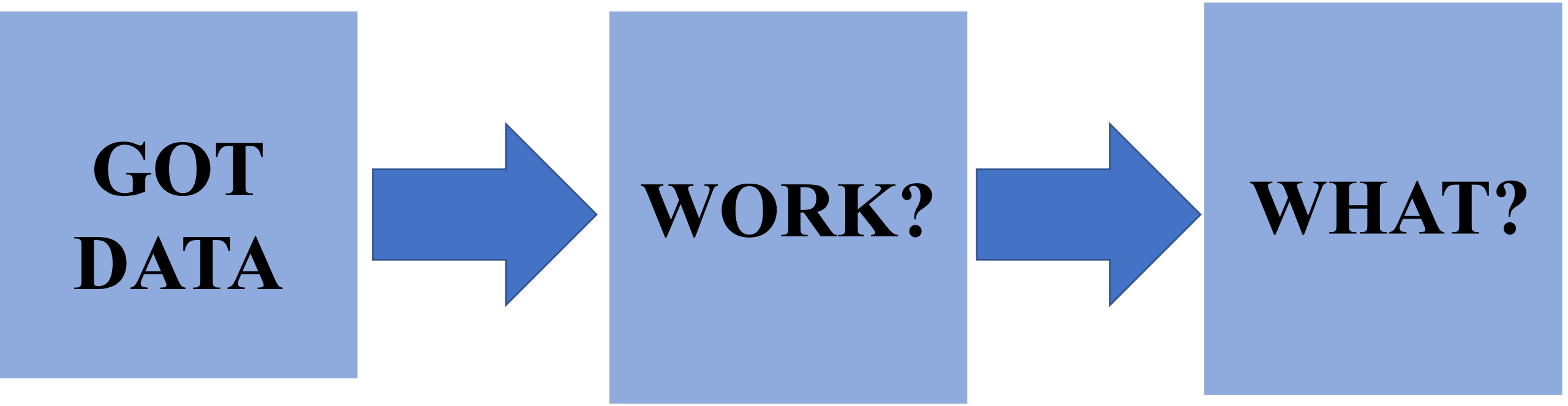
NYC Datascience PROJECT 3
Kaggle Zillow
$how us the money

