Principles of Data Science Coursework

Airbnb listings in London: Analysis of key characteristics driving revenue

Author: Miguel Bravo
Computational Notebook:
https://smcse.city.ac.uk/student/aczd014/MiguelBravoComputationalNotebook.html

Analysis Domain, Questions, Plan

Overview

Airbnb is an online platform connecting people wanting to rent out their properties with people seeking short-term accommodation (Parliament. House of Commons, 2018).

In London – as elsewhere in the world – Airbnb has seen extraordinary growth; from 2011 to 2015 the number of entire homes listed in London grew by over 4,000% (Parliament. House of Commons, 2018). This increase speaks to the clear benefits of Airbnb for hosts in London. Airbnb claims a typical host in London can expect to generate annual earnings of £3,000 (Airbnb Citizen, 2017).

However, with more listings comes increased competition. Research suggests that supply in Airbnb listings outstripped demand from February to June 2016 (RSA, 2016), indicating the London market is becoming oversaturated. As competition intensifies, there is a mounting need for hosts to configure their listings more effectively, to outcompete peers – a real challenge given the many complex factors driving a listing's performance.

This project seeks to take a data driven approach in addressing this challenge. By analysing the relationship between a listing's characteristics and its revenue relative to peers, this project aims to shed light on steps any host in London can take to outcompete peers.

Primary Goal

This project looks at Airbnb listings in London, analysing the relationship between their **'host-controllable' characteristics** and their **monthly revenue differential from their peer group average**. The goal is to explore which characteristics drive above average revenues – to inform practical changes a host can make to their listing to outcompete peers.

Terminology

- 'Host-controllable' characteristics of a listing are those a typical host is able to influence directly.
- Monthly revenue is estimated using the 'San Francisco' Occupancy Model devised by insideairbnb.com (explained in analysis notebook)
- Peer groups are all listings of the same room type, with the same number of bedrooms and bathrooms, and in the same neighbourhood.

Dataset

The dataset used is a snapshot of all London listings on Airbnb as at 6 October 2018 – sourced from insideairbnb.com.

Analysis Objectives

Milestones to achieve primary goal:

1. Perform exploratory data analysis to understand structure of data and inform further analysis
2. Select and train appropriate models to estimate a listing's revenue relative to peers based on its characteristics.
3. Examine most important characteristics driving revenue performance, as estimated by each model
4. Interpret practical insights arising from these estimates on ways a host can improve their listing's revenue performance.

Analysis Strategy

Steps to achieve analysis objectives:

1. **Data Preprocessing** – *preparing data for analysis*
   - Derive target and features
   - Deal with outliers and missing values
   - Select features with basic filtering methods

2. **Exploratory Data Analysis (EDA)** – *investigating structure of data*
   - Visualise feature distributions, and relationship with target
   - Make final tweaks to data from insights gleaned

3. **Modelling** – *building models to capture drivers of revenue*
   - Select, train and evaluate models
   - Simplify models through additional feature selection
   - Interpret models by exploring feature importance

4. **Reflection** – *summarising findings*
   - Summarise key findings of analysis
   - Reflect on whether analysis objectives have been met

Findings and Reflections

Before entering into the findings, it is important to define the variables used in the analysis (see figure 1).

