Springboard—DSC
Capstone Project 1
Predicting the Nightly Price of Toronto's Airbnb Listings

Final Report

George Tang

September, 2019

Table of Contents

List of Tables

List of Figures

1 Introduction

Airbnb, Inc. is an online marketplace and hospitality service brokerage company. What began as an idea of putting an air mattress in the living room and turning it into to bed-and-breakfast in 2007 "for a few bucks" has grown into an international business with annual revenue of over $2.6 billion in 2017. Members (hosts) uses the company's platform to list their properties to provide accommodation services, in which Airbnb receives commissions from each booking.

The biggest decision that hosts need to make is setting the prices for their listings. Hosts marking their prices too high or too low may risk driving potential customers away or shortchanging themselves. On the other hand, hosts that set prices based on the properties' locations and features along with competitors' prices can fully leverage the properties' true value and maximize their revenues.

While hosts can search the Airbnb's listings to get a reference rate, it is time consuming and often it is difficult to identify properties with features like the hosts' in their vicinity. In this project, we will analyze the historical listing information with data analytics, from which factors that affect price will be identified. We will also build machine learning models to predict the listing prices based on input such as host information, properties' features, booking policy, etc.

1.1 Objective

The objectives of is project are:

- To explore and analyze Airbnb's listings in Toronto, Canada
- To identify features that affect the prices of a nightly stay
- To develop machine learning models that predict the prices of a nightly stay based on relevant features
- Provide recommendations to hosts to increase their revenues

This report is divided into the following sections:

- Section 2: Data set description
- Section 3: Data cleaning and wrangling process
- Section 4: Data exploration and statistical analysis
- Section 5: Machine learning model development and evaluation
- Section 6: Recommendation to Airbnb hosts, and suggestions for future work

The programming codes used for this report can be found here.

1.2 Significance

By thoroughly examine Toronto's Airbnb market dataset, we will identify the important features that affect pricing, which the hosts can use as a reference to modify their properties or booking policy. We will also develop machine learning models that can be used by the hosts to set fair and competitive prices.

2 Dataset

2.1 Airbnb Listings Data

The dataset is obtained from the website Inside Airbnb. It is an independent, non-commercial website that allows users to explore how Airbnb is used in cities around the world. The dataset used in this project, referred to as "listings" thereafter, was collected on June 4, 2019.

The listings dataset consists of 20,769 listings (row), and 106 features (columns). Each row consists of a listing in the Greater Toronto area on June 4, 2019.

The features are divided into the following 8 categories:

1. Host information
2. Geographical information
3. Property information
4. Booking information and policy
5. Availability
6. Reviews
7. Airbnb listing information
8. Web scraping information

A list of the features is shown in

Appendix I. Most of the feature names are self-explanatory.

2.2 Toronto Geographical Information

Two Wikipedia pages (here and here) provide the information that links the listings' postal codes to their city names. The use of this information will be discussed in Section 3.4.1.

2.3 Mapquest API

As discussed in Section 3.4.1, some listings come with ambiguous geographical information. As such, their geographical information are obtained through Mapquest's API with their latitudes and longitudes as input.

3 Data Cleaning and Wrangling

The purpose the data cleaning and wrangling steps are:

- To ensure the all features are of the correct data type
- To ensure missing data are properly imputed
- To create potentially useful features
- To prepare the dataset for EDA and statistical analysis

3.1 Data Type Correction

The numerical features price, security deposit, cleaning fee, charge for extra people and host response rate are stored as string in the dataset, and as such, their date types are converted to numeric.

The datetime features last scraped, host since, calendar last scraped, first review and last review are stored as string in the dataset, and as such, their data types are converted to datetime.

3.2 Incorrect Price Data Elimination

The listings price is the focus of this study, and as such, its integrity is of utmost importance. There are four (4) listings with price of 0, which is unreasonable. Further investigation of the listings' website reveal that the prices are non-zero, which indicates data quality issue. Those listings are dropped from the dataset.

3.3 Missing Values Imputation

A list of features with missing values is shown in Appendix II

List of features with missing values. Overall, 54 of the 106 features consist of missing values, with counts from 1 (0.005%) to 20,769 (100%).

Only features that are considered potentially useful for data analysis will be imputed. For imputation of numeric feature, a binary feature with name *[variable]_NA* is created, where 1 and 0 represents missing and non-missing values, respectively. It may be useful if the reason for missing is systemic.

3.3.1 Numeric Features

The numeric features host listings count, number of bathrooms, host response rate, bedrooms and beds are imputed with their respective medians.

Missing security deposit and cleaning fee are due the hosts' decision to not include one, which is equivalent to a value of 0. As such, the missing values will be imputed with 0.

For review scores, the missing values are likely due to the facts that either the listings are new with few customers, or their customers did not leave a review score. The missing values are imputed with the feature median values.

3.3.2 Categorical Features

The categorical features host response time and superhost status are imputed with a new category "missing".

3.3.3 Datetime Features

The datetime features host since, first review, and last review are imputed with feature medians.

3.4 New Feature Creation

3.4.1 City Names

The dataset consists of two features, namely neighbourhood and cleansed neighbourhood, that provide the name of the neighhourhood for each listing. There are 140 unique values for each feature, which may be too granular for data analysis. Instead, grouping the listings with their city names may be more appropriate. The information is obtained with the feature zipcode, which is the postal code of the listing. Canada's postal code consists of six characters, where the first three characters are known as the Forward Sortation Area (FSA). The FSA is then matched with the list of cities discussed in Section 2.2.

For listings with erroneous FSA or unknown city names, their city names are obtained with Mapquests API as discussed in Section 2.3, with the listing's longitude and latitude information as input.

After these steps, there are two listings with missing city information which are removed. Additionally, the number of listings for the cities of Thornhill (7), Mississauga (2), Pickering (2), and Markham (1) are unusually low. Since all of those are cities of considerable size, it is likely that most of the listings of those cities are in other datasets, with only a small fraction of them included in this dataset. It makes the information of those cities non-generalizable. As such, we decided to remove the listings from these cities.

3.4.2 Indicator Variable for Amenities

The feature "amenities" contains a list of attributes provided by the host that the property contains. To further evaluate those amenities, a binary feature is created for each amenity, with '1' and '0' indicating the presence and absence of that amenity in a listing, respectively.

A list of amenities along with the percentage of listings with each amenity is shown in Appendix III. In total, there are 196 unique amenities. The rarest amenities are tennis court, brick oven, pool toys and hammock, each only available in 1 listing, while the most common amenities are wifi, heating, essentials, and smoke detector. Note that the amenities information should be used with caution because the information may not be complete. For instance, it is expected that hot water is provided in most listings; yet, it is only available to 59% of listings. It is possible that some amenities are so trivial that hosts did not bother to include them.

3.4.3 Host Verification

The feature "host_verification" consists of the methods using which a host is verified. In total there are 13 methods, including email address, phone number, and facebook id, among others. We create a new feature to capture the total number of methods through which a host is verified.

3.4.4 Days since Reference Day

The number of days since the recorded events can be a feature more useful than the dates. A reference date of 2019/6/27, which is the date the listings data was scrapped, is chosen.

4 Data Visualization and Analysis

4.1 Price

As summarized in Table 4-1, the nightly prices of show considerable variation. While the least expensive listing is $13, the most expensive listing is $13,422. This extraordinary offer is an "Art Collector's House" that "will have you living luxuriously just steps from Toronto's most stylish neighbourhood".

Table 4-1 Summary of price and log price distribution

Mean	143.35	4.66
Standard Deviation	234.24	0.72
Minimum	13.00	2.56
25%	64.00	4.16
50%	101.00	4.62
75%	160.60	5.08
Maximum	13,422	9.50

Since creating a machine learning model is one of our goals, the presence of outliers would reduce the generalizability of the models and reduce their performance. We therefore decided to cap the price with its 99th percentile value, which is $750. Listings with prices above this value will be excluded for further analysis.

Both the histogram (Figure 4-1(a)) and Q-Q plot (Figure 4-1(b)) show that, even after capping, the price is still highly right skewed. To reduce the influence of the skewness on the subsequent statistical analysis, we apply log transform to the price. As shown Figure 4-2, log price is relatively normal.

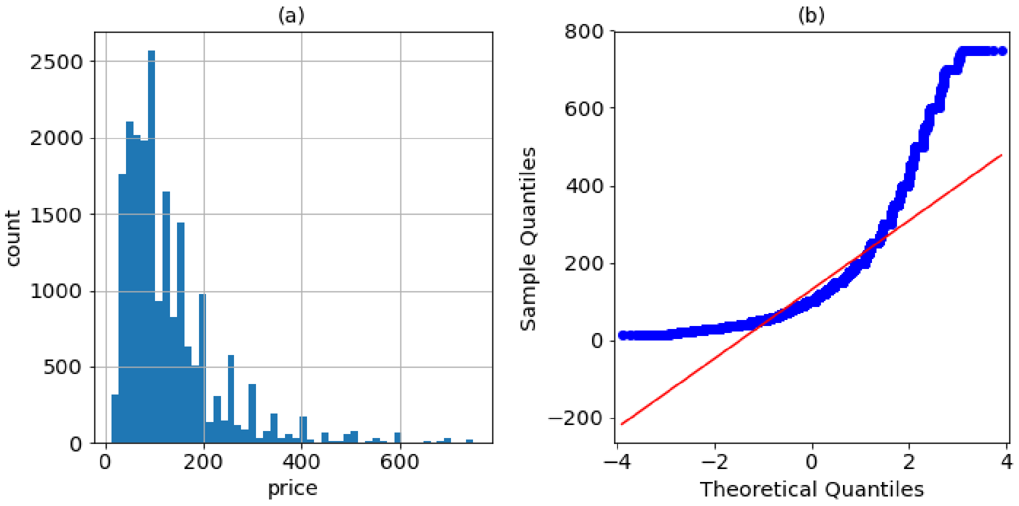


Figure 4-1 (a) Histogram and (b) Q-Q plot of price distribution (after capping)

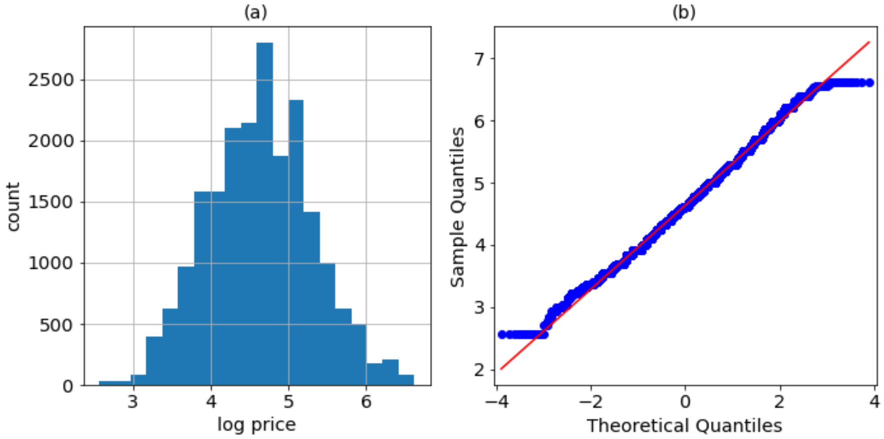


Figure 4-2 (a) Histogram and (b) Q-Q plot of log price distribution (after capping)

4.2 Geographical Information

The count of listings in each region is shown in Figure 4-3(a). Overall, Downtown Toronto has the most listings (39%), and the Toronto region (Downtown, Central, East and West) consists of the most combined listings. The York region (York, North York and East York) has the 2nd most combined listings, while the suburbs Scarborough (northeast of Toronto) and Etobicoke (west of Toronto) have the fewest listings.

The listings prices (Figure 4-3(b)) show considerable variation at each region. Downtown Toronto has the highest median price while Scarborough has the lowest median price.

