da3-a2-summary

Abigail Chen

Research Question

The task is to help a company operating small and mid-size apartments hosting 2-6 guests. The company is set to price their new apartments not on the market and the task is to help them make business decisions. The goal is to help them build a price prediction model.

Introduction

This project is a linear regression analysis on data about Airbnb rentals in Hawaii, USA. The dataset is taken from **Inside Airbnb Website** (http://insideairbnb.com/get-the-data.html). The relationship between price and room type, neighborhood along with other parameters will be used. The raw dataset has 24,294 observations.

Exploratory Data Analysis

First, the structure of the data was analysed. Then the various data types were changed to categorical from character format. The price had outliers and its distribution was right skewed. The "ln" was used and outliers were also removed from the list. So, for modeling our dependent variable was ln_price.

Variables Selection

Below is the list of shortlisted variables and the reasons behind their selection

1. **ln_price**: This is our dependent variable.
2. **room_type**: Entire home/apt has highest and shared room has lowest, tried using anova test to show that different room types have differnt affect on ln_price.
3. **neighbourhood_group**: Significant difference in ln_prices were seen based on neighbourhood_group, Honolulu has least and Maui has the highest
4. **minimum_nights**: The price seems to increase with increase in minimum nights , plot and anova test proves it.

Main model Linear Regression : lm

Formulas for seven lm models

$$lnPrices = \beta_0 + \beta_1(roomtype)$$

$$lnPrices = \beta_0 + \beta_1(neighbourhoodGroup)$$

$$lnPrices = \beta_0 + \beta_1(minimumNights)$$

$$lnPrices = \beta_0 + \beta_1(roomtype) + \beta_2(neighbourhoodGroup)$$

$$lnPrices = \beta_0 + \beta_1(roomtype) + \beta_2(minimumNights)$$

$$lnPrices = \beta_0 + \beta_1(neighbourhoodGroup) + \beta_2(minimumNights)$$

$$lnPrices = \beta_0 + \beta_1(minimumNights) + \beta_2(minimumNights) + \beta_3(roomtype)$$

External Validity

The data set was divided randomly into train and test in 80-20 percent ratio.

Alternative model: LASSO

Lasso regression is a classification algorithm that uses shrinkage in simple and sparse models(i.e model with fewer parameters). In Shrinkage, data values are shrunk towards a central point like the mean. Lasso regression is a regularized regression algorithm that performs L1 regularization which adds penalty equal to the absolute value of the magnitude of coefficients.

"LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is good for models showing high levels of multicollinearity or when you want to automate certain parts of model selection i.e variable selection or parameter elimination." so formula ln_price ~. we dont need preprocess and feature selaction. results of the Lasso model were not better but were close to Model 7. Sources looked into: https://www.geeksforgeeks.org/lasso-regression-in-r-programming/

below are the results of models on train data

Table 1: Models performance on train data

Model	Formula	Coefficients	R_squared	BIC	RMSE_Train	RMSE_K_Fold
Model 1	ln_price ~ room_type	4	0.0436848	28244.28	0.6685475	0.6686566
Model 2	ln_price ~ neighbourhood_group	4	0.1075511	27285.13	0.6458377	0.6460280
Model 3	ln_price ~ minimum_nights	8	0.0868160	27642.01	0.6532973	0.6535551
Model 4	ln_price ~ room_type + neighbourhood_group	7	0.1360116	26863.99	0.6354563	0.6355654
Model 5	ln_price ~ room_type + minimum_nights	11	0.1207339	27145.38	0.6410500	0.6414136
Model 6	ln_price ~ neighbourhood_group + minimum_nights	11	0.1642664	26440.74	0.6249793	0.6255038
Model 7	ln_price ~ neighbourhood_group + minimum_nights + room_type	14	0.1885714	26059.80	0.6158244	0.6165003
LASSO	ln_price ~.	11	0.1677446	NA	NA	0.6343082

Below are the results of models on test data

Table 2: Models performance on train data

Model	ME	RMSE	MAE	MPE	MAPE
Model_K 1	0.0069977	0.6615796	0.5067882	-1.255719	9.229519
Model_K 2	0.0033055	0.6420284	0.4866146	-1.229356	8.818409
Model_K 3	0.0005455	0.6474126	0.4992776	-1.321552	9.115025
Model_K 4	0.0046092	0.6344103	0.4823687	-1.157150	8.714107
Model_K 5	0.0017642	0.6387564	0.4935248	-1.241901	8.982829
Model_K 6	-0.0009688	0.6238330	0.4752401	-1.225922	8.612448
Model_K 7	0.0001930	0.6174927	0.4719185	-1.163068	8.530778

Conclusion

Ideally, lower RMSE and higher R-squared values are indicative of a good model. BIC and RMSE show that model 7 is best. LASSO did well but not better than model 7. Model 7 performs better on test data as well.