Predicting AirBnB Listing Price within New York City's Boroughs

EAS 503 | Group 7 | Gary Yu, Georgious Aland Feltama, Peter Pranata

Agenda

Introduction	Slide 3 - 4
Research Framework	Slide 5
Preliminary Analysis	Slide 6 - 9
Model Fitting	Slide 10 - 13
Model Validation	Slide 14 - 17
Results	Slide 18 - 19

Presence of Airbnb in NYC

Bronx
Manhattan
Queens
Staten Island
Brooklyn

NYC is the most visited city in the U.S

New York City has the 3rd largest active listings in the world

Approximately 48,000 active listings

1,143,036 reviews up to date

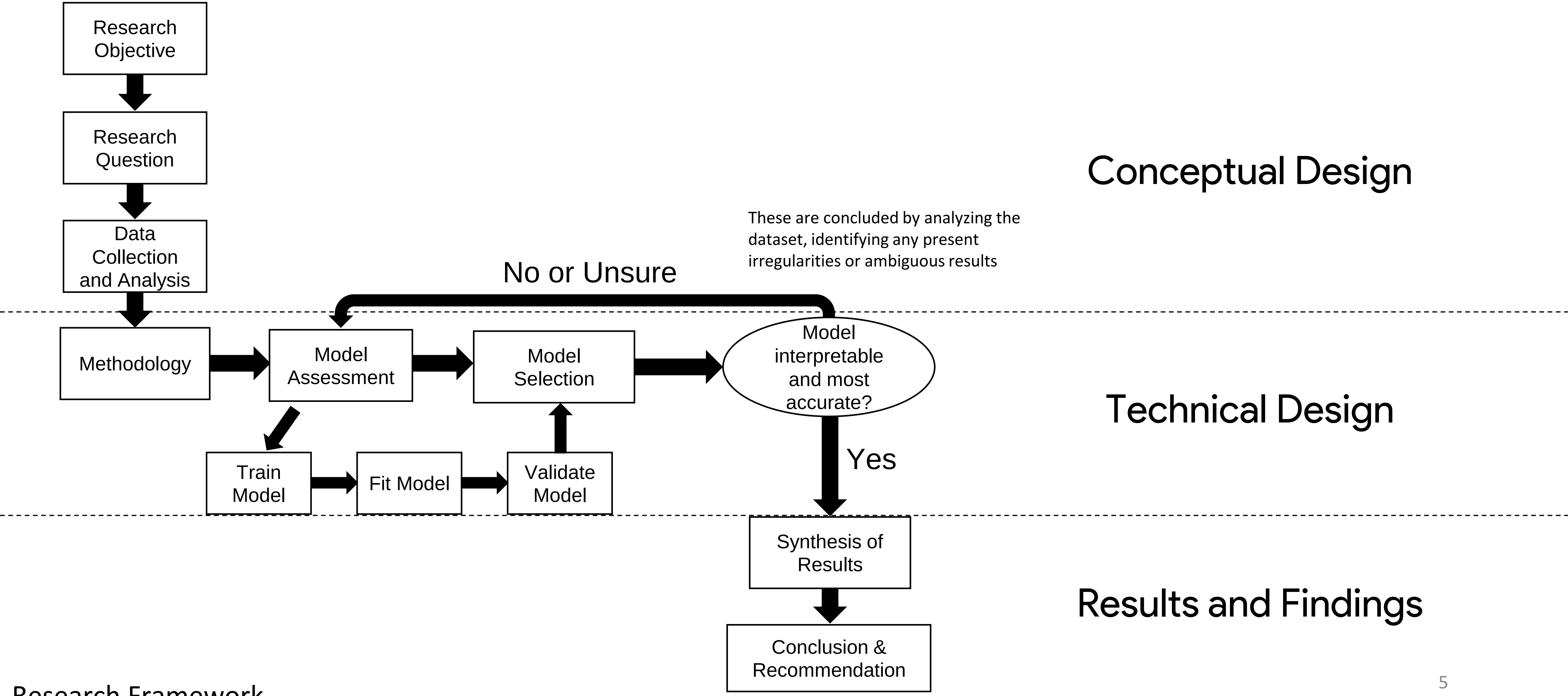
Motivation and Objective(s)

Airbnb pricing varies a lot; sometimes its cheaper than hotel, other times no

What **factors** within an **Airbnb** listing **dictates** its **nightly price** within the respective **neighborhood groups**?

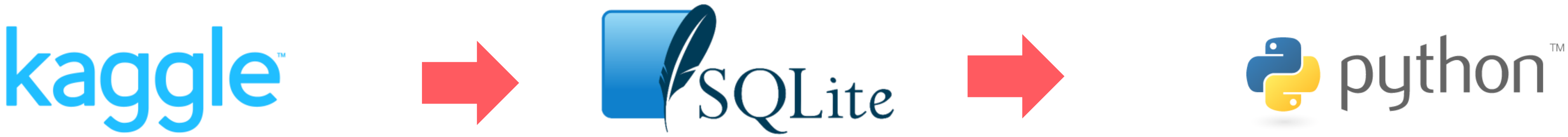
- Identify the relationship and differences between neighborhood groups
- Suggests which factors to improve if you want to increase listing price
- Provide insight on potential gaps in listing locations and business opportunities

Research Framework



These are concluded by analyzing the dataset, identifying any present irregularities or ambiguous results

Data Source



New York City Airbnb Open Data
https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Approximately 48,000 rows and 16 columns

