An Analysis of Airbnb's Prices and Superhosts of Boston

Duy Nguyen

November 30, 2024

1 Background

Airbnb was founded in 2008 by Brian Chesky, Joe Gebbia and Nathan Blecharczyk. It is a global platform that provides short-term lodging and experiences as an alternative to established hospitality business such as hotels. Airbnb has a strong global presence and operates in over 190 countries. As of Febuary 2024, there are a total of 7.7 million listings worldwide with the USA attributed for more than 2 million listings. However, this project will narrow the scope of litings to the city of Boston. As of June 2024, Boston has a total of 4,325 active or inactive listings.

The dataset utlilized in this project contain information on the 4,325 Airbnb's listings in Boston. This dataset is retrieved from https://insideairbnb.com/get-the-data/ and is not associated or endorsed by

Airbnb. The data is compiled from publicly available information on the Airbnb website by community members called Inside Airbnb. Inside Airbnb claimed the data has been verifed, cleansed, analyzed and aggregated. The dataset is very dense in information where each row contains represent one Airbnb listing and each column contains information about that listing. Information incluced are but not limited to: the host information, associated property information, price per night, reviews, ratings, avilability, and geographical enformation.

2 description of dataset

The dataset contains 4325 rows and 75 columns. Since there are too many columnns, we will list those which are relavant to this project. First we will list out the variables that are not self-explanatory and their descriptions below.

- accomodates: The maximum guest capacity of a listing.
- bathrooms, bedrooms, beds: the number of bathrooms, bedrooms and beds.
- room_type: One of Entire home/Apt, Private Room, Shared, or Hotel Room.
- host_id: The unique identifier of a host.

- property_type: More detailed description of room_type.
- price: daily price in local currency.

The following variables name are more self-explantory and do not require detailed description: host_response_rate, host_acceptance_rate, host_is_superhost, host_listings_count, latitude, longtitude, number _of_reviews, review_scores_rating.

Although the dataset is claimed to be cleaned, there are missing values found in different features. Some of which are crucial to our prediction process. In addition, the variable value of "price" is rendered as a string because of the symbol $ in front. Thus, we will address our handling of missing values and data preperation in the later section.

3 Research Questions

There are two main questions we are addressing in this project:

1. Which model would yield the best performance in predicting price?
    - Which features is the most impactful on the price of Airbnb?
2. Which model would yield the best performance in predicting whether a listing is a superhost?
    - Which features is the most impacful on predicting a host is a superhost?

Question 1 can be addressed with different regression models because the predictor, "price", is a continuous variable. We will build four different models ranging from less flexible to more flexible, they are: linear regression, lasso regression, ridge regression and random forest regression. To determine the best model, we will compare each of their test MSE. The model with the lowest MSE is the best performing model. As for important features, only the two models random forest and linear regression allows us to identify important features. Therefore, we will choose the better performing model for determining important features.

Question 2 is addressed with binomial classification models because we are predicting whether the host of a listing is a superhost. Similarly, we will build five models with varying flexibility, which are: logistic regression, lasso and ridge logistic regressionm, random forest classifier and support vector machine (SVM) with radial basis. The best model is determined by the model with the highest precision. Finding important features follow the same process as Question 1. We will show features importance based on the better performing model between logistic regression or random forest classifier.

4 Data Processing and Methods

4.1 Data Exploration

We will first explore our two response variables price and host_is_superhost. The variable price is the continous variable with a min of 25, median of

190, and max of 4786. There are 782 missing values in the variable. The distribution is skewed to the left as seen in the figure below.

