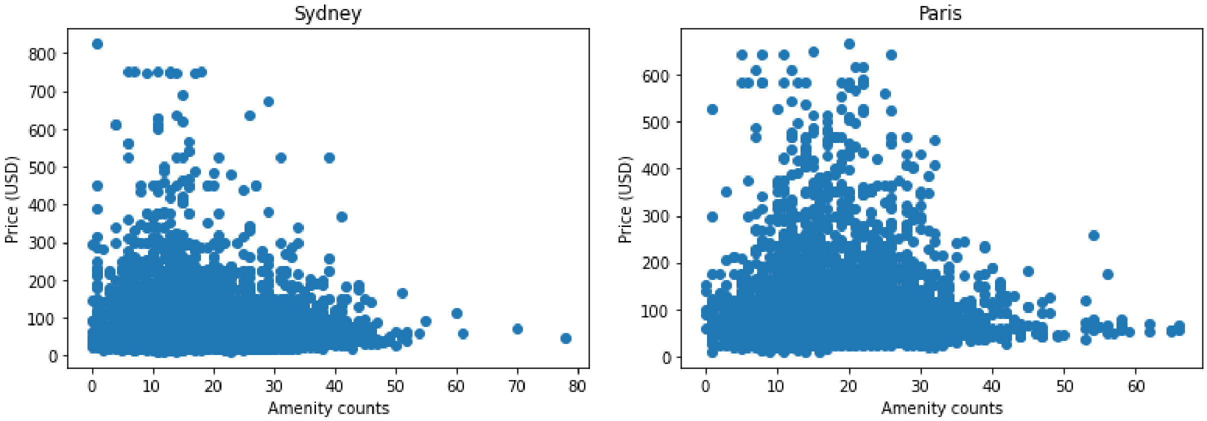
The relative differences in pricing is also worth noting. The average listing prices for all cities are between $20 to $100, suggesting that they are within the same order of magnitude. As expected, higher averages are associated to cities in countries with higher costs of living. As a product that provides visitors with lodging spaces, available amenities at the properties can be an influential factor in determining how prices are set. Number of amenities for

Figure 1: Price distribution for New York and Rio de Janeiro



each property was extracted from the lists of amenities within each listing and plotted against prices. There is a clear observation across all cities that there are a few low-price listings who provide high number of amenities. The most expensive listings in each city provide between 10 - 30 amenities. The trend suggests that some lower-priced hosts try to create a draw towards value by listing longer lists amenities. Scatter plots of amenity counts against listing prices for Sydney and Paris are shown in Figure 2. We see that for both cities, the more expensive listings often exhibit lower number of amenity counts.

Figure 2: Amenity count and price for Sydney and Paris



4 Feature Engineering

4.1 Dealing with Missing Values

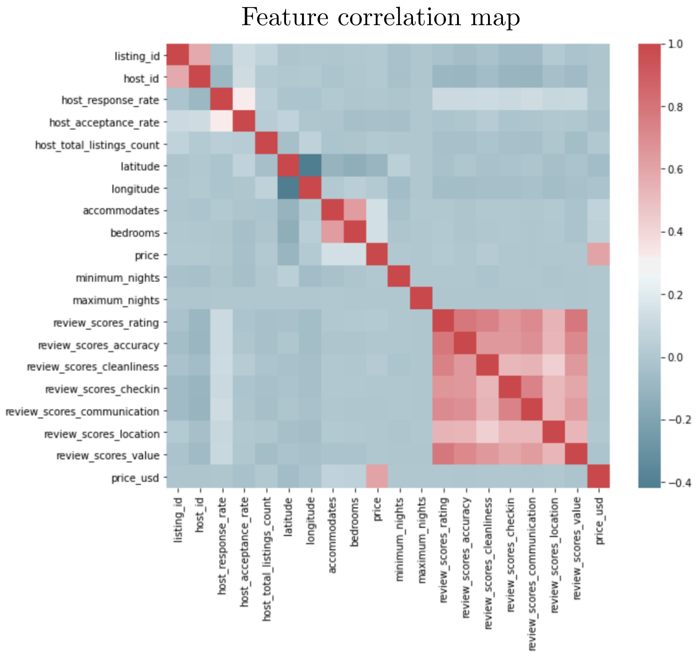
There are 279,712 observations in the Airbnb listings where 128,617 observations has at least one null value in the variables. The goal of the project is to investigate whether the review has an impact on the price of listing, so we only kept observations whose review score rating is not null. After that we used methods below to deal with other null values:

