



Predicting Prices for AirBnB

Shuhan Xia, Netra Pathak, Ashwin Tarikere

Motivation and Related Work



- AirBnB allows private individuals to rent out accommodation to tourists for a charge.
- Market price can be influenced by hundreds of attributes, while no two properties are the same.
- The company felt the need to design an algorithm that could suggest a price to the owner.
- They built their own tool using three main types of data: *Similarity*, *Recency*, *Location*. Actual number of variables in the thousands.

Our goal: To isolate the most important predictors and design a regression model to predict the listing price based on them.

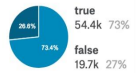
Some related work:

- Hill, “How much is your spare room worth?”, *IEEE Spectrum*, 2015.
- Teubner et al, “Price determinants on Airbnb: How reputation pays off in the sharing economy”, *J. Self-Governance & Management Economics*, 2017.
- Wang and Nicolau, “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com”, *Int. J. Hospitality Management*, 2017.

The data

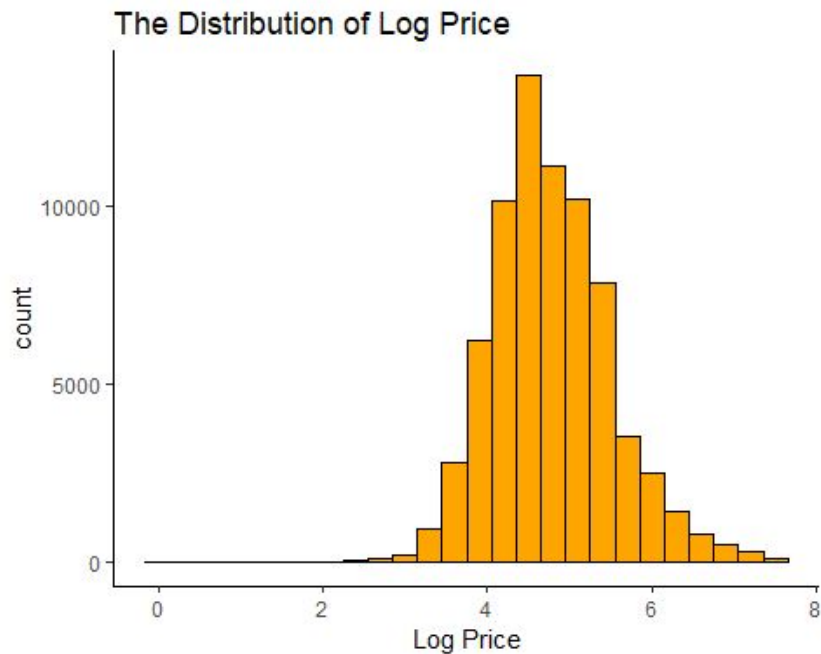
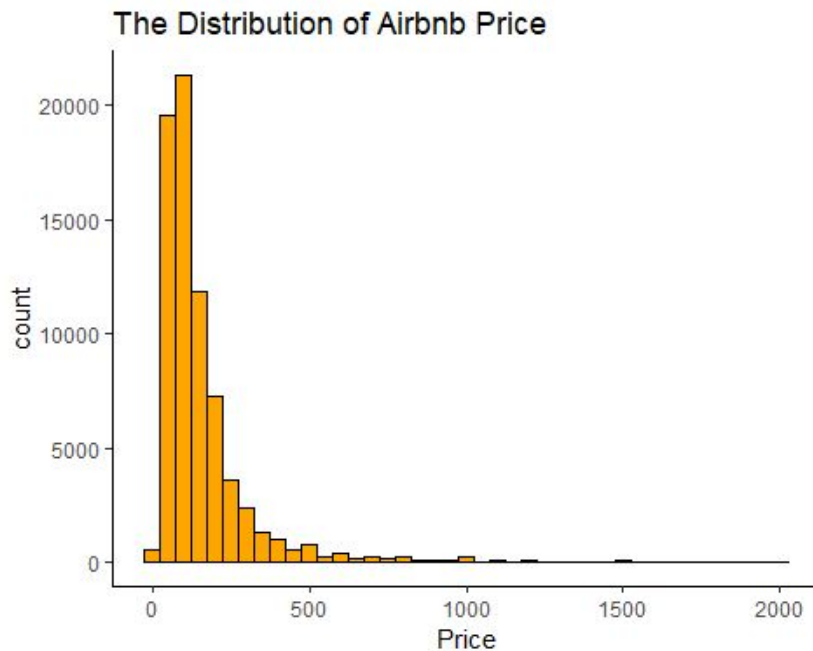
We obtained the dataset from Kaggle

- Over 70,000 observations
- 27 predictors, most of them categorical
- We split the dataset by random into two parts - one for training and the other for testing.

property_type	room_type	amenities	# accommodates	bathrooms	bed_type	cancellation_policy	cleaning_fee	city	description
Apartment 66% House 22% Other (33) 12%	Entire home/apt 56% Private room 41% Other (1) 3%	67122 unique values		1.0 78% 2.0 11% Other (15) 11%	Real Bed 97% Futon 1% Other (3) 2%	strict 44% flexible 30% Other (3) 26%		NYC 44% LA 30% Other (4) 26%	73479 unique values
Apartment	Entire home/apt	{ "Wireless Internet", "Air conditioning", "Kitchen", "Heating", "Family/kid friendly", "Essentials", "Hair dryer", "Iron", "translation missing: en.hosting_amenity_50" }	3	1.0	Real Bed	strict	True	NYC	Beautiful, sunlit brownstone 1-bedroom in the loveliest neighborhood in Brooklyn. Blocks from the promenade and Brooklyn Bridge Park, with their stunning views of Manhattan, and from the great shoppin...

Preliminary Analysis

Price is heavily right-skewed, so we took $\log(\text{Price})$ as the response variable.

