# BNB Smart Scan – ML Backed Benchmarking of Airbnb Listings

**GitHub URL: BNB SmartScan**

## Abstract

With a market valuation of USD$104B, Airbnb is the go-to platform for on-demand rentals. But with inflationary pressures weighing down on discretionary spending, the question of how to sustain growth post the pandemic becomes pressing. The paper tackles: (a) the key role of hosts and (b) how Airbnb can leverage on ML driven benchmarking and price prediction models to help property owners and their proxies calibrate price and amenity offerings.

## 1. Background and Motivation

### 1.1 Macro Environment

COVID-19 has changed the dynamics of travel globally, with shifts in guest behavior driven by the following:

**1.1.1 Prevalence of remote work:**

Guest behavior has shifted to reflect the increasing acceptance of flexible working arrangements in companies, with the McKinsey Global Institute expecting 20% of the global workforce to permanently shift to remote work for G&A function [1] and 3-5 times a week for more complex departments [2]. This has translated to a different type of traveler - one who is more likely to "work, stay and play" and stay longer at a particular place. Industry reports indicate that 28–30-day stays are the fastest growing category along with group travel both during and post the pandemic [3].

For the above reasons, property amenities essential to mid-size groups and remote work - particularly number of bedrooms, bathrooms, Wi-Fi [3] and access to travel facilities– have become important differentiators for guests.

**1.1.2 Price Sensitivity:**

Post the pandemic, inflation numbers have steadily risen hitting an all-time high in January – the highest since 1982. This is characterized by the cost for living essentials like fuel (+9.5% in Jan, +46.5% y.o.y), shelter (+0.3% in Jan, +4.4% y.o.y) and food (+0.9% in Jan, +7% y.o.y) rising as part of the "rapid cyclical acceleration in inflation in progress amidst an exceptionally tight labor market" [4] that is unlikely to abate anytime soon. As consumer purchasing power is diluted, the pot for discretionary spending lessens. This has important implications for the travel industry and Airbnb given its revenue model.

### 1.2 Revenue Model & Growth Drivers

Airbnb's business is driven by two (2) fee arrangements[5]: (a) host only (14-16% of booking price OR (b) shared (host: 3%, guest: 14.2%). Given its revenue structure, it is in the interest of Airbnb to ***strike a balance between "maximizing property utilization and booking price"***.

This objective closely aligns with that of hosts and their proxies, who want to squeeze out optimum profits from the listing by: (a) charging the best price possible while (b) keeping slack in property use to a minimum. In this sense, helping listing owners is analogous to sustaining the Airbnb growth engine – a key observation that defines the team's problem statement.

### 1.3 Problem Statement

Empowering hosts and listing owners to achieve the delicate balance between property utilization and optimal pricing is the main focus of this paper. In contrast to generic and vague tips such as 'considering the importance of location' or 'researching competitors', the team presents a more comprehensive framework that:

(a) Enables owners to get an **easy price estimate** given a listing's existing meta-features

(b) Facilitates better understanding of which property **features most influence price**

(c) Shows **comparable properties** that are closest a host's own listing

### 1.4 Solution Components

The project has 3 components to address the above:

(a) **NLP analysis of reviews** to extract customer sentiment and common themes associated to these sentiments.

(b) **Regression model** to output recommended pricing based on property features and customer sentiment. Emphasis is placed on explainability such that hosts can understand **which features contribute most to price**.

(c) **Recommender system** to identify top n accommodations that are closest to the host's property for comparison.

By enabling hosts to price and cater to guests' needs better, we create value for guests and increase booking rates. This consequently increases revenues and attracts more users and owners to fuel Airbnb's growth ambitions.