Data Preparation

Data Manipulation and Model Assessment

- sqlite3
- pandas
- numpy

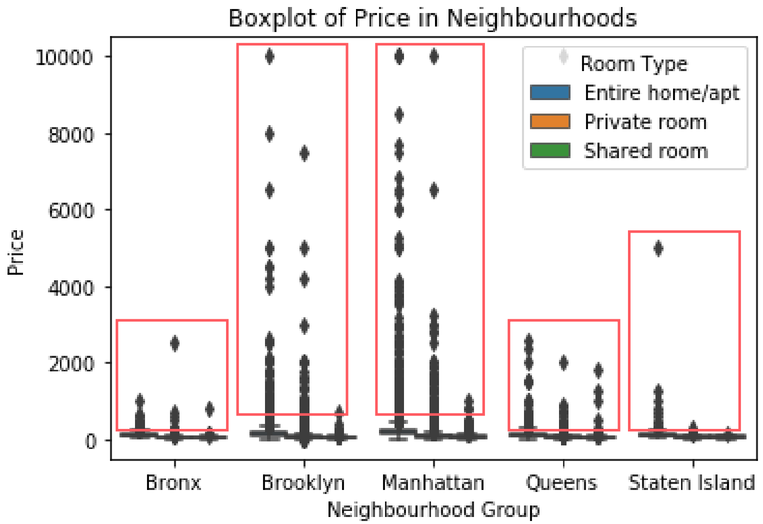
- seaborn
- matplotlib
- sklearn

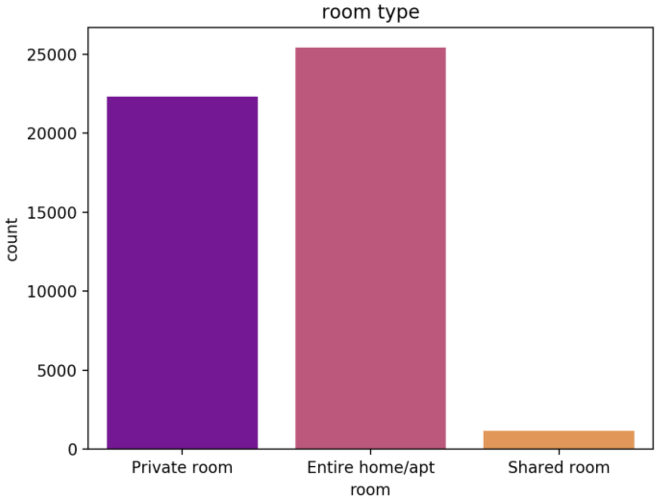
Initial Data Exploration

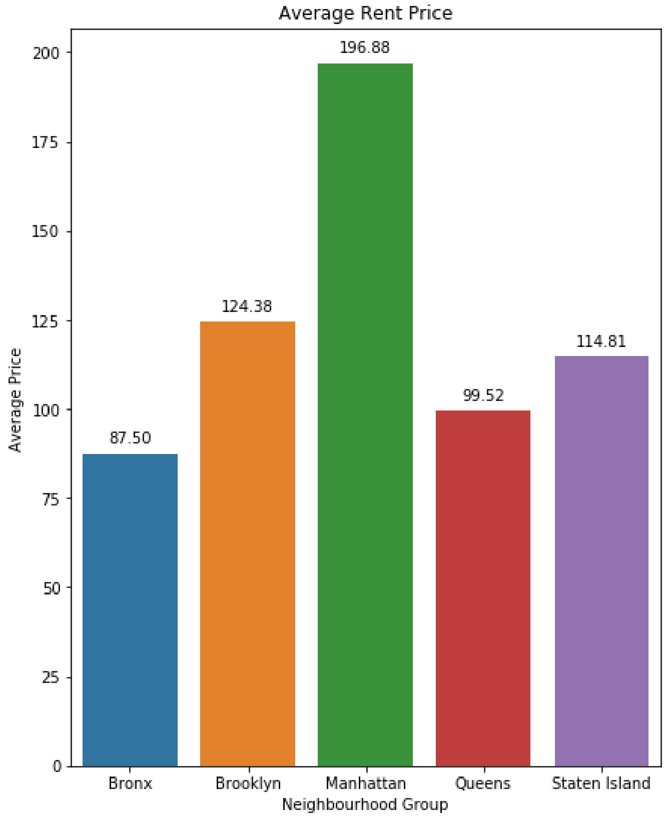
Converted data types to its appropriate format

