Price Analysis and Borough Prediction of Airbnb listings in New York City

Kwangmin Kim[1], Leiyu Yue[1] and Kathryn Addabbo[1]

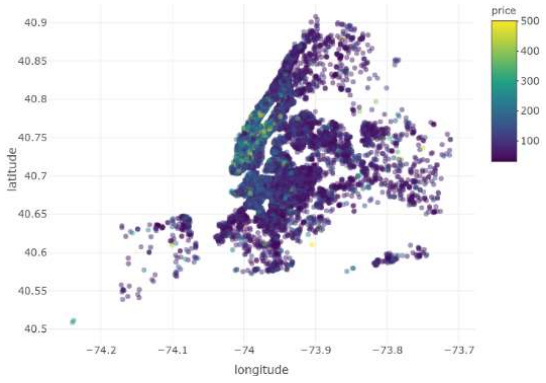
[1]Department of Biostatistics, Columbia University Mailman School of Public Health

1. Introduction

Countless tourists and students visit to explore all the attractions NYC has to offer as one of the largest cities in the world. Being a dynamic environment, the overall population of the city fluctuates daily. To handle the influx of individuals, information on available housing is extremely important. One of the most well-known housing companies is Airbnb. In this report, we analyze the Airbnb housing affordability and availability in each of New York City's boroughs using the company's released dataset in September 2017. The aim of this report is to provide clear visualization and meaningful interpretation of the data for convenience of visitors and all interested in the service of Airbnb.

This data is the dataset that has been used in the Data Science I course referenced by http://jeffgoldsmith.com/DSI/dataset_airbnb.html. This dataset is the result of merging "listings.csv.gz" with "listings.csv" from Inside Airbnb using the code from the referenced website. It contains a single data frame with 40,753 rows of data on 17 variables: id, review_scores_location, name, host_id , host_name, neighbourhood_group, neighbourhood, latitude, longitude , room_type, price , minimum_nights, number_of_reviews, last_review, reviews per month, calculated_host_listings_count, availability_365.

Figure 1: Price and Location in NYC



2. Exploratory Data Analysis

Figure 1 shows the price and locations in NYC. The closer to the yellow color is, the more expensive a location is, and vice versa. One of the covariates of interest (price) has values ranging from $10 to $10,000 per night. The analysis to values is restricted between $30 and $500 by the two reasons. From the standpoint of clear visualization, the range is so large, and the locations of the extremely high price are so rare that the locations would not be differentiable in color. From the perspective of the practical usage, considering that the average consumer does not have the appropriate funds to spend thousands of dollars a night on a room for a night, too much variation is not desirable in this analysis. The rating of each property was on a scale of 1 to 10, we changed the rating to a 5-star scale and assigned zeros to the properties with no ranking. We also assigned zeros to the rating and reviews per month column where there were NAs present. Housing is widely available through all five of the boroughs, however generally the closer to downtown Manhattan the property is, the more expensive the price per night as seen in Figure 1.

