Airbnb Pricing Project

Divya Sharma

Data Analysis Assignment 01

Divya Sharma (ds655)

Airbnb pricing in Asheville, NC

Section 1: Executive Report

Airbnb is a popular online platform that allows people to rent out their homes to travelers. Airbnb hosts set their own prices, but they can benefit from having accurate information about how much to charge. This project aims to develop a linear machine learning model to predict the price of Airbnb listings in Asheville, NC.

The model will be trained on a dataset of Airbnb listings that includes information such as the distance to downtown, the number of bedrooms and bathrooms, the amenities offered, and the reviews received. The model will then be used to predict the price of new Airbnb listings based on the same features.

This project has the potential to benefit both Airbnb hosts and guests. Airbnb hosts can use the model to set competitive prices that are likely to attract guests. Airbnb guests can use the model to find the best deals on rentals.

1.1 Benefits of this Project

The Airbnb price prediction project has a number of potential benefits, including:

- Helping Airbnb hosts to set competitive prices and improve their listings. By understanding how much guests are willing to pay for different types of listings and amenities, Airbnb hosts can set their prices accordingly. This can help them to attract more guests and increase their earnings.

- Helping Airbnb guests to find the best deals on rentals. By using the model to predict the price of different Airbnb listings, guests can find the best deals on rentals that meet their needs. This can help them to save money on their vacations.
- Conducting research on the impact of Airbnb on the local economy and community. The data collected for this project can be used to conduct research on the impact of Airbnb on the local economy and community. For example, researchers could look at how Airbnb rentals are distributed across different neighborhoods and how Airbnb prices compare to hotel prices.

1.2 The Model

The model that was used to analyze the Airbnb dataset is a linear regression model. Linear regression is a simple but powerful machine learning algorithm that can be used to predict the value of a variable (the price) based on the values of other variables (the information we have on the listings).

Linear Regression is a simple predictive model that can easily be interpreted

The variables that are included in the model are:

- `Room Details`: The room type (private room, shared room, entire home) and number of beds/bathrooms/bedrooms
- `Location`: Geographic information such as the distance to downtown Asheville, and the locality
- `Avalibility`: The availability of the listing
- `Reviews`: Information based on reviews provided to the listing
- `Host Details`: Host verification and contact availability details
- `Amenities`: Whether the listing has basic amenities such as AC, Parking, Wifi, Microwave, allows pets or not etc.

For a more detailed description of each variable that has been fed into the model, please refer to the technical report in section 2.

1.3 How the Model Accomplishes the Goal

The linear regression model works by fitting a line to the data. The slope of the line represents the relationship between the target variable (price) and each of the predictor variables. For example, the slope of the line that represents the relationship between price and distance to downtown would be positive, because listings that are closer to downtown are generally more expensive.

Once the model has been trained, it can be used to predict the price of new Airbnb listings by passing the values of the predictor variables to the model. For example, to predict the price

of a listing that is 1 mile from downtown, has 2 bedrooms, 1 bathroom, and offers wifi and parking, the model would use the following equation:

```
price = predicted_price =
  intercept
  + (importance of Room Details * Room Details)
  + (importance of Location * Location)
  + (importance of Avalibility * Avalibility)
  + (importance of Reviews * Reviews)
  + (importance of Host Details * Host Details)
  + (importance of Amenities * Amenities)
```

The values of the intercept and the slopes would be determined by the model during the training process.

It is important to note that linear regression is a simple model and may not be able to accurately predict the price of all Airbnb listings. However, it is a good starting point for developing a predictive model. The accuracy of the model can be improved by using more features and by using more complex machine learning algorithms. The model will learn the importance of each factor by training on a dataset of Airbnb listings. Once the model is trained, it can be used to predict the price of new Airbnb listings by passing the values of the factors to the model.

Example:

Imagine that you are an Airbnb host and you want to set a price for your listing. You could use the linear regression model to predict how much guests are willing to pay for your listing based on its distance to downtown, the number of bedrooms and bathrooms, the amenities offered, and the reviews received.

This information could help you to set a competitive price for your listing and maximize your earnings.

1.4 Metrics

To justify my model, I would use the following model metrics:

- **R-squared:** R-squared can be interpreted as the percentage of variation in the target variable (price) that is explained by the predictor variables (room type, bedrooms, dist_to_dt, etc.). R-squared is a measure of how well the model fits the data, and ranges from 0 to 1, with a higher value indicating a better fit. In this case, the R-squared value is 0.5865, which means that the model explains ~60% of the variation in the price of Airbnb listings.

- **F-statistic:** The F-statistic is a test of whether the model is a significant improvement over a simpler model, such as a model with no predictor variables. In this case, the F-statistic is 93.31 and the p-value is less than 2.2e-16, which means that the model is a significant improvement over a simpler model.

Assuming that the model has been trained on a dataset of Airbnb listings in Asheville, NC, we can predict that the price of a listing with 2 bedrooms, 2 bathrooms, and a distance to downtown of 1 mile would be around $200 per night. It is important to note that this is just a prediction. The actual price of the listing may vary depending on other factors, such as the time of year, the amenities offered, and the reviews received.

Section 2: Technical Report

Section 2.1 Model Selection

Section 2.2 Data Cleaning and EDA

The data used for this analysis is from Inside Airbnb, specifically, from here. The data contains basic details about Airbnbs listed in Asheville, North Carolina. The data dictionary for this dataset can be found here

There are 3,239 listings of Airbnbs which consist of mostly entire homes (87%), some Private Rooms (12%) and very few hotel rooms and shared rooms(~1%).

- **2.1.1 Location - adding Distance to downtown (dist_to_dt) based on Latitude and Longitude data**
  - Using the `distm()` function in the `geosphere` library, we can calculate the distance of the latitude and longitude of the Airbnb to the corresponding latitude and longitude of Downtown, Asheville. This gives us the `dist_to_dt` column which is highly significant while calculating the price
- **2.1.2 Cleaning the Price variable**
  - The `price` variable contains the price in comma separated USD values, so the data has to be cleaned and made numeric
- **2.1.3 Cleaning the bathrooms count**
  - The bathrooms column is text and contains a mix of numbers (1.5) and text (half) values. These are converted to the corresponding numeric values (0.5)
- **2.1.4 Host Verifications**
  - The Host Verifications contains json type formatted lists of combinations of email, work email, and phone. This is split into two binary columns - host_verification_email and host_verification_phone