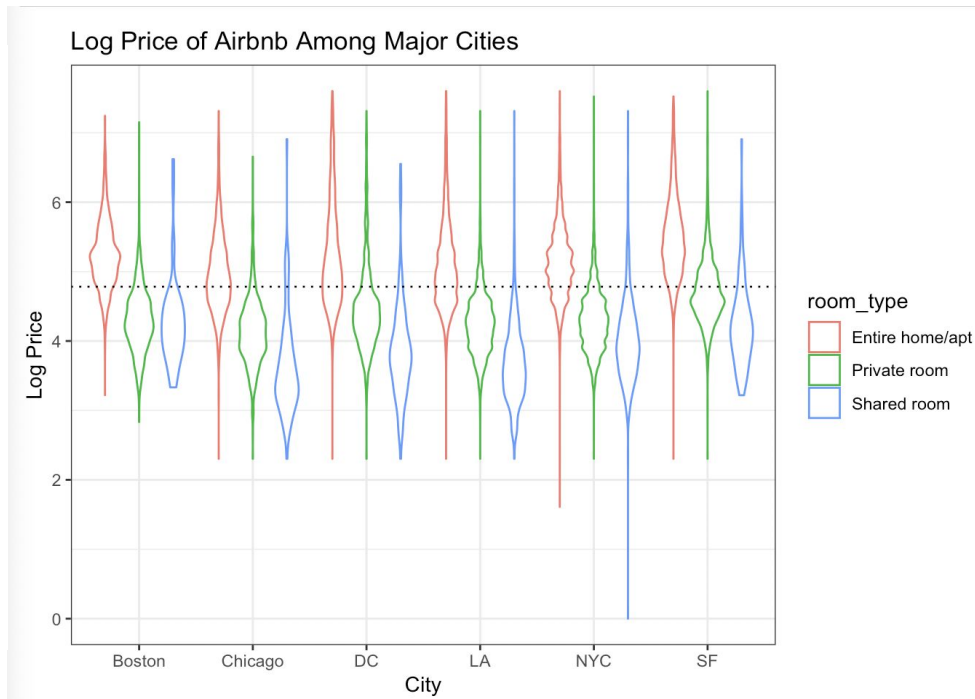
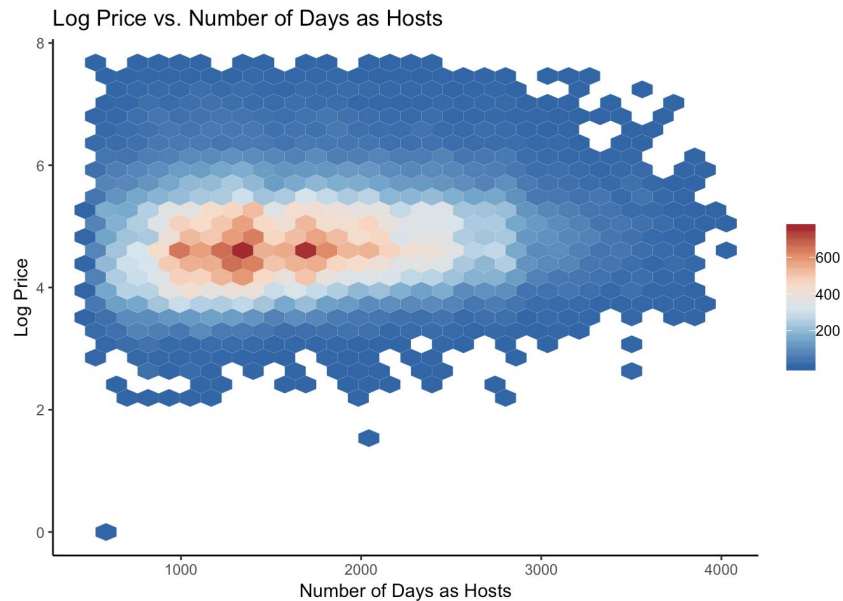


Data Cleaning

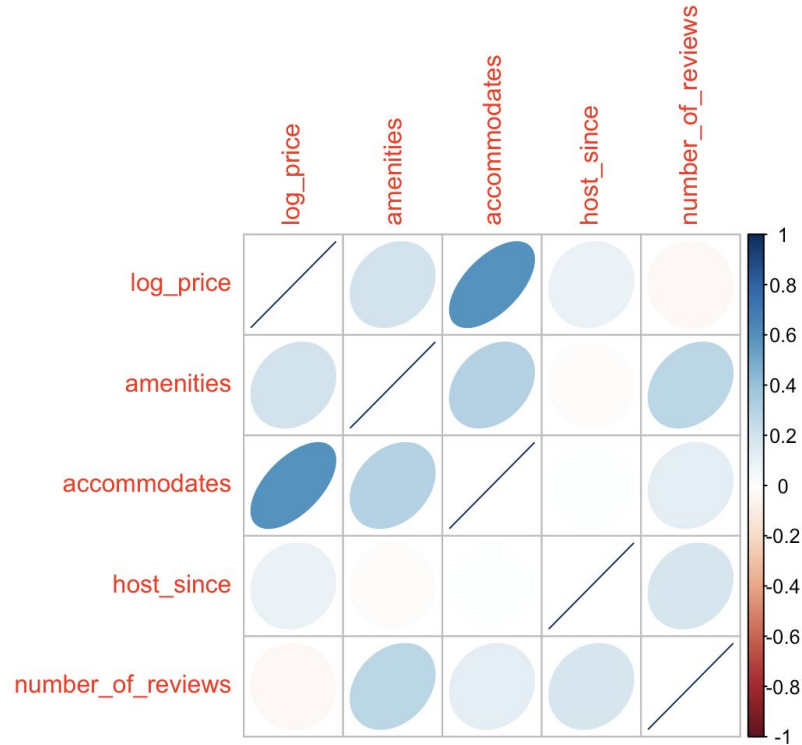


- Some variables, such as id and url are obviously irrelevant.
- Dropped variables which are hard to quantify, such as description and neighborhood.
- Dropped all location variables other than city name, such as latitude/longitude. We only had data from 6 cities.
- Dropped variables with too many NA values.
- Replaced the list of amenities by a simple count of the number of amenities.
- For categorical variables, combined certain factor levels with less observations in one. For example, the different types of strict cancellation policies into one factor.
- Dropped property types with <700 observations, such as castles, boats, tipis,..
- Changed numerical predictors such as bedrooms, etc. to factors.