Figure 2: Property Count

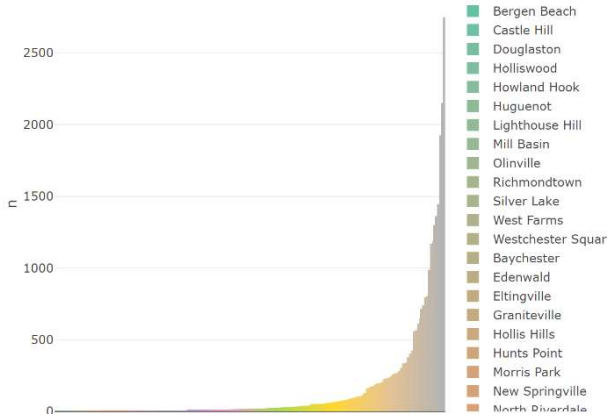


Figure 3: Top Neighborhood Prices

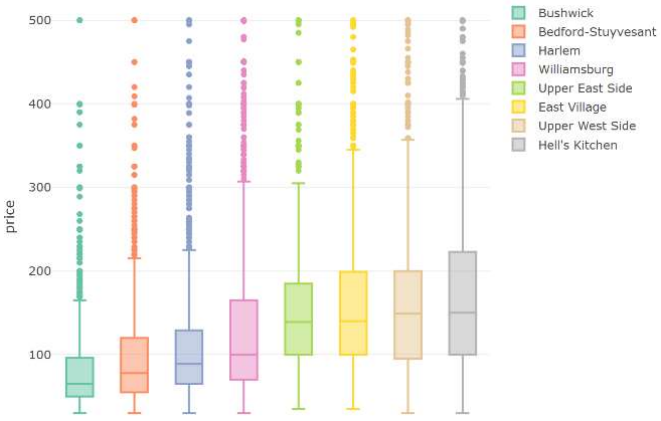
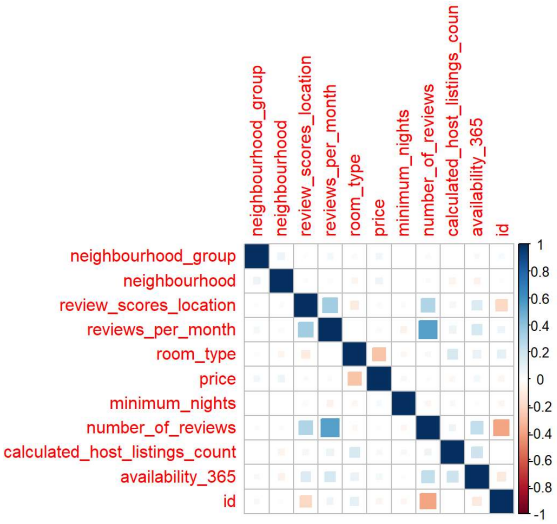


Figure 2, Property Count shows that the number of housing depending on the neighborhoods in NYC. It is expected that the more people demands the housing, the more housing will be supplied in the neighborhoods. It means the places are likely to be downtowns which are

popular with people. Of the properties, the majority are in Williamsburg and Bedford – Stuyvesant, both of which are in Brooklyn, which are displayed in the bar plot. To explore further, Figure 3 displays the price range of the top eight most populated neighborhoods, which range of average from $50 to $150 and maximums of $500 and presumably over. Interestingly, more popular neighborhoods have a wider range of housing price, which means that although it is a popular region for housing, housings with reasonable price could be found.

Figure 4: Correlation Plot of Covariates



To obtain an awareness of the relationships between the variables a correlation plot is created, which is viewed in Figure 4. Outside of the obvious correlation between number of reviews and reviews per month in addition to id, the important correlation to take into note is that price is correlated with room type. Customers are willing to pay more money to have an entire house to themselves for their stay as opposed to just renting out a room. Outside of this observation, there is very little correlation in the dataset, which is ideal for analysis.

3. Supervised Analysis

Multiple methods are used to investigate the relationship between price and the other covariates. Latitude, longitude, id and neighborhood and reviews per month are excluded from this analysis due to unique values and correlation between neighborhood and borough. For interpretability, regularization techniques are applied to this dataset. Simple linear regression

Figure 5: Model Test Errors

Linear	4316.674
Ridge	4436.158
Lasso	4436.618
PCR	4422.968
PLS	4014.465

yields a test error of 4316.674, an enormous number indicating poor fit. Our team moved to utilize shrinkage modeling methods such as ridge, lasso, principle component regression and partial least squares with the goal of identifying the lowest test error of those analyses. Each of the methods were assigned specific tuning parameters. Ridge regression with the tuning parameter 0.178 had a test error of 4436.158. This estimate is close to simple linear regression estimate due the ridge tuning parameter's proximity to zero, however it is slightly larger and thus a worse fit. Lasso was the next method observed; although the method is ideal for sparse models it compensates for variable selections unlike ridge regression. This methodology often reduces the variance at the cost of increased bias, which we see in the results.

With a best lambda of 0.28 the test error is 4436.618, this method is the worst of the three utilized thus far. PCR is a technique which derives a low dimensional set of features from a large set of variables, in which directionality of the data indicates which observations vary the most.