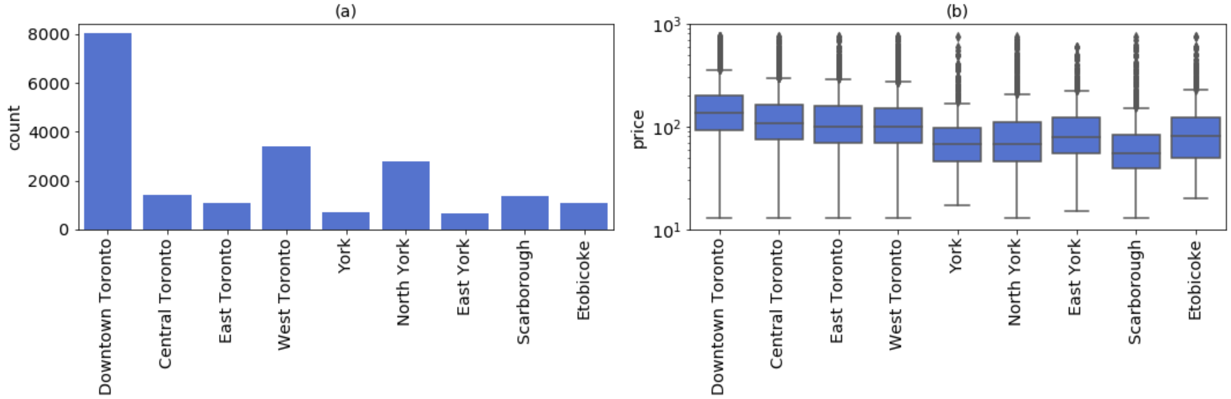


Figure 4-3 (a) Count of listings and (b) boxplot of price (log scale) in each region

4.3 Property Information

4.3.1 Property Type

The count of listings for each property type is shown in Figure 4-4. Overall, there are 30 different property types, with 16 of them with counts less than 10. The most common are Apartment, followed by Condominium and House.

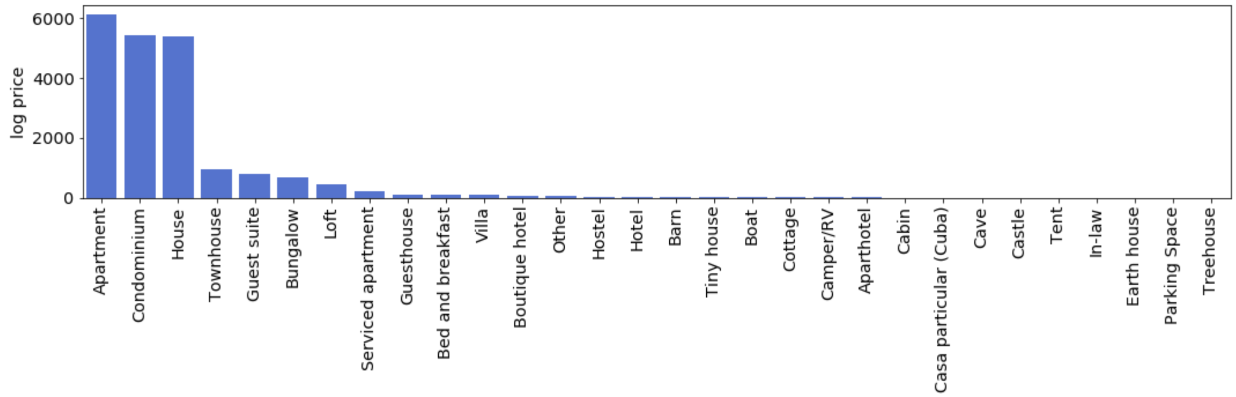


Figure 4-4 Count of listings for each property type

Bungalow is an interesting property type. According to Wikipedia: "Canada uses the definition of bungalow to mean a single-family dwelling that is one storey high". In other words, a bungalow is essentially a house. As such, this property type is assigned a value of "House".

To reduce granularity, property types with counts less than 5% of total listings are assigned a value of "Other". Figure 4-5 shows that condominium has the highest median price while house has the lowest median price.

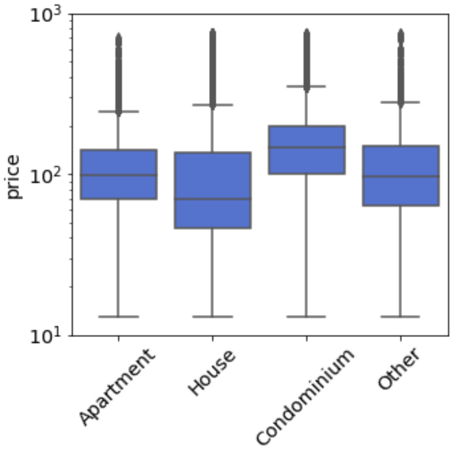


Figure 4-5 Price distribution for each property type

4.3.2 Room Type

The count of room type for each property type is shown in Figure 4-6(a). For houses, the most common room type is private room while for both apartment and condominium, the most common room type is entire home/apartment. Shared room is the least common for all property types.

As shown in Figure 4-6(b), the median price is the highest for the room type of entire home/apartment and lowest for shared room.

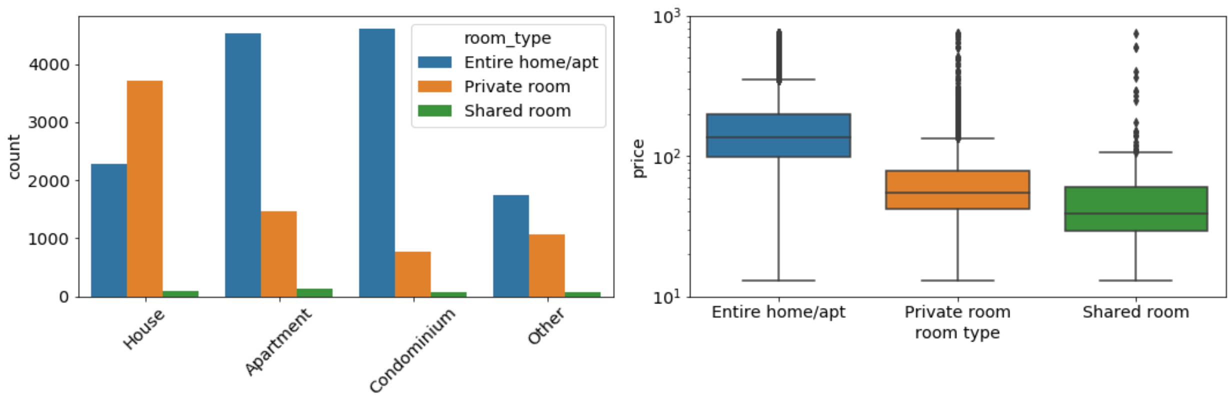


Figure 4-6 (a) Count of room type and (b) price distribution for each property type

4.3.3 Property Features

Features that fall into this category include accommodates, bathrooms, bedrooms, and beds. A heatmap of the correlations between the features and log price is shown in Figure 4-7. All features are somewhat correlated with log price. It is important to notice that the features are highly correlated with each other. It is expected because all these features are related to property size. A bigger property typically consists of more bedrooms, beds, bathrooms and subsequently higher number of accommodates, and vice versa.

High correlation among features can be an issue for linear regression models (Section 5.3) for statistical inference purposed due to an inflated variance of the target variable. For this project, the main goal is to build model with prediction accuracy which is not as affected by this correlation. Also, high correlation does not significantly affect the performance of non-parametric models (Sections 5.4 to 5.6). For simplicity, we will keep all these features when building all the models.

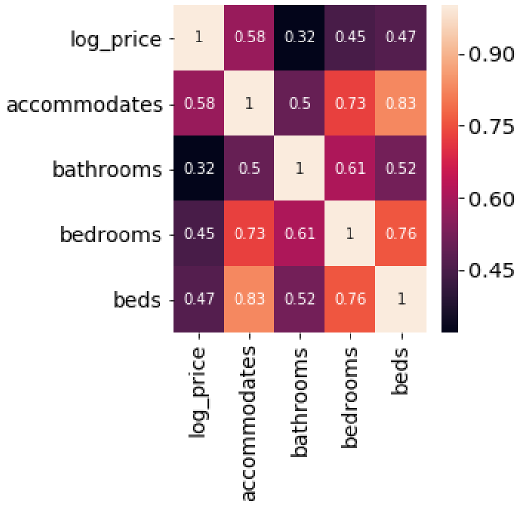


Figure 4-7 Correlations between property features and log price

4.3.4 Amenities

In creating their listings, hosts can detail the amenities available in their properties to attract customers. A total of 196 unique amenities is shown in the dataset. For this analysis, we pick the ones that (1) are intuitively non-trivial and may affect the price, and (2) have a balanced class, with we define as the majority class being less than 90% of the listings. The chosen amenities are “bathtub”, “pets allowed”, “pool”, “gym”, “family/kid friendly”, “private entrance”, “free parking on premises”, and “air conditioning”.

As shown in Figure 4-8, for the eight chosen amenities, the median price is higher for listings with the amenity.

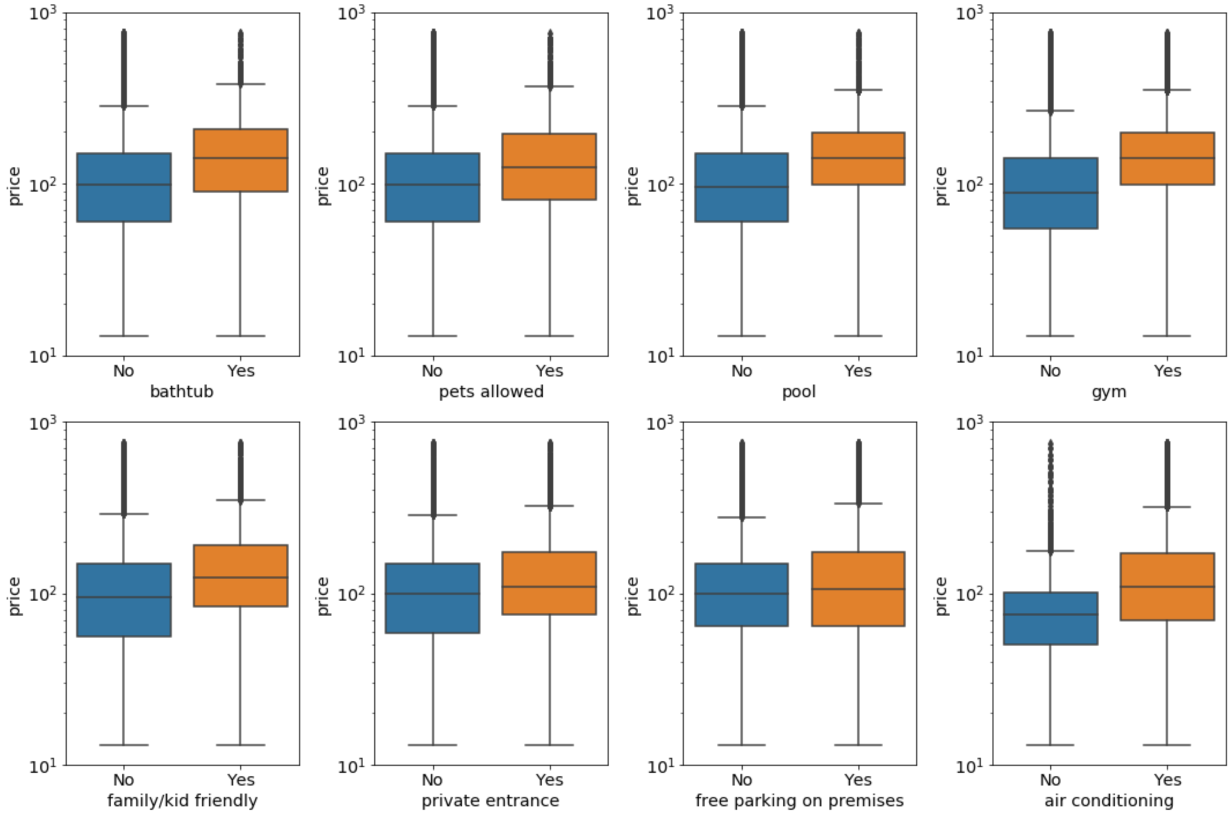


Figure 4-8 Distribution of log price for listings with and without amenities

4.4 Host Information

4.4.1 Superhost Status

According to Airbnb: “[s]uperhosts are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests”. As such, it is possible that a superhost status would command a higher price due to their good track records and reputations.

As shown in Figure 4-9(a), about 26% of listings are provided by superhosts. Overall, the median price is only slightly higher for listings provided by superhosts (Figure 4-9(b)). Note that we excluded the listings with missing superhost status because of its small proportion (< 1%).