Markson

Shu

In Suk

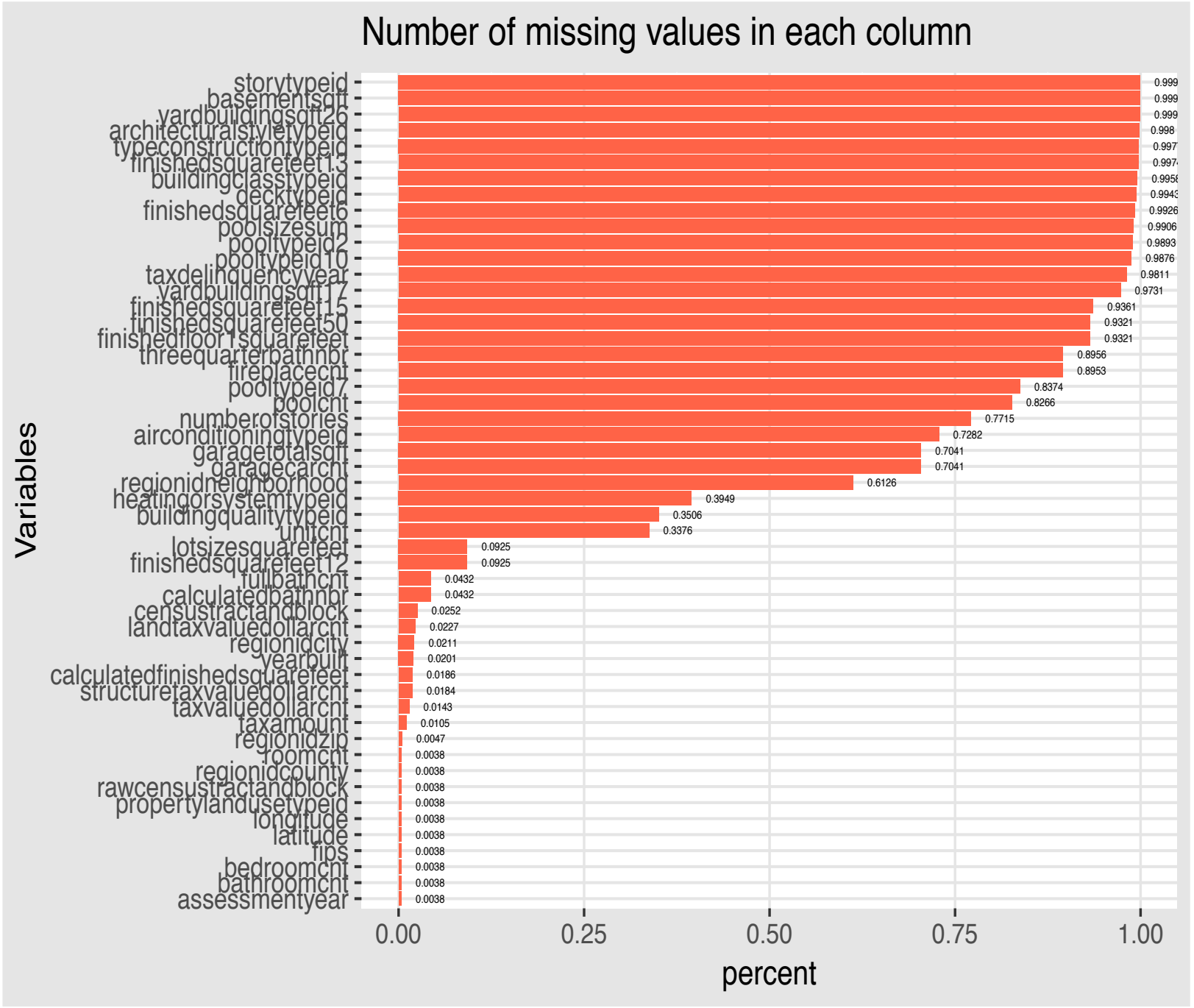

GOT DATA
WORK?
WHAT?

GOT DATA?

We have seen many missing values in the properties data set..

How many missing values are there for each feature? In fact, some features are missing nearly completely.

There are 29 out of 58 features having missing values over 70 percent of each column.



