Predicting Airbnb Prices in the US Most Visited Cities: Evidence from Machine Learning Techniques

Imisi Aiyetan, Magdana Kondaridze, Katelin Swanson

May 8, 2021

1 Introduction

Airbnb is a globally popular vacation rental website. Beginning in 2007, the company boasts of four million hosts and over 800 million guests. While large and bustling, the entire concept of their vacation rental properties is built on accessibility. Individuals can sign up a room, shared space, or entire home entirely online. Hosts are responsible for listing the available amenities their property possess. This substantiates the important question, how much do amenities listed influence an Airbnb rental price?

The objective of this project is to determine the value of the amenities included in a given Airbnb listing. In response to the COVID-19 pandemic and subsequently imposed restrictions regarding vacation rental use, it is paramount for Airbnb hosts to understand the impacts of their listing descriptors. This project proceeds utilizing machine learning methods and nontraditional text data to mine the importance of listed amenities. Unlike the existing studies that investigated airbnb price prediction by considering a single state or city, this current study contributes to these existing studies by examining most visited cities by travelers in eight states in the US, including Washington DC.

2 Methodology[1]

The techniques used to surmise the importance of amenity listings include a linear regression model, shrinkage estimation methods, and regression trees. We also include a brief sentiment analysis regarding the subjective nature of descriptors used.

2.1 Linear Regression

Linear regression assumes homoscedasticity, exogeneity of explanatory variables, independently and identically distributed error terms, and a matrix of explanatory variables that is of full rank.

$$min(Y - X\beta)'(Y - X\beta)$$

The optimization provides the solution:

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

The assumptions underlying linear regression are frequently violated in empirical practice. Throughout this project, we expect that they will be violated as well. Therefore, we use additional methods to conduct our analysis.

[1] Recommended book, Hansen (2021), was used for this section

2.2 Shrinkage Estimation Methods

2.2.1 Ridge Regression (RR)

To control variance and regularize he coefficients, RR is used as follows.

$$min(Y - X\beta)'(Y - X\beta)$$

such that

$$\sum_{j=1}^{k} \beta_i^2 \leq t$$

The solution is:

$$\widehat{\beta}_{RR} = (X'X + \lambda I_k)^{-1} X'Y$$

with the shrinkage parameter, making the solution non-singular even in the case of an non-invertible X'X.

$$\lambda > 0$$

This parameter is tuned to control the size of the coefficients and the amount of regularization. Not, as the shrinkage parameter approaches 0, our solution becomes the linear regression solution. We use K-fold cross-validation to select an optimal shrinkage parameter.

2.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Similar to RR, the LASSO relies on a shrinkage parameter to control for the amount of regularization in regression. LASSO penalizes coefficients, only giving predictive power to those substantially different from 0.

$$min(Y - X\beta)'(Y - X\beta)$$

such that

$$\sum_{j=1}^{k} \beta_i \leq t$$

2.2.3 Elastic Net

Elastic Net is a combination of both RR and LASSO. Utilizing weighted average of the penalties developed for RR and LASSO, the Elastic Net provides a third shrinkage estimation method for use in this project.

2.3 Regression Trees

This method of nonparametric regression relies on step-functions as a sufficient approximation for any functional form. Specifically, we utilize a Random Forest for parameter selection.

2.3.1 Random Forest

Following the steps as detailed by Friedman, Hastie, and Tibshirani (2008), this process reduces the correlation across bootstrapped regression trees. Subsequently, this procedure also reduces the variance across bootstrapped average, which is a desirable feature of this methodology.

3 Data Description

The dataset used for this study is publicly available on Kaggle (collected from InsideAirbnb.com, Karuna (2020)), which contains data on recent Airbnb property listings in eight major states of the US: California, Florida, Hawaii, Illinois, Nevada, New York, Tennessee, Washington, and also Washington DC. The dataset has various features about the hosts, locations, etc., and includes 158,202 entries in the raw dataset.

