Experiment 1: Random forest regression on yelp user information and their rating deviation

To investigate the relation between yelp user information and their rating deviation of a specific restaurant, I select 5 features to define a yelp user (num of friends, num of reviews, num of photos, living area, elite user or not) and fit with their rating deviation to the rating mean of one restaurant. Data were scraped from 71432 yelp users who rated 97 restaurants in atlanta in total.

y = user rating deviation =

I used binary vector for representing the user area (in Atlanta or not) and ‘elite2020’ label (elite user or not)

area\_AL:

if user in Atlanta 1

else 0

elite\_user:

if user is elite 1

else 0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| friends | number\_reviews | photos | area\_AL | elite\_user | rating\_average |
| 414 | 9 | 0 | 1 | 0 | 0.137585 |
| 0 | 13 | 6 | 0 | 0 | 0.137585 |
| 0 | 44 | 27 | 0 | 0 | 0.137585 |
| 545 | 1379 | 9924 | 0 | 1 | 0.137585 |
| 366 | 328 | 1102 | 0 | 1 | 0.137585 |
| 174 | 71 | 95 | 1 | 1 | -0.08993 |
| 45 | 47 | 246 | 0 | 1 | -0.08993 |
| 189 | 181 | 312 | 0 | 1 | -0.08993 |
| 21 | 71 | 317 | 0 | 1 | -0.08993 |

**part of codes:**

x\_train, x\_test, y\_train, y\_test = train\_test\_split( x\_data, y\_data, test\_size=0.30, shuffle=True, random\_state= 0)

rf = RandomForestRegressor(n\_estimators=20, random\_state=0)

rf.fit(x\_train, y\_train)

y\_pred = rf.predict(x\_test)

print('Mean Absolute Error:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('Mean Squared Error:', metrics.mean\_squared\_error(y\_test, y\_pred))

print('Root Mean Squared Error:', np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred)))

importances = rf.feature\_importances\_

print(importances)

**Error (MAE) and feature importance report:**

n\_estimator = 5:

Mean Absolute Error: 0.2244267073974103

[0.38660446 0.35181441 0.23128492 0.01882687 0.01146934]

n\_estimator = 10:

Mean Absolute Error: 0.2209094613515204

[0.39135136 0.34868485 0.22924052 0.02021638 0.01050689]

n\_estimator = 20:

Mean Absolute Error: 0.21921600351026954

[0.38880278 0.35205326 0.22884879 0.01966514 0.01063003]

n\_estimator = 30:

Mean Absolute Error: 0.21844714602669382

[0.3908963 0.34822915 0.22996935 0.0204014 0.0105038]

n\_estimator = 40:

Mean Absolute Error: 0.21816785556191073

[0.38874473 0.34877435 0.23146044 0.0205317 0.01048877]

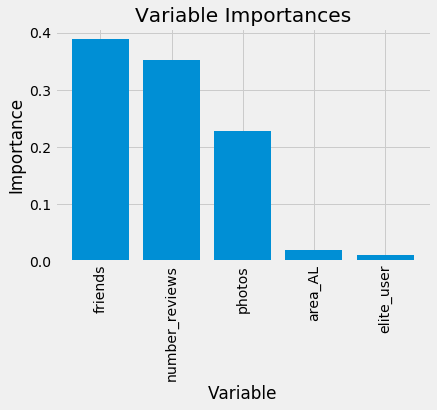


Figure 1. bar chart illustration of feature importance (random forest n\_estimator = 20).

