**Report**

This report aims to answer the question of which areas in New York command the highest client engagement for Airbnb and the highest return on investment (ROI) via a forecasting model.

**Executive Summary:**

This comprehensive analytical report elucidates the multifaceted determinants influencing the pricing schema of Airbnb listings in New York City for the fiscal year 2019. The insights provided are poised to refine pricing strategies, foster optimal market positioning, and pave the way for empirically informed decision-making processes.

**Introduction:**

In the sharing economy, Airbnb is a pioneering platform that has revolutionised global hospitality and lodging norms. This study thoroughly investigates the economic, sociological, and infrastructural factors influencing Airbnb listing prices in New York City.

**Methodology:**

As Alsghaier et al. (2017) suggested, a convergent analytical approach was employed, merging quantitative and qualitative data analysis (DA). The foundational statistical examinations were supplemented by advanced machine learning models, ensuring robustness in the findings using the following five steps:

***Exploratory DA***

Numeric data is analysed using descriptive statistics, while dominant categories are identified for categorical data. A heatmap indicates the correlation strength between numeric variables for better interpretation (Nguyen and Nguyen, 2021).

***Data Pre-processing:***

Before analysis, the data was cleaned, addressing missing values and retaining natural outliers. While the data quality was generally sound, categorical variables were encoded using sentiment scoring, one hot and frequency encoding. Notably, missing reviews could hint at new listings, but this was beyond the study's scope.

***Statistics-based D****A****:***

Lawani et al. (2017) performed a sentiment analysis on listing names to investigate the link between names and pricing. The results suggested that upscale locations tend to have more positive listing names. Chi-square, Pearson, and Spearman correlation tests similarly confirmed the positive relationship between location and pricing.

***Data Visualisation (****DV****):***

This report prioritised feature analysis guided by DV principles (Post et al., 2002) employing Random Forest (RF) and Gradient Boosting (GB) models to extract key feature importance metrics, facilitating the creation of a comprehensive overview. Price variations across New York City neighbourhoods were also investigated, uncovering potential links with the local economic landscape. This holistic approach yielded profound insights into the NYC Airbnb market dynamics and laid a robust foundation for further investigation.

***Unsupervised Machine Learning (UML) and Quantile Regression (QR):***

UML, specifically via clustering, was employed to discern patterns and trends, whereas GB was leveraged for its predictive prowess (Brown, 2021). Multivariate Regression (MR) was applied to predict the impact of each variable within its respective cluster. Based upon their salient features, the clustering of listings explained locations and price brackets that potentially provided heightened client engagement, as manifested in reviews.

**Key Findings:**

**A group of blue and black graphs

Description automatically generated with medium confidence**

Figure 1

Our results exhibit natural outlier values in "price," "minimum\_nights," "number\_of\_reviews" (and thus "reviews\_per\_month"), and "calculated\_host\_listings\_count." Notably, these variables are unevenly distributed. "Price" is right-skewed, and "neighbourhoods" and "room types" show uneven distributions (see Figure 2).

A graph of a number of people

Description automatically generated with medium confidence

Figure 2

Unsurprisingly, entire homes are, on average, more expensive than other types. In addition, prices in Manhattan and Brooklyn are the highest (median and mean). See Figure 3.A graph of average price

Description automatically generated with medium confidence

Figure 3

*Correlational Dynamics*

A screenshot of a graph

Description automatically generated

Figure 4

Spearman's Rank Correlation analysis revealed several associations within the dataset, including a correlation between location (longitude) and pricing and potential insights into listing tenure based on ID correlations. An intriguing observation emerged where properties with higher availability tended to accumulate more reviews, raising questions about review authenticity. Additionally, hosts with more listings received slightly more reviews, while longer minimum stay requirements were correlated with fewer reviews. These findings provide valuable insights into the dataset's characteristics and relationships, warranting further investigation.

*Categorical Variables and Price Associations*

A screenshot of a computer

Description automatically generated

Figure 5

Figure 5 displays the Chi-square test outcomes, revealing a significant relationship between the location variables, namely "neighbourhood" and "neighbourhood\_group," and the synthetic variable "price\_category." Furthermore, the analysis highlights a noteworthy association between the "host\_name" variable and "price\_category," suggesting that the number of listings attributed to a host is linked to the pricing category.

