Predicting Airbnb Rental Prices using Statistical Learning: A Data-Driven Approach

Aanchal Dusija, Ahmed Khair, Karan Uppal

Georgetown University

May 5th, 2023

Table of Contents

I. Introduction

The Airbnb dataset utilized in this project was created by collecting quarterly updates throughout the year 2022, in a seasonal manner. The data was gathered on specific dates, namely March 16th, December 15th, June 8th, and September 12th. Each update captures new listings and changes to existing listings. This is a comprehensive and dynamic source of data for statistical learning projects, as it enables the analysis of how the Airbnb market evolves over time and allows for the identification of trends that may be missed by looking at a single snapshot of the data.

Furthermore, the quarterly updates enable the capture of seasonal variations in the data, such as differences in the average price of Airbnb rentals during the summer months compared to the winter months. By collecting data at different times throughout the year, it is possible to identify these seasonal trends and adjust for them in the analysis, providing a more accurate picture of the Airbnb market.

The utilization of exploratory data analysis and statistical learning techniques allows us to gain some superficial insights from the Airbnb data. After an exploratory analysis, we do more sophisticated analyses like building a model to estimate listing price.

II. Data Science Problems

1. What machine learning models most accurately predict price? How do the models' accuracy vary between them?
2. What are the significant variables within the Airbnb Dataset in relation to price?

III. Data Description

The dataset provided is a collection of records for Airbnb listings in and around Austin, Texas. The data contains information on various aspects of each listing, such as the title of the listing, the type of room available (e.g., entire home, private room, shared room), the nightly price, the minimum number of nights required for a booking, and the availability of the listing. The dataset also includes neighborhood and latitude/longitude information, which can be used to gain a better understanding of the geographic distribution of the listings. This information may be useful for analyzing the supply and demand of Airbnb listings in different areas of Austin.

The dataset also includes information on the number of reviews for each listing, the date of the most recent review, and the average number of reviews per month. This information can be used to assess the popularity of individual listings and identify potential trends in customer satisfaction. The Airbnb dataset provides a rich source of data for analyzing various aspects of the Airbnb market in this area. It includes a wide range of information on individual listings and their hosts, as well as geographic and review data that can provide additional insights into the market.

The columns in our dataset are detailed on the next page:

Column Name	Description
id	Unique identifier for the Airbnb listing
name	Descriptive title of the listing
host_id	Unique identifier for the Airbnb host
host_name	Name of the Airbnb host
neighbourhood_group	The neighborhood group that the listing is located in (if applicable)
neighborhood	The neighborhood that the listing is located in
latitude	Latitude coordinates of the listing location
longitude	Longitude coordinates of the listing location
room_type	The type of room available for rent (e.g., Entire home/apt, Private room, Shared room)
price	The nightly price for the listing in US dollars
minimum_nights	The minimum number of nights required for booking
number_of_reviews	The total number of reviews for the listing
last_review	The date of the most recent review
reviews_per_month	The average number of reviews per month
calculated_host_listings_count	The number of listings that the host has on Airbnb
availability_365	The number of days the listing is available for booking within the next 365 days
number_of_reviews_ltm	The number of reviews in the last 12 months
license	The license number (if applicable) for the listing