1. Delete columns which has more than 80 percent missing values: district
2. Replace the missing values with the mean: bedrooms
3. For the informative missing values such as response rate and acceptance rate, we first replace the null values with 'unknown', treating them as a categorical value and then transform into dummy variables.
4. Group the values into bracket and treat the null value as an categorical value. For example, both host response rate and acceptance rate are skewed to the right. If the rate is smaller than 0.3 in the host response rate

column, it's a low response rate. If it's between 0.3-0.7, it's a medium response rate. If it's over 0.7, then it's a high response rate. The null value in host response rate is still null but is in a new bracket called unknown response rate.

The total number observations after dealing with the missing values is 188,199, which is large enough for regression modeling.

4.2 Correlation Analysis



Correlation Map Among Features

To run a regression analysis, the features need to be independent otherwise it will run into the multicollinearity issue. From the correlation map above, the review score rating is closely related to review score accuracy, review score cleanliness, review score check-in, review score communication, review score location and review score value. Therefore, we deleted all review score related features except review score rating. The longitude and latitude have high correlation as well but since city can also represent geographical information, we deleted the longitude and latitude.

5 Model Details

5.1 Linear Regression and Analysis

A preliminary linear regression was performed with all the processed features as inputs and price (in USD as output). The entire data set was used to fit the model to gauge the performance. As an evaluation metric, mean squared error was computed between predicted and real listing price. The preliminary model with all features provided a training error of 37,750. The high mean squared error suggests that simple linear regression is not capable of capturing the relationship between the features and predicted variable.

The p-value of each variables in the regression model was examined. Notably, no statistical significance was observed in variables such as maximum nights and the binary variables representing the city locations of New York and Sydney. Other features show p-values lower than a universally accepted threshold of 0.05, suggesting that they have significant relationship with the listing price.

Examination of coefficients for categorical variables provide some hints about their impact on pricing. There is no meaningful difference in the coefficients for the variable indicating whether a host is a "super host" or not.

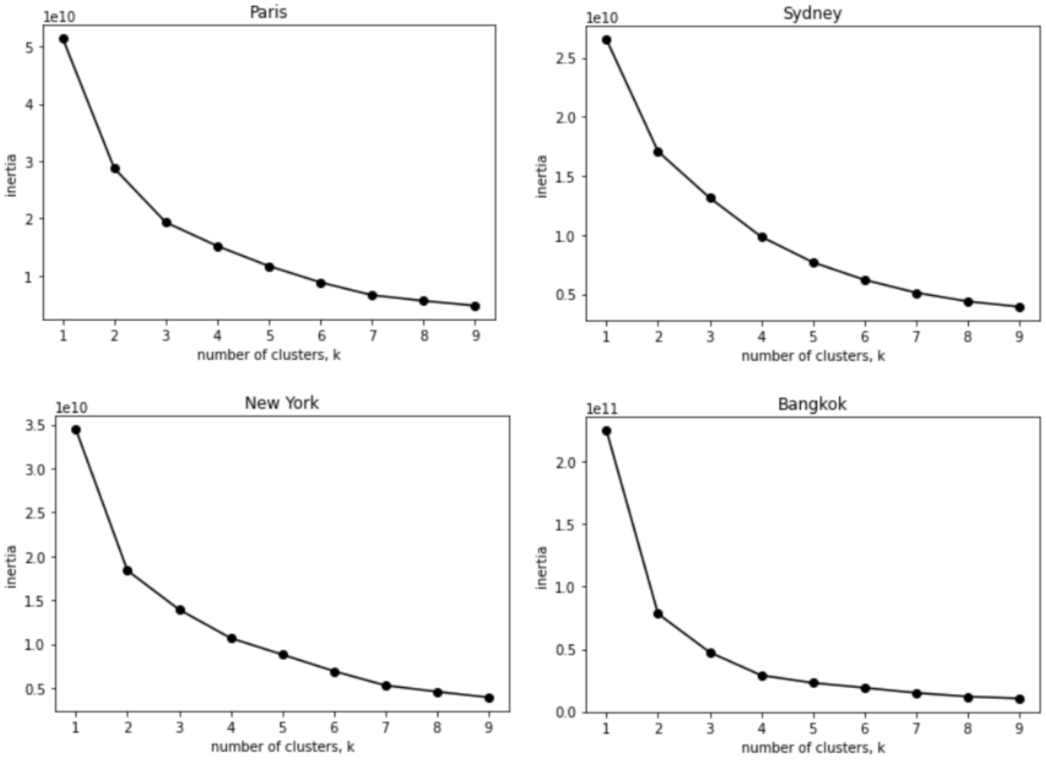
The coefficients of the regression model for numerical variables are also in line with expectation. The number of bedrooms and people that listings accommodate have positive coefficients in magnitudes that agree with how much price scales with capacity of accommodations. Host experience, measured as days since initial host registration, also has a positive coefficient. As mentioned earlier, counts of amenities in the listings are spread out among the range of pricing. The coefficient for count of amenities is positive and small in magnitude, suggesting that although the number does influence pricing positively, the effect is small to negligible. A positive coefficient for review scores rating (the overall score given in reviews) hints that higher prices are associated to higher scores given by visitors.