## 2. Data Sourcing and Processing

### 2.1 Data Source

Inside Airbnb[6] was the main data repository used. The site contains publicly available information compiled using various open-source technologies, including D3, Crossfilter, dc.js, Leaflet, Bootstrap, jQuery, Select2, Python, and PostgreSQL. For the scope of this project, the dataset used mainly focused on **London** - the city with the most Airbnb listings. There are 2 key files used:

**(a)** **Detailed listings file**: 66,641 unique listings with over 74 attributes across 5 main information categories - (a) host, (b) location, (c) property, (d) amenities, (e) ratings and reviews – were captured in the file albeit in varying states of usability. Further pre-processing was required before these could be utilized.

**(b)** **User reviews file**: 1,043,004 reviews are included in this data file with meta features on the listing id, reviewer id, date, and the description of the reviews submitted by the users.

In addition, **GoogleMaps API** was used to pull tube station details in London. This information is used downstream to engineer location features.

### 2.2 Exploratory Data Analysis

Initial data exploration provided direction on any required data transformations and analysis downstream. Some key ones are shared below:

- **Missing Values:** Of the original 74 columns, 42 attributes have missing rows. In fact, EDA reveals that all of the rows have at least 1 missing column value. This is a key observation later addressed during data pre-processing stage through the use of imputation.

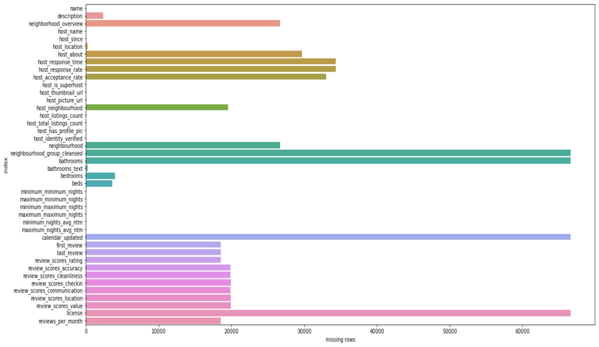

**Figure 1.** Missing values

- **Skewed Price:** As seen in the plot below, the dependent variable price is predominantly right skewed. This will need to be pre-processed before the model build stage

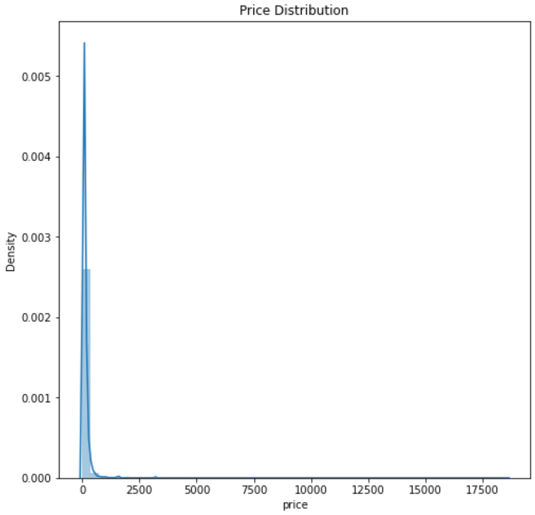

**Figure 2.** Price distribution

- **Properties per Host:** The team also delves into the typical number of properties under a host id. This gives a rough indication on portfolios managed by owners or their proxies. The visualization below shows a simple count plot for this.

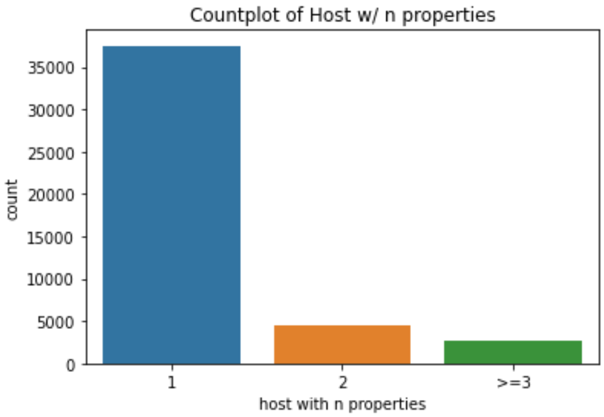

**Figure 3.** Properties per host

- **Old Timer Vs. New Hosts:** Initial analysis was done on the split of new and old hosts on Airbnb. An 'old' host is candidly defined as having been in the platform 1 year prior to the date of dataset i.e., Dec 5, 2020.