Variable	Description	Category	Data Type
revenue_var	**Estimated monthly revenue differential from peer group average**	**Target**	**Float**
accommodates	Maximum number of guests allowed to stay (as defined by the host)	Feature	Integer
amenities_hair_dryer	Hair dryer available	Feature	Binary (1=True, 0=False)
amenities_hangers	Hangers available	Feature	Binary (1=True, 0=False)
amenities_hot_water	Hot water available	Feature	Binary (1=True, 0=False)
amenities_lockbox	Lockbox available	Feature	Binary (1=True, 0=False)
amenities_long_term_stays_allowed	Long term stays allowed	Feature	Binary (1=True, 0=False)
amenities_parking	Parking available	Feature	Binary (1=True, 0=False)
amenities_self_check-in	Self check-in available	Feature	Binary (1=True, 0=False)
availability_30	Booking availability over next 30 days	Feature	Integer
availability_365	Booking availability over next 365 days	Feature	Integer
bed_type_airbed	Bed is an air bed	Feature	Binary (1=True, 0=False)
bed_type_real_bed	Bed is a real bed	Feature	Binary (1=True, 0=False)
beds	Number of beds available	Feature	Integer
calculated_host_listings_count	Host total number of listings	Feature	Integer
cancellation_policy_flexible	Full booking refund within limited period	Feature	Binary (1=True, 0=False)
cancellation_policy_moderate	Full booking refund within limited period (more strict)	Feature	Binary (1=True, 0=False)
cancellation_policy_strict_14_with_grace_period	Full refund if cancellation made within 48 hours of booking	Feature	Binary (1=True, 0=False)
cancellation_policy_strict/super_strict	50% refund if cancellation made prior to 30/60 days of check in	Feature	Binary (1=True, 0=False)
cleaning_fee	Cleaning fee	Feature	Float
days_as_host	Number of days since host registered on airbnb	Feature	Integer
days_from_host_to_first_review	Number of days between host first registering, and first review received	Feature	Integer
days_since_first_review	Number of days since listing received first review	Feature	Integer
days_since_last_review	Number of days since listing received last review	Feature	Integer
experiences_offered_business	Business experience offered	Feature	Binary (1=True, 0=False)
experiences_offered_family	Family experience offered	Feature	Binary (1=True, 0=False)
experiences_offered_none	No particular experience offered	Feature	Binary (1=True, 0=False)
experiences_offered_romantic	Romantic experience offered	Feature	Binary (1=True, 0=False)
experiences_offered_social	Social experience offered	Feature	Binary (1=True, 0=False)
extra_people	Cost of adding an extra person to the booking	Feature	Float
guests_included	Number of guests included in the booking	Feature	Integer
host_has_profile_pic	Host has profile picture	Feature	Binary (1=True, 0=False)
host_identity_verified	Host has identity verified	Feature	Binary (1=True, 0=False)
host_is_superhost	Host is a super host	Feature	Binary (1=True, 0=False)
host_name_included	Host name is included	Feature	Binary (1=True, 0=False)
host_response_rate	Percentage of messages which the host has responded to	Feature	Float
host_response_time_a_few_days_or_more	Host on average responds to a message after a few days or more	Feature	Binary (1=True, 0=False)
host_response_time_within_a_day	Host on average responds to a message within a day	Feature	Binary (1=True, 0=False)
host_response_time_within_a_few_hours	Host on average responds within a few hours	Feature	Binary (1=True, 0=False)
host_response_time_within_an_hour	Host on average responds within an hour	Feature	Binary (1=True, 0=False)
instant_bookable	Instant booking available	Feature	Binary (1=True, 0=False)
maximum_nights	Maximum number of nights for a booking	Feature	Integer
minimum_nights	Minimum number of nights for a booking	Feature	Integer
number_of_reviews	Total number of reviews received	Feature	Integer
price	Price per night	Feature	Float
property_type_apartment	Property is an apartment	Feature	Binary (1=True, 0=False)
property_type_other	Property is another type (not house)	Feature	Binary (1=True, 0=False)
require_guest_phone_verification	Guest phone verification required to make booking	Feature	Binary (1=True, 0=False)
require_guest_profile_picture	Guest profile picture required to make booking	Feature	Binary (1=True, 0=False)
requires_license	Listing requires licence	Feature	Binary (1=True, 0=False)
review_scores_accuracy	Median review score accuracy	Feature	Integer
review_scores_checkin	Review check-in score (aggregated)	Feature	Integer
review_scores_cleanliness	Review cleanliness score (aggregated)	Feature	Integer
review_scores_communication	Review communication score (aggregated)	Feature	Integer
review_scores_location	Review location score (aggregated)	Feature	Integer
review_scores_rating	Review score rating (aggregated)	Feature	Integer
review_scores_value	Review score value (aggregated)	Feature	Integer
reviews_per_month	Average number of reviews per month	Feature	Float
security_deposit	Security deposit amount	Feature	Float

Figure 1: Variable definitions table

The first noteworthy insight relates to the distribution of the target variable. After removing outliers it was found to have an approximately normal distribution.

