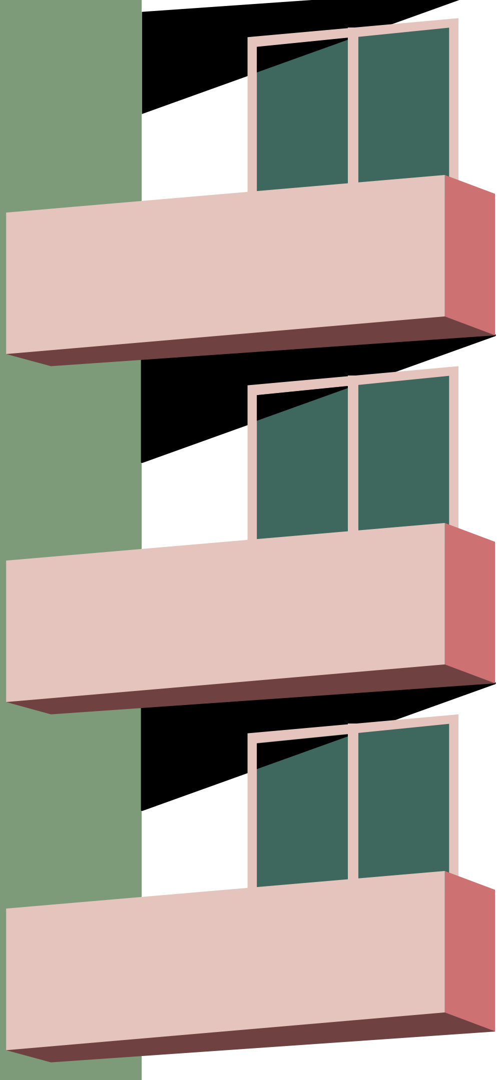
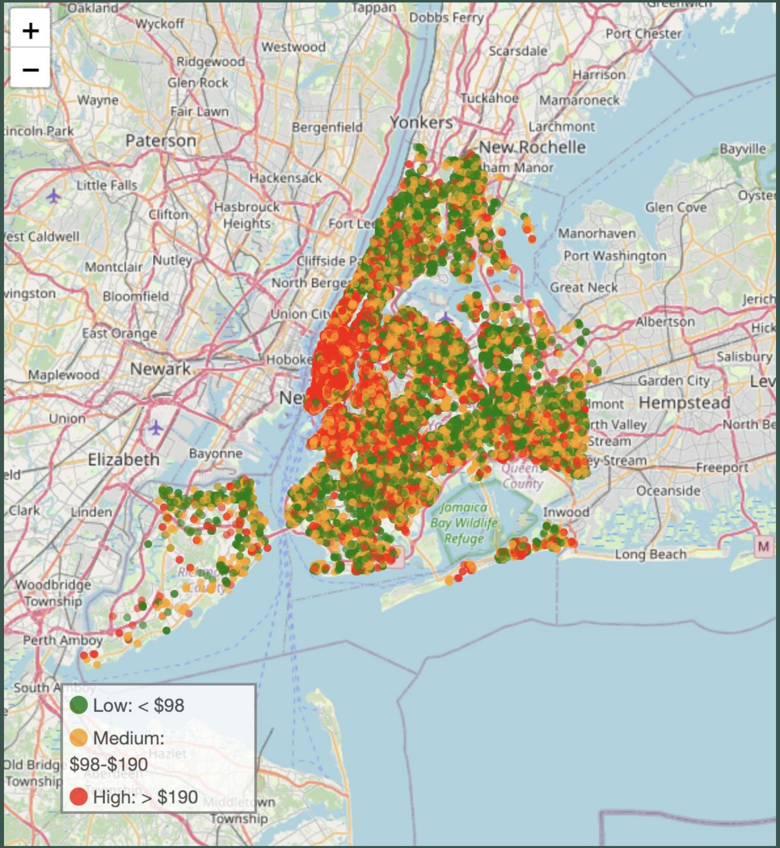
AIRBNB PRICE PREDICTION

Team 1: Jack Wang, Leo Yuan, Kenney Tran, Ryan Wu, Tanishka Gilara

01

PROBLEM & IMPORTANCE





OUR QUESTION AS A TEAM

"WHAT FACTORS CONTRIBUTE TO THE PRICE OF AN AIRBNB IN NEW YORK?"

PROPERTY A :

LOCATED IN BROOKLYN

ACCOMMODATES 2 PEOPLE,

1 BATHROOM,

1 BEDROOM,

1 BED,

4.14 AVERAGE REVIEW SCORE

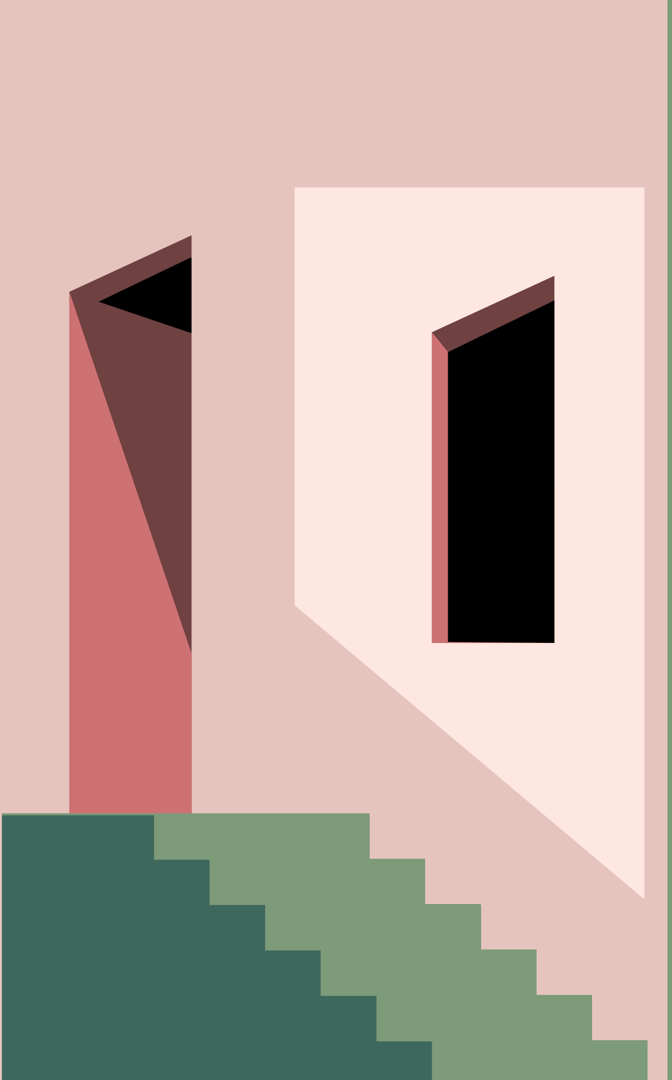
TAKE YOUR GUESS!

EXPLORING OUR REASONING

Why: Providing customers with a clearer understanding of factors influencing Airbnb prices

How: Building models to predict the most impactful factors on price

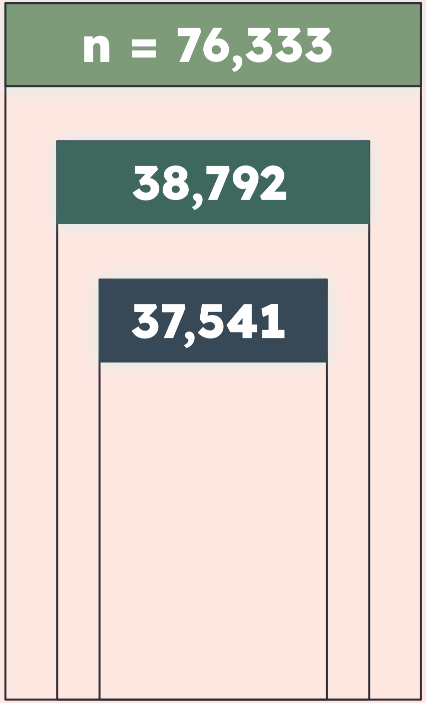
Target: Customers looking for better booking deals & fair value in New York (Initial)



02

DATA SUMMARY

DATASET OVERVIEW



ENTIRE DATASET

Comprises two New York datasets, with 2 categorical features and 17 numerical

MISSING VALUES

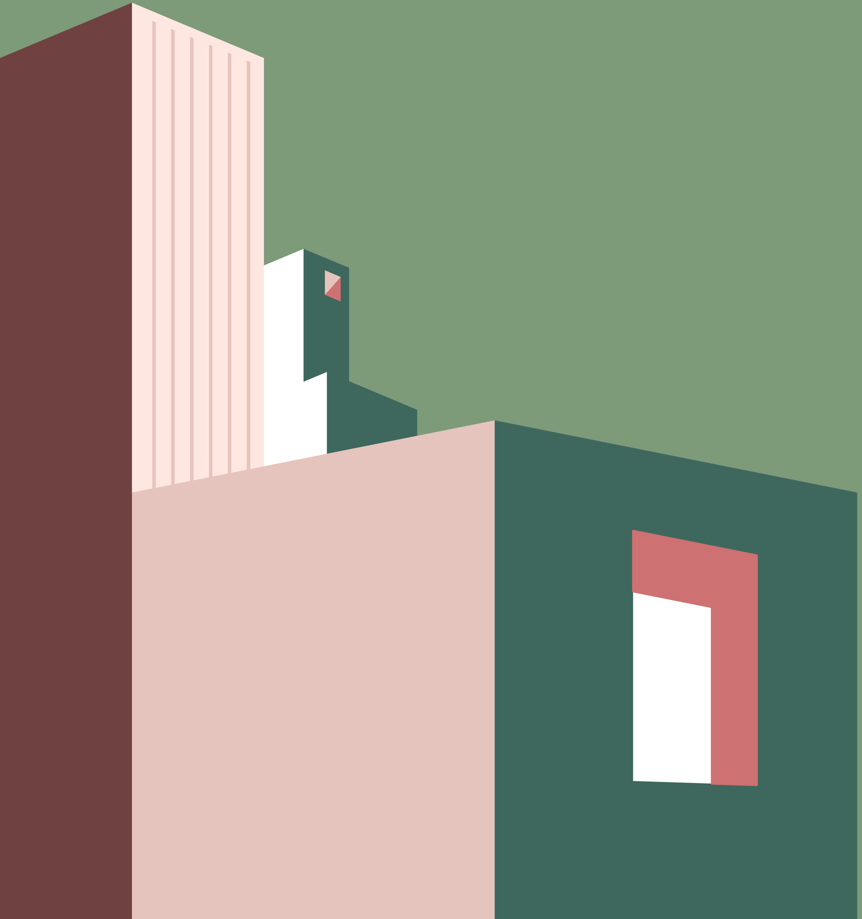
11 variables had missing values, ranging from 14k (18% of data) to 22k (30% of data)

IMPORTANT FEATURES

Data spanned 5 neighborhoods, and 5 metrics represented reviews of listing, 26 features post feature engineering