IV. Exploratory Data Analysis

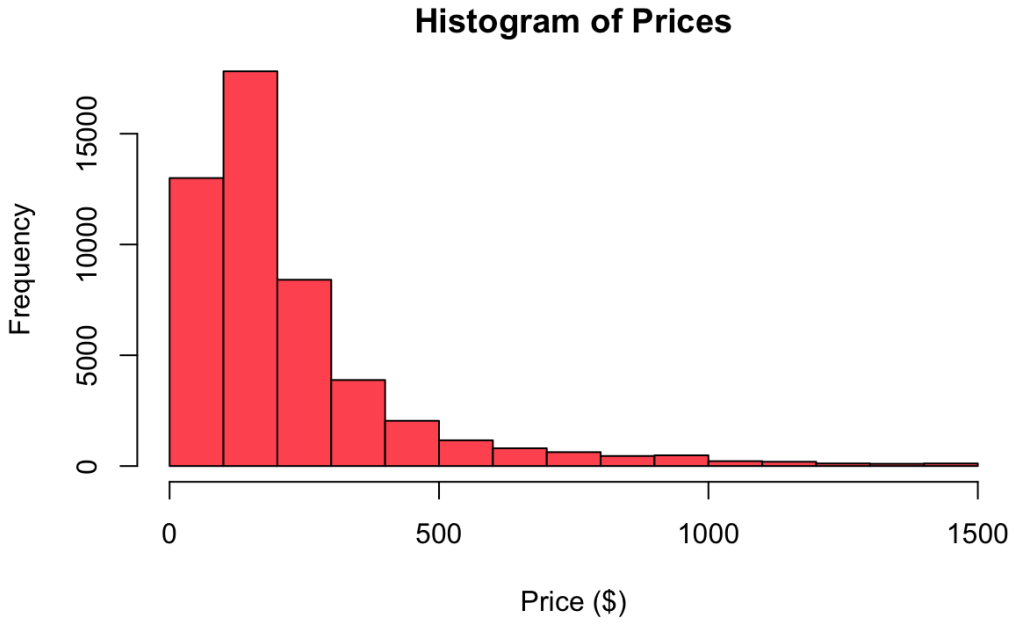

Figure 1: Histogram of Prices

This histogram gives the distribution of listing prices (after removing some outliers, like listings below $30/night and above $1500/night). Here, we see that prices are incredibly skewed right, with a long tail of some very expensive listings. Essentially, we needed to log transform this variable. The use of the natural logarithm transformation is a common technique to transform skewed data into a more symmetric form that can be better visualized and analyzed and often leads to better model performance.

Figure 2: Histogram of Log Prices

This histogram gives the resulting log-transformed price distribution. Here we see that the log-transformation accomplishes its goal of making the data more symmetrical, so we opt to use log-prices in our models.

