Predictive Analytics: Final Report

ECON 386 | May 22, 2020

Paige Junker, Ila Foskett, Margaret Fang, Tiffany Elamparo, Joseph Sunseri, Musaab Alhajeri

I. Executive Summary

For our final project, we decided to analyze a data set encompassing Airbnb listings throughout San Diego. On the official Airbnb website, detailed datasets are provided for most all cities that have Airbnb hosts, which is aggregated into comma-separated value files on Insider Airbnb. Our team chose to take a closer look at San Diego, as we all have a personal connection to this city and were interested in learning more about the attributes of specific listings depending on their location. The dataset that we analyzed was structured as cross sectional due to the fact that there was no time series component involved but instead was static, or taken at one point in time. We considered taking data from the month in which we began the project in March 2020, but we decided to analyze the data in January 2020 due to the effects of the COVID-19 pandemic. Stay-at-home orders have negatively affected Airbnb, which compelled to use data from a month that would more certainly have normal observations. When looking at the variables given in the dataset, we specifically focused on the quantitative variables that were attributes about the listings such as the number of bedrooms/bathrooms and how many people the location could accommodate. We also utilized the room type variable to take a closer look at how the size of the property impacted price, as well as analyzing the neighborhood variable to see trends in specific geographic locations. Additionally, we focused on property type, identifying if an Airbnb property is a house or not when we performed the classification task.

II. Preprocessing and Cleaning the Data

Initially, we had issues with exporting the data to R from excel due to the large size of the file. To reduce the size of the file, we began to eliminate variables that we considered irrelevant to our project or unable to process in a predictive analysis. For example, one of the variables was the description, which contained the entire description written by the host for each listing. After reducing the data to 20 variables we were interested in analyzing, we uploaded the file to the github repository and stored it as df_0 in the R script. To keep track of each step of the cleaning process, we created a new data frame with every altercation. We also used the dimension and view functions for each step in order to visually confirm our omissions. The first thing we noticed was the high volume of NA values for the square feet variable.The square feet variable would seem to be relevant in our model building, but according to the sum(is.na) function, we saw that 13569 of the observations had an NA value for this variable. Instead of deleting each observation that had an NA value, which would have left us with 162 observations, we decided to omit the square feet variable altogether by setting it equal to null. This left us with df_1, which is a dataset that excludes square feet.

```
> sum(is.na(df_0$square_feet))
[1] 13569
```

We continued to reduce our variables to analyze, such as setting host_is_superhost, host_identity_verified, is_location_exact to null. We had the idea of predicting price for the regression task in mind at the time, so we took this into consideration when eliminating variables. The elimination of these variables left us with df_2, which had 13731 rows and 15 columns. The next round eliminations included deleting the variables number_of_reviews_ltm and reviews_per_month for df_3. We already have a variable that addresses the number of reviews as well as other variables that analyze reviews, so we decided to omit these extra variables. We decided to include the id variable as an identifier for each observation.

Next, we dealt with the rest of the NA values in df_3. Using the subset function, we eliminated the observations that had NA values for host_response_rate and review_scores_rating. The new data frame without these observations was df_4, but we could see that there were still NA values by viewing df_4 and searching "NA". We were able to visually see that there were NA values for host_response_rate, so we omitted those observations under that variable for df_5. We repeated the process for the neighbourhood variable in df_6. At this point, we began the partitioning process of exporting the data to excel and dividing the training and testing data, but we realized there were some observations that had completely blank values for certain variables which should have been taken into consideration for the cleaning process. For df_6, we set all blank values to say "NA" and then used the is.na value to see exactly how many of these NA values were left. Instead of seeing which variables still had the NA values and deleting the observations using the subset function, we decided to use the na.omit function in order to mass delete these observations. This resulted in df_7, which is the final clean data frame we used for the regression and classification tasks.

df_0	13731 obs. of 20 variables
df_1	13731 obs. of 19 variables
df_2	13731 obs. of 15 variables
df_3	13731 obs. of 13 variables
df_4	11376 obs. of 13 variables
df_5	9515 obs. of 13 variables
df_6	9222 obs. of 13 variables
df_7	9151 obs. of 13 variables

Exploratory Analysis

Summary Statistics:

```
       id            host_response_rate neighbourhood       zipcode          property_type       room_type
 Min.   :    8488   Min.   :0.0000     Length:6406        Length:6406        Length:6406        Length:6406
 1st Qu.:14423546   1st Qu.:1.0000     Class :character   Class :character   Class :character   Class :character
 Median :23061448   Median :1.0000     Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :22878425   Mean   :0.9701
 3rd Qu.:33245520   3rd Qu.:1.0000
 Max.   :41349667   Max.   :1.0000
 review_scores_rating
 Min.   : 20.00
 1st Qu.: 94.00
 Median : 97.00
 Mean   : 95.28
 3rd Qu.: 99.00
 Max.   :100.00
```

accommodates	bathrooms	bedrooms	beds	price	number_of_reviews
Min. : 1.000	Min. :0.000	Min. : 0.000	Min. : 0.000	Min. : 0.0	Min. : 1.0
1st Qu.: 2.000	1st Qu.:1.000	1st Qu.: 1.000	1st Qu.: 1.000	1st Qu.: 88.0	1st Qu.: 6.0
Median : 4.000	Median :1.000	Median : 1.000	Median : 2.000	Median : 139.0	Median : 24.0
Mean : 4.766	Mean :1.512	Mean : 1.678	Mean : 2.514	Mean : 246.8	Mean : 51.7
3rd Qu.: 6.000	3rd Qu.:2.000	3rd Qu.: 2.000	3rd Qu.: 3.000	3rd Qu.: 250.0	3rd Qu.: 67.0
Max. :24.000	Max. :8.500	Max. :10.000	Max. :22.000	Max. :10000.0	Max. :838.0

After we finished cleaning the data and eliminating a large number of the variables, we ended up with 9151 rows, 13 columns, meaning 9151 observations and 13 variables. The above table shows us the summary statistics for the entire data set. The price variable ranges from $0 to $10,000, with the average price of a listing in San Diego being $246.80. A price of $10,000 appears to be an outlier, but we found 32 listings with prices at this amount. A price of $0 also appears as an outlier, but there are two listings with a price of $0 that have relatively normal values for the rest of the variables. This seems illogical to have a free listing, so this may be a mistake attributed to the original collector of that data. It was also interesting to see the number of bathrooms, bedrooms, and beds with a minimum value of 0, as one would believe that each listing would have at least one of each of these features. It could possibly be due to listings using descriptors such as “half-bathrooms” or “pull-out beds” for these variables, which may have been counted as zero rather than 0.5 or 1.The host response rate is a percentage, so for the variable to have a maximum of 1.0 is consistent with this fact. The ID variable was listed as a quantitative variable, but as we addressed earlier, it only exists in the dataset as an identifier.