For the initial data reshaping, we removed irrelevant and uninformative variables, such as id, host_id, host_since, host_is_superhost, last_review. We replaced missing values for beds, bedrooms, bathrooms, and reviews_per_month by zeros to avoid dropping some samples. In the raw data, the maximum price for a room is \$25,000 per night, which is suspiciously high (or maybe it is too luxurious). Therefore, to avoid any sampling error, we removed outliers (price $>$ \$1000) in the dataset. We also removed price = \$0, which in our opinion is unrealistic. After that we are left with 108,675 entries. In Table 1, we show summary statistics of some of the important variables such as price, bathrooms, bedrooms, beds, and number of reviews after the cleaning and reshaping of the data.

Figure 1 displays the number of Airbnb lists in each state, and it is easy to see that majority of Airbnbs in the dataset are from California. In Figure 2, we see that price distribution is right-skewed, and on average, Airbnb price is \$163 per night. Figure 3 displays price distribution by the states/territories (Panel A) and room type (Panel B). The average price is the highest in Hawaii (\$232.4 per night), followed by Tennessee (\$188.5 per night), and the lowest in Illinois (\$127.6 per night). In terms of room types, the average price per night is highest for a hotel room (\$232.4 per night) and the lowest for a shared room (\$58.6 per night). The most expensive neighborhood groups are Maui (\$276.6 per night), Kauai (\$276.4 per night), and Santa Cruz (\$263.5 per night), and the cheapest - Queens (\$93 per night) and Bronx (\$89 per night).

The nontraditional variables, the name of the listing and amenities, are analyzed using a natural language process. We started cleaning the text from all the unnecessary characteristics, such as numbers, punctuation, and we replaced upper case letters with lower. Figure 4 displays Word Clouds for listing name (Panel A) and amenities listed (Panel B). We see that the most used words in listing names are private studio rooms closer to the ocean/beach, which is not surprising, as most of the listings are from coastal states. In the amenities listed - all the safety detectors are the most used words. This makes sense, as, by law, all the units are required to have safety detectors installed. [2]

For the sentiment analysis, we generated new, polarity and subjectivity, variables for both listing names and amenities listed. Polarity, in sentiment analysis, identifies sentiment orientation, which can be positive, negative, or neutral. Subjectivity identifies personal feelings, views, or beliefs. Figure 5 shows sentiment analysis for Listing name (Panel A) and Amenities listed (Panel B) by states. On both figures, we see that most of the states are in the upper-right of the quadrant, which means that both listing names and amenities are mostly positive or neutral opinions.

4 Results and Discussions[3]

This study used machine learning techniques to predict the Airbnb prices based on the amenities included in a given Airbnb listing. Table 1 shows the performance metrics of the techniques, namely, Ridge Regression, Lasso Regression, Elastic Net Regression, and Random Forest. In terms of model performance, both training and validation (testing) splits were used to select the best model. Likewise, Mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), R^2 score, and

[2] At this stage, we did not use stop words. However, we plan to further clean the data in the next stages

[3] In the analysis, GitHub.com (2020) was used as a reference for coding

accuracy were used to evaluate the models. The results show that R^2 scores for the five techniques are low and relative similar under testing split. However, while linear, Ridge, Lasso and Elastic Net Regressions R^2 scores are low and relative similar, Random Forest produced a very strong R^2 scores under training split. This indicates that Random Forest feature importance analysis had made the most impact on improving the performance of the model. In terms of MAE, RMSE, and MAPE, Random Forest turned out to be the best performing model for both testing and training splits, with the lowest MAE, RMSE and MAPE. In this case, Random Forest predicts Airbnb prices relatively well over the training and testing sample.

Looking at the feature importance results produced by Random Forest, the results revealed that listings with more bedrooms, more bathrooms, and that can accommodate more persons tend to predict Airbnb prices compared to listings that are available for 356 days with good reviews, among other characteristics. This finding corroborates the study by Lawani, Reed, Mark, and Zheng (2019) that found similar results using an Econometric approach. Also, our findings are consistent with the previous studies on the hospitality industry (Costa (2013); de Oliveira Santos (2016); Espinet, Saez, Coenders, and Fluvià (2003); among others). Finally, figure 8 shows the results of Lasso and Elastic net Regressions. By penalizing the model for number predictors included in the model, the results show that more bedrooms, more bathrooms, and accommodation are determinants of Airbnb prices.