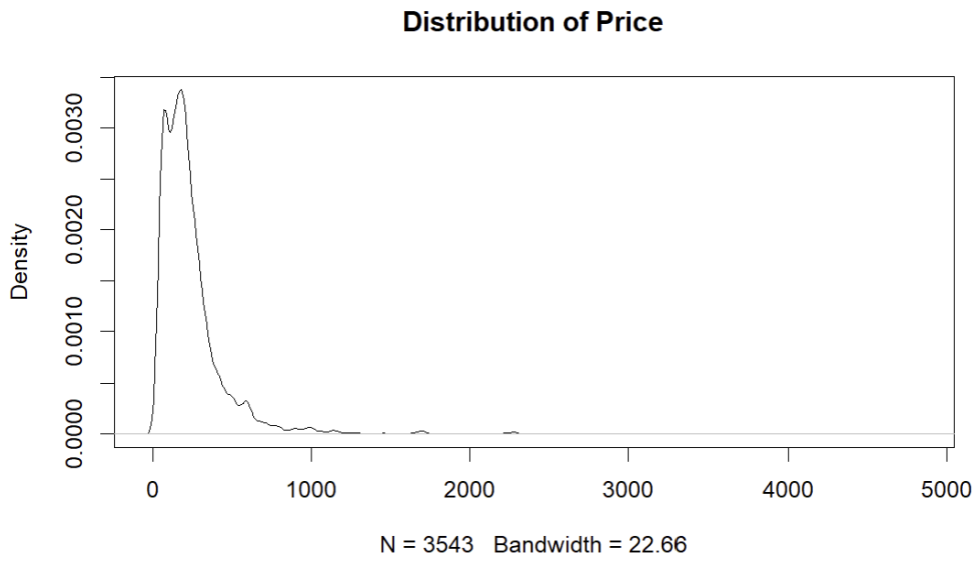

Figure 1: Distribution of response price

Our next response is host_is_superhost, a binary response variable with two factors "t" indicates a host is superhost, and "f" for not superhost. In distribution of superhost in Boston is 28.39$ are superhosts while 66.98% are not. The remaining 4.62% are missing values in the dataset.

Next we will be looking at the correalation between the continous variables that we think are important for our prediction. The correlation matrix is show in the figure below.

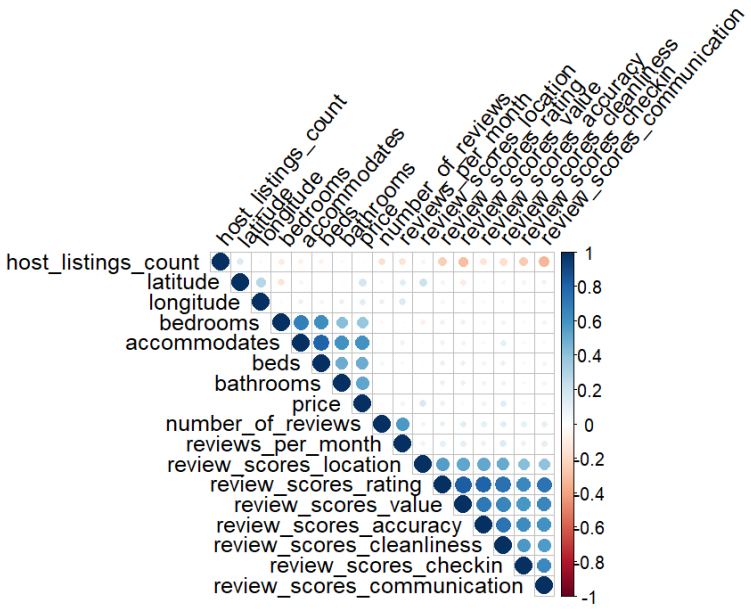

Figure 2: Correlation matrix of continuous features

The correalation mamtrix shows a moderate relationship between the features bedrooms, accomodates, and bathrooms. This makes sense given that a property has a larger property should be able to hold more bedrooms and bathrooms. Of course, bigger property also means more customers can be accomodated. Since these variables are not highly correlated thus multicollinearity is not an issue. The response price also have moderate correlation with aformentioned features. The reviews variable have a moderately high correlation with each other but no correalation with price. Again, since the reviews variables are not highly correalated multicollinearity issue is not a problem.