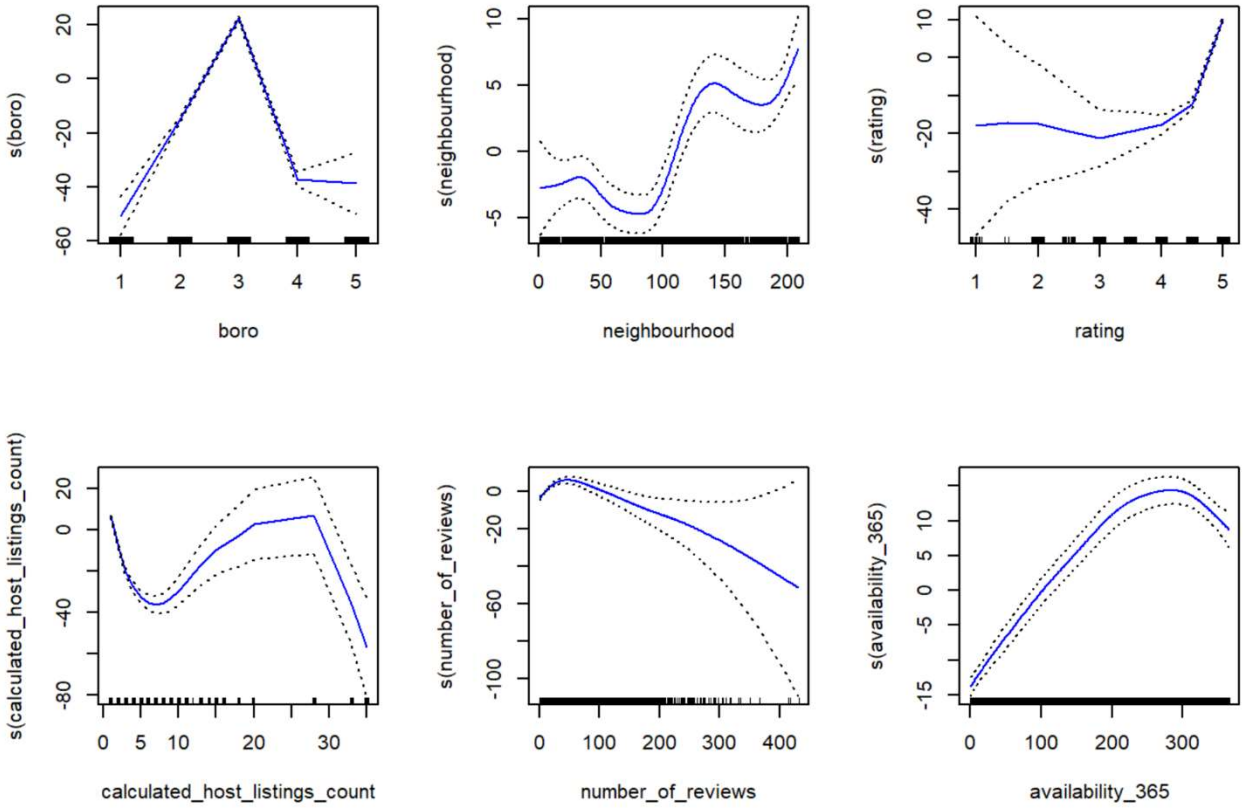
Figure 6: ANOVA

```
Analysis of Variance Table

Model 1: price ~ rating
Model 2: price ~ poly(rating, 2)
Model 3: price ~ poly(rating, 3)
Model 4: price ~ poly(rating, 4)
Model 5: price ~ poly(rating, 5)
Model 6: price ~ poly(rating, 6)
Model 7: price ~ poly(rating, 7)
  Res.Df       RSS Df Sum of Sq        F    Pr(>F)
1  30050 201703495
2  30049 200437508  1   1265987 190.0506 < 2.2e-16 ***
3  30048 200199628  1    237880  35.7107 2.315e-09 ***
4  30047 200154125  1     45503   6.8309  0.008964 **
5  30046 200143414  1     10710   1.6079  0.204803
6  30045 200141575  1      1839   0.2761  0.599243
7  30044 200132607  1      8968   1.3463  0.245942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

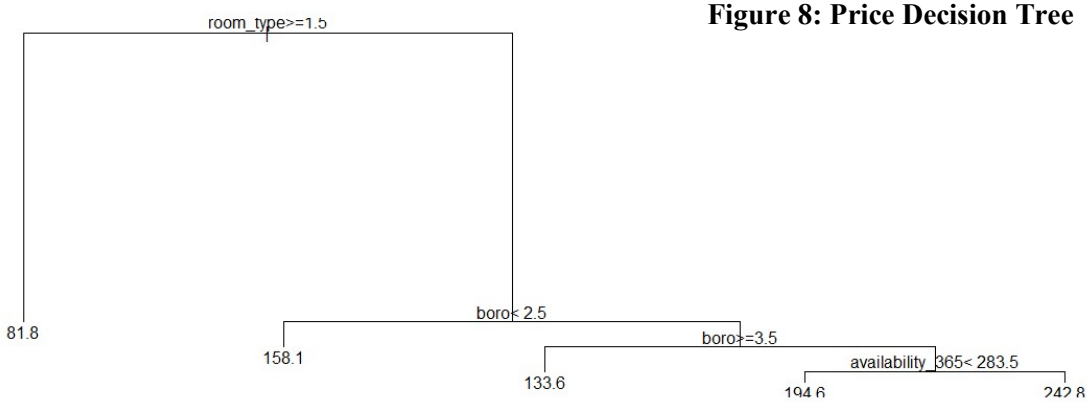
PCR yields a test error of 4422.968. The final method used is partial least squares which is the supervised alternative of PCR. PLS yields a test error of 4014.465, which is by far the best estimate of all the methods, therefore PLS delivers the best fitting model of the shrinkage methods as seen in Figure 5. However, the test error estimate is still rather large.

Figure 7: GAM results



For the price prediction, more flexible nonlinear methods are tried to be applied such as spline, GAM and regression methodologies. The optimal polynomial degree when selected by cross validation is 5. However, when evaluating by p value and RSS, the optimal degree is 4. This is decided by evaluating the best degrees of freedom using the ANOVA test, which shows a substantial decrease in RSS, as seen in Figure 6. Spline also shows the best fit for a degree of 4. The data was fit using GAM with six predictors, price is shown to have nearly linear positive

Figure 8: Price Decision Tree

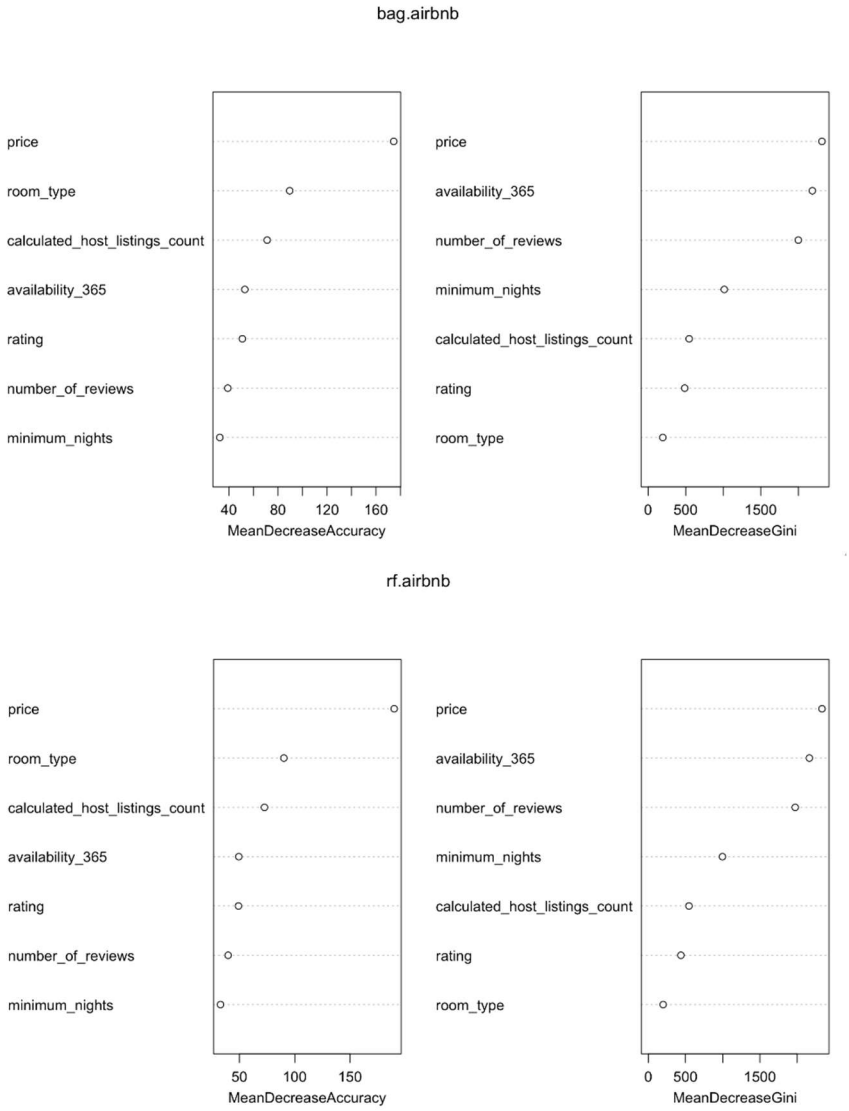


associations with availability and negative association with the number of reviews. Rating also appears to have a linearly increasing relationship with price, with Manhattan being the borough with the highest prices by far. These relationships can be viewed in Figure 7 above.