5 Conclusion

This study investigated the determinants of Airbnb prices based on amenities included in a given Airbnb listing. The machine learning techniques considered are Ridge Regression, Lasso Regression, Elastic Net Regression, and Random Forest. The results show that Random Forest turned out to be the best performing model for predicting Airbnb prices. In terms of the determinants, our results show that listings with more bedrooms, more bathrooms, and that can accommodate more persons are the main determinants of Airbnb prices.

Appendix

Tables

	price	**bathrooms**	**bedrooms**	**beds**	**num_of_rev**
mean	162.7073	1.3923	1.4211	2.0169	110.6969
std	142.6598	0.7544	1.0228	1.6439	274.7679
min	10.0000	0.0000	0.0000	0.0000	0.0000
25%	75.0000	1.0000	1.0000	1.0000	1.0000
50%	120.0000	1.0000	1.0000	2.0000	12.0000
75%	199.0000	2.0000	2.0000	3.0000	62.0000
max	999.0000	50.0000	27.0000	51.0000	1825.0000

Table 1: Data Description

	MAE	**RMSE**	R^2	**MAPE**	**Accuracy**
Linear	65.768	104.515	0.446	51.769	48.23%
Ridge	65.767	104.516	0.446	51.769	48.23%
Lasso	65.649	104.624	0.445	54.290	45.71%
Elastic Net	65.691	104.661	0.445	58.996	41.00%
Random Forest	57.319	95.683	0.536	42.939	57.06%

Table 2: Test Split

	MAE	**RMSE**	R^2	**MAPE**	**Accuracy**
Linear	66.444	106.478	0.450	51.610	48.39%
Ridge	66.444	106.478	0.450	51.610	48.39%
Lasso	66.345	106.589	0.449	54.213	45.79%
Elastic Net	66.375	106.608	0.449	58.945	41.05%
Random Forest	22.678	39.629	0.924	16.660	83.34%

