

Price prediction for Airbnb Listings in Bangkok

Bruno Helmechy

02/02/2021

Abstract

This report investigates how to predict the prices at which apartments in Bangkok can be rented out on Airbnb. Below I discuss a number of features taken into account, compare 6 prediction models of varying complexity, & after choosing & re-estimating my final model, I investigate expected performance across property types, the number of people accommodated, while investigating most important accommodation features. Viewers can access all project artifacts [here](#). Based on 5-fold cross validation, I found the expanded Random Forest model predicting log-transformed prices to perform best in terms of all metrics. After re-estimation, the final model boasted 47.2% R-squared, & performed with 17.8 dollar MAE & 32.6 dollar RMSE, ca. 65% of the average price during out-of-sample testing. Finally, model diagnostics showed the model to perform fairly consistently given accommodation size (2-6 people), neighbourhoods, & property types, however in each case, the model performs better for the most frequent sub-groups, i.e. for rentals accommodating 2-3 people (25.5-27.7 RMSE), apartments (26.7 RMSE), or the 2 most frequent neighbourhoods, Khlong Toei & Vadhana (0.55-0.57 Normalized RMSE).

Data, Cleaning & Feature Engineering: The raw dataset was downloaded from- & is available at AirBnB-s website (from [here](#)), & comprises over 19.7K observations & 74 Variables, summarizing all accommodations in the Bangkok area available for rental, as of 23rd-24th December, 2020. After keeping only apartments-, & condominiums hosting between 2-6 people, 9962 observations remained to clean for analysis. To manage this feature space, i.e. to keep what's important & drop what is not, while looking to maintain intuition throughout the process, I grouped variables into 7 subjects (excluding ID variables): **Host** information, **Geo-Spatial**-, & **Property** information, **Sales**-, & **Availability**-related data, & **Reviews**-, & **Listings** data.

Some of these groups proved rather not useful, e.g. from **Geo-spatial** variable group only the **neighbourhood_cleansed** variable is retained, as latitudinal-, & longitudinal metrics are indirectly already captured by the neighbourhood they fall into, while the raw **neighbourhood** variable held 500+ values, also posing a linguistic challenge. Similarly, **Listings**-group variables were in essence all dropped, as they contain information regarding listings of different room/apartment types, or contain information already present in **Host**-related variables. My substantive rationale for that is **a)** the number of listings attributed to a host above all are a proxy for the likelihood of him/her being a scammer, rather than an attribute impacting rentals' pricing, & **b)** even on AirBnb, this information is presented only as a host having either 1, or multiple listings online. As such, seeing as prices are rather market-, then inventory-based, a binary variable showing whether the host has only 1 listing seems sufficient. To be safe, the integer variable denoting how many listings a host has, is also kept, so a LASSO model could capture if either is less important by dropping it.

Speaking of **Host**-variables, I view their primary purpose is indicating whether a listing is a scam or not, as these are what both Airbnb & prospective guests use this information for. Thus, 6 of 8 Host-variables are logical / binary variables, while host verifications are a list of credential details (similarly to amenities being a list of objects in the rental property). Any single 1 of these credential details is not important, say whether a host has an email address or Facebook profile, however the total number of credentials may indicate hosts' credibility, & professionalism, so I simply counted the **number of verifications** a host provided.

Guest Reviews-related variables: As proxy for a hosts' experience & recent activity, Nr. of days since their 1st of last reviews were calculated. Ca. 34% NAs, for hosts without reviews - seen from the same observations also missing review scores, & Nr. of reviews per month. In terms of review ratings, only the overall review scores are kept, being an aggregation of sub-scores, & for Nr. of days since 1st & last reviews, instead of the median-, the maximum values were imputed, signifying the quasi-infinite number of days the host has had a review. To impute NAs for review scores, I took the arithmetic mean instead of the median. Though the mean is less robust to extreme values, there is some level of uncertainty arising in guests' minds, if seeing there are no reviews what so ever for a listing. I intend to take this into account by using the mean.

