**Linear Regression on LA Airbnb Data:**

**Predicting Airbnb Rental Price using Multiple Linear Regression**

Alexander Vaillant

College of Information Technology, Western Governors University

D214: MSDA Capstone

Dr. Daniel Smith

August 27, 2021

**Research Question**

**“**To what extent do the independent variables of Airbnb rentals predict the rental price in the Los Angeles Market?**”**

**Justification for Research Question**

According to the 2021 Airbnb Statistics Article released by Steve Deane, there are over “14,000 new hosts joining Airbnb each month in 2021” (Deane, 2021). Deane references Airbnb statistics from 2019 which deemed Los Angeles, California as one of the most popular cities for Airbnb in the US (Deane, 2021). To successfully enter a competitive market like Los Angeles, new hosts must know the financial impact that different features of an Airbnb rental have on its price. In their paper “Real Estate Value Prediction Using Linear Regression”, Ghosalkar and Dhage utilize linear regression to predict the value of real estate (Ghosalkar & Dhage, 2018). Multiple Linear Regression has been a proven method used to accurately predict price based on various features while accounting for their impact on the variance of price.

**Context**

The contribution of this study to the MSDA program and the Data Analytics field is to create a predictive model which approximates an Airbnb’s rental price so that a new host in the Los Angeles market may gauge a potential property’s affordability and revenue against competitors. With 32,241 listings in the Los Angeles market, price and the variables with influence on price play a crucial role in the revenue of an Airbnb. In this study, a Multiple Linear Regression model will be utilized to analyze the statistical significance of independent, or predictor, variables which have the most influence on an Airbnb’s rental price (dependent variable). When these highly influential predictor variables are known, a host may cater to those areas to attract customers. “Multiple regression allows for a relationship to be modeled between multiple independent variables and a single dependent variable where the independent variables are being used to predict the dependent variable” (Laerd Statistics, 2015). In their paper “Real Estate Value Prediction Using Linear Regression”, Ghosalkar and Dhage utilize linear regression to predict the value of real estate (Ghosalkar & Dhage, 2018). Like with real estate value, linear regression can be used to predict the rental price of an Airbnb rental.

**Null and Alternate Hypotheses**

The null hypothesis (H0) of this statistical analysis is that a statistically significant model cannot be created to predict the Airbnb rental price. The alternate hypothesis (Ha) is that a statistically significant model can be created to predict the Airbnb rental price. The acceptance or rejection of the null hypothesis will be decided based on the p-value of the Multiple Linear Regression model created.

**Data Collection**

**Description of Relevant Data**

The data needed to be collected for this study is publicly available through Inside Airbnb website. “Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world” (Cox, n.d.). The data was compiled from public information on the Airbnb website by Murray Cox. Before any data cleansing and removal is done, the Los Angeles dataset contains 32,241 rows with 74 columns and data sparsity of less than 10%. Host demographic and PII data will be removed before analysis.

This dataset is available through Inside Airbnb here: <http://insideairbnb.com/get-the-data.html>. The study will place a delimitation on the dataset by limiting the neighborhood group to only the city of Los Angeles. There are two other neighborhood groups with observations that will be removed. Another potential delimitation would be the number of features allowed in the model based on stepwise regression. The limitations of this dataset are the presence of host PII and the data sparsity of 10% in both categorical and continuous variables. These limitations can be worked around in the data cleansing process. After PII data is removed, the dataset contains the following 31 usable variables:

|  |  |  |
| --- | --- | --- |
| Variable | Type | Intention |
| Id | Continuous | Index |
| host\_response\_time | Categorical | Predictor/Independent |
| host\_response\_rate | Continuous | Predictor/Independent |
| host\_acceptance\_rate | Continuous | Predictor/Independent |
| host\_is\_superhost | Categorical | Predictor/Independent |
| host\_listings\_count | Continuous | Predictor/Independent |
| host\_has\_profile\_pic | Categorical | Predictor/Independent |
| host\_identity\_verified | Categorical | Predictor/Independent |
| neighbourhood\_group\_cleansed | Categorical | Predictor/Independent |
| room\_type | Categorical | Predictor/Independent |
| accommodates | Continuous | Predictor/Independent |
| bathrooms\_text | Continuous (once cleansed) | Predictor/Independent |
| bedrooms | Continuous | Predictor/Independent |
| beds | Continuous | Predictor/Independent |
| price | Continuous | Response/Dependent |
| minimum\_nights | Continuous | Predictor/Independent |
| maximum\_nights | Continuous | Predictor/Independent |
| has\_availability | Categorical | Predictor/Independent |
| availability\_30 | Continuous | Predictor/Independent |
| availability\_60 | Continuous | Predictor/Independent |
| availability\_90 | Continuous | Predictor/Independent |
| availability\_365 | Continuous | Predictor/Independent |
| number\_of\_reviews | Continuous | Predictor/Independent |
| review\_scores\_rating | Continuous | Predictor/Independent |
| review\_scores\_accuracy | Continuous | Predictor/Independent |
| review\_scores\_cleanliness | Continuous | Predictor/Independent |
| review\_scores\_checkin | Continuous | Predictor/Independent |
| review\_scores\_communication | Continuous | Predictor/Independent |
| review\_scores\_location | Continuous | Predictor/Independent |
| review\_scores\_value | Continuous | Predictor/Independent |
| instant\_bookable | Categorical | Predictor/Independent |

The data was compiled from public information on the Airbnb website by Murray Cox. “The data is available under a Creative Commons CC0 1.0 Universal (CC0 1.0) ‘Public Domain Dedication’ license” (Cox, n.d.). Based on the Creative Commons CC0 1.0 Universal license, a user can “copy modify, distribute and perform the work, even for commercial purposes, all without asking permission” (Creative Commons, n.d.). it can be used for commercial purposes. There are several other Airbnb datasets available on Kaggle with public domain licenses. For example, another dataset can be found at https://www.kaggle.com/kritikseth/us-airbnb-open-data with the CC0: Public Domain license. All host PII data will be removed at the start of the analysis to increase privacy.

The dataset will be downloaded from the Inside Airbnb website in .gz format. From there, the listings.csv.gz can be scraped with a few for loops in Python and the final listings.csv file will be exported to be cleansed of host PII data in excel. The sparsity in this dataset is less than 10%.

**Advantages and Disadvantages of Data-Gathering Methodology**

With the dataset publicly available on Insider Airbnb, an advantage is the cited Creative Commons Public Domain License. This allows the user to distribute and analyze the data for commercial purposes without legal consequences. Another advantage is the data compilation’s monthly frequency. In the current Covid pandemic, data and situations are undergoing continuous change, and having up-to-date data is pertinent for success in competitive markets.

In this published dataset, there are some removed and calculated features. This is a disadvantage without the presence of a fully completed data dictionary. Secondly, each feature contains missing values. Another disadvantage of this dataset is the inclusion of Airbnb host Personal Identifiable Information, or PII.

**Overcoming Any Challenges Encountered in Data Collection Process**

To overcome removed and calculated features, the dataset comes with a data dictionary to boost understanding of each feature. The data dictionary is thorough and explains most features. The challenge of missing values in features is overcome through K-Nearest Neighbors Imputation on continuous and binary dummy features transformed from categorical features. The final challenge of host PII is dealt with through the removal of any feature containing PII.

