Matthew Brigham
STA 685
Airbnb Price Prediction – Random Forest

The dataset is a collection of Airbnb listings from Austin, Texas. Each listing represents a single listing on the Airbnb platform and has many descriptive variables. Table 1 below shows a summary of the variables for each listing. There are qualitative and quantitative variables, as well as text data. In total, for the Austin dataset, there are 74 variables and 10809 observations. The dataset was compiled on September 15, 2021, via web scraping by an independent firm that is not affiliated with Airbnb and is not their competition.

Name	Text string name of the property given by the owners.
Description	Text data string describing the property created by the owners.
Host About	Text data string describing the owners, written by the owners.
Host Acceptance Rate	This is a percentage value reflecting the percent of applications to stay that are "accepted" by the host.
Host is Superhost	A binary variable that is True if the host is a superhost, else False.
Latitude	A latitude description of the location of the Airbnb (quantitative).
Longitude	A longitudinal description of the location of the Airbnb (quant.)
Property type	A text description of the property, e.g., "Entire residential home", "Private room in the residential home", "Entire guesthouse", etc.
Room Type	A text description of the type of room on the property, e.g., "Private room" or "Entire home/apt". There are only two types.
Accommodates	The number of visitors that the property can accommodate.
Bathrooms	The number of bathrooms.
Bedrooms	The number of bedrooms.
Amenities	A list of strings describing the amenities the property has, e.g., "Free street parking", "Air conditioning", "Patio or balcony", "Wifi", etc.
Price	Price of the unit on the date the data was scraped.

Table 1: A sample of the 74 variables in the data set with descriptions.

Dynamic pricing is the concept of variable pricing based on demand, or other factors. This concept is commonly employed by Uber, Airbnb, Amazon, airlines, and other e-commerce sites. Dynamic pricing improves profitability by manipulating prices based on a variety of market conditions, such as perceived demand, competition prices, and seasonality. The alternative to dynamic pricing is a constant price structure. One complaint of a constant price structure is that if the price is set too high, then few will buy the product. If the price is too low, then the company is missing out on profits. In either case, the company is not achieving optimal profitability/revenue. Having an automated dynamic pricing model would be beneficial to most companies based on these reasons.

The goal of this project is to accurately predict the price of the Airbnb listing. The prices in Airbnb are chosen at the discretion of the hosts, but they have the option to use Airbnb's Smart

Price Tool that considers over 70 different factors, and is dynamic over time. This project proposes that a random forest model be used on the Airbnb dataset to predict the price of the Airbnb. All quantitative variables, qualitative variables, and text data will be considered. There are several NLP text mining techniques that may be useful, such as TF-IDF. Other machine learning models, such as linear regression, and decision trees, will be explored and compared. If time allows, different dynamic pricing models may be explored to show how the random forest model can be applied in context. Dynamic pricing models are designed for the needs of the company. One example of a hypothetical pricing model for Airbnb is:

$$price(DBB) = \begin{cases} High & 0 < DBB < 10 \\ Med & 10 \leq DBB < 20 \\ Low & otherwise \end{cases}$$

Where DBB means the number of days before booking. This is interpreted as the closer the date is to the event, the higher the price is.

The paper titled "The Effect of Splitting on Random Forests" by Hermant Ishwaran (2014) may be useful in this application of the random forest. The paper discusses an improvement to pure random splitting in the random forest algorithm. A method of splitting is proposed that uses weighted splitting and is computationally efficient. The method is called ECP, or end-cut preference and is not new, however, a new argument is made in favor of its use. This method will be useful for noisy variables by making splits near the edge of noisy variables. According to the author, data with a large number of predictors will often have noisy candidate variables, making ECP splits useful. I suspect that the data may be noisy because there will be features that the dataset cannot grasp that may be guiding Airbnb visitors' decisions, such as visual cleanliness or design of the space that are difficult to quantify.

Random forests have been shown to work well for pricing models. Branda et al (2020) built a dynamic pricing model that utilized a random forest for bus ticket sales. Other individuals were able to build a random forest model for this Airbnb data set with good results (Luo et al, 2019). One problem to explore is how the price of the listing in the data set was determined. If the price may already have dynamic pricing effects incorporated, there may be a source of error when training a predictive model for the price.

References

Ishwaran, H. (2014). The effect of splitting on random forests. *Machine Learning*, *99*(1), 75–118. https://doi.org/10.1007/s10994-014-5451-2

Cox, M., & Morris, J. (n.d.). Retrieved October 19, 2021, from http://insideairbnb.com/get-the-data.html.

Branda, F., Marozzo, F., & Talia, D. (2020). Ticket sales prediction and dynamic pricing strategies in public transport. *Big Data and Cognitive Computing*, *4*(4), 36. https://doi.org/10.3390/bdcc4040036

Luo , Y., Zhou, X., & Zhou, Y. (2019). Predicting Airbnb Listing Price Across Different Cities .