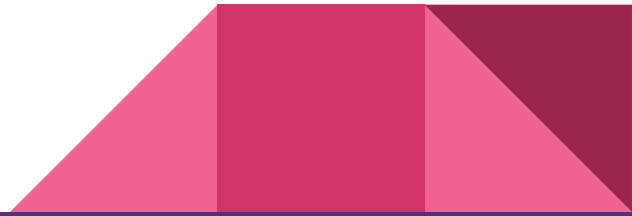
Price and Popularity Analysis of Airbnbs in NYC

Anbo Guo(ag6738)
Letao Hou (lh3163)
Yanchen Liu (yl4265)
Yuetong Zhou(yz6729)

Introduction

- Sharing economy market captures eyes
- Airbnb: a famous lodging company which provides vacation rental housing
- Increasing number of people entering the sharing market
- Our project is aiming for helping the hosts, especially new hosts to increase booking rates and earn more profit form the market
- Our main hypothesis is that seasonal trends exist, people prefer private properties over shared ones and location matters.



Data

- NYC airbnb daily open data from October 2017 to October 2018 & NYC Airbnb listings data (Airbnb Open Data from Inside Airbnb); contains detailed listings data for New York City
- Airbnb calendar price datasets of October 2017-2018 & 2021-2022

Methods and Results

- Correlation analysis and PCA
- Time Series analysis and prediction model
- Booking rates analysis and prediction model

Correlation Analysis

- Correlation between listing features and prices
- Correlation between listing features and booking rates

Correlation--prices

- Private properties are more expensive than the shared ones
- Property location and property size matter

	features	pri_cor
373	room_type_Private room	-0.277343
380	bathrooms_text_1 shared bath	-0.232763
337	property_type_Private room in apartment	-0.203694
346	property_type_Private room in house	-0.120518
352	property_type_Private room in townhouse	-0.079409
385	bathrooms_text_2 shared baths	-0.078217
382	bathrooms_text_1.5 shared baths	-0.065834
374	room_type_Shared room	-0.056850
14	calculated_host_listings_count	-0.056320
407	host_response_time_within an hour	-0.051955
13	instant_bookable	-0.051073
359	property_type_Shared room in apartment	-0.050788
46	host_neighbourhood_Bushwick	-0.048563
1	popularity	-0.045752
19	host_neighbourhood_Astoria	-0.041697
98	host_neighbourhood_Elmhurst	-0.039346
89	host_neighbourhood_East Flatbush	-0.036654
108	host_neighbourhood_Flushing	-0.035766
91	host_neighbourhood_East New York	-0.033321
28	host_neighbourhood_Bedford-Stuyvesant	-0.031916

	features	pri_cor
5	price	1.000000
2	accommodates	0.486090
3	bedrooms	0.421059
4	beds	0.357533
371	room_type_Entire home/apt	0.285180
384	bathrooms_text_2 baths	0.248369
395	bathrooms_text_5 baths	0.239291
390	bathrooms_text_3.5 baths	0.214501
386	bathrooms_text_2.5 baths	0.208358
332	property_type_Entire townhouse	0.189382
394	bathrooms_text_4.5 baths	0.155531
323	property_type_Entire condominium	0.139420
388	bathrooms_text_3 baths	0.139367
328	property_type_Entire house	0.130608
329	property_type_Entire loft	0.128376
251	host_neighbourhood_Soho	0.102206
348	property_type_Private room in resort	0.094399
289	host_neighbourhood_Upper West Side	0.093181
397	bathrooms_text_6 baths	0.090711
383	bathrooms_text_15.5 baths	0.085427

