Department of Data Science

DTSC 3010: Project Report on Machine Learning

Airbnb Price Regression

Author:
Gabriel Allen
Student ID: 900294613

Supervisor:
Dr. Nawa Raj Pokhrel

April 2024

Contents

1 Background

Airbnb is a peer to peer rental service that connects homeowners directly to renters looking for short and medium term residency. It was originally meant to fill the gap between hotels and apartments in terms of price and duration, and has become a vital source of income for millions of homeowners worldwide. As such, it is important that the pricing of Airbnbs remains economically viable for homeowners while being affordable to its users. One method of maintaining this balance for homeowners and users is providing full transparency of the factors that affect the pricing of units. This project aims to delineate the factors that determine pricing of Airbnbs specifically for the city of New Orleans. The input to my algorithms are [host-response-time, host-is-superhost, host-identity-verified, neighbourhood-cleansed, property-type, room-type, accommodates, bedrooms, beds, minimum-nights, maximum-nights, number-of-reviews, review-scores-rating, review-scores-accuracy, review-scores-cleanliness, review-scores-checkin, review-scores-communication, review-scores-location, review-scores-value, instant-bookable, reviews-per-month]. I then use various models to output a predicted [price].

2 Related Work

The problem of Airbnb price prediction using machine learning techniques, while not thoroughly saturated, has been explored before. In a paper published in 2021, Kalehbasti et al. [1] compare a wide variety of machine learning models to predict Airbnb pricing while maintaining a particular focus on sentiment analysis. While they only use data from New York City, they compare the results of a regression analysis using Ridge regression, k-means clustering, support vector regression, neural network, and gradient boosting tree ensemble techniques. This provides a balance between more traditional machine learning techniques and newer, state-of-the-art machine learning and deep learning techniques. In a similar paper published in 2019, Luo et al. [2] compare various tree ensemble techniques and neural networks to determine the optimal model for generalizing price regression across different geographical areas. Although this paper only compares performance between random forest, gradient boosting tree ensemble, and multi-layer perceptron neural network, they focus on analyzing data from New York City, Paris, and Berlin. Using data from different cities in different countries allows for the development of more generalized models that are accurate across different locations. Another paper published in 2017, Zhang et al. [3] focuses on intra-city factors by weighting geographical location within a single city. This study only compares a general linear model and a geographically weighted model and only includes data from the metro area of Nashville, Tennessee, but includes data pertaining to the distance a given Airbnb listing is from the nearest highway as well as the distance from the Nashville convention center. The selection of this specific data demonstrates their concentration on determining the location factors within a city that affect pricing. All of these studies use some form of linear model as a baseline to compare the performance of their more complex models. Similarly, this project will focus on comparing the performance of various implementations of linear regression with a tree ensemble regression technique. Specifically,

this project compares simple linear regression (using the number of bedrooms as a predictor), multiple linear regression, Ridge regression, Lasso regression, and random forest regression.

3 Data-set and Features

In this project I use a Kaggle dataset for Airbnb listings in the city of New Orleans, Louisiana. This dataset includes 6028 listings. After cleaning the dataset, which consisted of removing irrelevant columns and removing rows with null values, the dataset was trimmed down to 3715 listings. I employed cross-validation with every model so all of the data was used in training and testing. Additionally, I removed the 'host-since' and 'amenities' columns from the dataset, as they presented various issues during the data cleaning process. Dummy variables were created for columns with categorical data (ex. 'neighborhood-cleansed'), and at the end of the project I performed a feature importance analysis using the random forest model.

4 Exploratory Data Analysis

The pre-processing for this project consisted of removing columns with no apparent value (ex.host-id, host-name, etc.), removing rows with null values, creating dummy variables for categorical data, and casting the 'price' column to a float for model interpretation. I created a boxplot for 'price' to visualize the distribution of the target data, as well as a histogram for 'bedrooms' to visualize the distribution of the predictor I use in the simple linear regression model.