**Data Extraction and Preparation**

**Extraction and Preparation Process**

1. Import necessary libraries

Graphical user interface, text, application

Description automatically generated

1. Extract data from original downloaded .gz file

Text

Description automatically generated

1. Load raw data into pandas dataframe for EDA and cleaning

Graphical user interface, text, application, email

Description automatically generated

1. Select desired features without any host PII

Table

Description automatically generated

1. Perform Exploratory data analysis
   1. Check Initial Data Sparsity

A picture containing text

Description automatically generated

* 1. Descriptive Summary of the Data

Table

Description automatically generated

1. Limit data to only the City of Los Angeles (Delimitation #1)

Text

Description automatically generated

1. Check data sparsity after delimitation in case it eliminated missing values.

Table

Description automatically generated

1. Convert percentages and currency to Floats (for MLR Algorithm)

Table

Description automatically generated

1. Parse Bathrooms\_text column and convert to Float (for MLR Algorithm)

Graphical user interface, text, email

Description automatically generated

1. Convert Categorical Variables into Binary Dummy Variables

Table

Description automatically generated

1. Perform KNN Imputation

Table

Description automatically generated

1. Check final data sparsity

Table

Description automatically generated

1. Normalize data after KNN Imputation (Normality Assumption of MLR Algorithm)

Table

Description automatically generated

1. Define calculate\_VIF() function to (Minimize Multicollinearity)

Graphical user interface, text, application

Description automatically generated

1. Split dataframe into X and y values for VIF check

Graphical user interface, text

Description automatically generated

1. Use VIF function to Remove Multicollinearity

Table

Description automatically generated

1. Split Data into Training and Testing datasets (80-20; Model Evaluation)

Graphical user interface, text, application, email

Description automatically generated

1. Perform Stepwise Regression (Feature Selection)

Graphical user interface, text, application, email

Description automatically generated

1. Remove non-supported columns from training and testing dataset

Graphical user interface, table

Description automatically generated with medium confidence

1. Export Post-Stepwise Regression Training and Testing Datasets (ready for modeling).

Text

Description automatically generated

**Explanation of Tools and Techniques**

***Pandas***

The Pandas library was used to load data, create dataframes, export dataframes to csv files, describe the data, count values, check data sparsity, and create binary dummy variables.

***NumPy***

The NumPy library was used to transform dataframes to arrays, reshape arrays, and set the random seed for reproducible results.

***Gzip***

The Gzip library was used to extract the csv file from the downloaded raw .gz file.

***Shutil***

The Shutil library was used to extract the csv file from the downloaded raw .gz file.

***KNNImputer***

KNNImputer is a function from sklearn.impute. It was used to impute missing values in categorical and continuous features.

***Re***

The Re, or Regular Expression, library was used to remove symbols in the process of converting text and currency columns to floats.

***Seaborn***

The Seaborn library was used to plot the correlation heatmap of the data.

***Normalize***

Normalize is a function from sklearn.preprocessing. It was used to normalize the dataset at the dataset grain level rather than at the column level.

***LinearRegression***

LinearRegression is a function from sklearn.linear\_model. It was used in the Feature Selection process to instantiate the LinearRegression estimator.

***SequentialFeatureSelector***

SequentialFeatureSelector is a function from sklearn.feature\_selection. It was used to perform Forward Stepwise Regression for feature selection.

***Variance\_Inflation\_Factor***

Variance\_Inflation\_Factor is a function from statsmodels.stats.outliers\_influence. It was used to detect, remove, and minimize multicollinearity among the features.

**Justification of Tools and Techniques**

According to Brownlee, “Datasets may have missing values, and this can cause problems for many machine learning algorithms” (Brownlee, 2020). In preliminary exploratory data analysis, over half of the 32,241 observations have at least one feature with a missing value. Rather than remove half of the observations, this study utilizes the KNN algorithm to impute missing values. In his 2020 paper “kNN Imputation for Missing Values in Machine Learning”, Brownlee states that “the k-nearest neighbor (KNN) algorithm has proven to be generally effective” at predicting and imputing missing values (Brownlee, 2020). This imputation will account for both categorical and continuous variables.