03

PRE-PROCESSING



DATA PRE-PROCESSING METHODOLOGIES

Feature Engineering

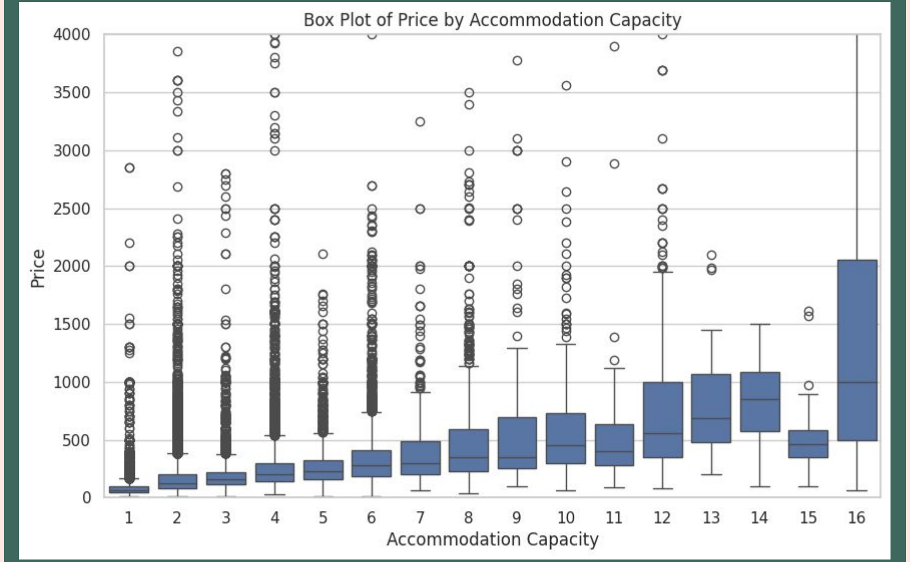
has_essentials: essentials, heating, air conditioner, dryer, washer etc
has_kitchen: refrigerator, coffee maker, dishwasher etc
has_entertainment: TV, workspace etc
has_safety: smoke, carbon mono etc
has_outdoor_space: backyard, patio

price (label): log transformation to reduce skewness, and to force normal distribution

high_price = price > median
low_price = *price < median*

Intuition:
Regression:
High p values
R^2 = 0.053
Regression (Log):
Low p values
R^2 = 0.458
Regression (IQR):
Medium p values
R^2 = 0.371

DATA PRE-PROCESSING METHODOLOGIES



Intuition: Manhattan Airbnb prices were almost double that of Queens, Bronx, Brooklyn, and Staten Island

Data Imputation

review metrics: cleanliness, location, value, accuracy, communication, checkin, and rating

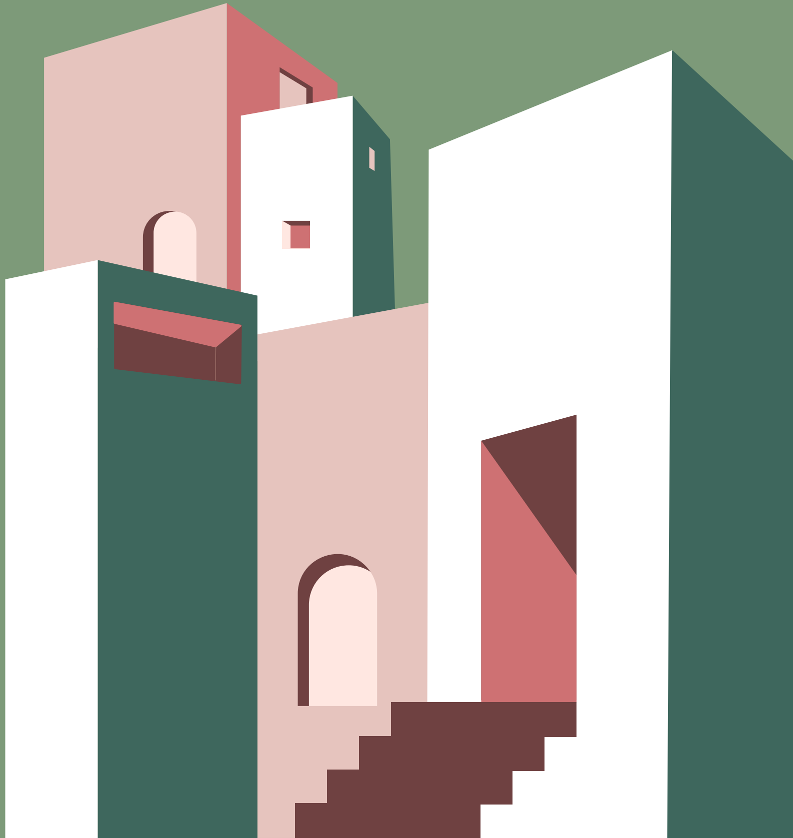
- Imputed grouping by neighborhood means

descriptive metrics: bedrooms, beds, bathrooms

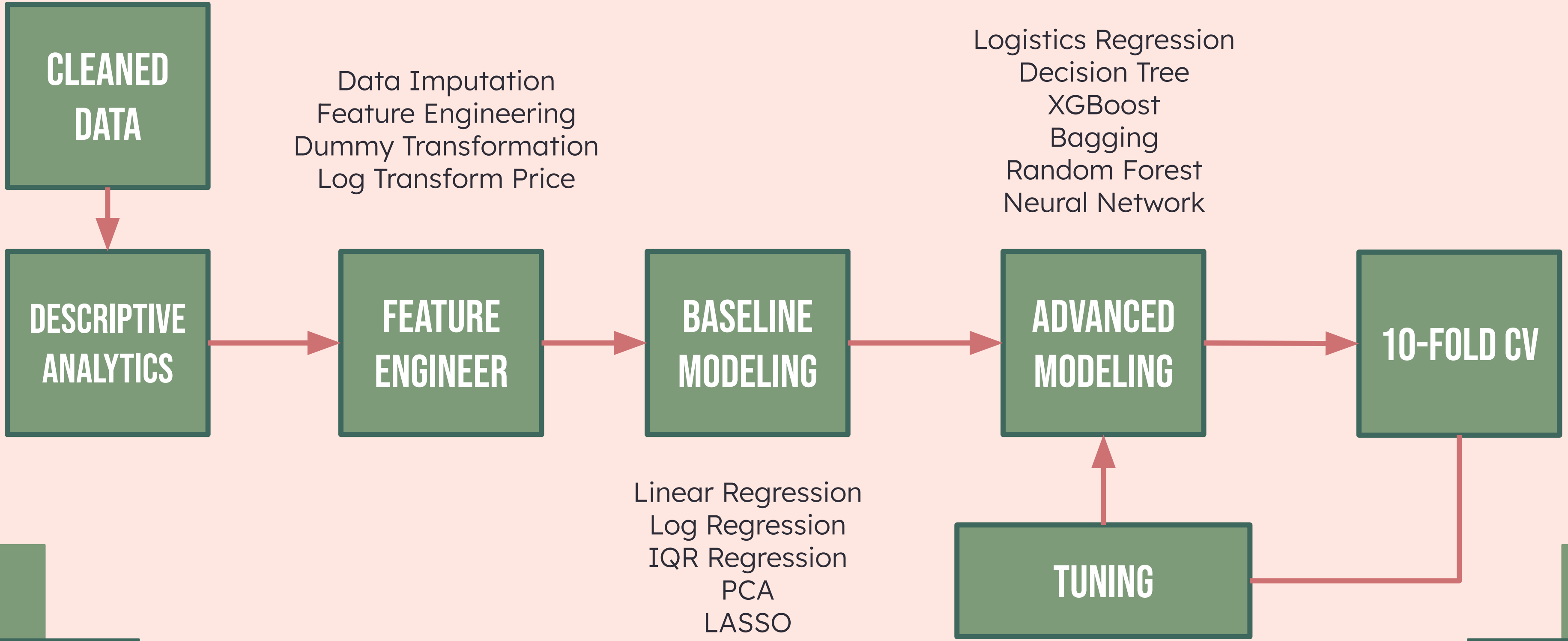
- Imputed grouping by 'accommodates' median to avoid outliers

04

MODELING APPROACH

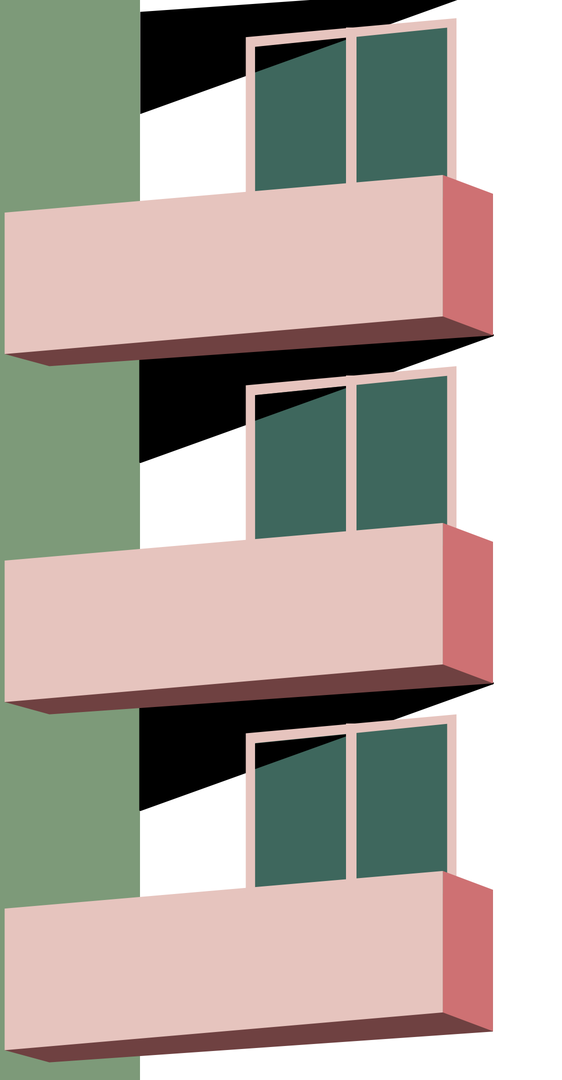


DATAFLOW


CLEANED DATA
DESCRIPTIVE ANALYTICS
Data Imputation
Feature Engineering
Dummy Transformation
Log Transform Price
FEATURE ENGINEER
BASELINE MODELING
Linear Regression
Log Regression
IQR Regression
PCA
LASSO
Logistics Regression
Decision Tree
XGBoost
Bagging
Random Forest
Neural Network
ADVANCED MODELING
10-FOLD CV
TUNING

05

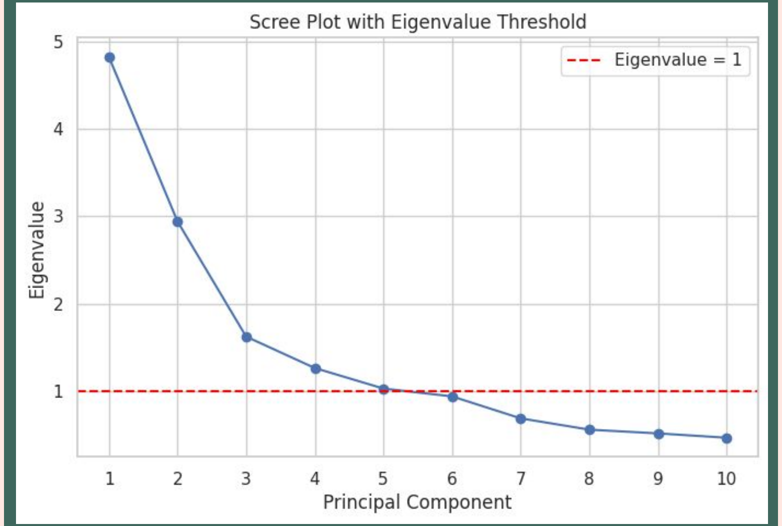
TUNING & MODEL EVALUATION



	LASSO	Bagging	Principal Components	Neural Network
Weakness	**Review variables** heavily co-linear, LASSO incorrectly drops	**Difficult to visualize,** black box	**No need** for dimension reduction	**Non-interpretable** for general use
Tuning	Lowest ***α*** yielded best outcomes	**Only reduces variance,** did not help in bias reduction	**Eigenvalue > 1** at 6 predictors	32,64 hidden layers, 0.1 learning rate
Overall Evaluation	Weak model, Ridge > Lasso	Great model performance, but weak interpretability	Unnecessary as we want **interpretability**, and the models are parsimonious	Without tuning, it performs similar to the **ensemble methods,** not many complex relationships to capture