Exploratory Data Analysis



Exploratory Data Analysis



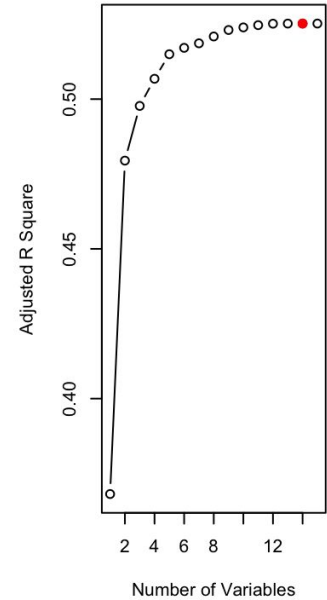
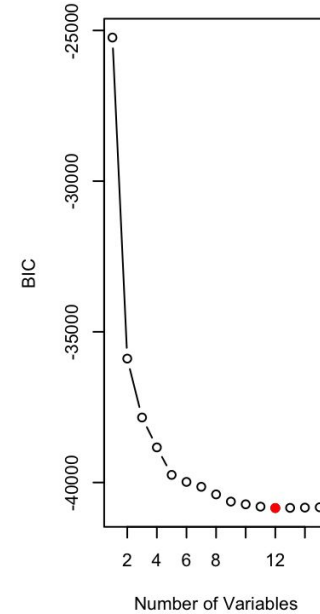
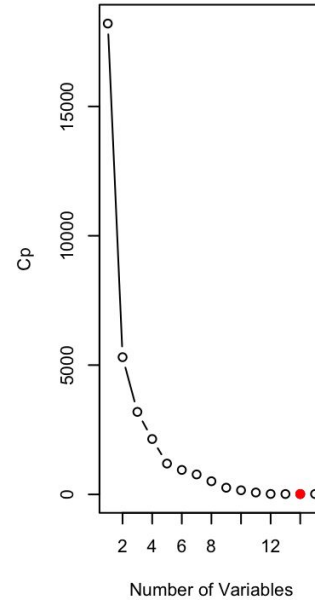
Methods



- Linear Regression Model
- Ridge Regression Model
- Elastic Net Regression Model
- Regression Tree Model
- Random Forest Model

Linear Regression Model

- Variables Selection
- Forward Stepwise Approach
- Criteria: Cp, BIC, Adjusted R²



Linear Regression Model

- Variables Selection
- Forward Stepwise Approach
- Criteria: Cp, BIC, Adjusted R^2
- Least Important Variable:
real_bed
- Followed by: property_type;
cancellation_policy

Selection Algorithm: forward

		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
1	(1)	"	"	"*	"	"	"	"	"	"	"	"	"	"	"	"
2	(1)	"	"	"*	"	"	"	"	"	"	"	"	"	"	"	"*
3	(1)	"	"	"*	"	"	"*	"	"	"	"	"	"	"	"	"*
4	(1)	"	"	"*	"	"	"*	"	"	"	"	"*	"	"	"	"*
5	(1)	"	"	"*	"	"	"*	"*	"	"	"	"	"	"	"	"*
6	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"	"	"*
7	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"	"	"*
8	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"	"	"*
9	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"	"	"*
10	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"	"*
11	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"*	"
12	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"*	"*
13	(1)	"	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"*	"*
14	(1)	"*	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"*	"*
15	(1)	"*	"	"*	"	"	"*	"*	"	"	"	"*	"*	"*	"*	"*