Removed rows with NA values

Generated basic plots to understand data





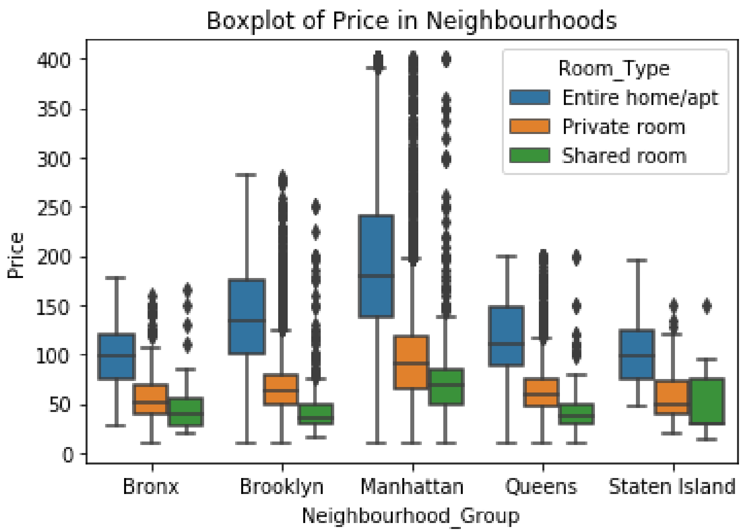


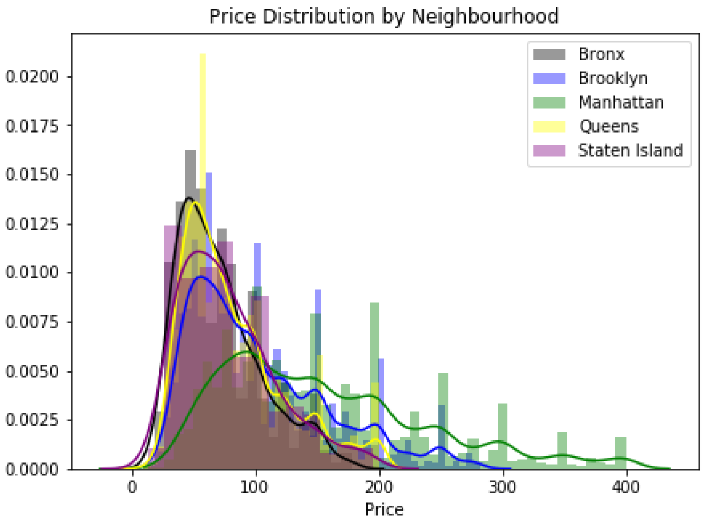
Outliers & Extreme Values

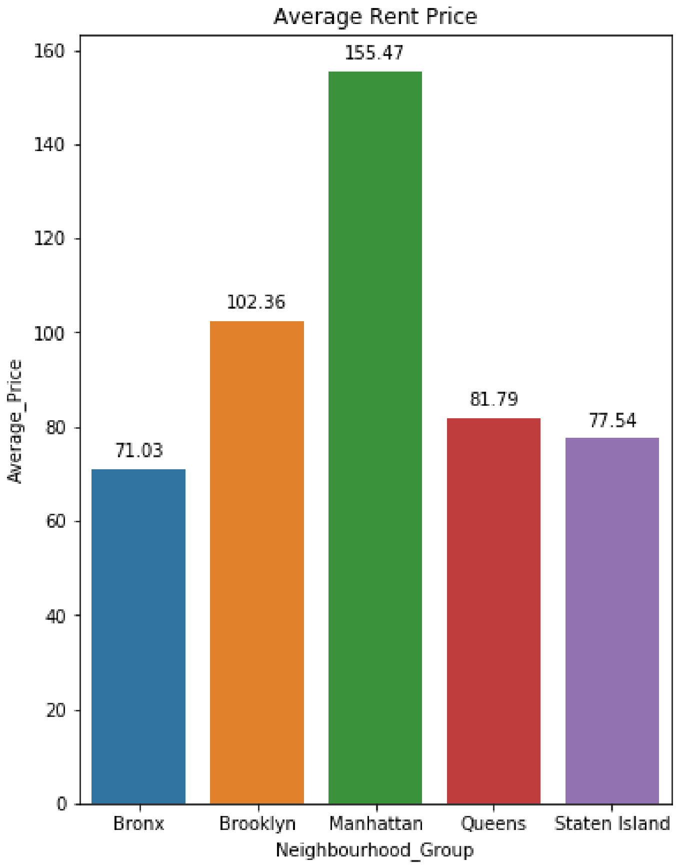
Extreme values present for all neighborhood groups

Sample 99th percentile of each neighborhood

Removed rows containing a price of 0 (system error)







k-Level Categorical Features

‘Neighborhood_Groups’ = 5 Levels

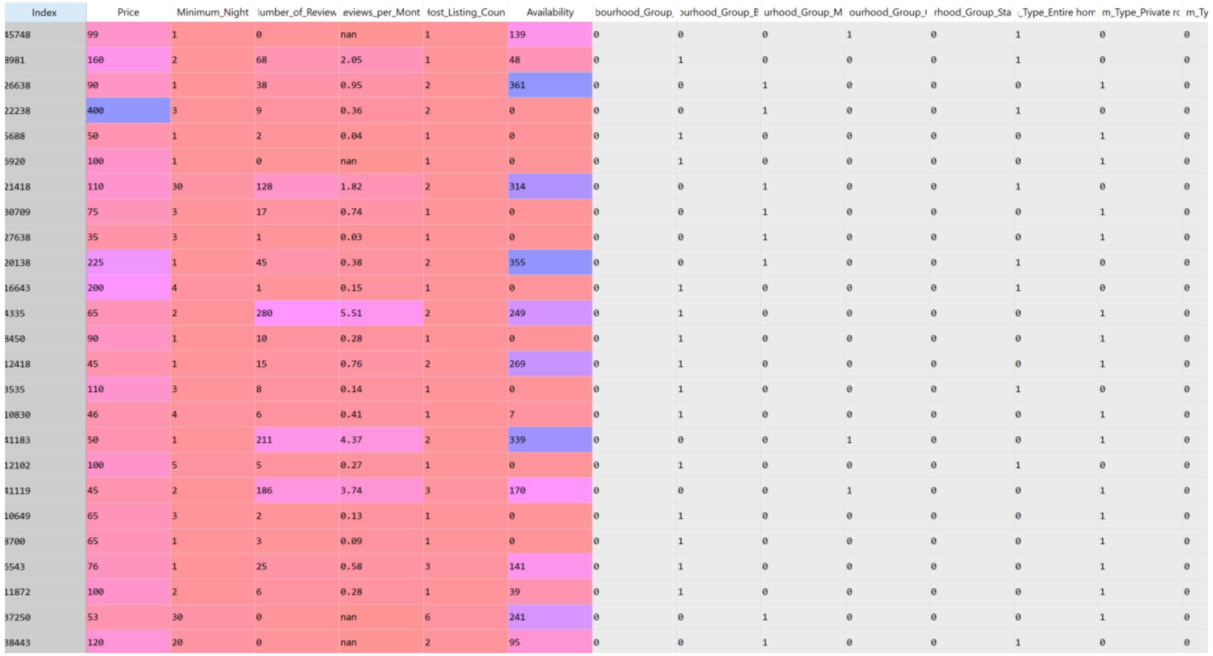
‘Room_Type’ = 3 Levels

“One-Hot Encoding”

- Converts multiple level categories into dummy columns
- Binary response (1 or 0)

