NYU | STERN

airbnb

Is the Price Reasonable?

Xinyi Gong, Lanyu Shang, Yang Sun

Role & Problem

Role & Problem:

- Role: Guest
- Problem: Is the price reasonable?
- Location: San Francisco

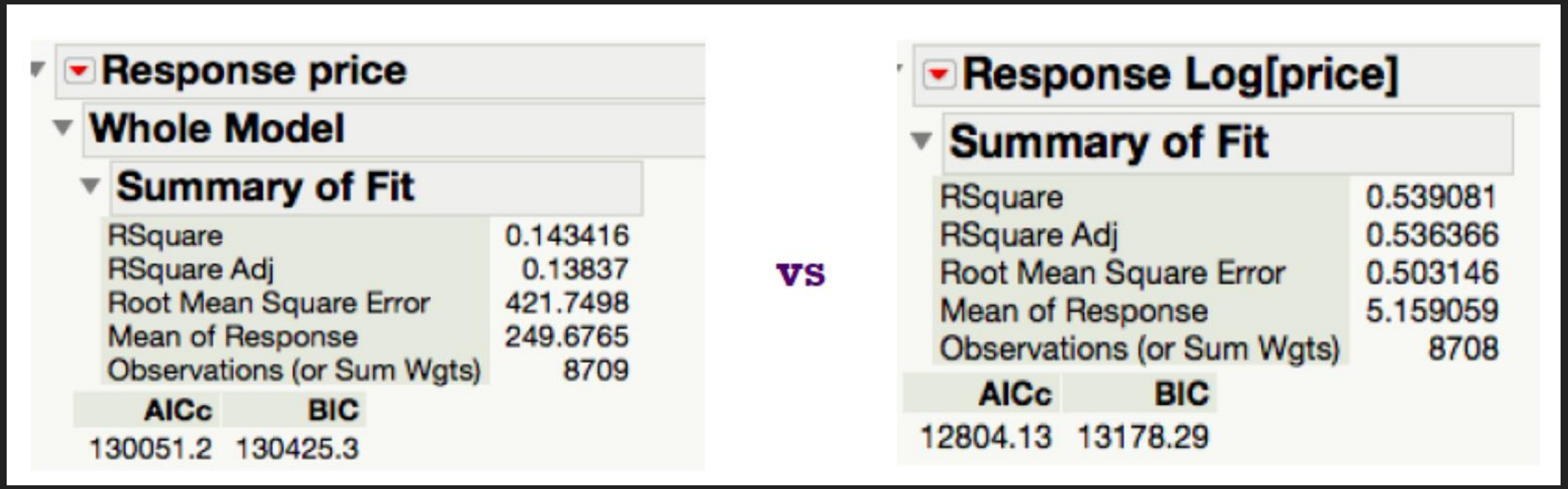
In-Sample Modeling

Data Overview

- "listing.csv": 8720 listings and 95 variables.
- Target Variable: "*price*"
- 33 predictors are selected.

Model Comparison

Linear Regression vs Semi-log Regression



Response price

Whole Model

Summary of Fit

RSquare	0.143416
RSquare Adj	0.13837
Root Mean Square Error	421.7498
Mean of Response	249.6765
Observations (or Sum Wgts)	8709

AICc	BIC
130051.2	130425.3

vs

Response Log[price]

Summary of Fit

RSquare	0.539081
RSquare Adj	0.536366
Root Mean Square Error	0.503146
Mean of Response	5.159059
Observations (or Sum Wgts)	8708

AICc	BIC
12804.13	13178.29

Factor Analysis

- Several predictors have similar meanings.
- Factor analysis particular useful when predictors are highly correlated.

Clustering

- Group similar records into the same cluster (useful in later classification algorithm).
- Find outliers to ensure model accuracy.

Final model features & coefficients

Summary of Fit

RSquare	0.538525
RSquare Adj	0.537138
Root Mean Square Error	0.45743
Mean of Response	5.128155
Observations (or Sum Wgts)	8010