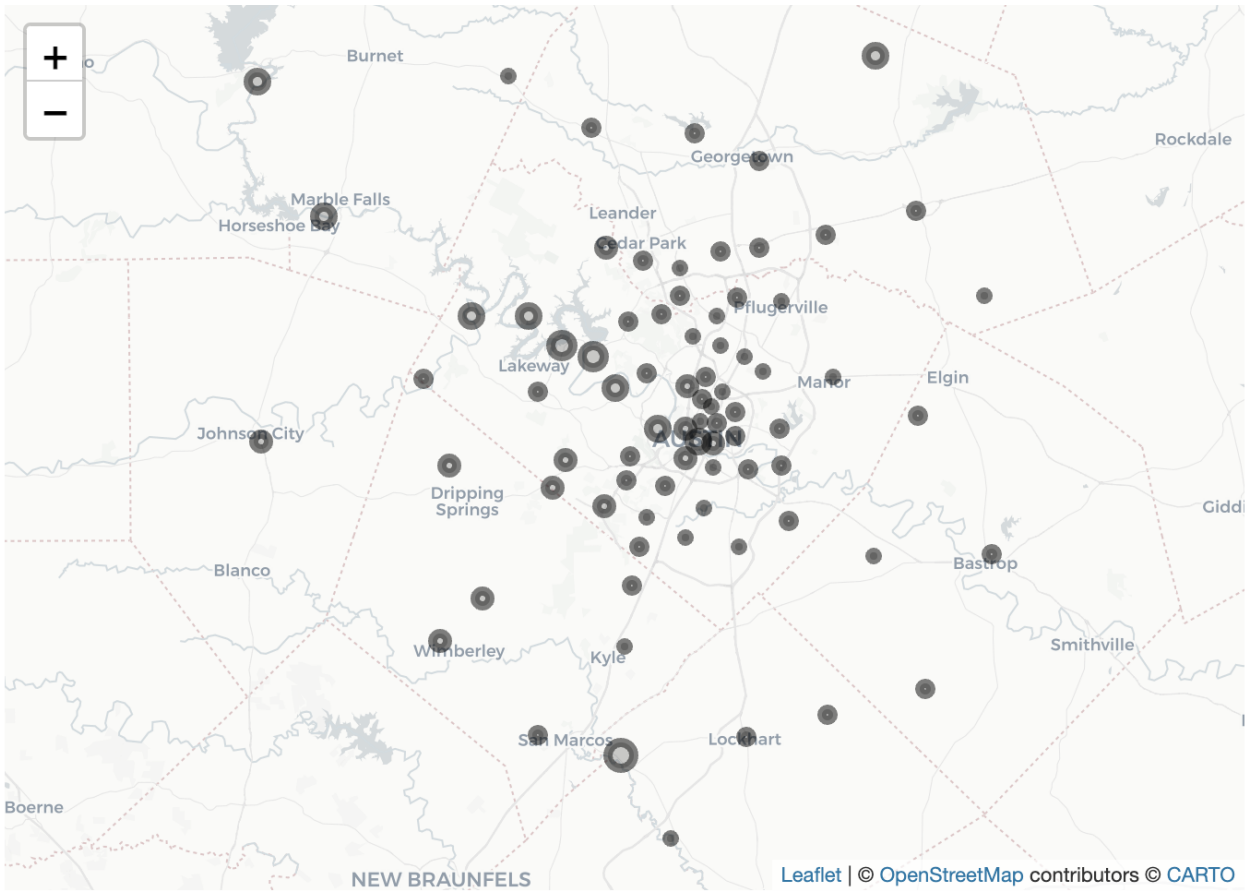

Figure 3: Airbnb Listings for Austin, TX

The plot is a geographic visualization of the neighborhoods in the Austin, Texas dataset, where each neighborhood is represented by a circular marker on the map. The markers are scaled by median listing price of each neighborhood. The Leaflet package is used to create an interactive map that displays the neighborhoods. This plot aims to visualize the spatial distribution of the listings and their corresponding prices across different neighborhoods. The map allows us to identify any geographic patterns or trends in the data, and can provide insights into the factors that influence listing prices in different locations.

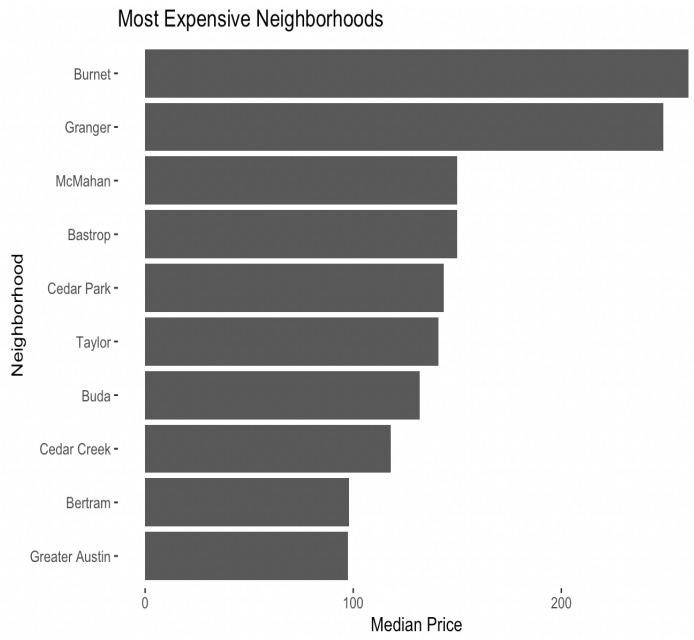

Figure 4: Table of median prices from top 10 listed neighborhoods within Austin, Texas