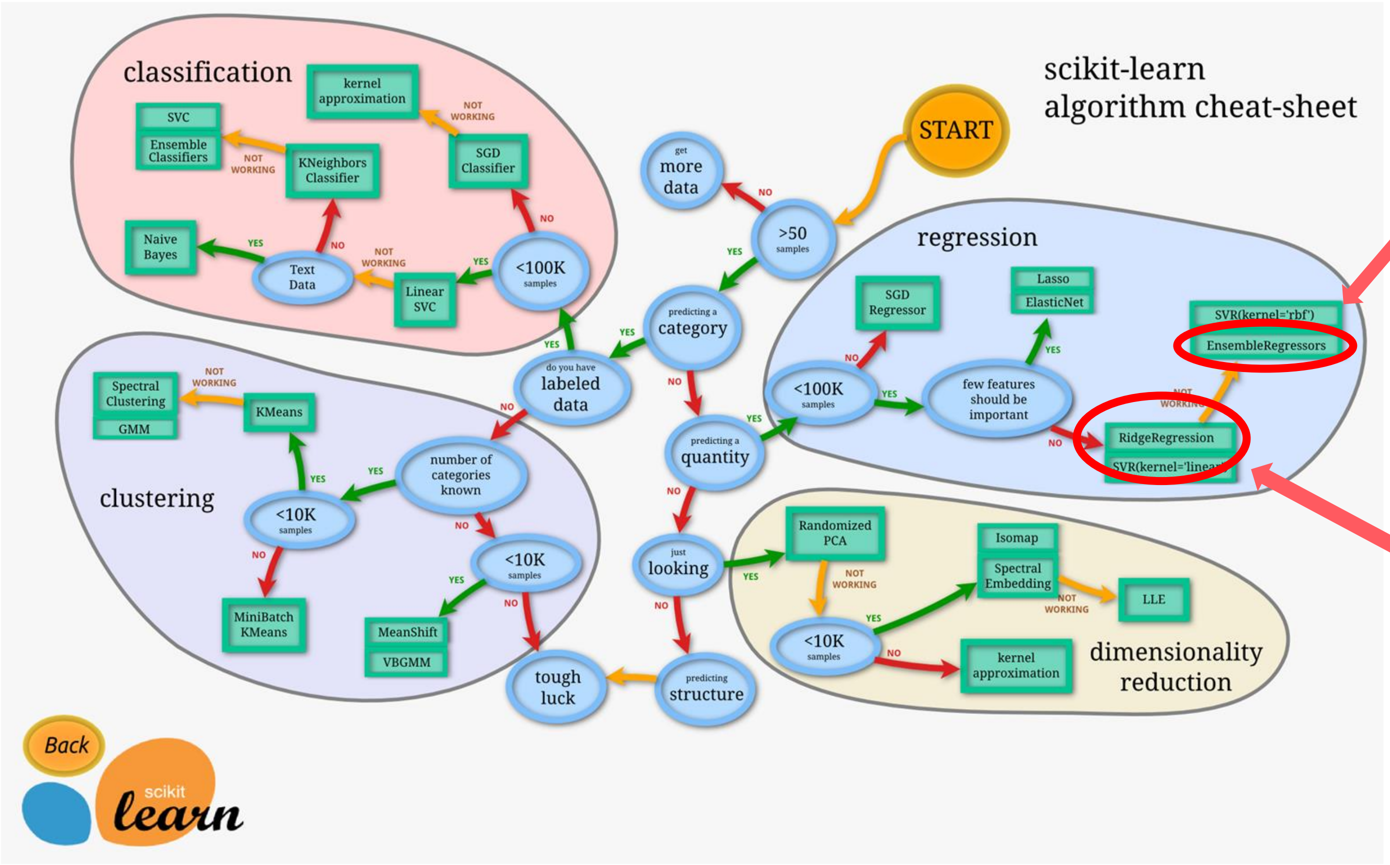
Features

Dummy Features



Index	Price	Minimum_Night	lumber_of_Review	eviews_per_Mont	lost_Listing_Coun	Availability	bourhood_Group_	ourhood_Group_B	urhood_Group_M	ourhood_Group_	rhood_Group_Sta	_Type_Entire hom	m_Type_Private rc	m_Ty
45748	99	1	0	nan	1	139	0	0	0	1	0	1	0	0
3981	160	2	68	2.05	1	48	0	1	0	0	0	1	0	0
26638	90	1	38	0.95	2	361	0	0	1	0	0	0	1	0
22238	400	3	9	0.36	2	0	0	0	1	0	0	1	0	0
5688	50	1	2	0.04	1	0	0	1	0	0	0	0	1	0
5920	100	1	0	nan	1	0	0	1	0	0	0	0	1	0
21418	110	30	128	1.82	2	314	0	0	1	0	0	1	0	0
30709	75	3	17	0.74	1	0	0	0	1	0	0	0	1	0
27638	35	3	1	0.03	1	0	0	0	1	0	0	0	1	0
20138	225	1	45	0.38	2	355	0	0	1	0	0	1	0	0
16643	200	4	1	0.15	1	0	0	1	0	0	0	1	0	0
4335	65	2	280	5.51	2	249	0	1	0	0	0	0	1	0
3450	90	1	10	0.28	1	0	0	1	0	0	0	0	1	0
12418	45	1	15	0.76	2	269	0	1	0	0	0	0	1	0
3535	110	3	8	0.14	1	0	0	1	0	0	0	1	0	0
10830	46	4	6	0.41	1	7	0	1	0	0	0	0	1	0
41183	50	1	211	4.37	2	339	0	0	0	1	0	0	1	0
12102	100	5	5	0.27	1	0	0	1	0	0	0	1	0	0
41119	45	2	186	3.74	3	170	0	0	0	1	0	0	1	0
10649	65	3	2	0.13	1	0	0	1	0	0	0	0	1	0
3700	65	1	3	0.09	1	0	0	1	0	0	0	0	1	0
5543	76	1	25	0.58	3	141	0	1	0	0	0	0	1	0
11872	100	2	6	0.28	1	39	0	1	0	0	0	0	1	0
37250	53	30	0	nan	6	241	0	0	1	0	0	0	1	0
38443	120	20	0	nan	2	95	0	0	1	0	0	1	0	0

Model Exploration



Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Train Model using **Random Forest**