AICc	BIC
10228.79	10410.31

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.6076363	0.234367	11.13	<.0001*
host_response_time[a few days or more]	0.078334	0.041978	1.87	0.0621
host_response_time[N/A]	0.2049233	0.016141	12.70	<.0001*
host_response_time[within a day]	-0.015854	0.018256	-0.87	0.3852
host_response_time[within a few hours]	-0.038932	0.015729	-2.48	0.0133*
host_is_superhost[f]	-0.089616	0.015249	-5.88	<.0001*
host_identity_verified[f]	0.0237336	0.012213	1.94	0.0520
room_type[Entire home/apt]	1.141381	0.037536	30.41	<.0001*
room_type[Private room]	0.5698977	0.037204	15.32	<.0001*
bathrooms	0.155993	0.013081	11.93	<.0001*
bedrooms	0.1845283	0.010899	16.93	<.0001*
beds	0.0788507	0.009406	8.38	<.0001*
security_deposit	1.6821e-5	1.975e-5	0.85	0.3945
guests_included	0.0058403	0.005822	1.00	0.3158
availability_365	0.0005147	4.178e-5	12.32	<.0001*
instant_bookable[f]	-0.002534	0.014027	-0.18	0.8566
reviews_per_month	-0.049811	0.004042	-12.32	<.0001*
cancellation_policy[flexible]	-0.468438	0.188179	-2.49	0.0128*
cancellation_policy[moderate]	-0.552193	0.188131	-2.94	0.0033*
cancellation_policy[strict]	-0.535849	0.187956	-2.85	0.0044*
cancellation_policy[super_strict_30]	-1.202577	0.22172	-5.42	<.0001*
review_scores_rating	0.0050617	0.001842	2.75	0.0060*
review_scores_cleanliness	0.07296	0.011707	6.23	<.0001*
review_scores_location	0.1152075	0.011071	10.41	<.0001*
review_scores_value	-0.072432	0.013074	-5.54	<.0001*

Recall

- The constructed regression model helps us to predict a reasonable price, given guests' preferences.
- We will apply multiple machine learning algorithms to better classify if a listed price is reasonable.

Out-Sample Modeling

Pre-processing: Target

Target:

{“underpriced”, “reasonable”, “overpriced”}

- Step 1: K-Means Clustering
  - Optimal CCC:

    29 Clusters
  - “**price_benchmark**”:

    mean price in each cluster

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	10	92.2042	
K-Means Clustering	11	97.1637	
K-Means Clustering	12	104.144	
K-Means Clustering	13	120.493	
K-Means Clustering	14	118.861	
K-Means Clustering	15	124.404	
K-Means Clustering	16	146.099	
K-Means Clustering	17	137.174	
K-Means Clustering	18	140.085	
K-Means Clustering	19	152.163	
K-Means Clustering	20	153.969	
K-Means Clustering	21	153.363	
K-Means Clustering	22	134.079	
K-Means Clustering	23	151.401	
K-Means Clustering	24	159.832	
K-Means Clustering	25	158.64	
K-Means Clustering	26	148.694	
K-Means Clustering	27	163.027	
K-Means Clustering	28	150.166	
K-Means Clustering	29	165.604	Optimal CCC
K-Means Clustering	30	146.585	

Pre-processing: Target

Target:

{“underpriced”, “reasonable”, “overpriced”}

- Step 2: Percentage Difference =

$$\frac{\text{“price”} - \text{“price\_benchmark”}}{\text{“price\_benchmark”}}$$

Quantiles

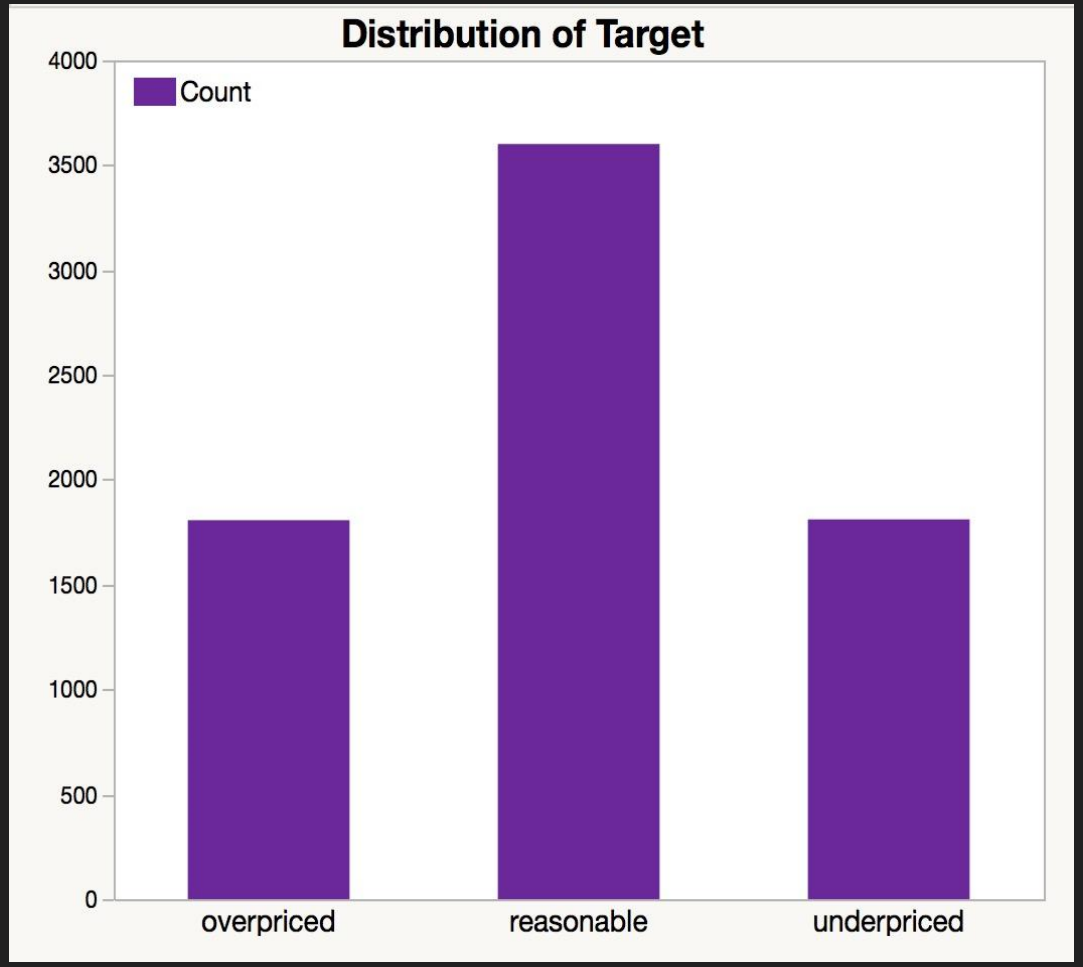
100.0%	maximum	13.142620232173
99.5%		3.9985802591282
97.5%		1.7020447906524
90.0%		0.6817951435285
75.0%	quartile	0.1779424956329
50.0%	median	-0.154223051797
25.0%	quartile	-0.416705552963
10.0%		-0.562529164722
2.5%		-0.697287697772
0.5%		-0.789129632558
0.0%	minimum	-0.94728240814