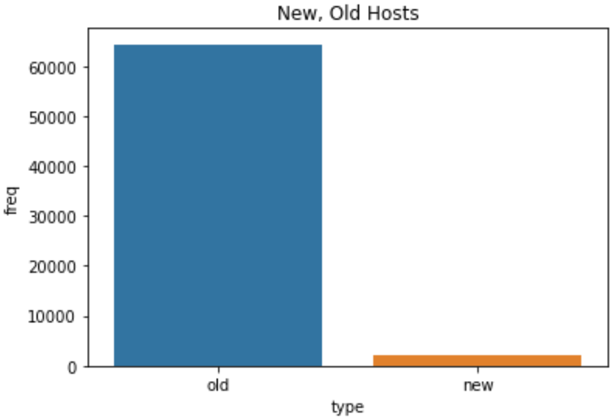

**Figure 4.** New vs Old Host Counts

- **Extreme Outlier Columns:** Boxplots were utilized to check for extreme values. Some notable ones include: (a) bedrooms, (b) minimum nights, (c) availability, (d) review_scores_rating, and (e) reviews per month.

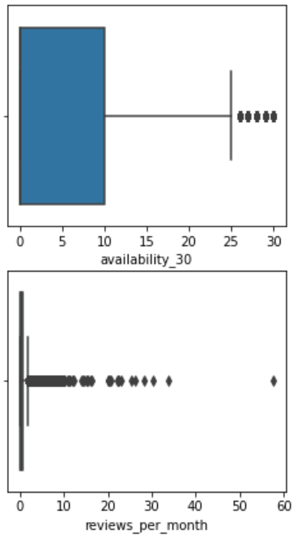

**Figure 5.** Extreme outlier columns

Many of the above insights from data exploration are actioned in the succeeding sections for pre-processsing and feature engineering.

### Data Pre-processing and Cleaning

Aside from trivial cleaning activities like data type broadcasting, further data processing was done on both the listing and reviews dataset. Each transformation is enumerated and discussed as follows:

#### 2.3.1 Listings Data Pre-processing

(a) **Dropping of non-value add columns:** This included attribute columns of the following nature:

a. Columns deemed irrelevant to predicting price (*listing_url, scrape_id, etc.*) or contain only zero values (*calendar_updated, bathrooms*)

b. Columns that present information already contained in other columns (*calendar_updated, bathrooms, beds, etc.)*

**(b) Removal of inactive properties**: Properties listed prior to December 6, 2020 but with number of reviews less than equal to 0.09 (25th percentile) are considered inactive and were dropped from the analysis. The underlying rationale is simple. Said properties are no longer updated and information from reviews or price may unnecessarily influence predictions for a more current listing.

There is a 45% reduction in the number of observations pre and post the transformation with counts 66,641 and 36,705 respectively. This makes intuitive sense as some owners may have stopped accepting guests to mitigate COVID transmission risk.

**(c) Extracting text type amenity features:** Further processing was done for 2 columns 'bathroom text' and 'amenities' to extract key information such as the number of bathrooms, whether or not there was bathroom sharing as well as the full list of amenities available. These were encoded as numerical or boolean values.

**(d) Filling missing values:** All rows have missing values in one or more of 42 columns. For simplicity's sake, missing value were imputed with '0'.

**(e) Winsorization of upper threshold outliers:** Outliers especially on the upper threshold were noted for bedrooms, minimum_nights, maximum nights and minimum_nights_avg_ntm. Data capping was done with max cap set at the 95th percentile.

**(f) Log transformation of price:** Plotting the price distribution, a right skewed plot can be seen. To remedy this issue, log transformation is used as it is more interpretable as percentage change in price vis-à-vis other transformation techniques.