Table 3: Train Split

Graphs

Figure 1: Number of Airbnb Lists in each State/Territory

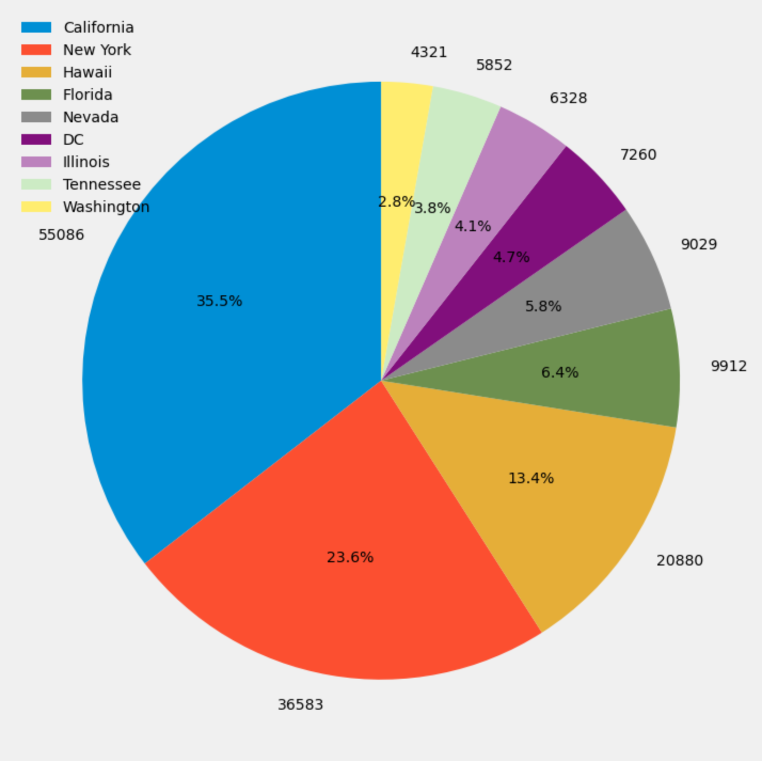
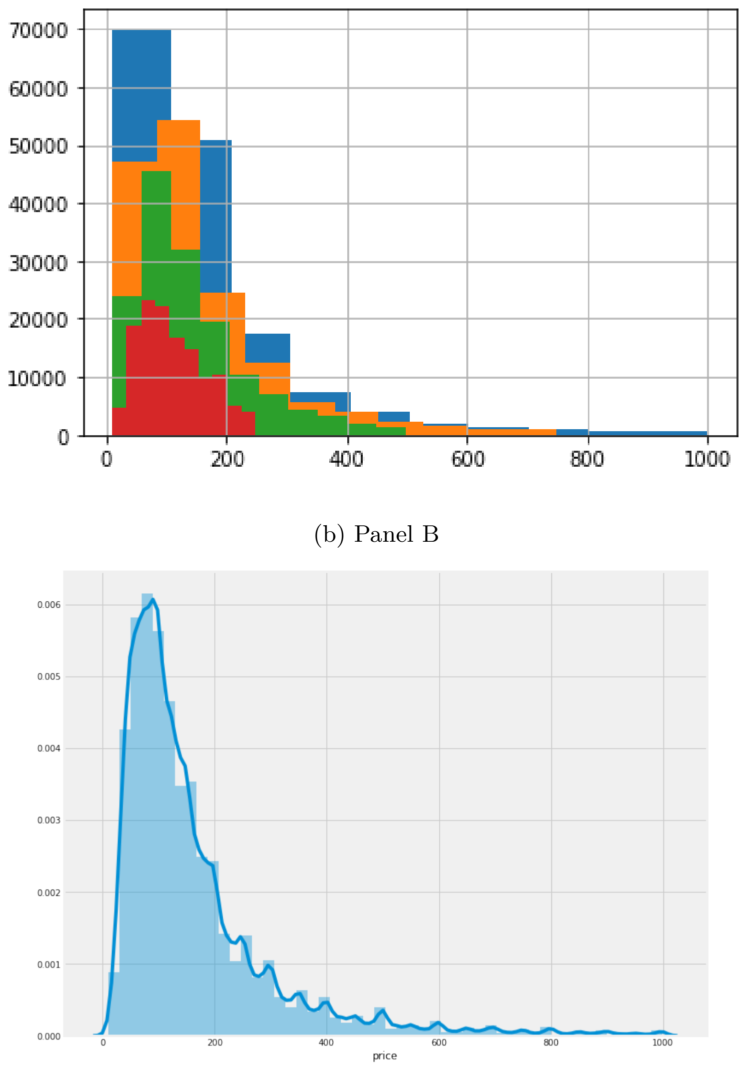