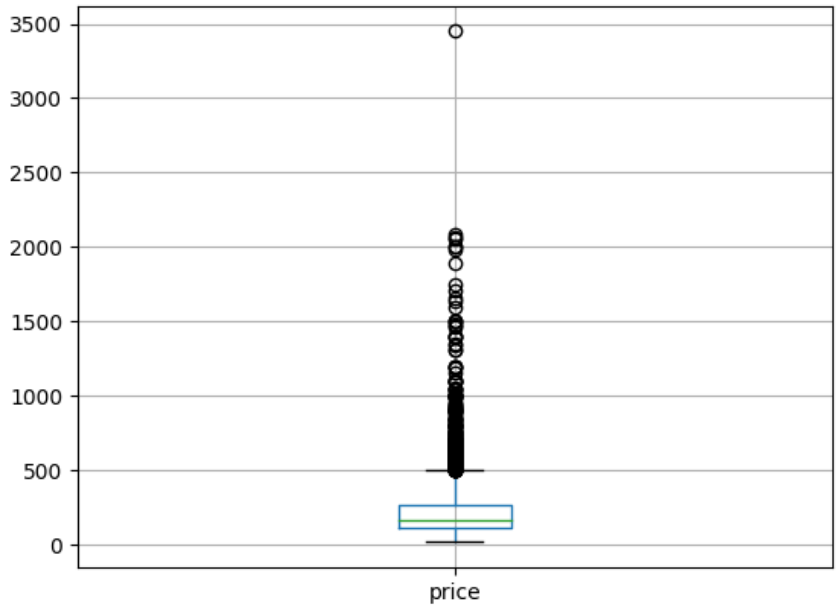

Figure 1: Boxplot of the 'price' target variable

The majority of the listing prices were between 0$ and 500$, with multiple outliers reaching above the 1,000$ range.

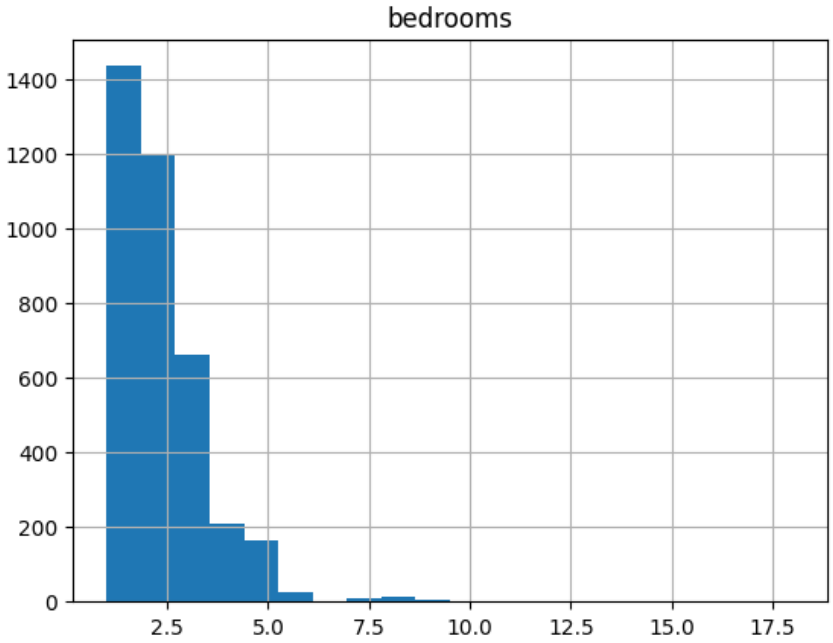

Figure 2: Histogram of the 'bedroom' column

The distribution of bedrooms per listing is reasonably skewed to the right, with the vast majority of listing containing between 1 and 5 bedrooms.

5 Methods

This project compares the performance of a total of 5 models. The first model, simple linear regression, compares the relationship between only 2 variables (number of bedrooms and price in this case). The underline mechanism behind simple linear regression involves using the least squares criterion (the process of minimizing the sum of squared error) to find the optimal coefficients for the slope(B1) and intercept(B0) in the equation 'y=B0+B1X1'. The second model, multiple linear regression, takes this concept a step further by using the same methodology to find the optimal coefficients for the equation 'y = B0+B1X1+B2X2+...+BnXn'. The third and fourth models, Ridge and Lasso regression respectively, are one step up in complexity from multiple linear regression. Ridge and Lasso regression both operate under the same underline mechanism as simple and multiple linear regression, but with the addition of a penalty term. This penalty term, usually referred to as 'alpha', reduces the slope of the model, which in turn reduces the model's sensitivity to variation and helps avoid over-fitting. In Ridge regression, alpha is calculated by squaring the slope of the model, where it is then added to the sum of squared residuals. In Lasso regression, alpha is calculated by taking the absolute value of the slope where it is then added to the sum of squared residuals. The fifth and final model implemented in this project is the random forest regression model. The three most important concepts to understand the mechanism behind this model are Gini impurity, decision trees, and ensembling. Put simply, Gini impurity is a mea-

surement of the likelihood of incorrectly classifying data, if that data was randomly classified according to the distribution of class labels from the data set. Higher Gini impurity means a higher chance of incorrectly classifying the data. Decision trees are essentially a series of conditions that attempt to minimize the Gini impurity of a dataset by splitting the data at particular points. Each split or 'branch' in the tree reflects a specific condition upon which the data is separated. In a well designed decision tree, following a path from the 'root' of the tree to a leaf node should cleanly separate the data into distinct categories where it can be assigned a class (in the case of classification problems) or a continuous value (in the case of regression problems). The final important concept necessary for understanding random forest regression is the ensemble technique. Ensembling is simply the process of averaging together the results of multiple different models to reduce the variance of a model's predictions. It is important to note that models used in ensembling must have different splitting criteria or they will not perform their function in reducing variance due to the models having the same or similar predictions. Random forest regression is an ensemble of decision trees (the reason its called a 'forest') with the added condition that the splits in each tree are performed at random to reduce the bias that may form when operating solely on the principle of minimizing Gini impurity.