4.2 Missing Value Analysis

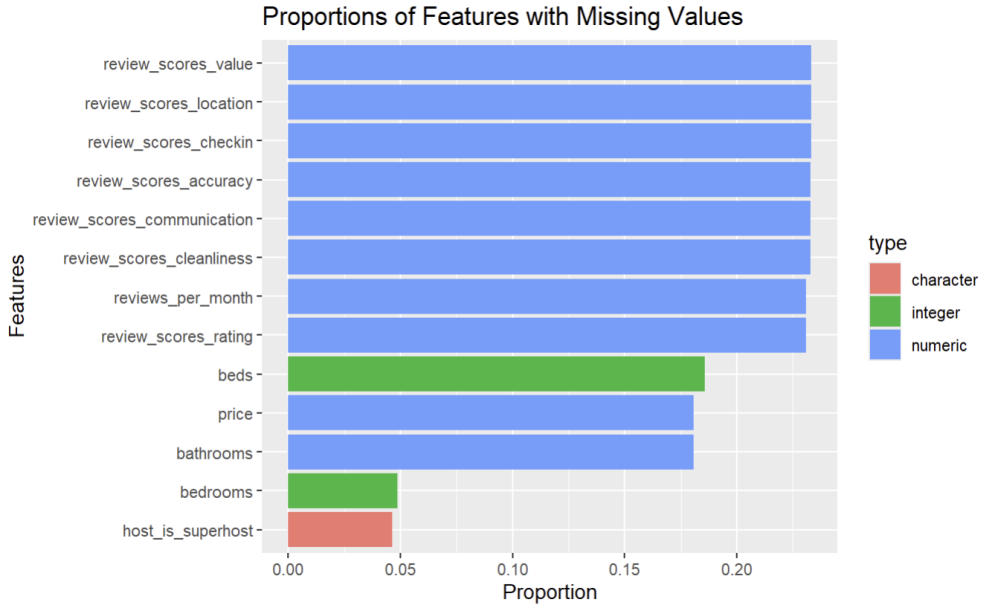

Figure 3: Features missing values and their frequencies.

There are 15 features in the dataset with missing values. From Figure 1, there are two features "neighbourhood_group_cleansed" and "calendar_update" consist entirely of missing values. In other words, they provide no information and will be removed from the dataset thusly. The remaining 13 features only partially have missing values. Removing rows with missing values is unwised because we would lose valuable information especially when proportion of missing values are not high. As a result, our approach to rectify this problem is to impute the missing values. This process will be discussed further in the Imputation section.

Notice from Figure 3, there are features sharing the same frequency of

missing values, indicating they might have missing value on the same rows. Applying the same method on the remaining 13 features missing values. From Figure 1, we can easily identified four groups having missing value on the same rows consisting of three pairs of features, and one triplet of features. The other five features do share rows but insigficant amount.

4.3 Data Imputation

In general, for each method, we impute only the predictors but not the response. This is done to give each imputation a fair chances when we compare them in later section. Naturally, more complicated imputation would be able to capture the relationship between features better than their cruder counterpart. Thus, more complicated imputation would have better prediction.

We will create two imputed datasets with three different imputation methods. The first is a relatively simple method. We impute the features with missing values with the feature's median if it is continuous. For categorical feature, we impute with the mode. The second method is using principal component analysis (PCA) for imputation. Since PCA only makes sense with continuous variable, we only perform imputation on the continuous features but use the mode to impute categorical response. PCA imputation is performed with the help of the package missMDA. We first find the optimal number of the dimensions when running PCA on the dataset. Then we impute the missing values based on that value. Finally, we will perform KNN imputation with $k = 5$ using the VIM package.

4.4 Data Processing for Machine Learning Algorithms

As mentioned above, one of the response variable "price" is encoded as a character in the original dataset. Hence, we use the function gsub to remove all strings from the numerical value. Then we cast the response to a numeric. The other response, "host_is_superhost", is cast to factor for the random forest classification and SVM algorithms. We made the choice of reducing the number of features used from 75 to 21 that we determined to be useful for prediction.. This narrows the number of predictor from 75 to 21.

All imputed datasets will be split into a 80-20 training and testing sets. All of the models will be fitted on the training and tested on the testing set. In the case of models that have tunning parameters such as the regularizations models and SVM. We will perform 10-fold cross-validation on those models to determine the parameters that yield the best performing model. The regularized model only have one tunning paramater λ. On the other hand, SVM using radial basis kernel have two paramenters *cost* and γ.

5 Result Analysis

In this section, we will report the result we have from building models with the three imputed datasets. For regression model, we will be reporting their training and test MSE, then suggest the model with the lowest testing MSEs. In addition, we will also report the important features based either on the best performing model. For classification, we will also report their training

and testing precision. Then based on the model with the best precision, a corresponding matrix will be shown. Finally, we will also report important features.