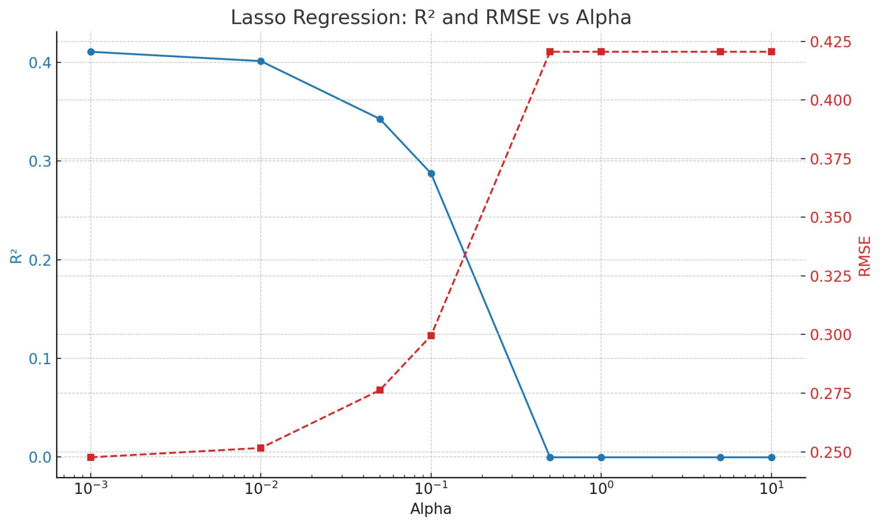
HYPERPARAMETERS TUNING

PCA



Principal components reduce **predictors to 6,** and misses out on **predictive power**

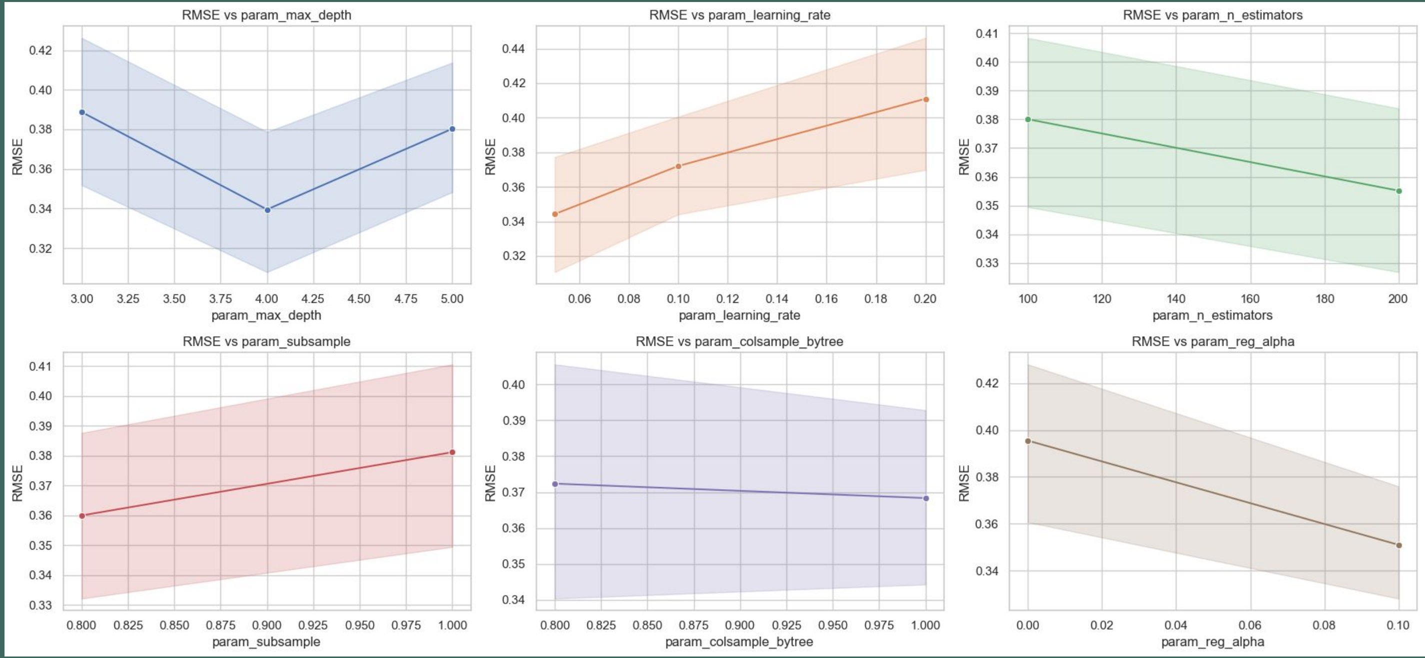
Lasso Regression Tuning



Lowest ***a*** performs worse than base log transformed regression, indicating no need to **underfit our model**

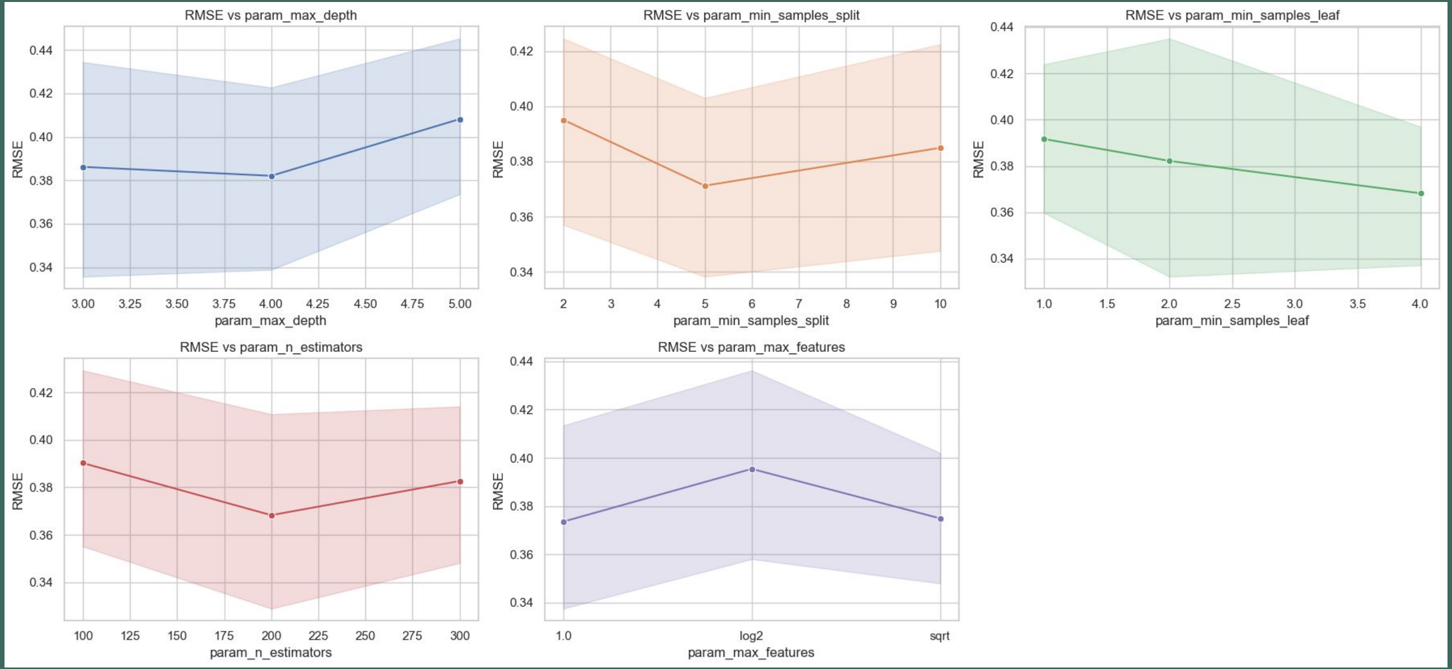
HYPERPARAMETERS TUNING CONT.

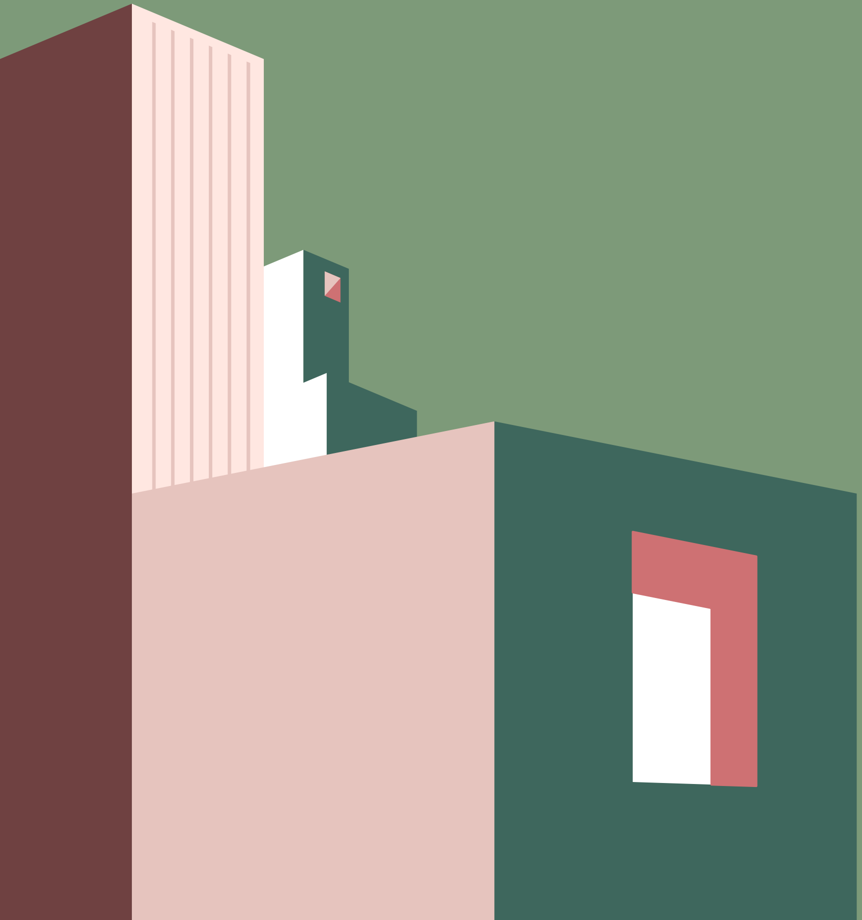
XGBoost Decision Tree



HYPERPARAMETERS TUNING CONT.

Random Forest





06

MODEL PERFORMANCE

Weak / Punching Bag Models

Linear Regression
$R^2 = 0.053$

(Log) Linear Regression
$R^2 = 0.458$

Lasso Regression
$R^2 = 0.411$
$\alpha = 0.001$

Decision Tree
$R^2 = 0.465$
RMSE: 0.225

Algorithm Performance (Cross Validated)

Logistics Regression
L2 Penalty
10-Fold CV
CS = 10

ROC AUC = 0.86
Accuracy: 79%
F1 Score: 80%

XGBoost Decision Tree
n = 200
Rate = 0.2
Depth = 5
$\lambda = 1$

$R^2 = 0.6625$
RMSE = 0.1418

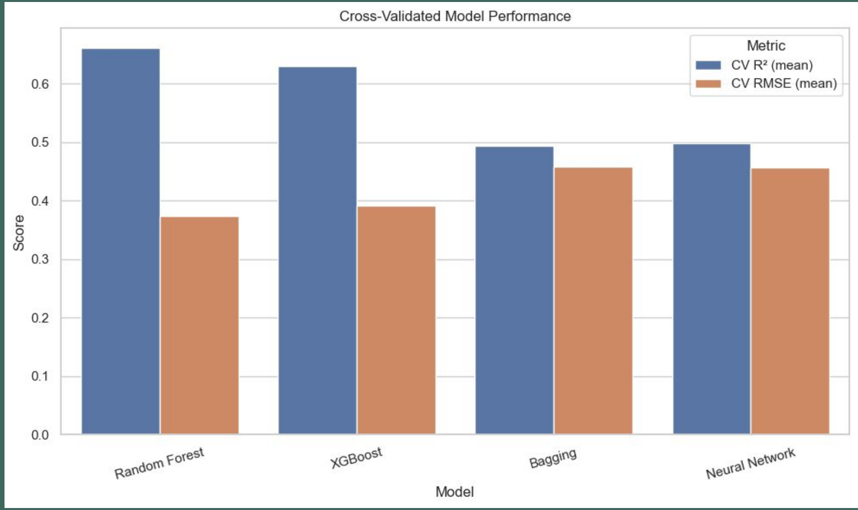
Random Forest
n = 300
Depth = none
Features = 1.0

$R^2 = 0.7084$
RMSE = 0.1226

Neural Network
Hidden = (32, 64)
Logistic activation

$R^2 = 0.6137$
RMSE = 0.1624

STRONGEST MODEL?