Correlation and covariance matrices:

Correlation matrix:

	accommodates	bathrooms	bedrooms	beds	price	number_of_reviews	review_scores_rating
accommodates	1.00000000	0.71430367	0.82975545	0.86577626	0.24344899	-0.13914042	-0.07152821
bathrooms	0.71430367	1.00000000	0.78783049	0.68856648	0.23849057	-0.18180778	-0.02939068
bedrooms	0.82975545	0.78783049	1.00000000	0.80514646	0.23119945	-0.17657245	-0.03509141
beds	0.86577626	0.68856648	0.80514646	1.00000000	0.22126036	-0.14047748	-0.06979975
price	0.24344899	0.23849057	0.23119945	0.22126036	1.00000000	-0.10010015	-0.01278704
number_of_reviews	-0.13914042	-0.18180778	-0.17657245	-0.14047748	-0.10010015	1.00000000	0.08183312
review_scores_rating	-0.07152821	-0.02939068	-0.03509141	-0.06979975	-0.01278704	0.08183312	1.00000000

Covariance Matrix:

	accommodates	bathrooms	bedrooms	beds	price	number_of_reviews	review_scores_rating
accommodates	9.865620	1.9105403	3.1687509	5.3976303	484.39496	-32.17992	-1.4827146
bathrooms	1.910540	0.7251406	0.8156798	1.1638373	128.65059	-11.39969	-0.1651729
bedrooms	3.168751	0.8156798	1.4782625	1.9430588	178.07052	-15.80768	-0.2815753
beds	5.397630	1.1638373	1.9430588	3.9397669	278.20717	-20.53106	-0.9143387
price	484.394956	128.6505880	178.0705156	278.2071670	401290.50380	-4669.10438	-53.4585707
number_of_reviews	-32.179916	-11.3996889	-15.8076803	-20.5310599	-4669.10438	5421.74153	39.7664215
review_scores_rating	-1.482715	-0.1651729	-0.2815753	-0.9143387	-53.45857	39.76642	43.5547994

By finding the correlation between price and these variables, we were able to pinpoint which to focus on. All of the variables within the correlation and covariance matrices are

quantitative variables, while property type, room type, neighborhood are qualitative. We were most interested in price, as we had decided to predict this variable for the regression task. The highest correlation with price was accommodates at 0.243, so we took this into consideration when building our models. The lowest correlation with price was review scores rating at -0.013. In the covariance matrix, the highest correlation with the residuals of price exist with the residuals of the number of reviews at -4669.104.. The lowest correlation with the residuals of price is with the residual of review scores rating at -53.459. It was interesting to see that there were two different variables that had the highest correlation with price for the matrices.

III. Model Proposals - Regression Task

For the regression task, we developed models that would predict the price of airbnb listing given certain variables. We mainly took the quantitative variables into consideration, such as the number of reviews or host response rate. We found certain classes under the neighborhood variable had a high level of significance, so some of the models below include dummy variables of neighborhoods with significance such as Corridor, GrantVille, and La Jolla.

Model 1

Price = -62.375 + 28.967(accommodates) + 88.450(bathrooms) + 3.649(bedrooms) + 0.558(beds) - 0.499(number of reviews)+ 0.581(review scores rating)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1029.8  -101.1   -50.0     8.3  9806.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -62.3747   112.4294  -0.555    0.579
accommodates        28.9668     5.4691   5.296 1.22e-07 ***
bathrooms           88.4502    14.8500   5.956 2.72e-09 ***
bedrooms             3.6486    13.4382   0.272    0.786
beds                 0.5576     8.1003   0.069    0.945
number_of_reviews   -0.4948     0.1059  -4.670 3.07e-06 ***
review_scores_rating  0.5812     1.1649   0.499    0.618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 610.9 on 6399 degrees of freedom
Multiple R-squared:  0.07101,   Adjusted R-squared:  0.07014
F-statistic: 81.52 on 6 and 6399 DF,  p-value: < 2.2e-16
```

```
                          2.5 %      97.5 %
(Intercept)        -282.7738818 158.0245708
accommodates         18.2455154  39.6880831
bathrooms            59.3391270 117.5612067
bedrooms            -22.6947281  29.9919554
beds                -15.3217715  16.4369687
number_of_reviews    -0.7024819  -0.2871037
review_scores_rating -1.7024210   2.8648108
```

In this initial model, we used a majority of the quantitative variables provided in our dataset. This included accommodates, number of bathrooms, number of bedrooms, number of beds, amount of reviews, and the overall review rating. After running the model, it was apparent that only three out of the six variables were statistically significant as well as the intercept term which was negative at -62.375. The bathroom variable had the largest effect on price, as increasing the number of bathrooms by +1 would result in a +88.45 increase in price. Thinking about this in terms of practicality, it is difficult to make sense of a negative price as a starting point, which affected our approach when creating the remaining models. The in-sample error for this model was $610.90 with a multiple R-squared of 0.07. This value means that only 7% of the variation in the price variable is explained by the model we created which is notably lower than

desired. Looking at the out-of-sample error, we got $618.49, which in the following models we worked to further reduce in addition the generalization error of 7.59. Lastly, when looking at the confidence intervals for each of the variables, we can be 95% confident that the true mean of the model parameters falls between the numbers depicted in the screenshot above.

Model 2

Price = 0 + 29.640(accommodates) + 88.262(bathrooms) - 0.513(number of reviews)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1017.9  -103.2   -52.9     6.3  9805.0

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
accommodates      29.63975    3.42737   8.648  < 2e-16 ***
bathrooms         88.26164   11.33264   7.788 7.87e-15 ***
number_of_reviews -0.51336    0.09404  -5.459 4.96e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 610.7 on 6403 degrees of freedom
Multiple R-squared:  0.1933,    Adjusted R-squared:  0.193
F-statistic: 511.6 on 3 and 6403 DF,  p-value: < 2.2e-16
```

```
                      2.5 %      97.5 %
accommodates      22.9209662  36.3585317
bathrooms         66.0458713 110.4774039
number_of_reviews -0.6976955  -0.3290146
```

In this model, we used the variables accommodates, bathrooms, and number of reviews to predict price. These were the variables that were significant between the 0.01% and .001% levels in model one. The intercept term was taken out and the model was forced through the origin as an attempt to decrease our in-sample error and out-of-sample error. Removing the intercept term and all negative values makes sense for predicting the price variable, as we should concentrate on values we could realistically receive. This method did work, as we were able to lower the in-sample residual error from $610.9 in model one to $610.7 in model two. Accommodates and bathrooms have positive intercept values, which implies a direct relationship between price and these two predictors. As accommodates and bathrooms increase, price increases as well. The number of bathrooms had the largest effect on the price, as increasing the bathrooms by 1 causes the price to experience a +$88.26 increase. The negative intercept of -0.51 for number of reviews implies an inverse relationship between this variable and price. As the number of reviews increases by +1, price decreases by -$0.70. This is an interesting observation, as it seems like a higher number of reviews would result in a higher price. The p-value of model 2 is < 2.2e-16, which is below the 0.05 significance level used to reject the null hypothesis. The F-statistic shows that there is joint significance in model two, and the multiple R-squared value of 0.1933 means that 19.33% of the variation in prices is explained by the variation in accommodates, bathrooms, and number of reviews. According to the confidence interval function output, we can be 95% confident that the true mean for each model parameter exists between 22.92 and 36.36 for accommodates, 66.05 and 110.48 for bathrooms, and -0.70 and -0.33 for number of reviews.