5.1 Median Imputation

5.1.1 Question 1

Algorithm	Training MSE	Testing MSE
Linear Regression	22926	15336
Lasso	23057	15439
Ridge	23090	15481
Random Forest	4926	10522

Table 1: Four regression models with training and testing MSE

From Table 1, we observe that random forest is the best at predicting a listing price. We also note that the regularized help us reduce the MSE over the normal linear regression. This is not always the case, as we see in the later questions. The lower MSE of random forest in comparison to the linear models suggest non-linear relationship between the predictors and the response. Since random forest also outperformed linear regression, we will look at the top ten most important feature via the mean decrease in node impurity, as seen in the table below.

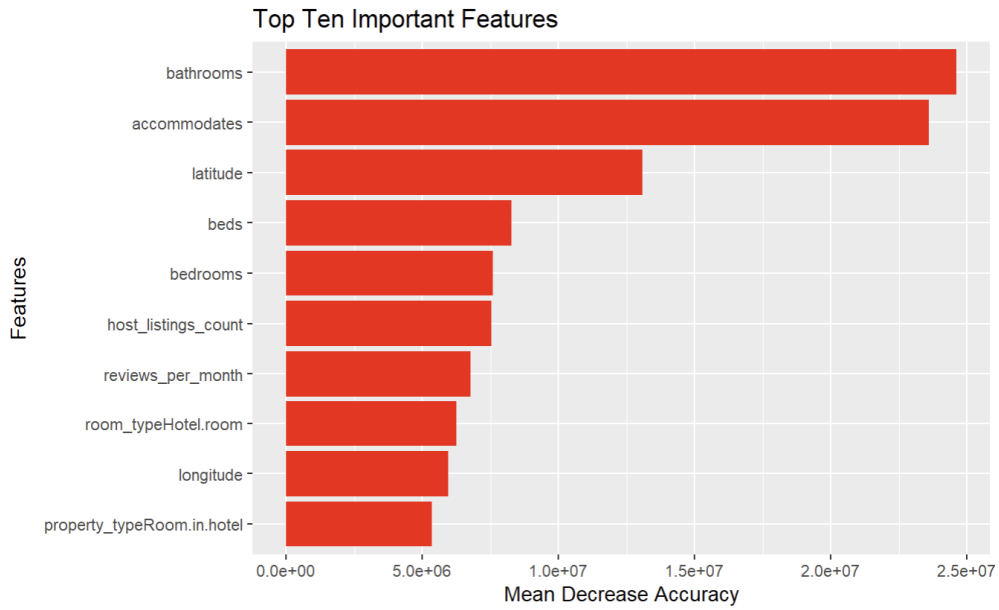

Figure 4: Features missing values and their frequencies with median imputation.

The top ten variables here make sense as these features are usually associate with price of listing. For instances, features that increase the capacity of the property will tend to also increase the price. The longititude and latitude will reflect geographical location of a property that would cost more or less. Interestingly, it seems that a property that is similar to hotel will also affect the price. Our guess is that there is a demand for property to be designed similar to a hotel. However, the number of listing for a host does not make sense here.

Algorithm	Training Precision	Testing Precision
Logistic Regression	0.8	0.7593
Lasso	0.7968	0.7581
Ridge	0.7714	0.7565
Random Forest	0.9968	0.8217
SVM	0.9978	0.7967

Table 2: Four classification models with training and testing precision

Similar to above, the random forest classification is the best performing model, follow by SVM and the three linear models. The non-linear relationship between predictors and response are also suggest here where the two non-linear models perform much better. As we alluded, regularized model does not always outperformed the non-regularized. Here we can see that logistic regression performs similar to ridge and outperform lasso. Since, the random forest classification performs is the best performer we will

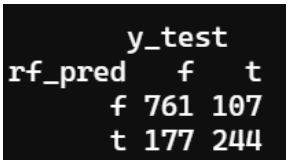

Figure 5: Confusion matrix of Random Forest with median imputation.

