



CAN MACHINE LEARNING IMPROVE MY AIRBNB?

Joseph Pearson

January 2019

London DSI 7



Airbnb Stats:

- 640,000 hosts globally
- 4.0 million listings globally
- 150 million Airbnb users worldwide
- 700,000 Airbnb Guests in 2017 in Berlin alone.



Airbnb Stats:

- 640,000 hosts globally
- 4.0 million listings globally
- 150 million Airbnb users worldwide
- 700,000 Airbnb Guests in 2017 in Berlin alone.

How do you make sure you capitalize on this?

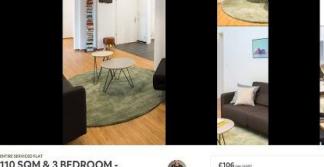
Increase your Occupancy Rate

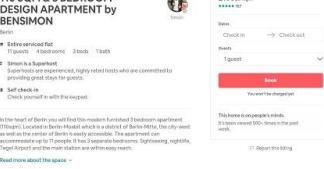
The Hypothesis . . .

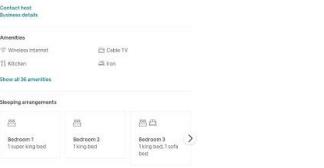


Occupancy Rate can be predicted and therefore optimized based on listing information and NLP of reviews

A search bar at the top left contains the word "Search". To its right are buttons for "Become a host", "Help", and "Sign up".

A large central image shows the interior of a modern apartment. It features a living room with a dark sofa, a round green rug, and a small white coffee table. In the background, there's a dining area with a wooden table and chairs, and a kitchen area. A bookshelf filled with books is visible on the left.

A smaller image on the right shows a different view of the apartment, possibly a bedroom or another part of the living space. It includes a bed, a nightstand, and some decorative items.

A third smaller image at the bottom right shows a close-up of a room, likely a bathroom or a small study, with a desk and some plants.

**ENTIRE SERVICED FLAT
110 SQM & 3 BEDROOM - DESIGN APARTMENT by BENISONIM**

A circular profile picture of a woman with short brown hair, smiling. Below it is her name, "BENISONIM".

£10€ per night
A week = 101€

Check-in → Check-out
Week 1 guest

Book
You've got the right spot!

In the heart of Berlin you will find the modern & minimal 3 bedroom apartment (110sqm). Located in Berlin-Mitte which is a district of Berlin Mitte, the city center. The apartment is located in a quiet residential building, very close to the subway, recommended up to 10 people. It has 3 separate bedrooms. Brightening, nightlife, Tokyo style interior, modern kitchen and within walking reach.

Read more about the spot...

[Cancel host](#) [Business details](#)

Anemones

wireless internet Cable TV
 Washer Iron

[Show all 26 amenities](#)

Sleeping arrangements


Bedroom 1
1 double bed


Bedroom 2
1 single bed


Bedroom 3
1 triple bed, 1 sofa bed

Accessibility

Disabled parking spot

Availability

Updated today

A calendar for January 2019 showing dates from Monday to Sunday. The 19th is highlighted in yellow, indicating it's the current date.

157 Reviews 



Accuracy	Communication	Cleanliness	Location	Check-in	Value
					

 **Herry**
January 2019

This is a great apartment - clean and immaculate. Simon was easy to contact, very responsive and helpful. The location is excellent and the neighborhood provided. The subway station is just round the corner and we found it easy to get around from there. Some of us had trouble

 **Herry**
December 2018

Largest of the spaces with easy check-in and very attentive host. Would recommend.

The Data . . .



Data from <http://insideairbnb.com/berlin/> in csv format scrapped from AirBnb

Reviews . . .

listing_id	id	date	reviewer_id	reviewer_name		comments	length	language
0	2015	69544350	2016-04-11	7178145	Rahel	Mein Freund und ich hatten gute gemütliche vie...	88	de
1	2015	69990732	2016-04-15	41944715	Hannah	Jan was very friendly and welcoming host! The ...	29	en
2	2015	71605267	2016-04-26	30048708	Victor	Un appartement tres bien situ� dans un quartie...	47	fr
3	2015	73819566	2016-05-10	63697857	Judy	It is really nice area, food, park, transport ...	10	en
4	2015	74293504	2016-05-14	10414887	Romina	Buena ubicaci�n, el departamento no est� orden...	53	es

- 401963 reviews from April 2017 to November 2018

Listings . . .

- All the rest of the Listing information
 - 22500 unique listings

The Cleaning . . .



- Many columns contained mainly null values or very few non-zero values and were dropped.
- Listings with null values dropped
- Review lengths calculated and language of review each review determined and non-english reviews dropped
- Dropped reviews with < “50 days occupied/month” (I will explain this later)
- Dummified or Binarized many features

The Preprocessing . . .