**For regression on data after fake date detection (n\_estimator = 20):**

Mean Absolute Error: 0.22143776405341808

[0.38474885 0.34113219 0.23889948 0.02485705 0.01036242]

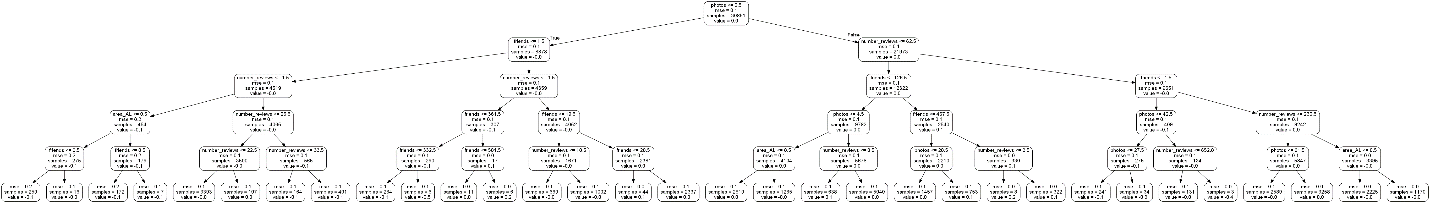
Variable: friends Importance: 0.38

Variable: number\_reviews Importance: 0.34

Variable: photos Importance: 0.24

Variable: area\_AL Importance: 0.02

Variable: elite\_user Importance: 0.01



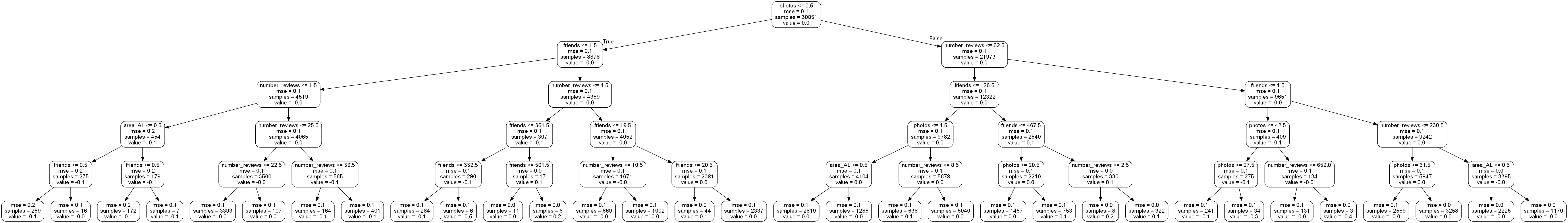


Figure 2. Decision tree plotting for my random forest regressor model.

Experiment 2: Using random forest regression to predict restaurants’ average rating with restaurant features (num of reviews, neighborhood, price level, restaurant type) from 2351 restaurants in Atlanta:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| review\_count | average\_rating | neighborhood\_Buckhead | neighborhood\_East Lake | neighborhood\_Grant Park | price\_$ | price\_$$ | price\_$$$ | price\_$$$$ | type\_african | type\_american | type\_asian |
| 6  +737 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 53 | 4.5 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 995 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 749 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 5 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 434 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1341 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 143 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 225 | 4 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