For another useful interpretation methodology, tree-based method is used for price interpretation segmenting the predictor space into several simple regions. Figure 8 demonstrates that the regression tree above was pruned with an optimal value of 5 calculated to be 7274.2.

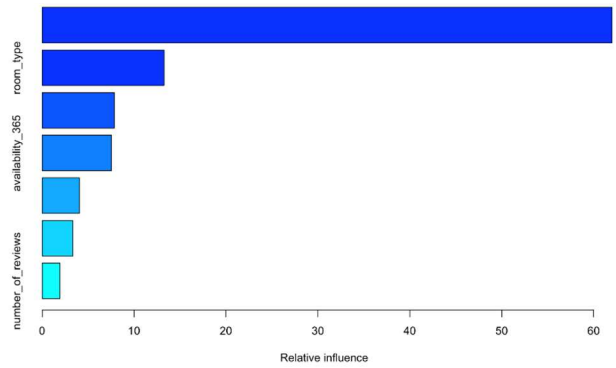
The other investigation being completed with this dataset is whether we can predict the borough the property is located using the other covariates of the dataset. Methods used for this analysis include classification tree, bagging, random forest, boosting, LDA and SVM. The goal of random forest is to improve the performance of decision trees by averaging the variance of the trees. This method creates a strong model by balancing the bias variance tradeoff and yields a test error of 0.42932. Fitting

Figure 9: Bagging & Random Forest Covariate Results



a classification tree with a cross validated pruning parameter of tree size 3 yields a test error of 0.43445. Because this is a simple elementary analysis of the relationship between the borough and the covariates, bagging is used to fit several models on repeatedly sampled different subsets of the data with replacement. The corresponding models are then averaged and give a calculated test error of 0.4083. The analysis also indicates that price and availability are the most important variables in the analysis. Bagging and random forest summary plots of the accuracy and significance of the covariates can be viewed in Figure 8.

Figure 10: Boosting Covariate Influence



When using the sequential model building method of boosting, the calculated test error is 0.3156 with the most important variable being price followed by room type, rating and availability. Figure 10 displays a bar plot of the influence of the covariates on the response of borough.

An attempt at support vector machine was made using a sample of 200 observations from the original dataset, which separates the sample using hyperplanes. This method is very useful when the distribution of the data is unknown. To increase the accuracy of this method, the svm model was tuned with the parameter for linear type with an optimal cost of 0.01 and a gamma of 0.001 which calculates a test error of 0.52749. The model was then tuned with a radial type

Figure 11: SVM Prediction Results