Model 3

Price = 0 + 33.803(accommodates) + 42.167(bedrooms) - 0.542(number of reviews) + 0.452(review scores rating)

```
Residuals:
   Min     1Q Median     3Q    Max
-791.0 -108.2  -54.4   10.0 9789.9

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
accommodates        33.8028     4.3465   7.777 8.59e-15 ***
bedrooms            42.1674    11.3565   3.713 0.000207 ***
number_of_reviews   -0.5420     0.1059  -5.117 3.19e-07 ***
review_scores_rating  0.4516     0.1638   2.758 0.005830 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 612.4 on 6402 degrees of freedom
Multiple R-squared:  0.1889,    Adjusted R-squared:  0.1884
F-statistic: 372.8 on 4 and 6402 DF,  p-value: < 2.2e-16
```

	2.5 %	97.5 %
accommodates	25.2822739	42.3232707
bedrooms	19.9048339	64.4300250
number_of_reviews	-0.7495761	-0.3343473
review_scores_rating	0.1306351	0.7726477

I ran a variety of combinations of independent variables to see how they interacted with each other and how they would affect price. I noticed that bedrooms and bathrooms often run together would make either one insignificant, which could be possible multicollinearity. Ultimately, I decided to use accommodates, number of bedrooms, number of reviews, and review scores rating as independent variables with an intercept of 0, the coefficients are all statistically significant. All variables except for the variable “number of reviews” had a positive impact on price. Accommodates had the largest contribution to price, where an increase in 1 accommodates would increase price by $33.80. The number of reviews, however, had a negative effect on price where for every review written, the price would drop by $0.54. Some variables were more significant than others, where accommodates, bedrooms, and number of reviews were significant at close to the 100% confidence level while review score ratings were significant at 99%. We can also see how each of the coefficients were within the confidence intervals shown in table 3, which further solidifies the fact that all variables used were significant towards price. The RMSE was the lowest (which will be discussed later in validation) and can be supported by the findings here in terms of determining whether or not this model was a good fit. The Adjusted R-squared was fairly low (~19%) and could have resulted from the nature of our dataset being skewed.

Model 4

Price = 0 + 29.744 (accommodates) + 84.396 (bathrooms) - 0.510 (number of reviews) + 247.419 (Grantville) + 59.908 (La Jolla)

```
Residuals:
   Min     1Q Median     3Q    Max
-989.5 -102.5  -51.7    6.1 9808.3

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
accommodates       29.7435     3.4281   8.676  < 2e-16 ***
bathrooms          84.3956    11.4801   7.351 2.20e-13 ***
number_of_reviews  -0.5095     0.0940  -5.420 6.19e-08 ***
grantville        247.4187   117.6556   2.103   0.0355 *
lajolla            59.9080    29.6449   2.021   0.0433 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 610.4 on 6401 degrees of freedom
Multiple R-squared:  0.1944,    Adjusted R-squared:  0.1938
F-statistic: 308.9 on 5 and 6401 DF,  p-value: < 2.2e-16
```

```
                      2.5 %      97.5 %
accommodates      23.023203  36.4638176
bathrooms         61.890816 106.9004223
number_of_reviews -0.693735  -0.3251861
grantville        16.774391 478.0629280
lajolla            1.794037 118.0218975
```

In this model, we are predicting price by using accommodates, bathrooms, number of reviews, and neighborhoods Grantville and La Jolla to explain the variation in price. As it is appeared above, accommodates, bathrooms, and number of reviews variables are significant at the 0.001% level, but Grantville and La Jolla variables are significant at the 0.05% level and both of those variables are binary variables, meaning the variables state that either the property is located in that neighborhood or not. Also, the intercept term was taken off the model because we forced it to the origin, since price can’t be negative so we thought it would be more realistic to force the model to the origin, which is an attempt to decrease our in-sample and out-of-sample error. Therefore, we believe that our attempt to lower the generalization error was a success, since we were able to lower the in-sample error to $610.40, and the out-of-sample error to $615.59, which resulted in a generalization error of $5.19. Furthermore, accommodates variable shows a positive coefficient which implies a direct positive relationship with price, meaning every increase in the number of accommodations results in a higher price by $29.74, bathrooms variable shows a positive coefficient as well which also implies a direct positive relationship with price, meaning that every increase in number of bathrooms results in a higher price by $84.40, number of reviews variable shows a negative coefficient which implies an inverse negative relationship with price, meaning that every decrease in the number of reviews results in an increase in price by $0.51 which was not expected, Grantville variable shows a positive coefficient which implies a direct positive relationship with price, meaning that if the property is located in GrantVille the price would increase by $247.42, finally, La Jolla shows a positive coefficient which implies a direct positive relationship with price, meaning that if the property is located in La Jolla the price of the property would increase by $59.91. Moreover, the R-squared came out to be 0.1944 which is a measure of the model fit, also it means that the variation in price is 19.44% explained by the independent variables in the model. Lastly, according to the

confidence intervals viewed above, we can be 95% confident that the true mean of the model parameters falls between the numbers shown above for each variable.