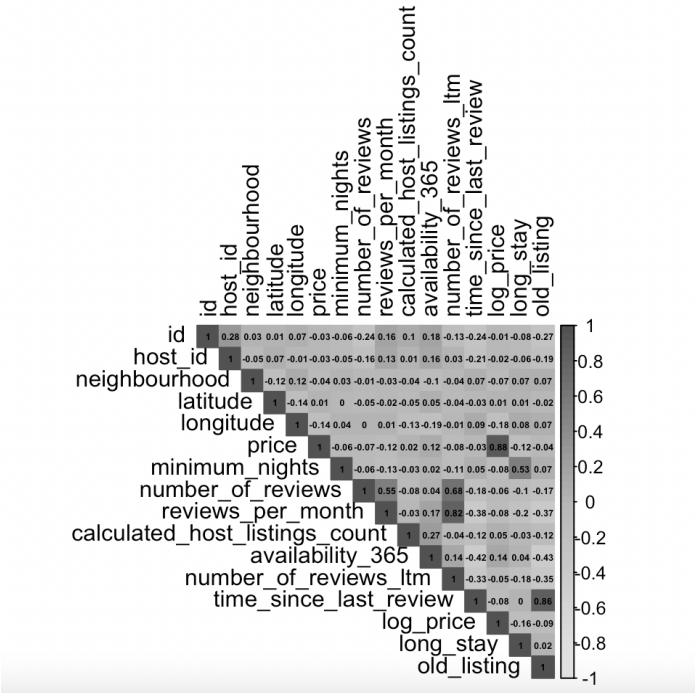

Figure 5: Correlation matrix of confounding variables used within our models

As shown above, *Figure 4* portrays a bar plot of the median prices of listings in the top 10 most expensive neighborhoods in the dataset using the ggplot2 package. This plot is a part of the exploratory data analysis process that aims to identify the most expensive neighborhoods in the dataset and visualize the differences in median prices between them. The bar plot provides a clear and concise representation of the relative prices in each neighborhood, which can help inform pricing strategies for hosts and provide insights into the demand for listings in different areas.

Figure 5, offers a look at a correlation matrix that aims to examine the relationships between the numeric variables in the dataset. The correlation matrix and plot can help identify any strong positive or negative correlations between variables, which can inform feature selection and model building in later stages of the analysis. Importantly, it can help us avoid multicollinearity in any linear regression models. The use of a custom color palette makes the plot more visually appealing and easier to interpret.

Figure 5: Median price of property listings from June 2022 - March 2023

The graph displays fluctuations in property prices over time. It suggests that listing prices experienced a drop over several months from March to December. During this period, we also see a stark drop in the number of listings from ~14k to ~10k in December. As a result, we may be seeing that "peak holiday season" in Austin is in the summer, which causes high demand for Airbnb units, which leads to a higher prices and number of listings. During the winter, people are not traveling to Austin as much, leading to lower demand and prices for Airbnb units.

Additionally, the city hosts various events throughout the year, such as the South by Southwest (SXSW) music festival, which attracts many music artists over a 10-day period. This event is a major contributor to the surge in property prices in March as individuals book relevant listings on Airbnb. Austin also hosts other popular events, including the Austin City Lights music festival and the Circuit of the Americas Grand Prix, both of which occur in October, along with the Austin Film Festival and Texas Book Festival (Macias, 2019). In April, the famous Austin Food and Wine Festival takes place as well (Macias, 2019).

V. Statistical Methods and Results

Algorithm	RMSE	R2	Error
Linear Regression	0.6046127	0.3763792	0.1178848
Polynomial Regression	0.7460255	0.0505914	0.1454569
Spline Regression	0.5942247	0.3976403	0.1158594
Bagging	0.6341104	0.3154346	0.1236362
Random Forest	0.4400912	0.6622657	0.1291104
XG Boost	0.6043712	0.3598725	0.0702312

Table 1 : Error Performance indicators used to evaluate models using R^2 and RMSE

1. Linear Regression

Our initial base model was a linear regression. Starting with a linear regression is a good way to identify which variables have a strong linear relationship with the outcome variables (Log Price). We also use the linear model in a technique known as feature selection, where significant variables are chosen for our following models. Ultimately, we see that the linear model fits the data relatively poorly, as it had a low test R^2 and high test RMSE at 0.38 and 0.61, respectively.

From this model we were able to grab the listed variables for our other models (variable dataset names in parenthesis):

1. Room Type (room_type)
2. Reviews per month (reviews_per_month
3. Total number of Reviews (number_of_reviews)
4. Neighborhood zip area codes (neighborhood)
5. Number of times previously listed (old_listing)
6. Listings with long-term stays (long_stay)
7. Date (date)
8. Days available throughout the year (avilability_365)
9. Number of reviews in the last 12 months (number_of_reviews_ltm)
10. Time Since Last Review (time_since_last_review)
11. Calculated Host Listing Count (calculated_host_listings_count)