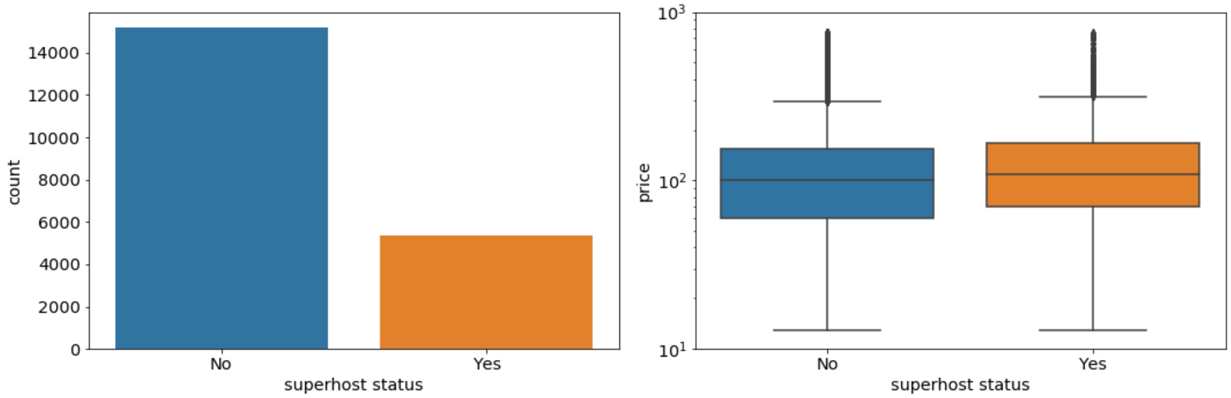


Figure 4-9 (a) Count of listings and (b) distribution of log price for superhost status

4.5 Booking Policy

4.5.1 Cleaning Fee

Cleaning fee is a one-time, non-refundable fee charged by the host, regardless of the duration of stay. It is an interesting feature because it is part of the revenue from each booking and can be used as a pricing strategy. From guests' point of view, long-term guests may not mind a higher cleaning fee if the nightly rate is reasonable. On the other hand, short-term guests may find listings with cleaning fee relatively minimal to the nightly rate more attractive. Nonetheless, for this project, cleaning fee will be used solely as a predictor for price.

The plot of price vs log cleaning fee is shown in Figure 4-10. For visualization, log transformation is applied to cleaning fee due to its high skewness. The value of 1 is added prior to the transformation to take care of the 0 values. As shown in the figure, there are two distinct populations of cleaning fee. On the left, the cleaning fee is 0, where on the right, there is a positive correlation between log price and log cleaning fee.

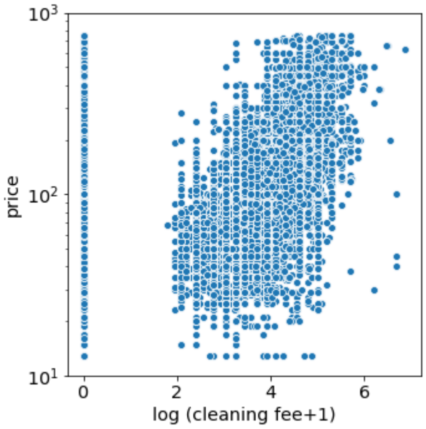


Figure 4-10 Log price vs log cleaning fee

4.5.2 Minimum Nights of stay

As shown in Figure 4-11(a), most listings require minimum nights of stay of 3 nights or less. There is no obvious difference among them in term of price (Figure 4-11(b)).

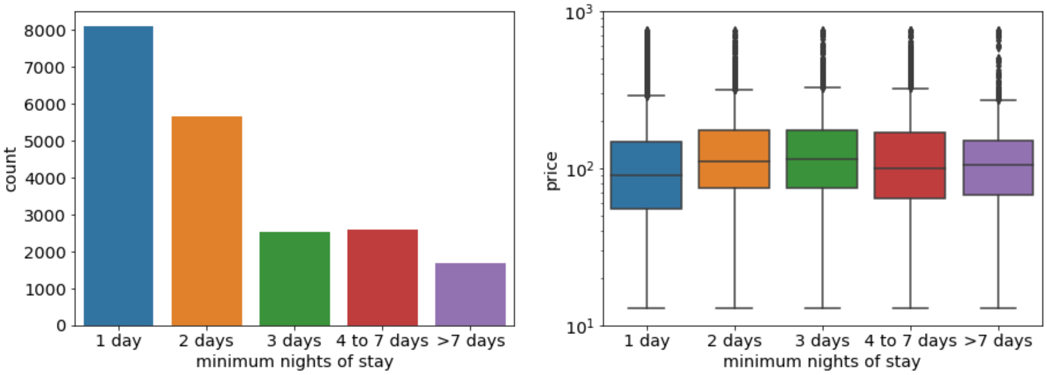


Figure 4-11 (a) Count of listings and (b) distribution of log price for minimum nights of stay

4.5.3 Other numeric Features

Features that fall into this category include security deposit, cleaning fee, included number of guests, charge for extra people, minimum and maximum nights of stay.

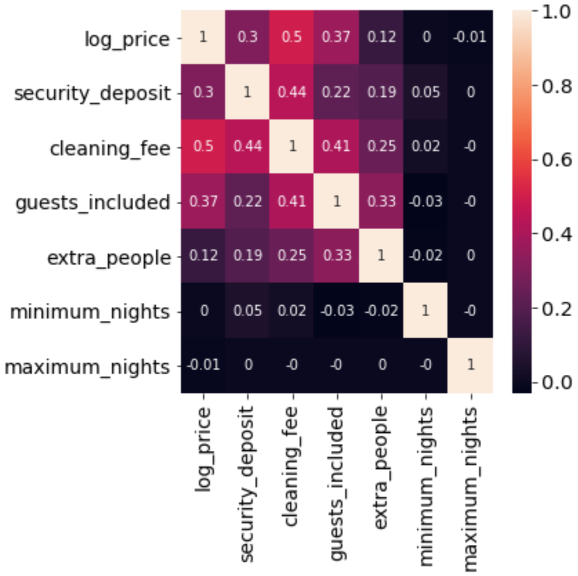


Figure 4-12 shows that security deposit, cleaning fees and included number of guests are correlated with log price.

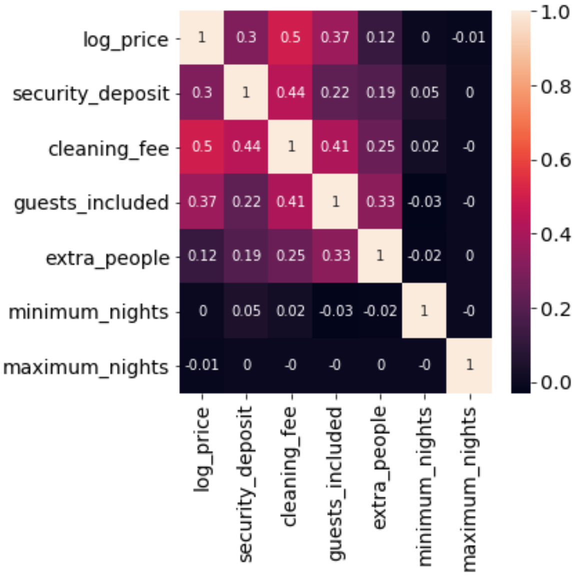


Figure 4-12 Correlations between booking policy features and log price

4.6 Availability

The dataset provides information on the listings' availabilities for the next 30, 60, 90 and 365 days. Figure 4-13(a) shows that most of listings have no availability for the next 30 days. Figure 4-13(b) shows that while the availabilities are highly correlated with each other, they are not correlated with log price. It is important to note that, however, features not significantly correlated with log price does not imply that they are not useful in explaining price, since interactions among features are not considered.

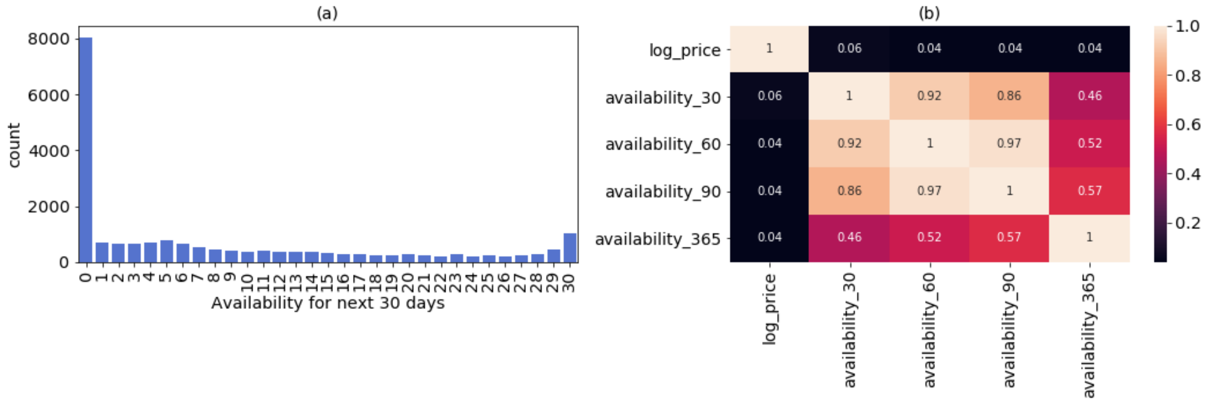


Figure 4-13 (a) Availability count of next 30 days; (b) Correlation between availability and log price

4.7 Review Scores

Airbnb allows customers to provide review scores of their experiences on various categories, including accuracy, cleanliness, check in, communication, and value. The number of reviews received by each listing, and the number of reviews of the last twelve months (ltm) are also included. Figure 4-14(a) shows that most the review are scores are close to the full score of 100, with few below 80. Figure 4-14(b) shows that there is not significant difference in price distribution at the different score groups. is not the neither the number of reviews nor review scores are correlated with log price. Figure 4-15 shows that there is no significant correlation between review score information and log price.

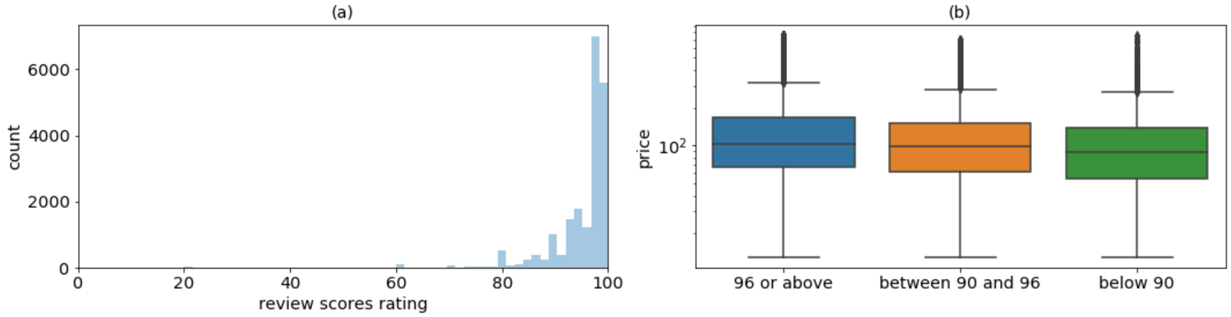


Figure 4-14 (a) Review scores rating histogram; (b) Price distribution by review score groups

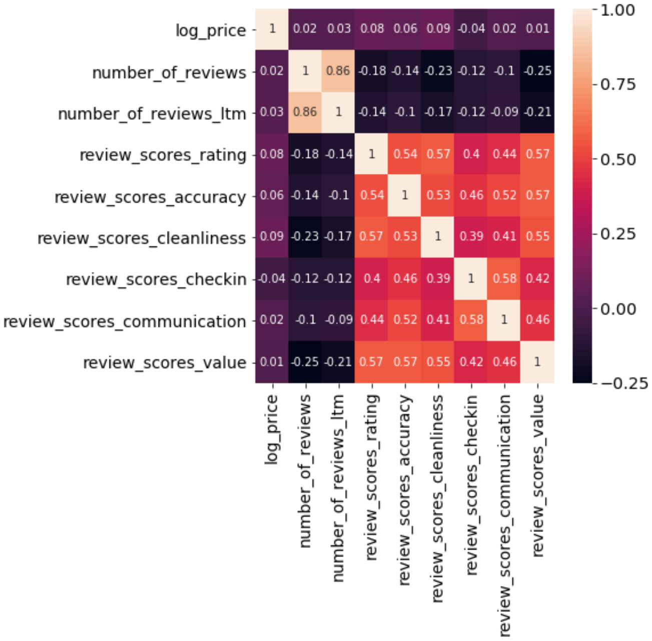


Figure 4-15 Correlation between review score information and log price

4.8 Statistical Analysis of Categorical Features

In this section, we will conduct hypothesis tests on selected categorial features. The null hypothesis is that the distributions of log price are the same among all categories within a feature, where the alternative hypothesis is that the distributions are not the same. Since the log price is not normal, the Mann Whitney U (MWU) test will be used for features with binary response (1/0, or t/f, etc), where Kruskal-Wallis (KW) test will be used for multi-category response. The hypothesis tested by both MWU test and KW test is whether the samples from the different categories are taken from the same population. If they are not, then the feature may useful for price prediction.