Linear Regression Model

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.522e+00  3.161e-02  143.052 < 2e-16 ***
## property_typeOthers  2.997e-02  1.075e-02   2.789  0.005285 **
## property_typeCondominium  1.369e-01  1.131e-02  12.102 < 2e-16 ***
## property_typeHouse -3.950e-02  5.788e-03  -6.825  8.89e-12 ***
## property_typeLoft   1.508e-01  1.612e-02   9.357 < 2e-16 ***
## property_typeTownhouse -2.890e-02  1.388e-02  -2.083  0.037299 *
## room_typePrivate room -5.898e-01  5.590e-03 -105.518 < 2e-16 ***
## room_typeShared room -1.030e+00  1.309e-02 -78.624 < 2e-16 ***
## amenities         4.947e-03  3.269e-04  15.134 < 2e-16 ***
## accommodates      7.330e-02  1.795e-03  40.838 < 2e-16 ***
## bathrooms1       1.383e-01  2.826e-02   4.894  9.89e-07 ***
## bathrooms1.5     1.853e-01  2.957e-02   6.266  3.73e-10 ***
## bathrooms2       2.637e-01  2.908e-02   9.067 < 2e-16 ***
## bathrooms> 2     4.633e-01  3.041e-02  15.236 < 2e-16 ***
## cancellation_policymoderate -5.270e-02  5.813e-03  -9.067 < 2e-16 ***
## cancellation_policystrict -1.528e-02  5.379e-03  -2.842  0.004489 **
## cleaning_feeTrue  -6.350e-02  5.121e-03  -12.401 < 2e-16 ***
## cityChicago      -3.054e-01  1.312e-02 -23.284 < 2e-16 ***
## cityDC           4.286e-02  1.205e-02   3.557  0.000375 ***
## cityLA          -1.266e-01  1.031e-02 -12.276 < 2e-16 ***
## cityNYC         -1.425e-02  1.006e-02  -1.417  0.156537
## citySF           3.155e-01  1.176e-02  26.826 < 2e-16 ***
## host_identity_verifiedt -3.885e-02  4.759e-03  -8.163  3.33e-16 ***
## host_since       5.168e-05  3.454e-06   14.965 < 2e-16 ***
## instant_bookablet -5.136e-02  4.800e-03 -10.699 < 2e-16 ***
## number_of_reviews -9.401e-04  5.838e-05 -16.104 < 2e-16 ***
## bedrooms1       7.429e-02  7.857e-03   9.455 < 2e-16 ***
## bedrooms2       2.517e-01  1.015e-02  24.803 < 2e-16 ***
## bedrooms3       4.468e-01  1.451e-02  30.794 < 2e-16 ***
## bedrooms>3      6.387e-01  2.064e-02  30.943 < 2e-16 ***
## beds2          -9.519e-03  6.272e-03  -1.518  0.129071
## beds3          -4.101e-02  1.023e-02  -4.009  6.11e-05 ***
## beds>3         -1.513e-01  1.331e-02 -11.368 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4784 on 54982 degrees of freedom
## Multiple R-squared:  0.5569, Adjusted R-squared:  0.5567
## F-statistic: 2160 on 32 and 54982 DF, p-value: < 2.2e-16
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.527e+00  3.160e-02  143.250 < 2e-16 ***
## room_typePrivate room -5.981e-01  5.466e-03 -109.422 < 2e-16 ***
## room_typeShared room -1.028e+00  1.311e-02  -78.372 < 2e-16 ***
## amenities      5.084e-03  3.241e-04  15.688 < 2e-16 ***
## accommodates    7.441e-02  1.796e-03  41.425 < 2e-16 ***
## bathrooms1     1.357e-01  2.831e-02   4.795  1.63e-06 ***
## bathrooms1.5   1.794e-01  2.962e-02   6.057  1.39e-09 ***
## bathrooms2     2.658e-01  2.913e-02   9.124 < 2e-16 ***
## bathrooms> 2   4.555e-01  3.046e-02  14.955 < 2e-16 ***
## cleaning_feeTrue -7.216e-02  4.938e-03 -14.614 < 2e-16 ***
## cityChicago    -3.049e-01  1.316e-02 -23.175 < 2e-16 ***
## cityDC         2.781e-02  1.202e-02   2.314  0.0207 *
## cityLA        -1.429e-01  1.022e-02 -13.988 < 2e-16 ***
## cityNYC       -12.194e-02  1.003e-02  -2.186  0.0288 *
## citySF        3.077e-01  1.176e-02  26.160 < 2e-16 ***
## host_identity_verifiedt -4.155e-02  4.763e-03  -8.724 < 2e-16 ***
## host_since     5.162e-05  3.455e-06  14.940 < 2e-16 ***
## instant_bookablet -5.030e-02  4.816e-03 -10.443 < 2e-16 ***
## number_of_reviews -1.041e-03  5.785e-05 -17.995 < 2e-16 ***
## bedrooms1      7.151e-02  7.868e-03   9.089 < 2e-16 ***
## bedrooms2     2.462e-01  1.012e-02  24.319 < 2e-16 ***
## bedrooms3     4.292e-01  1.442e-02  29.768 < 2e-16 ***
## bedrooms>3    6.101e-01  2.054e-02  29.701 < 2e-16 ***
## beds2        -1.172e-02  6.288e-03  -1.865  0.0622 .
## beds3        -4.531e-02  1.026e-02  -4.416  1.01e-05 ***
## beds>3       -1.572e-01  1.334e-02 -11.781 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4802 on 54989 degrees of freedom
## Multiple R-squared:  0.5535, Adjusted R-squared:  0.5533
## F-statistic: 2727 on 25 and 54989 DF, p-value: < 2.2e-16
```

Linear Regression Model

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.153e+00  3.471e-02  119.640 < 2e-16 ***
## property_typeApartment    -2.997e-02  1.075e-02   -2.789  0.005285 **
## property_typeCondominium    1.069e-01  1.507e-02    7.096  1.30e-12 ***
## property_typeHouse    -6.948e-02  1.125e-02   -6.178  6.53e-10 ***
## property_typeLoft    1.208e-01  1.895e-02    6.376  1.84e-10 ***
## property_typeTownhouse    -5.887e-02  1.703e-02   -3.456  0.000549 ***
## room_typePrivate room    -5.898e-01  5.590e-03  -105.518 < 2e-16 ***
## room_typeShared room    -1.030e+00  1.309e-02  -78.624 < 2e-16 ***
## amenities    4.947e-03  3.269e-04   15.134 < 2e-16 ***
## accommodates    7.330e-02  1.795e-03   40.838 < 2e-16 ***
## bathrooms1    1.383e-01  2.826e-02   4.894  9.89e-07 ***
## bathrooms1.5    1.853e-01  2.957e-02   6.266  3.73e-10 ***
## bathrooms2    2.637e-01  2.908e-02   9.067 < 2e-16 ***
## bathrooms> 2    4.633e-01  3.041e-02  15.236 < 2e-16 ***
## cancellation_policyflexible    5.270e-02  5.813e-03    9.067 < 2e-16 ***
## cancellation_policystrict    3.742e-02  5.124e-03    7.303  2.86e-13 ***
## cleaning_feeTrue    -6.350e-02  5.121e-03  -12.401 < 2e-16 ***
## cityBoston    3.054e-01  1.312e-02  23.284 < 2e-16 ***
## cityDC    3.483e-01  1.177e-02  29.585 < 2e-16 ***
## cityLA    1.788e-01  9.998e-03  17.886 < 2e-16 ***
## cityNYC    2.912e-01  9.757e-03  29.842 < 2e-16 ***
## citySF    6.209e-01  1.149e-02  54.047 < 2e-16 ***
## host_identity_verifiedt    -3.885e-02  4.759e-03   -8.163  3.33e-16 ***
## host_since    5.168e-05  3.454e-06   14.965 < 2e-16 ***
## instant_bookablet    -5.136e-02  4.800e-03  -10.699 < 2e-16 ***
## number_of_reviews    -9.401e-04  5.838e-05  -16.104 < 2e-16 ***
## bedrooms1    7.429e-02  7.857e-03    9.455 < 2e-16 ***
## bedrooms2    2.517e-01  1.015e-02  24.803 < 2e-16 ***
## bedrooms3    4.468e-01  1.451e-02  30.794 < 2e-16 ***
## bedrooms>3    6.387e-01  2.064e-02  30.943 < 2e-16 ***
## beds<=1    4.101e-02  1.023e-02    4.009  6.11e-05 ***
## beds2    3.149e-02  9.172e-03    3.433  0.000597 ***
## beds>3    -1.103e-01  1.155e-02   -9.547 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4784 on 54982 degrees of freedom
## Multiple R-squared:  0.5569, Adjusted R-squared:  0.5567
## F-statistic: 2160 on 32 and 54982 DF, p-value: < 2.2e-16
```