Pre-processing: Target

Target:

{“underpriced”, “reasonable”, “overpriced”}

- Step 3: Create Target
  - “Underpriced”: lower than 1st quantile
  - “Reasonable”: between 1st and 3rd quantile
  - “Overpriced”: higher than 3rd quantile



Pre-processing: Text

“review2.csv”

- Remove Reviews:
  - ✓ Not in English
  - ✓ Automated: "This is an automated posting"
  - ✓ Invalid
- Combine reviews associated with the same “listing id” and merge with listing records.
- Lowering letters, stemming, removing punctuations, stopwords and numbers.

Text Mining

Comparing Frequent Terms

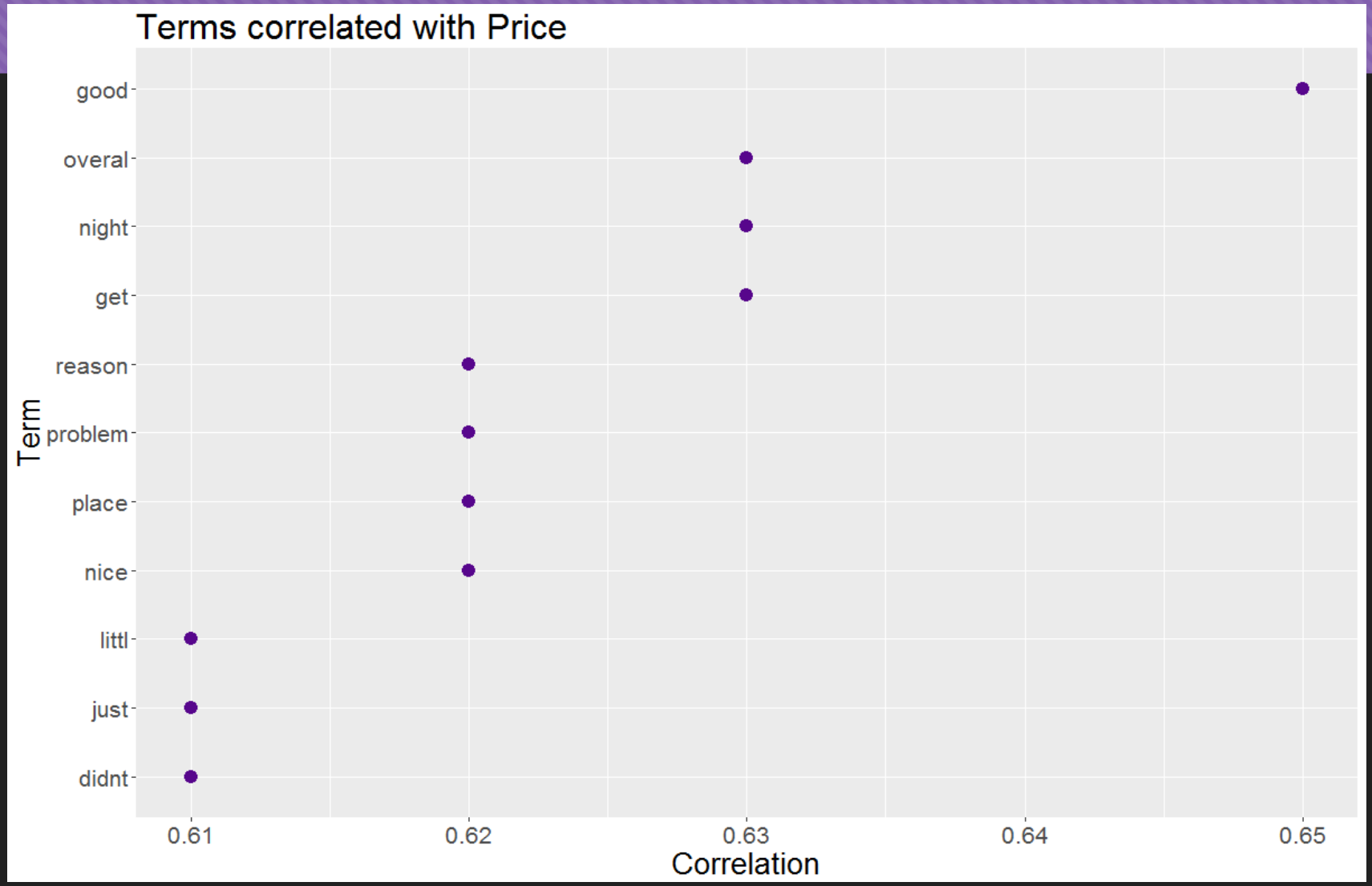
Reasonable

Overpriced

Underpriced

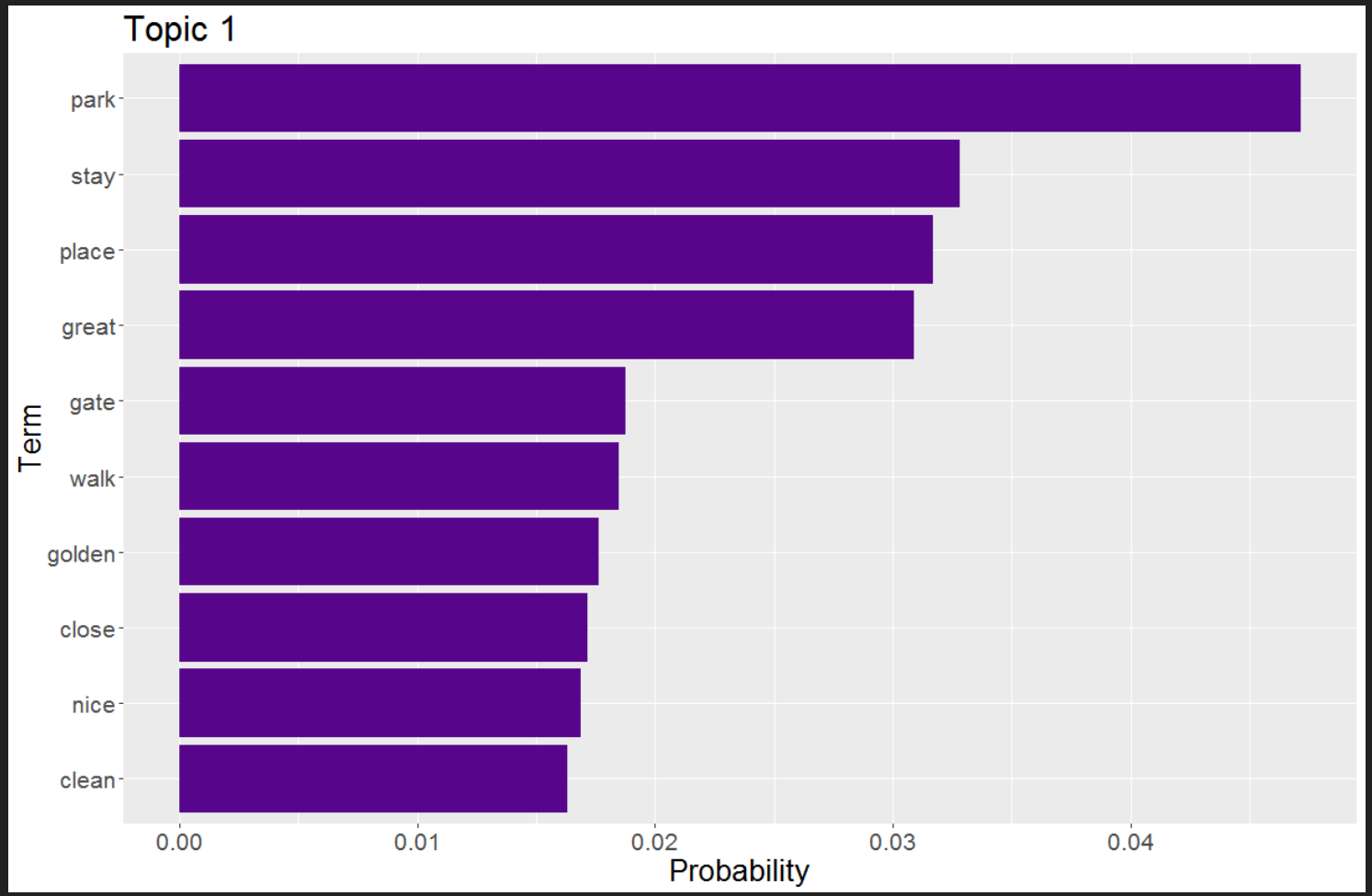
Terms Correlated in “Reasonable”

“Good” is the highest correlated word, which indicates that customers are fairly satisfied.