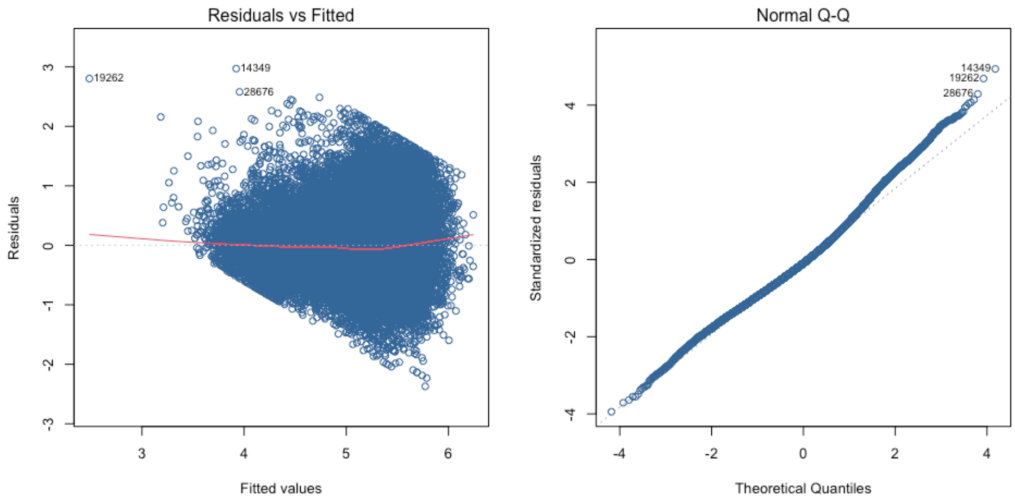

Figure 7: Fitted and residual plot and Quantile plot

As we can see from the graphs above, the residuals vs fitted residuals plot shows an increased spread as the fitted values increase. This suggests potential heteroscedasticity. This inherent character of the model tells us that the significance tests for the regression coefficients of the model may be invalid. We can also note that the QQ plot of the standardized residuals suggests that the residuals are normally distributed, which is a major assumption behind regression.

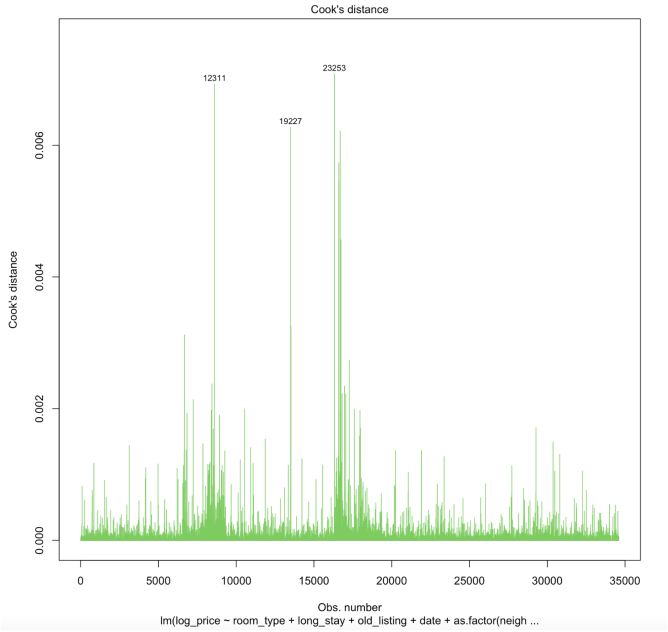

Figure 8: Cook distance plot from linear regression model

Figure 8 is an important note to our linear regression as this diagnostic tool shows the measure of how much the regression coefficient would change if the observations were removed from the dataset. Since most of the leverage points had a small cook distance measure it suggests that the model is not influential.

2. Polynomial Regression

Adding polynomial terms to a regression allows for a nonlinear relationship between our explanatory variables and log price. If the squared terms of the numeric variables in the polynomial regression are statistically significant, this suggests that a nonlinear relationship between that explanatory variable and log price should be explored.

The results of the polynomial regression suggest that the following numeric variables have a nonlinear relationship with log-price: reviews_per_month, availability_365, number_of_reviews_ltm, time_since_last_review, and calculated_host_listing_count. Since we see that the majority of our numeric variables may be nonlinear, we next choose to implement a spline regression.

3. Spline Regression

The spline regression method utilizes knots (breakpoints) to allow for a nonlinear relationship between the explanatory variables and the predictor (log price). Since the earlier polynomial regression suggests most of our numeric variables had a nonlinear relationship with log price, we allow for knots in all our numeric variables in the spline regression. We also include our non-numeric variables (i.e. dummies like long_stay and factors like room_type).