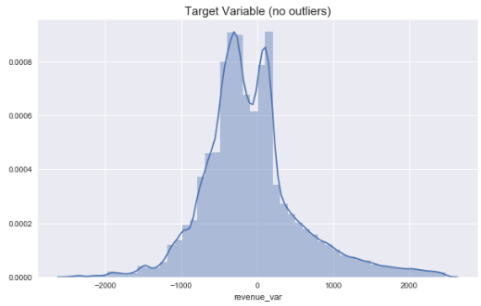

Figure 2: Histogram plot of target variable

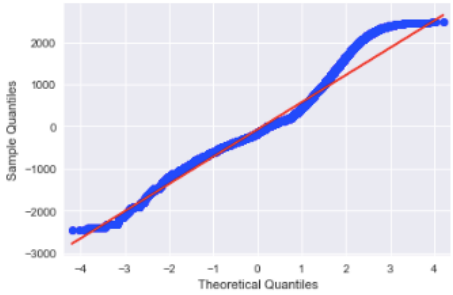

Figure 3: Q-Q plot of target variable

This suggests the variable is reliable and free from sampling bias – adding credibility to the analysis.

The next key finding was that the features and target were not very strongly correlated. Visual exploration of each feature versus the target failed to reveal many clear signals.

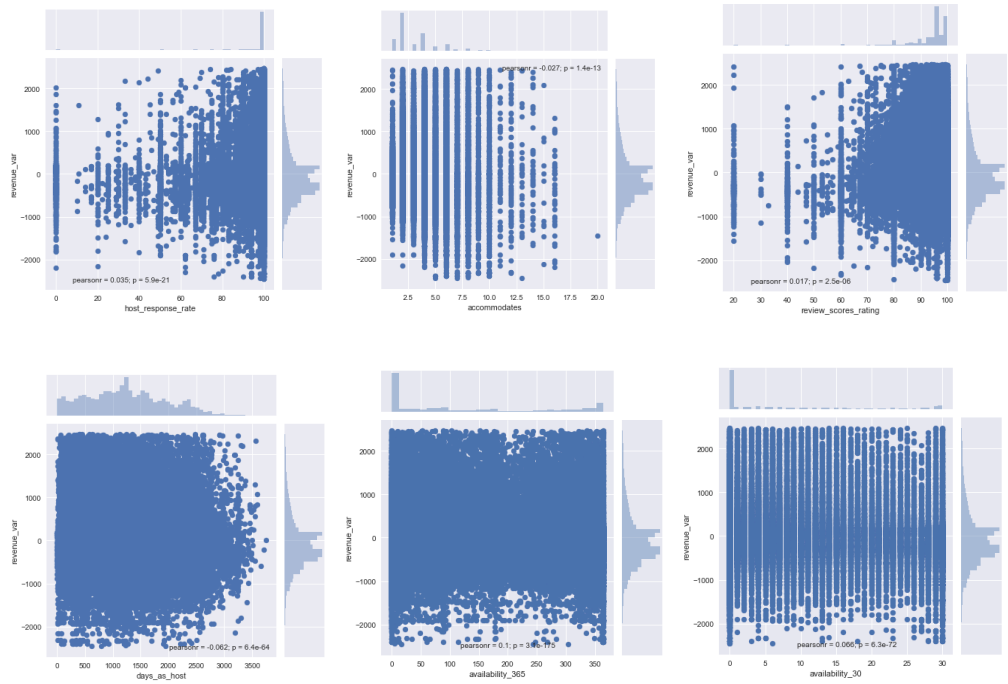

Figure 4: Sample of scatter plots between features and target

This informed the following modelling strategy:

- Avoid regression models – unlikely to work well with data
- Set up a binary classification problem – predicting whether a listing's revenue will over or under earn versus their peer average. Likely to offer best chance of leveraging weak signals
- Train one parametric and one non-parametric classifier and investigate different insights offered by each model.

Modelling

Logistic Regression Classifier (parametric)

The model was trained multiple times and performed best using a subset of 20 features (see figure 5) – offering similar accuracy compared with using the full feature space, while also performing well under cross validation (see figure 6) relative to other feature sets.

```
feature count:  20

['days_since_last_review', 'review_scores_accuracy', 'days_from_host_to_first_review', 'review_scores_rating', 'revi
ew_scores_cleanliness', 'accommodates', 'calculated_host_listings_count', 'review_scores_checkin', 'review_scores_lo
cation', 'guests_included', 'beds', 'cleaning_fee', 'availability_365', 'days_as_host', 'review_scores_communication
', 'days_since_first_review', 'security_deposit', 'price', 'availability_30', 'review_scores_value']

              precision    recall  f1-score   support

           0       0.79      0.73      0.76     14215
           1       0.64      0.71      0.68      9827

avg / total       0.73      0.72      0.72     24042
```