**(g) One Hot Encoding:** All categorical features were one hot encoded in preparation for model build and training.

Post data preprocessing, 36,705 rows of the original 66,641 listings remain. This is mainly due to the inactive listings dropped.

### 2.3.1 Reviews Data Pre-processing

(a) **Dropping of irrelevant reviews:** Dataset is scanned for any comments that are either duplicates or have less than 5 words. These observations are dropped from the dataset.

**(b) Dropping of non-English reviews:** Langdetect is also used to detect the language used for reviews. Reviews that are not in English are dropped as VADER accuracy is questionable when applied to non-English text.

**(c) Lemmatization:** This is done without cleaning as the VADER package is known to work well with all caps, emojis, punctuations and the like.

## 2.3 Feature Engineering

On top of the existing features, additional ones were engineered as they are hypothesized to be relevant when predicting price and grouping similar properties together.

### 2.4.1 Listings: Host and Property

- **Host Duration:** Listing owners and their proxies who have been on the platform longest are believed to have more experience managing properties, hence better able to calibrate their properties to align with changes in user preferences and market price changes.

- **Properties in London:** This is to be distinguished from total properties that the hosts have - one which might be reflected in the platform numbers. This is an intermediary attribute required to derive the succeeding bullet, 'managed_property'.

- **Managed Property:** From total properties the host has in London, those managing more than 2 properties (95th percentile) are considered professionally managed. Listings under such a host id portfolio are labelled accordingly. The hypothesis is that such properties are better calibrated and would be priced optimally vis-à-vis those that are not.

- **Occupancy Rate:** These are derived values from the different property availability metrics, forward looking n days into the future (availability_30, availability_60, etc.). To make these figures more comparable, available days was divided by n future days to compute for the occupancy metric across different time periods. The occ_30D figure was used as proxy for short-term, 60-90D for medium term and 365D for long term. The intent is to see how occupancy affects price and whether or not the relationship may prove to be counterintuitive.

- **Latitude and Longitude Transformation:** Further transformation is done on the coordinate data to better represent 3D space, hence the creation of variables x, y, z which are more representative of actual distances when mapped to a perfect sphere.

### 2.4.2 Listings: Location

- **Nearest Station:** Tube stations closest to the listings are determined, as proximity to such facilities are deemed important for traditional travelers or remote workers needing to commute back to the office when required.
- **Walking Distance:** This attribute indicates the number of stations within 'walking distance' from the listing. 'Walking distance' refers to anything with a 10 minute reach with walking speed of around 1.33 m/s.
- **Station Distance**: This is defined to be the distance to the nearest station in the listing.

### 2.4.3 Reviews: Sentiment Analysis

VADER was used to process comments and derive sentiments behind the review. Net sentiment score was calibrated as follows: (a) **positive** if value > 0.05, (b) **negative** if value <= -0.05 and (c) **neutral** if otherwise.

### 2.4 Feature Selection & Final Feature Set

Overall, there were a total of 114 attributes from both the original and the derived feature set. In order to reduce data dimensionality, feature selection was employed using the **SelectFromModel package** from sklearn. This effectively reduced the feature from 114 to 65. Variable definitions alongside the models that are used in are set out in a data dictionary hosted on the project GitHub page.

## 3. Machine Learning Models

### 3.1 High Level Flow

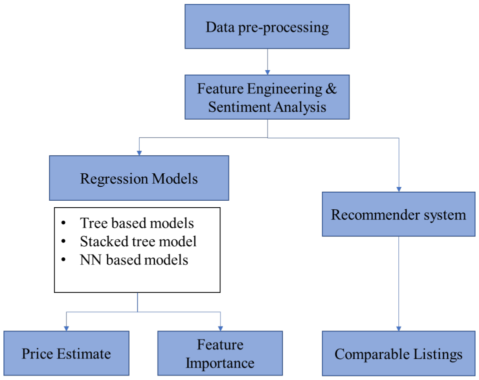

**Figure 6.** High level flow

The above diagram shows how the different use cases flow into one another. The choice of models for both the regression and recommender system are detailed in the succeeding subsections.