6 Result Discussion

The primary metric used to evaluate the models' performance in this project is root mean square error (RMSE). This metric is calculated by averaging the squared error (actual value - predicted value) of a given model and then taking the square root. This metric is particularly useful when interpreting a model's performance because the units are the same as the target variable. This metric was further averaged using 10 fold cross-validation to get a more reliable RMSE value. The RMSE values for each model are reported below.

Model	RMSE
Simple Linear Regression	148.21
Multiple Linear Regression	137.48
Ridge Regression	136.06
Lasso Regression	138.25
Random Forest Regression	137.34

The best performing model was Ridge regression, followed closely by random forest regression, multiple linear regression, Lasso Regression, and finally simple linear regression. The results of the RMSE analysis defied my expectations as I hypothesized that random forest would exhibit by far the best performance, followed by Ridge, Lasso, and multiple linear regression, and after a sharp drop-off, simple linear regression. However, the error values of the all models were relatively close to one another and much higher than I anticipated. After considering several factors, I believe my models have under-fit the data. Removing columns such as 'amenities' and 'bathrooms' from the original dataset due to the pre-processing issues may have seriously impacted the regression analysis. Furthermore, I neglected to scale the

continuous data columns which may also have had a serious impact. Looking at Figure 3 and 4, it can be surmised that penalizing the model had a dramatic impact on the performance of the regression analysis. While some amount of penalizing was beneficial for Ridge regression, significant penalizing of the model lead to an increase in error.

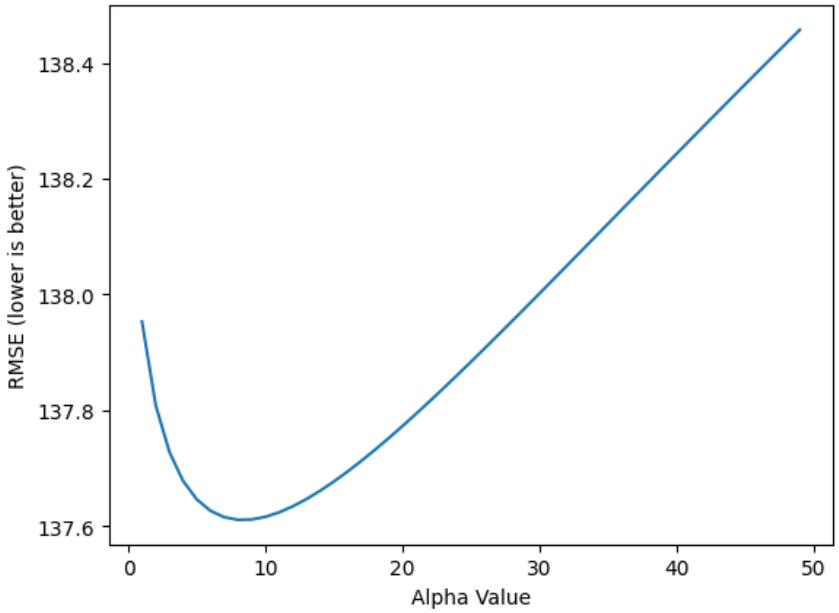

Figure 3: Plot of different alpha values in Ridge regression

In the case of Lasso regression, any penalty whatsoever lead to a sharp increase in error. This signals that there is a significant amount of variation that is not explained by the model.