Correlation--booking rates

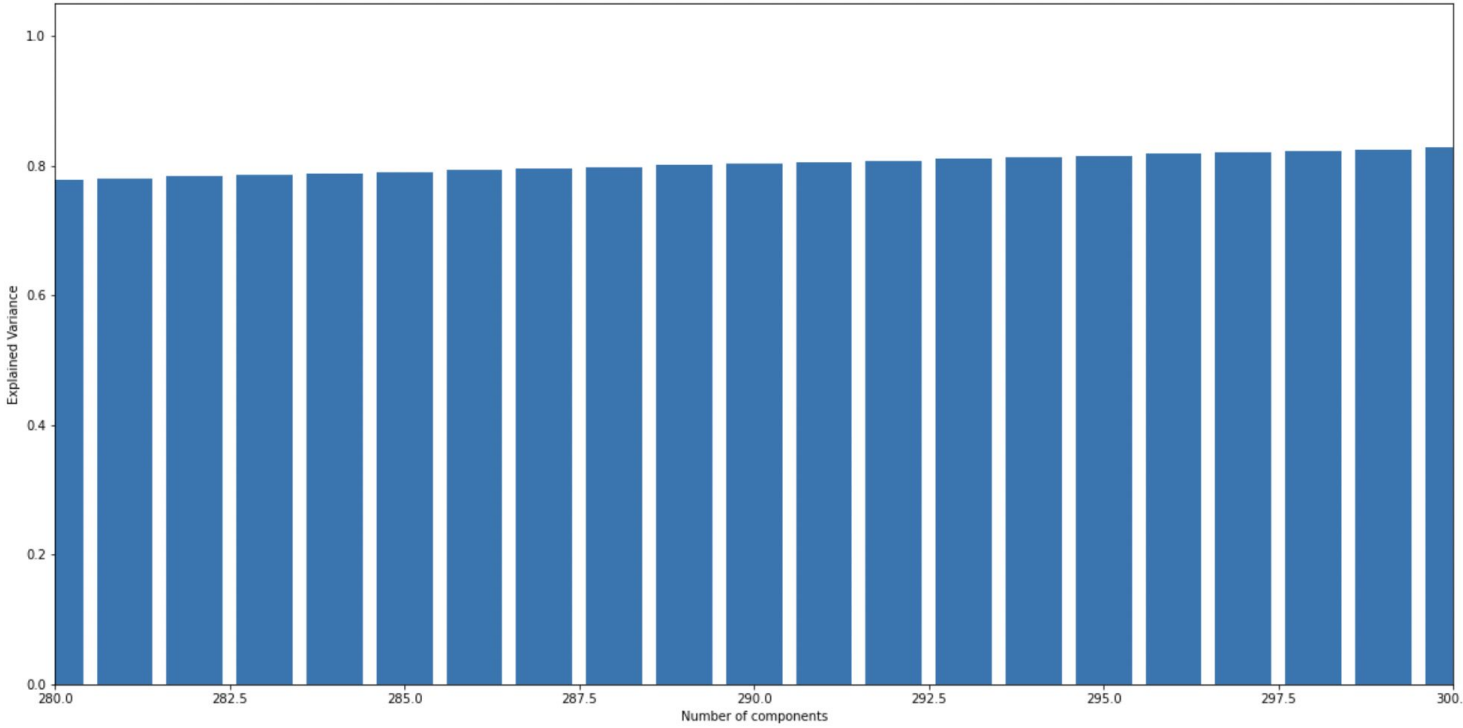
- Guests prefer private properties over shared ones
- Property location matters

	features	bora_cor
14	calculated_host_listings_count	-0.037390
6	minimum_nights	-0.030905
288	host_neighbourhood_Upper East Side	-0.026678
406	host_response_time_within a few hours	-0.022905
7	maximum_nights	-0.017764
380	bathrooms_text_1 shared bath	-0.016634
381	bathrooms_text_1.5 baths	-0.016296
19	host_neighbourhood_Astoria	-0.012707
91	host_neighbourhood_East New York	-0.011998
374	room_type_Shared room	-0.011497
89	host_neighbourhood_East Flatbush	-0.010674
388	bathrooms_text_3 baths	-0.010673
167	host_neighbourhood_Long Island City	-0.010346
283	host_neighbourhood_Tribeca	-0.010084
404	host_response_time_a few days or more	-0.010074
47	host_neighbourhood_Cambridge	-0.009112
386	bathrooms_text_2.5 baths	-0.008568
98	host_neighbourhood_Elmhurst	-0.008274
377	bathrooms_text_0 shared baths	-0.008214
389	bathrooms_text_3 shared baths	-0.008181

	features	bora_cor
0	booking_rate	1.000000
1	popularity	0.095921
15	host_neighbourhood_Allerton	0.068734
308	host_neighbourhood_Windsor Terrace	0.039493
106	host_neighbourhood_Flatiron District	0.038395
9	host_is_superhost	0.038181
372	room_type_Hotel room	0.034426
46	host_neighbourhood_Bushwick	0.028523
407	host_response_time_within an hour	0.027780
28	host_neighbourhood_Bedford-Stuyvesant	0.024469
253	host_neighbourhood_South Beach	0.023754
186	host_neighbourhood_Midtown East	0.021085
162	host_neighbourhood_Lefferts Garden	0.020827
353	property_type_Private room in villa	0.019319
127	host_neighbourhood_Greenwich Village	0.018718
13	instant_bookable	0.018376
378	bathrooms_text_1 bath	0.017814
358	property_type_Room in serviced apartment	0.016830
385	bathrooms_text_2 shared baths	0.016102
355	property_type_Room in boutique hotel	0.016041

