**Linear Regression on LA Airbnb Data:**

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

***Executive Summary***

Alexander Vaillant

College of Information Technology, Western Governors University

D214: MSDA Capstone

Dr. Daniel Smith

August 27, 2021

**Statement of Problem and Hypotheses**

**Research Question**

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

**Problem**

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. As rental price is a continuous, dependent variable, Multiple Linear Regression can be utilized to analyze the data. This study sought to produce a statistically significant Multiple Linear Regression model which predicts Airbnb rental price. With this trained model, new Airbnb hosts can evaluate their potential properties for higher profit margins.

**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.

**Summary of Data Analysis Process**

The raw data file was downloaded from Insider Airbnb and parsed in Python. All variables containing host PII were removed, and the raw data was initially explored. The data was limited to only the neighborhood of “City of Los Angeles”. To use KNN Imputation and build the Multiple Linear Regression model, many of the variables needed to be cleansed. The categorical variables were transformed into binary dummy variables. Any missing values were imputed using KNN Imputation. The resulting data was normalized and a VIF check was performed to minimize instances of Multicollinearity between variables. The data was split into training and testing datasets and feature selection using Forward Stepwise Regression was performed. The most impactful features were extracted and used to limit the training and testing datasets. The Multiple Linear Regression model was trained using the cleansed training dataset. Independent variables and the overall model were checked for p-values under 0.05 to show statistical significance. The model was evaluated using Mean Squared Error, both regular and Adjusted R-Squared, and residual plots on unseen (testing) data. Recommendations and direction proposals were created from the model summary and evaluation metrics.

**Outline of Findings**

There was a statistically, significant model found, and the null hypothesis was rejected in favor of the alternate hypothesis. The final, statistically significant model consisted of 10 independent variables and one constant to predict Rental Price. Both regular and Adjusted R-Squared values were above 0.80 indicating that the model explains over 80% of the variance in Rental Price. There are five variables with absolute coefficients over one: Accommodates, Bedrooms, Host\_Is\_Superhost\_T, Room\_type\_Private room, and Room\_type\_Shared room. These five variables have a stronger impact on an Airbnb’s Rental price. Although the model was statistically significant, there was a large Condition Number found which suggests the lingering presence of Multicollinearity or other numerical errors in the final model.

**Explanation of Technique & Tools’ Limitations**

**KNN Imputation**

Although KNN Imputation can be executed easily with the default hyper parameters, it can be computationally expensive with large amounts of data. Also, it requires the input data to be continuous variables. Any categorical variables must be transformed into continuous variables prior to KNN Imputation.

**Scaling (Normalization)**

Normalization is a standard, scaling practice in Regression analysis. However, the type and level of scaling used heavily depends on the dataset. Making these executive decisions on how to scale the data is above the beginner user’s ability level. This study tried other scaling methods such as Standardization and MinMax scaling by column.

**Multiple Linear Regression**

Limitations of Multiple Linear Regression include that there are “assumptions on linearity” and the distributions of the data and it is “prone to multicollinearity” (Waseem 2021). As seen in the final model, there is still a potential presence of multicollinearity. Non-parametric methods of analysis that do not make assumptions on the distributions of the data may provide better prediction results. Also, PCA could have been used to reduce multicollinearity.

**Variance Inflation Factor (VIF)**

VIF shows how highly correlated a variable may be with other variables, but this method does not explicitly indicate which other variables it may be highly correlated with in the dataset. Domain knowledge would further help with minimizing multicollinearity using VIF.

**Stepwise Regression**

This feature selection method can be affected by multicollinearity. VIF was used before to reduce multicollinearity. However, it may not always be successful.

**Summary of Proposed Actions**

**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**

***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.

**Expected Benefits of Study**

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. According to her article “Machine Learning Consulting Rates: What to Expect?”, Kavita Ganesan found the average data science or machine learning freelancer charged about $100 per hour for consulting projects (Ganesan, n.d). With small projects lasting over 50 billable hours, a freelance can make $5000 for a week’s worth of work. However, with a streamlined version of this final model, Airbnb hosts can pay a monthly subscription fee to use the model. Similar analytics companies charge members ~$100 per month for analytics of popular markets like Los Angeles. For new hosts, a $100 price tag may deter them from purchasing the services. However, by targeting this demographic and making the pricing more affordable, the new subscription service stands to make a substantial profit while increases the profits of their members with affordable, insight analytics used to reduce common, beginner host mistakes.

**References**

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

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

Ganesan, K. (n.d.). *Machine Learning Consulting Rates: What to Expect*. Opinosis Analytics. <https://www.opinosis-analytics.com/blog/machine-learning-consulting-rates/#.YTKftI5KjZk>.

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