*Feature importance*

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Table 1

Using the RF and GB (Li, 2020) model, Table 1 showcases the feature importance derived from predictive analyses, revealing that variables including location, room type, availability and minimum nights are the primary determinants of price, highlighting their crucial role in influencing economic valuations. Surprisingly, the (auto increment) IDs are considered essential, yielding a possibility for the duration a listing stays on the platform to be relevant.

*Quantile pricing forecast:*

Employing a relatively wide prediction interval (0.2 for the lower and 0.8 for the upper alpha) has yielded a prediction accuracy of approximately 60%. Consequently, achieving a substantially higher level of performance through feature engineering or hyperparameter tuning appears improbable. Instead, it is advisable to explore the possibility of augmenting the dataset with additional and diverse variables to enhance predictive capabilities.

A screenshot of a computer

Description automatically generated

Table 2

A black text on a white background

Description automatically generated

*Geographical Price Dispersion*

By applying k-means clustering iteratively on combinations of all variables, various demarcations into discernible clusters emerged, typified by distinct economic brackets and characteristic attributes, as shown in Figure 2.

A map of a city

Description automatically generated with medium confidenceA map of different colored dots

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer program

Description automatically generated

Figure 6

Interestingly, clustering the variables 'name' (listing sentiment score), 'calculated\_host\_listings\_count', 'room\_type\_Entire home/apt', 'reviews\_per\_month' produces a similar pattern.

**Conclusions**

As the above results have shown, the intricate tapestry of determinants influencing Airbnb pricing is punctuated by tangible assets like amenities and spatial considerations and intangible elements like neighbourhood prestige. It is this complex dynamic that is imperative for a structured pricing strategy. For example, desirable locations lead to higher demand, prices, limited availability, and more reviews. However, costly listings tend to have fewer reviews and greater availability.

Using Multivariate Regression with clustering to assess variable effects on price is effective but explains only a fraction of the variance. Non-linear models perform better, yet Generalised Linear Models (GLM) with a preferred Gamma distribution yield low coefficients, and Gaussian Distribution occasionally produces unlikely results.

Importantly, due to the low dimensionality of the dataset, temporal conditions were ignored. For example, a review can only come after a price has been set. Depending on the purpose of the model, this can lead to problems.

**Recommendations**

Price predictability needs improvement (currently around 60% accuracy) to make this approach helpful. This limitation is likely due to the circumscribed data dimensionality needing more details like property size, amenities, and age. Once price predictability improves, multivariate models can forecast variable impacts in specific clusters, benefiting hosts with revenue-boosting recommendations and increasing Airbnb's commissions. In conclusion, Airbnb is poised to benefit from developing sophisticated analytical tools which can be harnessed to offer real-time, data-driven pricing recommendations to hosts, thereby elevating both profitability and stakeholder satisfaction.

**References:**

Alsghaier, H., Akour, M., Shehabat, I. and Aldiabat, S., 2017. The importance of big data analytics in business: a case study. American Journal of Software Engineering and Applications, 6(4), pp.111-115.

Brown, S. (2021). Machine learning, explained. [online] MIT Sloan. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

Campbell, D., 2001. The long and winding (and frequently bumpy) road to successful client engagement: one team's journey. System Dynamics Review: The Journal of the System Dynamics Society, 17(3), pp.195-215.

Lawani, A., Reed, M.R., Mark, T. and Zheng, Y., 2019. Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. Regional Science and Urban Economics, 75, pp.22-34.

Li, Q., Wen, Z. and He, B., 2020, April. Practical federated gradient boosting decision trees. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 4642-4649).

Nguyen, G., Kim, D. and Nguyen, A., 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. Advances in Neural Information Processing Systems, 34, pp.26422-26436.

Post, F.H., Nielson, G. and Bonneau, G.P. eds., 2002. Data visualisation: The state of the art.

Reinhold, S. and Dolnicar, S., 2018. Airbnb’s business model. Peer-to-peer accommodation networks, p.27.

**Appendix:**

\*\*\*\*\*Code is submitted as a separate notebook on Turnitin\*\*\*\*\*