The results of the tests are shown in

Appendix IV. For almost all features, the null hypothesis is rejected, i.e. log price does not have time same distribution among all categories in each feature. It is important to note that, however, with such a large sample size (> 10,000), the null hypothesis will almost always be rejected. As such, this hypothesis testing may not create the most useful insights. A better approach would be to use machine learning models to identify features that are useful for predicting nightly price which will be the focus of the next Section.

5 Machine Learning

In this section, we will examine various machine learning models, evaluate their performances, recommend the best model, and identify the most important features that affect price.

Log price will be used as the target variable used in this study. Models built with log price as compared to price as the target has a few unique features. First, errors in predicting expensive listings and cheap listings affect the performance metrics equally. Also, with log transformation, the predicted price will always be non-negative, which is important for linear regression models.

For this regression problem, we examine the most widely used models, namely linear regression, random forest, gradient boosting and extreme gradient boosting (XGB).

5.1 Data Preprocessing

Data preprocessing consists of two steps. First, the data is split into the training and test datasets with 75/25 split. Then, the training data is normalized, i.e., for each feature, the mean and standard deviation are transformed to 0 and 1, respectively. The normalization parameters determined by the training dataset are then applied to normalizing the test data.

5.2 Performance Metrics

The performance metrics considered in this study are R^2, root mean squared error (RMSE), and mean absolute percentage error (MAPE).

R^2 is a measure of the proportion of data variation explained by the model.

RMSE is the square root of the average of squared difference between the actual value and the predicted value.

MAPE is the average of the absolute percentage difference between the actual value and the predicted value.

5.3 Linear Regression

Linear regression is the simplest and widely used regression approach. It is a parametric model, which assumes a linear relationship between the features and the target variable. Mathematically, the model is written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

where $\hat{y}$ is the predicted target value, $x_1 \ldots x_k$ are the features, β_0 is the intercept, and $\beta_1 \ldots \beta_k$ are the coefficients to be estimated.

A linear regression model is created with all the features with no interactions (i.e. no $x_i x_j$ terms). The model is then used to make predictions for both the training data and the test data, with the performance metrics summarized in Table 5-1. Since the two sets of metrics are similar, the model does not overfit the training data. The test metrics are used as the benchmark for the more advanced model as discussed in the coming sections.

Table 5-1 Performance metrics for linear regression model on training and test data

	R^2	RMSE	MAPE (%)
Training data	0.685	0.380	6.36
Test data	0.654	0.396	6.54

The diagnostic plots of the model's performance in prediction with the test data are shown in Figure 5-1. While the residuals are normally distributed, the residual shows an increasing trend with log price, which indicates the model tends to underpredict the price as price increases. Also, at the low end of price, there are outliers where the model significantly overpredicts.

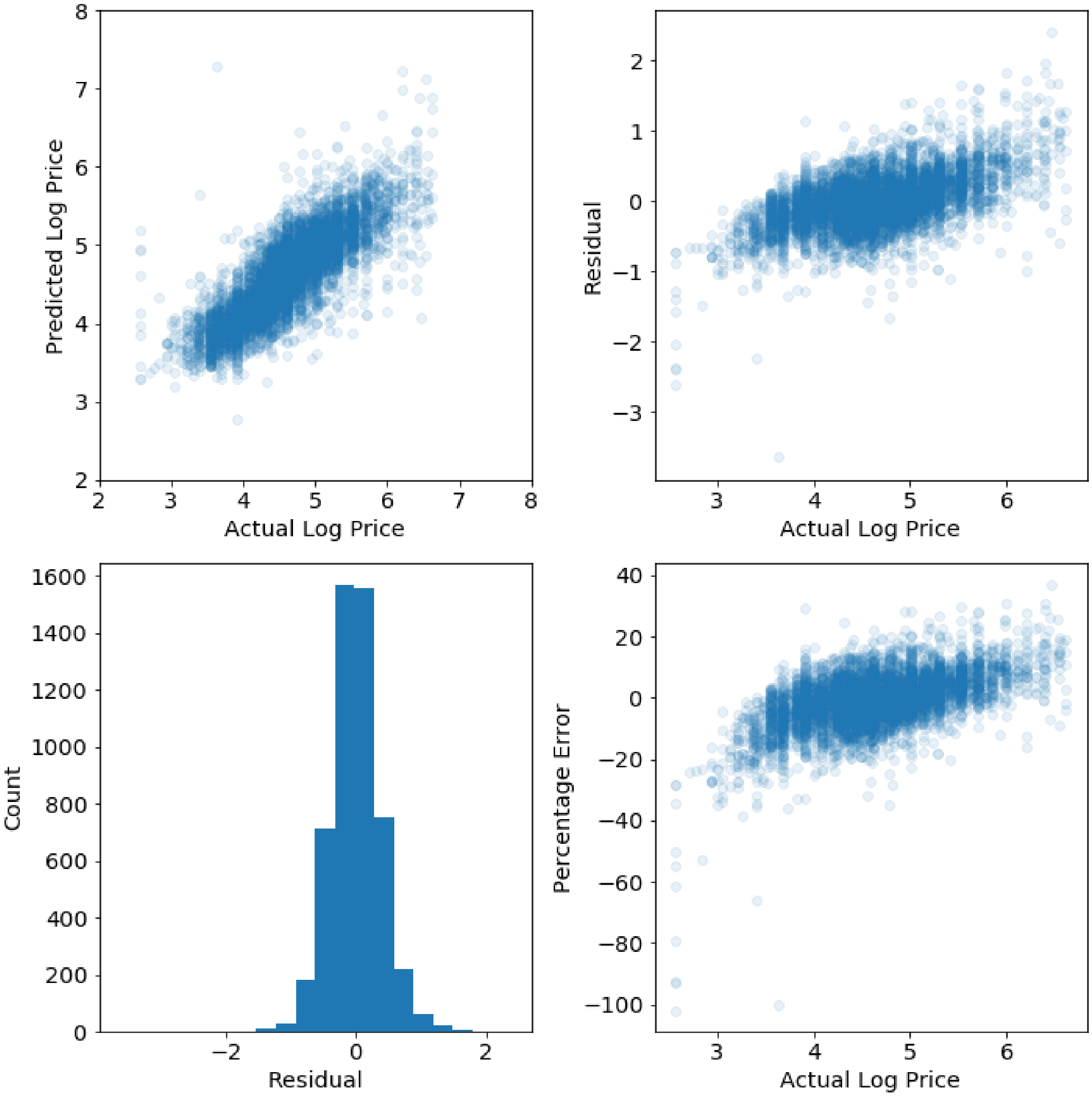


Figure 5-1 Diagnostic plots for the linear regression model with the test data

5.4 Random Forest Regressors

Random forests are a family of non-parametric ensemble models. In these models, a specified number of regression trees are created, each with a subsample of features and data. The prediction is then made with averaging the prediction results of each tree.

Model tuning is used to determine the optimal hyperparameters for the best model performance. Typically, it is done with cross validation (which will be covered in the next section). For random forest, however, we make use of the unique feature known as out-of-bag (OOB) scoring. Since only a subset of all data is used to build each tree, unused data can be used to validate the performance for that tree.

The two hyperparameters considered are: the number of trees (estimators), and the proportion of features selected for each tree. As shown in Figure 5-2, the options for feature proportions are (1) “none”: all features are considered to build each tree: (2): “log2”, only log$_2$(n) of features are randomly chosen to build each tree (n is the number of features), and (3) “sqrt”: only the square root of n of features are randomly chosen to build each tree. The figure shows the plot of the out-of-bag R^2 score at different number of estimators. At each number of trees, the model is evaluated with OOB samples, i.e., samples not used to build that tree. The figure shows that the best feature selection is “none”, i.e. all features are used for each tree, as it consistently has the highest OOB R^2 score. Also, while performance increases with the number of trees, no significant improvement is shown above around 300. As such, we build the random forest model with unlimited feature selection and number of estimators of 300.

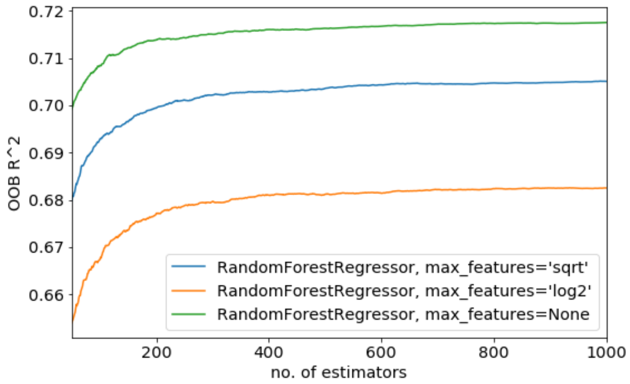


Figure 5-2 Random forest model tuning result

Table 5-5 shows that the random forest model is a significant improvement to the linear regression model. The diagnostics plots for the model is shown in

	test	number of categories	rejected_null_hypothesis
property_type_simple	Kruskal-Wallis	4	yes
room_type	Kruskal-Wallis	3	yes
bed_type	Kruskal-Wallis	5	yes
city_fsa	Kruskal-Wallis	9	yes
cancellation_policy	Kruskal-Wallis	4	yes
instant_bookable	MannWhitneyU	2	yes
is_business_travel_ready	MannWhitneyU	1	NA

host_response_time	Kruskal-Wallis	4	yes
host_is_superhost	MannWhitneyU	2	yes
host_has_profile_pic	MannWhitneyU	2	yes
host_identity_verified	MannWhitneyU	2	yes
elevator	MannWhitneyU	2	yes
hair dryer	MannWhitneyU	2	yes
pool	MannWhitneyU	2	yes
laptop friendly workspace	MannWhitneyU	2	yes
lockbox	MannWhitneyU	2	yes
paid parking on premises	MannWhitneyU	2	yes
shampoo	MannWhitneyU	2	yes
hot tub	MannWhitneyU	2	yes
air conditioning	MannWhitneyU	2	yes
luggage dropoff allowed	MannWhitneyU	2	no
safety card	MannWhitneyU	2	yes
carbon monoxide detector	MannWhitneyU	2	yes
refrigerator	MannWhitneyU	2	yes
stove	MannWhitneyU	2	yes
cable tv	MannWhitneyU	2	yes
free parking on premises	MannWhitneyU	2	yes
cooking basics	MannWhitneyU	2	yes
private entrance	MannWhitneyU	2	yes
hot water	MannWhitneyU	2	yes
family/kid friendly	MannWhitneyU	2	yes
patio or balcony	MannWhitneyU	2	yes
bathtub	MannWhitneyU	2	yes
lock on bedroom door	MannWhitneyU	2	yes
washer	MannWhitneyU	2	yes
internet	MannWhitneyU	2	yes
private living room	MannWhitneyU	2	no
paid parking off premises	MannWhitneyU	2	yes
dishes and silverware	MannWhitneyU	2	yes
fire extinguisher	MannWhitneyU	2	yes
tv	MannWhitneyU	2	yes
dishwasher	MannWhitneyU	2	yes
first aid kit	MannWhitneyU	2	no
microwave	MannWhitneyU	2	yes
iron	MannWhitneyU	2	yes
oven	MannWhitneyU	2	yes
host greets you	MannWhitneyU	2	yes
coffee maker	MannWhitneyU	2	yes
no stairs or steps to enter	MannWhitneyU	2	yes
bed linens	MannWhitneyU	2	yes

gym	MannWhitneyU	2	yes
long term stays allowed	MannWhitneyU	2	yes
pets allowed	MannWhitneyU	2	yes
extra pillows and blankets	MannWhitneyU	2	yes
garden or backyard	MannWhitneyU	2	yes
dryer	MannWhitneyU	2	yes
translation missing: en.hosting_amenity_50	MannWhitneyU	2	yes
translation missing: en.hosting_amenity_49	MannWhitneyU	2	yes
keypad	MannWhitneyU	2	yes
self check-in	MannWhitneyU	2	yes
hangers	MannWhitneyU	2	yes

Appendix V.

Table 5-2 Performance metrics for random forest and linear regression models on the test data

	R^2	RMSE	MAPE (%)
Random Forest	0.708	0.364	5.87
Linear Regression	0.654	0.396	6.54

Feature importance score indicates how useful a feature is in the construction of a tree, typically measured as in increase in performance metrics such as mean squared error. For ensemble models like random forest, the score for each feature is averaged across all trees. Feature importance for the optimal Random Forest Regressor model is shown in Figure 5-3. The room type (entire home/apt) is by far the most important feature, followed by the number of bathrooms, cleaning fee and location (Downtown Toronto). No amenities make it to the top 20 list.