↓

Cross-Validate & Evaluate

↓

Synthesize Results

Train data using **Ridge Regression**

↓

Cross-Validate & Evaluate

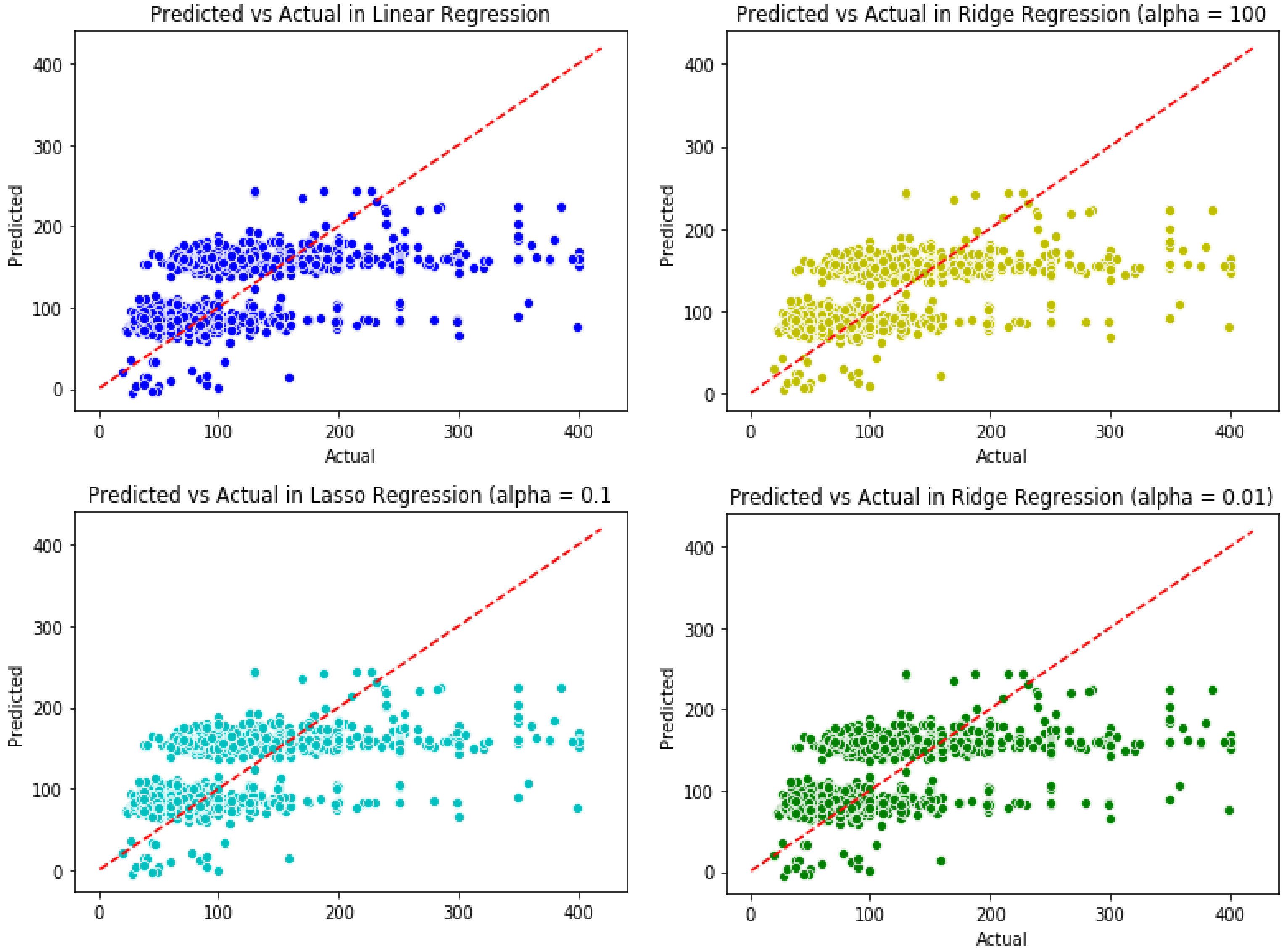
↓

Synthesize Results

Regression Models

Sampled 10% of entire data (~4,000 rows)

75% Training set
25% Testing set



Predicted vs Actual in Linear Regression

Predicted vs Actual in Ridge Regression (alpha = 100

Predicted vs Actual in Lasso Regression (alpha = 0.1

Predicted vs Actual in Ridge Regression (alpha = 0.01)

Accuracy and Validation

	Accuracy	R-sq
Linear	55.740437	0.326937
Ridge (a=0.01)	55.740380	0.326938
Ridge (a=100)	55.138659	0.325496
Lasso	55.700691	0.326868

No significant differences!

Very similar capability and accuracy

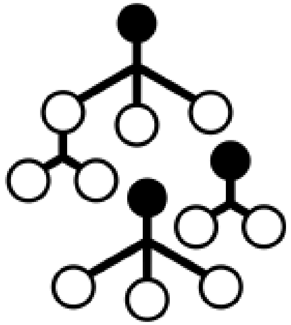
Next: Ensemble Regressor Methods

Training on Random Forest

Due to computing power constraint

Sampled 10% of entire data (~4,000 rows)