Model 5

Price= 0 - 0.494 (number of reviews) + 30.763(accommodates) + 94.431 (bathrooms) -49.201(House)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1094.3  -103.5   -48.4     9.5  9793.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
number_of_reviews -0.49376    0.09421  -5.241 1.65e-07 ***
accommodates      30.76260    3.44612   8.927  < 2e-16 ***
bathrooms         94.43144   11.51504   8.201 2.86e-16 ***
House            -49.20144   16.58194  -2.967  0.00302 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 610.3 on 6402 degrees of freedom
Multiple R-squared:  0.1945,    Adjusted R-squared:  0.194
F-statistic: 386.4 on 4 and 6402 DF,  p-value: < 2.2e-16
```

```
                      2.5 %      97.5 %
number_of_reviews -0.6784472  -0.3090826
accommodates      24.0070564  37.5181512
bathrooms         71.8581078 117.0047645
House            -81.7075971 -16.6952890
```

In this model, we analyzed variables number_of_reviews, accommodates, bathrooms, and house. We controlled for the property type of house, where observations listed as a house received a 1 and all other property types receiving a 0. We look further into the house variable during the classification task, but for this model, it appears to have the second largest effect on price. If the observation were a house, it decreased the price by -49.20, which was an interesting effect to see. It would seem that a house, a larger property, would have a positive effect on price, but the summary function shows us that the opposite is true. Some listings had property names such as “Bungalow” or “Townhouse”, and it is possible that the specificity of these listings would have a more positive impact on price rather than identifying as only “house”. The first three variables were significant in model 2, so we wanted to see the price effect of Airbnb houses versus other property types offered (apartment, single room, etc.) in model 5. The number of reviews reacted as it had in the other models-- an +1 increase in the number of reviews leads to a slight decrease of -0.494 in the price. The accommodates variable shows us that if a property can accommodate +1 person, the price will increase by $30.76. The bathrooms variable is similar, and an extra bathroom in a property will increase the price by $94.43.

Model 6

Airbnb Price =0-79.516(Privateroom) - 166.118(Sharedroom) + 23.903 (accommodates) + 110.722(bathrooms) - 0.431(number of reviews)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1060.2   -97.9   -37.2    17.8  9803.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
Privateroom       -79.516     19.293  -4.122 3.81e-05 ***
Sharedroom       -166.118     76.095  -2.183   0.0291 *
accommodates       23.903      3.651   6.548 6.29e-11 ***
bathrooms         110.722     12.361   8.958  < 2e-16 ***
number_of_reviews  -0.431      0.096  -4.489 7.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 609.8 on 6401 degrees of freedom
Multiple R-squared:  0.1959,    Adjusted R-squared:  0.1953
F-statistic: 311.9 on 5 and 6401 DF,  p-value: < 2.2e-16
```

```
                      2.5 %      97.5 %
Privateroom       -117.3357974 -41.6962433
Sharedroom        -315.2906119 -16.9463674
accommodates        16.7469293  31.0597183
bathrooms           86.4906155 134.9527391
number_of_reviews   -0.6191791  -0.2427938
```

In model 6, we chose to examine the variables accommodate, bathrooms, number of reviews, private room, and shared room. In order to decide which dummy variables from room type we should use, we ran a regression prediction for price using room type. From the summary statistics, both private room and shared room appeared significant with t-values greater than 1.96 and, shared room being significant between the 0.001% and 0.01% levels and private room being significant between the 0% and 0.001% levels.. Although shared room was not significant between the 0% and 0.001% levels. it was still included in the model because of the relatively large adjusted R-squared of 0.1953. We controlled for private room and shared room in model 6 by creating dummy variables for both room types. The other variables, accommodate, bathrooms, and number of reviews, were significant between the 0% and 0.001% levels and had t-values greater than 1.96. In order to decrease in and out of sample error, this model was forced through the origin, removing the intercept term. Amongst these five variables, private room, shared room, and number of reviews have a negative relationship with price while accommodates and bathrooms have a positive relationship with price. A positive relationship means that an increase in a particular variable would positively affect price, whereas a negative relationship means that an increase in a particular variable would negatively affect the price. In the case of the variable bathroom, for each additional bathroom a property has, Airbnb price is predicted to increase by $110.72 on average holding all else constant. The variable private room shows us that if a posting is a private room, rather than a shared or hotel room, Airbnb’s price is predicted to decrease by $79.52 on average holding all else constant. Similarly, shared room shows us that if a posting is a shared room, Airbnb’s price is predicted to decrease by $166.118 on average holding all else constant. These results are consistent with what we predicted. When

examining the confidence interval, we can be 95% certain that the true mean exists between -117.34 and -41.70 for private room, -315.29 and -16.95 for shared room, 16.75 and 31.06 for accommodates, 86.49 and 134.95 for bathrooms and -0.62 and -0.24 for number of reviews.

IV. Validating the Regression Models

In order to validate our models, we created a training and testing partition for our clean data. The clean data, df_7, was exported from R to Excel after installing the "open.xlsx" package. We used the write.xlsx function to export the data frame as a sheet in Excel. Using the random number generator function, we reassigned each of the observations to a new number. We reordered the rows in ascending order of the new numbers, and then took the top 30% of the 9151 observations to put into a separate excel file. This is our testing data set which consists of 2745 observations. Exactly 30% of 9151 is 2745.3, but we rounded down to logically have whole observations. The extra one observation was left in the training data which consists of 6406 observations. We saved these two excel sheets as comma-separated value files and pushed it to the Github repository. We have three datasets called random, training, and testing. Random is the entire dataset in the randomized order, training is the bottom 70% of that data, and testing is the top 30% of that data. These datasets were uploaded to GitHub through the terminal window, and we stored them all as separate variables with the same name on our R script. This partition is what we used for building our models and validating them with the root-mean-square-error function. We also refer to these datasets for the classification task.

Model 1

In model one, the RMSE is 618.4991, so the out-of-sample error is $618.50. Comparatively, the in-sample error is $610.90 proving to be $7.59 less than the out-of-sample error.

Model 2

The RMSE for model two is 618.1637. Compared to the in-sample error value of $610.70, the out-of-sample error is $7.46 higher at $618.16.

Model 3

The RMSE is 619.5536. The in-sample error is $612.50, which is $7.05 lower than the out-of-sample error of $619.55.

Model 4

The RMSE for this model is 615.59, meaning an out-of-sample error of $615.59, with an in-sample error of $610.40. The difference between the two is $5.19.

Model 5

The RMSE for model 5 is 617.457, meaning an out-of-sample error of $617.46. The in-sample error is $610.30, both of which are similar to the other models. The difference between the two is $7.16, which is slightly higher than model 4 and in line with our other models.

Model 6

The RMSE for model 6 is 617.818, seemingly in line with the other models. This out-of-sample error of $617.82 can be compared to the in-sample error or residual standard error of $609.80. Since the out-of-sample error is $8.02 higher than the in-sample error the model performs better in-sample then out-of-sample.

Best Model

Out of the six models that we ran, we determined that model 4 was the best model due to its lowest out-of-sample error value of $615.59. Although model 6 had the lowest in-sample error of $609.80, deciding the best model on the standard of having the lowest out-of-sample error was better, making model 4 the most favorable. It makes sense to prioritize low out-of-sample error, as the model's ability to work on datasets outside of which it was created truly determines what qualifies as a good model. This relates to the generalization error, as having a small out-of-sample error indicates the model is a good approximation of the data generating process.