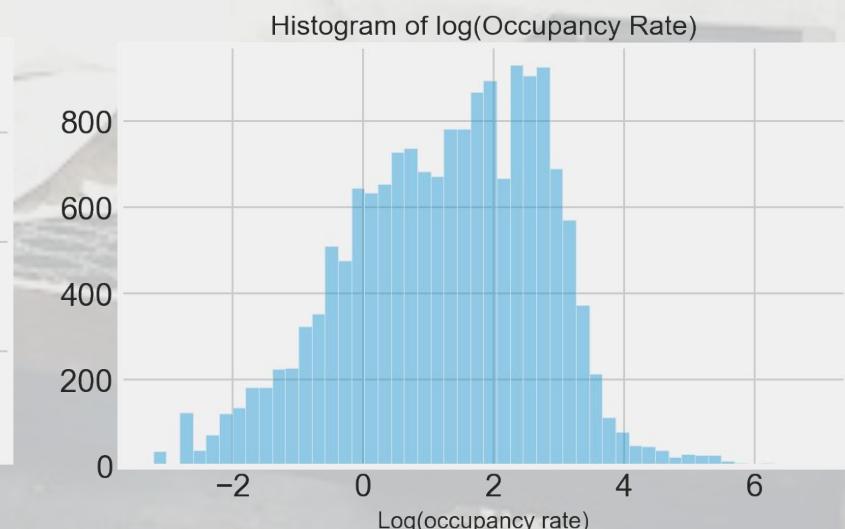
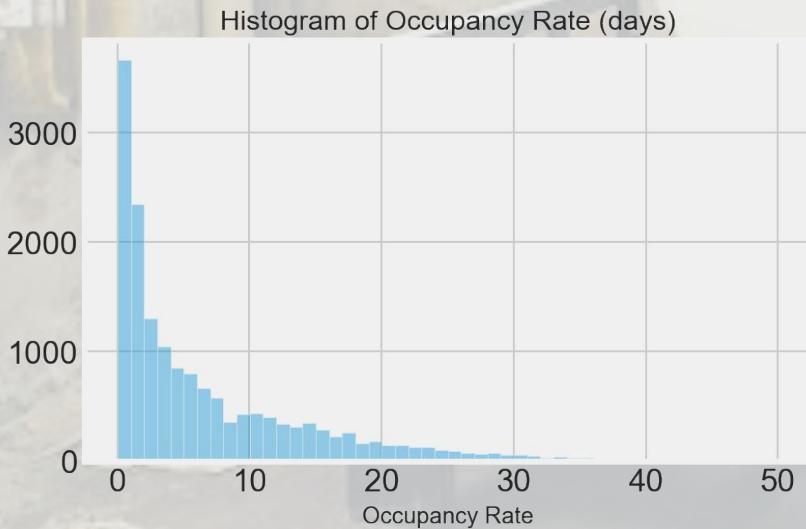
- Language of host determined and added as additional feature
- Vader Sentiment calculated for each review and then aggregated by listing
- CountVectorizer, TFiDF, Textacy doc_to_terms + Vectorizer, Textacy Topic Modeling
- Reviews and listing data-frames merged for modelling
- Final Dataset:
 - 16,500 listings,
 - 244 features (not including word counts and topics)

The Target . . .



Occupancy Rate:

$$\frac{\text{Reviews_per_Month} \times 2.4 \text{ days}}{0.5}$$

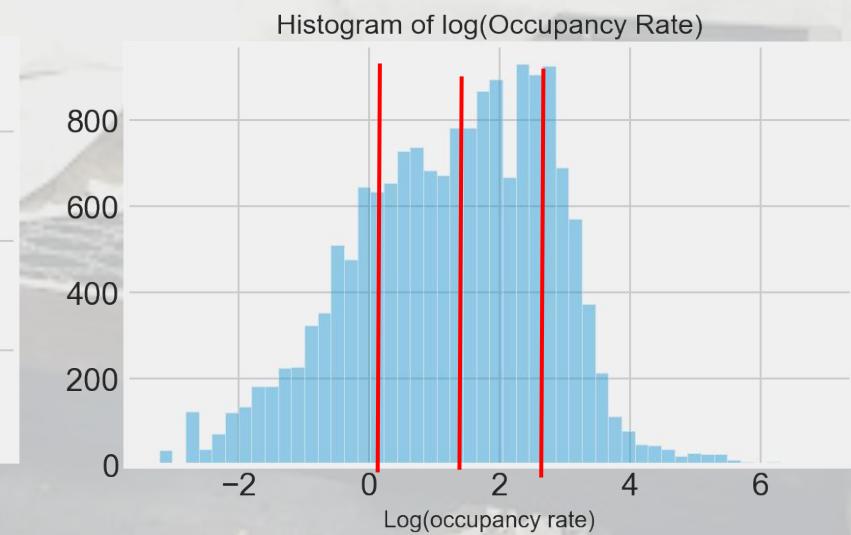
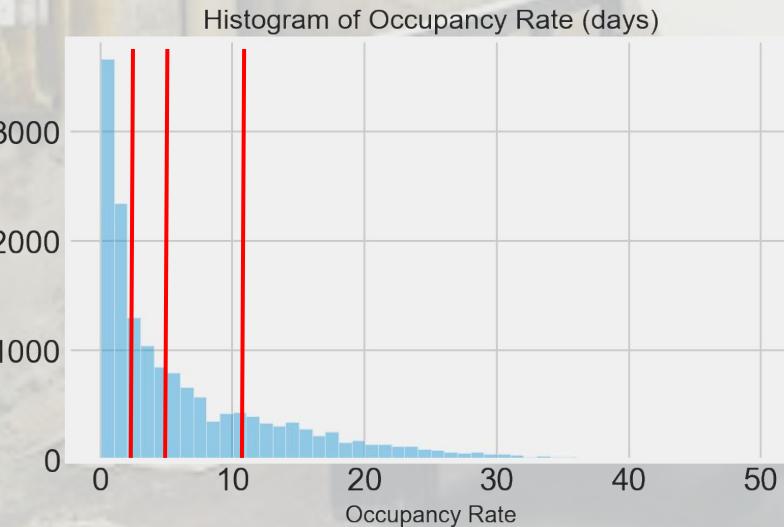


The Target . . .