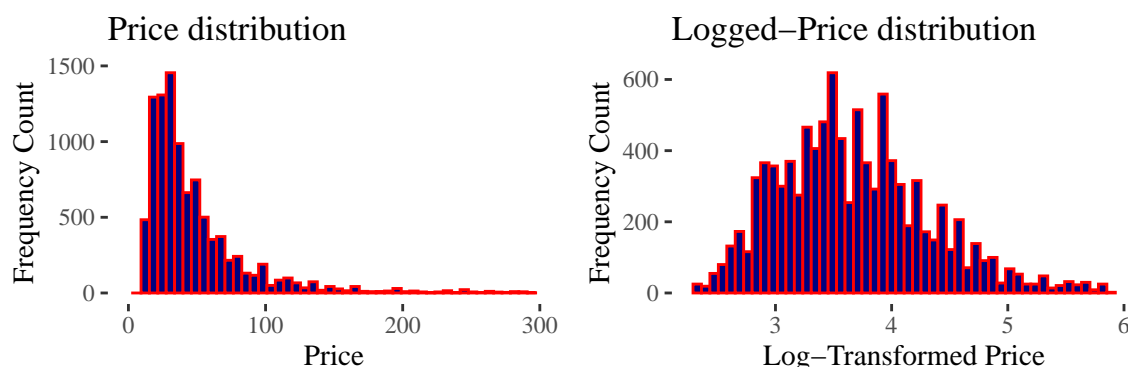
Sales & Availability-related variables are what hoteliers would consider most obviously relating to prices, & besides price, also include minimum-maximum stay-restrictions, as well as the number of available room nights in the coming periods of different length. I argue the number of room nights a host has sold more closely relates to what price he asks for, rather than the number of room nights he has available in the future, given the confidence it may instill. So, I transformed these to denote the number of rooms sold in future periods of different lengths, & eventually reduced these to factor variables. With stay-restrictions, I 1st wrote a function to calculate how identical these columns are, allowing me eventually to retain only minimum stay nights, as maximum nights contained many extreme-, & missing values.

Property Variables The final, & perhaps most important feature engineering decision is regarding **Property**-related variables, most strongly contributing the dropping observations, & 48 dummy variables standing for amenities. To obtain data only for apartments, I 1st filtered the `room_type` variable to only include entire homes or apartments, & also removed listings where the property type was described as e.g. Room, Castle, hostel, or treehouse & other properties of similarly different nature. This still left some non-trivial property types, like Bungalows, Guest Suites, Lofts & Villas, but I view these as legitimate alternatives in potential guests' minds, whereas I hazard, the type of guest looking to book a Castle or a Treehouse is materially different then someone looking to stay in apartments. Once filtered for these property types, listings priced above 350 USD were also dropped, resulting in the cleaned dataset holding 9866 observations.

The greatest cleaning challenge however, is of handling the **amenities** column. After creating a dummy-variable table of all possible amenities, I wrote 3 functions to handle this variable: 1st to coerce columns whose names match certain sets, & combinations of keywords; 2nd to remove duplicate columns, & 3rd to check the frequency of Yes-es in the remaining dummy columns, latter looking to remove amenities with imbalanced values, by myself defined with having less then 1% Yes values. All this resulted in reducing the number of amenities dummies from 262 to 48.

Modelling

Eventually, I grouped my variables into **Property**-, **Host**-, **Reviews**-, & **Amenities**-related, seeing these as the varying degrees, & order of additional information a guest might be looking for, to decide whether to stay at a given listed rental or not. For OLS & LASSO models, I also incorporated interactions between key property-related variables & all amenities, & predicted log-transformed price in 5 of 6 models, seeing how skewed it is. Please see the finalized variable groups, & derived predictor groupings in bulleted lists below:



- **Property-related:** Property Type, Nr. of Guests accommodated (level, log-,level squared, log-squared), Nr. of Beds-, Bedrooms & Bathrooms, Neighbourhood Cleansed, Has Availability, Nr. of Rooms Sold in coming 30-, 60-, 90-, & 365 days.
- **Host-related:** Host is owner, Nr. of hosts' listings & verifications, & whether he greets you, accept every reservation, has a profile pic, & his/her identity is verified.
- **Reviews-related:** Nr. of Reviews (Total & per month), Average Review score, Nr. of days since 1st & last reviews (Level- & Log form), & Flag-variables for missing observations on review per month, score rating, & Nr. of days since last review variables.
- **Amenities-related:** 48 dummy variables indicating whether a collection of amenities are present.
- **Interactions:** I hazard the presence of certain amenities affects rental prices to varying degrees depending on property characteristics, e.g. though wifi is a given necessity in any context, it's perhaps even more important in larger accommodations to keep children at peace. Thus, all 48 amenities were interacted with core property features, Nr. of beds / bedrooms / bathrooms & people accommodated.
- **Predictor Group 1:** Property Variable group
- **Predictor Group 2:** Predictor Group 1 + Host-, & Reviews variable groups
- **Predictor Group 3:** Predictor Group 2 + Amenities
- **Predictor Group E:** Predictor Group 3 + Interactions