75% Training set
25% Testing set



Ran model using standard & automatic predictors/parameters

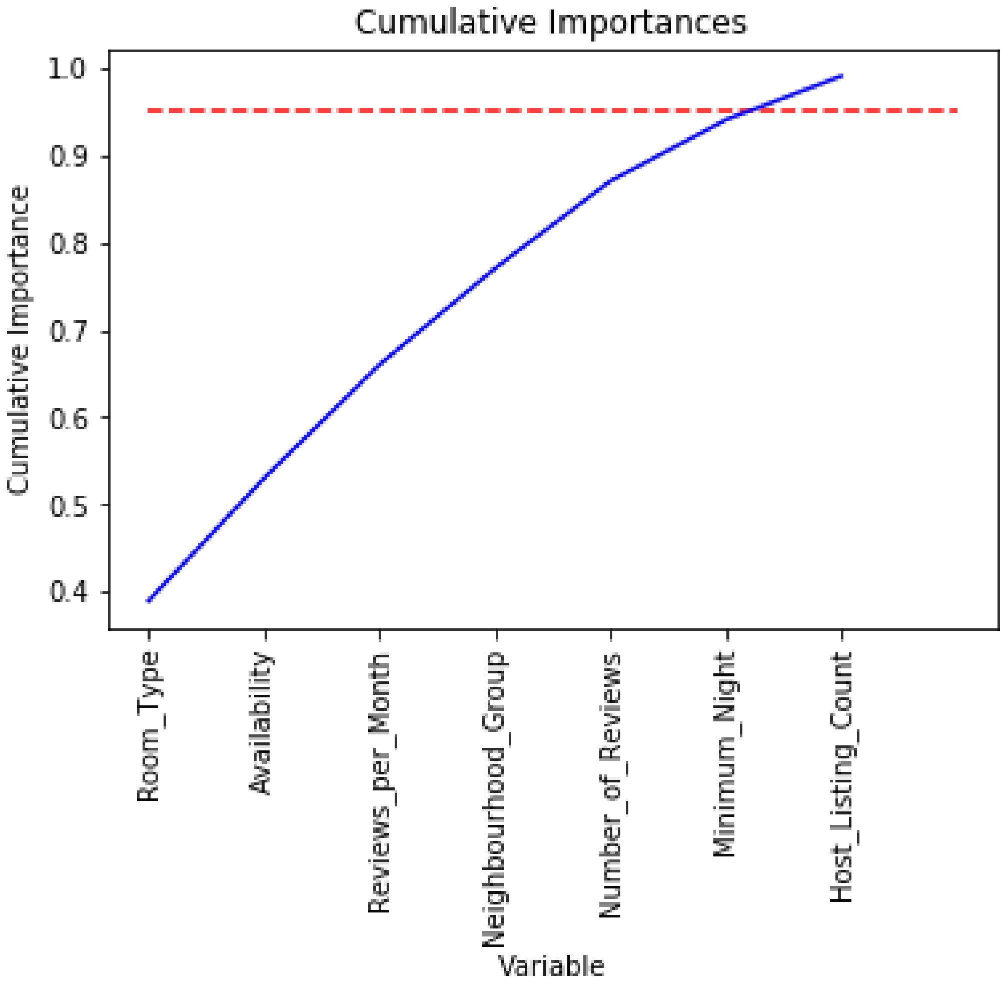
- 1000 trees

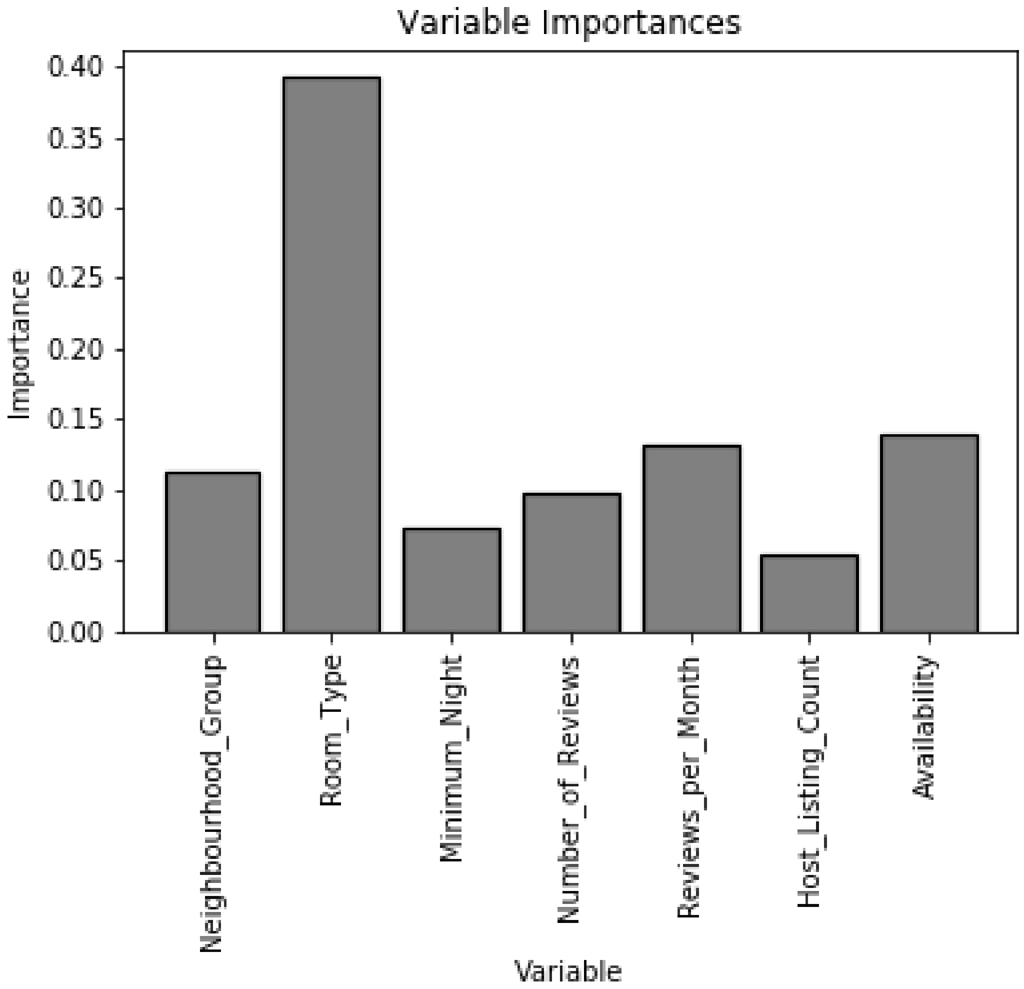
Accuracy: 62.73%

- Accuracy is calculated by 100 - MAPE

We'll try to improve it

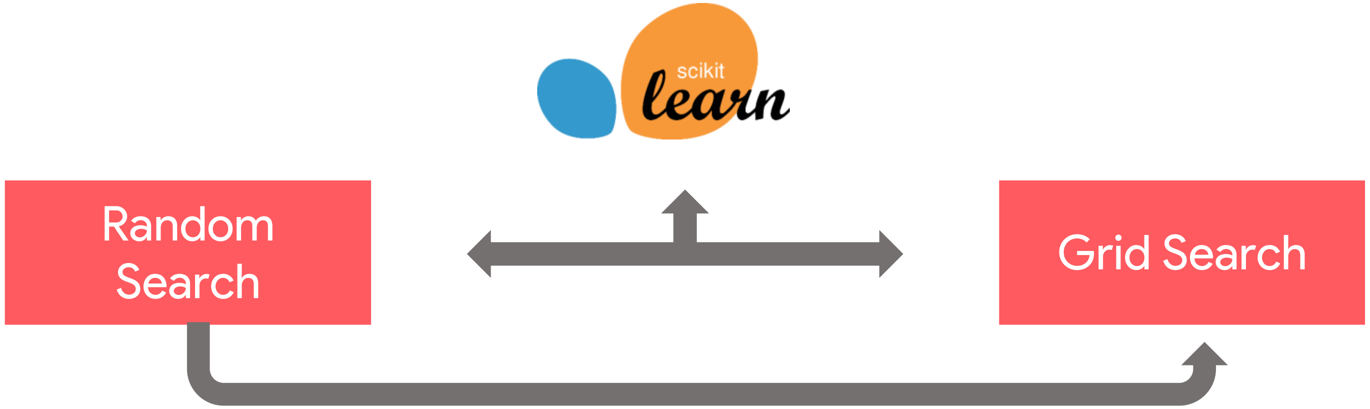
Variable Importance in RF





Hyperparameter Tuning

"It is a parameter whose value is set before the learning process begins"