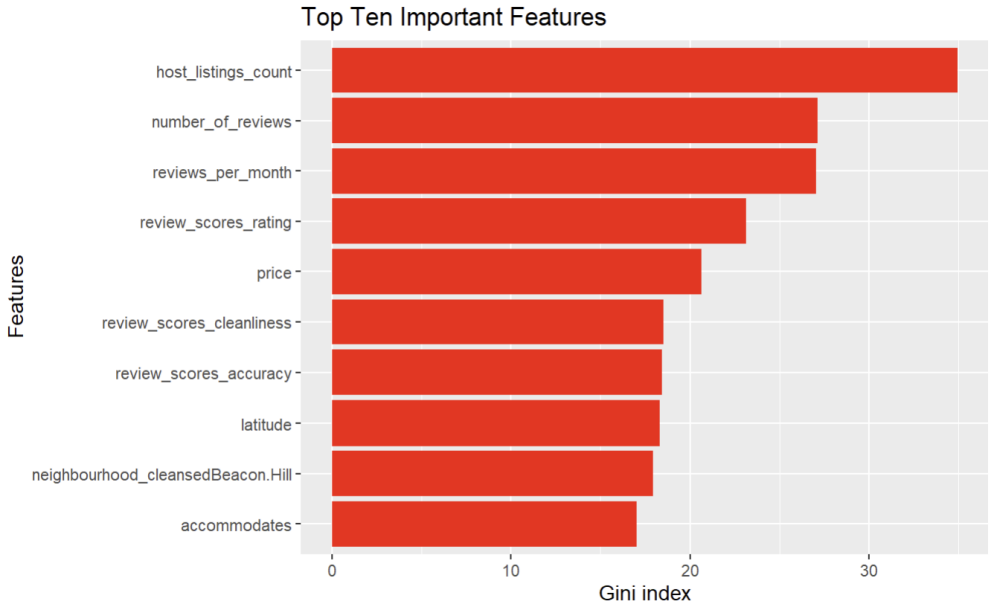

Figure 6: Features missing values and their frequencies with median imputation.

Interestingly, the most important feature is the number of listings a host has. However the other features are less suprising such as number of reviews, reviews per month, ratings, reviews for clealiness. It seems a host have a better chance to be superhost when they own more properties. Next, it is advantageous for host to been doing AirBnB longer and have more customers coming in. Finally, being good hosts by providing a good experience to customers, maintaining clealiness will help out as well.

5.2 PCA Model

5.2.1 Question 1

Similar to the Question 1 in the above section, we will make a table of each model with their respective MSE.

Algorithm	Training MSE	Testing MSE
Linear Regression	22081	19282
Lasso	22184	19984
Ridge	26246	19439
Random Forest	4379	13576

Table 3: Four regression models with training and testing MSE.

Random forest is the best performance here as well. The same pattern holds for the regularized model where both outperform the linear regression and the lasso regression is also the better performer. There are no chances to the features which are the top ten important features. However, their order slightly chances but the top three features still kept position.

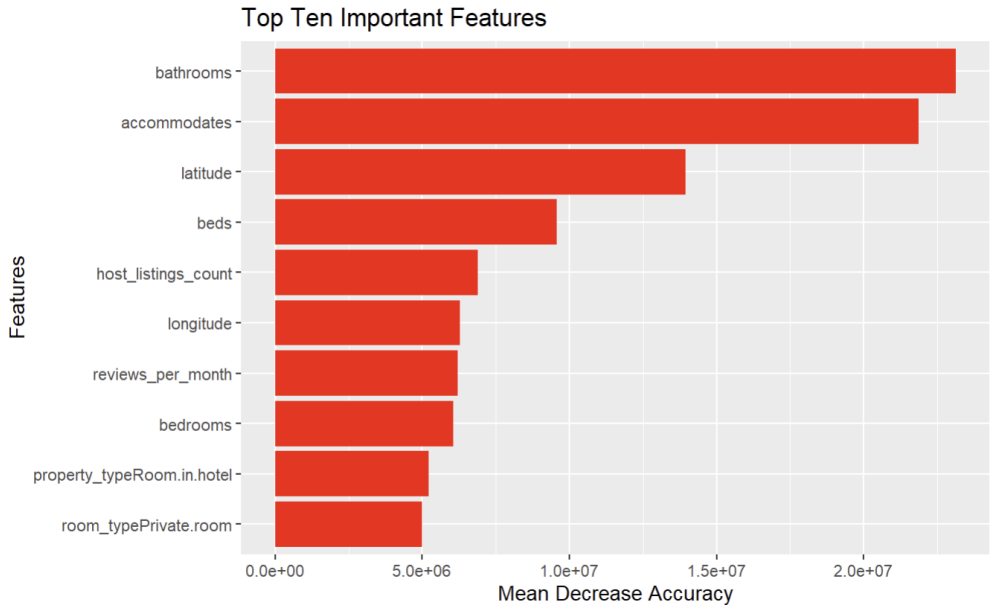

Figure 7: Features missing values and their frequencies with PCA imputation.