Occupancy Rate:

$$\frac{\text{Reviews_per_Month} \times 2.4 \text{ days}}{0.5}$$

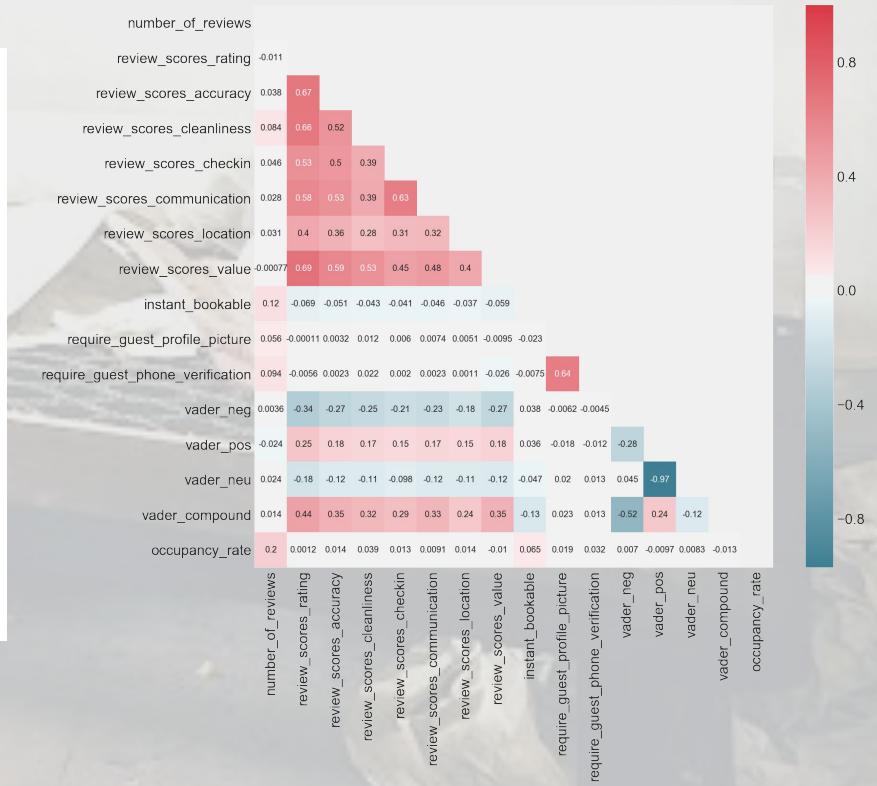
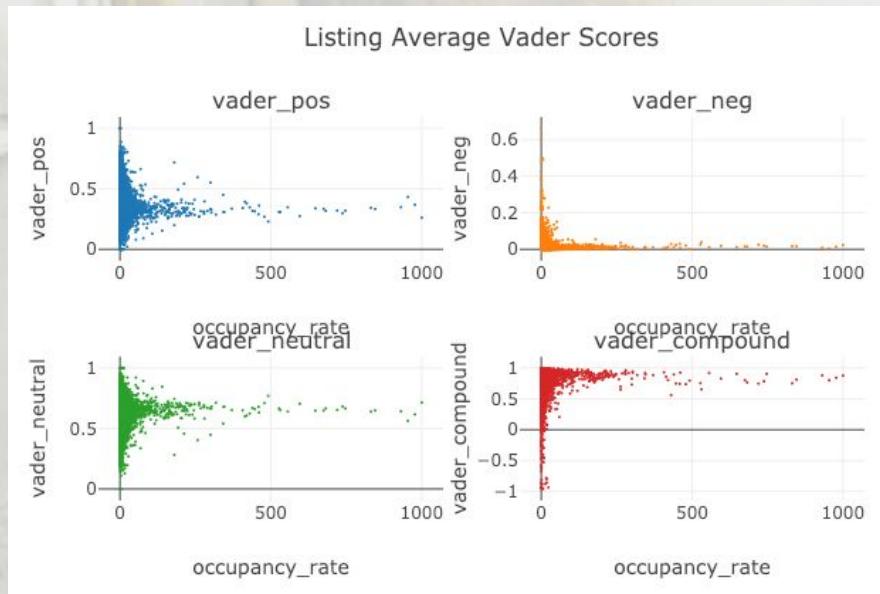


Classification target (quartiles):