Figure 2: Price Distribution

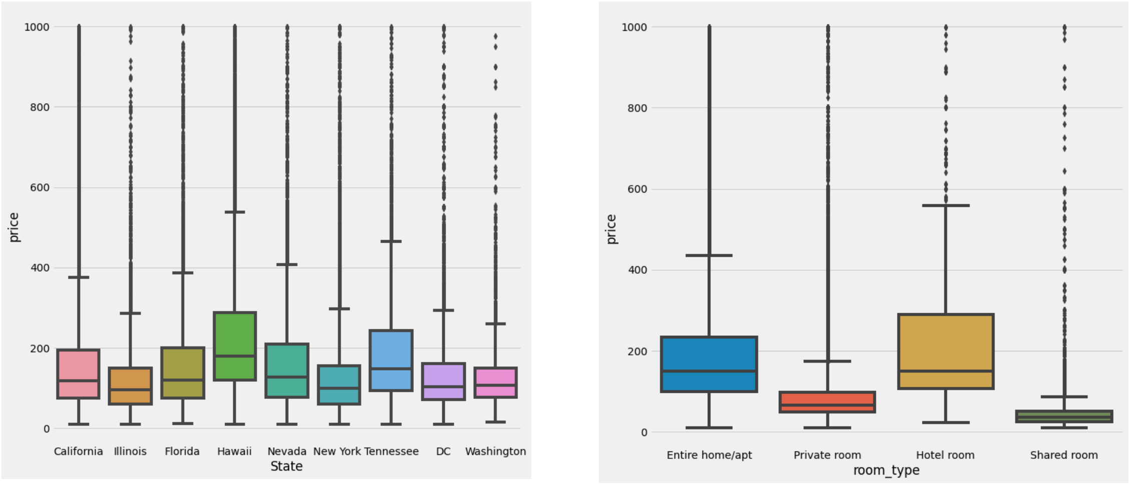
(a) Panel A



(b) Panel B

Note: You can see price distribution for price <250, <500, <750, and <1000, in red, green, orange and blue colors, respectively

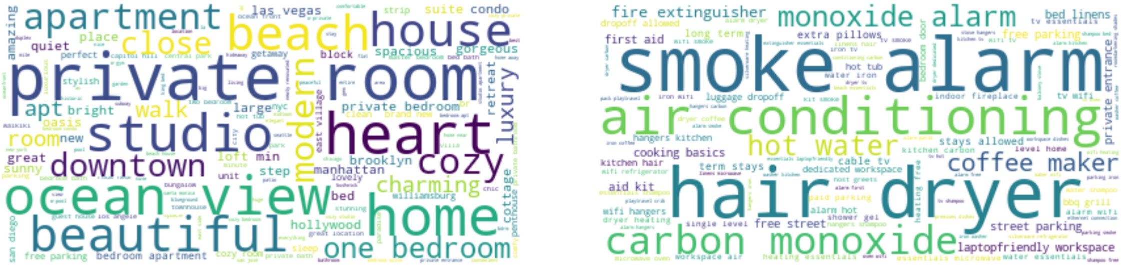
Figure 3: Price Distribution by State (Panel A) and by Room Type (Panel B)



(a) Panel A

(b) Panel B

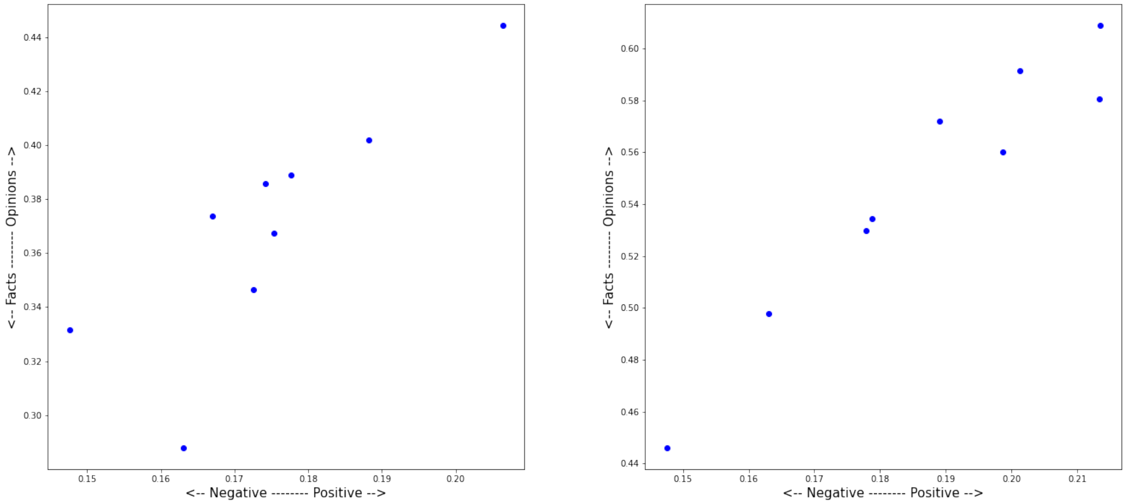
Figure 4: WordClouds for Listing name (Panel A) and Amenities (Panel B)