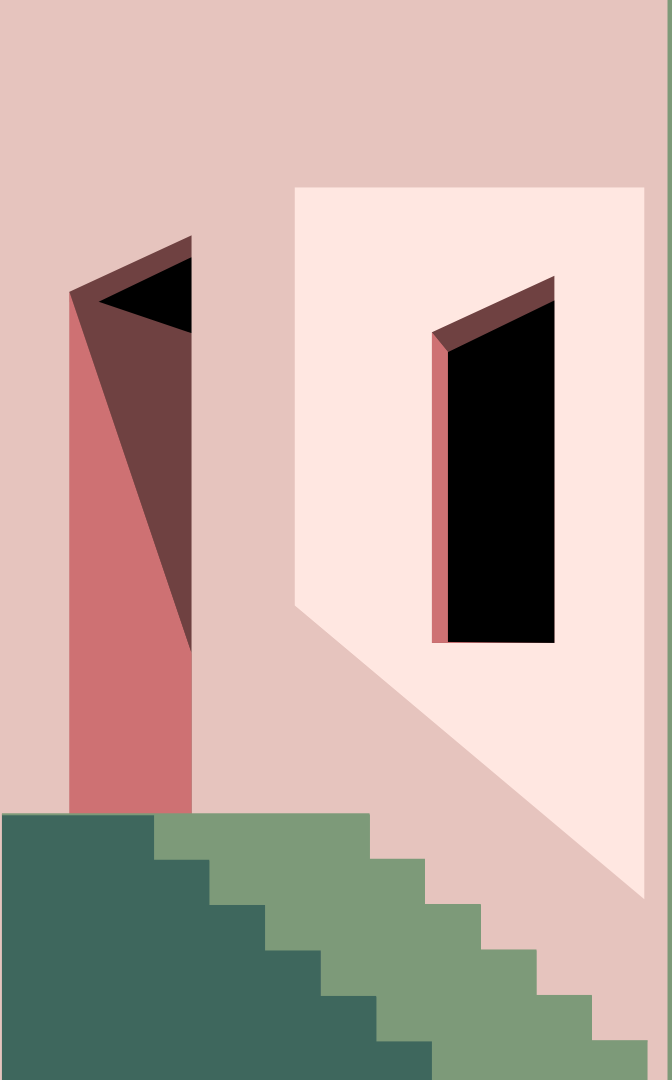
Model strength should not be the only parameter when deciding on the ultimate algorithm as interpretability and generalizability is crucial

XGBoost - Robust to missing values and non-linearities

Logistics Regression - Interpretable and strong for predicting if price > median

Random Forest - Most powerful model, but not ideal for general use

Log Transformed Regression - Ideal for interpretation, but weak model predictive power



07

CHALLENGES & NEXT STEPS

CHALLENGES

Challenges	What happened?	How would we address?
Finding a powerful but interpretable model	Models used have are less interpretable due to the black box models being the most powerful	Model Stacking (utilizing Random Forest coefficients in a Neural Network) Finely tuned Decision Trees
Interpreting locational data beyond neighborhood splitting	We interpreted latitude and longitude as is, which isn't robust enough to capture interactions	Geography remains a key predictor, and kriging can enhance accuracy
Lack of dimensions and features for predictions	We lacked features such as age of property, condition, as well as attractions near each listing	Merge dining / attractions datasets to add additional dimensionality and improve predictive power

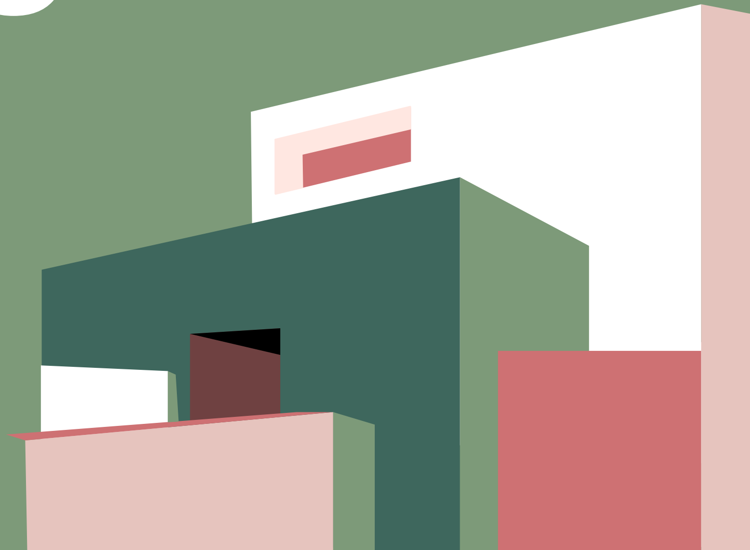
Predicted

Actual

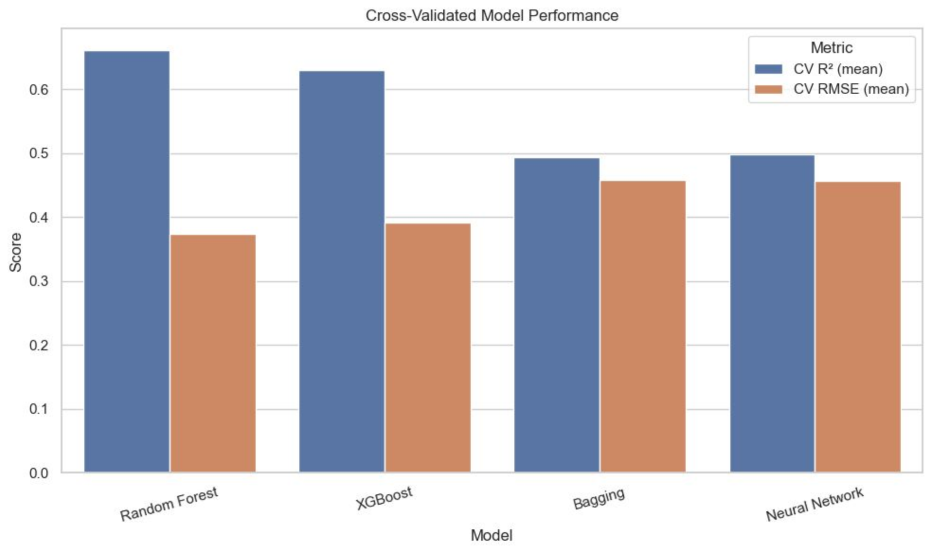
Model	Predicted Price
Random Forest	$90.65
XGBoost	$76.77
Bagging	$85.18
Neural Network	$166.93

$95

THANK YOU

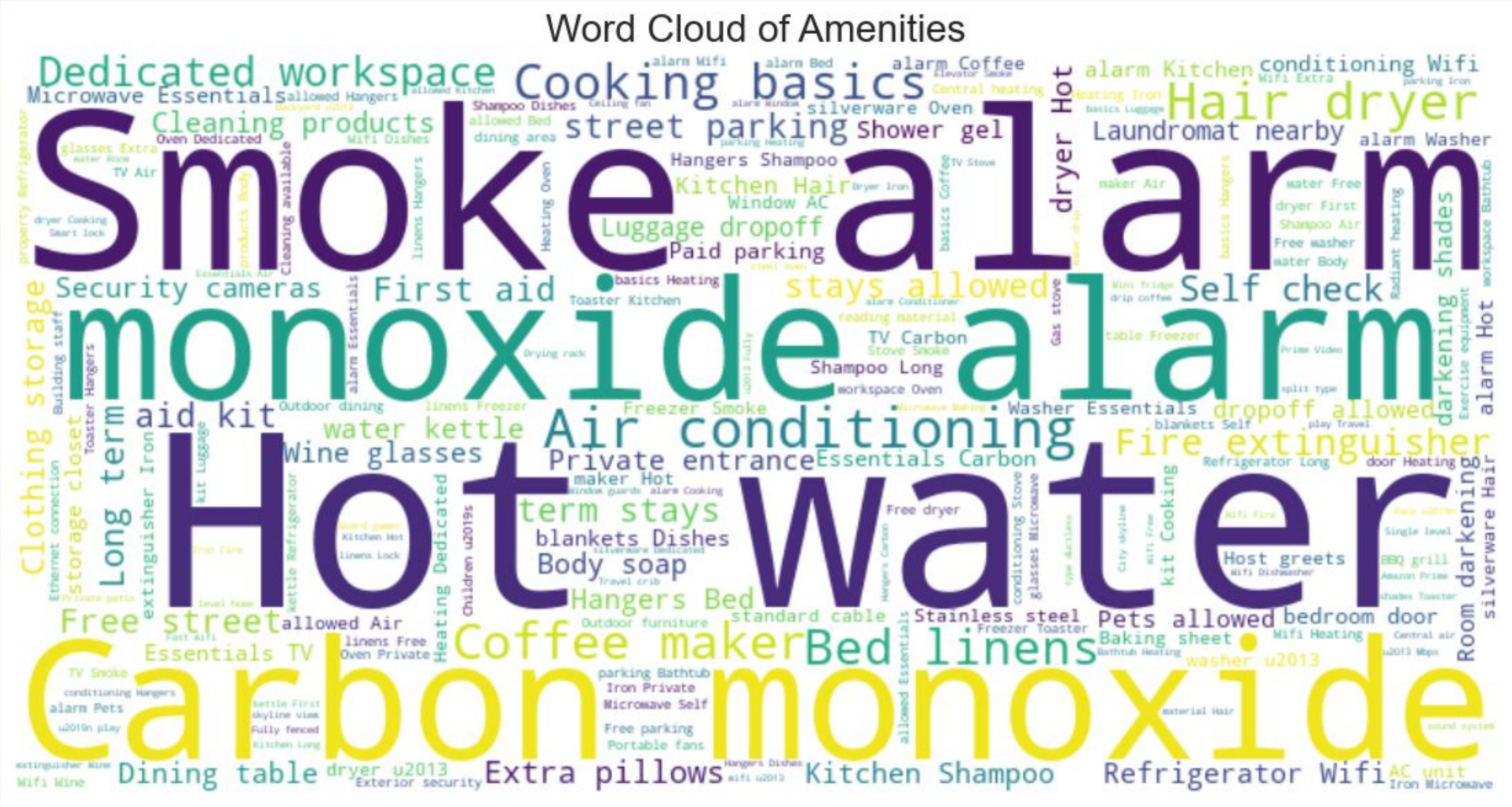


APPENDIX A

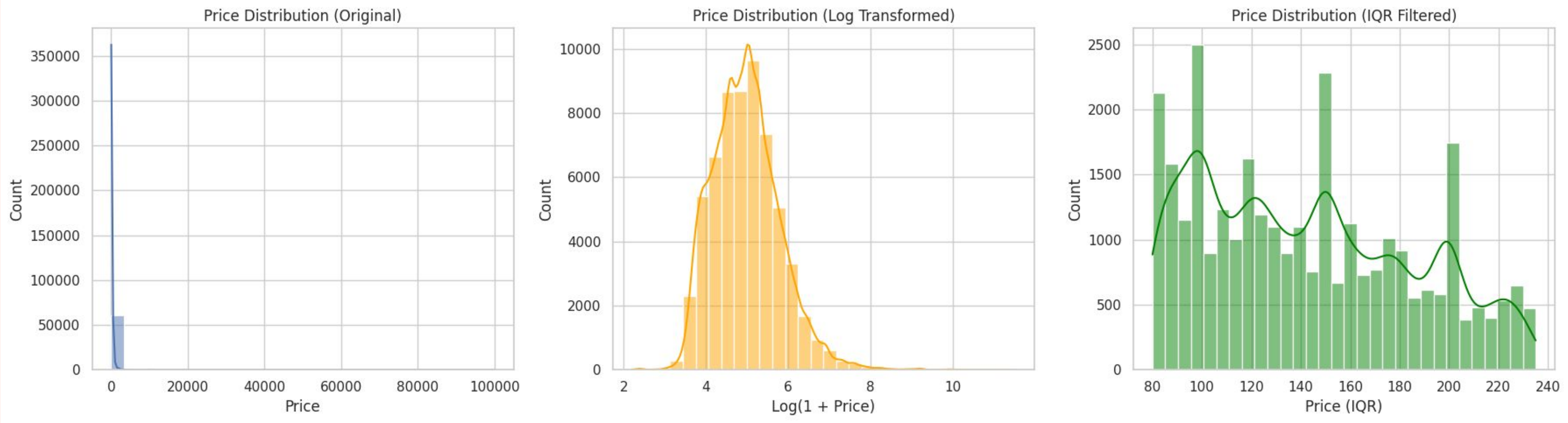


	Model	CV R^2 (mean)	CV R^2 (std)	CV RMSE (mean)	CV RMSE (std)
0	Random Forest	**0.661764**	0.056405	**0.373892**	**0.019373**
1	Bagging	0.494240	0.048128	0.458728	0.007174
2	Boosting	0.629595	0.045716	0.392085	0.010915
3	Neural Network	0.498783	0.039881	0.456948	0.004293

