Airbnb Pricing in Ashville, NC

Revanth Chowdary Ganga (rg361)

1. Executive Report

1.1 Introduction

Airbnb is a platform which allows the owners("**Hosts**") of residential properties to list them online to other people (*"**Guests**"*) for temporary stays. The hosts get to decide most of the components like the rules, amenities provided and the price for the listing etc. This project aims to develop a model to assist the hosts in deciding a price for their listings in Ashville, North Carolina.

The Data used in this project is from InsideAirbnb and contains the details of some Airbnb listings in Ashville, NC. The Data is as of: ***2023-Jun-18*** which has **3,239** observations (rows) and ***75*** variables (columns)

This data is used post cleaning to analyze and determine the factors(variables) which could impact the price of a listing. The selected factors are then used to generate a model which can assist the hosts in deciding a price for their listings in Ashville.

1.2 Method

The model was developed in multiple stages as follows:

1.2.1 Data Preparation

To ensure proper analysis of data and generation of the model, the data first had to be cleaned and modified as per the requirements. some operations like deletion of non-required columns, creation of helper columns, empty value cleaning were performed to prepare the data for analysis.

1.2.2 Analysis

Once the Data has been cleaned and prepared as mentioned in the previous step, the Data was explored to see both the individual variable wise characteristics and for any patterns or relationship between variables, possible duplication(redundancy) or correlation of data between variables. post this process the relevant data was selected for using in the model

1.2.3 Modelling

For this project, **Linear Regression** was used to generate the price of a listing based on the other characteristics of the listing. Linear Regression works by trying to fit a equation which would represent the relationship between the output and input variables.In this project, the output variable is the price of the listing, The input variables were selected as per the analysis process mentioned above and are as follows(all w.r.t to the listing of interest):

- Host: Verification Status, Super-host status, email verification, response rate
- Property: Room type, No. of bedrooms, bathrooms and beds. No. of guests it can accommodate
- location: distance to downtown Ashville (calculated), neighborhood
- Other: number of reviews, average ratings, selected amenities, minimum nights

1.3 Results & Conclusion

By using the selected input variables, a Linaer Regression model was made which can be used to generate the listing price based on the values of the input variables.

The generated model has an $\mathbf{R}^2$ value of 0.64, $\mathbf{R}^2$ metric is used as a measure to see the effectiveness of the model, it ranges from 0 to 1 and the closer it is to 1, the better the model is.

Sample Output:

as per the model a listing which has a **verified superhost** with a **response rate** of 70% can list his **entire apartment** property at a distance of **1.2 mi** to downtown Ashiville in the **neighbourhood** with code "28732" which can **accomodate** 4 guests and has amenities like **TV** and **parking** with 2 each of **Bathrooms, Beds and Bedrooms** for approximtely 100$ provided the property has 50 **reviews** and average reviews in the range 3.5 to 4.5.

While the model is able to generate a price based on the inputs, it can be further improved by having additional data points such as time-related information, uniform listing of amenities etc.

2. Technical Report

2.1 Introduction

The raw Data provided had **3239** observations and **75** variables.

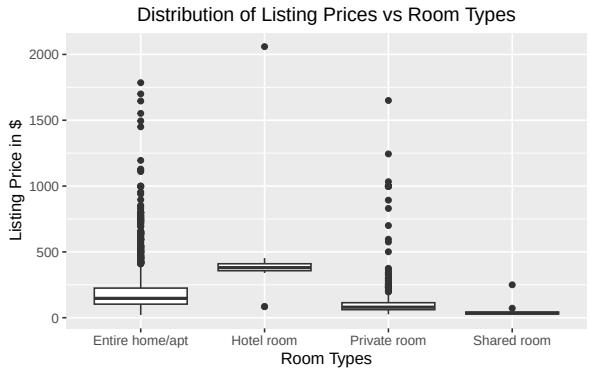
The Following Data cleaning and preparation operations were performed

- removal of columns which have identification or unique details which are not required for the model such as the columns with the listing or host IDs and URLs etc.
- extraction of numerical values from columns: The columns like Price and number of Bathrooms were converted from Text to Numeric data by extracting the numeric information from them by using functions like **grepl**, **parse_number** etc. so that they could be used more effectively in the model
- New variables were calculated to be used in the model:
  - distance to downtown: was calculated using the **distm** function and the Latitude and Longitude information provided in the data
  - amenities: new columns (start with "amn_") were created to check if a selected amenity was present (represented by 1) or not (represented by 0) in a listing so that these parameters can also be used for the modelling
  - host verification: the detail about host e-mail and phone verification were stored in separate columns to be used in the modelling. For e-mail verification, it was considered to be verified even if one of the personal or work emails were verified in order to reduce complexity
- Null Values in the columns were treated using one of the following 2 approaches:
  - Imputing: Null values were imputed with default values (e.g. host response rate was considered as 0 if absent) or by assumed logic (e.g. number of beds was considered equal to number of bedrooms if absent)
  - Dropping: Rows with Null data in Bedrooms column were dropped for the category of "entire apt" since it was not possible to come up with a logic as in the case of pvt room or hotel where 1 can be assumed.