PCA

- Choose n_components = 290 for 80% explained variance
- (10163,408)
  to
  (10163, 290)



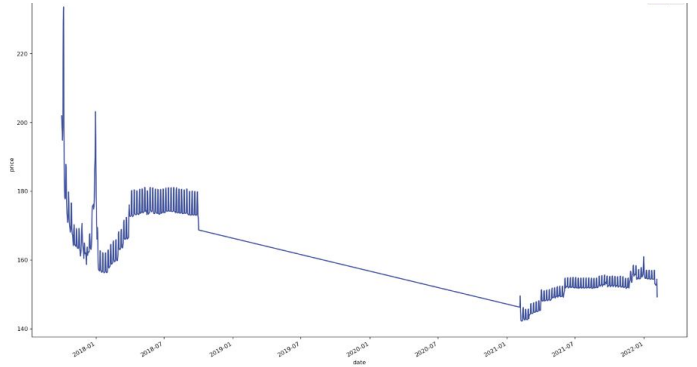
Time Series analysis and prediction model

Data cleaning and time series

	listing_id	date	available	price
0	2515	2018-10-01	t	$99.00
1	2515	2018-09-30	t	$89.00
2	2515	2018-09-29	t	$99.00
3	2515	2018-09-28	t	$99.00
4	2515	2018-09-27	t	$99.00
...	...	...	...	...
13464016	47939451	2022-01-31	f	$61.00
13464017	47939451	2022-02-01	f	$61.00
13464018	47939451	2022-02-02	f	$61.00
13464019	47939451	2022-02-03	f	$61.00
13464020	47939451	2022-02-04	f	$62.00

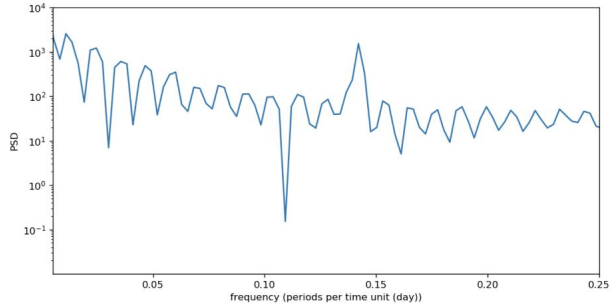
19044658 rows × 4 columns

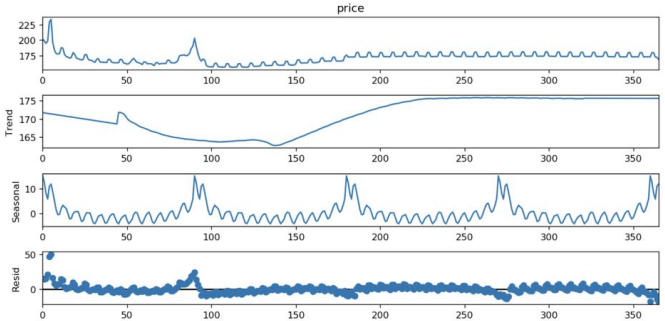
date	2017-10-02	2017-10-03	2017-10-04	2017-10-05	2017-10-06	2017-10-07	2017-10-08	2017-10-09	2017-10-10	2017-10-11	...
listing_id											
2515	NaN	NaN	NaN	NaN	NaN	NaN	NaN	99.0	99.0	NaN	...
2539	150.0	150.0	150.0	150.0	99.0	99.0	150.0	150.0	150.0	150.0	...
2595	198.0	198.0	198.0	198.0	198.0	NaN	NaN	198.0	198.0	198.0	...
3330	NaN	NaN	NaN	NaN	NaN	NaN	70.0	70.0	70.0	70.0	...
3647	150.0	150.0	150.0	150.0	150.0	150.0	150.0	150.0	150.0	150.0	...
...	...	...	...	...	...	...	...	...	...	...	...
48033101	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
48033611	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
48038944	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
48039640	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
48039776	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...