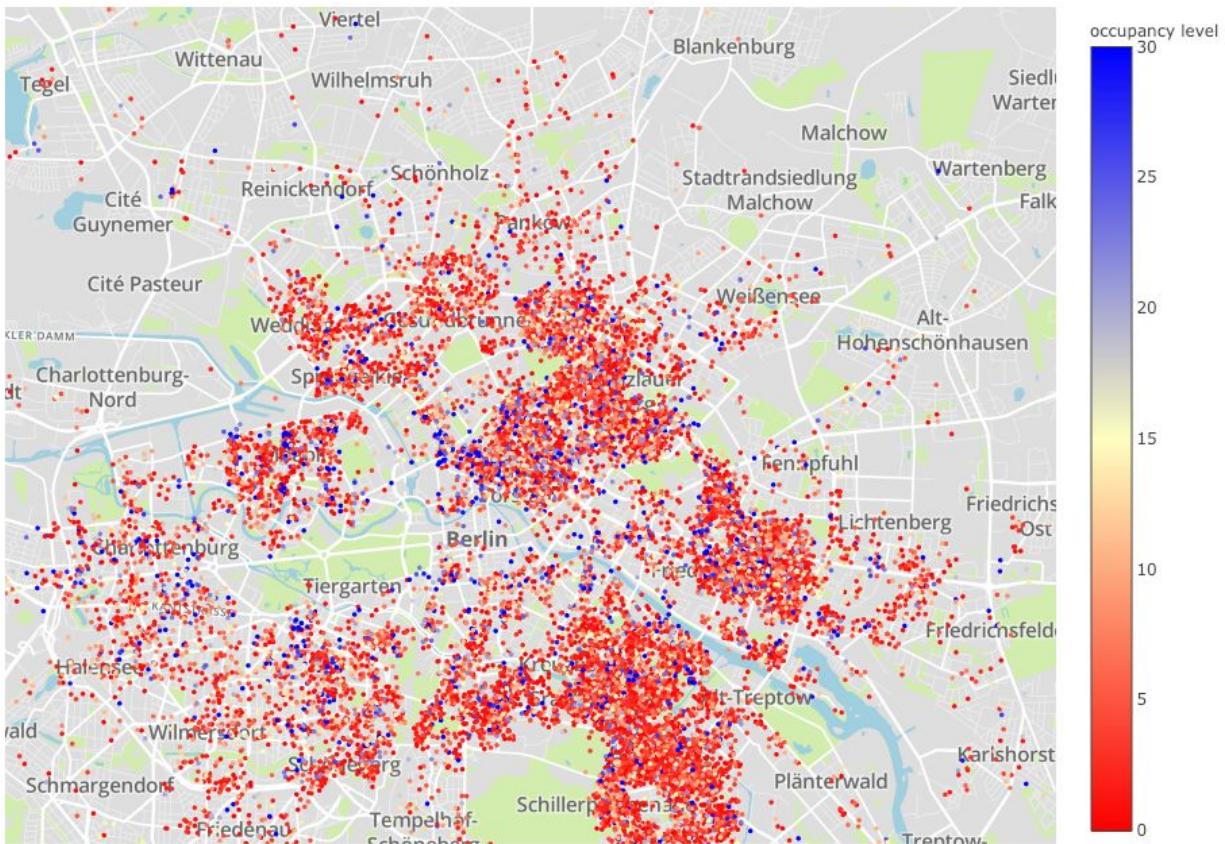
Occupancy Rate = ['low' < 1.2 days < 'med_low' < 3.96 days 'med_high' < 10.8 days < 'high']

log(Occupancy Rate) = ['low' < 0.182 < 'med_low' < 1.376 'med_high' < 2.379 days < 'high']

The EDA . . .



AirBnbs in Berlin



The EDA . . .

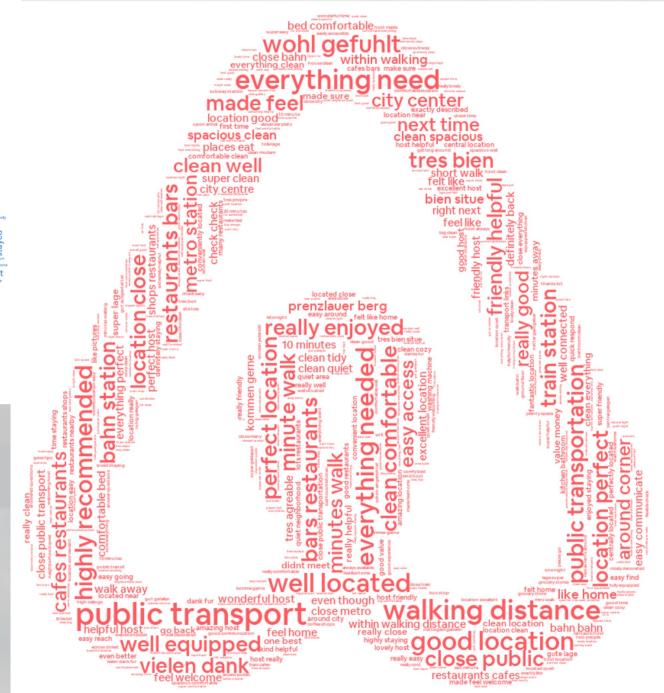
Positive Words



Negative Airbnb reviews word cloud



Positive n_grams (2 & 3)



The EDA...