Random Search

Grid Search

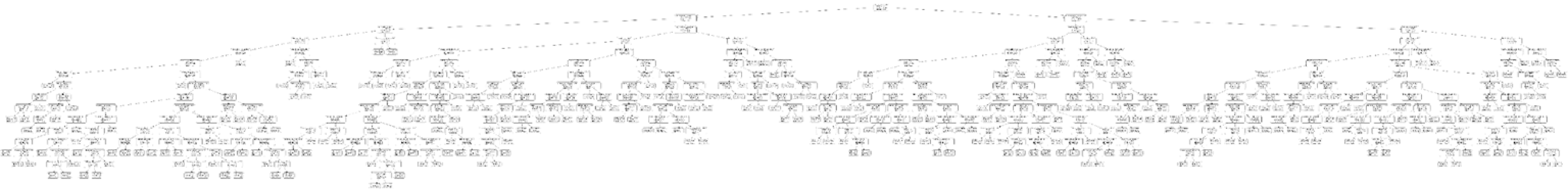
Train model with random combinations of the parameters and see which iterations/set is the best

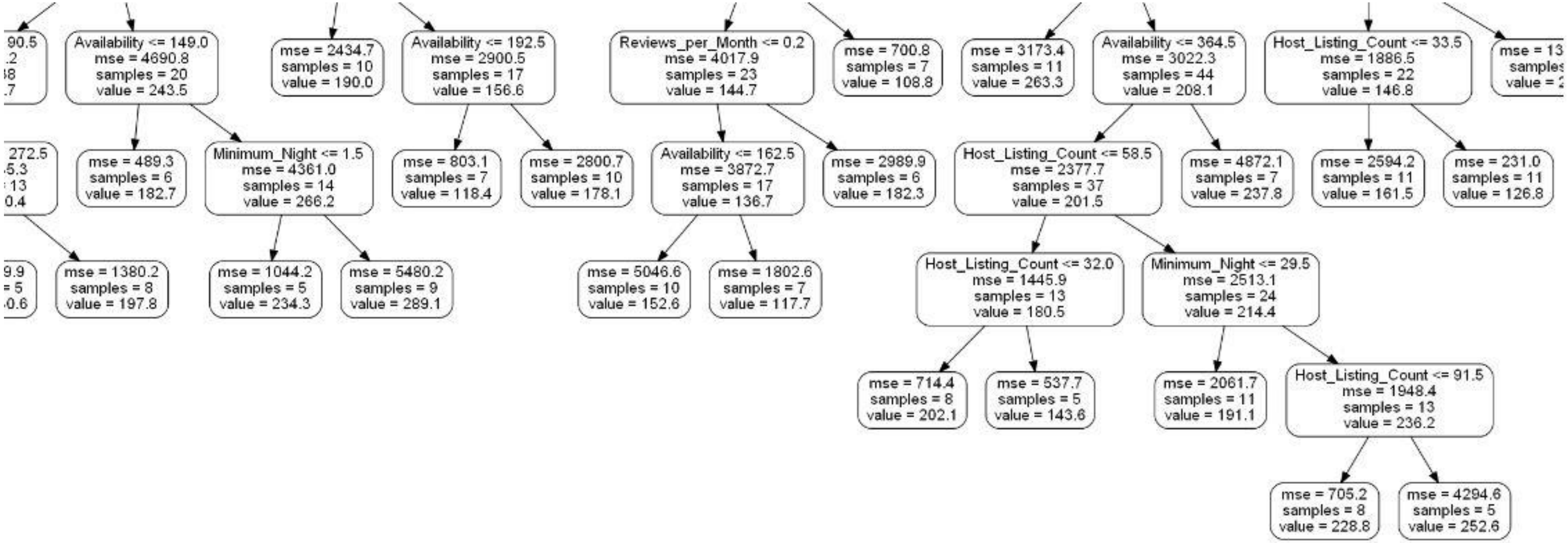
Utilize Random Search to narrow down parameters for Grid Search

Using the selected optimized parameters obtained from Random Search, re-train model using every single combination of hyperparameter values

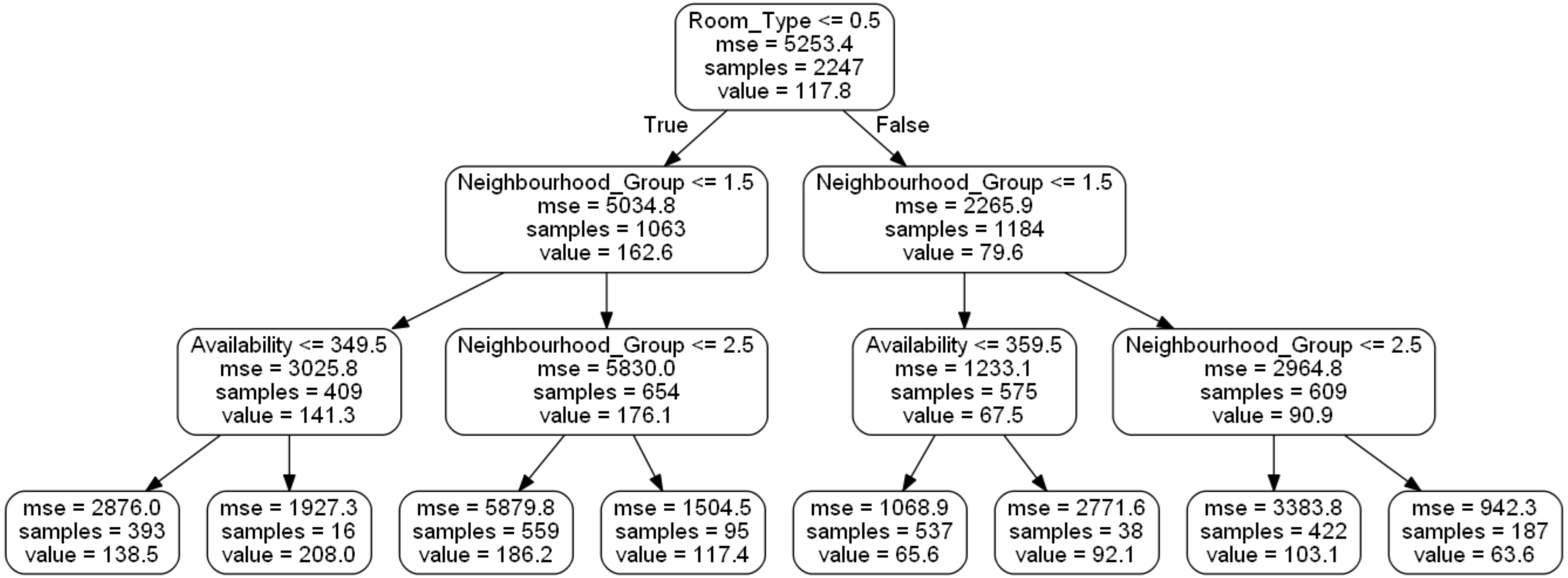
Best Grid RF Accuracy: 64.89%
Improvement of 2.16%