```
> table(airbnb_svm_sample_test$boro , test.tune.pred1)
               test.tune.pred1
                Bronx Brooklyn Manhattan Queens Staten Island
Bronx              16       88        94     21            12
Brooklyn          159     2256      3233    425            67
Manhattan          95     2568      4082    228            41
Queens             66      540       608    141            49
Staten Island       5       34        36     10             2
> table(airbnb_svm_sample_test$boro , test.tune.pred)
               test.tune.pred
                Bronx Brooklyn Manhattan Queens Staten Island
Bronx               0        2       229      0             0
Brooklyn            0       48      6092      0             0
Manhattan           0       33      6981      0             0
Queens              0       16      1388      0             0
Staten Island       0        0        87      0             0
```

parameter, with optimal cost as 100 and gamma as 0.5. This yields a test error of 0.56326 while LDA achieves a measure of 0.5699. LDA is ideal for dimensionality reduction and viewing what variables are important to distinguish the differences in the response variable. Although a ROC curve cannot be plotted due to the multiple levels of the response variable. There is theory being investigated currently to plot categorical variables of multiple levels however nothing is proven at this time.

Of all the results investigating variation of price, boosting achieves the best fit of the data and classification tree produces the worst performance. With the multiple predictors in the data, boosting is easier to tune and unlikely to overfit the data compared to the other methods. In conclusion, the two analyses conducted on this dataset don't reproduce ideal prediction values. This is most likely due to the skewed nature of the data and that most of the properties are in Manhattan and have the highest price overall.

4. Conclusion

In our project, out team tries to find the best model for interpretation on price on housing by applying the regularization methods. As a result, the PLS method reports the least test error for interpretability. For classifying the categorical variable "borough" as a main variable, classification tree, bagging, random forest boost, and svm are used. Classification tree has the worst performances. After the reduction of bias-variance tradeoff, boost has the least test error. Price is the most important variable to predict determining a housing with Manhattan having the highest price. Room type and availability are as well of importance.