How to Handle the Missing Data?

The more missingness a feature has, the less important it will be.

We removed the less important features.

Set a threshold?

29 columns were removed.

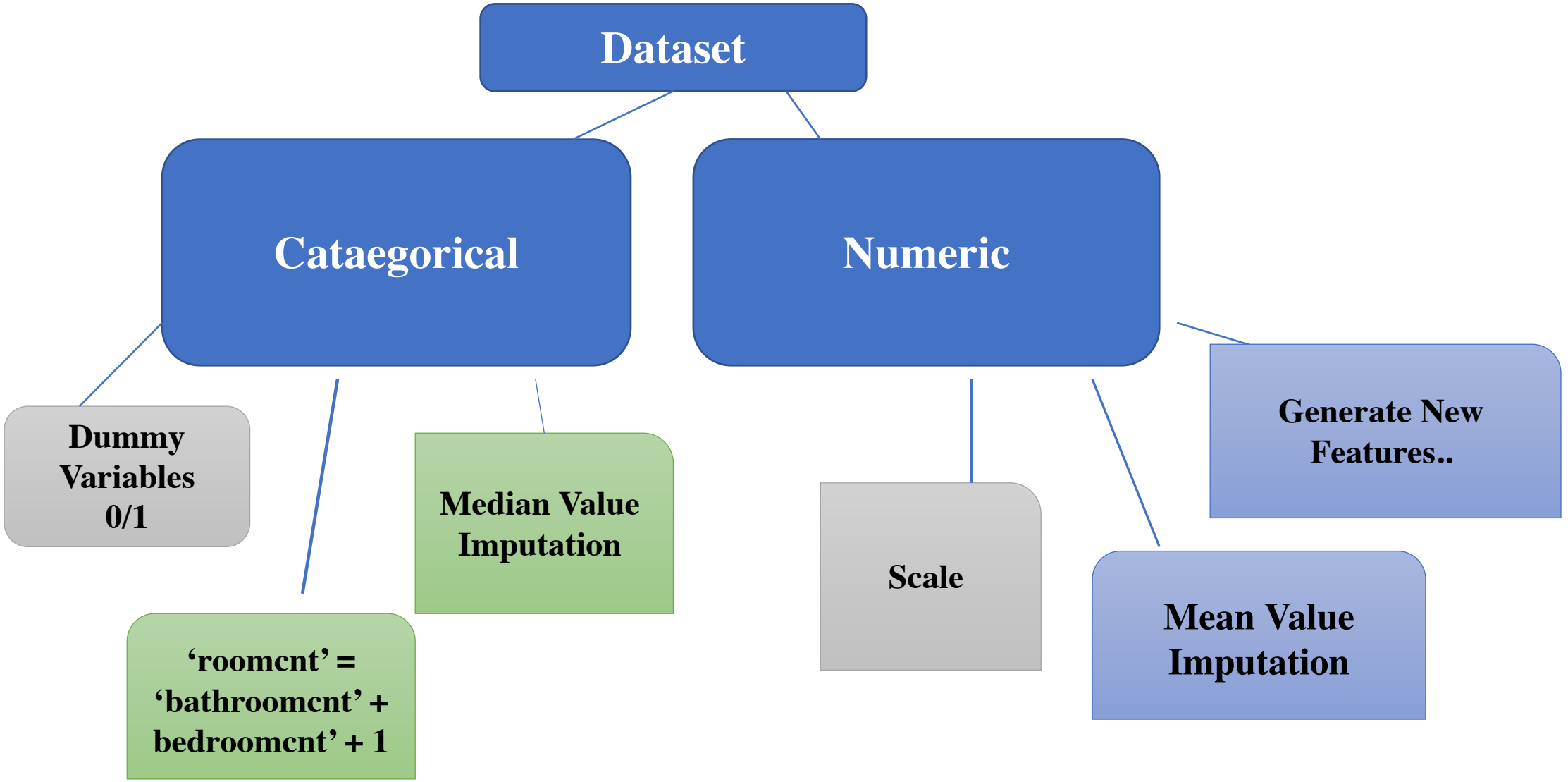
We also removed features that have duplicated meanings.