This regression shows improvement from the initial linear regression. The test R^2 is slightly higher (0.398 vs 0.376), and the test RMSE is lower (0.59 vs 0.60).

4. Bagging

Bagging is a random forest methodology that utilizes bootstrapping techniques to create multiple regression trees from the data, which in turn reduces variance and improves performance. This is an ensemble methodology.

Our bagging model performed poorly, however. It had lower test R^2 and considerably higher test RMSE than either the Spline Regression or the simple Linear Model. This suggests that we should not use this model to predict Airbnb listings prices.

5. XGBoost

XGBoost uses a variety of regularization techniques to prevent overfitting, including L1 and L2 regularization, early stopping, and tree pruning. It also includes a built-in feature selection mechanism that can automatically identify the most important features in the dataset, which can help improve the accuracy of the model and reduce the risk of overfitting.

Despite this, the XGBoost performed poorly. It only outperformed the Bagging model, and had a lower test R^2 and higher test RMSE than both the Spline Regression and the simple Linear Model.

6. Random Forest

A Random Forest model was applied in our last step to explore the performance based on the predictions of the decision trees. A Random Forest is similar to Bagging in its approach, with the key difference that each "tree" only considers a *subset* of the explanatory variables. This works to "decorrelate" the trees compared to Bagging, and often leads to better model performance.

From the error indicators, the Random Forest model had a 0.66 test R^2 value and 0.44 test RMSE, which were both much better than all our other models. Given this performance, we draw the conclusion that the Random Forest model is the best model for predicting Airbnb listing prices.

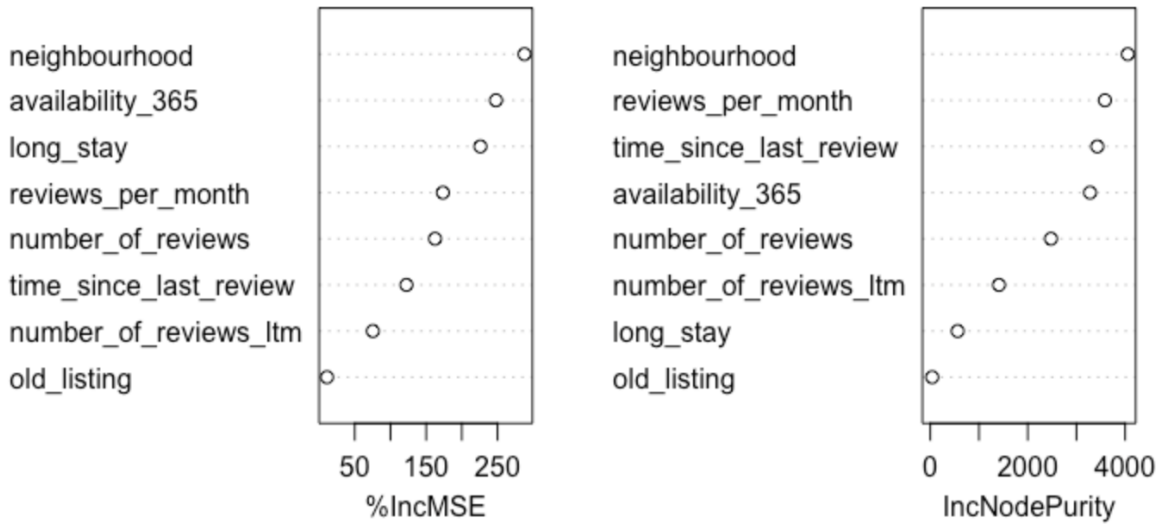

Figure 9: Random Forest model indicating percent increase of MSE scores when the variable was added to the model and Node Purity for dependent variables

Now that we've selected Random Forest as our desired model methodology, we further explore the relationship between our explanatory variables and price through the Random Forest "importance scores."

In *Figure 9*, the "%incMSE" shows how much more poorly the model performs when the variable in question is removed. A higher value indicates the model is "worse off" without that variable, suggesting that the variable is important to the model. Ultimately, we see that Neighbourhood, availability_365, and long_stay are "most important" using this metric.