Figure 5: Classification report for 20-feature Logistic Regression Model

```
accuracy score for each run:  [ 0.67764512  0.73757892  0.75387783  0.73891558  0.70665752  0.72326699
  0.71489362  0.72354152  0.66794784  0.57597804]

average accuracy overall:  0.702030296923

accuracy variance:  0.00240584244442
```

Figure 6: Result of 10-fold cross validation on 20-feature Logistic Regression model

These 20 features were selected through Principal Component Analysis (PCA), and were features contributing most heavily to the first 3 principal components in the data, based on their factor loadings (see figure 7).

They therefore captured a significant amount of the data's variance, in particular the key signals leveraged by the model for classification – given the similar performance achieved compared to using the entire feature space.

What are these signals? The model's feature coefficients provide some insight (see figure 8). In logistic regression, a feature's coefficient indicates its impact on the likelihood of success as the feature increases (all else held equal).

The top 3 features with the highest coefficients are: 'cleanliness review score', 'location review score', and 'number of guests included'. This can be interpreted in plain English as offering the following recommendation:

A listing will see the highest rise in its chances of 'out-earning' peers by: 1) improving its perceived cleanliness; 2) improving perceptions of its location; and 3) increasing the minimum number of guests allowed per booking

This example illustrates the useful insights this model can provide to help hosts improve their listing's revenue[1]. Importantly, these are insights a host can take action on, since they are based on 'host-controllable' characteristics.