2.2 Method

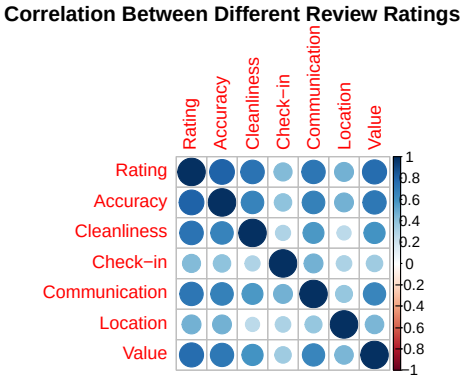
Post the cleaning process mentioned above, the different variables were explored to view their distributions, and possible correlation with other variables.

The price (outcome variable) had the following distributions when split according to the room types:



The Hotel room and Shared room had very few observations which resulted in the very narrow distribution except the few outliers. Combining or dropping these room types was considered as a solution but they were left as-is so that the model will be able to predict prices for these categories as well if required.

A correlation matrix was plotted to study the relation between different ratings provided by the guests:



During the analysis it was found that many of the properties had their host phone-numbers verified and had amenities like wi-fi etc. , since these were available in almost all the properties, their significance would be lesser, so they were not considered for the modelling.

Post model creation **VIF** function was used to check and verify if there was any multi-collinearity between the input variables for the model.

The selected variables were then used for the modelling. For this project, Linear Regression was chosen for its robustness(to noise and outliers) and simplicity(easy and fast implementation and processing).

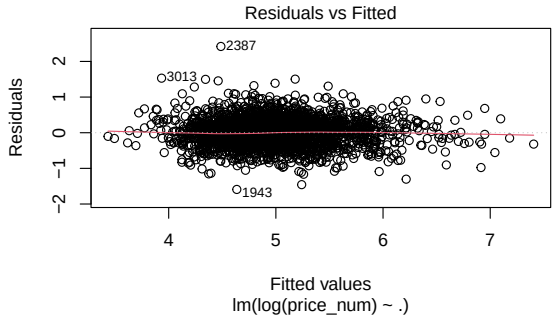
One additional benefit of Linear Regression is its interpretability, by knowing the relative imporatance of the factors from the model, the hosts can also get an idea of what changes they can make so as to get a higher value for their listing.

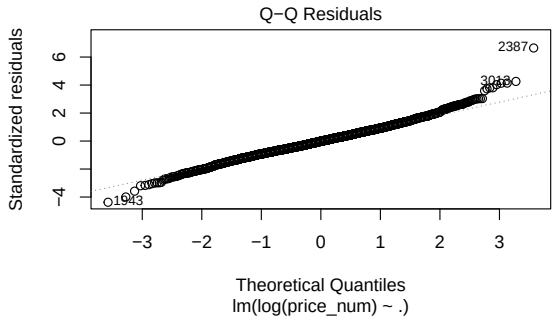
Once the model was generated the diagnostic plots were studied to validate the model, and the following 2 changes were made to correct the errors:

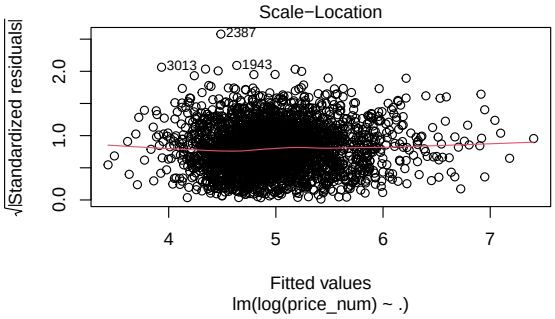
1. the outcome variable(price) was changed to log scale
2. one of the observations which was having high leverage was dropped

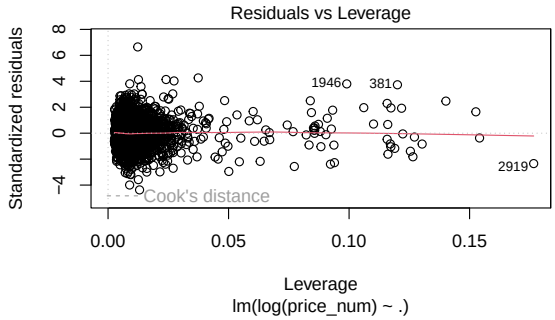
The new model was verified again using the diagnostic plots and no violations of the assumptions for Linear Regression were present.

Note: The predicted output price from the model will now be in log-scale and so has to be converted back to linear scale before using it









2.3 Conclusion

The model developed has a R^2 of 0.636 and F-Statistic of 156.4 which imply that the model is moderately effective in predicting the price as it is able to explain around 64% of the variation in price based on the input variables. The following changes or data will help in the analysis and improving the model:

- Hosts should have a pre-made list of amenities to choose from so that there is no variation and the impact of amenities can be used in a more effective way. e.g. in the current process of free-text entry "pool" can be confused between swimming pool or a pool table.
- The current data does not have information with respect to time, if this data was available additional analysis can be performed to improve the model, e.g. impact of weekends/holidays or gap between booking and check-in on the price.

Note: Since the Dataset only contains data for Ashville, the model should only be used for the listings in Ashville.