A Foodie's Quest to Predict Yelp Restaurant Ratings

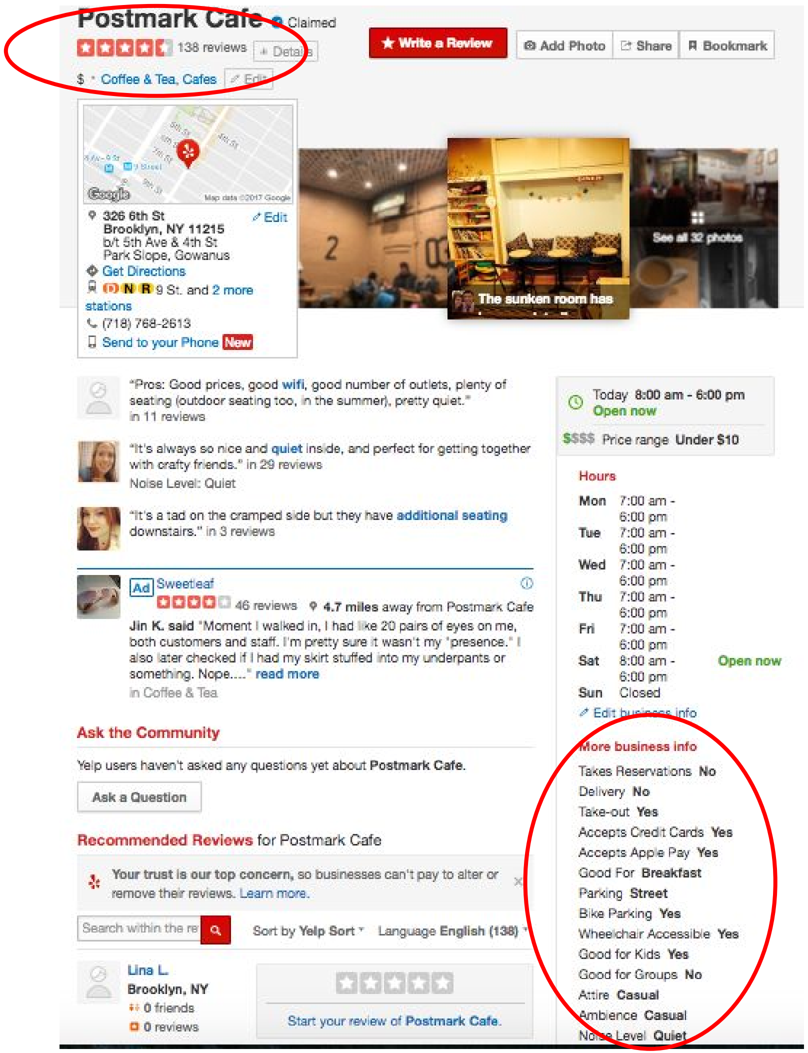
Lina Lavitsky
General Assembly ~ May 24, 2017

Project Overview

Can I predict Yelp ratings for restaurants using business characteristics?

Applications:

- Helpful for restaurant owners to know what features affect their ratings



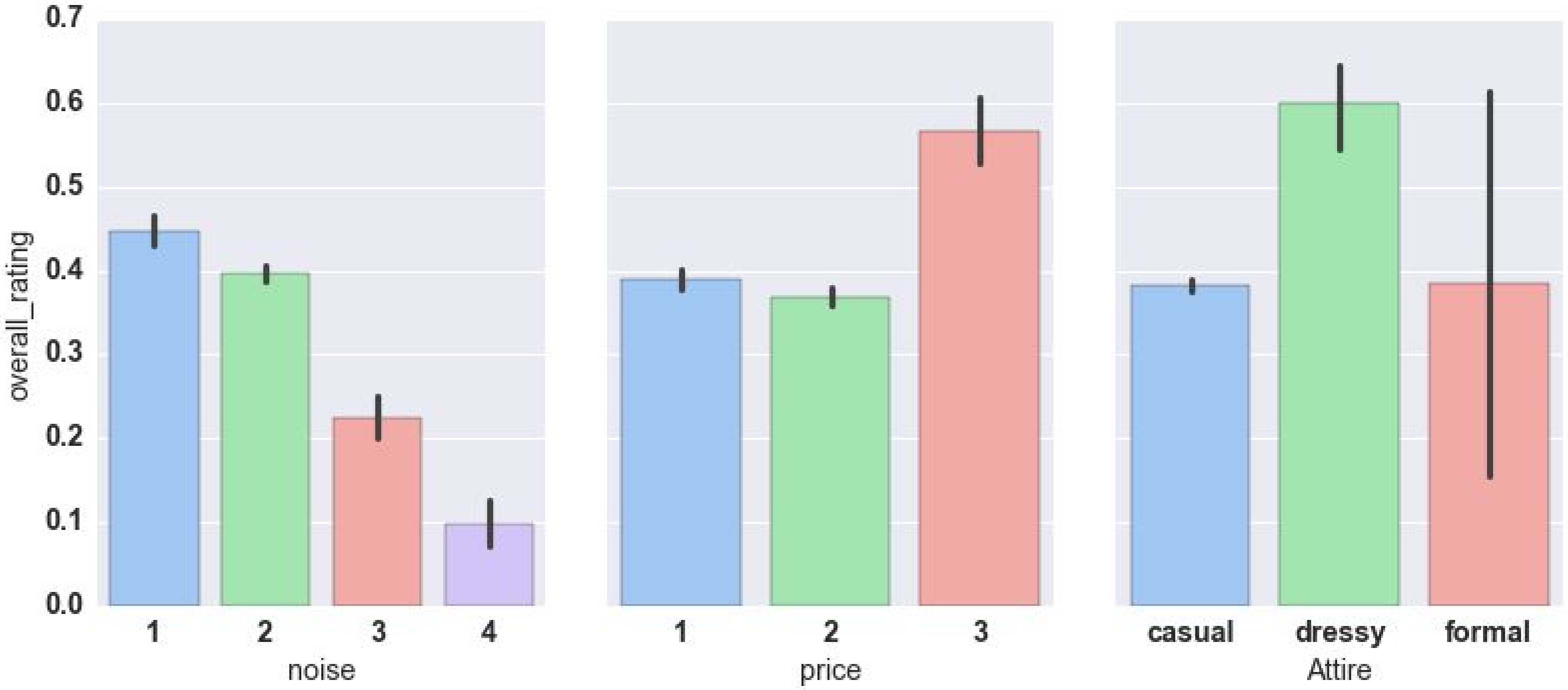
Data

- 72000 businesses of which 22000 are restaurants
- 89 different features, of which I isolated the ones that I thought could be most helpful in predicting rating review:

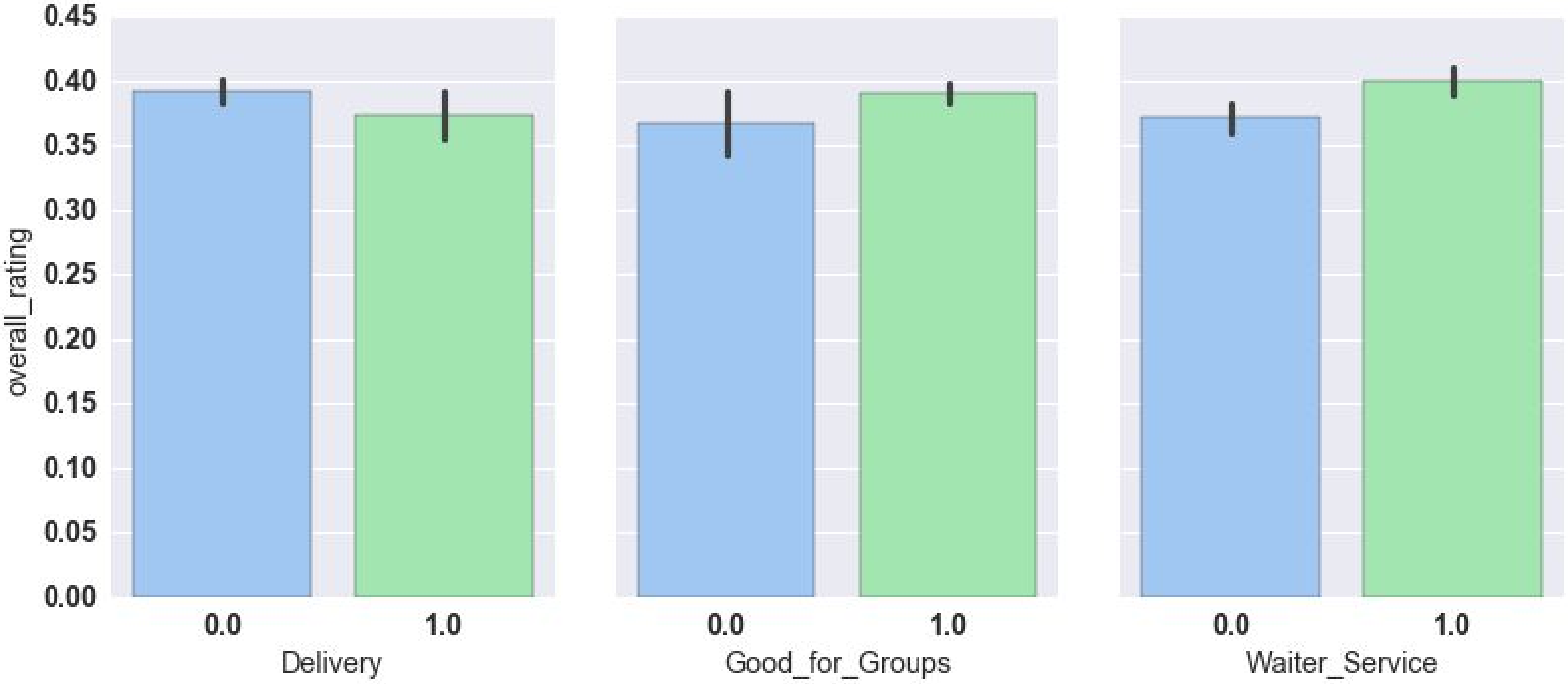
```
cols_to_keep = ['stars','state', 'review_count','Noise Level', 'Attire', 'Price_Range',
'Delivery', 'Good_for_Groups', 'Waiter_Service', 'Number_of_Checkins', 'Take_Out',
'Number_of_Tips']
```

- To simplify the question and make the distribution more even, I created a new variable in which I placed above median ratings into “good rating” bucket and median and below ratings into “not good rating” bucket
- Kept the 5 largest states (some states only had a few values each)
- Got rid of rows with null values in the columns I cared about
- Combined “$$$” and “$$$$” into one price category because there were fewer of those values
- Final data: 14,325 restaurants