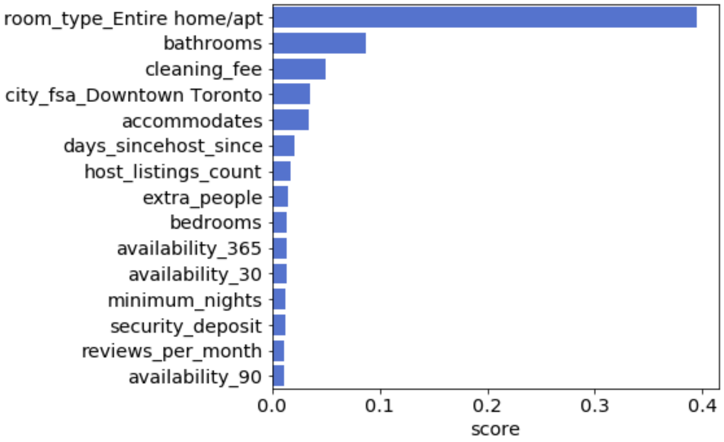


Figure 5-3 Feature importance of random forest model

5.5 Gradient Boosting

Gradient boosting is another family of non-parametric ensemble models. It differs from random forest in that instead of building trees simultaneously, trees are built sequentially with each tree fit to the residuals of the preceding tree. The advantage of this approach is that the predictive power of weaker features can be more fully utilized.

To obtain the optimal hyperparameters for the model, grid search along with cross validation is used. The tuned hyperparameters are learning rate, max depth of trees, and fraction of features used for each tree. The best parameters are: learning rate = 0.1, max depth = 5, number of estimators = 500, features = ‘sqrt’.

Table 5-3 shows that the gradient model is a significant improvement to the linear regression model. The diagnostics plots for the model is shown in Appendix VI.

Table 5-3 Performance metrics for gradient boosting and linear regression models on the test data

	R^2	RMSE	MAPE (%)
Gradient Boosting	0.726	0.352	5.73
Linear Regression	0.654	0.396	6.54

Feature importance for the gradient boosting model is shown in Figure 5-4. Interestingly, the most important features are the duration since the host's first hosting, and the number of days since the first and last review. Intuitively, those are features not related to the listings. Also, compared to the random forest model, the scores are more evenly distributed among features.

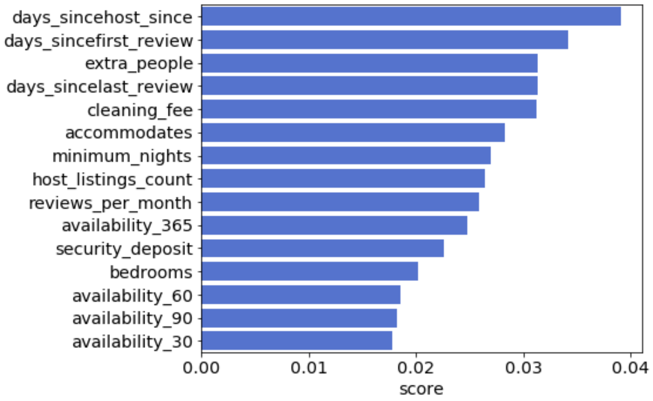


Figure 5-4 Feature importance of gradient boosting model

5.6 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a recent advancement to the gradient boosting model. Essentially, it adds a regularization component to the algorithm to limit the complexity of trees to prevent overfitting.

Model tuning is performed with cross validation along with grid search. The tuned hyperparameters include maximum depth of each tree and fraction of features included for each tree. The best parameters are learning_rate = 0.1, gamma = 0.2 (regularization parameter), max depth = 5, number of trees = 500, and 50% of all features randomly considered for each tree.

Table 5-4 Performance metrics for XGBoost and linear regression models on the test datashows that the XGBoost model is a significant improvement to the linear regression model. Cross-validation results and diagnostics plots for the model is shown in Appendix VII.

Table 5-4 Performance metrics for XGBoost and linear regression models on the test data

	R^2	RMSE	MAPE (%)
XGBoost	0.733	0.348	5.63
Linear Regression	0.654	0.396	6.54

The top 20 most importance features for the XGBoost model are shown in Figure 5-4. The most important features are the room types (entire home/apt, private room). Location (Downtown Toronto) and number of accommodates are also important features. Contrarily to the other models, three amenities (elevator, tv, gym) makes the top 20 list.

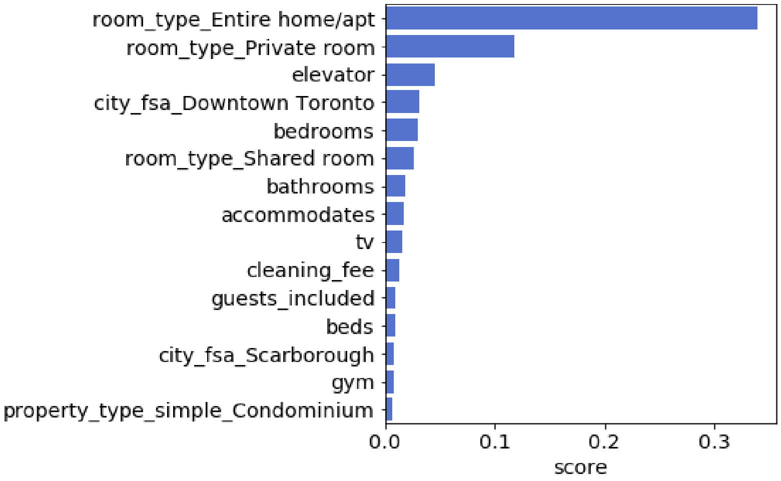


Figure 5-5 Feature importance of XGBoost model

5.7 Model Comparison

The performance metrics for the four models are summarized in Table 5-5. The performances of the advanced models are very close, and all of them shows significant improvement to the linear regression model. Overall, the XGBoost model shows the best result in terms of the three metrics.

Table 5-5 Summary of performance metrics of the machine learning models (log price as target)

	Linear Regression	**Random Forest**	**Gradient Boosting**	**XGBoost**
R^2	0.654	0.708	0.726	0.732
RMSE	0.396	0.364	0.352	0.348
MAPE	6.538	5.868	5.736	5.628

Since the ultimate objective is to predict price, we use the models to predict the price for the test data, with the performance metrics shown in Table 5-6. XGBoost model is the best model. The RMSE is $60.3 and the average absolute percentage error is 27.4%.

Table 5-6 Summary of performance metrics of the machine learning models (price as target)

	Linear Regression	**Random Forest**	**Gradient Boosting**	**XGBoost**
R^2	0.468	0.607	0.638	0.647
RMSE	73.74	63.39	60.81	60.10
MAPE	32.20	28.55	27.79	27.31

A diagnostics plot for actual price prediction with XGBmodel is shown in Figure 5-6. The residual is right skewed, and the model tends to underpredict at higher prices, due to the scarcity of listings at that price range. Also, at the low end of price, the model tends to significantly overpredict the price, with percentage error of up to -1,200% (i.e. predicted price is about 12x the actual price). Those listings are worth further investigation.

Based on this analysis, there is a 95% chance the difference between the actual price and the predicted price (residual) is between -$79 and $152, i.e. overpredicting by $79 or underpredicting by $152.

Also, there is a 95% chance the absolute percentage difference is below 72%.

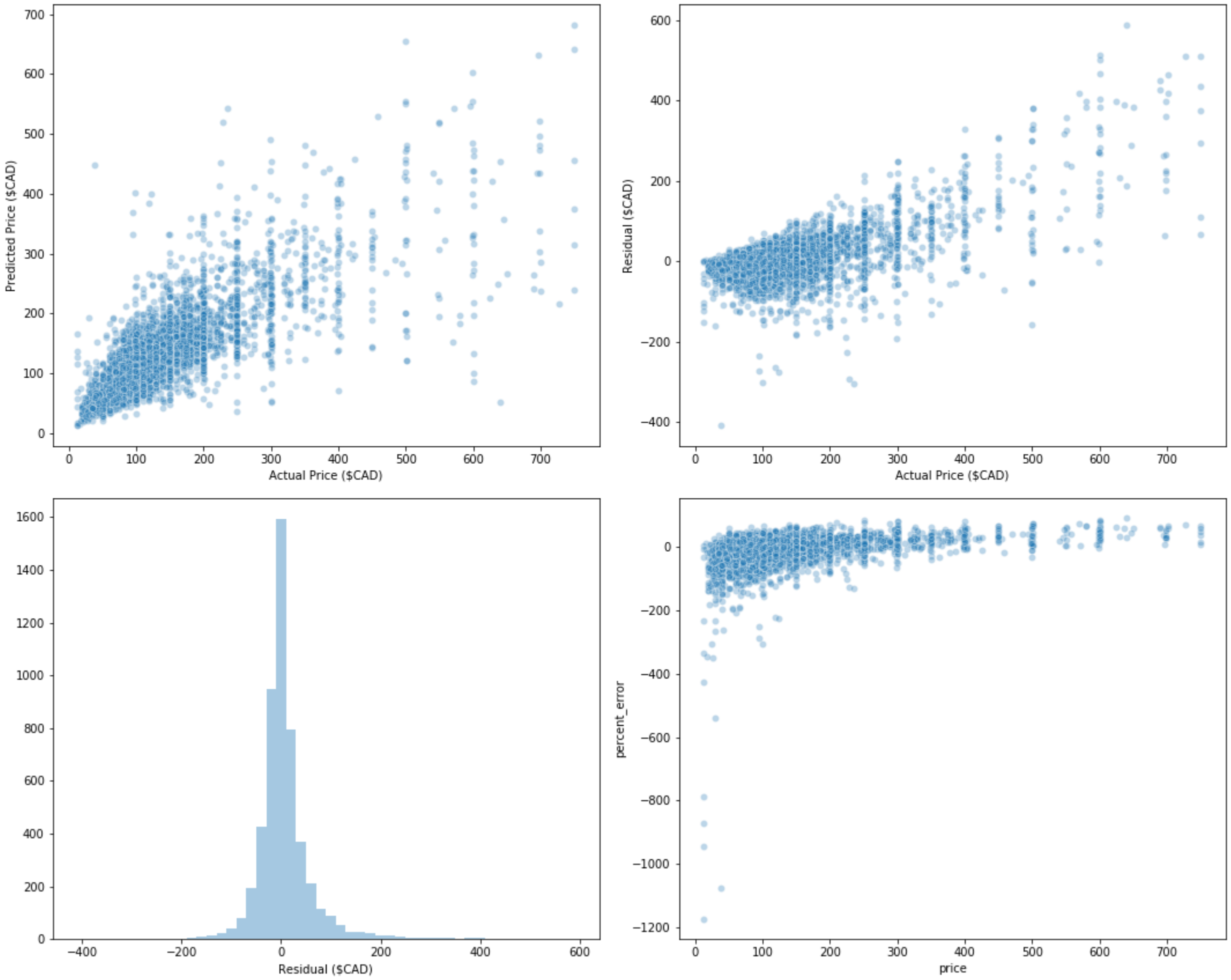


Figure 5-6 Diagnostic plots for XGBoost model (price as target)

6 Recommendations and Future Work

In this project, we thoroughly examined Toronto's Airbnb listings data through visualization and statistical analysis. We have also developed machine learning models that allows hosts to decide on their nightly prices for their listings.

We have the following recommendations for hosts and investors of the Airbnb market:

1. If the listing desired price and price suggested by the model are significantly different, hosts are encouraged to identify the cause of the discrepancies. One approach is to compare their listings to those with prices similar the predicted prices to identify any significant differences among the properties. Hosts can then either raise the price (if priced too low) to increase revenue or lower the price (if priced too high) to increase competitiveness.

2. Based on our best model, room type is the most important factor. Hosts can consider offering the entire house or apartment instead of private rooms or shared rooms to command a higher price.

3. Property characteristics such as the number of bedrooms, bathrooms and beds, which ultimately decide the number of accommodates the property can host, are important features. Hosts can consider upgrading their properties to increase the number of accommodates, which will allow them to set higher price.

We propose the following future works to derive further insight in the market and to develop models with better performances:

Enable listings comparison
As a follow up to Recommendation #1, we can create a product that assist hosts to compare their listings to others based on price, features and/or locations. Hosts can use this information to adjust their prices or modify their properties to increase competitiveness and maximize revenue.