n\_estimators = 100

Mean Absolute Error: 0.5956120883758985

Variable: review\_count Importance: 0.4387

Variable: type\_fastfood Importance: 0.0624

Variable: neighborhood\_Buckhead Importance: 0.0507

Variable: price\_1 Importance: 0.0427

Variable: neighborhood\_Westside / Home Park Importance: 0.0348

Variable: neighborhood\_West End Importance: 0.0315

Variable: price\_2 Importance: 0.0253

Variable: neighborhood\_Morningside / Lenox Park Importance: 0.022

Variable: type\_asian Importance: 0.0197

Variable: type\_american Importance: 0.0194

Variable: neighborhood\_West Paces Ferry / Northside Importance: 0.0191

Variable: type\_meat Importance: 0.0188

Variable: neighborhood\_Ormewood Park Importance: 0.0186

Variable: type\_mexican Importance: 0.0175

Variable: type\_cafe Importance: 0.0164

Variable: neighborhood\_Kirkwood Importance: 0.015

Variable: neighborhood\_East Lake Importance: 0.0141

Variable: type\_european Importance: 0.0137

Variable: type\_bar Importance: 0.0132

Variable: type\_caribbean Importance: 0.012

Variable: type\_others Importance: 0.0112

Variable: type\_breakfast Importance: 0.0111

Variable: type\_bakery Importance: 0.0103

Variable: price\_3 Importance: 0.0095

Variable: type\_specificDiet Importance: 0.0089

Variable: neighborhood\_Poncey - Highland Importance: 0.0071

Variable: type\_african Importance: 0.0055

Variable: neighborhood\_Reynoldstown Importance: 0.0054

Variable: type\_mediterranean Importance: 0.0041

Variable: type\_salad Importance: 0.004

Variable: neighborhood\_Old Fourth Ward Importance: 0.0039

Variable: type\_foodmarket Importance: 0.0038

Variable: type\_buffet Importance: 0.0034

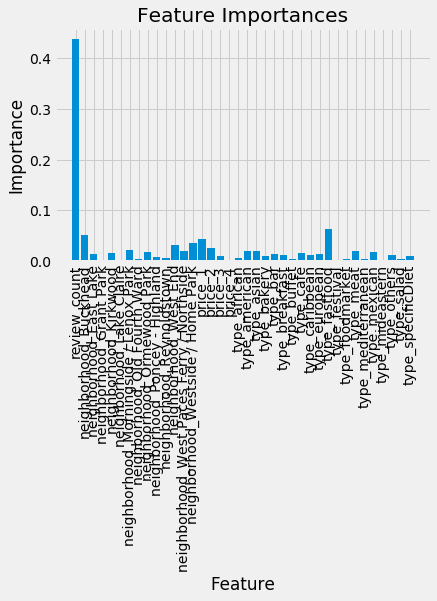
Variable: price\_4 Importance: 0.0024

Variable: neighborhood\_Lake Claire Importance: 0.0014

Variable: type\_festival Importance: 0.001

Variable: type\_mideastern Importance: 0.001

Variable: neighborhood\_Grant Park Importance: 0.0005



Because the importance of review numbers to restaurant’s rating is a lot, I dropped this feature and fit with random forest regression model again:

n\_estimators = 100

Mean Absolute Error: 0.5900246939434137

Variable: type\_fastfood Importance: 0.1677

Variable: neighborhood\_Buckhead Importance: 0.0957

Variable: price\_$ Importance: 0.0875

Variable: price\_$$ Importance: 0.0754

Variable: neighborhood\_West End Importance: 0.0563

Variable: neighborhood\_Westside / Home Park Importance: 0.0369

Variable: neighborhood\_Morningside / Lenox Park Importance: 0.0299

Variable: type\_mexican Importance: 0.0299

Variable: neighborhood\_West Paces Ferry / Northside Importance: 0.0278

Variable: neighborhood\_Ormewood Park Importance: 0.0257

Variable: neighborhood\_Kirkwood Importance: 0.0251

Variable: neighborhood\_East Lake Importance: 0.0232

Variable: type\_breakfast Importance: 0.023

Variable: type\_meat Importance: 0.0229

Variable: type\_others Importance: 0.0226

Variable: type\_asian Importance: 0.0206

Variable: type\_american Importance: 0.0199

Variable: price\_$$$ Importance: 0.0193

Variable: neighborhood\_Poncey - Highland Importance: 0.019

Variable: type\_specificDiet Importance: 0.0165

Variable: type\_bakery Importance: 0.0163

Variable: type\_cafe Importance: 0.0163

Variable: type\_bar Importance: 0.0148

Variable: type\_european Importance: 0.0148

Variable: type\_caribbean Importance: 0.0147

Variable: type\_african Importance: 0.0118

Variable: price\_$$$$ Importance: 0.0092

Variable: type\_mediterranean Importance: 0.0081

Variable: type\_buffet Importance: 0.0077

Variable: type\_foodmarket Importance: 0.0076

Variable: neighborhood\_Reynoldstown Importance: 0.0072

Variable: neighborhood\_Old Fourth Ward Importance: 0.0067

Variable: type\_salad Importance: 0.0059

Variable: neighborhood\_Lake Claire Importance: 0.0058

Variable: type\_festival Importance: 0.0042

Variable: type\_mideastern Importance: 0.0025

Variable: neighborhood\_Grant Park Importance: 0.0019