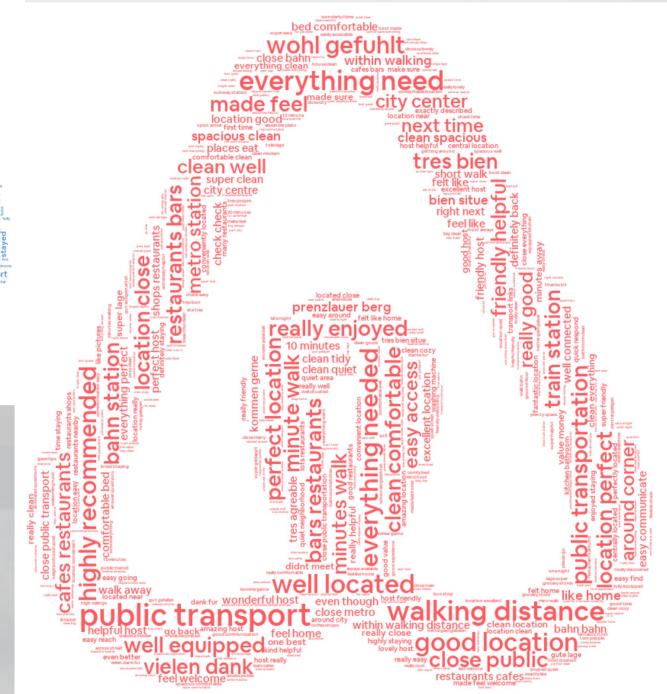
Positive Words



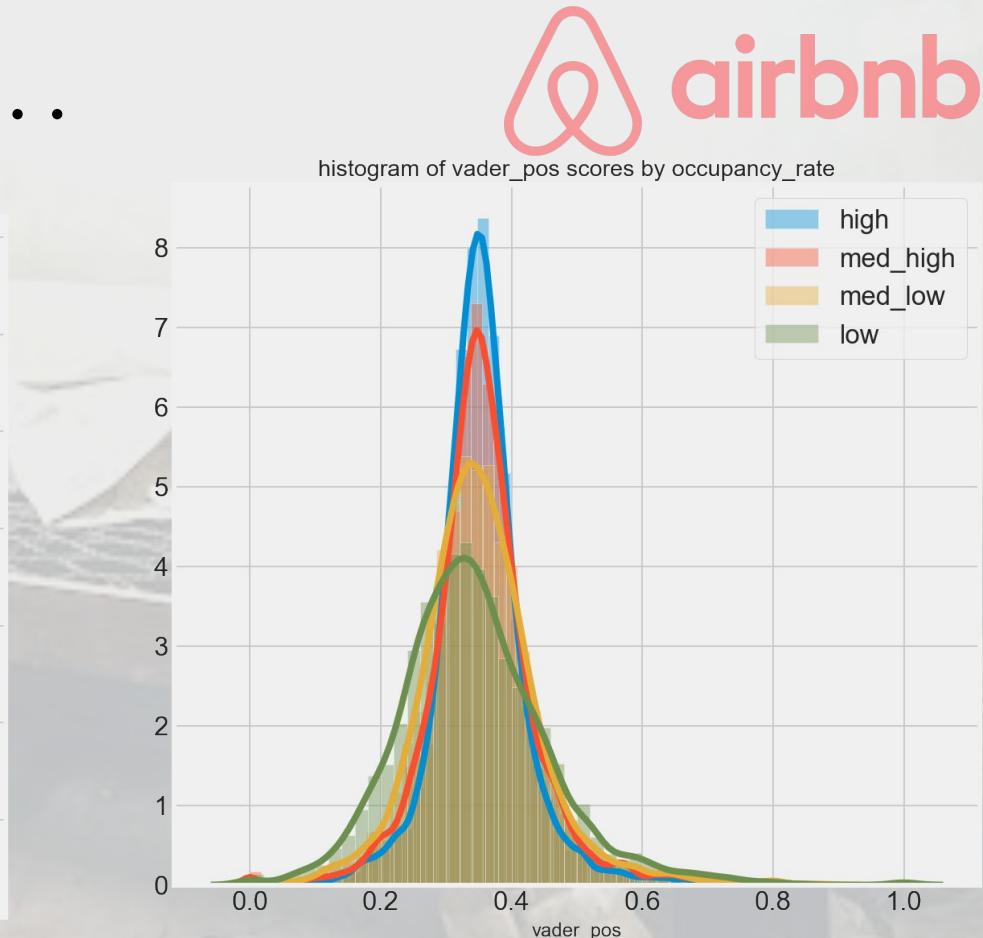
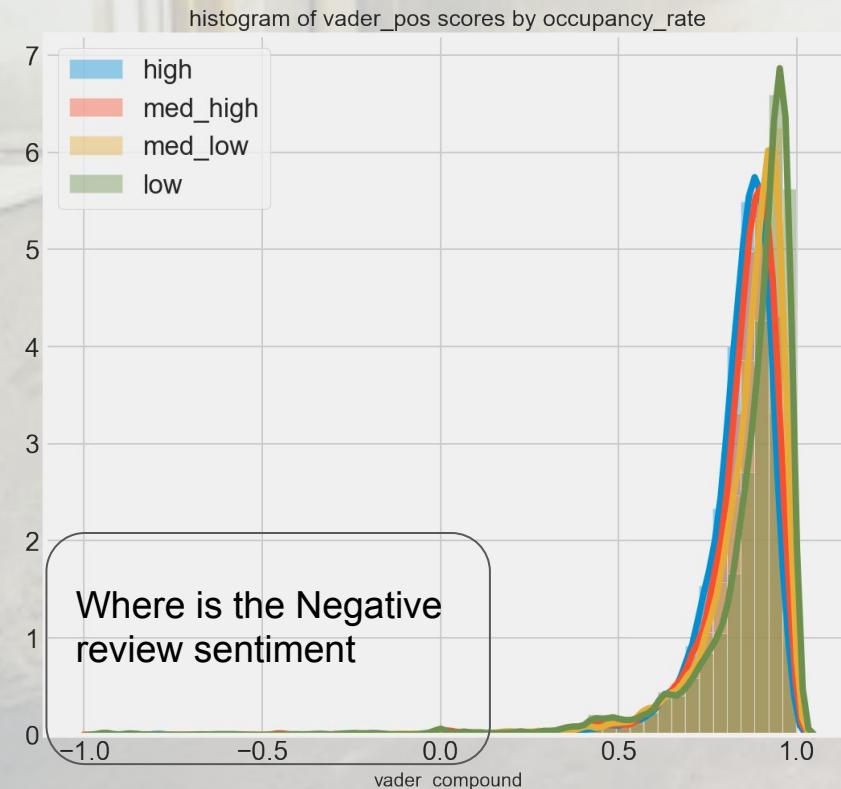
Negative Airbnb reviews word cloud