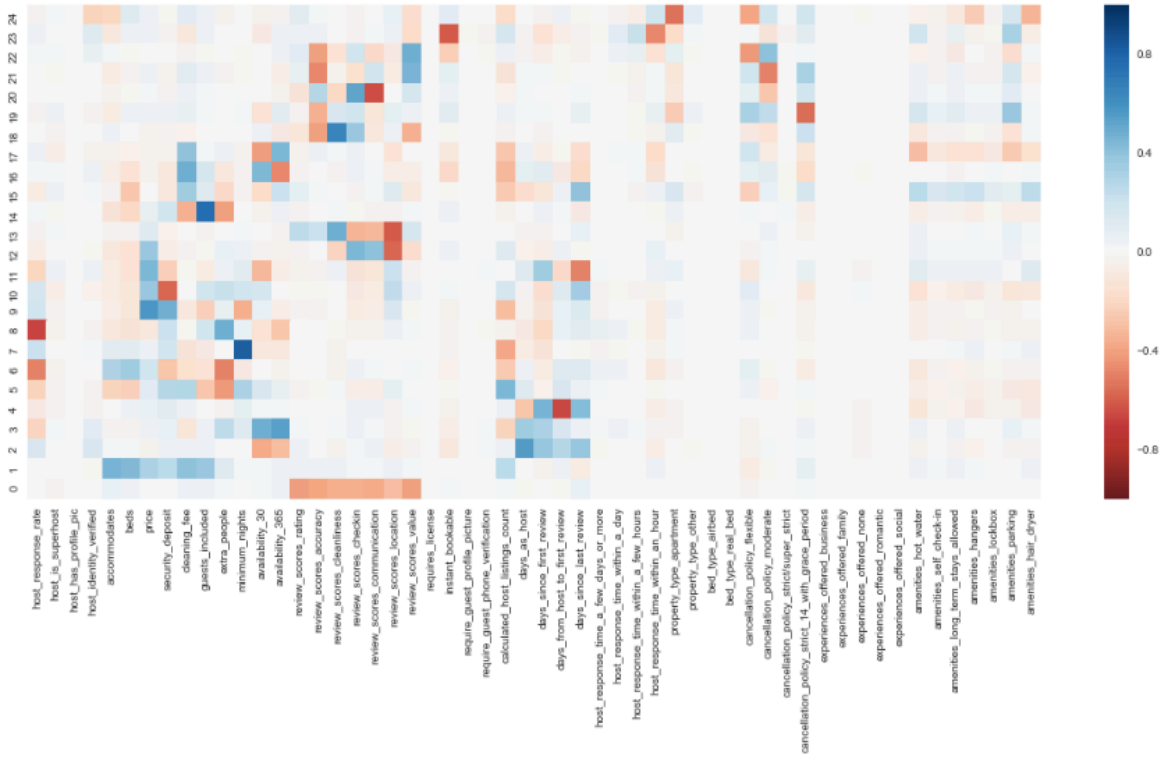

Figure 7: Heat Map showing feature loadings for each principal component

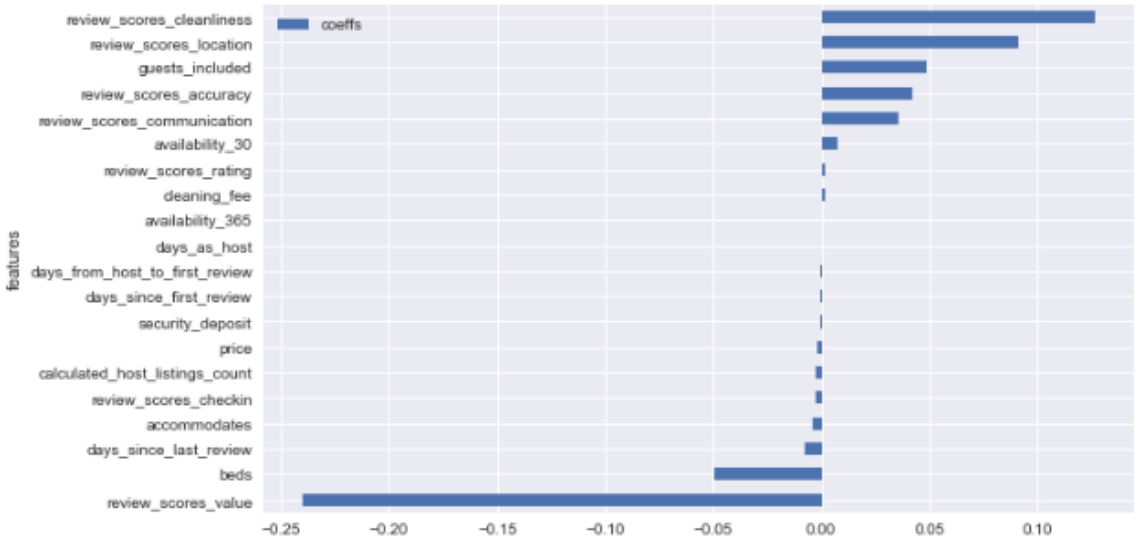

Figure 8: Logistic Regression feature coefficients

Random Forest Classifier (non-parametric)

[1] Assuming the model is correct. Logistic regression makes linearity assumptions which do not appear to hold for this dataset – casting doubt on the reliability of the results. This would need to be investigated further.

This model was trained as per the same approach as the logistic regression model. After a number of iterations, performance was eventually optimised by a subset of 23 features (see figure 9) – accounting for the first 6 principal components in PCA.

```
feature count:  23

['days_since_last_review', 'days_from_host_to_first_review', 'review_scores_rating', 'beds', 'extra_people', 'days_a
s_host', 'price', 'review_scores_value', 'calculated_host_listings_count', 'review_scores_cleanliness', 'review_scor
es_checkin', 'review_scores_accuracy', 'host_response_rate', 'minimum_nights', 'days_since_first_review', 'review_sc
ores_location', 'guests_included', 'cleaning_fee', 'availability_365', 'review_scores_communication', 'security_depo
sit', 'availability_30', 'accommodates']

              precision    recall  f1-score   support

           0       0.84      0.85      0.85     14240
           1       0.78      0.76      0.77      9795

avg / total       0.81      0.82      0.81     24035
```

Figure 9: Classification report for 23-feature random forest model

```
accuracy score for each run:  [ 0.69439868  0.79338276  0.77869303  0.76029654  0.72964438  0.73870658
  0.75381024  0.78648908  0.77481807  0.81049162]

average accuracy overall:  0.762073097908

accuracy variance:  0.00106083187243
```

Figure 10: Cross validation results for 23-feature random forest model

Under cross validation this feature set also came out on top in average accuracy and accuracy variance, compared to other candidates.

The signals picked up by this model can be interpreted by looking at the model's feature importances (see figure 11). This is a feature's relative contribution towards the model's overall ability to discriminate between target classes.