Fitting a linear regression to a data set with numerous one-hot encoded features is far from ideal. However, it enables primitive analyses of coefficients and statistical significance for some (mostly numerical) variables. A benchmark error performance metric was also achieved from the linear model. A natural next step is to implement non-linear models in order to further explore the relationship between individual features and listing pricing.

5.2 Clustering of Selected Cities' Listings

To narrow down the analysis, four cities within the data set were selected for modelling based on their continent and number of observations. New York, Paris, Bangkok and Sydney were represented by the highest number of listings in their respective continent. K-means clustering was first performed in order to divide the listings of each city into clusters based on the transformed features. The number of clusters, k, for each city was selected based on the "elbow method" and the relative decrease in inertia with an increase in k (Kavyazin, 2019). The resulting number of clusters for New York, Paris, Bangkok and Sydney were 3, 4, 3 and 4 respectively.

Figure 3: Inertia and number of clusters



Paired scatter plots and mean values of key features were examined for each city. The key takeaways from clustering the listings are as follow:

- The cluster with most expensive listing average in each city has the lowest average number of listings. Vice-versa is true for the cluster with the cheapest average listing.
- There is an inverse relationship between review scores rating and prices in Paris. The relationship is the opposite for Sydney. There is no clear relationship between the two for New York and Bangkok.
- Generally, the bigger the property (measured by number of bedrooms), the higher review score rating the listing receives.
- Host experience (time since hosts first register) has a positive correlation with listing prices in Bangkok and New York. Negative correlations are observed for Paris and Sydney.

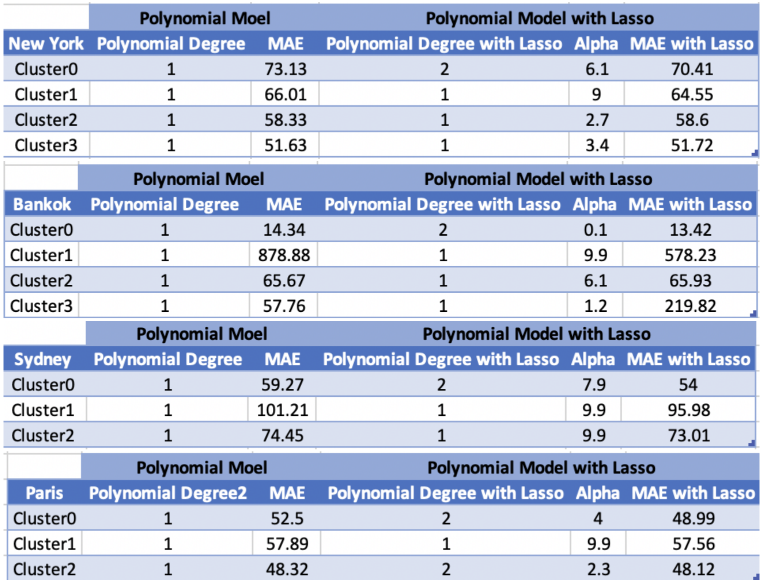
- In Sydney, more responsive hosts have cheaper listings, shorter hosting experience and lower review scores rating. The opposite relationships are observed in Paris. No clear trend was observed for either Bangkok or New York.

