Analysis and Modeling of Los Angeles Airbnb Listing Prices

Date: Sep 19 2023

Author: Minghan Yang, 3831285746

1. Introduction

The primary aim of this report is to analyze, model, and predict the Airbnb property listing prices in Los Angeles. The dataset for this analysis has been sourced from Inside Airbnb, containing a range of features, including several categorical variables with many levels.

2. Data Preprocessing

a) Loading and Initial Exploration

The dataset was loaded from the listings.csv file. An initial examination of the price column was done, where it was then converted to a numeric datatype for ease of analysis.

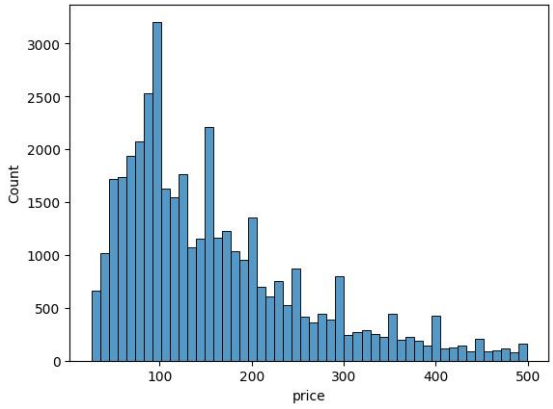
b) Price Distribution Analysis

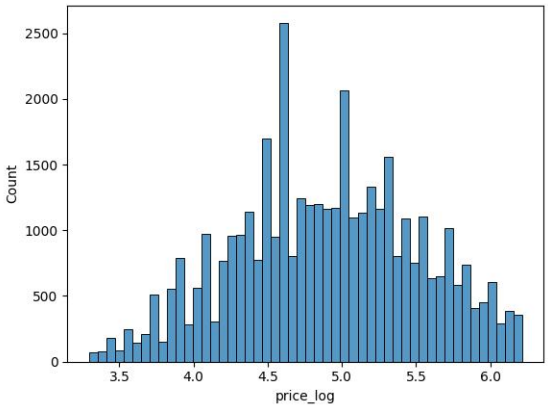
Used the describe function to show the statistical value of the price, I find there are some outlier values in the price dataset, to address this imbalance and make the distribution more symmetrical, prices outside the range of $25 to $500 were removed.

```
count    44464.000000
mean       279.030969
std        880.129338
min          0.000000
25%         89.000000
50%        148.000000
75%        250.000000
max      90150.000000
Name: price, dtype: float64
```

```
count    39979.000000
mean       159.886816
std        100.372385
min         26.000000
25%         88.000000
50%        133.000000
75%        206.000000
max        499.000000
Name: price, dtype: float64
```

I plotted the hist graph of the price and revealed that its distribution was right-skewed. To deal with the skewed, I log-transformed the price as the price_log, then the new hist graph closely resembled a normal distribution.





c) Feature Selection

I tested many different value with the correlation heatmap, then choose the best

performance ones as the model features.
The final selected features for modeling process were bathrooms, bedrooms, beds, property type, accommodates, neighbourhood, review_scores_rating and instant bookable.