Positive n_grams (2 & 3)



The Issue with NLP . . .



The Issue with NLP . . .



Sophie

August 2017



[REDACTED] However it is visually clean but rather dusty and dirty closer up. You must be quite from 10pm and Claudio is not there as it first appears on the listing. [REDACTED] as we had a small mix up with checking in [REDACTED]
[REDACTED]

Is this a Negative Review (?) . . .

The Issue with NLP . . .



Sophie

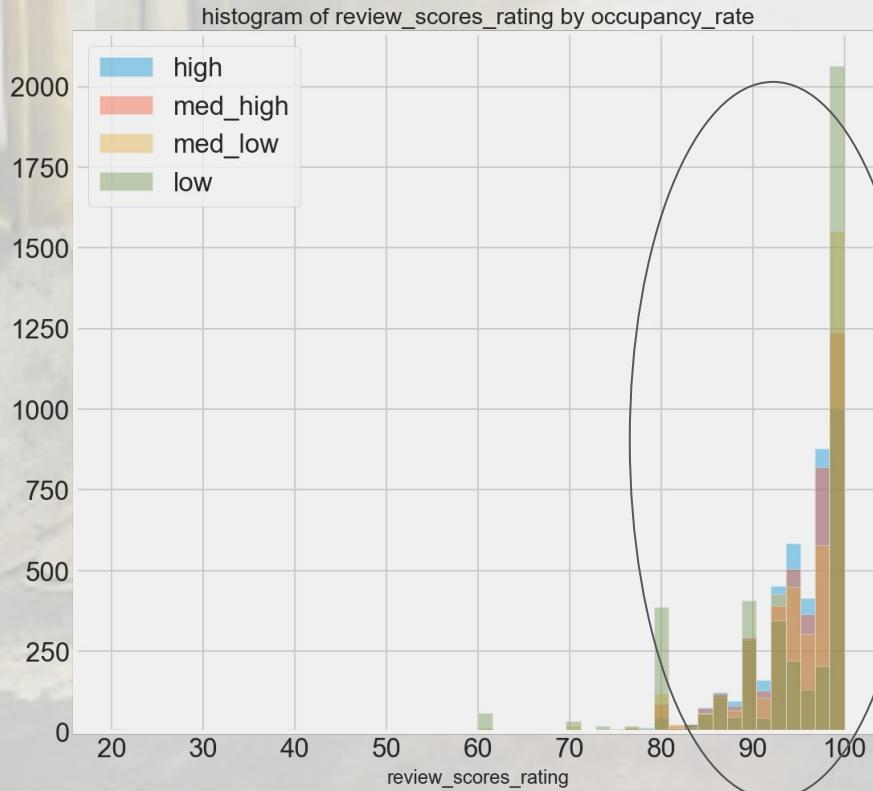
August 2017



Room is a lot bigger and nicer in real life than on pictures. However it is visually clean but rather dusty and dirty closer up. You must be quite from 10pm and Claudio is not there as it first appears on the listing. Andy was incredibly helpful as we had a small mix up with checking in and all members were extremely friendly and lovely to meet. Overall we had a really lovely stay.

Is this a Negative Review . . . or a Positive review?

And Issue with AirBnb Reviews



The Models . . .