APPENDIX B


Word Cloud of Amenities
Smoke alarm
monoxide alarm
Hot water
Carbon monoxide
Dedicated workspace
Cooking basics
Hair dryer
Microwave Essentials
Cleaning products
street parking
Shower gel
Laundromat nearby
alarm Washer
silverware Oven
alarm Kitchen
conditioning Wifi
alarm Coffee
Hangers Shampoo
Kitchen Hair
Window AC
Luggage dropoff
Paid parking
Security cameras
First aid
stays allowed
Self check
dryer Hot
TV Carbon
Shampoo Long
aid kit
Air conditioning
water kettle
Wine glasses
Private entrance
Essentials Carbon
Fire extinguisher
dropoff allowed
Washer Essentials
Freezer Smoke
darkening shades
Clothing storage
storage closet
Long term
term stays
blankets Dishes
Body soap
Hangers Bed
Host greets
Room darkening
Free street
allowed Air
Coffee maker
Bed linens
Essentials TV
Stainless steel
Pets allowed
bedroom door
Baking sheet
washer u2013
Dining table
Extra pillows
Kitchen Shampoo
Refrigerator Wifi
Free parking
Portable fans
Micromave Self
Iron Private
parking Bathtub
standard cable
Outdoor dining
Heating Dedicated
kit Cooking
silverware Hair
alarm Hot
AC unit

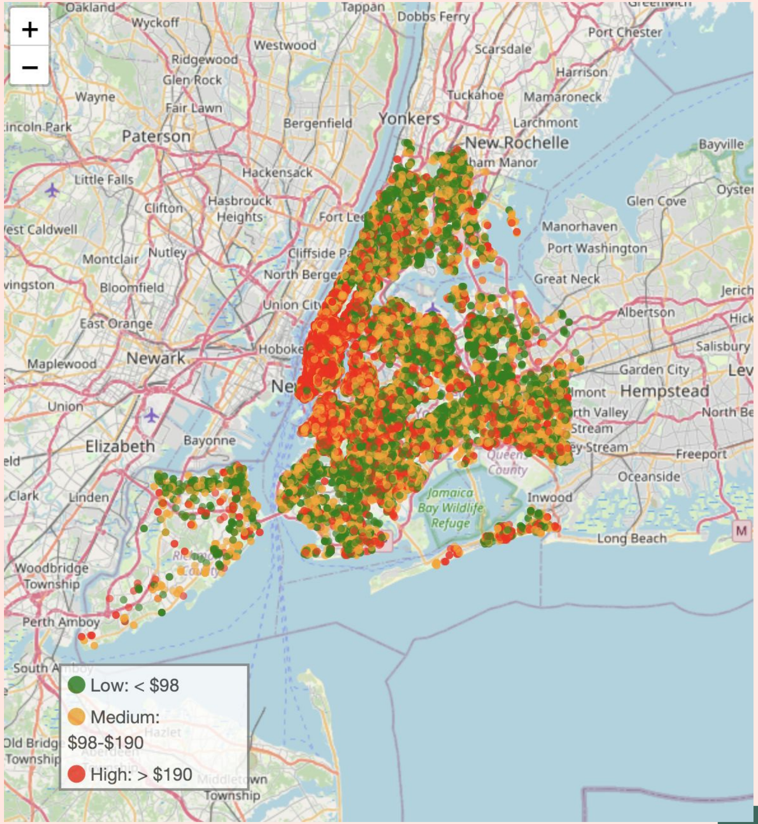
APPENDIX C



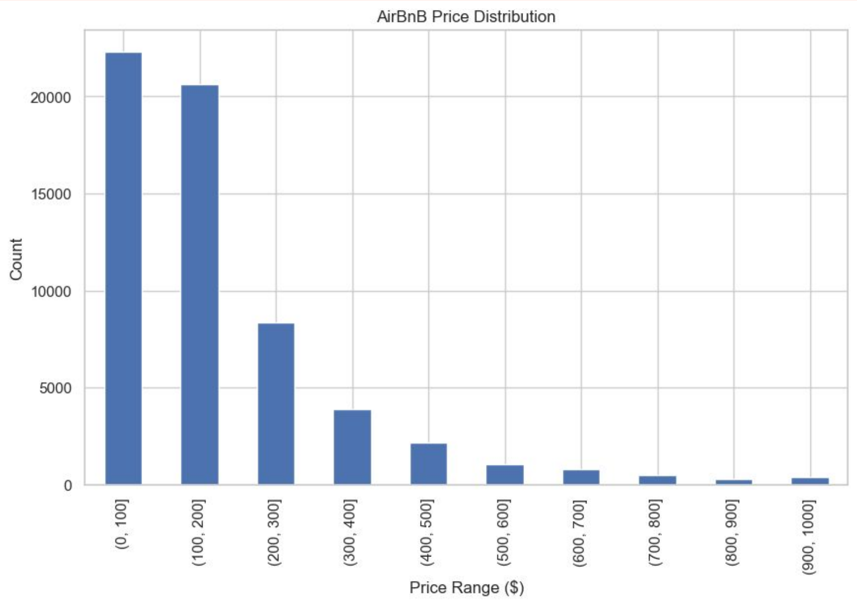
Price Distribution (Original)
Count
350000
300000
250000
200000
150000
100000
50000
0
0
20000
40000
60000
80000
100000
Price
Price Distribution (Log Transformed)
Count
10000
8000
6000
4000
2000
0
2
4
6
8
10
Log(1 + Price)
Price Distribution (IQR Filtered)
Count
2500
2000
1500
1000
500
0
80
100
120
140
160
180
200
220
240
Price (IQR)

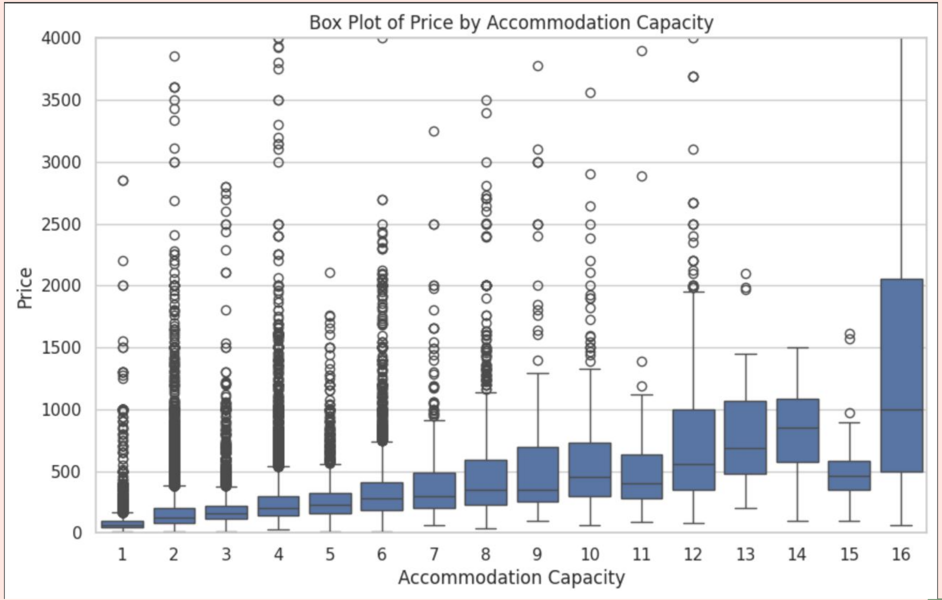
APPENDIX D


Average Airbnb Price by Neighborhood Group
Average Price ($)
0
50
100
150
200
250
300
Manhattan
Brooklyn
Queens
Staten Island
Bronx
Neighborhood Group

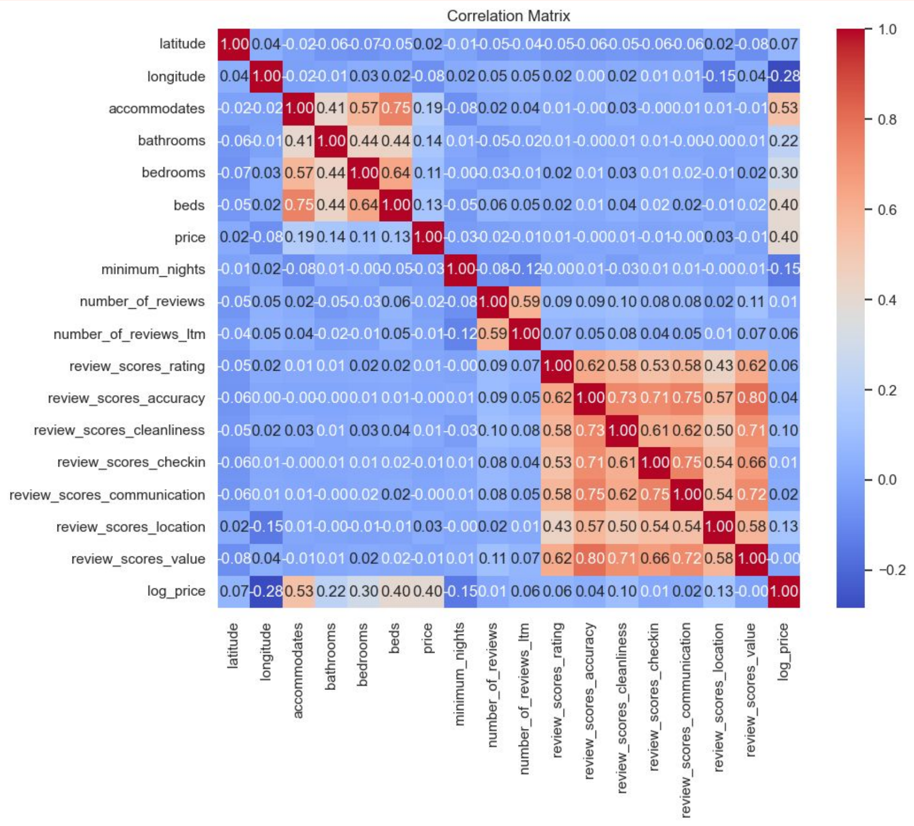

Low: < $98
Medium:
$98-$190
High: > $190
Yonkers
New Rochelle
Paterson
Newark
Elizabeth
Hempstead
Jamaica Bay Wildlife Refuge
Long Beach
Oceanside
Perth Amboy