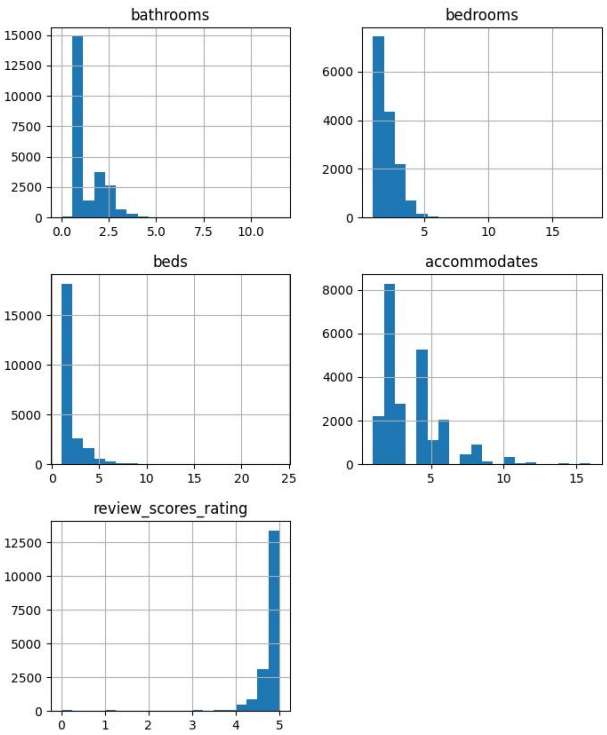
d) Data Cleaning

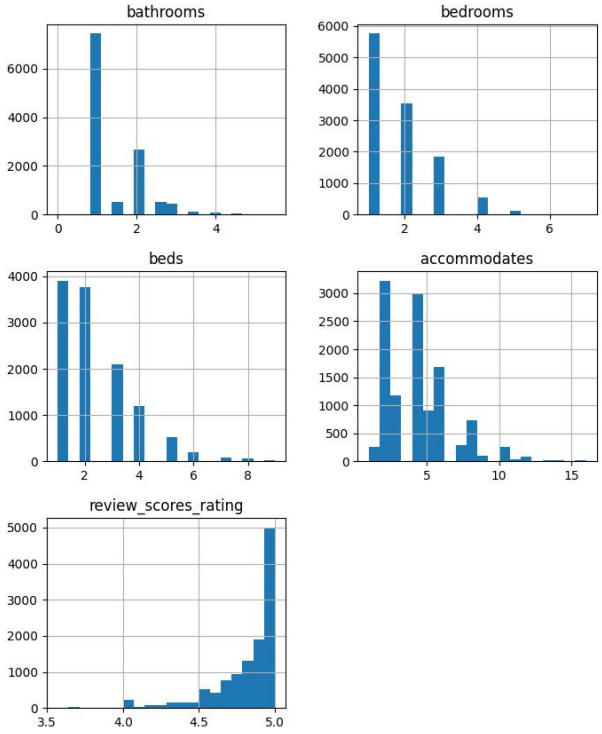
Various data cleaning steps were undertaken:
- Extracted numerical values from the `bathroom_text` column to fill the `bathroom` column.
- Minimized the number types in property_type : Aggregated less frequent (which <1000) property types into an "others" category, the final property types are 8 types.

```
property_type
Entire rental unit               10776
Entire home                       7683
Private room in home              7298
Entire guesthouse                 2405
Private room in rental unit       2172
Entire condo                      1732
Entire guest suite                1296
Entire townhouse                   691
Name: count, dtype: int64
```

```
property_type
Entire rental unit               10776
Entire home                       7683
Private room in home              7298
Other                             6617
Entire guesthouse                 2405
Private room in rental unit       2172
Entire condo                      1732
Entire guest suite                1296
Name: count, dtype: int64
```

- Addressed missing values by filling with mode value for features such as 'bathrooms', 'bedrooms', 'beds', and 'review_scores_rating'. Rows with missing 'neighbourhood' values were dropped.
- Outliers were identified using histogram plots and removed. The criteria used included bathrooms < 6, bedrooms < 8, beds < 10, and review_scores_rating > 3.5.

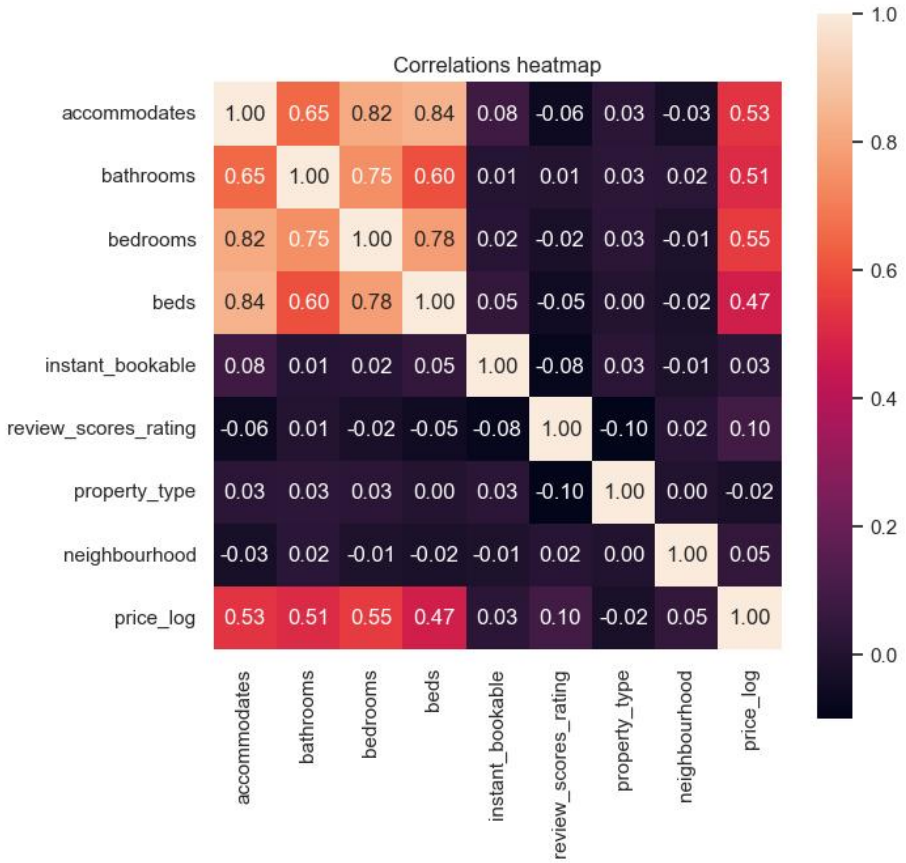




- Labeled categorical data the neighbourhood, property_type, and instant_bookable by LabelEncoder function.

A heatmap was plotted to evaluate correlations between selected features.

Red squares imply high relevance, purple squares imply irrelevance, based on the heatmap I find that the feature of bathrooms, bedrooms, beds and accommodates have a high correlation of the price_log, and review_scores_rating, neighbourhood and instant_bookable have a low correlation of the price_log, and property_type performance is bad, maybe should delete it (but it's a requirement).



e) Data Splitting

The dataset was split into training and testing sets with 70% for training and 30% for testing, and random state = 0.

f) Data Encoding and Scaling

Categorical variables were encoded using Label Encoding, and numerical features were standardized using the StandardScaler.

3. Regression Modeling

Using grid search, the best hyperparameters for each model were determined:

a) Multiple Linear Regression

- RMSE: 0.3966

b) Random Forest

- Best Parameters: {'max_depth': 10, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
- RMSE: 0.3610

c) Decision Tree

- Best Parameters: {'criterion': 'friedman_mse', 'max_depth': 10, 'min_samples_leaf': 10, 'min_samples_split': 2}
- RMSE: 0.3778

d) Support Vector Regression

- Best Parameters: {'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}
- RMSE: 0.3616

4. Results and Conclusion

The Random Forest model, with an RMSE of 0.3610, emerged as the best-performing model. Using this model, the predicted price for a 3-bedroom, 2-bathroom, instant bookable, Entire home/apt listing that accommodates 5 persons was approximately $321.10.

Overall, the Random Forest model provided a robust solution for predicting Airbnb listing prices in Los Angeles, demonstrating its potential in real-world applications for both hosts and guests to gauge appropriate pricing structures.