### 3.2 Price Regression

Price regression objectives are straightforward – it is nothing more than (a) maximizing cross validated R-squared while (b) giving hosts a better understanding of factors that contribute most to price, (c) both locally and globally. This last portion is equally important with price prediction accuracy since this gives hosts more insight on what property features are most correlated to price.

### 3.2.1 Data Preparation

Before model training, data was split into train and test sets with test size set to 20% of the total data.

A final check was also done to detect data skew. 7 variables (*reviews per month, review score rating, minimum nights, maximum nights, minimum nights average, station distance and walking distance*) were identified and column contents were transformed using PowerTransform's Yeo-Johnson method.

### 3.2.1 Price Prediction Baseline

An untuned OLS was used as baseline, passing in all features from the pre-processed dataset sans any feature selection with reported R2 value of 0.682 and MSE of 191.

The complete measures are shown in Table 1. It is worth noting that the adjusted R2 of 0.682 is exactly the same as the regular R2, with an F-score probability of less than 0.05, indicating significance of the selected features, as well as the statistical significance of the results.

**Table 1.** Baseline OLS results and significance

| | R2 | ADJ. R2 | P(F-STAT) |
|---|---|---|---|
| GLOBAL | 0.682 | 0.682 | 0.00 |

| FEATURE | COEF | P>\|T\| |
|---|---|---|
| ENT. HOME/APT | 95.9 ± 0.2 | 0.00 |
| Z | 0.784 ± 0.021 | 0.00 |
| X | 0.757 ± 0.021 | 0.00 |

| | | |
|---|---|---|
| ENT. HOME/APT | 0.271 ± 0.006 | 0.00 |
| ACCOMMODATES | 0.144 ± 0.005 | 0.00 |
| BATHROOM_SHARED | -0.14 ± 0.005 | 0.00 |
| HOST_TOTAL_LIST. | 0.119 ± 0.003 | 0.00 |
| BEDROOMS | 0.116 ± 0.004 | 0.00 |
| Y | -0.11 ± 0.003 | 0.00 |

### 3.2.2 Model Training

Two families of models were trained for the price regression use case.

(a) Ensemble and boosting tree based methods were chosen for performance and explainability of top features driving said price predictions. Stacking the different base level 0 models are also explored with LGBM as the base level 1 model to see if there is any improvement in performance.

(b) In contrast, neural net methods were selected for their flexibility and ability to model complex, non-linear relationships in large datasets such as the one the team is exploring in this paper.

Specific models explored under each family are tabulated as follows.

**Table 4.** Explored models by category and complexity

| MODEL FAMILY | MODEL NAME |
|---|---|
| TREE BASED | 1.1 LightGBM Regressor<br>1.2 XGB Regressor<br>1.3 RandomForest Regressor |
| STACKING | 1.4 Stacked Regressor |
| NEURAL NET BASED | 2.1 Deep Neural Net<br>2.2 Wide Neural Net |

### 3.2.3 Model Tuning

Hyperparameter tuning was done for each model using a 2-stage approach:

(a) First, by doing RandomizedSearch CV to narrow down the search region for best parameters

(b) Followed by GridSearch CV to determine the final best parameter to use in the model.

### 3.2.4 Model Evaluation and Selection

Each model is retrained using the best parameters from the above exercise and evaluated using 5-fold cross validated R2 and MSE to mitigate overfitting risk. Run results are summarized as follows:

**Table 3.** Model evaluation and selection

| MODEL NAME | TEST R2 |
|---|---|
| 1.1 LightGBM Regressor | 0.7924 |
| 1.2 XGB Regressor | 0.7855 |
| 1.3 RandomForest Regressor | 0.7630 |
| 1.4 Stacked Regressor | 0.7954 |
| 2.1 Deep Neural Net | 0.7478 |
| 2.2 Wide Neural Net | 0.7282 |

From the above, the stacked regressor built on top of the three (3) tree based models, performs the best.