V. Model Proposals - Classification Task

For the classification task, we predicted whether the listing was a house or not from the property_type variable. We created a dummy variable for house so that every entry that had the house property type received a 1, and every other entry that was not listed as such received a 0. One of the outputs details the confidence interval for each model parameter, as well as a value next to each of these pairs. This value is the beta estimate evaluated through the exponent function that raises e to the value of the beta estimate, which assists in our analyses of the odds ratio.

Model 1

House = -3.835 - 0.229(bathrooms) + 1.262(bedrooms) - (0.0003)price - 0.167(accommodates) + 0.023(review scores rating)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2277  -0.8296  -0.6662   1.0338   2.7823

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.8350745  0.4800645  -7.989 1.36e-15 ***
bathrooms          -0.2299910  0.0578412  -3.976 7.00e-05 ***
bedrooms            1.2619798  0.0586426  21.520  < 2e-16 ***
price              -0.0003179  0.0001139  -2.791  0.00526 **
accommodates       -0.1665819  0.0183132  -9.096  < 2e-16 ***
review_scores_rating  0.0228132  0.0049324   4.625 3.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 7172.4  on 6400  degrees of freedom
AIC: 7184.4

Number of Fisher Scoring iterations: 5
```

```
                                    2.5 %     97.5 %
(Intercept)          0.02159973 0.008279305 0.05440173
bathrooms            0.79454077 0.709226383 0.88974313
bedrooms             3.53240815 3.152626238 3.96755103
price                0.99968219 0.999423238 0.99986738
accommodates         0.84655349 0.816479656 0.87725984
review_scores_rating 1.02307538 1.013403662 1.03319709
```

```
      FALSE TRUE
FALSE  3870 1395
TRUE    359  782

               Accuracy : 0.7262
                 95% CI : (0.7151, 0.7371)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3101

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9151
            Specificity : 0.3592
         Pos Pred Value : 0.7350
         Neg Pred Value : 0.6854
             Prevalence : 0.6602
         Detection Rate : 0.6041
   Detection Prevalence : 0.8219
      Balanced Accuracy : 0.6372

       'Positive' Class : FALSE
```

```
      FALSE TRUE
FALSE  1609  634
TRUE    160  342

               Accuracy : 0.7107
                 95% CI : (0.6934, 0.7277)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : 9.784e-14

                  Kappa : 0.2917

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9096
            Specificity : 0.3504
         Pos Pred Value : 0.7173
         Neg Pred Value : 0.6813
             Prevalence : 0.6444
         Detection Rate : 0.5862
   Detection Prevalence : 0.8171
      Balanced Accuracy : 0.6300

       'Positive' Class : FALSE
```

For this model, we aimed to predict whether the listing was a house or not based on the bathroom, bedroom, price, accommodates, and review score variables. After running this model, all of the variables resulted in having the highest level of significance besides price, which was significant at the 0.01% level. Taking a closer look at these variables, the only ones that have a positive relationship are the number of bedrooms and the review score rating at 1.262 and 0.023 respectively. This can be interpreted as the more reviews and higher number of bedrooms, the more likely it is that the listing is a house. Yet, the remaining variables all elicited a negative output which is counter-intuitive to what we predicted. It would be easy to assume that if a property has more bathrooms, a higher accommodation number and a higher price, then it would most likely be a house since it is a bigger space. Even though the betas for these negative relationships are relatively low, they could potentially be explained by the fact that there are different categories for property type and some could be misclassified. Something we mentioned earlier is that there are the categories of “townhouse” or “guesthouse” that are specific labels, and these are excluded from counting as the house variable. It’s possible that the inclusion of these other specific categories would fix this discrepancy. The confidence interval tells us that we are 95% confident that the true average of the model parameters exists between the value pairs above. The highest odds ratio was with bedrooms, saying that if the number of bedrooms increased by +1, the odds of the observation being a house would increase by +

The in-sample error after running the classification matrix resulted in 0.274, found by taking 1 minus the accuracy shown above. For out-of-sample error, the same steps were taken except using a confusion matrix for the testing data. This resulted in an error of 0.289, which we focused on trying to lower in the following models.

Model 2

House = -3.520 -2.229(entirespace) + 0.015(accommodates) - 0.460(bathrooms) + 1.342(bedrooms) + 0.001(number of reviews) + 0.022(review scores rating) + 0.664(host response rate)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7819  -0.7677  -0.4867   1.0252   2.8200

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.5200092  0.5594443  -6.292 3.13e-10 ***
entirespace          -2.2293450  0.0855466 -26.060  < 2e-16 ***
accommodates          0.0149270  0.0196793   0.759  0.44814
bathrooms            -0.4600937  0.0592066  -7.771 7.79e-15 ***
bedrooms              1.3418296  0.0623095  21.535  < 2e-16 ***
number_of_reviews     0.0012117  0.0004279   2.831  0.00464 **
review_scores_rating  0.0216076  0.0051483   4.197 2.70e-05 ***
host_response_rate    0.6635427  0.2843459   2.334  0.01962 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 6432.1  on 6398  degrees of freedom
AIC: 6448.1

Number of Fisher Scoring iterations: 5
```

```
                                2.5 %     97.5 %
(Intercept)          0.02959916 0.009716198 0.08717557
entirespace          0.10759888 0.090904032 0.12712781
accommodates         1.01503901 0.976558536 1.05491210
bathrooms            0.63122447 0.561840394 0.70863513
bedrooms             3.82603728 3.390239477 4.32839229
number_of_reviews    1.00121241 1.000371516 1.00205212
review_scores_rating 1.02184278 1.011742541 1.03238341
host_response_rate   1.94165895 1.123582688 3.43464496
```

```
       FALSE TRUE
FALSE  3498  859
TRUE    731 1318

               Accuracy : 0.7518
                 95% CI : (0.741, 0.7623)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4388

 Mcnemar's Test P-Value : 0.001448

            Sensitivity : 0.8271
            Specificity : 0.6054
         Pos Pred Value : 0.8028
         Neg Pred Value : 0.6432
             Prevalence : 0.6602
         Detection Rate : 0.5461
   Detection Prevalence : 0.6801
      Balanced Accuracy : 0.7163

       'Positive' Class : FALSE
```

```
       FALSE TRUE
FALSE  1429  393
TRUE    340  583

               Accuracy : 0.733
                 95% CI : (0.716, 0.7494)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.4101

 Mcnemar's Test P-Value : 0.05477

            Sensitivity : 0.8078
            Specificity : 0.5973
         Pos Pred Value : 0.7843
         Neg Pred Value : 0.6316
             Prevalence : 0.6444
         Detection Rate : 0.5206
   Detection Prevalence : 0.6638
      Balanced Accuracy : 0.7026

       'Positive' Class : FALSE
```