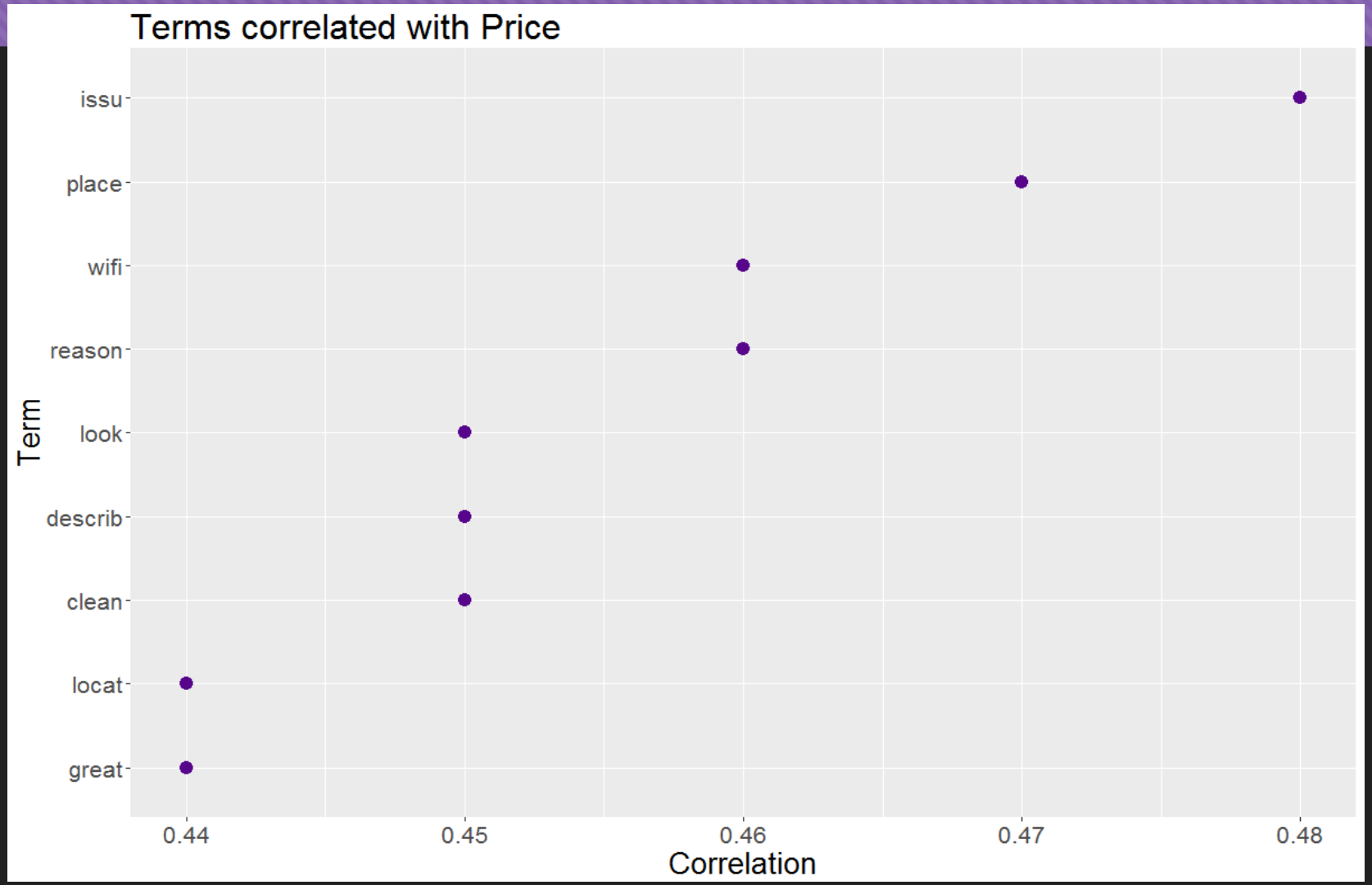
Top Topic in “Reasonable”

About the distance to the scenic spots.