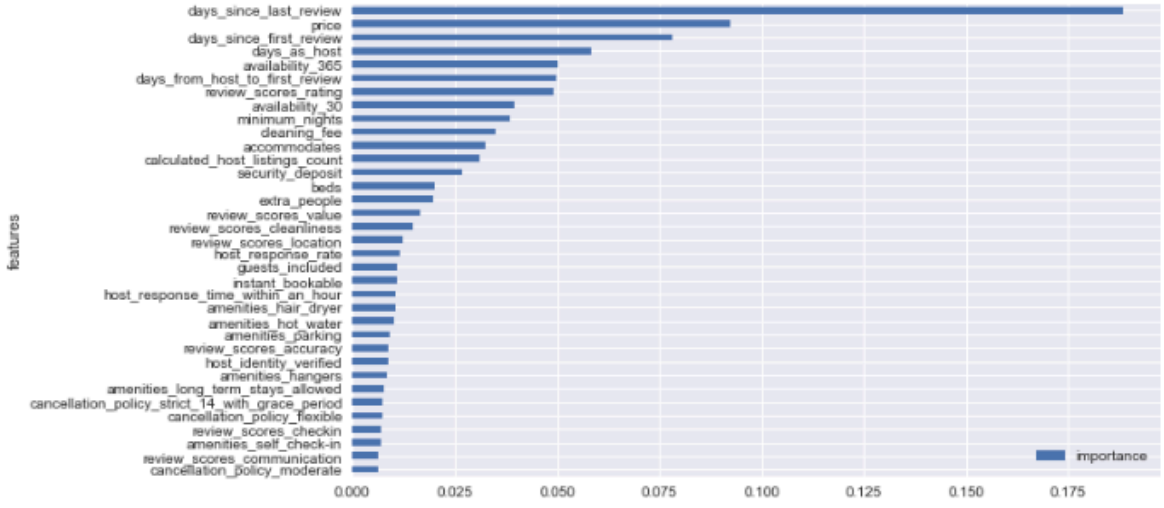

Figure 11: Random forest feature importances

The most important features were 'days since last review', followed by 'price', and 'days since first review' – i.e. these were the characteristics which helped the model the most in distinguishing between classes.

Unlike logistic regression, it is not possible to infer anything more specific about the impact of these features on the target. Further visual investigation is required to help answer this (see figure 12).