For model 2, we incorporated the variables of entirespace, accommodates, bathrooms, bedrooms, number_of_reviews, review_scores_rating, and host_response_rate in our prediction. Entirespace is a dummy variable created from the room_type variable that gives every observation labeled as “Entire home / apt” a 1 and all other labels a 0. All of these variables are significant between the 0% and .001% levels except for the accommodates variable, which lost all of its significance with this combination of variables. Accommodates has a relatively high correlation with bathrooms (0.714) and bedrooms (0.830), so this may have had an effect on its significance in this model in terms of multicollinearity. The summary statistics tell us that increasing the number of accommodations by 1 increases the probability of the observation’s property being a house by 0.015, which is a small amount. It is interesting how a +1 increase in the number of reviews results in an even smaller increase of +0.001, yet is still significant at the .001% level. After using the exponential function on the betas, increasing the number of accommodations by one would result in a +0.015 increase in the odds of the observation being a house. It appears the bedroom variable had the largest effect on the odds of an entry being categorized as a house, as +1 in the number of bedrooms increased the odds by +2.826. This makes sense logically, as we would expect an increased number of bedrooms to have a higher

chance of that property being a house. The confidence interval for each variable from the training data is at the 95% level, telling us that we can be 95% confident that the true mean of each model parameter exists between the paired values listed above. The first output from the confusion matrix function is from testing the model on the training data, which resulted in 75.18% accuracy. The p-value is <2.2e-16, which is below the 0.05 significance level used to reject the null hypothesis. The amount of true positives and negatives exceeded the amount of false positives and negatives as shown in the confusion matrix. 82.7% sensitivity and 60.5% specificity tells us that we had more values that were not categorized as a house versus the values that were categorized as a house. The second output from the confusion matrix function is from running the model on the testing data, which resulted in the expected lower accuracy of 73.3%. The testing data produced arguably close results, as the p-value is <2.2e-16, 0.808 sensitivity, and 0.597 specificity.

Model 3

House = -5.028 + 2.178(privateroom) - 0.014(accommodates) + 1.121(bedrooms) + 0.001(number of reviews) + 0.019(review scores rating)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6667  -0.8056  -0.4879   1.0620   2.6582

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.0280577  0.4975600 -10.105  < 2e-16 ***
privateroom          2.1782707  0.0838216  25.987  < 2e-16 ***
accommodates        -0.0140267  0.0193213  -0.726 0.467858
bedrooms             1.1213904  0.0546687  20.512  < 2e-16 ***
number_of_reviews    0.0014186  0.0004243   3.344 0.000827 ***
review_scores_rating 0.0193577  0.0050777   3.812 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 6467.6  on 6400  degrees of freedom
AIC: 6479.6

Number of Fisher Scoring iterations: 5
```

```
      FALSE TRUE
FALSE  1366  306
TRUE    403  670

               Accuracy : 0.7417
                 95% CI : (0.7249, 0.758)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4487

 Mcnemar's Test P-Value : 0.0003117

            Sensitivity : 0.7722
            Specificity : 0.6865
         Pos Pred Value : 0.8170
         Neg Pred Value : 0.6244
             Prevalence : 0.6444
         Detection Rate : 0.4976
   Detection Prevalence : 0.6091
      Balanced Accuracy : 0.7293

       'Positive' Class : FALSE
```

```
      FALSE TRUE
FALSE  3365  708
TRUE    864 1469

               Accuracy : 0.7546
                 95% CI : (0.7439, 0.7651)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4624

 Mcnemar's Test P-Value : 9.254e-05

            Sensitivity : 0.7957
            Specificity : 0.6748
         Pos Pred Value : 0.8262
         Neg Pred Value : 0.6297
             Prevalence : 0.6602
         Detection Rate : 0.5253
   Detection Prevalence : 0.6358
      Balanced Accuracy : 0.7352

       'Positive' Class : FALSE
```

```
                                2.5 %     97.5 %
(Intercept)          0.006551523 0.002425846  0.0170845
privateroom          8.831021234 7.498918836 10.4163811
accommodates         0.986071234 0.949317229  1.0240389
bedrooms             3.069118617 2.760294104  3.4201047
number_of_reviews    1.001419617 1.000586312  1.0022527
review_scores_rating 1.019546282 1.009605480  1.0299201
```

Using the same independent variables in model 3 of the regression task, we find accommodates to be insignificant. In other words, accommodates has no effect on the type of property we are trying to determine. Model 3 also includes the dummy variable privateroom from room type, where all entries that have private room listed under room type receive a 1 and all other entries receive a 0. When we raised e to the beta parameters, all coefficients returned positive values, which contributes to an increase in the odds of the property being a house. The largest effect was seen with the privateroom variable, as a value of 1 for privateroom would

increase the odds of the observation being a house by 7.831. Looking at the confusion matrices based on training and testing data, we can interpret the reference values in data vs. prediction first by determining the accuracy of our algorithm. With training data, we had 976 observations put into the algorithm and resulted in 670 of them being correctly identified as true (68% specificity or the true positive rate). In our sample, we had 1,769 observations and correctly rejected 1,366 of them. This results in 77.22% sensitivity, where over ¾ of our data was correctly rejected. Overall, this model correctly predicts the outcome of the data about 74% of the time. The model did a better job of rejecting observations than accepting them. As for our testing data, we had 2,177 observations as input and resulted in 1,469 of them being correctly identified (67% specificity). In our sample, we had a total of 4,229 observations and correctly rejected 864. Sensitivity comes out to 79%, where a great majority of our data was therefore accurately rejected. Overall, this model correctly predicts the outcome of our testing data about 75% of the time. Once again, the testing confusion matrix did a better job of rejecting observations than accepting them.

Model 4

House = -3.666 + 2.258 (private room or not) - 0.424 (bathrooms) + 0.789 (host_response_rate) + 1.318 (bedrooms)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7061  -0.7619  -0.5061   1.0631   2.7974

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.66605    0.29299 -12.512  < 2e-16 ***
privateroom       2.25771    0.07984  28.279  < 2e-16 ***
bathrooms        -0.42378    0.05896  -7.188 6.58e-13 ***
host_response_rate  0.78914    0.28595   2.760  0.00578 **
bedrooms          1.31803    0.05057  26.063  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 6435.9  on 6401  degrees of freedom
AIC: 6445.9

Number of Fisher Scoring iterations: 5
```

```
       FALSE TRUE
FALSE  3442  786
TRUE    787 1391

               Accuracy : 0.7544
                 95% CI : (0.7437, 0.7649)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4528

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8139
            Specificity : 0.6390
         Pos Pred Value : 0.8141
         Neg Pred Value : 0.6387
             Prevalence : 0.6602
         Detection Rate : 0.5373
   Detection Prevalence : 0.6600
      Balanced Accuracy : 0.7264

       'Positive' Class : FALSE
```

```
       FALSE TRUE
FALSE  1387  342
TRUE    382  634

               Accuracy : 0.7362
                 95% CI : (0.7193, 0.7527)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4297

 Mcnemar's Test P-Value : 0.1472

            Sensitivity : 0.7841
            Specificity : 0.6496
         Pos Pred Value : 0.8022
         Neg Pred Value : 0.6240
             Prevalence : 0.6444
         Detection Rate : 0.5053
   Detection Prevalence : 0.6299
      Balanced Accuracy : 0.7168

       'Positive' Class : FALSE
```

```
                                  2.5 %      97.5 %
(Intercept)          0.02557725 0.01420293  0.04489597
privateroom          9.56116735 8.18294491 11.19040957
bathrooms            0.65457106 0.58290527  0.73448644
host_response_rate   2.20150270 1.27130166  3.91072714
bedrooms             3.73606809 3.38713436  4.12991419
```