Removal of multicollinearity is an assumption of Multiple Linear Regression models. According to Jong Hae Kim, “Diagnostic tools of multicollinearity include the variance inflation factor (VIF)” (Kim 2019). This study utilizes VIF with a threshold of 10 to reduce multicollinearity among the predictor variables.

***Advantages and Disadvantages of Tools and Techniques***

Advantages of using KNN Imputer would be the ease of the default hyper-parameters, how it uses an algorithm to impute missing values based on the nearest neighbors, and there are no assumptions about the data needed. The resulting missing values would be a more accurate imputation over a generic mean value imputation. One disadvantage would be that with the large amount of data processed, it is computationally expensive. The KNN Imputation stage takes a few minutes to run.

**Analysis**

**Description of Analysis Technique and Process**

Python was used for the creation, exploration, and evaluation of the Multiple Linear Regression model. After data cleansing, the dataset was randomly split into training and testing sets of 80% and 20% size, respectively. The training set is used in the model fitting phase while the testing set is used in the model evaluation phase. Forward Stepwise regression was used to identify the most impactful variables of the dataset on an Airbnb’s rental price. There were 29 initial independent variables, excluding the ID field. Forward Stepwise regression built a Linear Regression model by adding one variable at a time based on the explained variance. Without any restrictions on the number of variables to include, the Forward Stepwise Regression selected 10 independent variables. The Multiple Linear Regression model uses the 10 independent variables to predict the outcome of the Airbnb Rental’s price. The resulting model summary indicates that the model and each of the 10 independent variables are statistically significant with a p-value extremely close to zero.

***Calculations and Outputs***

Table

Description automatically generated

Please note that the p-value, which is shown as Prob (F-statistic), is 0, statistically significant, in favor of rejecting the null hypothesis. Also, each independent variable has a p-value under 0.05, which indicates that each independent variable and the constant are statistically significant. The R-Squared and Adjusted R-Squared values are approximately 0.82 or 82% explained variance.

Chart, scatter chart

Description automatically generated

Please note the above residual plot of the training set and that the residuals do not appear to be normally distributed.

Graphical user interface, text, application, email

Description automatically generated

Please note that the model was evaluated on unseen, unknown data using the metrics of Mean Squared Error and R-Squared. The final model explained about 82% of the variance in Price and had a low Mean Squared Error value of ~0.015.

Chart, scatter chart

Description automatically generated

Please note the above residual plot of the testing set and that the residuals do not appear to be normally distributed.

**Justification of Analysis Technique**

According to a Udacity India Article by Prince Patel, “the main reason for using Python would be readability, versatility and easiness” (Patel, 2018). Since this study is meant to be accessible to new hosts and Multiple Linear Regression is “usually the first machine learning algorithm that every data scientist comes across”, it affords the newer data enthusiasts an easier route to utilize this study in commercial practice (Agarwal, 2018). According to Laerd Statistics, Multiple Linear Regression would be a viable method for this study as it is “used to predict a continuous dependent variable based on multiple independent variables” (Laerd Statistics, 2015).

***Advantages and Disadvantages of Analysis Technique***

According to Mohammad Waseem, the advantages of Multiple Linear Regression include that it is “easier to implement and interpret, performs well on linear data, and handles overfitting well when used with dimensionally reduction techniques” (Waseem 2021). Disadvantages of Multiple Linear Regression are that there are “assumptions on linearity” and the distributions of the data and it is “prone to multicollinearity” (Waseem 2021).

**Data Summary and Implication**

**Summarize implications by discussing results with 1 limitation**

Table

Description automatically generated

Please note the above image shows the coefficients of the final Multiple Linear Regression Model on normalized data. Each coefficient represents a variable’s one-normalized unit effect on price when all other variables remain the same. For example, for each one-normalized unit increase in the number of bedrooms, the rental price is increased by ~19.56 given all other variables remain the same. Variables with negative coefficients have a negative impact on price when all other variables remain the same. The five most impactful variables on price given the size of their coefficients are accommodates, bedrooms, host\_is\_superhost\_true, room\_type\_Private room, and room\_type\_Shared room.