APPENDIX E

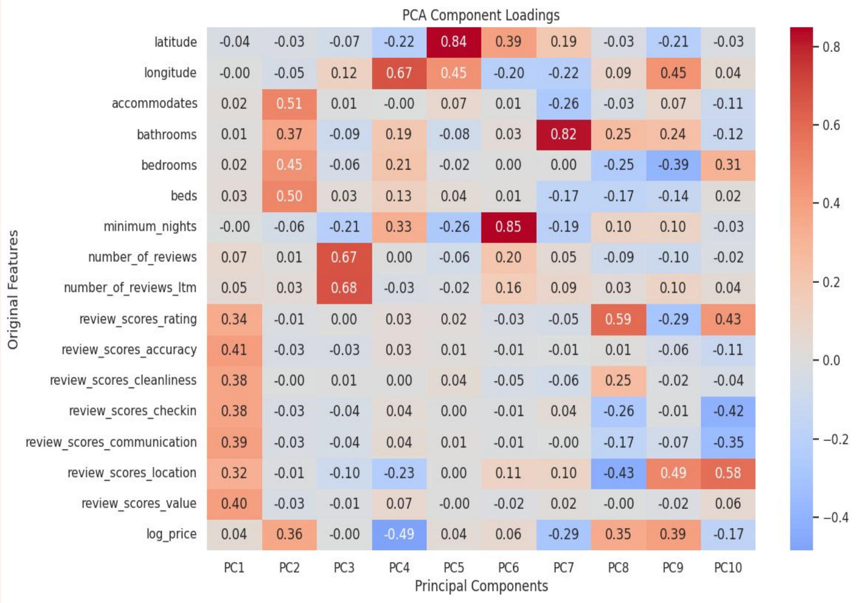

AirBnB Price Distribution
Count
20000
15000
10000
5000
0
(0, 100]
(100, 200]
(200, 300]
(300, 400]
(400, 500]
(500, 600]
(600, 700]
(700, 800]
(800, 900]
(900, 1000]
Price Range ($)

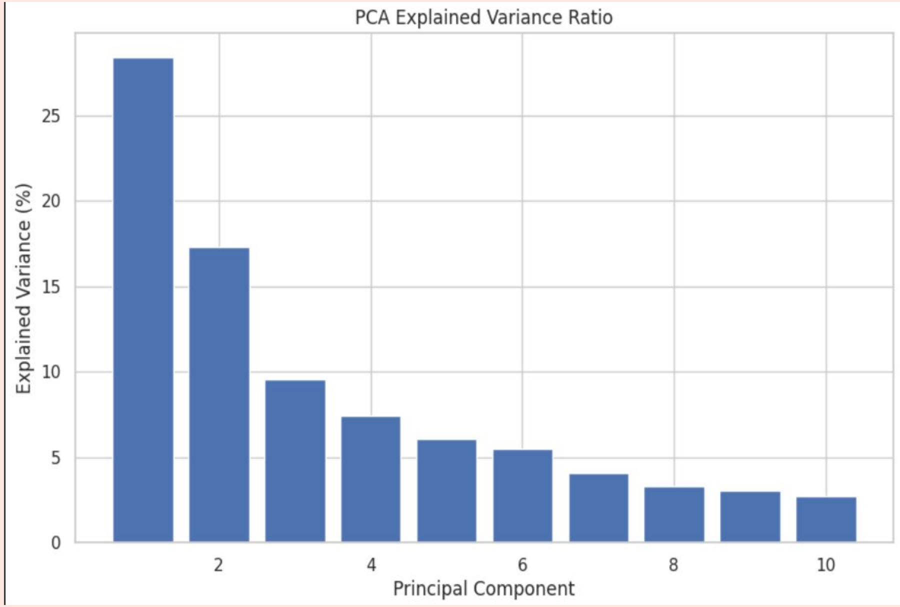

Box Plot of Price by Accommodation Capacity
Price
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Accommodation Capacity

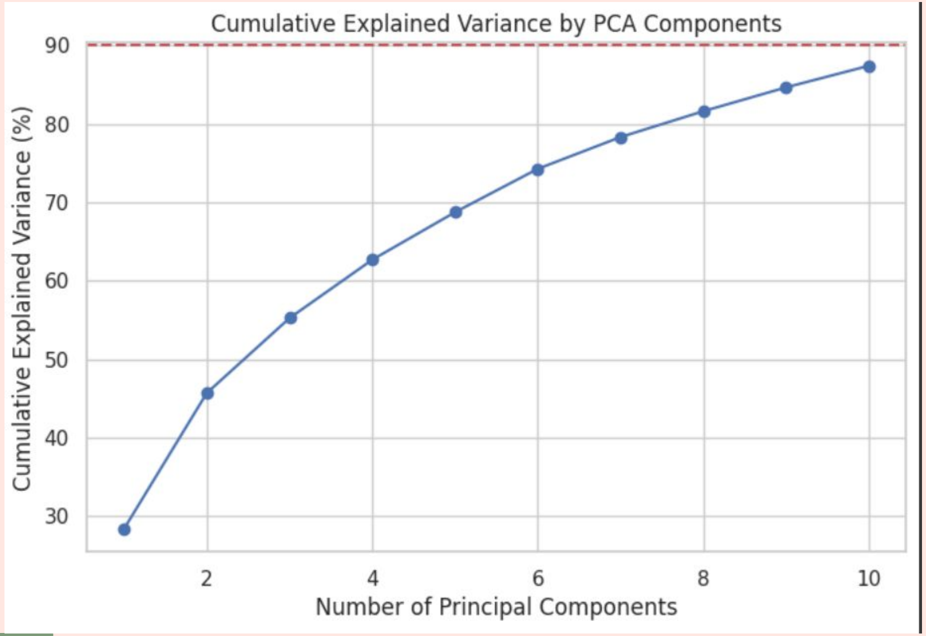
APPENDIX F

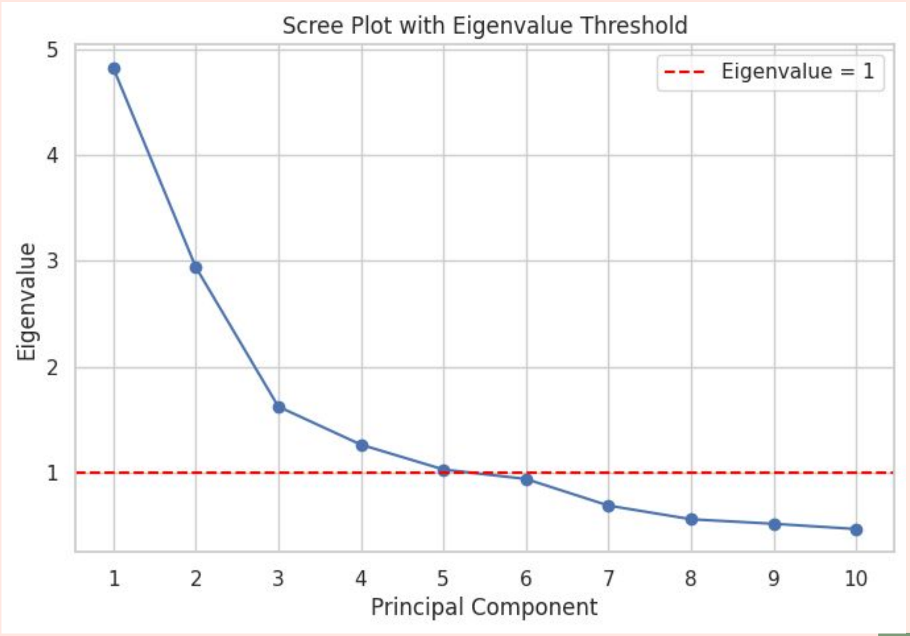

Correlation Matrix
latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
minimum_nights
number_of_reviews
number_of_reviews_ltm
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
log_price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

APPENDIX G


PCA Component Loadings
Original Features
Principal Components
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10
latitude -0.04 -0.03 -0.07 -0.22 0.84 0.39 0.19 -0.03 -0.21 -0.03
longitude -0.00 -0.05 0.12 0.67 0.45 -0.20 -0.22 0.09 0.45 0.04
accommodates 0.02 0.51 0.01 -0.00 0.07 0.01 -0.26 -0.03 0.07 -0.11
bathrooms 0.01 0.37 -0.09 0.19 -0.08 0.03 0.82 0.25 0.24 -0.12
bedrooms 0.02 0.45 -0.06 0.21 -0.02 0.00 0.00 -0.25 -0.39 0.31
beds 0.03 0.50 0.03 0.13 0.04 0.01 -0.17 -0.17 -0.14 0.02
minimum_nights -0.00 -0.06 -0.21 0.33 -0.26 0.85 -0.19 0.10 0.10 -0.03
number_of_reviews 0.07 0.01 0.67 0.00 -0.06 0.20 0.05 -0.09 -0.10 -0.02
number_of_reviews_ltm 0.05 0.03 0.68 -0.03 -0.02 0.16 0.09 0.03 0.10 0.04
review_scores_rating 0.34 -0.01 0.00 0.03 0.02 -0.03 -0.05 0.59 -0.29 0.43
review_scores_accuracy 0.41 -0.03 -0.03 0.03 0.01 -0.01 -0.01 0.01 -0.06 -0.11
review_scores_cleanliness 0.38 -0.00 0.01 0.00 0.04 -0.05 -0.06 0.25 -0.02 -0.04
review_scores_checkin 0.38 -0.03 -0.04 0.04 0.00 -0.01 0.04 -0.26 -0.01 -0.42
review_scores_communication 0.39 -0.03 -0.04 0.04 0.01 -0.01 -0.00 -0.17 -0.07 -0.35
review_scores_location 0.32 -0.01 -0.10 -0.23 0.00 0.11 0.10 -0.43 0.49 0.58
review_scores_value 0.40 -0.03 -0.01 0.07 -0.00 -0.02 0.02 -0.00 -0.02 0.06
log_price 0.04 0.36 -0.00 -0.49 0.04 0.06 -0.29 0.35 0.39 -0.17
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


PCA Explained Variance Ratio
Explained Variance (%)
0
5
10
15
20
25
Principal Component
2
4
6
8
10


Cumulative Explained Variance by PCA Components
Cumulative Explained Variance (%)
Number of Principal Components
90
80
70
60
50
40
30
2
4
6
8
10


Scree Plot with Eigenvalue Threshold
Eigenvalue = 1
Eigenvalue
Principal Component
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10