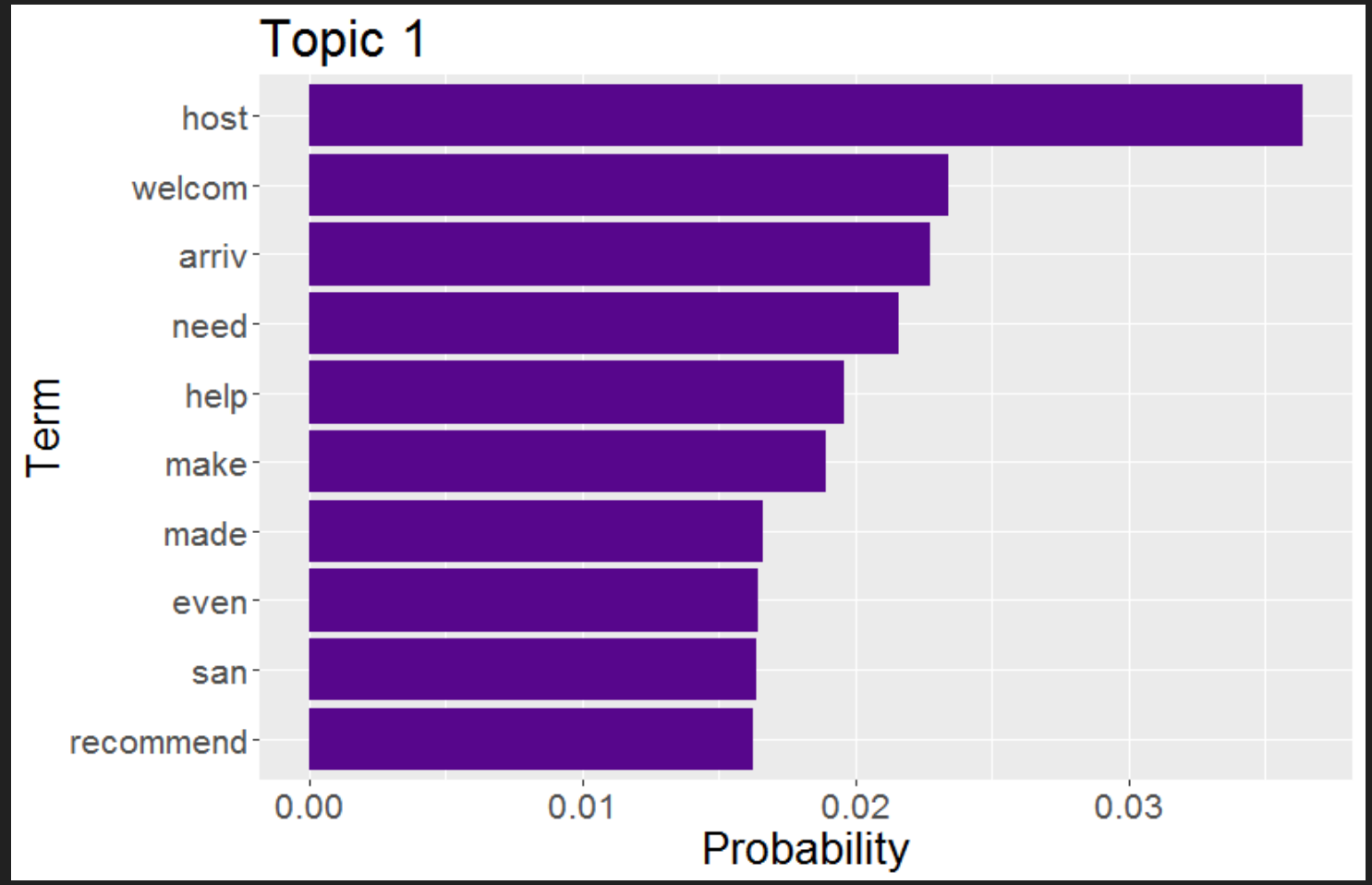
Terms Correlated in “Overpriced”

“Issue” is the highest correlated word with “price”, which may cause the disagreement on price.