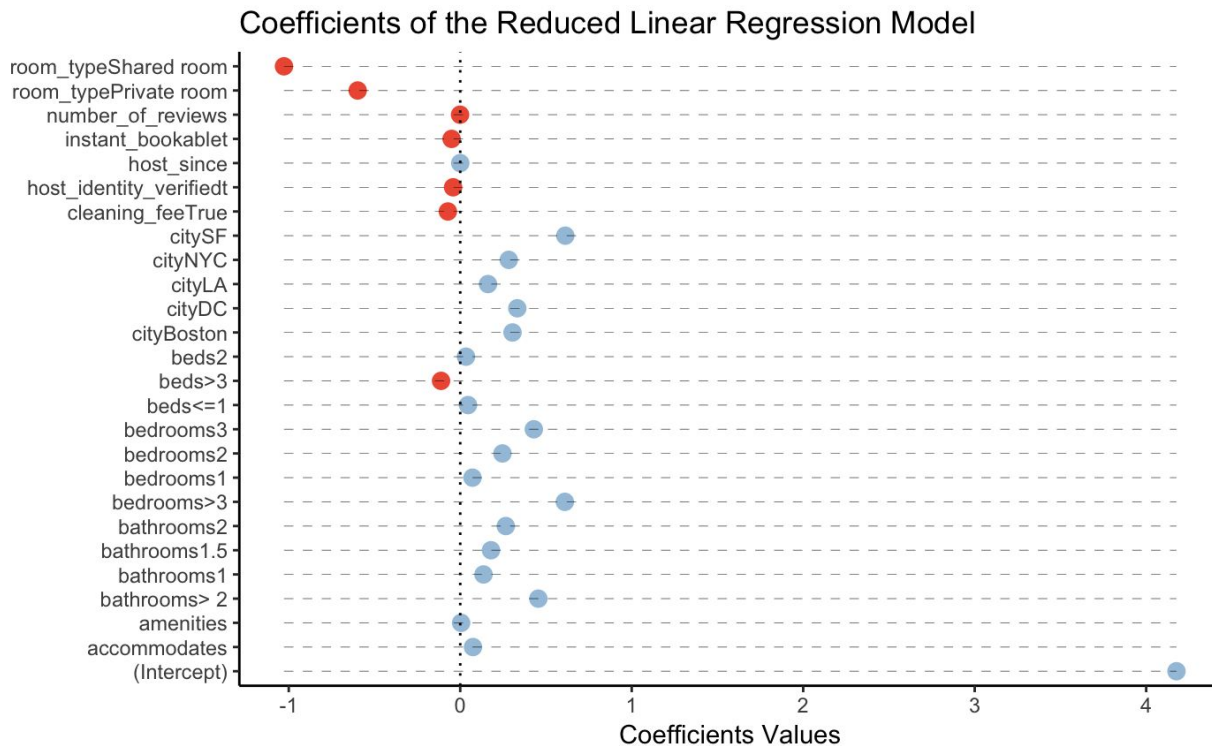
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.177e+00  3.359e-02  124.336 < 2e-16 ***
## room_typePrivate room    -5.981e-01  5.466e-03  -109.422 < 2e-16 ***
## room_typeShared room    -1.028e+00  1.311e-02  -78.372 < 2e-16 ***
## amenities    5.084e-03  3.241e-04   15.688 < 2e-16 ***
## accommodates    7.441e-02  1.796e-03   41.425 < 2e-16 ***
## bathrooms1    1.357e-01  2.831e-02   4.795  1.63e-06 ***
## bathrooms1.5    1.794e-01  2.962e-02   6.057  1.39e-09 ***
## bathrooms2    2.658e-01  2.913e-02   9.124 < 2e-16 ***
## bathrooms> 2    4.555e-01  3.046e-02  14.955 < 2e-16 ***
## cleaning_feeTrue    -7.216e-02  4.938e-03  -14.614 < 2e-16 ***
## cityBoston    3.049e-01  1.316e-02  23.175 < 2e-16 ***
## cityDC    3.327e-01  1.174e-02  28.336 < 2e-16 ***
## cityLA    1.620e-01  9.882e-03  16.392 < 2e-16 ***
## cityNYC    2.830e-01  9.726e-03  