### 3.3 Price Feature Importance

Model agnostic explainability methods were implemented on the best preforming model. This not only facilitates a better understanding of the model, but also shows the user: (a) which features and (b) in what magnitude the said features are responsible for the increase or decrease in predicted listing price.

Overall, the following insights, the variables that most explain price are those that are:

(a) related to available space or number of rooms

(b) describe either the host's experience or status

(c) describe the listing location and availability, booking limitations and amenities available

(d) user ratings or reviews

Drilling down on the methodology, three (3) different levels of explainability were implemented using:

(a) **Permutation Feature Importance** for global absolute importance

(b) **Global SHAP** for direction and magnitude

(c) **Local SHAP** for explainability of specific listings.

### 3.3.1 Permutation Feature Importance

Permutation feature importance calculates the relative importance of a feature based on the decrease in model score when said feature is removed. As such, it provides a global overview of the absolute impact each feature has on model score.

Based on the following chart, it is clear  that the number of people the room accommodates, as well as the type of room (home/apartment) are the most important features in influencing price. Other important features are location (y, x, z), as well as how many listings the host has, indicating a more professional owner.

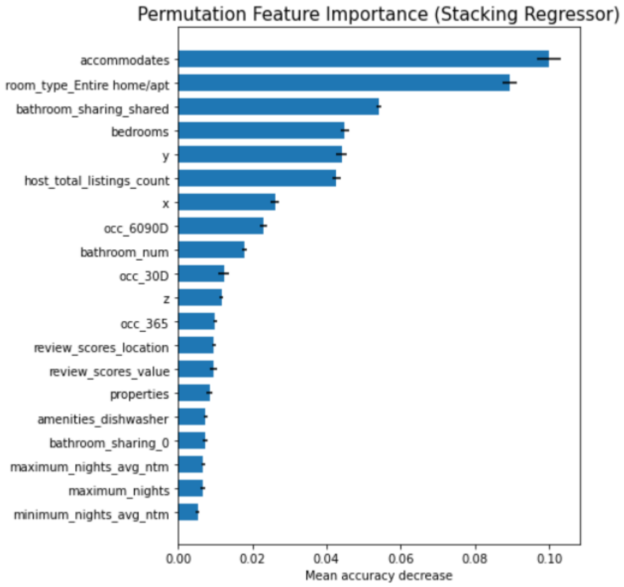

**Figure 7.** Permutation feature importance

### 3.3.2 SHAP Global Explainability & Magnitude

In order to get more insight on not only global absolute impact of each feature, but also direction, magnitude, and their relationship, we implemented the SHAP algorithm. The following graph shows the top 20 features for the regressor model.

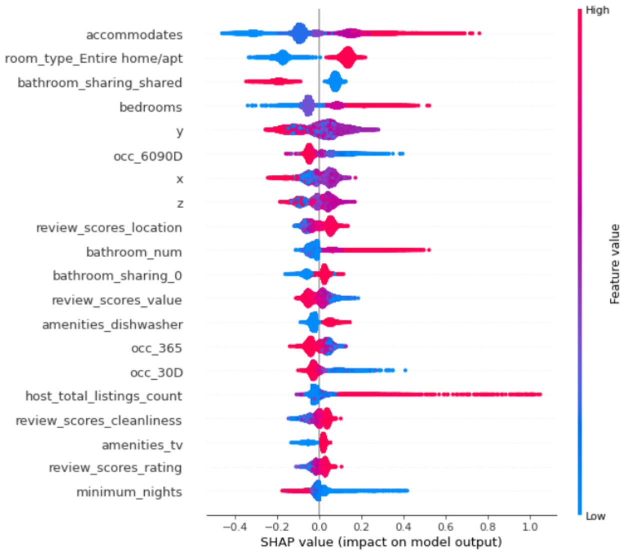

**Figure 8**. Global SHAP explainability

Looking at the graph, we can observe a similar story to PFI in terms of global importance ranking. However, we can gather some additional information.