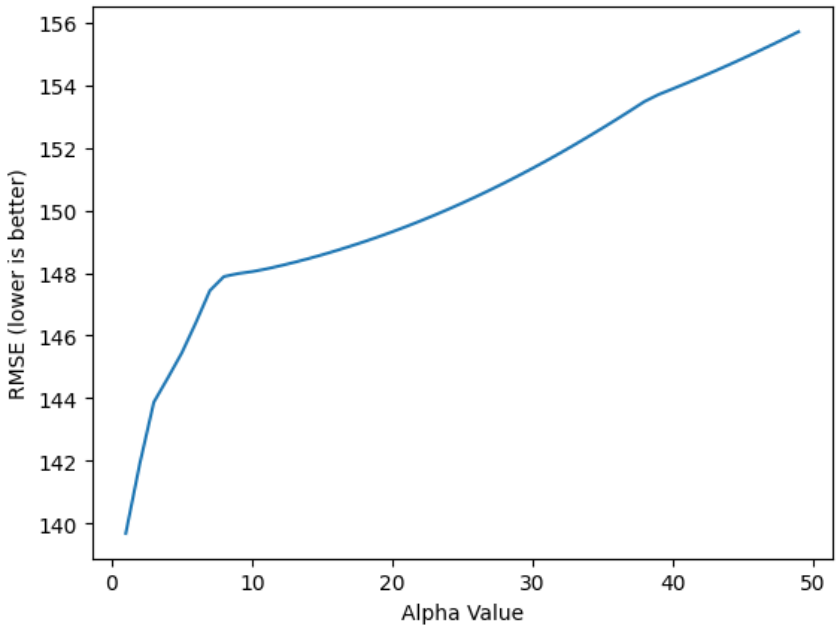

Figure 4: Plot of different alpha values in Lasso regression

The last signal that something may be amiss with the manner in which I supplied the data to the models is that the RMSE for the random forest model converged after around 20 estimators even though the error was still extremely high. This means that adding computational resources to the regression analysis did not significantly aid in reducing the error of the mode, which leads me to believe the model is operating from flawed data.

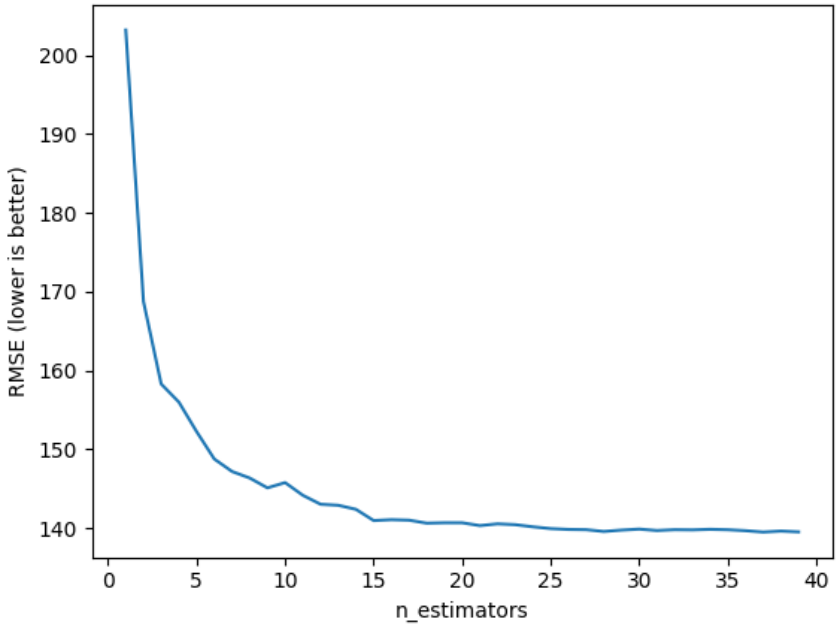

Figure 5: Plot of different numbers of estimator/tree values in random forest regression

The table below demonstrates the results of a feature importance analysis based on the random forest model. The most important feature by far, according to the random forest model, is the number of bedrooms. This singular variable accounts for nearly half of the variable importance in the random forest model and is closely correlated with the next two most important variables; the number of people a given listing accommodates, and the number of beds in a given listing. All other variables in the model had less than 5% importance and were excluded from the table as such.

Features	**Importance**
accommodates	0.08
bedrooms	0.44
beds	0.06
reviews_per_month	0.05

7 Conclusion/Future Work

In summary, regardless of this project's failure to fully meet my expectations for performance, it is clear that more complex forms of linear regression and tree ensemble techniques are better suited to complex regression analysis problems than simple linear regression. This is primarily due to their ability to factor in multiple variables, reduce variance, and minimize bias. Even working with flawed data, the 3 versions of multiple linear regression models and random forest model were able to outperform the simple linear model. To improve the performance of the models in this project it is vital to revisit the data pre-processing stage and correct several mistakes I made that likely had a major impact. For future work, I would like to explore the more niche applications of each of these models to see where they can be most effectively employed and to explore the differences between them in a more defined manner. More specifically, I would be interested in exploring the application of regression techniques to other areas of economic importance to provide more transparency to consumers and empower them to make the most financially informed decisions possible.

References

[1] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. Airbnb price prediction using machine learning and sentiment analysis. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 173–184. Springer, 2021.

[2] Yuanhang Luo, Xuanyu Zhou, and Yulian Zhou. Predicting airbnb listing price across different cities, 2019.

[3] Zhihua Zhang, Rachel JC Chen, Lee D Han, and Lu Yang. Key factors affecting the price of airbnb listings: A geographically weighted approach. *Sustainability*, 9(9):1635, 2017.

A Appendix