5.2.2 Question 2

Similar pattern to the previous question 2 is also observed here with a slight variation. The table of the four models and their precision are shown below.

Algorithm	Training Precision	Testing Precision
Logistic Regression	0.7968	0.7486
Lasso	0.7982	0.7381
Ridge	0.7996	0.7509
Random Forest	1	0.8563
SVM	0.9996	0.8021

Table 4: Four classification models with training and testing precision.

The order of performance remains the same with the median's result.

The order of best to worst performer is still random forest, SVM and the three linear models . Regularization in the case does not always help because logistic regression perform better than the lasso but slightly worse than ridge. We will observe table of top ten features below.

	y_test	
rf_pred	f	t
f	860	128
t	78	223

Figure 8: Confusion matrix of random forest with PCA imputation.

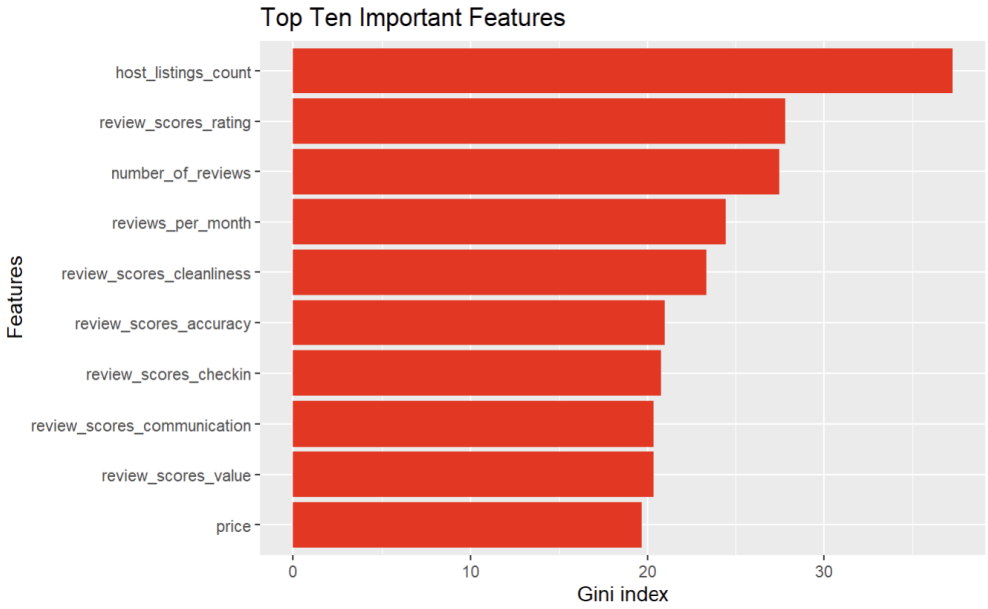

Figure 9: Features missing values and their frequencies with PCA imputation.

The feature importance for the PCA imputation is very similar to the median's with the most important feature being the same. However, the

important features are even more driven here. However, this does not change our view on the criteria for being a superhost as stated in the median's result.

5.3 KNN Imputation

5.3.1 Question 1

Algorithm	Training MSE	Testing MSE
Linear Regression	22990	14280
Lasso	23153	14157
Ridge	23158	14287
Random Forest	4613	9206

Table 5: Four regression models with training and testing MSE

Similar to above, the random forest model is the best performer. However, the performance for each model is much better than the other imputed datasets. We will look at the model important below.

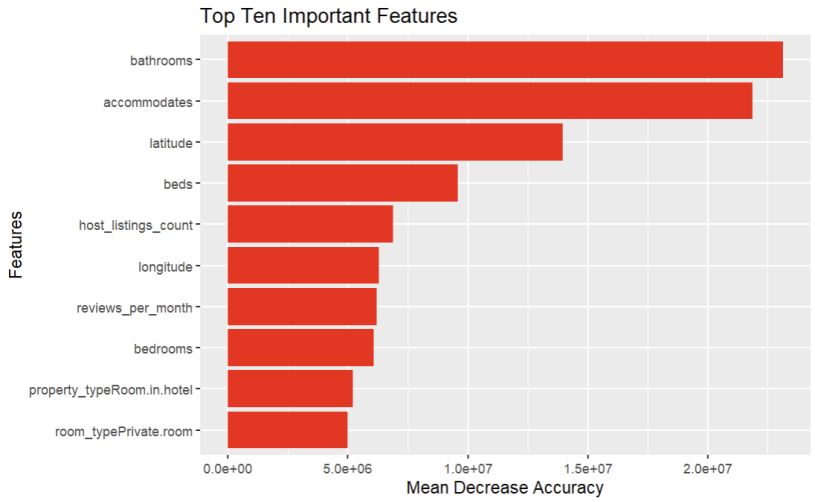

Figure 10: Features missing values and their frequencies with KNN imputation.