Optimized Tree Diagram





Pruned Tree



Room_Type <= 0.5
mse = 5253.4
samples = 2247
value = 117.8

True

False

Neighbourhood_Group <= 1.5
mse = 5034.8
samples = 1063
value = 162.6

Neighbourhood_Group <= 1.5
mse = 2265.9
samples = 1184
value = 79.6

Availability <= 349.5
mse = 3025.8
samples = 409
value = 141.3

Neighbourhood_Group <= 2.5
mse = 5830.0
samples = 654
value = 176.1

Availability <= 359.5
mse = 1233.1
samples = 575
value = 67.5

Neighbourhood_Group <= 2.5
mse = 2964.8
samples = 609
value = 90.9

mse = 2876.0
samples = 393
value = 138.5

mse = 1927.3
samples = 16
value = 208.0

mse = 5879.8
samples = 559
value = 186.2

mse = 1504.5
samples = 95
value = 117.4

mse = 1068.9
samples = 537
value = 65.6

mse = 2771.6
samples = 38
value = 92.1

mse = 3383.8
samples = 422
value = 103.1

mse = 942.3
samples = 187
value = 63.6

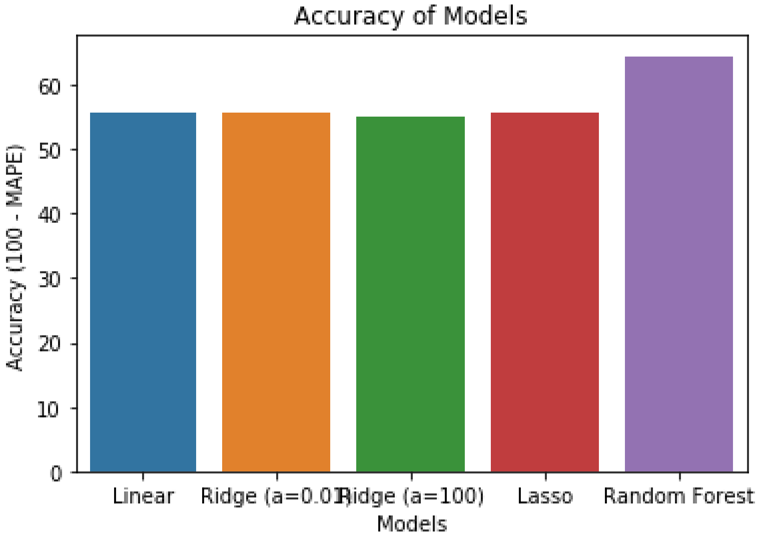
Number of Trees = 10
Max Depth = 3

Accuracy: 62.53%
R-sq: 43.99%

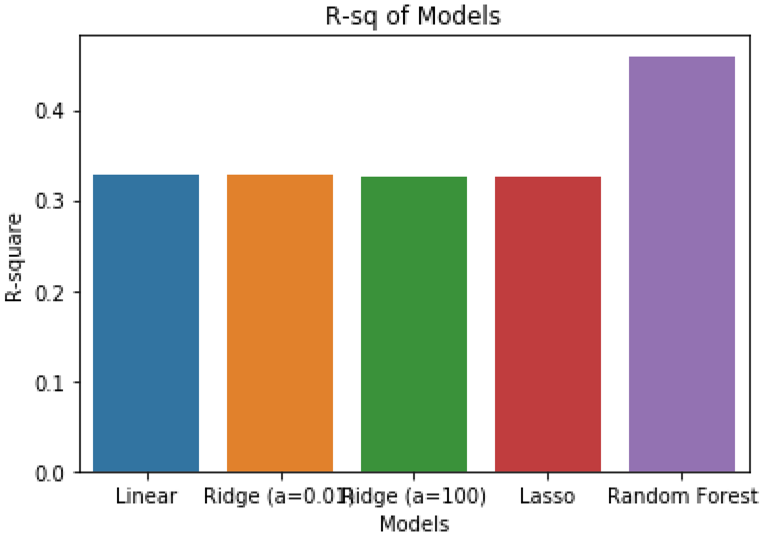
→

Easier to interpret despite the slight loss of accuracy

Synthesis of Results



Accuracy of Models



R-sq of Models

Random Forest yields the best **Accuracy** and **R-sq** compared to the regression models

There is no significant differences between the regression models

Best Model Accuracy & R-sq:

- **Accuracy: 64.89%**
- **R-sq: 45.96%**

Conclusion

More features is needed to increase model accuracy

Room type has the highest impact in determining the listing price of Airbnb in NYC

Hosts does not necessarily increase their price based on higher reviews

- Having more reviews doesn’t mean the listing will be more expensive
- Number of reviews is a small factor which means price is inelastic to reviews

Utilize model to provide pricing suggestions for their customers and business partners
Given the basic features (room type, location, etc), they can predict future outcome