(a) Panel A

(b) Panel B

Figure 5: Sentiment Analysis



(a) Panel A

(b) Panel B

Figure 6: Goodness of Fit of price

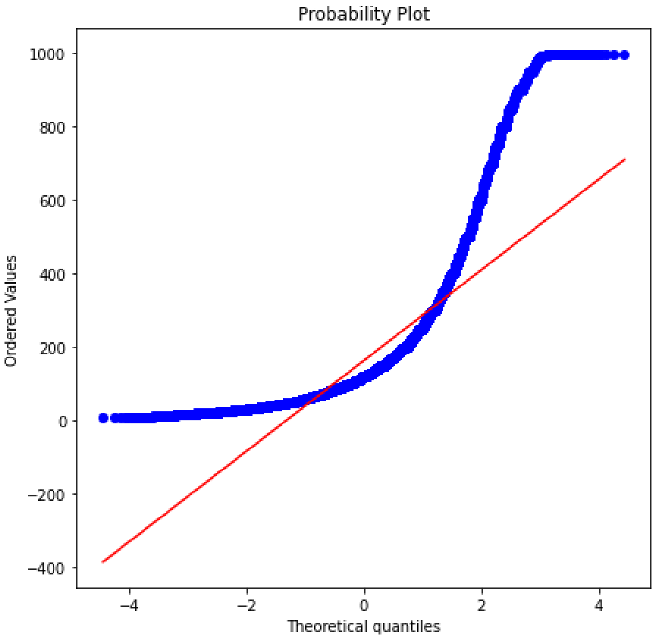


Figure 7: Linear Model Prediction