For each city, each cluster will be modelled separately since the differences in feature values such as the property classes (price, size) and host experiences can be obstacles in achieving well-generalized models. The dataset does not contain potential protected attributes that can directly lead to discrimination or unfairness. However, it should be recognized that differences in factors such as relative locations within each city and pricing brackets could lead to an unfair model output if all listings were assessed together. Clustering has another benefit in addressing this issue. Conducting modeling on different clusters ensures that listings are assessed in their respective groups.

5.3 Polynomial Regression

Since each cluster have different feature values, we used polynomial regression on each cluster in each region. We tried two methods. The first method is to apply polynomial on all of the features. We used the cross validation method to find the optimal polynomial degree. This is important because the degree of the polynomial dramatically increases the number of input features, resulting in not generalizing well with unseen dataset. The second method is to utilize Lasso regression to select the features whose coefficients are not zero and then ran the model on selected features. The advantage of using Lasso is that it will shrink the coefficients that are not correlated to the label to exactly zero, selecting the most important features to fun in the model and leading to lower mean absolute error. The hyper-parameters in the second methods are alpha which is the magnitude of the regularization and the polynomial degree. We used the GridSearch method with cross validation equal to 5 to select the best combination. The model results are described below:

Figure 4: Polynomial regression results



	Polynomial Moel		Polynomial Model with Lasso		
New York	Polynomial Degree	MAE	Polynomial Degree with Lasso	Alpha	MAE with Lasso
Cluster0	1	73.13	2	6.1	70.41
Cluster1	1	66.01	1	9	64.55
Cluster2	1	58.33	1	2.7	58.6
Cluster3	1	51.63	1	3.4	51.72

	Polynomial Moel		Polynomial Model with Lasso		
Bankok	Polynomial Degree	MAE	Polynomial Degree with Lasso	Alpha	MAE with Lasso
Cluster0	1	14.34	2	0.1	13.42
Cluster1	1	878.88	1	9.9	578.23
Cluster2	1	65.67	1	6.1	65.93
Cluster3	1	57.76	1	1.2	219.82

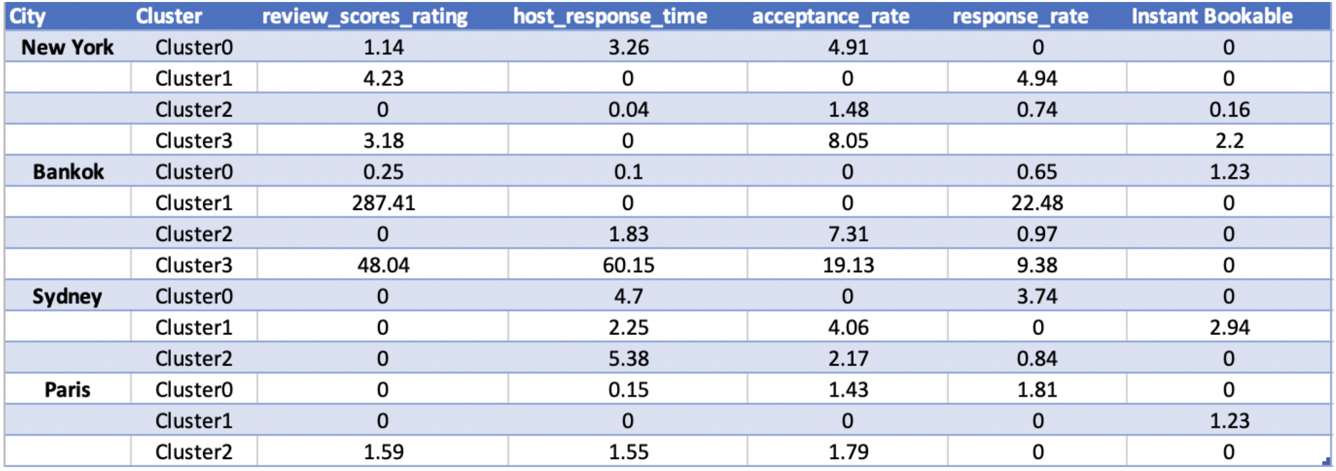
	Polynomial Moel		Polynomial Model with Lasso		
Sydney	Polynomial Degree	MAE	Polynomial Degree with Lasso	Alpha	MAE with Lasso
Cluster0	1	59.27	2	7.9	54
Cluster1	1	101.21	1	9.9	95.98
Cluster2	1	74.45	1	9.9	73.01