Group listings into different price categories
Given the high variation in prices, it may be appropriate to split the dataset into different prices groups (e.g. economy and luxury groups). This way, we can perform EDA and build machine learning models that are more generalizable to each category.

Apply text analytics to the text features and customer review
Features such as properties' text description may contain valuable information for model development. natural language processing (NLP) can be used for this analysis. Also, NLP can be used to analyse customer's reviews to identify their preferences and what factors contributed to their good experiences. With this information, hosts can improve their properties to make them more attractive, which may ultimately lead to more business and increase revenue.

Understand pricing dynamics
In this project, we only look at listing prices at one specific day. In reality, prices are typically higher during weekends and holidays, and lower during weekdays. Time series analysis can be used to understand pricing dynamics. The information can be used by hosts to adjust their listing prices at different time of year to maximize revenue.

Appendix

Appendix I

Features List

Host information	**Property Information**	**Booking information and policy**
host_id	name	price
host_url	summary	weekly_price
host_name	space	monthly_price
host_since	description	security_deposit
host_location	experiences_offered	cleaning_fee
host_about	neighborhood_overview	guests_included
host_response_time	notes	extra_people
host_response_rate	transit	minimum_nights
host_acceptance_rate	access	maximum_nights
host_is_superhost	interaction	minimum_minimum_nights
host_thumbnail_url	house_rules	maximum_minimum_nights
host_picture_url	street	minimum_maximum_nights
host_neighbourhood	neighbourhood	maximum_maximum_nights
host_listings_count	neighbourhood_cleansed	minimum_nights_avg_ntm
host_total_listings_count	neighbourhood_group_cleansed	maximum_nights_avg_ntm
host_verifications	city	requires_license
host_has_profile_pic	state	license
host_identity_verified	zipcode	jurisdiction_names
calculated_host_listings_count	market	instant_bookable
calculated_host_listings_count_entire_homes	smart_location	is_business_travel_ready
calculated_host_listings_count_private_rooms	country_code	cancellation_policy
calculated_host_listings_count_shared_rooms	country	require_guest_profile_picture
	latitude	require_guest_phone_verification
	longitude	
	is_location_exact	
	property_type	
	room_type	
	accommodates	
	bathrooms	
	bedrooms	
	beds	
	bed_type	
	amenities	
	square_feet	

Availability	**Airbnb listing information**	**Reviews**	**Web scraping information**
calendar_updated	id	number_of_reviews	scrape_id
has_availability	listing_url	number_of_reviews_ltm	last_scraped
availability_30	thumbnail_url	first_review	
availability_60	medium_url	last_review	
availability_90	picture_url	review_scores_rating	
availability_365	xl_picture_url	review_scores_accuracy	
calendar_last_scraped		review_scores_cleanliness	
		review_scores_checkin	
		review_scores_communication	
		review_scores_location	
		review_scores_value	
		reviews_per_month	

Appendix II

List of features with missing values

Feature	**Missing Count**	**Missing Percentage**
thumbnail_url	20765	100
medium_url	20765	100
host_acceptance_rate	20765	100
neighbourhood_group_cleansed	20765	100
xl_picture_url	20765	100
jurisdiction_names	20763	99.99036841
license	20761	99.98073682
square_feet	20609	99.24873585
monthly_price	18987	91.43751505
weekly_price	18678	89.94943414
notes	10967	52.81483265
host_about	8757	42.17192391
access	7986	38.45894534
interaction	7716	37.15868047
neighborhood_overview	7291	35.11196725
transit	7086	34.12472911
house_rules	6576	31.66867325
space	5774	27.80640501
host_response_time	4984	24.00192632
host_response_rate	4984	24.00192632
security_deposit	4902	23.60703106
review_scores_location	4287	20.64531664
review_scores_value	4284	20.63086925
review_scores_checkin	4282	20.62123766
review_scores_accuracy	4281	20.61642186
review_scores_communication	4279	20.60679027
review_scores_cleanliness	4279	20.60679027
review_scores_rating	4271	20.56826391
last_review	3982	19.17649892
reviews_per_month	3982	19.17649892
first_review	3982	19.17649892
cleaning_fee	3384	16.29665302
host_neighbourhood	2520	12.13580544
summary	664	3.197688418
zipcode	365	1.757765471
description	346	1.66626535
market	37	0.178184445
state	28	0.134842283

beds	23	0.110763304
host_location	19	0.09150012
bathrooms	15	0.072236937
bedrooms	8	0.038526366
host_identity_verified	5	0.024078979
host_has_profile_pic	5	0.024078979
host_total_listings_count	5	0.024078979
host_listings_count	5	0.024078979
host_picture_url	5	0.024078979
host_thumbnail_url	5	0.024078979
host_is_superhost	5	0.024078979
host_since	5	0.024078979
host_name	5	0.024078979
city	1	0.004815796
neighbourhood	1	0.004815796
name	1	0.004815796

Appendix III

List of Amenities

Amenity	Count	Percentage of Listings
brick oven	1	0.004819
pool toys	1	0.004819
tennis court	1	0.004819
hammock	1	0.004819
washer / dryer	2	0.009638
alfresco bathtub	2	0.009638
mobile hoist	2	0.009638
heat lamps	2	0.009638
private gym	2	0.009638
ground floor access	2	0.009638
pool cover	2	0.009638
ceiling hoist	3	0.014457
private pool	3	0.014457
touchless faucets	3	0.014457
standing valet	3	0.014457
air purifier	4	0.019276
fax machine	4	0.019276
outdoor kitchen	4	0.019276
mountain view	5	0.024095
projector and screen	5	0.024095
private hot tub	5	0.024095
bidet	6	0.028914
steam oven	6	0.028914
sauna	6	0.028914
stand alone steam shower	7	0.033733
private bathroom	7	0.033733
heated towel rack	7	0.033733
jetted tub	7	0.033733
fire pit	8	0.038552
double oven	8	0.038552
beach view	8	0.038552
wine cooler	10	0.04819
mudroom	11	0.053009
amazon echo	11	0.053009
shared hot tub	11	0.053009
high-resolution computer monitor	11	0.053009
ski-in/ski-out	11	0.053009
electric profiling bed	12	0.057829
murphy bed	14	0.067467
warming drawer	15	0.072286

sun loungers	15	0.072286
mini fridge	17	0.081924
hbo go	17	0.081924
shared pool	18	0.086743
firm mattress	19	0.091562
outdoor parking	20	0.096381
printer	21	0.1012
dvd player	21	0.1012
pool with pool hoist	22	0.106019
shared gym	26	0.125295
day bed	29	0.139752
exercise equipment	32	0.154209
heated floors	33	0.159028
gas oven	34	0.163848
shower chair	37	0.178305
kitchenette	39	0.187943
ceiling fan	39	0.187943
bathtub with bath chair	42	0.2024
table corner guards	42	0.2024
fixed grab bars for toilet	46	0.221676
pillow-top mattress	48	0.231314
terrace	50	0.240952
sound system	53	0.255409
formal dining area	54	0.260228
rain shower	54	0.260228
espresso machine	56	0.269867
other pet(s)	57	0.274686
memory foam mattress	57	0.274686
soaking tub	57	0.274686
convection oven	69	0.332514
balcony	69	0.332514
walk-in shower	72	0.346971
central air conditioning	74	0.356609
baby monitor	74	0.356609
outdoor seating	76	0.366247
en suite bathroom	78	0.375885
breakfast table	91	0.438533
beach essentials	92	0.443352
fireplace guards	100	0.481904
fixed grab bars for shower	104	0.501181
smart tv	129	0.621657
beachfront	137	0.660209
netflix	141	0.679485

roll-in shower	145	0.698762
changing table	147	0.7084
window guards	151	0.727676
hot water kettle	161	0.775866
ev charger	181	0.872247
baby bath	182	0.877066
stair gates	195	0.939714
outlet covers	232	1.118018
pocket wifi	299	1.440894
game console	299	1.440894
babysitter recommendations	313	1.508361
toilet paper	328	1.580647
bath towel	328	1.580647
body soap	328	1.580647
bedroom comforts	333	1.604742
bathroom essentials	333	1.604742
wide clearance to shower	387	1.86497
toilet	387	1.86497
crib	396	1.908342
full kitchen	400	1.927618
children’s dinnerware	404	1.946894
cat(s)	420	2.023999
disabled parking spot	421	2.028818
dog(s)	473	2.279408
suitable for events	480	2.313142
extra space around shower and toilet	509	2.452894
wide doorway to guest bathroom	564	2.717941
waterfront	631	3.040817
accessible-height toilet	658	3.170932
smoking allowed	673	3.243217
high chair	755	3.638379
accessible-height bed	778	3.749217
building staff	831	4.004626
children’s books and toys	867	4.178112
pack ’n play/travel crib	877	4.226302
wide entryway	908	4.375693
smart lock	915	4.409426
cleaning before checkout	933	4.496169
wide entrance	1032	4.973254
extra space around bed	1050	5.059997
pets live on this property	1076	5.185292
room-darkening shades	1094	5.272035
doorman	1142	5.503349

flat path to guest entrance	1160	5.590092
wheelchair accessible	1165	5.614187
wide entrance for guests	1308	6.303311
wide hallways	1308	6.303311
ethernet connection	1312	6.322587
single level home	1375	6.626187
lake access	1378	6.640644
buzzer/wireless intercom	1542	7.430967
other	1627	7.840586
bbq grill	1646	7.932148
breakfast	1850	8.915233
well-lit path to entrance	1938	9.339309
24-hour check-in	1958	9.43569
free street parking	2015	9.710375
indoor fireplace	2051	9.883861
keypad	2136	10.29348
translation missing: en.hosting_amenity_49	2251	10.84767
no stairs or steps to enter	2337	11.26211
bathtub	2475	11.92714
private living room	2592	12.49096
paid parking on premises	2599	12.5247
pets allowed	2604	12.54879
garden or backyard	2660	12.81866
safety card	2702	13.02106
translation missing: en.hosting_amenity_50	2799	13.48851
hot tub	3115	15.01132
host greets you	3239	15.60889
lockbox	3359	16.18717
pool	3819	18.40393
luggage dropoff allowed	3913	18.85692
patio or balcony	4587	22.10496
cable tv	4735	22.81818
paid parking off premises	4759	22.93383
internet	4998	24.08559
extra pillows and blankets	5835	28.11913
dishwasher	5875	28.31189
lock on bedroom door	6244	30.09012
gym	6295	30.33589
long term stays allowed	6471	31.18404
family/kid friendly	6560	31.61293
private entrance	7006	33.76223
coffee maker	7184	34.62002
self check-in	7229	34.83688

bed linens	7875	37.94998
first aid kit	7924	38.18611
cooking basics	8121	39.13546
oven	8174	39.39087
free parking on premises	8327	40.12819
stove	8521	41.06308
microwave	8671	41.78594
elevator	8671	41.78594
dishes and silverware	8700	41.92569
refrigerator	9431	45.44841
fire extinguisher	10742	51.76618
hot water	12253	59.04776
iron	14644	70.57009
tv	14662	70.65684
hair dryer	15306	73.7603
laptop friendly workspace	15611	75.23011
dryer	16542	79.71664
shampoo	16602	80.00578
carbon monoxide detector	16655	80.26119
washer	16797	80.9455
hangers	17590	84.767
air conditioning	17728	85.43203
kitchen	19127	92.17387
smoke detector	19520	94.06776
essentials	19736	95.10867
heating	20091	96.81943
wifi	20366	98.14467