30%

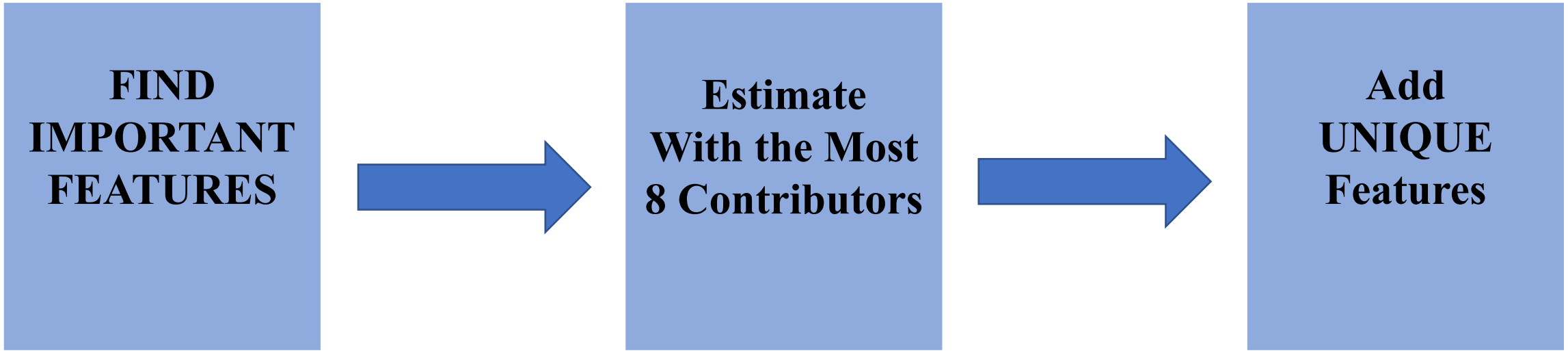
Missing Values by Column

Variables	Total	PercentageOfMissingness
basementsqft	2983589	0.9995
storytypeid	2983593	0.9995
yardbuildingsqft26	2982570	0.9991
architecturalstyletypeid	2979156	0.998
typeconstructiontypeid	2978470	0.9977
finishedsquarefeet13	2977545	0.9974
buildingclasstypeid	2972588	0.9958
decktypeid	2968121	0.9943
finishedsquarefeet6	2963216	0.9926
poolsizesum	2957257	0.9906
pooltypeid2	2953142	0.9893
pooltypeid10	2948278	0.9876
taxdelinquencyyear	2928753	0.9811
yardbuildingsqft17	2904862	0.9731
finishedsquarefeet15	2794419	0.9361
finishedfloor1squarefeet	2782500	0.9321
finishedsquarefeet50	2782500	0.9321
threequarterbathnbr	2673586	0.8956
fireplacecnt	2672580	0.8953
pooltypeid7	2499758	0.8374
poolcnt	2467683	0.8266
numberofstories	2303148	0.7715
airconditioningtypeid	2173698	0.7282
garagecarcnt	2101950	0.7041
garagetotalsqft	2101950	0.7041
regionidneighborhood	1828815	0.6126
heatingorsystemtypeid	1178816	0.3949
buildingqualitytypeid	1046729	0.3506
unitcnt	1007727	0.3376
finishedsquarefeet12	276033	0.0925
lotsizesquarefeet	276099	0.0925
calculatedbathnbr	128912	0.0432
fullbathcnt	128912	0.0432
censustractandblock	75126	0.0252
landtaxvaluedollarcnt	67733	0.0227
regionidcity	62845	0.0211
yearbuilt	59928	0.0201
calculatedfinishedsquarefeet	55565	0.0186
structuretaxvaluedollarcnt	54982	0.0184