Table

Description automatically generated

Please note the above image shows the MLR Model’s output summary. Each independent variable and constant have a p-value of approximately 0 and are statistically significant. The final model has a p-value of nearly 0 and is statistically significant. The p-value supports a rejection of the Null Hypothesis and acceptance of the Alternate Hypothesis that a statistically significant model can be created to predict the Airbnb rental price.

The R-Squared and Adjusted R-Squared values are within a close range of each other, and this indicates that the model is not overfitting. The Condition Number is large and in the notes section, there is an indication of either strong multicollinearity or other numerical problems. VIF was used to minimize multicollinearity. However, there is still multicollinearity present in the dataset. Multiple Linear Regression requires assumptions about parameters and distributions. With the presence of multicollinearity, any further analysis may be best explored using non-Parametric analysis methods.

**Recommendations**

Based on the results of this study, it is recommended that a new Airbnb host in the Los Angeles market should focus on locating properties that can be rented entirely, accommodate more people, and have a larger number of bedrooms while striving to become a superhost on Airbnb. In a 12-month period, Airbnb superhosts “must maintain a 90% response rate, 1% cancellation rate, and 4.8 overall rating after completing at least 10 trips” (Airbnb n.d).

**Proposals for Future Study of Dataset**

***Proposal One***

Given the lingering presence of Multicollinearity after the VIF check and the large condition number, this study proposes using a non-Parametric method such as KNN to further analyze the dataset. The KNN algorithm does not require strict assumptions of the underlying data and would avoid the issue of multicollinearity.

***Proposal Two***

If Parametric analysis is necessary, this study also proposes using another feature selection method such as Principal Component Analysis for dimensionality reduction. Principal Component Analysis would reduce the number of features while further accounting for multicollinearity.

***Proposal Three***

Further analysis should be done on the various neighborhoods in the Los Angeles market available in this dataset. The dataset was delimited to only include the neighborhoods within the City of Los Angeles. A variety of neighborhoods and locations may present other findings of impactful variables.

**References**

Agarwal, A. (2018, October 5). *Linear Regression using Python*. Towards Data Science. <https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2#>.

Airbnb (n.d). *How do I become a Superhost*. Airbnb. <https://www.airbnb.com/help/article/829/how-do-i-become-a-superhost>.

Brownlee, J. (2020, June 24). *kNN Imputation for Missing Values in Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/>.

Cox, M. (n.d.). *Get the Data*. Inside Airbnb. <http://insideairbnb.com/behind.html>.

Creative Commons. (n.d.). *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication*. Creative Commons. <https://creativecommons.org/publicdomain/zero/1.0/>.

Deane, S. (2021, January 26). *2021 Airbnb Statistics: Usage, Demographics, and Revenue Growth*. Stratos Jet Charters Inc. <https://www.stratosjets.com/blog/airbnb-statistics/>.

Ghosalkar, N., & Dhage, S. N. (2018). *Real Estate Value Prediction Using Linear Regression*. Semantic Scholar. <https://www.semanticscholar.org/paper/Real-Estate-Value-Prediction-Using-Linear-Ghosalkar-Dhage/f2308f0a4f0981801b518b9ca2152bcb4c797ad7>.

Kim, J. H. (2019, July 15). *Multicollinearity and misleading statistical results*. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/>.

Laerd Statistics (2015). *Multiple regression using SPSS Statistics*. Laerd Statistics. <https://statistics.laerd.com/premium/spss/mr/multiple-regression-in-spss-20.php>.

Patel, P. (2018, March 8). *Why Python is the most popular language used for Machine Learning*. Medium. <https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77>.

Waseem, M. (2021, July 15). *How to Implement Linear Regression for Machine Learning?* Edureka. <https://www.edureka.co/blog/linear-regression-for-machine-learning/>.