Appendix IV

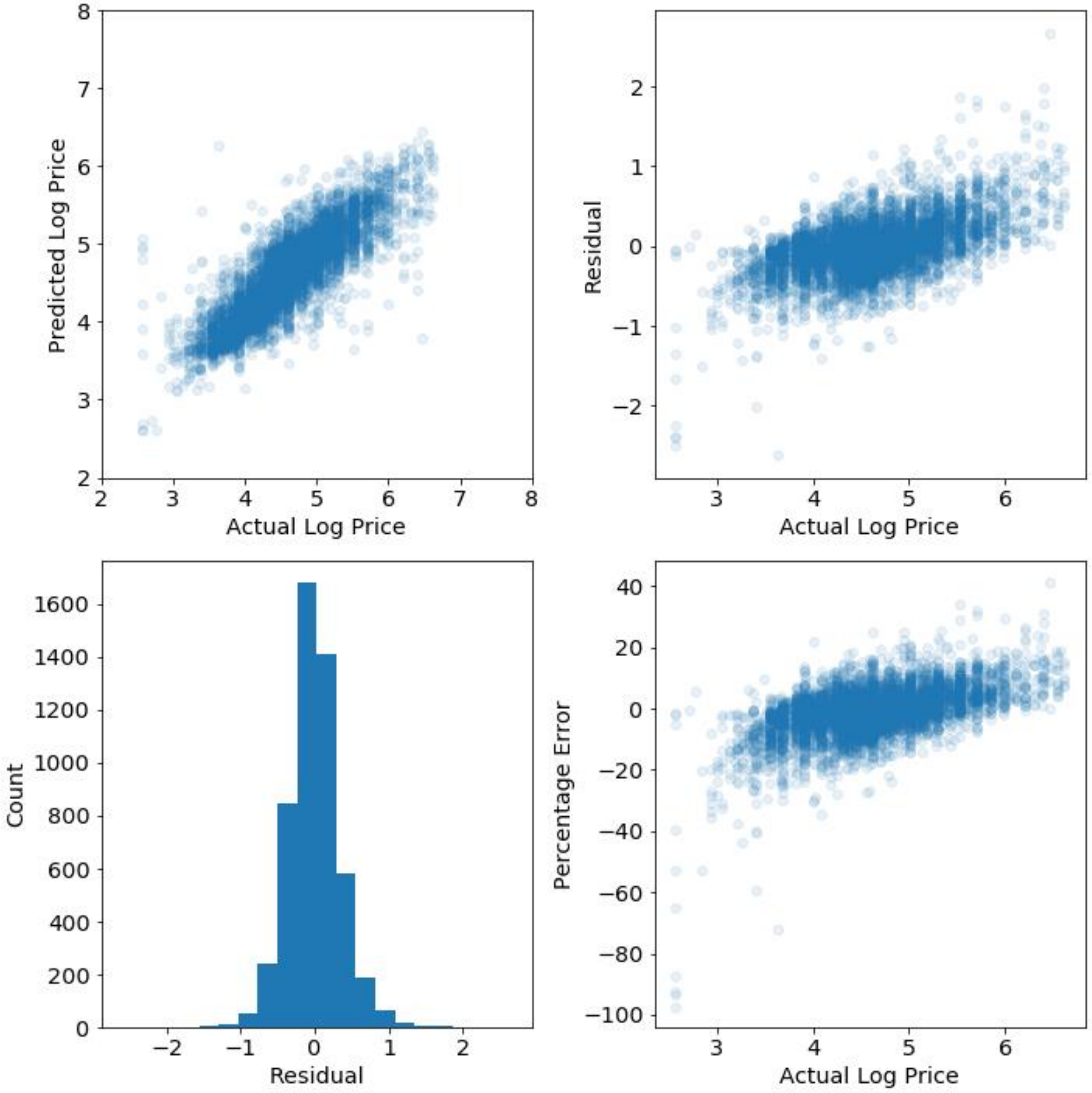
Hypothesis Test Results for Categorical Features

	test	number of categories	rejected_null_hypothesis
property_type_simple	Kruskal-Wallis	4	yes
room_type	Kruskal-Wallis	3	yes
bed_type	Kruskal-Wallis	5	yes
city_fsa	Kruskal-Wallis	9	yes
cancellation_policy	Kruskal-Wallis	4	yes
instant_bookable	MannWhitneyU	2	yes
is_business_travel_ready	MannWhitneyU	1	NA
host_response_time	Kruskal-Wallis	4	yes
host_is_superhost	MannWhitneyU	2	yes
host_has_profile_pic	MannWhitneyU	2	yes
host_identity_verified	MannWhitneyU	2	yes
elevator	MannWhitneyU	2	yes
hair dryer	MannWhitneyU	2	yes
pool	MannWhitneyU	2	yes
laptop friendly workspace	MannWhitneyU	2	yes
lockbox	MannWhitneyU	2	yes
paid parking on premises	MannWhitneyU	2	yes
shampoo	MannWhitneyU	2	yes
hot tub	MannWhitneyU	2	yes
air conditioning	MannWhitneyU	2	yes
luggage dropoff allowed	MannWhitneyU	2	no
safety card	MannWhitneyU	2	yes
carbon monoxide detector	MannWhitneyU	2	yes
refrigerator	MannWhitneyU	2	yes
stove	MannWhitneyU	2	yes
cable tv	MannWhitneyU	2	yes
free parking on premises	MannWhitneyU	2	yes
cooking basics	MannWhitneyU	2	yes
private entrance	MannWhitneyU	2	yes
hot water	MannWhitneyU	2	yes
family/kid friendly	MannWhitneyU	2	yes
patio or balcony	MannWhitneyU	2	yes
bathtub	MannWhitneyU	2	yes
lock on bedroom door	MannWhitneyU	2	yes
washer	MannWhitneyU	2	yes
internet	MannWhitneyU	2	yes
private living room	MannWhitneyU	2	no

paid parking off premises	MannWhitneyU	2	yes
dishes and silverware	MannWhitneyU	2	yes
fire extinguisher	MannWhitneyU	2	yes
tv	MannWhitneyU	2	yes
dishwasher	MannWhitneyU	2	yes
first aid kit	MannWhitneyU	2	no
microwave	MannWhitneyU	2	yes
iron	MannWhitneyU	2	yes
oven	MannWhitneyU	2	yes
host greets you	MannWhitneyU	2	yes
coffee maker	MannWhitneyU	2	yes
no stairs or steps to enter	MannWhitneyU	2	yes
bed linens	MannWhitneyU	2	yes
gym	MannWhitneyU	2	yes
long term stays allowed	MannWhitneyU	2	yes
pets allowed	MannWhitneyU	2	yes
extra pillows and blankets	MannWhitneyU	2	yes
garden or backyard	MannWhitneyU	2	yes
dryer	MannWhitneyU	2	yes
translation missing: en.hosting_amenity_50	MannWhitneyU	2	yes
translation missing: en.hosting_amenity_49	MannWhitneyU	2	yes
keypad	MannWhitneyU	2	yes
self check-in	MannWhitneyU	2	yes
hangers	MannWhitneyU	2	yes

Appendix V

Diagnostics plots of the random forest model

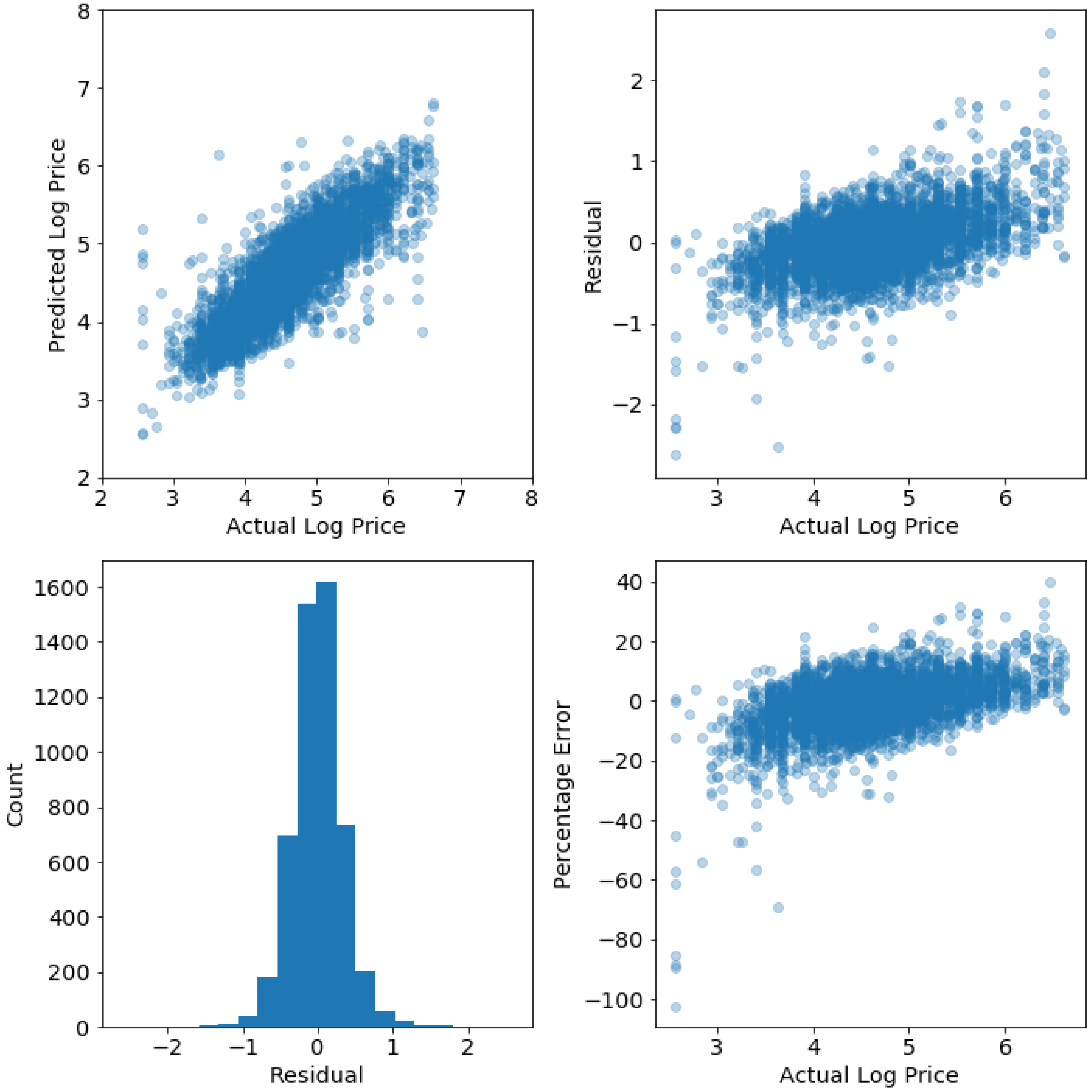


Appendix VI

Cross validation results and diagnostics plots of the gradient boosting model

learning_rate	max_depth	max_features	n_estimators	mean_test_score
0.01	2	None	100	0.45218663
0.01	2	None	300	0.60134258
0.01	2	None	500	0.63843038
0.01	2	sqrt	100	0.31312793
0.01	2	sqrt	300	0.54126925
0.01	2	sqrt	500	0.60822787
0.01	2	log2	100	0.23064191
0.01	2	log2	300	0.46067712
0.01	2	log2	500	0.55392031
0.01	5	None	100	0.54522917
0.01	5	None	300	0.6850974
0.01	5	None	500	0.70903576
0.01	5	sqrt	100	0.45827076
0.01	5	sqrt	300	0.64815196
0.01	5	sqrt	500	0.68695402
0.01	5	log2	100	0.38251483
0.01	5	log2	300	0.60581988
0.01	5	log2	500	0.66110435
0.1	2	None	100	0.67668559
0.1	2	None	300	0.71030264
0.1	2	None	500	0.71827995
0.1	2	sqrt	100	0.65508041
0.1	2	sqrt	300	0.6985305
0.1	2	sqrt	500	0.7096035
0.1	2	log2	100	0.6291184
0.1	2	log2	300	0.68637351
0.1	2	log2	500	0.69942518
0.1	5	None	100	0.72651805
0.1	5	None	300	0.73416848
0.1	5	None	500	0.73469309
0.1	5	sqrt	100	0.70727963
0.1	5	sqrt	300	0.7294817
0.1	5	sqrt	500	0.73477511
0.1	5	log2	100	0.69285992
0.1	5	log2	300	0.72112479
0.1	5	log2	500	0.7283724
1	2	None	100	0.68681277
1	2	None	300	0.68031137
1	2	None	500	0.66609217

1	2	sqrt	100	0.65998806
1	2	sqrt	300	0.66564543
1	2	sqrt	500	0.66502615
1	2	log2	100	0.61666074
1	2	log2	300	0.63528382
1	2	log2	500	0.63418953
1	5	None	100	0.53294102
1	5	None	300	0.47112981
1	5	None	500	0.46101269
1	5	sqrt	100	0.47111021
1	5	sqrt	300	0.40938343
1	5	sqrt	500	0.3832228
1	5	log2	100	0.50522974
1	5	log2	300	0.44515904
1	5	log2	500	0.41638298


Predicted Log Price
Actual Log Price
2
3
4
5
6
7
8
Residual
Actual Log Price
−2
−1
0
1
2
3
4
5
6
Count
Residual
0
200
400
600
800
1000
1200
1400
1600
−2
−1
0
1
2
Percentage Error
Actual Log Price
−100
−80
−60
−40
−20
0
20
40
3
4
5
6