	Polynomial Moel		Polynomial Model with Lasso		
Paris	Polynomial Degree2	MAE	Polynomial Degree with Lasso	Alpha	MAE with Lasso
Cluster0	1	52.5	2	4	48.99
Cluster1	1	57.89	1	9.9	57.56
Cluster2	1	48.32	2	2.3	48.12

The polynomial regression with the selected features performs better than the polynomial regression with all the features based on Mean Absolute Error. For polynomial regression on all the features, the optimal polynomial degree is always 1 and it's probably because there are many categorical features, which is very hard for regression modeling. The MAE for cluster 1 in Bangkok is very relatively large because there are really few data points in this cluster.

Using the lasso regression, we can also investigate whether the review system has a determinant effect on the price of housing. The review system contains two main categories: review score and host behaviors. The host behaviors can be whether the host respond and accept to requests in a timely manner and whether they provide instant booking service. If the coefficients of these feature are zero in the cluster, then it means that the features are not important in determining the price of housing. If the review system works well, we expect that the review scores and the features related to host behavior should not be zero in the lasso regression result. The output of the lasso regression are summarized int the table below.

Figure 5: Coefficients of Features Related to Host Behavior from Lasso Regression



City	Cluster	review_scores_rating	host_response_time	acceptance_rate	response_rate	Instant Bookable
New York	Cluster0	1.14	3.26	4.91	0	0
	Cluster1	4.23	0	0	4.94	0
	Cluster2	0	0.04	1.48	0.74	0.16
	Cluster3	3.18	0	8.05		2.2
Bankok	Cluster0	0.25	0.1	0	0.65	1.23
	Cluster1	287.41	0	0	22.48	0
	Cluster2	0	1.83	7.31	0.97	0
	Cluster3	48.04	60.15	19.13	9.38	0
Sydney	Cluster0	0	4.7	0	3.74	0
	Cluster1	0	2.25	4.06	0	2.94
	Cluster2	0	5.38	2.17	0.84	0
Paris	Cluster0	0	0.15	1.43	1.81	0
	Cluster1	0	0	0	0	1.23
	Cluster2	1.59	1.55	1.79	0	0

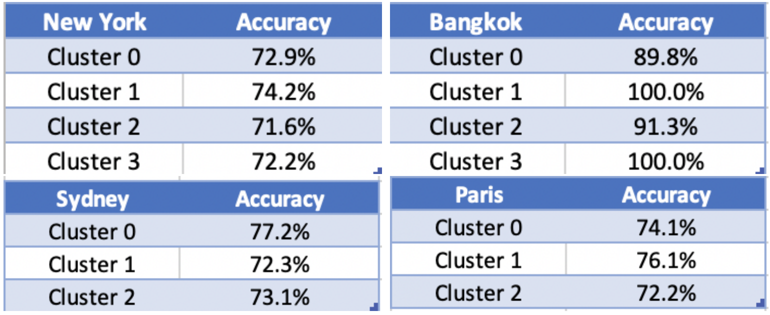
Compared to number of bedrooms and accommodates whose coefficients are around 50, the magnitude of the coefficients of review score and host response time and acceptance rate are very small in the Lasso regression in each cluster and in each region besides the cluster 3 in Bankok. Therefore, we can conclude that the review score and host behavior don't have much impact on the price of housing and the review system is not efficient for Airbnb.