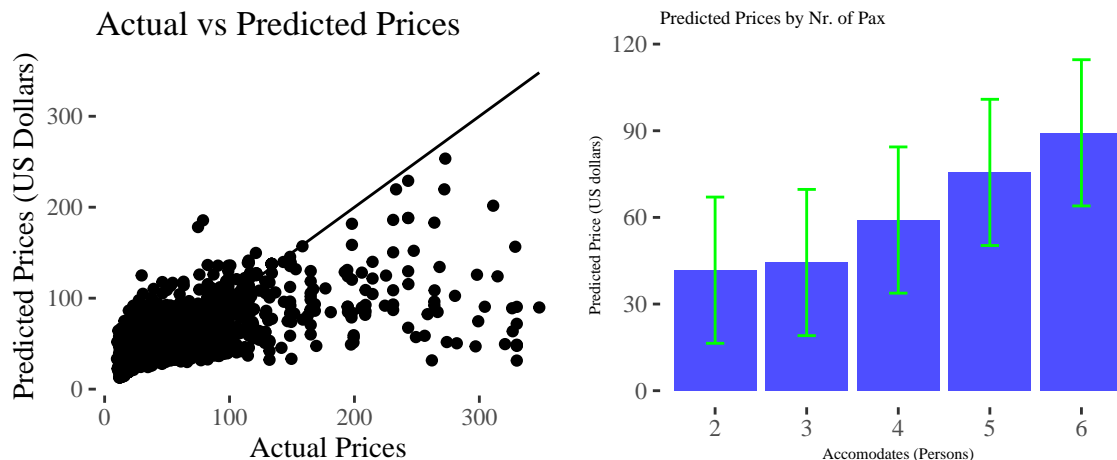
Finally, please see the created Predictor groups above. In the 4 best models summarized below, LASSO utilizes Predictor Group E, the simplified Random Forest (rf_model_1) used Predictor Group 2, & the rest Predictor Group 3. Due to prices' strong skewness shown earlier, all models predict log-prices, which were subsequently transformed, to obtain price predictions & calculate MAE & RMSE.

models	Rsqr	MAE	RMSE	RMSE_norm
ols_model	0.437	19.035	33.7	0.678
lasso_model	0.478	17.958	32.227	0.648
rf_model_1	0.528	10.169	22.346	0.449
rf_model_2	0.56	8.875	20.194	0.406
rf_model_2_level	0.46	18.416	31.902	0.642

As can be seen above, the extended Random Forest model (rf_model_2) beat all other candidates, with 5-fold cross-validated performance of R-squared of 56%, & a RMSE of 20.2 USD, ca. 40% of the average listing price. As such, I recommend the extended Random Forest to estimate the clients' future listings' profitability. What's left is re-estimating the chosen model on the complete training dataset & test expected future performance ex-sample, utilizing data reserved only for this stage, 30% of the cleaned dataset. As a sanity check, I also ran the level-price predicting model:

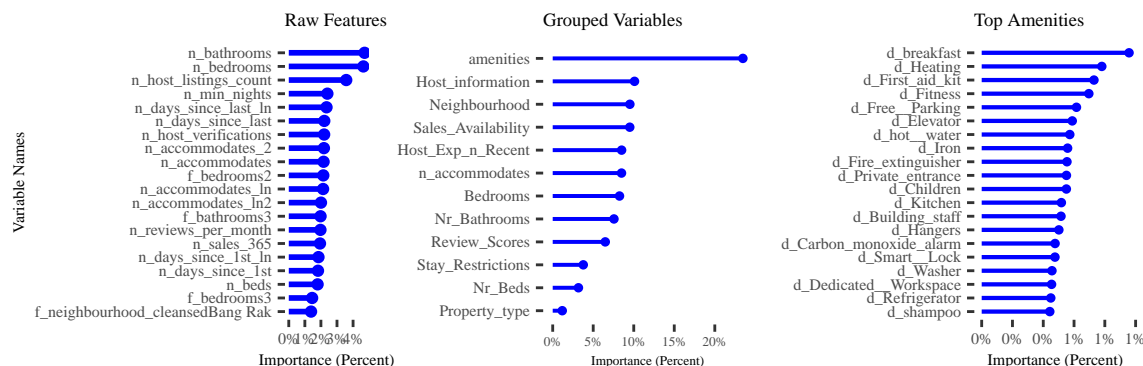
modelname	Rsqr	MAE	RMSE	RMSE_norm
Final Full Random Forest Model	0.561	17.3	33.2	0.67
Full Random Forest w Level Price	0.472	17.8	32.6	0.656