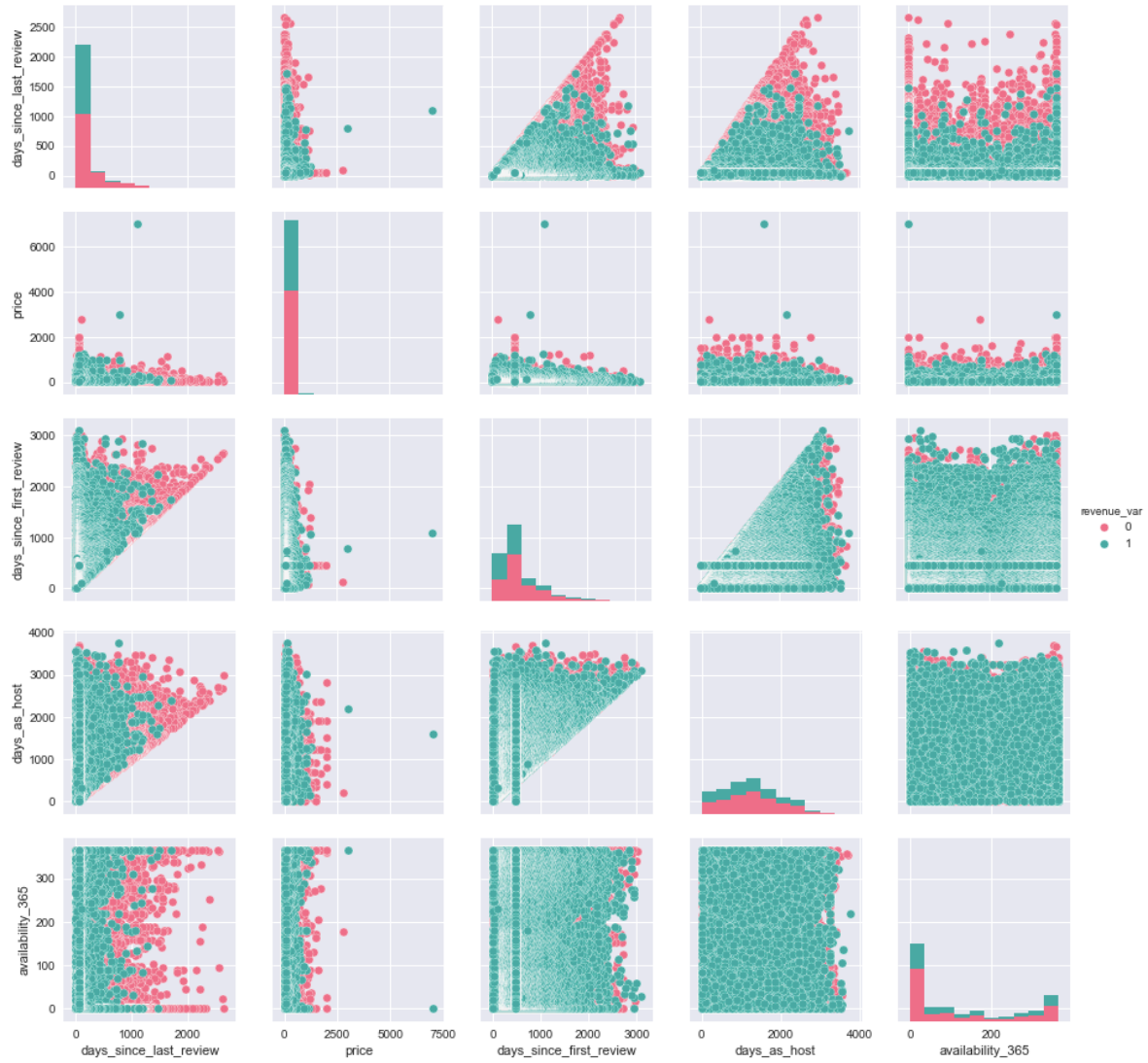

Figure 12: Scatter plot grid of random forest most important features

This visual clearly highlights the discriminative power of 'days since last review', explaining why it was assigned top importance by the model. Listings receiving their last review over 3 years ago (quite understandably) mostly tend to be 'under-earners'.

Price also acts as a pretty good discriminator. Listings priced above the £1,000 mark, mostly tend to be 'under-earners' as well.

These two observations can be interpreted as offering the following recommendation:

A listing can expect to earn more than peers by: 1) ensuring it receives regular reviews; and 2) keeping a listing's price reasonably low and never exceeding £1,000

This is a good example of a clear and actionable insight captured by this model, which can be leveraged by hosts to improve their listing's revenue performance.

Conclusion

To conclude it is worth evaluating to what extent the initial analysis goal and objectives have been met.

The primary goal has been addressed in that the analysis has identified specific characteristics driving revenue, and how this can be translated into clear 'data-driven' recommendations to improve listing revenue.

However a few issues were encountered casting doubt on the analysis results, the main one being unexpectedly high levels of noise in the data. The main consequences of this were:

- Powerful explanatory models such as linear regression were ruled out as unsuitable for the data.
- The accuracy of the models used – between 70-80% – falls short of real world standards.
- The results of logistic regression – which makes linear assumptions about the data – are open to question, and require validation.

One potential explanation for this noisiness is the data source used. Insideairbnb.com scrapes the data without the consent of Airbnb, raising questions around data accuracy and quality (though it is a widely used data source). As a next step, it would be worth exploring ways to source official data from Airbnb to validate and enrich the results of this analysis.

References

1. Parliament. House of Commons (2018). *The growth in short-term lettings (England)*. (HC 8394, 2018-19). [Accessed 8 December 2018]. Available at: http://researchbriefings.files.parliament.uk/documents/CBP-8395/CBP-8395.pdf
2. Airbnb (2018). *Airbnb Press Room: Fast Facts*. [Accessed 8 December 2018]. Available at: https://press.airbnb.com/en-uk/fast-facts/
3. Airbnb Citizen (2017). *Airbnb UK Insights Report*. [Accessed 8 December 2018]. Available at: https://www.airbnbcitizen.com/a-look-at-the-impact-of-home-sharing-across-the-uk/
4. Residential Landlords Association (2016). *The Bedroom Boom: Airbnb and London*. [Accessed 8 December 2018]. Available at: https://research.rla.org.uk/wp-content/uploads/The-Bedroom-Boom-Airbnb-and-London.pdf