5.4 Random Forest

Besides regression models, We also implemented a random forest model to predict the expected price range. The advantage of a random forest model is that due to the bootstrapping method the model uses, it will alleviate overfitting by taking the majority vote of multiple trees. Therefore, it will help us obtain a higher accuracy prediction result. When training the model, we first converted the prices in training set from a numeric variable to an ordinary variable by changing it to 'low price', 'medium price' and 'high price' based on the quantile statistics. For example, if the price is smaller than or equal to 25% quantile, then it is counted as low price room; if the price is between 25-75 % quantile, then it is considered in the 'medium price' class. Otherwise, it is in the high price category.

After data preparation, we used random search to tune the hyperparameters of maximum depth of each tree, maximum features and criterion. The final training result for each cluster and city is as follows:

Figure 6: Random Forest Accuracy Result



New York	Accuracy
Cluster 0	72.9%
Cluster 1	74.2%
Cluster 2	71.6%
Cluster 3	72.2%

Bangkok	Accuracy
Cluster 0	89.8%
Cluster 1	100.0%
Cluster 2	91.3%
Cluster 3	100.0%

Sydney	Accuracy
Cluster 0	77.2%
Cluster 1	72.3%
Cluster 2	73.1%

Paris	Accuracy
Cluster 0	74.1%
Cluster 1	76.1%
Cluster 2	72.2%

The overall accuracy of New York and Sydney are around 70%, but the validation accuracy is high for Bangkok because there are only few points in cluster 1 and 3 for Bangkok city. Overall, the model accuracy is higher than the regression models.

In order to test the impact of review system on price, we checked the feature importance based on feature permutations. The permutation feature importance computes the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the price, thus the drop in the model score is indicative of how much the model depends on the feature [2]. The result is shown as graphs in the appendix page.

According to the feature importance graphs, we can see that for all cities, on average, the most important features are the number of bedrooms, the number of customers the room can accommodate and location(longitude, latitude). Comparing with these features, the importance score of reviews is much lower. Therefore, we conclude that the review system does not have the biggest impact on the price of Airbnb.

6 Conclusion

It is important for any platform business to have a credible review system to maintain the equilibrium between the buying and selling parties (guests and hosts for home-sharing platform). Airbnb listing data for New York, Paris, Bangkok and Sydney were selected to examine whether the review system has an influence on how prices are set. Since multiple property tiers are likely to exist within each city, k-means clustering was first performed to divide properties into groups for further modeling. The properties in the same tier have similar features such as the number of bedrooms and the number of properties that the host has.

Then regression and classification models were performed to predict the price of housing and to investigate whether the review system is efficient. The first model was polynomial regression with 5-fold cross validation to select the optimal polynomial degree. However, since there are too many categorical features, the optimal polynomial degree is always 1. Therefore, the Lasso regression was used to select the features that are correlated with the housing price and then we ran the polynomial regression on selected features. The GridSearch and 5-folded cross validation were utilized to select the optimal polynomial degree and degree of regularization. From the feature importance of Lasso regression, we concluded that either the review score or the features related to the host behavior have an impact on the price of housing. Besides the polynomial regression, we also tried random forest to see the feature importance based on feature permutations. From the feature permutations, the most important features are the properties of room instead of the review score, leading to the same conclusion that the review system doesn't influence the price of Airnbnb much.

Airnbnb which serves as a two-sided market platform should ensure that the house price reflects the customer feedback. If the house price is extremely high but the customer review is very bad, then the high price will attract less booking and customers feel the property they are looking is over-priced. The review system can balance the over-pricing and customer's feedback, producing a conductive booking environment. The contribution of the report is that Airbnb lacks a review system which is something they can improve on in the future.