For example, it is possible to see how having a home increases the price by a larger magnitude than not containing a shared bathroom. It is also possible to observe how a listing that there is a small range in where a listing can accommodate an average number of clients without much drop/gain in price.

### 3.3.3 Local Explainability

As a value added business proposition, SHAP can also be used to give insights to the user on what determines the price of a particular listing. This allows the host access to additional insights insights that may be masked by the global explainability models.

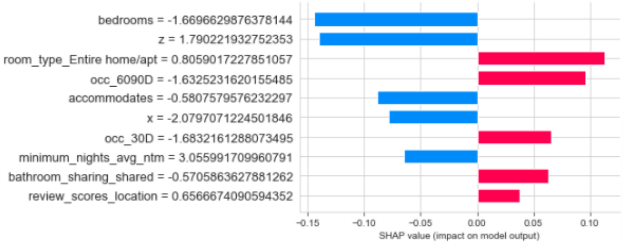

**Figure 9.** Local SHAP explainability

Figure 9 shows the feature importance for a specific listing with its (scaled) feature values. At the same time, it tells the user if this specific feature's value is increasing or decreasing the listing price. Some features may be hard for a user to adjust, such as the number of bathrooms. However, other features, such as the minimum number of nights, could be easily changed.

In this specific example, it is clear that the location and size of the property is lowering overall value. However, location review scores seem to influence the value positively. This may indicate that the property is in a well located but low income neighborhood.

### 3.4 Recommender System

#### 3.4.1 Recommender System Rationale

The bargain mentality is strong in the travel industry as it is in most of the goods in the discretionary spending pool. So aside from price prediction and related feature importance, it is important to benchmark a listed property against listings that are most similar to it. Intent-wise, the recommender model is meant to simulate guests trying to shop around like listings to get best value for money.

#### 3.4.2 Processing Approach

Logic-wise, the recommender system code follows the below flow:

(a) The dataset is filtered to only include features that are likely to be searched for by guests when looking for an Airbnb (e.g., neighborhoods, station_distance, amenities_kitchen, etc.)

(b) Columns which are relevant from a result perspective but used for the feature comparison are dropped (i.e., id, host_id, log_price).

(c) Due to memory constraints, data is processed in 4 chunks and a similarity matrix is generated relative to the host property using cosine similarity. The top 10 from each chunk are added to a shortlist.

(d) From the shortlist of 40 most similar properties, the top 10 are extracted and displayed to the host as the properties to benchmark against.

#### 3.4.3 Recommender Model Evaluation

Given the absence of user interaction data (e.g., clicks) relative to the top10 recommended items, it is not possible to come up with an evaluation metric given the current datasets available. In lieu of this, a future approach to be taken is proposed – the first from the host side, the second from the guest side.

(1) **From host side**: Track host interaction activity with outputted recommendation, such as:

   a. Tracking how many of the recommended properties the host actually clicks through to review

   b. Asking host to rate the quality of reviewed property recommendations based on a 5 star rating scheme.

(2) **From guest side**: From top 10 lists recommended, assess how many of these appear based on guest searches.

## 4. Business Application and Delimitation

More than model accuracy, the ability to integrate outputs back into operations and translate these into tangible business impact is more important. In this section, the team delves into how to operationalize our 2 models in a way that benefits the hosts and by consequence Airbnb.

### 4.1 Business Implementation

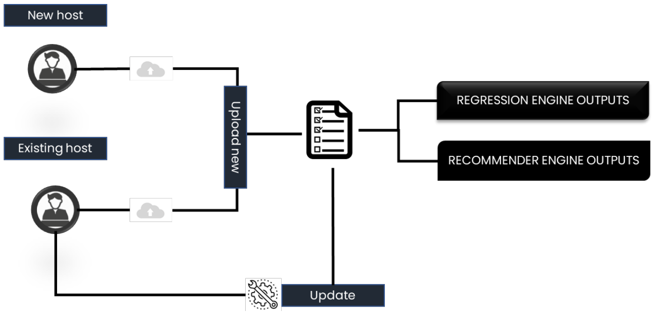

**Figure 10.** Business integration flow

Both new and existing hosts need to ensure that all property meta features are up to date for their listings. These would then be taken by both the regression and recommender engines to produce three (3) types of outputs:

(a) **Listing price recommendations:** Essentially, this will output a market calibrated price that can quickly allow hosts to determine if they under or overcharging relative to the rest of the market given the property's current feature set.

