Airbnb Pricing Project

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
- Helping Airbnb guests to find the best deals on rentals
- Conducting research on the impact of Airbnb on the local economy and community

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)

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.

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, based on the inputs for a new airbnb, the model would use the following equation to predict the price:

```
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)
```

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 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.

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.). We have the R-Square as 0.58
- **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. We have the F-statistic as 93.31

Section 2: Technical Report

Section 2.1 Model Selection

We have chosen linear regression over other models for predicting Airbnb prices because of the following reasons:

- **Simplicity:** Linear regression is a relatively simple model to understand and implement. This makes it a good choice for cases where interpretability is important.
- **Interpretability:** The coefficients of a linear regression model can be interpreted directly as the effect of each input variable on the output variable. This can be helpful for understanding the factors that drive Airbnb prices and for making predictions.
- **Efficiency:** Linear regression models are typically very efficient to train and predict with. This makes them a good choice for applications where speed is important, such as real-time Airbnb pricing.
- **Robustness:** Linear regression models are relatively robust to outliers and noise in the data. This makes them a good choice for real-world Airbnb datasets, which may contain outliers due to factors such as seasonal demand and special events.

In addition to these general advantages, linear regression has also been shown to be effective for predicting Airbnb prices in a number of studies. For example, a study by Airbnb found that linear regression was able to predict Airbnb prices with an accuracy of over 90%.

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
- **2.1.5 True/False columns**
  - Columns like has_availability, host_identity verified, is_superhost have values 't' and 'f' for true and false, and also contain blanks. These have been converted into 1s (for true) and 0s (for false and blanks)
- **2.1.6 Amenities**
  - The Amenities column has json formatted lists of amenities. We have taken some of the most relevant Amenities such as Wifi, Parking, Air Conditioning, Kitchen, Pet friendliness, Microwave, Refrigerator, TV, and Heating, and created columns such as has_wifi, has parking etc with 1s and 0s
  - *The data showed that all the listings have TVs, so the entire column was coming as 1, so we removed that column from this analysis*

EDA

The Price ranges from $14/Night to $2,059/Night. The average price is $180/Night (indicated by the red line in the chart).

Distribution of Price

```
![Distribution_of_Price.png](https://github.com/DivyaSharma0795/IDS702_Data_Analysis_Assignme
```

Section 2.3 Inputs to the model

- `Room type`: The type of Airbnb listing, such as private room, entire home, or shared room.
- `Bedrooms`: The number of bedrooms in the Airbnb listing.
- `Dist_to_dt`: The distance to downtown Asheville.
- `Bathrooms_numeric`: The number of bathrooms in the Airbnb listing.
- `Accommodates`: The maximum number of guests that the Airbnb listing can accommodate.
- `Beds`: The number of beds in the Airbnb listing.
- `Minimum_nights`: The minimum number of nights that guests can stay in the Airbnb listing.
- `Has_availability`: Whether or not the Airbnb listing is available on the date that the prediction is being made.
- `Number_of_reviews`: The number of reviews that the Airbnb listing has received.
- `Review_scores_rating`: The average rating of the reviews that the Airbnb listing has received.
- `Reviews_per_month`: The average number of reviews that the Airbnb listing receives per month.
- `Review_scores_location`: The average rating of the Airbnb listing's location.
- `Review_scores_value`: The average rating of the Airbnb listing's value.
- `Review_scores_communication`: The average rating of the Airbnb listing's communication.
- `Review_scores_checkin`: The average rating of the Airbnb listing's checkin process.
- `Review_scores_cleanliness`: The average rating of the Airbnb listing's cleanliness.
- `Review_scores_accuracy`: The average rating of the Airbnb listing's accuracy.
- `Host_has_profile_pic`: Whether or not the Airbnb host has a profile picture.
- `Host_identity_verified`: Whether or not the Airbnb host's identity has been verified.
- `Host_verification_email`: Whether or not the Airbnb host's email address has been verified.
- `Host_verification_phone`: Whether or not the Airbnb host's phone number has been verified.
- `Host_has_profile_pic`: Whether or not the Airbnb host has a profile picture.
- `Host_identity_verified`: Whether or not the Airbnb host's identity has been verified.

- `Host_is_superhost`: Whether or not the Airbnb host is a Superhost.
- `Has_ac`: Whether or not the Airbnb listing has air conditioning.
- `Has_parking`: Whether or not the Airbnb listing has parking.
- `Has_wifi`: Whether or not the Airbnb listing has wifi.
- `Has_kitchen`: Whether or not the Airbnb listing has a kitchen.
- `Has_pets`: Whether or not the Airbnb listing allows pets.
- `Has_microwave`: Whether or not the Airbnb listing has a microwave.
- `Has_refrigerator`: Whether or not the Airbnb listing has a refrigerator.
- `Has_heating`: Whether or not the Airbnb listing has heating.

Section 2.4 Model Performance

The linear regression model has an R-squared value of 0.58, which indicates that the model explains 58% of the variability in the response variable. The Residual Standard error value of 91.03 suggests that the model's predictions are off by an average of $91.03. The F-statistic of 93.31 and the p-value of $<$ 2.2e-16 indicating that the model is statistically significant and that at least one of the predictor variables is significantly related to the response variable. The median residual of -3.72 indicates that the median difference between the predicted and actual values is $3.72, which suggests that the model is reasonably accurate. Overall, the model's performance is moderate, with room for improvement.

2.4.2 Charts

The diagnostic plots for the model are as below -

1. Residuals vs Fitted Values This chart is used to check the linearity and homoscedasticity assumptions of a linear regression model by plotting the residuals against the fitted values.

2. Q-Q Plot This chart is a graphical technique used to compare the distribution of a the predicted to the actual distribution by plotting the quantiles of the predicted data against the quantiles of the actual distribution.

3. Scale-Location Plot This plot is used to check the homoscedasticity assumption of a linear regression model by plotting the square root of the absolute standardized residuals against the fitted values.

4. Reiduals vs Leverage Plot This plot is used to check for influential observations in the linear regression model by plotting the standardized residuals against the leverage of each observation.

Section 2.5 Conclusion

Based on the linear regression model used to predict Airbnb prices, with an R-squared value of 0.58, we can conclude that the model explains 58% of the variability in the response variable. While this is not a perfect fit, it suggests that the model is moderately effective in predicting Airbnb prices. However, there may be factors like seasonality that are excluded in the current methodolgy that can help identify the prices more accurately