Appendix VII

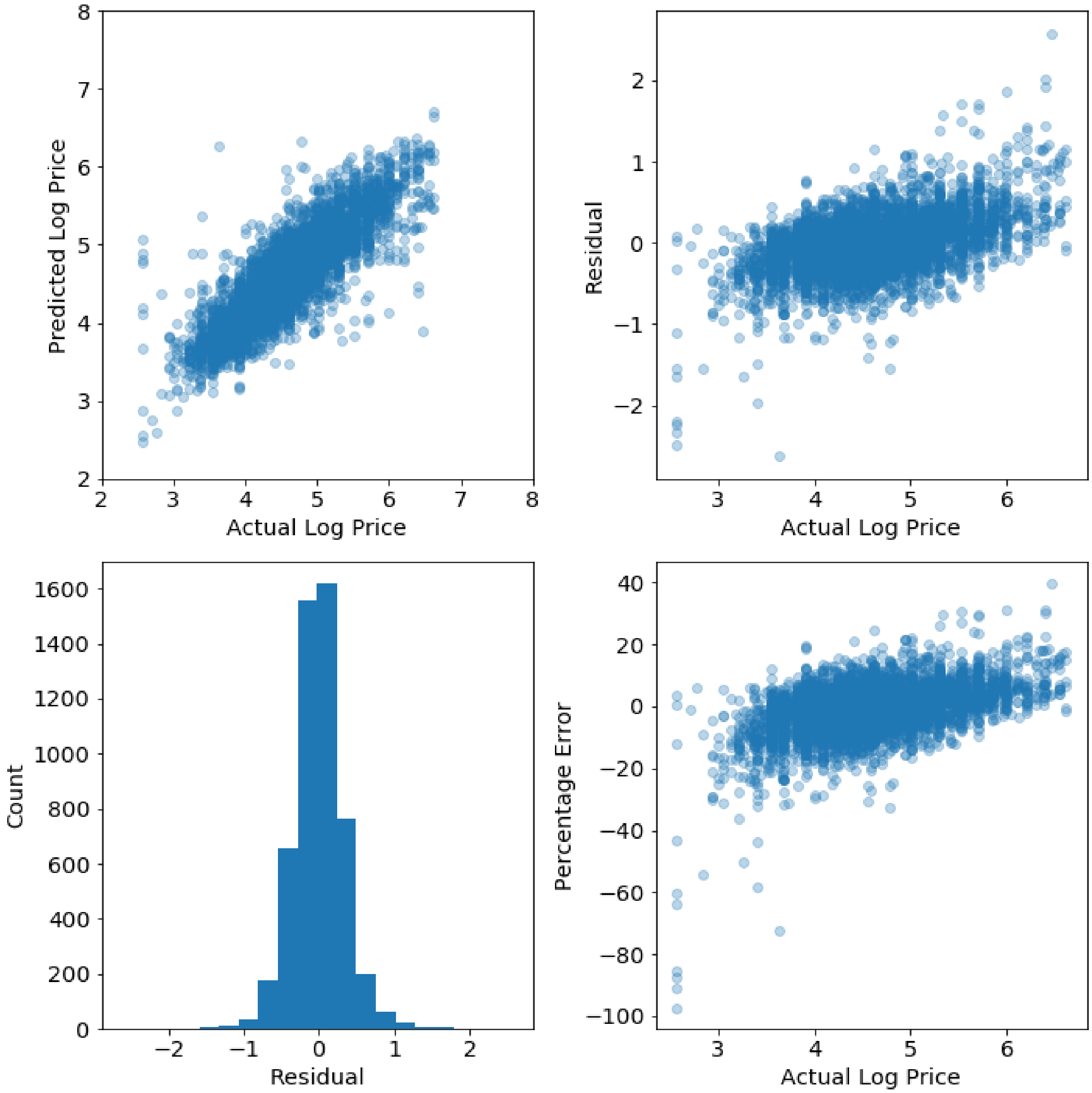
Cross validation results and diagnostics plots of the XGBoost model

colsample_bytree	eta	gamma	max_depth	n_estimators	mean_test_score
0.5	0.1	0	2	100	0.67397337
0.5	0.1	0	2	300	0.70879394
0.5	0.1	0	2	500	0.71744329
0.5	0.1	0	5	100	0.72558007
0.5	0.1	0	5	300	0.73666229
0.5	0.1	0	5	500	0.73842746
0.5	0.1	0.1	2	100	0.67397337
0.5	0.1	0.1	2	300	0.70877104
0.5	0.1	0.1	2	500	0.71737119
0.5	0.1	0.1	5	100	0.72533264
0.5	0.1	0.1	5	300	0.7359967
0.5	0.1	0.1	5	500	0.73779592
0.5	0.1	0.2	2	100	0.67397337
0.5	0.1	0.2	2	300	0.7088596
0.5	0.1	0.2	2	500	0.71739492
0.5	0.1	0.2	5	100	0.72627996
0.5	0.1	0.2	5	300	0.73783134
0.5	0.1	0.2	5	500	0.73939973
0.5	0.3	0	2	100	0.67397337
0.5	0.3	0	2	300	0.70879394
0.5	0.3	0	2	500	0.71744329
0.5	0.3	0	5	100	0.72558007
0.5	0.3	0	5	300	0.73666229
0.5	0.3	0	5	500	0.73842746
0.5	0.3	0.1	2	100	0.67397337
0.5	0.3	0.1	2	300	0.70877104
0.5	0.3	0.1	2	500	0.71737119
0.5	0.3	0.1	5	100	0.72533264
0.5	0.3	0.1	5	300	0.7359967
0.5	0.3	0.1	5	500	0.73779592
0.5	0.3	0.2	2	100	0.67397337
0.5	0.3	0.2	2	300	0.7088596
0.5	0.3	0.2	2	500	0.71739492
0.5	0.3	0.2	5	100	0.72627996
0.5	0.3	0.2	5	300	0.73783134
0.5	0.3	0.2	5	500	0.73939973
0.5	0.5	0	2	100	0.67397337
0.5	0.5	0	2	300	0.70879394
0.5	0.5	0	2	500	0.71744329
0.5	0.5	0	5	100	0.72558007

0.5	0.5	0	5	300	0.73666229
0.5	0.5	0	5	500	0.73842746
0.5	0.5	0.1	2	100	0.67397337
0.5	0.5	0.1	2	300	0.70877104
0.5	0.5	0.1	2	500	0.71737119
0.5	0.5	0.1	5	100	0.72533264
0.5	0.5	0.1	5	300	0.7359967
0.5	0.5	0.1	5	500	0.73779592
0.5	0.5	0.2	2	100	0.67397337
0.5	0.5	0.2	2	300	0.7088596
0.5	0.5	0.2	2	500	0.71739492
0.5	0.5	0.2	5	100	0.72627996
0.5	0.5	0.2	5	300	0.73783134
0.5	0.5	0.2	5	500	0.73939973
0.75	0.1	0	2	100	0.6758631
0.75	0.1	0	2	300	0.70972711
0.75	0.1	0	2	500	0.7177479
0.75	0.1	0	5	100	0.72612493
0.75	0.1	0	5	300	0.7373368
0.75	0.1	0	5	500	0.73883607
0.75	0.1	0.1	2	100	0.6758631
0.75	0.1	0.1	2	300	0.70960376
0.75	0.1	0.1	2	500	0.71784504
0.75	0.1	0.1	5	100	0.72531454
0.75	0.1	0.1	5	300	0.7363893
0.75	0.1	0.1	5	500	0.73786588
0.75	0.1	0.2	2	100	0.6758631
0.75	0.1	0.2	2	300	0.70957844
0.75	0.1	0.2	2	500	0.71760125
0.75	0.1	0.2	5	100	0.7259983
0.75	0.1	0.2	5	300	0.73660824
0.75	0.1	0.2	5	500	0.73756131
0.75	0.3	0	2	100	0.6758631
0.75	0.3	0	2	300	0.70972711
0.75	0.3	0	2	500	0.7177479
0.75	0.3	0	5	100	0.72612493
0.75	0.3	0	5	300	0.7373368
0.75	0.3	0	5	500	0.73883607
0.75	0.3	0.1	2	100	0.6758631
0.75	0.3	0.1	2	300	0.70960376
0.75	0.3	0.1	2	500	0.71784504
0.75	0.3	0.1	5	100	0.72531454
0.75	0.3	0.1	5	300	0.7363893

0.75	0.3	0.1	5	500	0.73786588
0.75	0.3	0.2	2	100	0.6758631
0.75	0.3	0.2	2	300	0.70957844
0.75	0.3	0.2	2	500	0.71760125
0.75	0.3	0.2	5	100	0.7259983
0.75	0.3	0.2	5	300	0.73660824
0.75	0.3	0.2	5	500	0.73756131
0.75	0.5	0	2	100	0.6758631
0.75	0.5	0	2	300	0.70972711
0.75	0.5	0	2	500	0.7177479
0.75	0.5	0	5	100	0.72612493
0.75	0.5	0	5	300	0.7373368
0.75	0.5	0	5	500	0.73883607
0.75	0.5	0.1	2	100	0.6758631
0.75	0.5	0.1	2	300	0.70960376
0.75	0.5	0.1	2	500	0.71784504
0.75	0.5	0.1	5	100	0.72531454
0.75	0.5	0.1	5	300	0.7363893
0.75	0.5	0.1	5	500	0.73786588
0.75	0.5	0.2	2	100	0.6758631
0.75	0.5	0.2	2	300	0.70957844
0.75	0.5	0.2	2	500	0.71760125
0.75	0.5	0.2	5	100	0.7259983
0.75	0.5	0.2	5	300	0.73660824
0.75	0.5	0.2	5	500	0.73756131
1	0.1	0	2	100	0.67659306
1	0.1	0	2	300	0.70968638
1	0.1	0	2	500	0.71800765
1	0.1	0	5	100	0.72717776
1	0.1	0	5	300	0.7369464
1	0.1	0	5	500	0.73796215
1	0.1	0.1	2	100	0.67659306
1	0.1	0.1	2	300	0.70970532
1	0.1	0.1	2	500	0.7181258
1	0.1	0.1	5	100	0.72720653
1	0.1	0.1	5	300	0.73663063
1	0.1	0.1	5	500	0.73688886
1	0.1	0.2	2	100	0.67659306
1	0.1	0.2	2	300	0.70969967
1	0.1	0.2	2	500	0.71819138
1	0.1	0.2	5	100	0.72680574
1	0.1	0.2	5	300	0.7352969
1	0.1	0.2	5	500	0.73529691

1	0.3	0	2	100	0.67659306
1	0.3	0	2	300	0.70968638
1	0.3	0	2	500	0.71800765
1	0.3	0	5	100	0.72717776
1	0.3	0	5	300	0.7369464
1	0.3	0	5	500	0.73796215
1	0.3	0.1	2	100	0.67659306
1	0.3	0.1	2	300	0.70970532
1	0.3	0.1	2	500	0.7181258
1	0.3	0.1	5	100	0.72720653
1	0.3	0.1	5	300	0.73663063
1	0.3	0.1	5	500	0.73688886
1	0.3	0.2	2	100	0.67659306
1	0.3	0.2	2	300	0.70969967
1	0.3	0.2	2	500	0.71819138
1	0.3	0.2	5	100	0.72680574
1	0.3	0.2	5	300	0.7352969
1	0.3	0.2	5	500	0.73529691
1	0.5	0	2	100	0.67659306
1	0.5	0	2	300	0.70968638
1	0.5	0	2	500	0.71800765
1	0.5	0	5	100	0.72717776
1	0.5	0	5	300	0.7369464
1	0.5	0	5	500	0.73796215
1	0.5	0.1	2	100	0.67659306
1	0.5	0.1	2	300	0.70970532
1	0.5	0.1	2	500	0.7181258
1	0.5	0.1	5	100	0.72720653
1	0.5	0.1	5	300	0.73663063
1	0.5	0.1	5	500	0.73688886
1	0.5	0.2	2	100	0.67659306
1	0.5	0.2	2	300	0.70969967
1	0.5	0.2	2	500	0.71819138
1	0.5	0.2	5	100	0.72680574
1	0.5	0.2	5	300	0.7352969
1	0.5	0.2	5	500	0.73529691


Predicted Log Price
8
7
6
5
4
3
2
2
3
4
5
6
7
8
Actual Log Price
Residual
2
1
0
−1
−2
3
4
5
6
Actual Log Price
Count
1600
1400
1200
1000
800
600
400
200
0
−2
−1
0
1
2
Residual
Percentage Error
40
20
0
−20
−40
−60
−80
−100
3
4
5
6
Actual Log Price