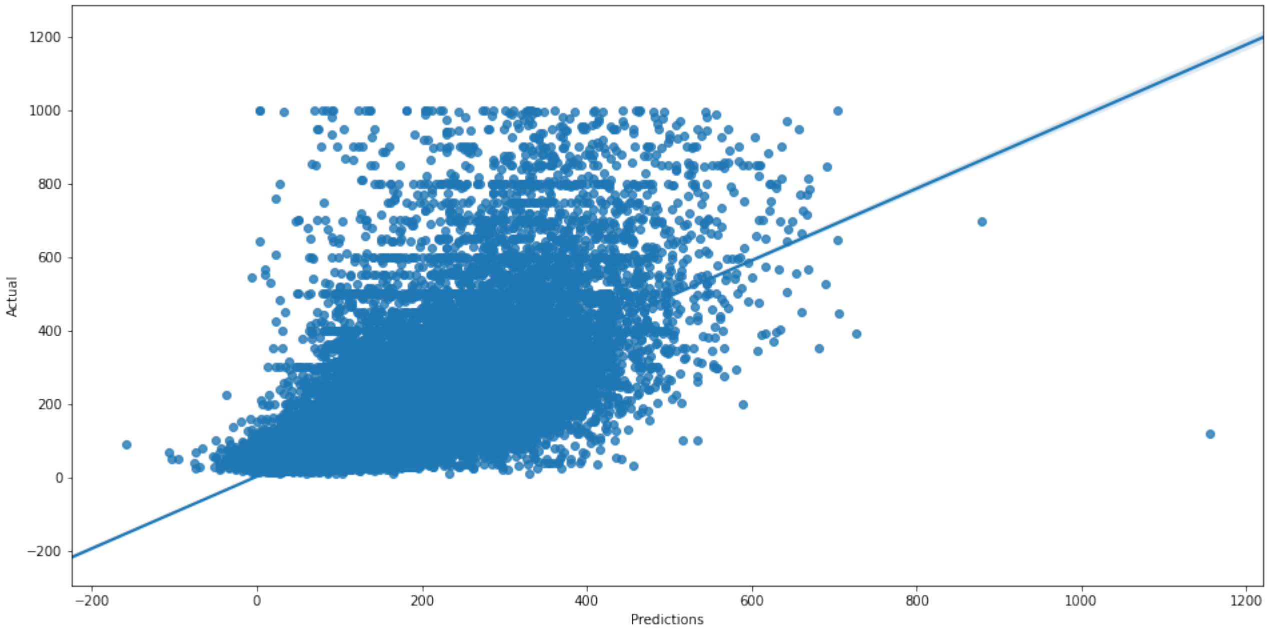
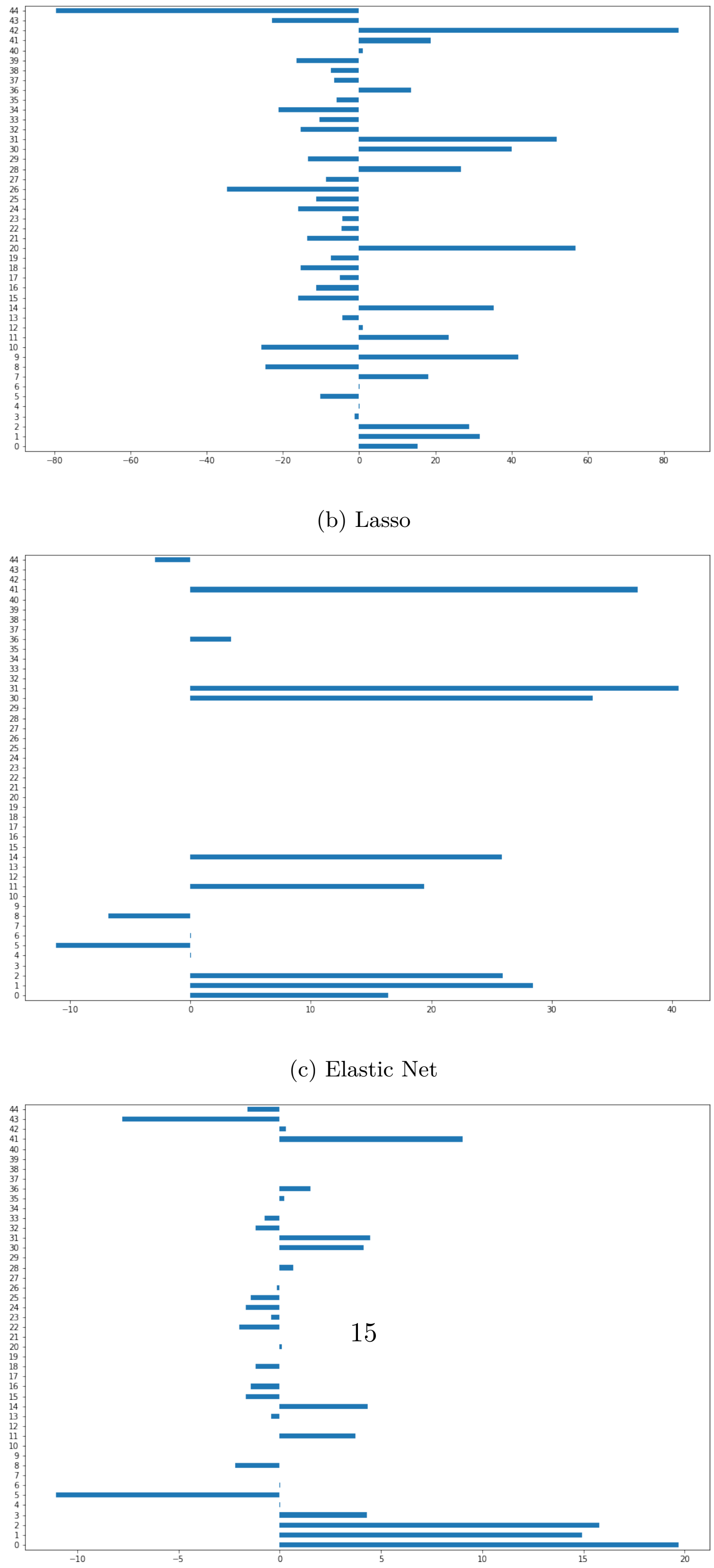


Figure 8: Important Coefficients for Ridge Regression, Lasso Regression, and Elastic Net