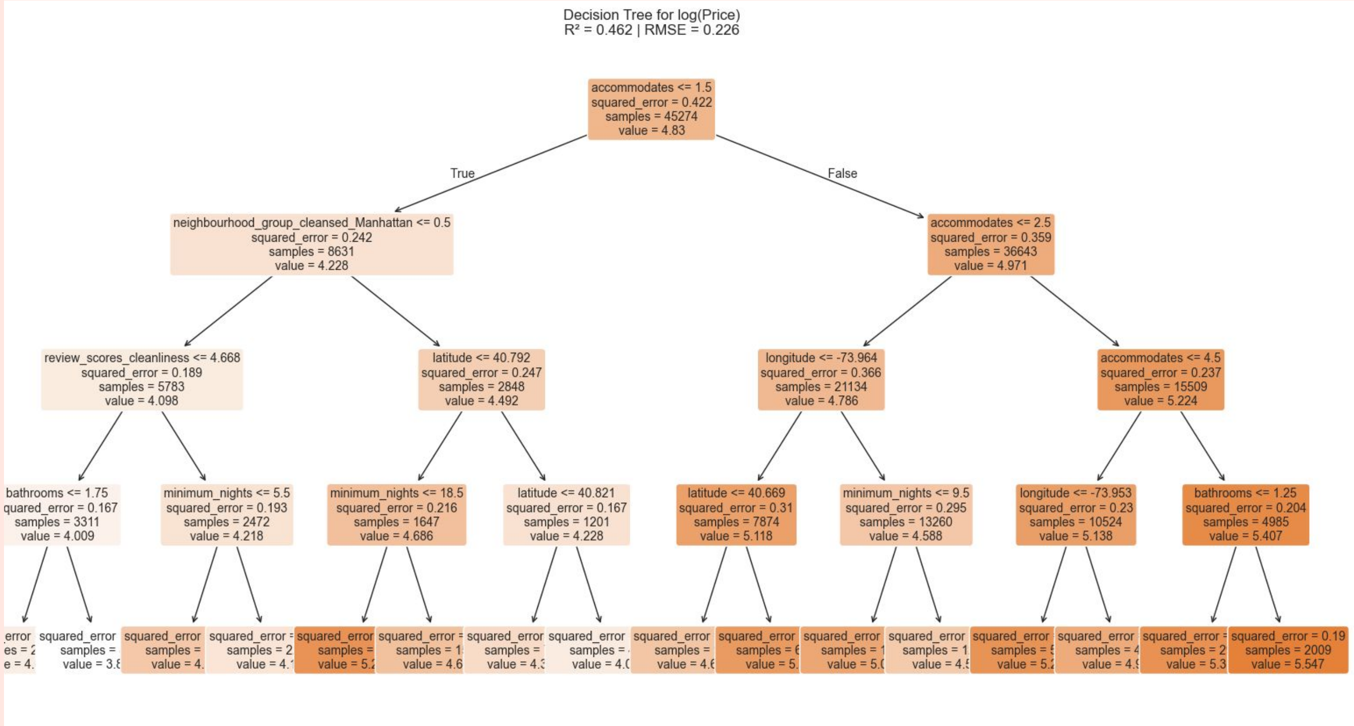
APPENDIX H

OLS Regression Results

Dep. Variable:	price	R-squared:	0.053
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	137.1
Date:	Wed, 02 Apr 2025	Prob (F-statistic):	0.00
Time:	14:06:00	Log-Likelihood:	-4.9022e+05
No. Observations:	61333	AIC:	9.805e+05
Df Residuals:	61307	BIC:	9.807e+05
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	221.8306	2.892	76.696	0.000	216.162	227.500
latitude	-29.2233	4.845	-6.031	0.000	-38.720	-19.726
longitude	-41.7832	4.970	-8.407	0.000	-51.524	-32.042
accommodates	116.3429	4.511	25.792	0.000	107.502	125.184
bathrooms	61.9079	3.361	18.418	0.000	55.320	68.496
bedrooms	2.3224	3.943	0.589	0.556	-5.406	10.051
beds	-15.6121	4.781	-3.265	0.001	-24.983	-6.241
minimum_nights	-11.3286	2.952	-3.838	0.000	-17.114	-5.543
number_of_reviews	-8.0826	3.620	-2.233	0.026	-15.177	-0.988
number_of_reviews_ltm	1.1835	3.596	0.329	0.742	-5.865	8.232
review_scores_rating	6.6948	3.928	1.704	0.088	-1.005	14.394
review_scores_accuracy	-1.5352	5.821	-0.264	0.792	-12.945	9.874
review_scores_cleanliness	9.5353	4.594	2.076	0.038	0.531	18.540
review_scores_checkin	-13.1672	4.752	-2.771	0.006	-22.481	-3.854
review_scores_communication	3.1500	5.116	0.616	0.538	-6.877	13.177
review_scores_location	9.8628	3.916	2.519	0.012	2.187	17.538
review_scores_value	-3.2963	5.514	-0.598	0.550	-14.105	7.512
has_essentials	-0.5742	3.060	-0.188	0.851	-6.571	5.423
has_kitchen	-22.5669	3.002	-7.518	0.000	-28.450	-16.683
has_entertainment	2.6956	3.055	0.882	0.378	-3.293	8.684
has_safety	-5.9358	2.982	-1.990	0.047	-11.781	-0.091
has_outdoor_space	10.7167	2.929	3.659	0.000	4.976	16.457
neighbourhood_group_cleansed_Brooklyn	-48.9820	10.492	-4.668	0.000	-69.547	-28.417
neighbourhood_group_cleansed_Manhattan	19.5861	9.392	2.085	0.037	1.178	37.994
neighbourhood_group_cleansed_Queens	-15.8019	7.310	-2.162	0.031	-30.129	-1.475
neighbourhood_group_cleansed_Staten Island	-32.6685	4.355	-7.502	0.000	-41.203	-24.134

APPENDIX I


Decision Tree for log(Price)
R² = 0.462 | RMSE = 0.226
accommodates <= 1.5
squared_error = 0.422
samples = 45274
value = 4.83
True
False
neighbourhood_group_cleansed_Manhattan <= 0.5
squared_error = 0.242
samples = 8631
value = 4.228
accommodates <= 2.5
squared_error = 0.359
samples = 36643
value = 4.971
review_scores_cleanliness <= 4.668
squared_error = 0.189
samples = 5783
value = 4.098
latitude <= 40.792
squared_error = 0.247
samples = 2848
value = 4.492
longitude <= -73.964
squared_error = 0.366
samples = 21134
value = 4.786
accommodates <= 4.5
squared_error = 0.237
samples = 15509
value = 5.224
bathrooms <= 1.75
squared_error = 0.167
samples = 3311
value = 4.009
minimum_nights <= 5.5
squared_error = 0.193
samples = 2472
value = 4.218
minimum_nights <= 18.5
squared_error = 0.216
samples = 1647
value = 4.686
latitude <= 40.821
squared_error = 0.167
samples = 1201
value = 4.228
latitude <= 40.669
squared_error = 0.31
samples = 7874
value = 5.118
minimum_nights <= 9.5
squared_error = 0.295
samples = 13260
value = 4.588
longitude <= -73.953
squared_error = 0.23
samples = 10524
value = 5.138
bathrooms <= 1.25
squared_error = 0.204
samples = 4985
value = 5.407
squared_error = 0.19
samples = 2009
value = 5.547

APPENDIX J

	precision	recall	f1-score	support
0	0.79	0.82	0.80	6021
1	0.78	0.75	0.77	5298
accuracy			0.79	11319
macro avg	0.79	0.78	0.78	11319
weighted avg	0.79	0.79	0.79	11319

ROC AUC: 0.8605

Neural Net R^2: 0.6137
Neural Net RMSE (log): 0.1624

APPENDIX K

```
=== Best Parameters XGBoost===
{'colsample_bytree': 0.8, 'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 200, 'reg_alpha': 0, 'reg_lambda': 1.5, 'subsample': 1.0}

=== Test Performance ===
R²: 0.6609
RMSE (log scale): 0.1425
```

	feature	importance
2	accommodates	0.2961251
5	beds	0.10844432
4	bedrooms	0.10109886
6	minimum_nights	0.07528877
22	neighbourhood_group_cleansed_Ma	0.07158521
1	longitude	0.06391023
3	bathrooms	0.033237386
0	latitude	0.032904126
8	number_of_reviews_ltm	0.025478369
11	review_scores_cleanliness	0.023847632

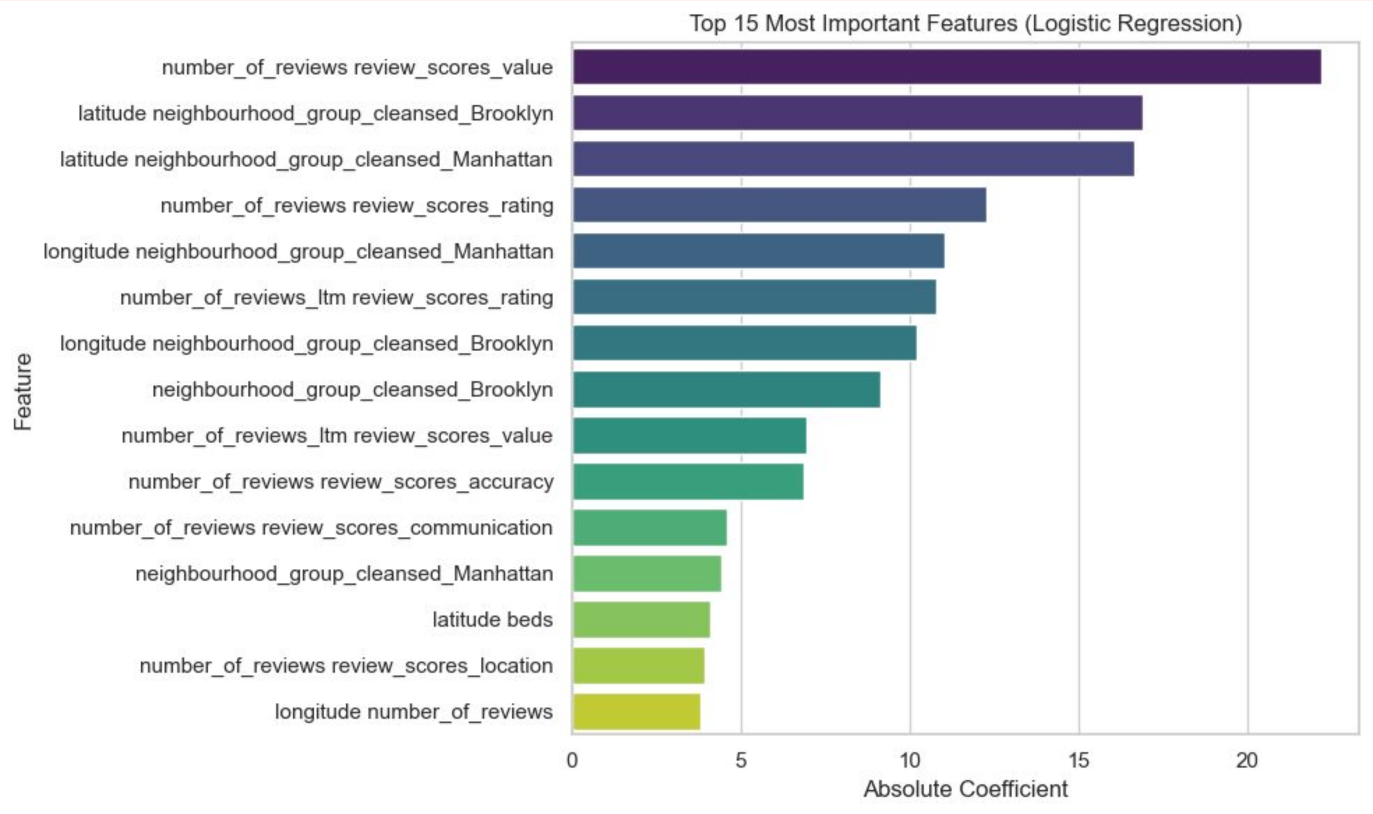
APPENDIX L

```
=== Best Parameters Random Forest===
{'max_depth': None, 'max_features': 1.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}

=== Test Performance ===
R²: 0.7084
RMSE (log scale): 0.1226
```