In this model we are predicting whether the property is a house or not, we are using privateroom, bathrooms, host_reponse_rate, bedrooms. We chose to use these variables as independent variables in explaining the model because of the high level of significance these variables hold. The privateroom variable describes whether the room type is private or not and it holds a level of significance of less than 0.001% Bathrooms variable describes the number of bathrooms the property holds and its level of significance is less than 0.001%. The host_response_rate variable describes the adequacy of the host responses to the tenants and its

level of significance is between 0.01% and 0.001%. Bedrooms variable describes how many bedrooms the property holds and its level of significance is less than 0.001%. Moreover, the intercept has a coefficient of -3.666 which means that the probability of the property being not a house is higher than being a house. The Privateroom variable has a coefficient of 2.258 which means that it has a positive linear relationship with the dependent variable house, and as the probability of the room type being private gets higher the higher the probability of it being a house. However, the bathrooms variable has a coefficient of -0.424, which means that bathrooms variable has a negative relationship with the house variable, and as the number of bathrooms increases the probability of the property being a house is lower. This is the opposite of our prediction before constructing the model. Also, the host_response_rate has a coefficient of 0.789, which means that host_response_rate has a positive relationship with typehouse variable, and as the host_response_rate increases the probability of the property being a house increases. Lastly, bedrooms variable has a coefficient of 1.318, which means that it has a positive relationship with the typehouse variable, and as the number of bedrooms increases the probability of the property being a house increases. Additionally, this model has an accuracy of 73.62% of achieving the correct prediction on the testing dataset, 75.44% of achieving the correct prediction on the training dataset. The model achieved a sensitivity of 78.41% using the testing dataset, which means that the model correctly predicted 1387 out of 1769 of properties being not a house. Also, it achieved a specificity of 64.96% using the testing dataset, which means that the model correctly predicted 634 out of 976 of the properties being a house. Lastly, the confidence intervals next to the exponent function values of the beta parameters tells us that we can be 95% that the true mean of the model parameters exists between those paired values above.

Model 5

House= -4.428 + 0.026(review scores rating)+ 0.001(number of reviews) + 0.797(bedrooms)-0.001(price)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1896  -0.7974  -0.6954   1.0438   3.7122

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.4282143  0.4859021  -9.113  < 2e-16 ***
review_scores_rating  0.0257944  0.0050050   5.154 2.55e-07 ***
number_of_reviews     0.0008637  0.0003903   2.213   0.0269 *
bedrooms              0.7971062  0.0340691  23.397  < 2e-16 ***
price                -0.0006563  0.0001565  -4.194 2.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 7281.0  on 6401  degrees of freedom
AIC: 7291

Number of Fisher Scoring iterations: 6
```

```
       FALSE TRUE
FALSE  3857 1348
TRUE    372  829

               Accuracy : 0.7315
                 95% CI : (0.7205, 0.7423)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3286

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9120
            Specificity : 0.3808
         Pos Pred Value : 0.7410
         Neg Pred Value : 0.6903
             Prevalence : 0.6602
         Detection Rate : 0.6021
   Detection Prevalence : 0.8125
      Balanced Accuracy : 0.6464

       'Positive' Class : FALSE
```

```
       FALSE TRUE
FALSE  1612  622
TRUE    157  354

               Accuracy : 0.7162
                 95% CI : (0.6989, 0.733)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : 7.828e-16

                  Kappa : 0.3067

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9112
            Specificity : 0.3627
         Pos Pred Value : 0.7216
         Neg Pred Value : 0.6928
             Prevalence : 0.6444
         Detection Rate : 0.5872
   Detection Prevalence : 0.8138
      Balanced Accuracy : 0.6370

       'Positive' Class : FALSE
```

```
                                    2.5 %     97.5 %
(Intercept)          0.01193578 0.00452242 0.03039444
review_scores_rating 1.02612993 1.01628404 1.03642535
number_of_reviews    1.00086408 1.00009101 1.00162364
bedrooms             2.21911004 2.07812407 2.37458008
price                0.99934390 0.99902952 0.99963374
```

In this model, we are using the review scores rating, the number of reviews, the bedrooms, and the price of the property in order to predict whether or not an Airbnb listing is a house. The review scores rating, bedrooms, and price variables were all significant between the 0% and .001% levels. The final variable, number of reviews, had a p-value of .0269 which is also statistically significant. Besides the intercept term, the number of bedrooms has the biggest effect on determining whether or not it is a house, which intuitively makes sense as homes usually have more bedrooms than other Airbnb listings and the relationship between the two raises the probability. For each additional bedroom that a property had, the probability that it is a house raised +0.78. In terms of the odds ratio, when the number of bedrooms increases by +1, the odds of the observation being a house increases by +1.22. The price variable has a negative relationship with housing, signifying that the probability of the observation being a house decreases by -0.0007 with a +1 increase in price. This was an interesting occurrence, as we would expect the chance of the observation being a house to increase along with the price. For the confusion matrix of the training data, there were 829 true positives compared to 372 false positives. There were 3857 true negatives and 1348 false negatives. This relates to the 91.2% sensitivity and the 38.1% specificity, as our model rejected more values than accepted them. The testing data experienced similar trends for the confusion matrix, as we also rejected more values than accepted, and had more accurate true/false values than inaccurate true/false values. The out-of-sample error was 0.284 and an in-sample error was 0.268. These are extremely similar to the other models, and slightly worse than model 3. When examining the confidence interval, we are 95% confident that the true mean value for review scores rating is between 1.016 and 1.036, number of reviews is between 1.00009 and 1.00, bedrooms is between 2.078 and 2.37, and price is between 0.999 and 0.9996.

Model 6

House = -1.642 - 0.173(accommodates) - price - 0.224(bathrooms) + 1.267(bedrooms)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6416769  0.0635430 -25.836  < 2e-16 ***
accommodates -0.1731972  0.0182653  -9.482  < 2e-16 ***
price        -0.0003219  0.0001146  -2.810 0.004954 **
bathrooms    -0.2236975  0.0575589  -3.886 0.000102 ***
bedrooms      1.2662939  0.0585537  21.626  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8211.6  on 6405  degrees of freedom
Residual deviance: 7195.6  on 6401  degrees of freedom
AIC: 7205.6

Number of Fisher Scoring iterations: 5

                                2.5 %    97.5 %
(Intercept)  0.1936550 0.1708269 0.2191540
accommodates 0.8409718 0.8111717 0.8713926
price        0.9996781 0.9994192 0.9998649
bathrooms    0.7995569 0.7140952 0.8948642
bedrooms     3.5476801 3.1667894 3.9839872
```

```
          FALSE TRUE
FALSE  3894 1413
TRUE    335  764

               Accuracy : 0.7271
                 95% CI : (0.716, 0.738)
    No Information Rate : 0.6602
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3088

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9208
            Specificity : 0.3509
         Pos Pred Value : 0.7337
         Neg Pred Value : 0.6952
             Prevalence : 0.6602
         Detection Rate : 0.6079
   Detection Prevalence : 0.8284
      Balanced Accuracy : 0.6359

       'Positive' Class : FALSE
```

```
          FALSE TRUE
FALSE  1614  647
TRUE    155  329

               Accuracy : 0.7078
                 95% CI : (0.6904, 0.7248)
    No Information Rate : 0.6444
    P-Value [Acc > NIR] : 1.094e-12

                  Kappa : 0.2812

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9124
            Specificity : 0.3371
         Pos Pred Value : 0.7138
         Neg Pred Value : 0.6798
             Prevalence : 0.6444
         Detection Rate : 0.5880
   Detection Prevalence : 0.8237
      Balanced Accuracy : 0.6247

       'Positive' Class : FALSE
```