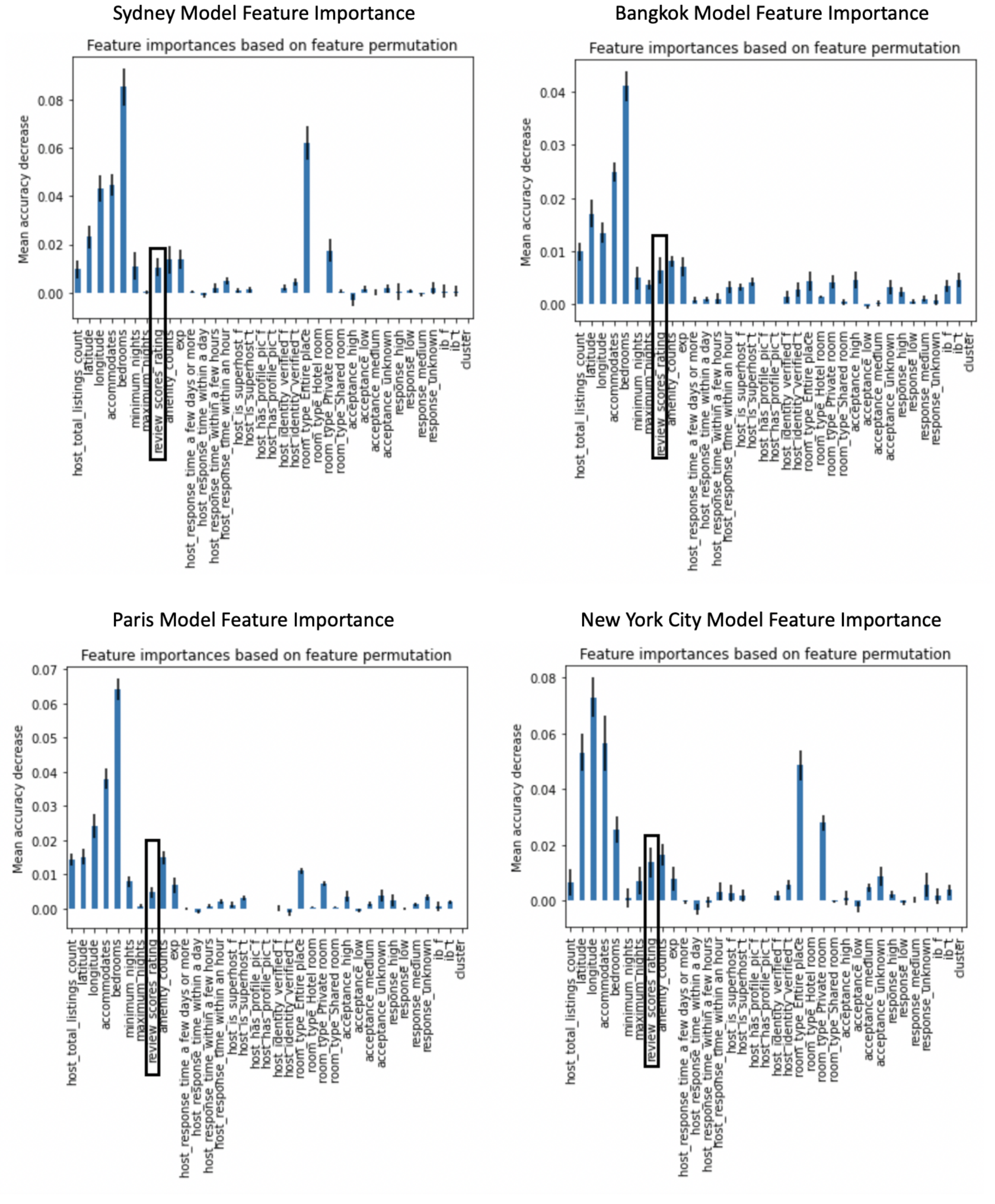
References

https://www.kaggle.com/mysarahmadbhat/airbnb-listings-reviews.

https://scikit-learn.org/stable/auto$_examples/ensemble/plot_forest_importances.html$

Kavyazin, D. (2019, Feb 20). Principal Component Analysis and k-means Clustering to Visualize a High Dimensional Dataset, from https://medium.com/@dmitriy.kavyazin/principal-component-analysis-and-k-means-cluster ing-to-visualize-a-high-dimensional-dataset-577b2a7a5fe2

Figure 7: Random Forest Feature Importance



Sydney Model Feature Importance

Bangkok Model Feature Importance

Paris Model Feature Importance

New York City Model Feature Importance