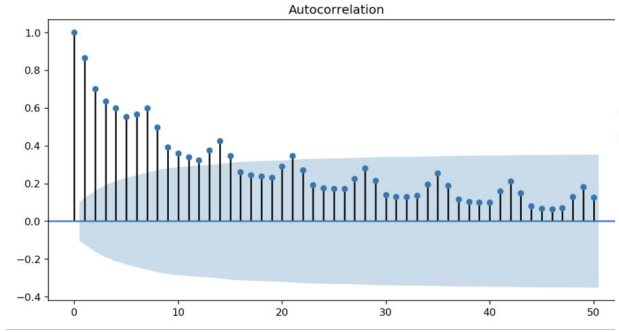
Time Series analysis and prediction model

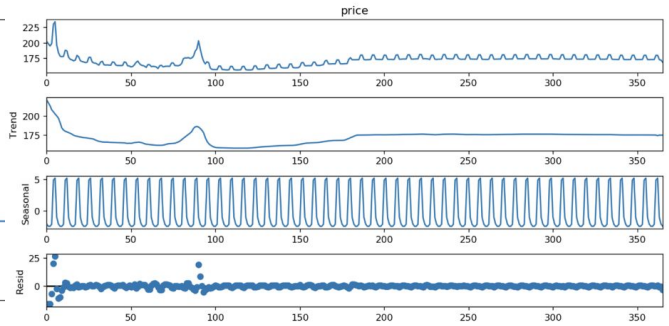
Signal periodogram test and ACF test



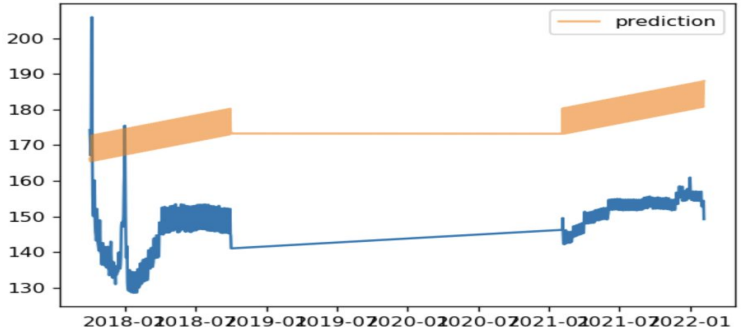


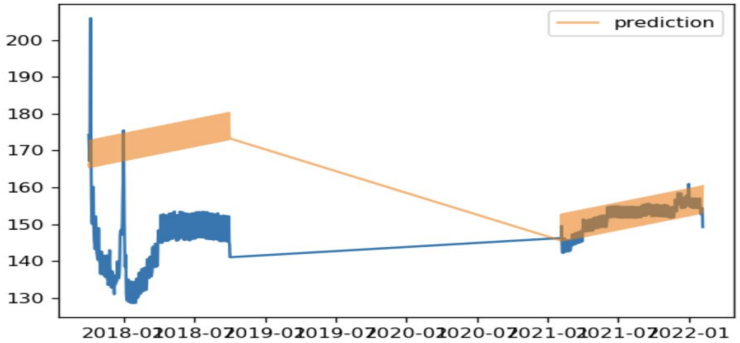
Strongest period length = 91.5

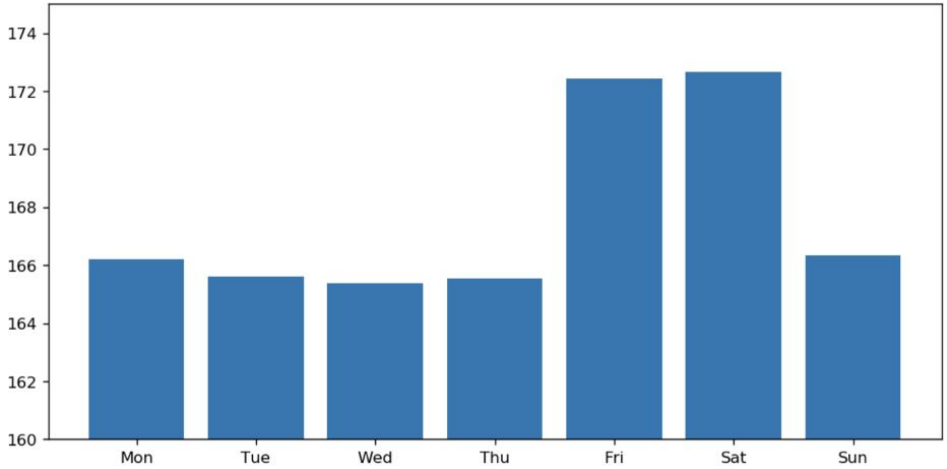




Time Series analysis and prediction model





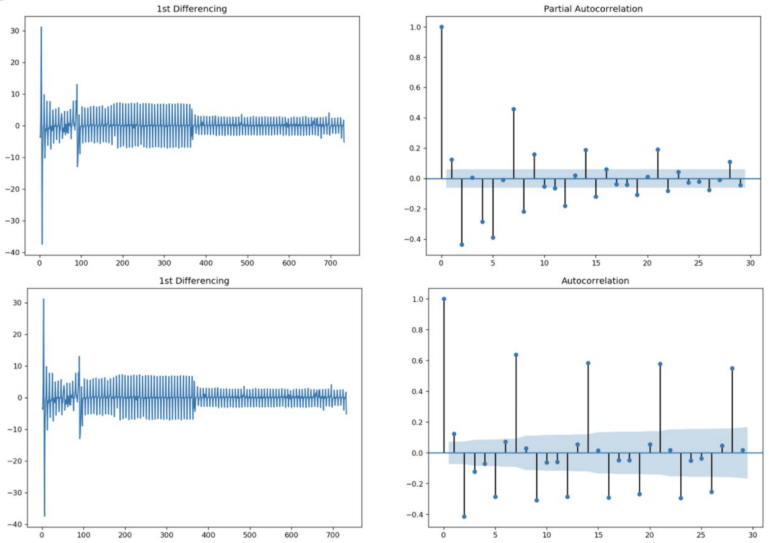


Time Series analysis and prediction model

ARIMA(1,1,3)

```
ADF Statistic: -2.704897
p-value: 0.073195
Critical Values:
        1%: -3.440
        5%: -2.866
        10%: -2.569
```

```
1st order differencing
ADF Statistic: -6.551318
p-value: 0.000000
2nd order differencing
ADF Statistic: -11.994548
p-value: 0.000000
```



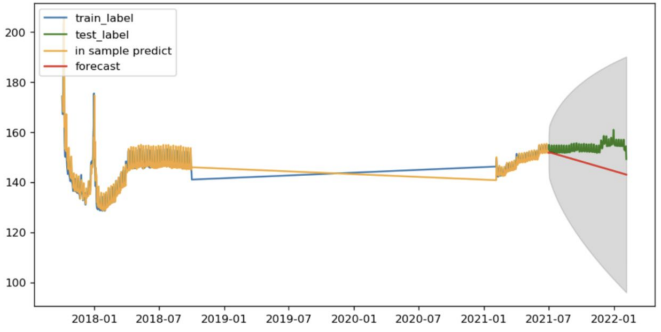
ARIMA Model Results

Dep. Variable:	D.price	No. Observations:	513
Model:	ARIMA(1, 1, 3)	Log Likelihood	-1348.560
Method:	css-mle	S.D. of innovations	3.350
Date:	Fri, 30 Apr 2021	AIC	2709.120
Time:	20:03:45	BIC	2734.561
Sample:	1	HQIC	2719.092

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0413	0.071	-0.584	0.559	-0.180	0.097
ar.L1.D.price	0.1722	0.223	0.772	0.440	-0.265	0.609
ma.L1.D.price	-0.0127	0.221	-0.058	0.954	-0.445	0.420
ma.L2.D.price	-0.5996	0.048	-12.597	0.000	-0.693	-0.506
ma.L3.D.price	0.0053	0.136	0.039	0.969	-0.261	0.272

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	5.8078	+0.0000j	5.8078	0.0000
MA.1	-1.2945	+0.0000j	1.2945	0.5000
MA.2	1.2881	+0.0000j	1.2881	0.0000
MA.3	112.4518	+0.0000j	112.4518	0.0000