In this model we chose to examine the variables accommodate, price, bathrooms, and bedrooms and examine their significance in determining if an Airbnb property was a house. The variables accommodate, bathrooms, and bedrooms were significant between the 0% and 0.001% levels and the price variable was significant between the 0.001% and 0.01% levels. Amongst these four variables, accommodates, price, and bathrooms have a negative relationship with housing while bedrooms have a positive relationship with price. One surprising relationship was price. The model shows that as price increases, the chance of an Airbnb listing being a house goes down. This is contrary to our hypothesis that an increase in price would indicate a higher chance of a property being a house. The bathroom variable has a coefficient of -0.224, meaning that for every additional bathroom a property has, the chance of it being a house goes down by 0.224. The accommodates variable has a coefficient of -0.173, meaning that for every additional person a listing accommodates, the chance of it being a house goes down by 0.173. The bedrooms variable has a coefficient of 1.267, meaning that for every additional bedroom, the chance of it being a house goes down by -1.267. From the confusion matrices, we were able to see that the amount of true positives and negatives exceeds the amount of false positives and negatives, which corresponds with the accuracy percentages of 72.71% for the training data and 70.78% for the testing data. When examining the confidence interval, we can be 95% certain that the true mean exists between 0.811 and 0.872 for accommodates, 0.999 and 0.999 for price, 0.714 and 0.895. for bathrooms, and 3.167 and 3.984 for bedrooms.

VI. Validating the Classification Models

For validating the classification task, we utilized the same training and testing partition as we did for the regression task. When we calculated the two confusion matrices to find in and out-of-sample error, training data was used to find in-sample error and testing data was used to

find out-of-sample error. We found these error values by computing 1 - accuracy for both matrices.

Model 1

The in-sample error is 27.38% and the out-of-sample error is 28.93%.

Model 2

The in-sample error is 24.82% and an out-of-sample error of 26.7%.

Model 3

The in-sample error is 24.54% and the out-of-sample error is 25.83%.

Model 4

The in-sample error is 24.56% and the out-of-sample error is 26.38%.

Model 5

The in-sample error is 26.72% and the out-of-sample error is 28.34%.

Model 6

The in-sample error is 27.29% and the out-of-sample error is 29.22%.

Best Model

We decided that the best way to choose the best model is to compare the in and out-of-sample errors, and the model with the lowest out-of-sample error would have higher accuracy in predicting whether the property is a house or not. Therefore, out of our six models, we determined that the best model was model 3 since it has the lowest out-of-sample error of 25.83% and lowest in-sample error of 24.54% as well. This means that model 3 has the highest percentage of accuracy (74.17%) when predicting the testing dataset.

VII. Conclusion

The 12 models above analyzed the Airbnb listings throughout San Diego. These models used datasets pulled from the official Airbnb website that was reflective of listings throughout San Diego. The datasets we chose were cross-sectional and although both quantitative and qualitative variables existed, our group chose to focus on the quantitative variables.

For the regression task, we chose to analyze six different models that focused on different variables' ability to predict the price of an Airbnb listing. In order to decrease in and out of sample error, some of these models were forced through the origin, removing the intercept term. Additionally, dummy variables were used in order to most accurately predict price. Specifically, Grant Ville and La Jolla in model 4 which both proved to be positively correlated

with price, house in model 5 which has a negative correlation with price, and shared and private room in model 6, which similarly are negatively correlated. From these six models, we concluded that model 4 was the best model due to its lowest out-of-sample error value. We concluded this despite model 6 having the lowest in-sample error.

The classification task focused on predicting whether or not a San Diego Airbnb listing was a house or not. A dummy variable was created for house that was used as the dependent variable in the six regressions. Additionally, model 2 and model 4 included dummy variables in their independent variables. Model 2 used entire space to predict whether or not a listing was a house and model 4 used private room. Something interesting to note is that when Model 5 and 6 used price to predict whether or not a listing was a house it showed that the price variable has a negative relationship with housing. This was not in line with our predictions. From these six models, we concluded that model 3 was the best model because it had the lowest error and therefore has the highest accuracy in predicting whether the property was a house or not.

Overall, the analysis showed that the models made from our chosen variables did a better job of predicting whether an Airbnb listing was a house or not than predicting price. If we compare the actual price with the predicted price from the best model, the first entry had a predicted price of $201.11 and an actual price of $150. The linear regression models could maybe see more improvement by including the square feet variable if we were to redo the entire regression task to see better results. On the other hand, the models for the classification task have relatively high accuracy percentages.

Through this predictive analytics project, we gained insight on San Diego airbnb listings and the interaction between the data of these listings. Certain results surprised us, such as the negative relationship between price and the house dummy variable. It was interesting to see what neighborhoods in San Diego would be significant. Through this project, we were also successfully developed models to solve both a regression and classification problem. We gained further familiarity with R as well as a better understanding of the concepts we learned during lecture. The project required a balance between knowledge of linear and logistic regressions and logic, as making sure our results intuitively made sense is part of how we can verify certain results. This was a beneficial analysis to complete, as we can apply the skills we learned to future analytics projects.