29.092 < 2e-16 ***
## citySF    6.126e-01  1.149e-02  53.310 < 2e-16 ***
## host_identity_verifiedt    -4.155e-02  4.763e-03   -8.724 < 2e-16 ***
## host_since    5.162e-05  3.455e-06   14.940 < 2e-16 ***
## instant_bookablet    -5.030e-02  4.816e-03  -10.443 < 2e-16 ***
## number_of_reviews    -1.041e-03  5.785e-05  -17.995 < 2e-16 ***
## bedrooms1    7.151e-02  7.868e-03    9.089 < 2e-16 ***
## bedrooms2    2.462e-01  1.012e-02  24.319 < 2e-16 ***
## bedrooms3    4.292e-01  1.442e-02  29.768 < 2e-16 ***
## bedrooms>3    6.101e-01  2.054e-02  29.701 < 2e-16 ***
## beds<=1    4.531e-02  1.026e-02    4.416  1.01e-05 ***
## beds2    3.359e-02  9.205e-03    3.649  0.000264 ***
## beds>3    -1.118e-01  1.159e-02   -9.650 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4802 on 54989 degrees of freedom
## Multiple R-squared:  0.5535, Adjusted R-squared:  0.5533
## F-statistic: 2727 on 25 and 54989 DF, p-value: < 2.2e-16
```