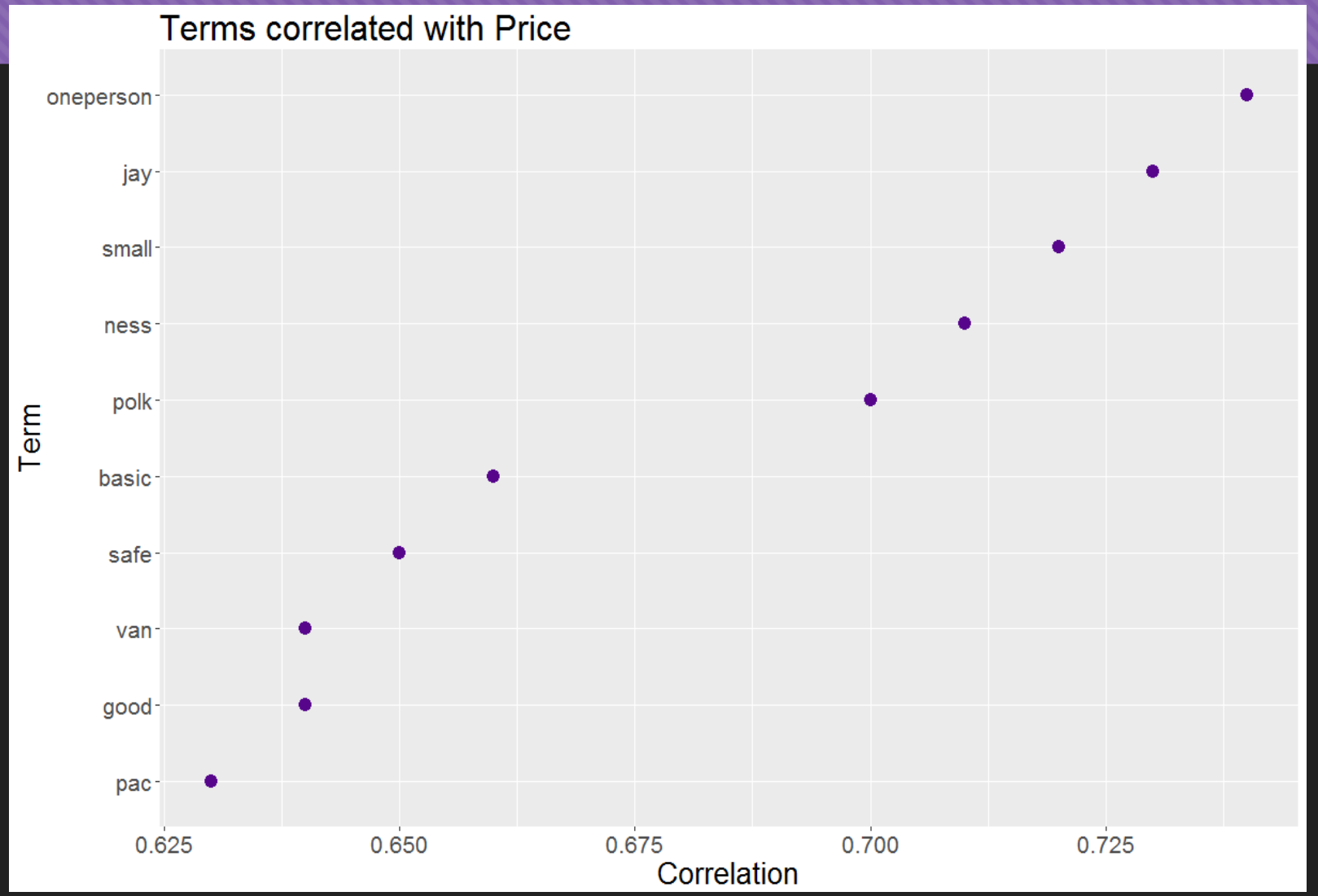
Top Topic in “Overpriced”

About the customer services (could be the reason for the extra fees).



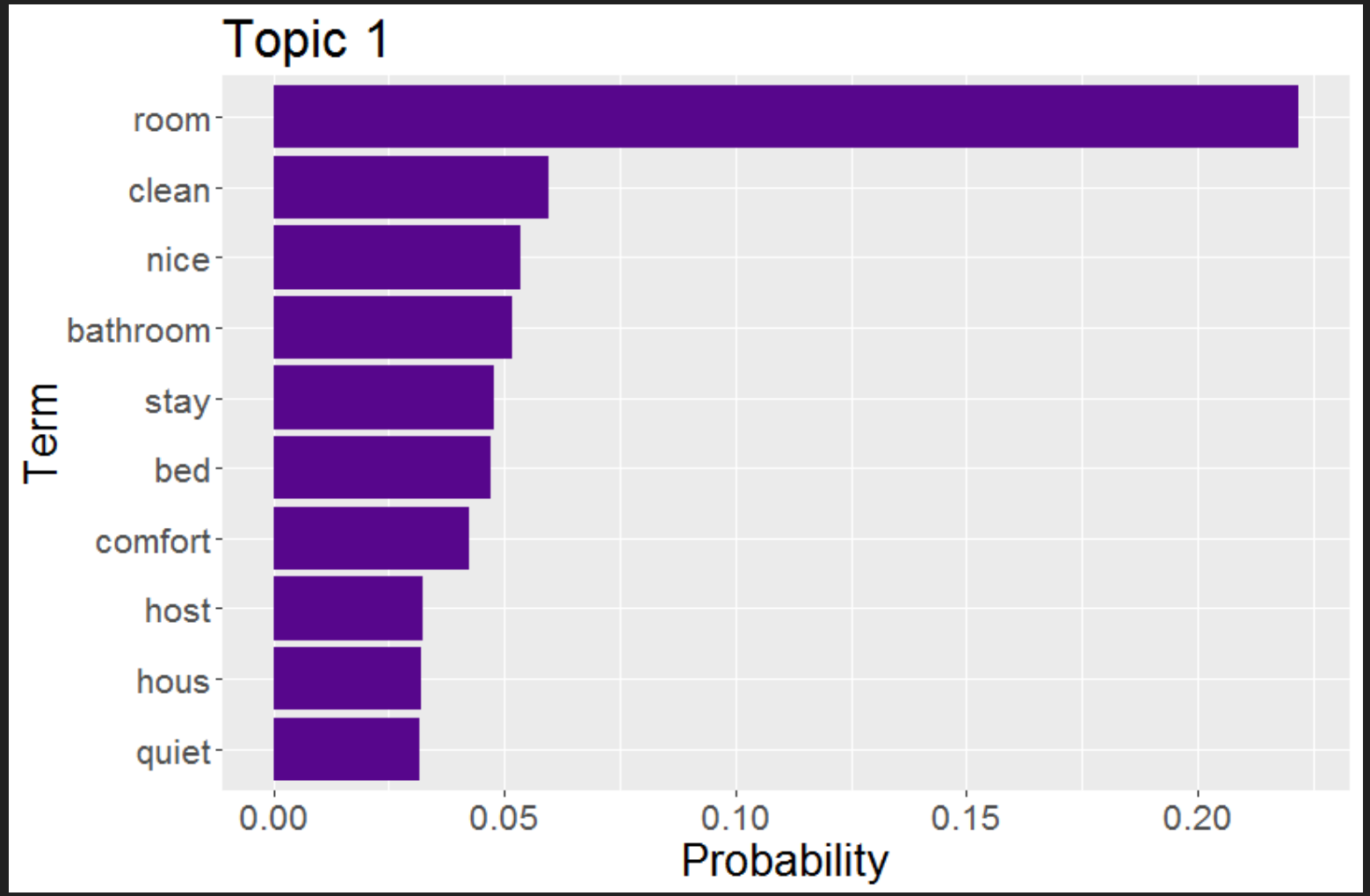
Terms Correlated in “Underpriced”