Regression

- Linear Regression
- LASSO Regression
- Ridge Regression
- Gradient Boosting
- Support Vector Machine
- Decision Tree
- Random Forest

Classification

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Support Vector Machine



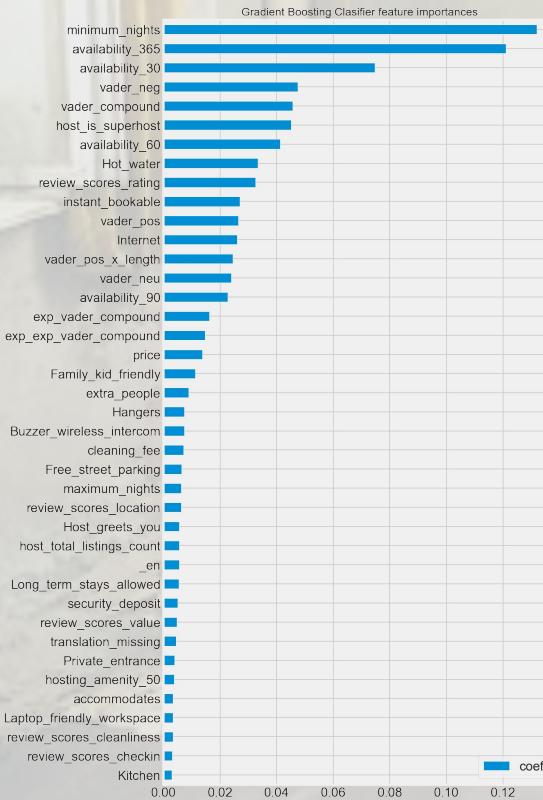
The Features . . .

Features tried:

- Listing Features only with Vader sentiment
- CountVectorisor & TFidF on review comments with and without listing features
- CountVectorisor & TFidF on Description column with and without listing features
- Textacy vectorisor on review comments with and without listing features
- Topics from Textacy topic models with Listing Features
- TruncatedSVD() on all above combinations

The Best Feature Combination : Listing Features with Vader Sentiment scores

The Classification Results . . .



GradientBoostingClassifier()

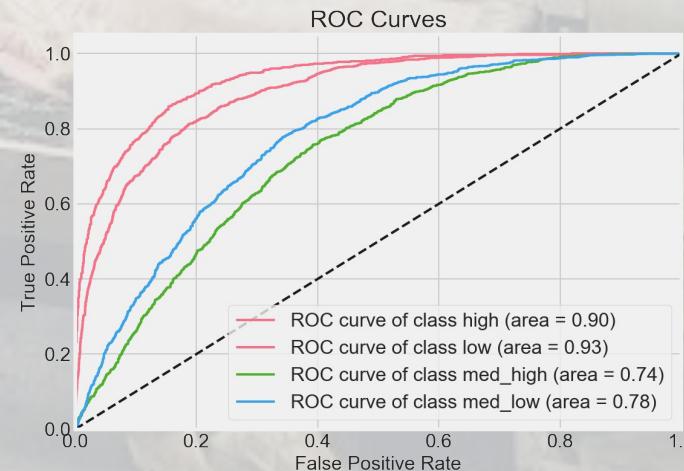
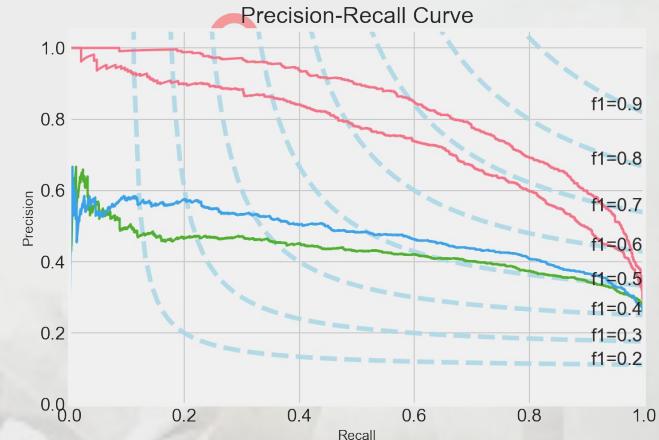
Test score: 0.59882899252978

Predicted high low med_high med_low All

Actual	high	low	med_high	med_low	All
high	873	14	267	91	1245
low	17	957	62	224	1260
med_high	369	73	514	283	1239
med_low	93	233	261	622	1209
All	1352	1277	1104	1220	4953

precision recall f1-score support

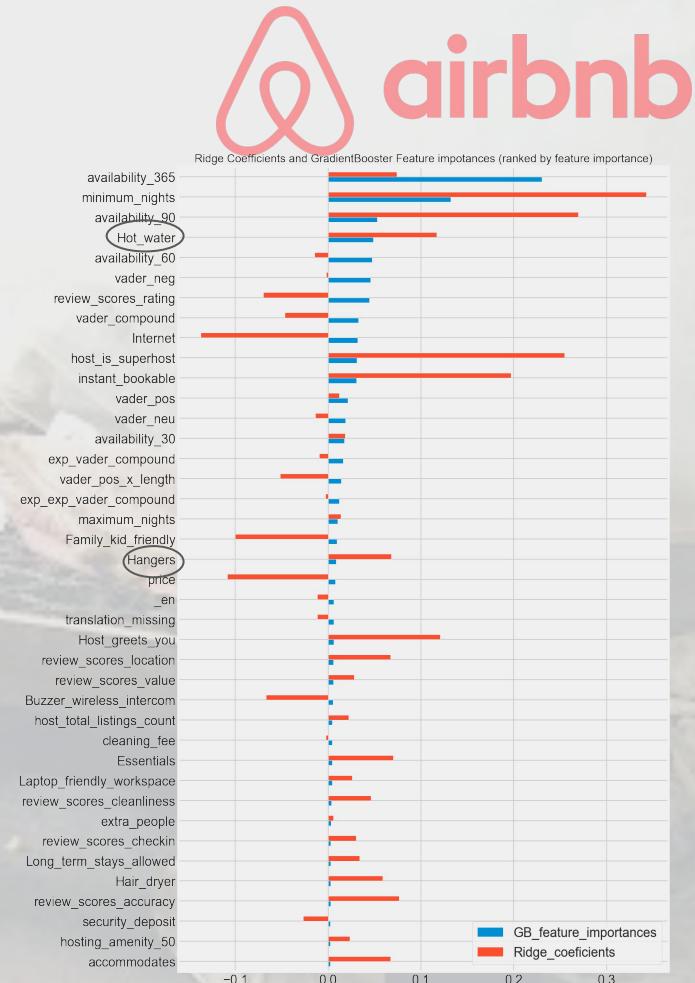
	high	low	med_high	med_low	All
high	0.65	0.70	0.67	1245	
low	0.75	0.76	0.75	1260	
med_high	0.47	0.41	0.44	1239	
med_low	0.51	0.51	0.51	1209	
micro avg	0.60	0.60	0.60	4953	
macro avg	0.59	0.60	0.59	4953	
weighted avg	0.59	0.60	0.60	4953	