All the important features are similar to the random forest models above.

5.3.2 Question 2

Algorithm	Training Precision	Testing Precision
Logistic Regression	0.7863	0.7796
Lasso	0.7923	0.7727
Ridge	0.7274	0.7797
Random Forest	1	0.8417
SVM	1	0.7944

Table 6: Four Alogrithms and Their Precision

Unlike the regressionn models, this KNN imputation dataset does not yield a better performance on each model when compare to the PCA.

	y_test	
rf_pred	f	t
f	861	136
t	77	215

Figure 11: Confusion matrix of random forest with KNN imputation.

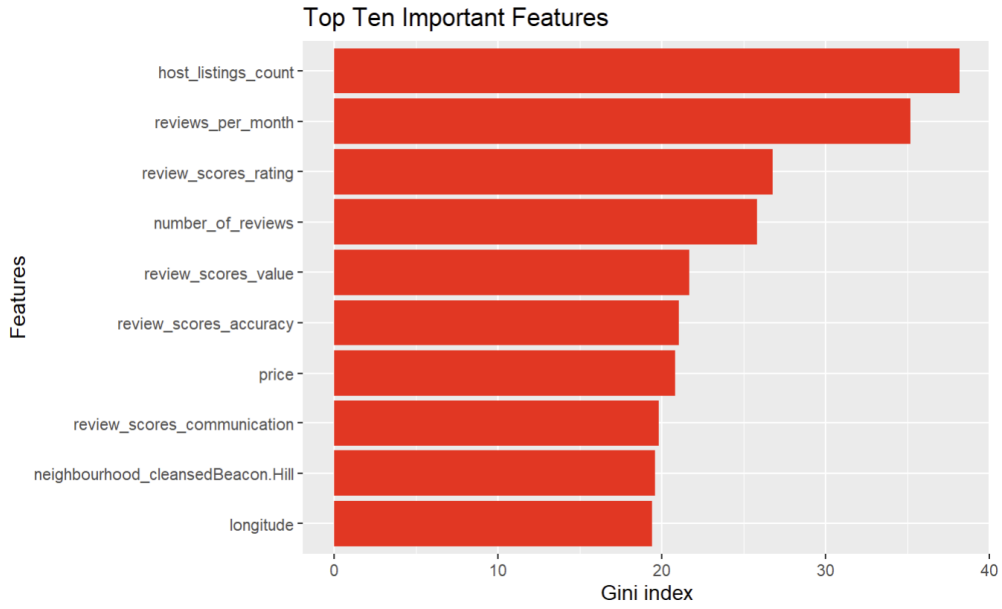

Figure 12: Features missing values and their frequencies with KNN imputation.

The KNN imputation has the same important features as the median's thus our analysis for superhost holds.

6 Conclusion

From observing the performance of the models with both imputed dataset, we see that the ensemble models in both classification and regression is the best performing model. For both imputed datasets, the classification and regression models performance are close to each other. In general, the classification models perform much more equally than their regression counterparts. In general, the more complicated imputation dataset out performs the crude one, i.e the median/mode imputation. In terms of the best performing model, for predict price, we would suggest the KNN imputed dataset with the random forest regression. For classification, our best performing model is the random forest on the PCA imputation.

In term of praticallity, our regression models would not be suitable for deployment. Even with the best model, the mean error is around $96. Such large error would make our price prediction useless. Furthermore, there is a consistent problem with the regression models for all three imputed datasets. The three linear models training MSEs are always higher the their testing's counterparts. We have yet to discover the underlying issue but we will work in the future.

In constrast, we would recommend to use the best performing classification model because it has a very high precision of around 85%. However, this model and its respective counterparts with different imputation method also have the same issue. These model prediction has lower false negative then

false positive. That is, if this model predicts "t" on an observation, we are less likely to be sure if that observation is actually "t".