Terms such as “oneperson”, “small” and “basic” are likely to be related with a single person’s room, which has small space and basic amenity.



Top Topic in “Underpriced”

Terms such as “clean”, “nice” and “quiet” are positive words, which may show that hosts’ listed prices are below customers’ expected range.



Conclusion for Text Mining

- The frequent terms in these 3 classes are not significantly different (many positive words).
- Sentiment analysis could be used to predict rating, instead of prices (our dependent variable).
- Therefore, use rating for the price prediction could be enough.

Multiclass Classification

Machine Learning

Variables

- Target:

  {“underpriced”, “reasonable”, “overpriced”}
- 10 Nominal Variables
- 19 Continuous Variables

Methods

- Logistic Regression
- Bootstrap Forest
- Neural Networks

Logistic Regression

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3729	0.2826	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6166	0.5086	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6520	0.7484	$\sum$ -Log(ρ[j])/n
RMSE	0.4628	0.4735	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.4031	0.4095	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.2863	0.3165	$\sum$ (ρ[j]≠ρMax)/n
N	5715	1504	n

Bootstrap Forest

- Parameters by default

Number of Trees in the Forest	100
Number of Terms Sampled per Split:	7
Bootstrap Sample Rate	1
Minimum Splits per Tree:	10
Maximum Splits per Tree	2000
Minimum Size Split:	7

Measure	Training	Validation
Entropy RSquare	0.4469	0.2134
Generalized RSquare	0.6916	0.4103
Mean -Log p	0.5751	0.8206
RMSE	0.4403	0.5447
Mean Abs Dev	0.4155	0.5167
Misclassification Rate	0.1601	0.3777
N	5715	1504

Neural Networks

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	3	0	0
Second	0	0	0

Validation

target

Measures	Value
Generalized RSquare	0.299515
Entropy RSquare	0.1458293
RMSE	0.5604209
Mean Abs Dev	0.5277727
Misclassification Rate	0.4308511
-LogLikelihood	1340.2208
Sum Freq	1504

Confusion Matrix

Actual	Predicted Count		
target	overpriced	reasonable	underpriced
overpriced	118	226	28
reasonable	98	565	81
underpriced	28	187	173

Model Comparison

Measures of Fit for target

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Training	Fit Nominal Logistic		0.3729	0.6166	0.652	0.4628	0.4031	0.2863	5715
Training	Bootstrap Forest		0.4469	0.6916	0.5751	0.4403	0.4155	0.1601	5715
Training	Neural		0.1996	0.3883	0.8322	0.5421	0.5108	0.4024	5715
Training	Neural		0.2269	0.4299	0.8039	0.5327	0.4948	0.3811	5715
Training	Bootstrap Forest		0.2570	0.4731	0.7726	0.5321	0.5146	0.3340	5715
Validation	Fit Nominal Logistic		0.2259	0.4291	0.8076	0.4735	0.4095	0.3165	1504
Validation	Bootstrap Forest		0.2134	0.4103	0.8206	0.5447	0.5167	0.3777	1504
Validation	Neural		0.1458	0.2995	0.8911	0.5604	0.5278	0.4309	1504
Validation	Neural		0.1802	0.3578	0.8552	0.5464	0.5064	0.3969	1504
Validation	Bootstrap Forest		0.1842	0.3643	0.8511	0.5621	0.5442	0.3983	1504

Discussion

Discussion

- Generally, sentiment analysis is used to predict ratings, so we cannot apply "review words" and ratings simultaneously to make classifications
- The predictive power of sentiment analysis and ratings are "overlapped".
- Hierarchical clustering seems better at searching for outliers; k-means cluster can group similar records to make labels for classification.

NYU | STERN

airbnb

THANK YOU