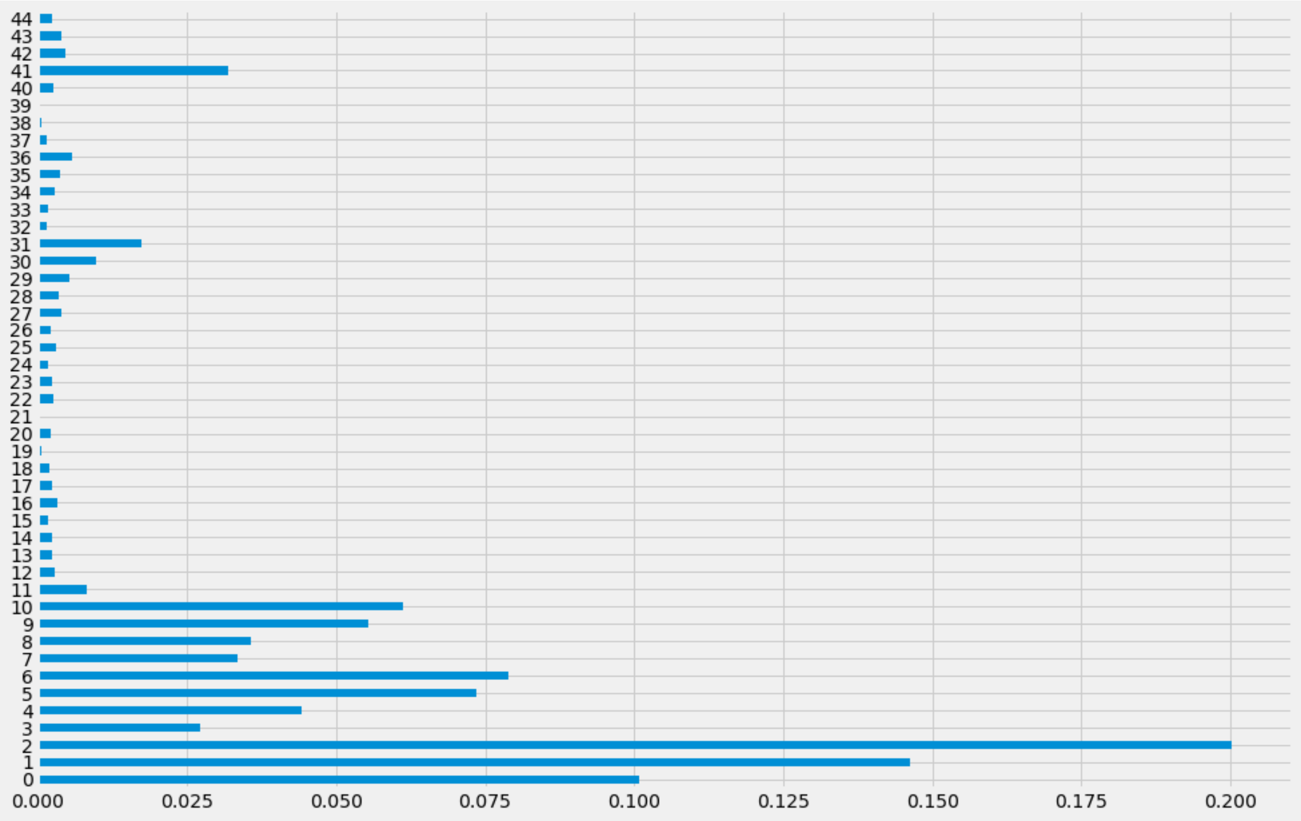
(a) Ridge



(b) Lasso

(c) Elastic Net

Figure 9: Important Coefficients for Random Forest



References

Costa, J. C. C. (2013). Price formation and market segmentation in seaside accommodations. *International Journal of Hospitality Management*, *33*, 446–455.

de Oliveira Santos, G. E. (2016). Worldwide hedonic prices of subjective characteristics of hostels. *Tourism Management*, *52*, 451–454.

Espinet, J. M., Saez, M., Coenders, G., & Fluvià, M. (2003). Effect on prices of the attributes of holiday hotels: a hedonic prices approach. *Tourism Economics*, *9*(2), 165–177.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

GitHub.com. (2020). *Machine learning.* `https://github.com/`.

Hansen, B. E. (2021). Econometrics.

Karuna, K. (2020). *Airbnb us dataset.* `https://www.kaggle.com/kavithakaruna/airbnb-us-dataset`. (Accessed: 02.20.2021)

Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of airbnb reviews in boston. *Regional Science and Urban Economics*, *75*, 22–34.