"IncNodePurity" is another metric that we can use to deduce the importance of an explanatory variable. Using this metric, we see that neighbourhood, reviews_per_month, time_since_last_review, and availability_365 are the most important variables. Ultimately, we can conclude that neighbourhood, reviews_per_month, and availability_365 are the most important variables to our model, as they perform well in both importance measures.

VI. Conclusions

In conclusion, we see that our initial linear, polynomial, and spline regressions offered poor performance in regards to predicting price based on Airbnb listings data with RMSE values of 0.60, 0.75, and 0.59 respectively. This suggests that classical regression methods are inappropriate to predict price, even when nonlinearity is allowed for.

Moreover, we performed three additional models involving Bagging, Xgboost and Random Forest Techniques. These techniques were used to identify predictability of the log price variable in a manner using ensemble methods instead of application of mathematical functions between independent and dependent variables.

Two of our three ensemble methods, Bagging and XGBoost, performed similarly to the earlier models (Linear, Polynomial, and Spline). They had test RMSE values of 0.63 and 0.60, respectively.

On the other hand, Random Forest performed exceptionally well in comparison to all models. Random Forest may have performed better due to its ability to reduce the variance of the predictions, as this ensemble method is unique in that it applies an additional step in building the decision tree by randomly selecting subsets of the feature to build the model.

Additionally, we can consider these models relevant as model performance issues regarding over-fitting and under-fitting are reduced since Xgboost incorporates regularization techniques, and Bagging and Random Forest primarily focuses on reducing the variance of the prediction while also increasing the diversity of the individual models.

Beyond this analysis, we need to take into account that there are market forces that are not being accounted for in response to how prices are determined within each individual listing, especially in the case of analyzing one city, Austin, Texas. Price determination may be a seasonal trend that is unique to the market it's in. There is no doubt that the city of Austin is a growing and populated city, yet Airbnb may be offering listings that only appeal to tourism. Additionally, we must account for seasonal trends behind how prices are determined, and not by their inherent property characteristics.

VII. Future Work

For predicting Airbnb rental prices using statistical learning, there are several potential ways to improve the accuracy of the models. One way is to add more sophisticated models to the analysis, such as neural networks, which are known for their ability to capture complex relationships in data. Neural networks can potentially improve the predictive accuracy of the model by uncovering patterns and trends that may not be captured by other models.

Another way to improve the accuracy of the model is to add more data to the analysis. In particular, adding more data from different dates can help capture seasonal variations in rental prices. For example, prices may be higher during peak tourist seasons or during major events in the area. By adding more data from different dates, the model can better capture these variations and provide more accurate predictions. We can also add more variables to

improve our predictions. For instance, we can incorporate more information about the properties themselves (i.e. square footage, # of rooms, # of bathrooms, etc).

Finally, adding data on economic conditions such as demand and supply can also help improve the accuracy of the model. This data can include factors such as local employment rates, population growth, and the availability of short-term rental properties. By incorporating this data into the analysis, the model can better account for factors that may affect rental prices and provide more accurate predictions.

In conclusion, by incorporating additional models and data, the accuracy of the model for predicting Airbnb rental prices can be improved. This can lead to more informed pricing decisions for hosts and a better experience for guests.

VIII. References

1. Gibbs, C., Guttentag, D., Gretzel, U., Yao, L., & Morton, J. (2018, January 8). *Use of dynamic pricing strategies by Airbnb hosts*. International Journal of Contemporary Hospitality Management. Retrieved May 5, 2023, from https://www.emerald.com/insight/content/doi/10.1108/IJCHM-09-2016-0540/full/html?skipTracking=true
2. Inside Airbnb. (n.d.). *Get the Data*. Inside Airbnb. Retrieved April 25, 2023, from http://insideairbnb.com/get-the-data/
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Resources - ISLR second edition*. An Introduction to Statistical Learning. Retrieved April 25, 2023, from https://www.statlearning.com/resources-second-edition
4. Macias, R. (2019, June 3). Austin's Top Annual Events. TripSavvy. Retrieved April 30, 2023, from https://www.tripsavvy.com/top-annual-events-in-austin-texas-4582569

IX. Appendix

Github Repository: https://github.com/ahmedkhair1/ANLY-512-Final-Project.git