Model Implications



Please see the re-estimated random forest models' ex-sample performance summary statistics above. Visibly, the level-price prediction model boasts higher MAE, yet lower RMSE meaning, though it is less right on average, it is also less frequently very wrong. For this reason, as well as due to improving ex-sample, i.e. not being overfitted, my final recommendation is the level-price expanded Random Forest model. Please see its performance, plotting predicted prices scattered around the line of actual prices. Visibly, the chosen model tends to predict right on average up to ca. 100 USD, however is under predicting prices above 125 USD on every occasion. From a business perspective, this means the chosen model can confidently estimate price ranges' lower bounds, more useful to evaluate investment decisions' profitability.

Variable Importance



Above, I'd like to point out the most important features for the chosen model, & as such, to determining an Airbnb listings' price. The 3 plots above 1st show raw features, where one can see Nr. of bedrooms & bathrooms in a rental, as well as how many other apartments does a host have are most important, each contributing over 4% to the model. Apartments have 1000s of features however, so the 2nd plot groups all features used into 12 aspects, aiming to recommend prospective hosts what aspects to focus on: 1st it's visible that amenities are overwhelmingly the most important aspect of pricing according to our model (20+ %), while Hosts' trustworthiness & the location are 2nd & 3rd (ca. 10% each). Interestingly, the type of property in question has very little impact. Finally, as Amenities are the most important aspect to pricing, a top 20 list comprises the 3rd plot. Apparently, after the most important meal of the day, 2 safety / security-related amenities are most important, while fitness & parking possibilities round out the top 5.

Var.1	RMSE	Mean.price	RMSE.price
Apartment size			
2	27.7	42	0.66
3	25.5	44	0.58
4	42.8	60	0.71
5	43.2	72.3	0.6
6	40.6	85.2	0.48
Type			
apartment	26.8	48.3	0.55
condominium	33.2	46.4	0.72
Neighbourhood			
Khlong Toei	24.9	43.8	0.57
Vadhana	35	63.3	0.55
Huai Khwang	24.2	37.7	0.64
Ratchathewi	39	54.9	0.71
Sathon	21.8	48.1	0.45
All	32.6	49.6	0.66

Partial Dependencies: On a final note, I'd like to discuss how the recommended model performs across perhaps the 3 most obviously important property features: Nr. of People accommodated, Apartment type & Neighborhoods. As can be seen, there is a distinct jump in model performance above 3 people accommodated, to 40+ USD RMSE vs 25.5-27.7 for 2-3 people accommodated. Apartments are also distinctly performing better vs condominiums, possibly due to their less varied nature. Finally, the most hardly comparable feature, yet the most important since Conrad Hilton tossed the phrase, viewer can see the model performs significantly better the overall in 3 of the top5 neighbourhoods in the dataset, of which Vadhana seems to be a rather expensive area, meaning the model performs better than average where it seems more important.

Conclusion

This report investigated how to best predict Airbnb listings' prices in Bangkok. 6 models were tested, from which the expanded Random Forest model predicting level prices emerged as best, with ex-sample performance of 17.80 USD Mean Absolute Error, & 32.60 USD Root Mean Square Error (ca. 66% of the Average rental Price). Though the model consistently under predicts prices above 125 USD, it distinctly outperforms the benchmark OLS model. The chosen Random Forest model also outlined a number of key features prospective hosts should focus on satisfying to the best of their abilities. Amenities seem to be overwhelmingly the most important, followed by the quality & quantity of information the host has provided, while the specific property type seems to be only minimally important. On a final note, readers may also see that the proposed model performs better than average where it matters most, for accommodations hosting 2-3 people, for traditional apartments, & for the 2 most frequented neighbourhoods in Bangkok.