(b) **Feature highlights**: More than price, the regression engine should also be able to output features that are: (1) highly regarded in the overall market (global explainability) and (2) gives hosts an idea around which current features are most responsible for the price the listing is able to command in the market (local explainability). This insight will guide owners on how to best allocate

(c) **Competitor List**: The recommender engine on the other hand, outputs the top 10 most similar properties. This gives the hosts the ability to do some due diligence on potential competitors, benchmark pricing and amenities and take any necessary corrective actions to bring their listings up to speed with the rest of the identified cohort.

### 4.2 Business Impact and Caveats

By equipping hosts with the means for data-driven decision making, listings are almost always guaranteed to auto-calibrate in terms of features and pricing as owners try to get maximum profit by striking the balance between booking rate and optimum list pricing.

Granted, there are a couple caveats to implementation. The team outlines some of the critical ones and potential ways to work around them:

(1) A subset of hosts may prove to be initially resistant towards changes or suggestions. Hence, Airbnb can give out incentives in the early stage to get majority of hosts on the service. This can be discontinued downstream when owners become convinced of its business value.

(2) Time effects are not sufficiently modeled here. For example, price seasonality is not considered due to issues with data availability. In like manner, the influence of features on price may change over time. Hence, users should be given flexibility to run the model as an on demand service to better bridge the gap.

(3) Speaking of user flexibility, the recommender service can be tweaked to allow hosts to easily choose features used for comparison. This will allow flexibility in their analysis especially when choosing which direction to allocate limited renovation or amenity funds into.

With this data driven and grassroots approach, the team is confident that Airbnb can continue to sustain, if not, outperform previous performance and output numbers commensurate to its premium valuation in the market.

## 5. References

[L] Ryan H., Abishek S., Samir S., Edward W. (2021, May 14). Virtually possible: Getting remote work right for G&A functions. Mckinsey.com. Retrieved April 1, 2022, from https://www.mckinsey.com/business-functions/operations/our-insights/virtually-possible-getting-remote-work-right-for-g-and-a-functions

[2] Susan L., Anu M., James M., Sven S. (2020, November 23). What’s next for remote work: An analysis of 2,000 tasks, 800 jobs and nine countries. Mckinsey.com. Retrieved April 1, 2022, from https://www.mckinsey.com/featured-insights/future-of-work/whats-next-for-remote-work-an-analysis-of-2000-tasks-800-jobs-and-nine-countries

[3] Tobin, M. (2021, November 10). Airbnb Adds New Tools in Bet Remote Work Is Here to Stay. Bloomberg.com. Retrieved February 26, 2022, from https://www.bloomberg.com/news/articles/2021-11-09/airbnb-announces-new-tools-in-bet-on-remote-work-boom

[4] Jeff, C. (2022, February 10). Inflation surges 7.5% on an annual basis, even more than expected and highest since 1982. CNBC.com. Retrieved February 11, 2022, from https://www.cnbc.com/2022/02/10/january-2022-cpi-inflation-rises-7point5percent-over-the-past-year-even-more-than-expected.html

[5] Airbnb Help Centre. Airbnb service fees. (n.d.). Airbnb. Retrieved February 12, 2022, from https://www.airbnb.com.sg/help/article/1857/airbnb-service-fees

[6] Inside Airbnb. Adding data to the debate. (n.d.). Inside Airbnb. Retrieved February 12, 2022, from http://insideairbnb.com/get-the-data/