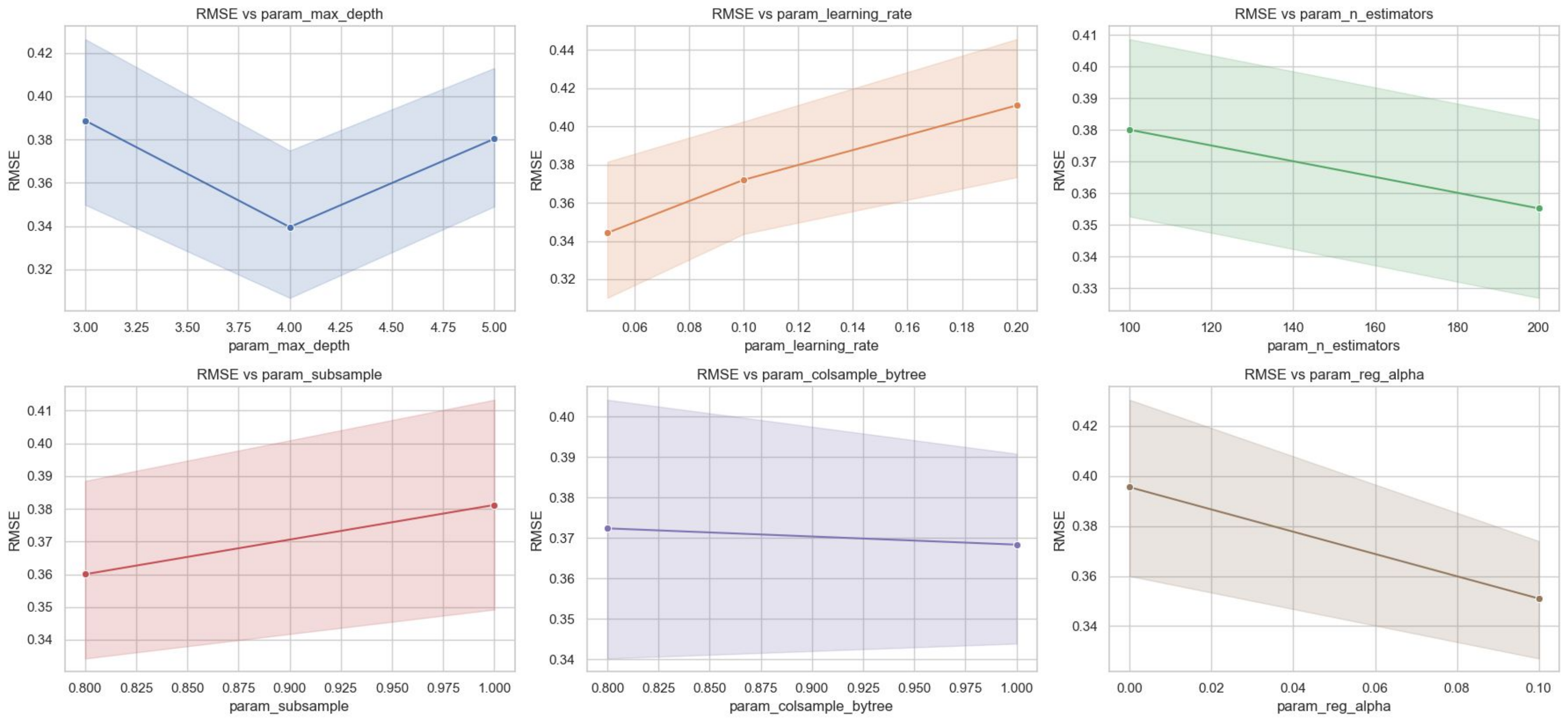
	feature	importance
2	accommodates	0.30297005215806017
1	longitude	0.20557052975783094
0	latitude	0.14281440184570726
6	minimum_nights	0.052459237961821066
11	review_scores_cleanliness	0.033958953842321185
7	number_of_reviews	0.02975290881249307
8	number_of_reviews_ltm	0.026047732685952518
4	bedrooms	0.02370654450463369
3	bathrooms	0.02343301151045217
15	review_scores_value	0.021154202745674564

APPENDIX M

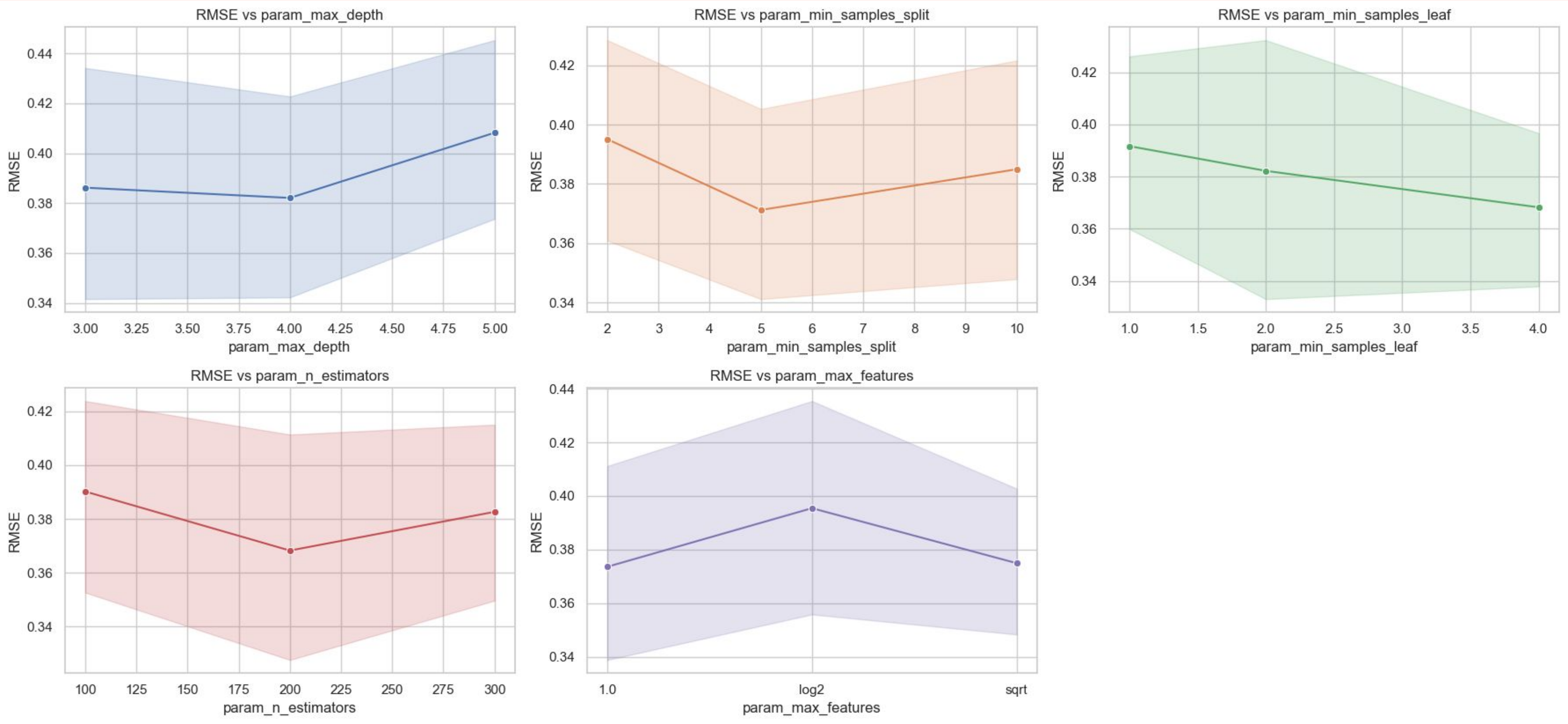


Top 15 Most Important Features (Logistic Regression)
Feature
number_of_reviews review_scores_value
latitude neighbourhood_group_cleansed_Brooklyn
latitude neighbourhood_group_cleansed_Manhattan
number_of_reviews review_scores_rating
longitude neighbourhood_group_cleansed_Manhattan
number_of_reviews_ltm review_scores_rating
longitude neighbourhood_group_cleansed_Brooklyn
neighbourhood_group_cleansed_Brooklyn
number_of_reviews_ltm review_scores_value
number_of_reviews review_scores_accuracy
number_of_reviews review_scores_communication
neighbourhood_group_cleansed_Manhattan
latitude beds
number_of_reviews review_scores_location
longitude number_of_reviews
0
5
10
15
20
Absolute Coefficient

APPENDIX N


RMSE vs param_max_depth
RMSE
param_max_depth
RMSE vs param_learning_rate
RMSE
param_learning_rate
RMSE vs param_n_estimators
RMSE
param_n_estimators
RMSE vs param_subsample
RMSE
param_subsample
RMSE vs param_colsample_bytree
RMSE
param_colsample_bytree
RMSE vs param_reg_alpha
RMSE
param_reg_alpha

APPENDIX O

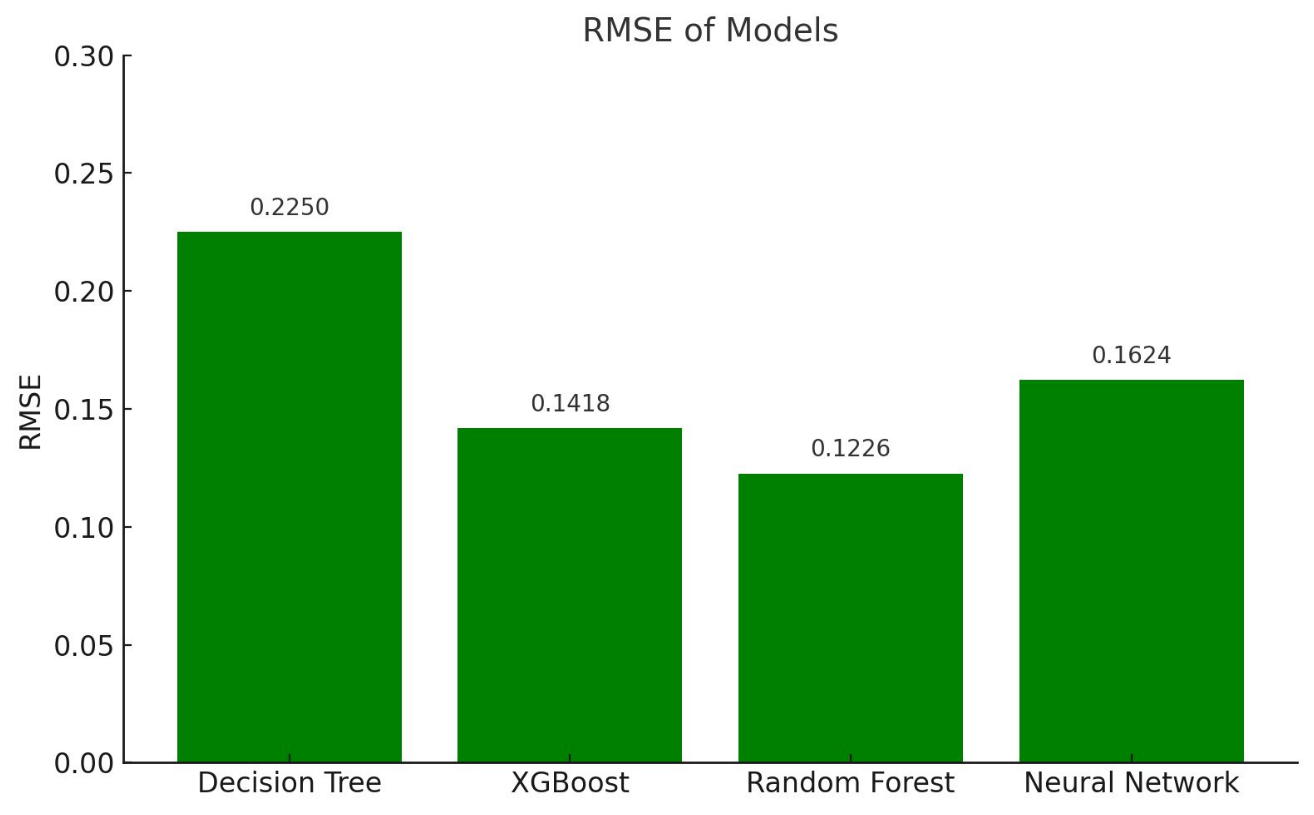


RMSE vs param_max_depth
RMSE
0.44
0.42
0.40
0.38
0.36
0.34
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00
param_max_depth
RMSE vs param_min_samples_split
RMSE
0.42
0.40
0.38
0.36
0.34
2
3
4
5
6
7
8
9
10
param_min_samples_split
RMSE vs param_min_samples_leaf
RMSE
0.42
0.40
0.38
0.36
0.34
1.0
1.5
2.0
2.5
3.0
3.5
4.0
param_min_samples_leaf
RMSE vs param_n_estimators
RMSE
0.42
0.40
0.38
0.36
0.34
100
125
150
175
200
225
250
275
300
param_n_estimators
RMSE vs param_max_features
RMSE
0.44
0.42
0.40
0.38
0.36
0.34
1.0
log2
sqrt
param_max_features

APPENDIX P

	Model	# Actual Price	# Predicted Price
0	Random Forest	95.0	90.65
1	XGBoost	95.0	76.7699966430664
2	Bagging	95.0	85.18
3	Neural Network	95.0	166.93

APPENDIX O


RMSE of Models
RMSE
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.2250
0.1418
0.1226
0.1624
Decision Tree
XGBoost
Random Forest
Neural Network