Linear Regression Model

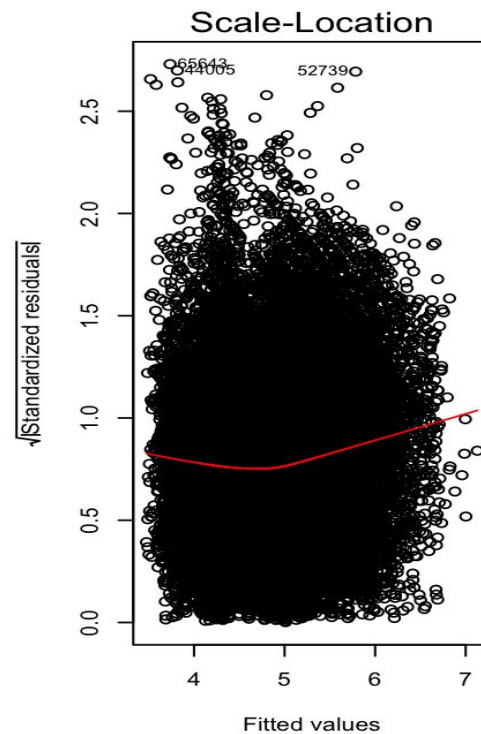
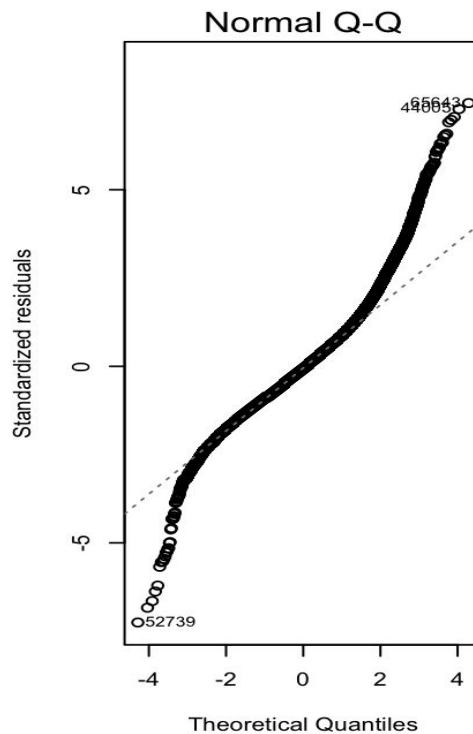
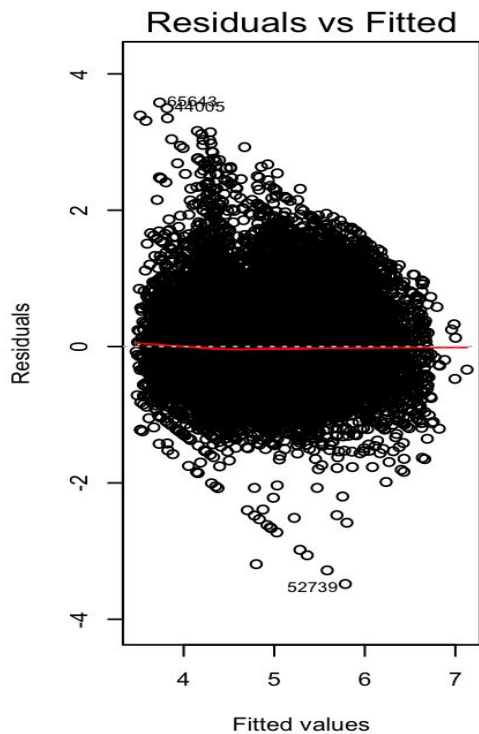
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	54982	12583				
2	54989	12680	-7	-97.034	60.571	< 2.2e-16 ***

- The coefficients of the property_type and cancellation_policy variables are all significant
- There is only trivial improvement for the model fitting

Linear Regression Model



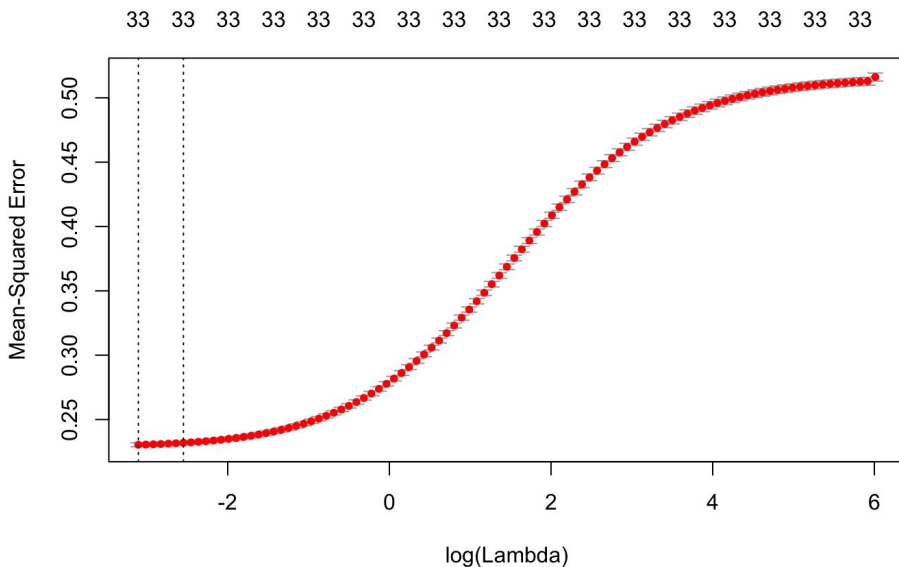
Linear Regression Model



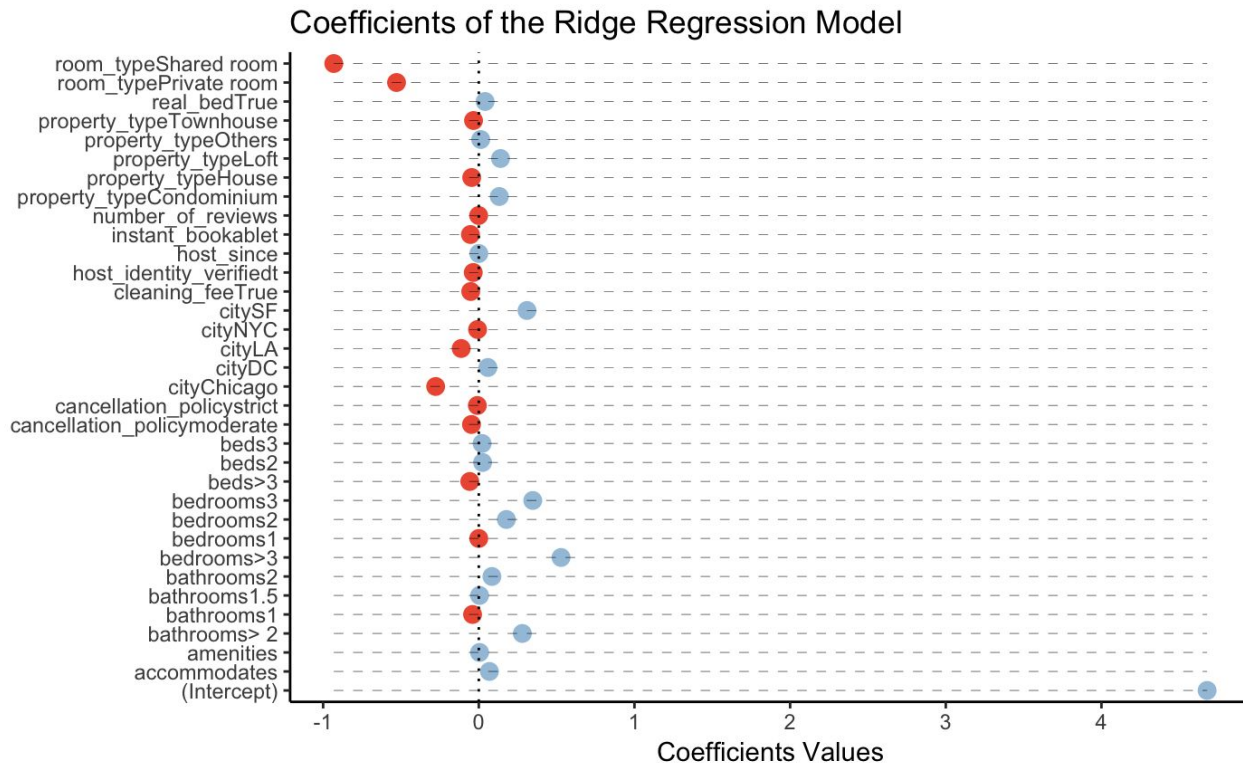
Ridge Regression Model

- A regularization method that adds penalty term
- Ridge regression coefficient estimates minimize:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_j \beta_j^2.$$

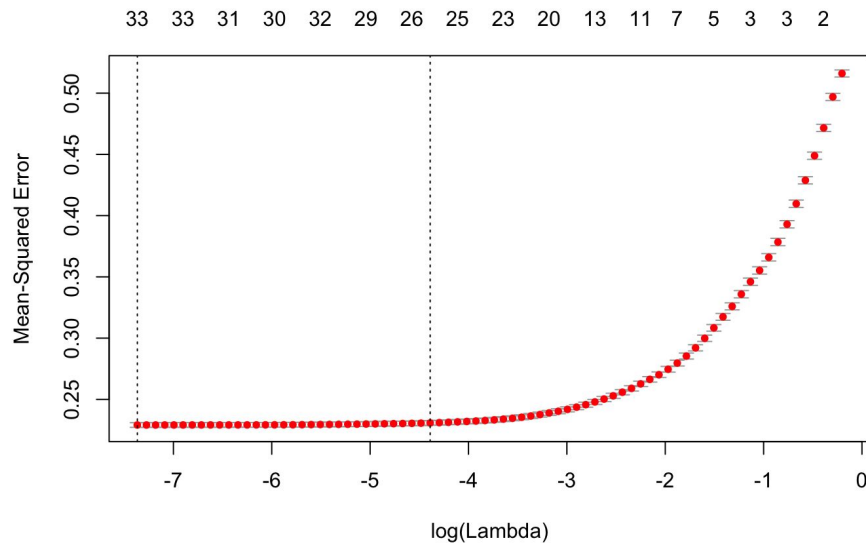


Ridge Regression Model

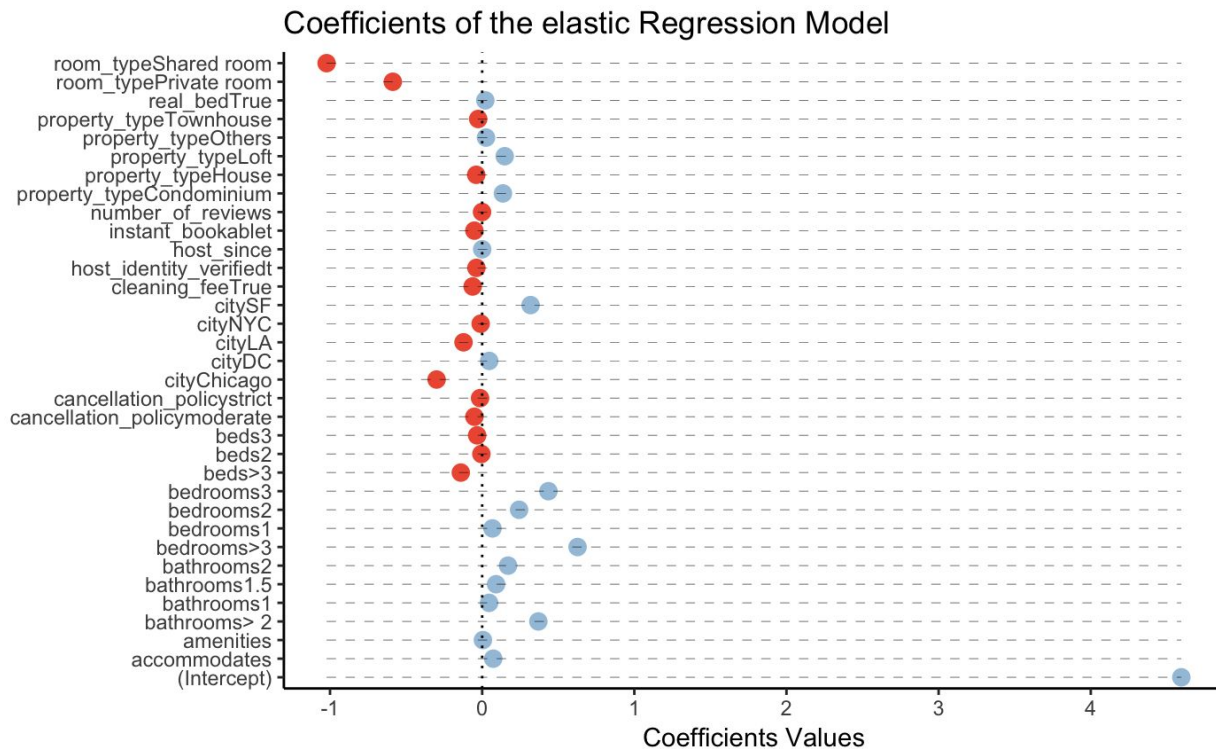


Elastic Net Regression Model