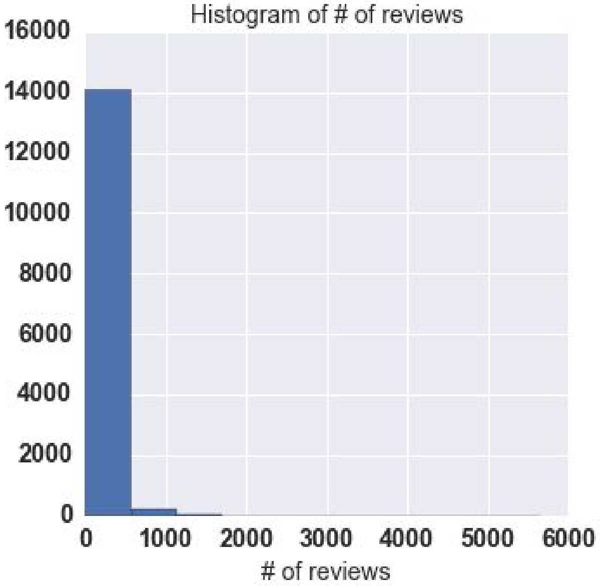
Exploring the data

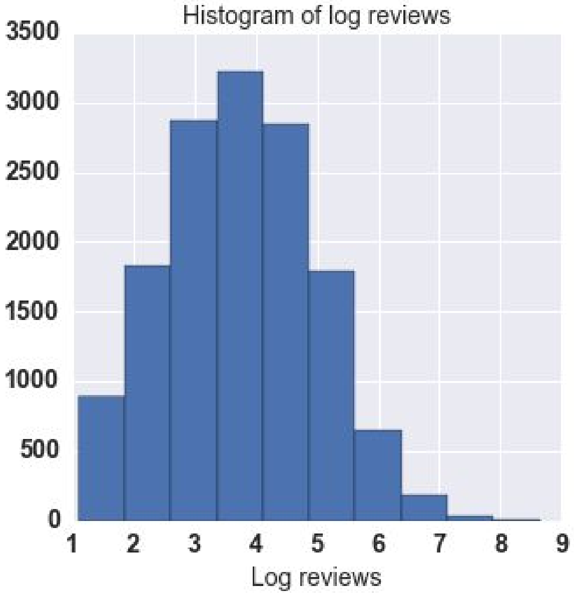


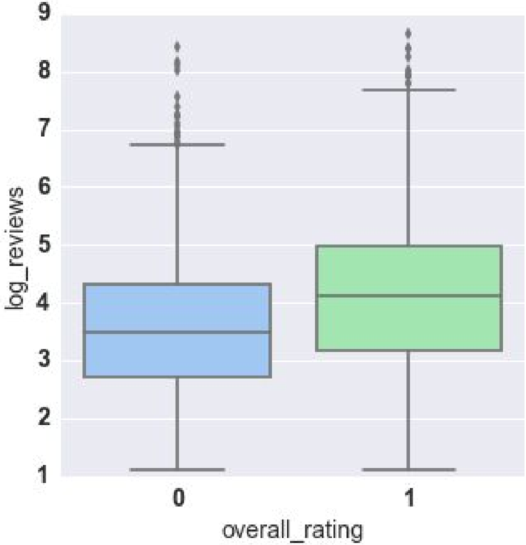
Exploring the data



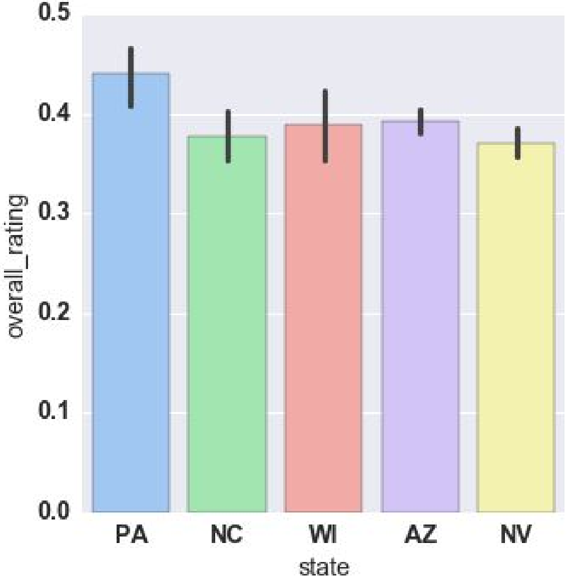
Exploring the data

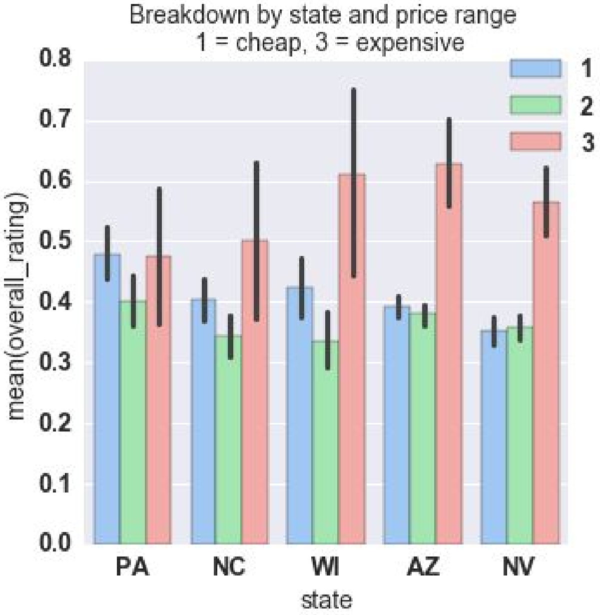


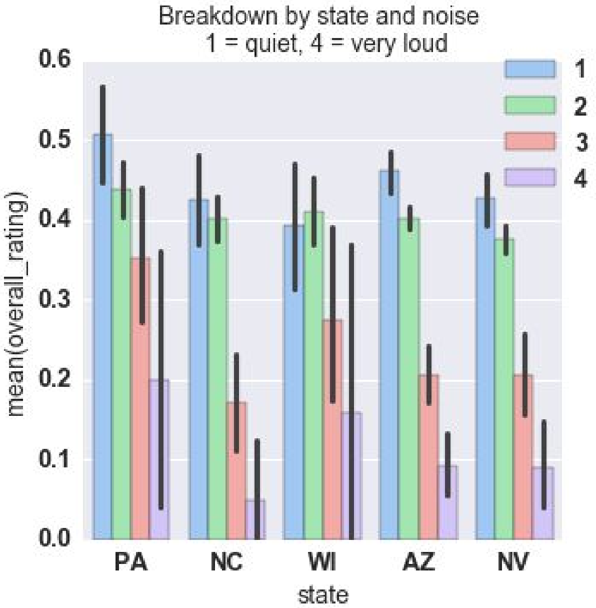




Exploring the data







Logistic Regression Results - Expanded Model

Logit Regression Results

Dep. Variable:	overall_rating	No. Observations:	14325
Model:	Logit	Df Residuals:	14309
Method:	MLE	Df Model:	15
Date:	Sat, 20 May 2017	Pseudo R-squ.:	0.06840
Time:	15:28:29	Log-Likelihood:	-8912.0
converged:	True	LL-Null:	-9566.4
		LLR p-value:	7.221e-270

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
log_reviews	0.4981	0.018	27.669	0.000	0.463	0.533
deliv_1.0	-0.0806	0.047	-1.700	0.089	-0.173	0.012
groups_1.0	-0.0097	0.065	-0.149	0.882	-0.137	0.118
waiter_1.0	-0.1075	0.046	-2.350	0.019	-0.197	-0.018
takeout_1.0	-0.1620	0.096	-1.685	0.092	-0.350	0.026
price_2	-0.4236	0.045	-9.323	0.000	-0.513	-0.335
price_3	-0.0924	0.126	-0.735	0.462	-0.339	0.154
noise_1	2.0456	0.180	11.345	0.000	1.692	2.399
noise_2	1.5246	0.178	8.581	0.000	1.176	1.873
noise_3	0.7927	0.193	4.112	0.000	0.415	1.170
attire_1	-0.2740	0.147	-1.864	0.062	-0.562	0.014
state_AZ	-0.1616	0.082	-1.983	0.047	-0.321	-0.002
state_NC	-0.0535	0.094	-0.572	0.567	-0.237	0.130
state_NV	-0.4448	0.085	-5.243	0.000	-0.611	-0.279
state_PA	0.2345	0.099	2.367	0.018	0.040	0.429
intercept	-3.0195	0.258	-11.688	0.000	-3.526	-2.513