Time Series analysis and prediction model

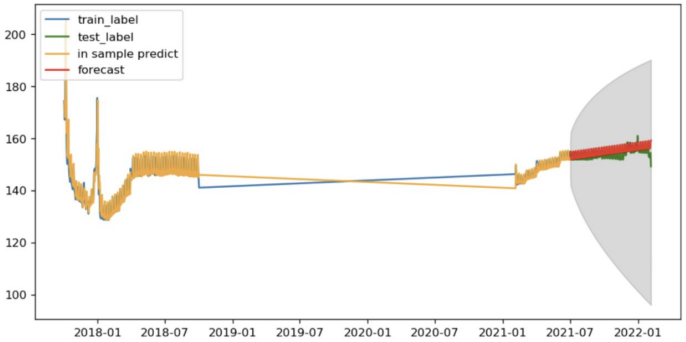
SARIMA(1,1,3,7)

SARIMAX Results

Dep. Variable:	price	No. Observations:	514
Model:	SARIMAX(1, 1, 3)x(1, 1, [1], 7)	Log Likelihood	-1132.905
Date:	Fri, 30 Apr 2021	AIC	2279.809
Time:	20:22:34	BIC	2309.395
Sample:	0 - 514	HQIC	2291.413
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2279	0.120	1.904	0.057	-0.007	0.463
ma.L1	0.0386	0.115	0.337	0.736	-0.186	0.263
ma.L2	-0.3330	0.036	-9.316	0.000	-0.403	-0.263
ma.L3	-0.1041	0.040	-2.627	0.009	-0.182	-0.026
ar.S.L7	0.0725	0.038	1.907	0.056	-0.002	0.147
ma.S.L7	-0.6241	0.037	-17.021	0.000	-0.696	-0.552
sigma2	5.1225	0.099	51.652	0.000	4.928	5.317

Ljung-Box (Q):	75.77	Jarque-Bera (JB):	20172.63
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.12	Skew:	-0.38
Prob(H) (two-sided):	0.00	Kurtosis:	33.92



1. Booking rates analysis and prediction model

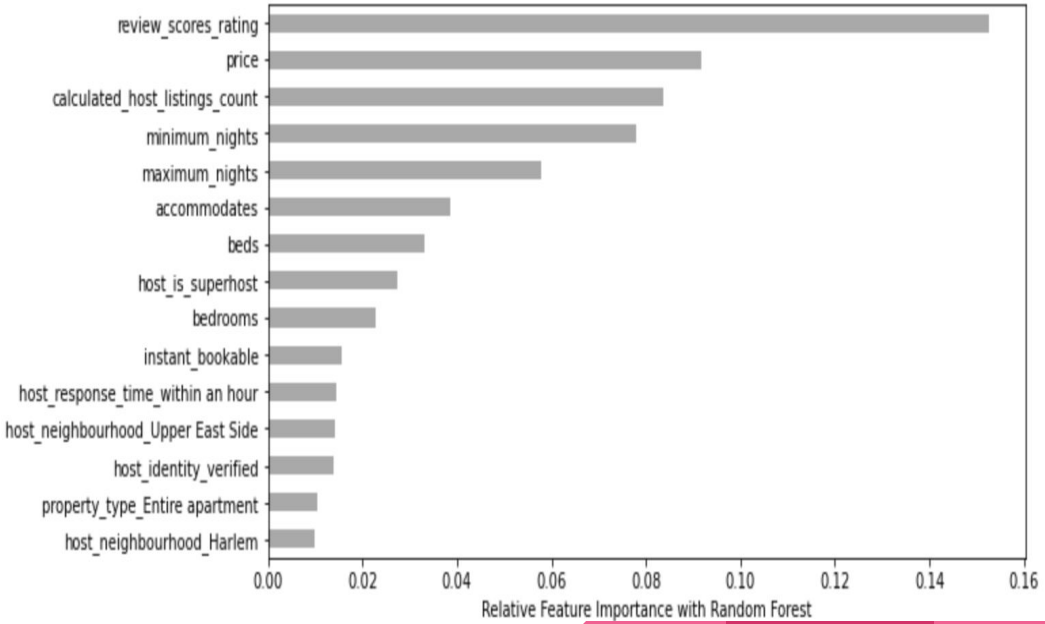
1.Baseline Model : Logistic Regression

2. Random Forest

- Feature importance
- Hyperparameter tuning

3.XGBoost

- Hyperparameter tuning
- Best Model with 0.78 accuracy



Conclusion

1. some features are of vital importance to increase booking rate(responding time, location, min/max nights)
2. some suggestions for the hosts :

- Know the seasonal trends on pricing
- Properties in good locations with high booking rates can rise prices accordingly
- New hosts can use our model to predict popularity and price of properties