- Lasso is not adopted because it will shrink the model by letting some coefficients to be exactly zero.
- A regularization method that linearly combines the penalties of the lasso and ridge methods

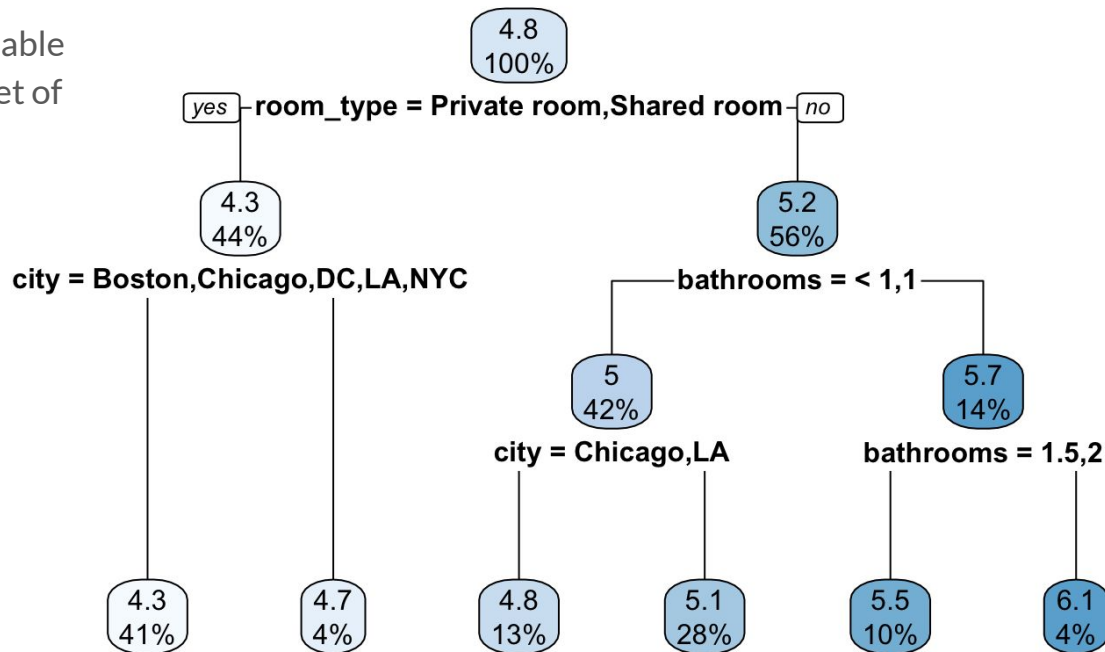


Elastic Net Regression Model



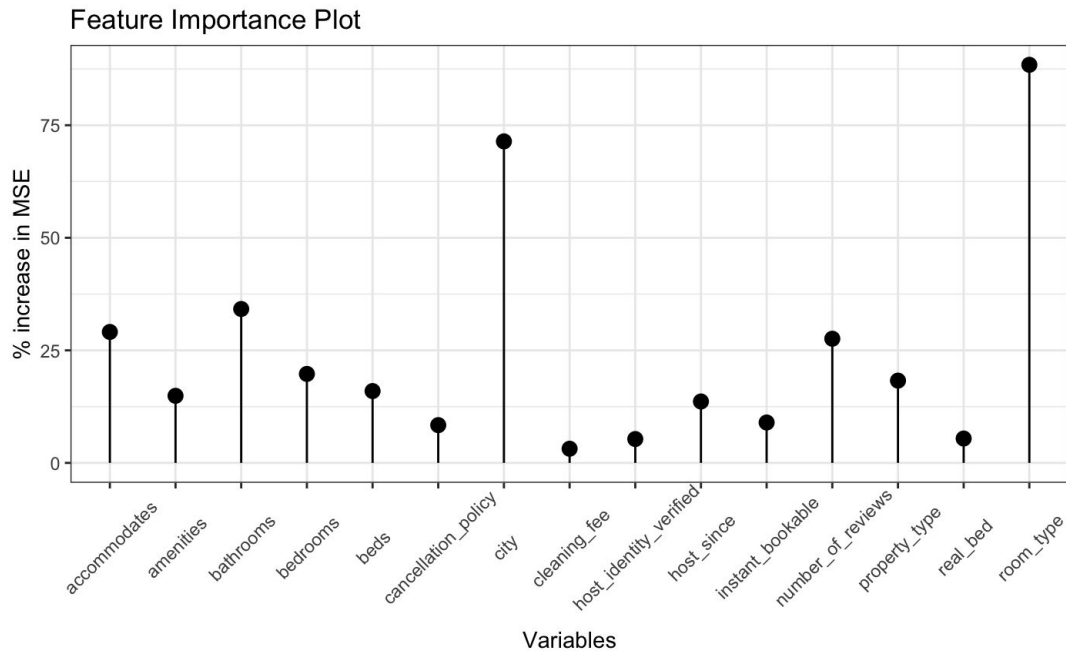
Regression Tree Model

- Works top-down, choosing a variable at each step that best splits the set of items.



Random Forest Model

- An ensemble of regression trees
- Avoid highly correlated predictions
- The result is not easy to interpret
- Variable importance plot



Results



The Mean Square Error of the test dataset for each method

- **Linear Regression Model:** 0.2263 for the full model; 0.2282 for the reduced model
- **Ridge Regression Model:** 0.2277
- **Elastic Net Regression Model:** 0.2264
- **Regression Tree Model:** 0.2578
- **Random Forest Model:** 0.2039