Logistic Regression Results - Simplified Model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         overall_rating   No. Observations:                14325
Model:                          Logit   Df Residuals:                    14317
Method:                           MLE   Df Model:                            7
Date:                Sat, 20 May 2017   Pseudo R-squ.:                 0.06719
Time:                        15:29:28   Log-Likelihood:                -8923.6
converged:                       True   LL-Null:                       -9566.4
                                        LLR p-value:                2.204e-273
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
log_reviews    0.4922      0.017     28.887      0.000         0.459     0.526
price_2       -0.4741      0.038    -12.386      0.000        -0.549    -0.399
noise_1        2.0118      0.180     11.196      0.000         1.660     2.364
noise_2        1.5045      0.177      8.493      0.000         1.157     1.852
noise_3        0.7907      0.192      4.110      0.000         0.414     1.168
state_NV      -0.3059      0.041     -7.415      0.000        -0.387    -0.225
state_PA       0.3589      0.067      5.384      0.000         0.228     0.489
intercept     -3.5949      0.184    -19.487      0.000        -3.956    -3.233
==============================================================================
```

After gridsearch:
ROC_AUC = 0.67

Odds Ratio:

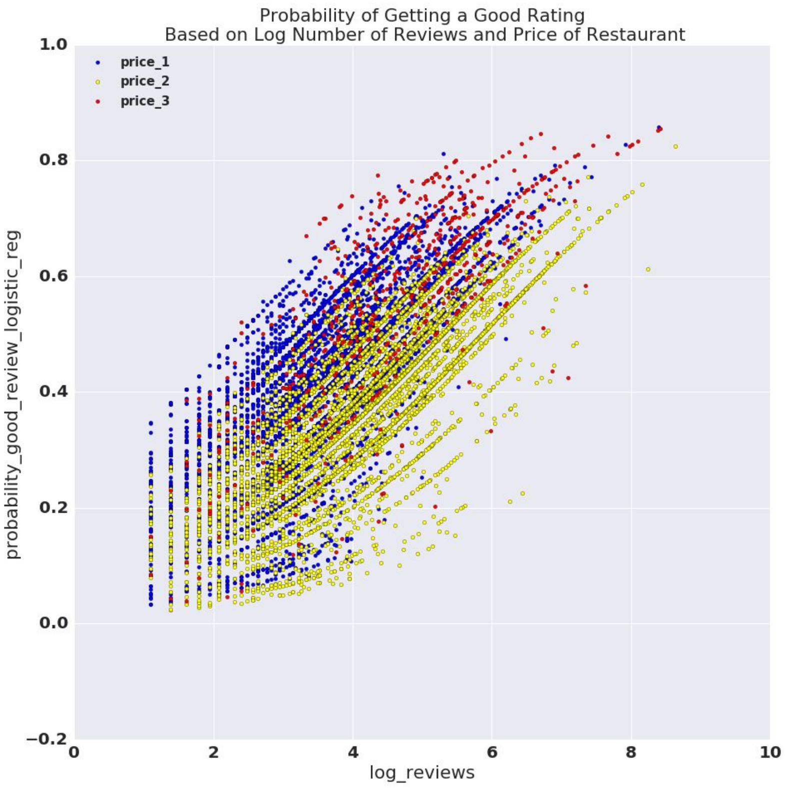
price_2	0.622435
noise_1	7.476790
noise_2	4.501933
noise_3	2.204840
state_NV	0.736481
state_PA	1.431698

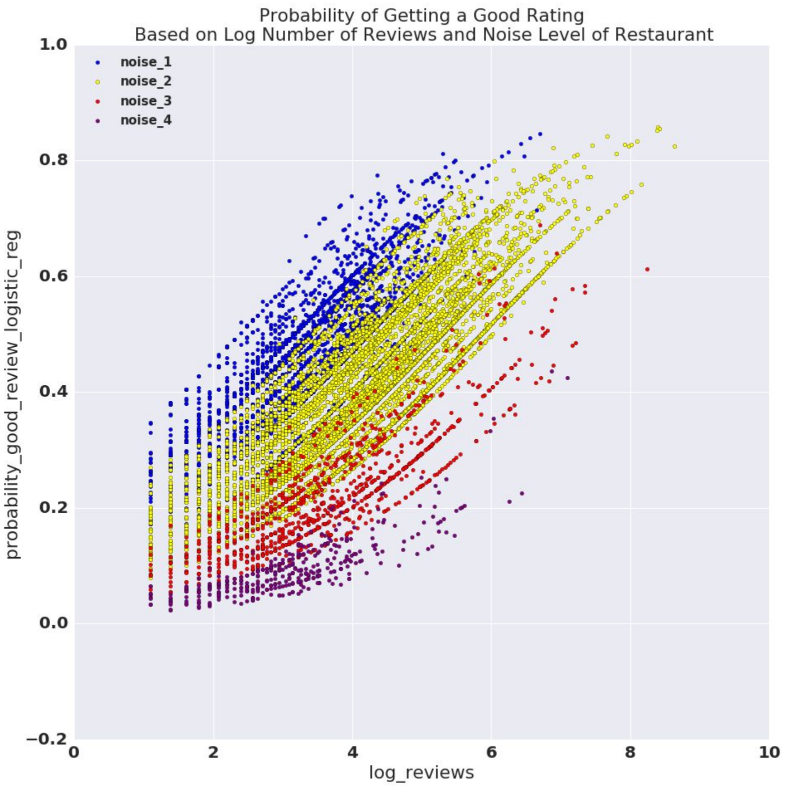
Random Forest Model

n trees: 1, CV AUC [0.56481514 0.56790506 0.53308935],
Average AUC 0.555269854439
n trees: 11, CV AUC [0.60284402 0.6044822   0.56650098],
Average AUC 0.591275731953
n trees: 21, CV AUC [0.60835283 0.60477434 0.56921241],
Average AUC 0.594113194136
n trees: 31, CV AUC [0.60694837 0.60245463 0.56526492],
Average AUC 0.591555975863
n trees: 41, CV AUC [0.60650619 0.60501957 0.56891859],
Average AUC 0.593481449823
n trees: 51, CV AUC [0.6077047   0.60309878 0.56968566],
Average AUC 0.593496376828
n trees: 61, CV AUC [0.60625067 0.60361555 0.56885687],
Average AUC 0.592907696157
n trees: 71, CV AUC [0.60848918 0.60535004 0.56916224],
Average AUC 0.594333820187
n trees: 81, CV AUC [0.60654662 0.60545404 0.56815328],
Average AUC 0.593384650726
n trees: 91, CV AUC [0.6085261   0.60698959 0.56774952],
Average AUC 0.594421736393

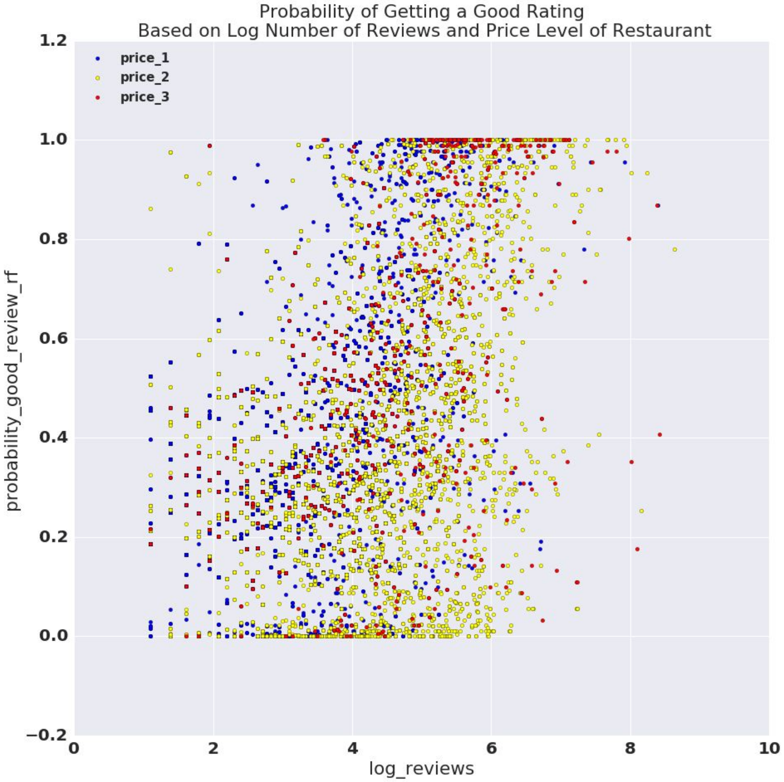
	Features	Importance Score
0	log_reviews	0.790777
5	price_2	0.025441
1	deliv_1.0	0.023570
3	waiter_1.0	0.021824
2	groups_1.0	0.019122
7	noise_1	0.015856
8	noise_2	0.014770
13	state_NV	0.014271
4	takeout_1.0	0.013811
11	state_AZ	0.013804
12	state_NC	0.011004
9	noise_3	0.010232