The Important Features . . .

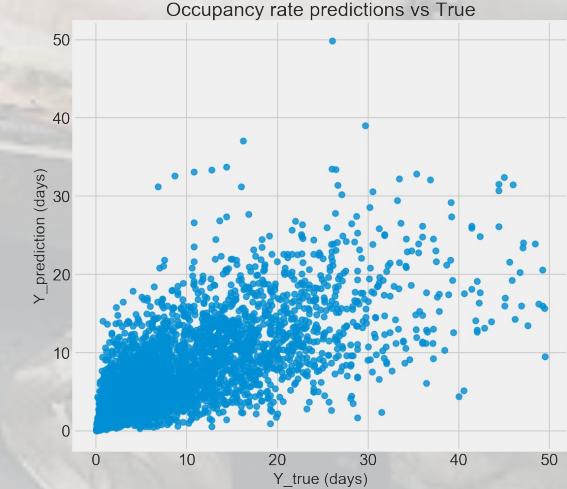
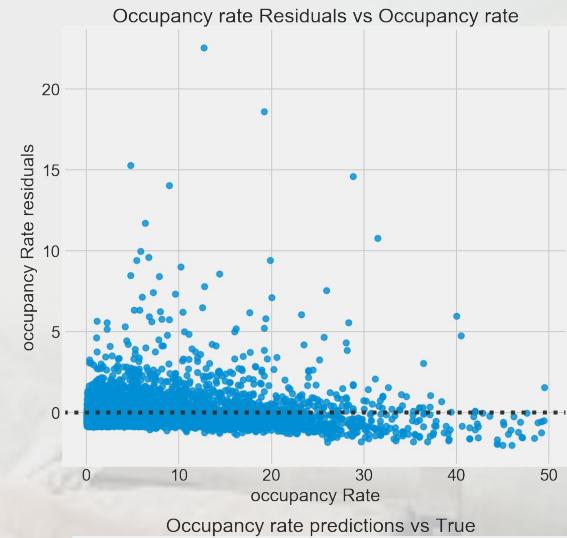
The Top 40 Features all pass the sense check

- “Hangers” stand out to me. Why are they important?
- “Hot_water” . . . ?



The Regression Results . . .

Regression Model Results			
	Gradient Boosting Regressor	Support Vector Machine	Tuned Ridge Linear Regression
Cross-validated score on training set	0.715	0.554	0.470
RMSE (average error in $\log(\text{days})$)	0.748	0.94	1.059
Score on unseen test set	0.734	0.582	0.469



What could be improved?



- Better occupancy information
 - Actual nights booked vs Actual nights available
- Where are the negative reviews?
- Add listing descriptions and headlines and the Host descriptions into the models
- Additional factors such as distance to public transport and tourist attractions.

What Next? . . .

- Other cities
- Web app ...
 - Like this one I “created” yesterday

The image shows a laptop screen displaying the AIRDNA software interface. The interface is designed for Airbnb hosts to manage their properties. It features a sidebar with various tabs such as Overview, Pricing, Occupancy, Seasonality, Revenue, Rental Analysis, Top Properties, Guests, and Rentalizer. The main content area shows two property listings:

- Laguna Beach Private Getaway**: Room in Boutique Hotel, Laguna Beach, 92651. Studio, 1 bath, 6 guests. ADR: \$233, OCC: 93%, RevPAR: \$216. 4.6 (223 reviews).
- Ocean Views from Ocean Avenue**: Room in Boutique Hotel, Santa Monica, 90407. 1 bed, 1 bath, 3 guests. ADR: \$189, OCC: 50%, RevPAR: \$94. 4.5 (171 reviews).

Below the listings is a "REVENUE POTENTIAL TREND" chart showing monthly performance from November 2017 to November 2018. The chart includes data points for REVENUE POTENTIAL (\$84K vs \$94K), OCCUPANCY (93% vs 62%), AVERAGE DAILY RATE (\$233 vs \$328), REVPAR (\$216 vs \$203), and LENGTH OF STAY (2.0d vs 3.2d). The chart also indicates a peak in revenue potential around July/August 2018.



Thankyou . . .

The Questions . . .