14	state_PA	0.009770
6	price_3	0.008230
10	attire_1	0.007518
15	intercept	0.000000

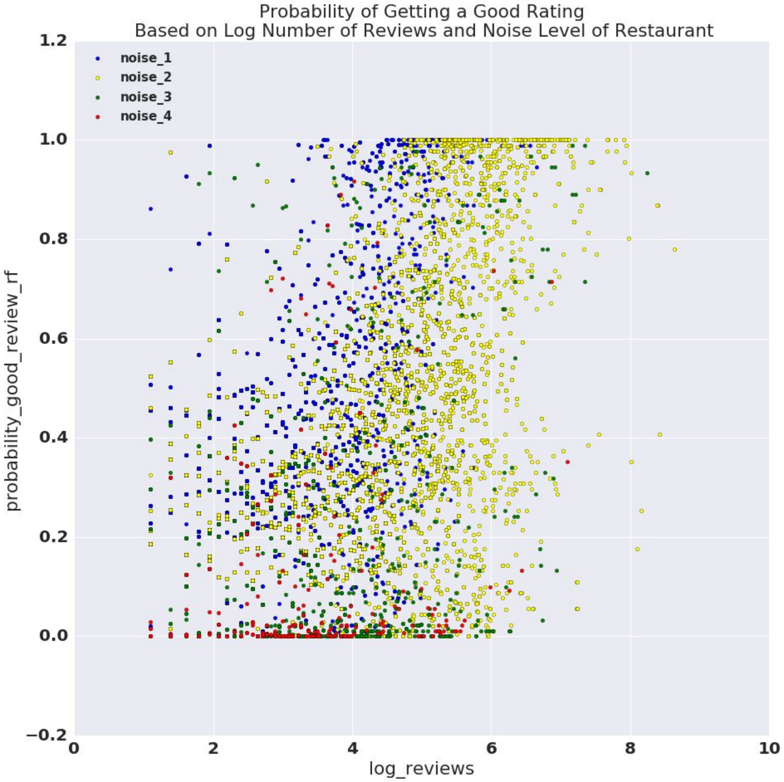
Visualizing Results - Logistic Regression



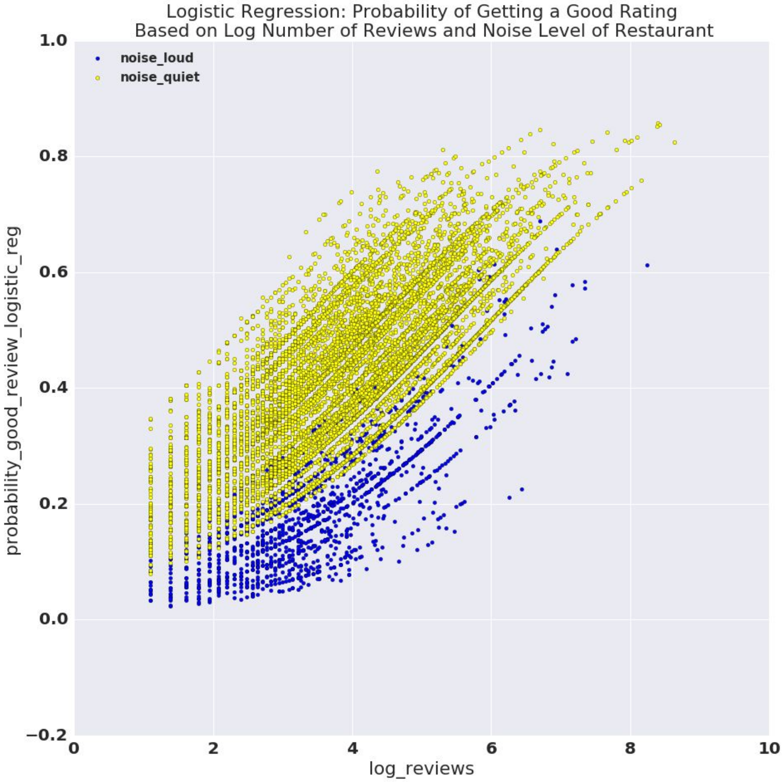


Visualizing Results - Random Forest





Visualizing Results - Comparison



Bonus: Can we predict Reviews Using Text Comments?

- Downloaded yelp user reviews json file
- Had 4,153,151 text reviews; randomly sampled 10,000 of them
- ALso created an “overall rating category”
- Ran CountVectorizer and TfidfVectorizer, then fit and transformed the data, and then used a Random Forest Classifier
  - Count Vectorizor: converts collection of text into matrix of features
  - Term Frequency and Inverse Document Frequency: provides count of features as well as uniqueness of features
- Count Vectorizer Average AUC: 0.81
- Term Frequency - Inverse Document Frequency Average AUC: 0.82

Future work...

- Can use multinomial logistic regression model to run it use all the different rating categories instead of the two that I used
- Can combine the text reviews and business features into one model to see if predictive power improves
- Can run using different kind of cuisines as features
- Can also incorporate user profiles to see different types of users rate the same restaurant in different ways
- Partial dependence plots to more clearly see results from random forest models

Thank you!