Featuring Engineering

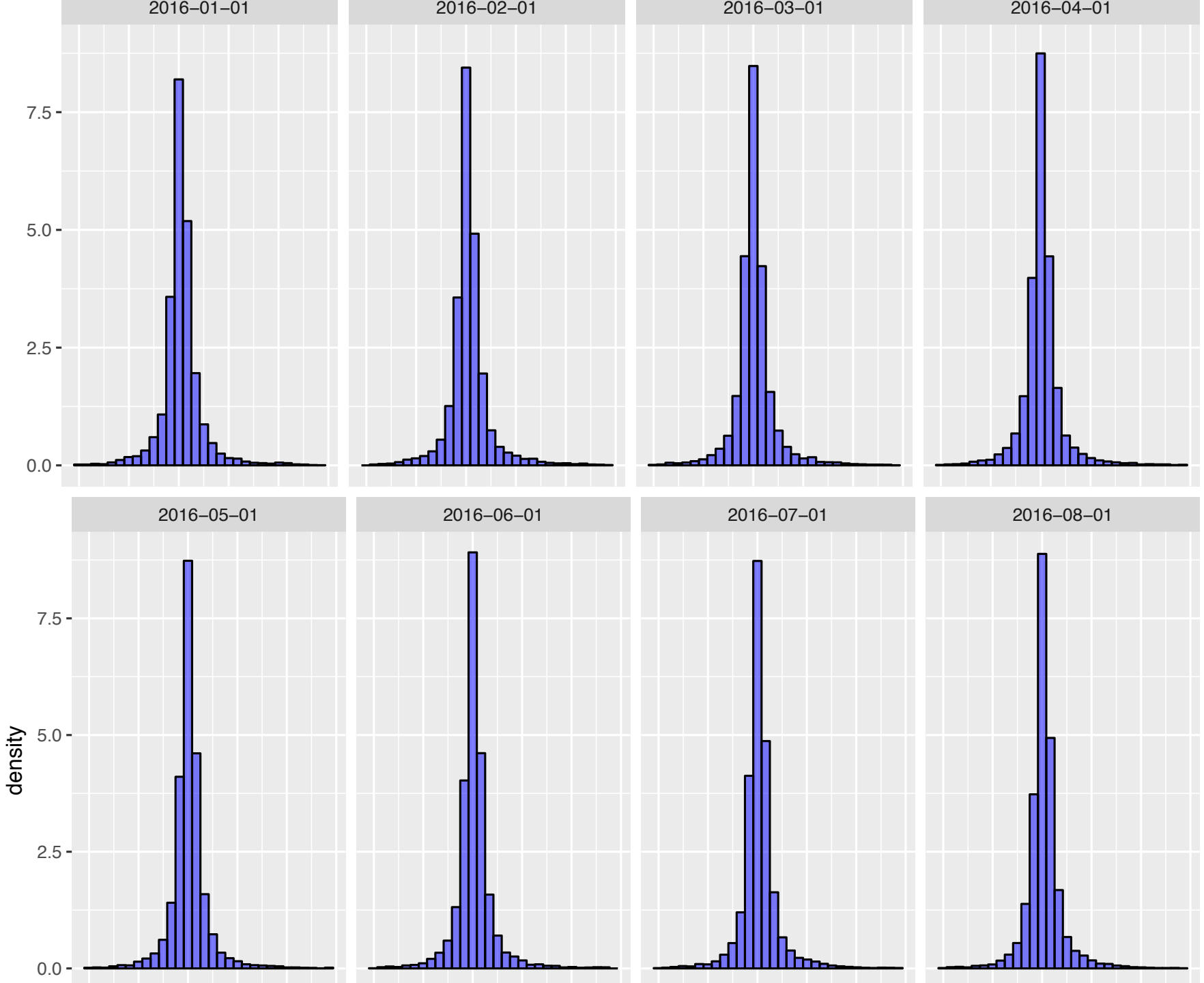


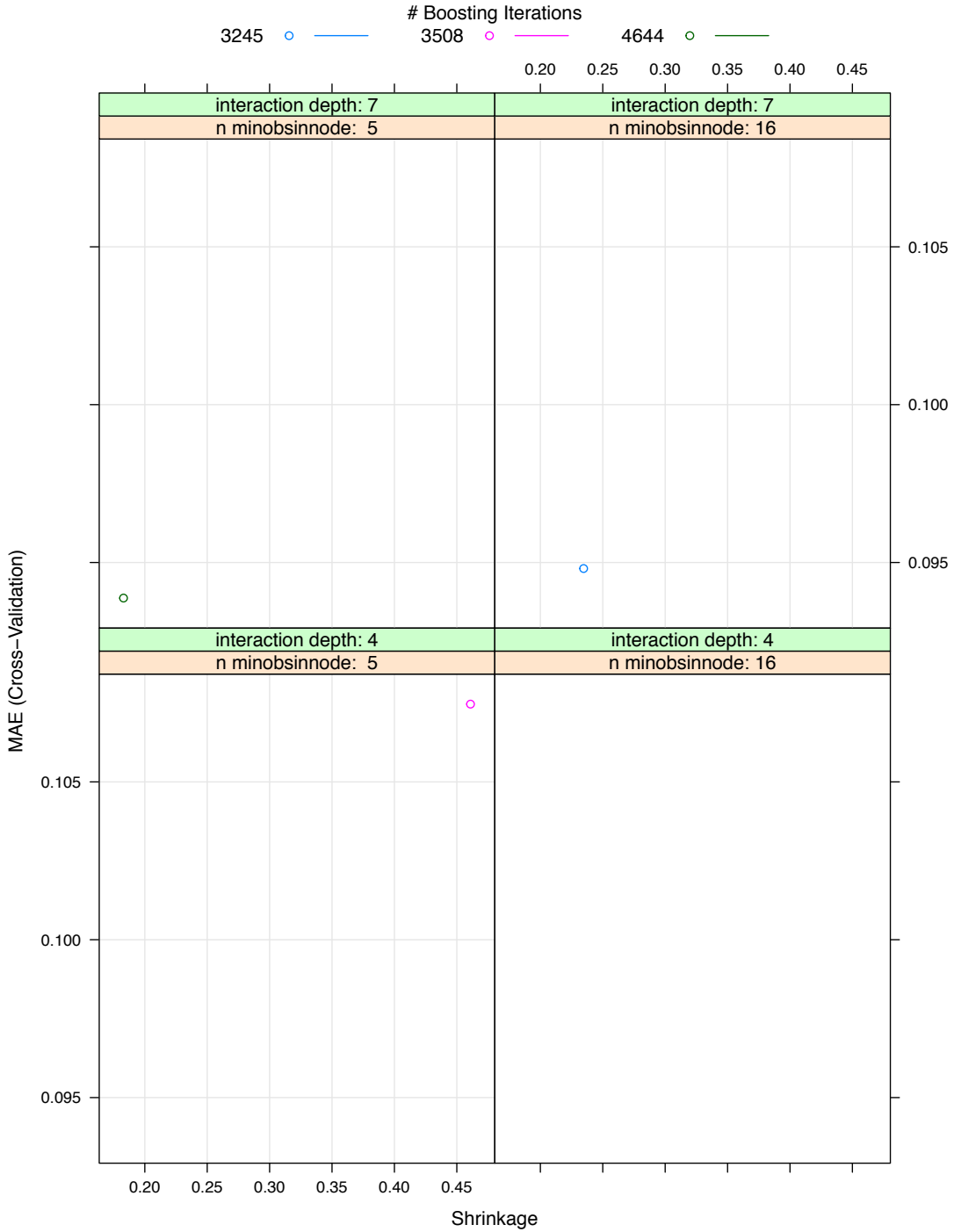
WORK?

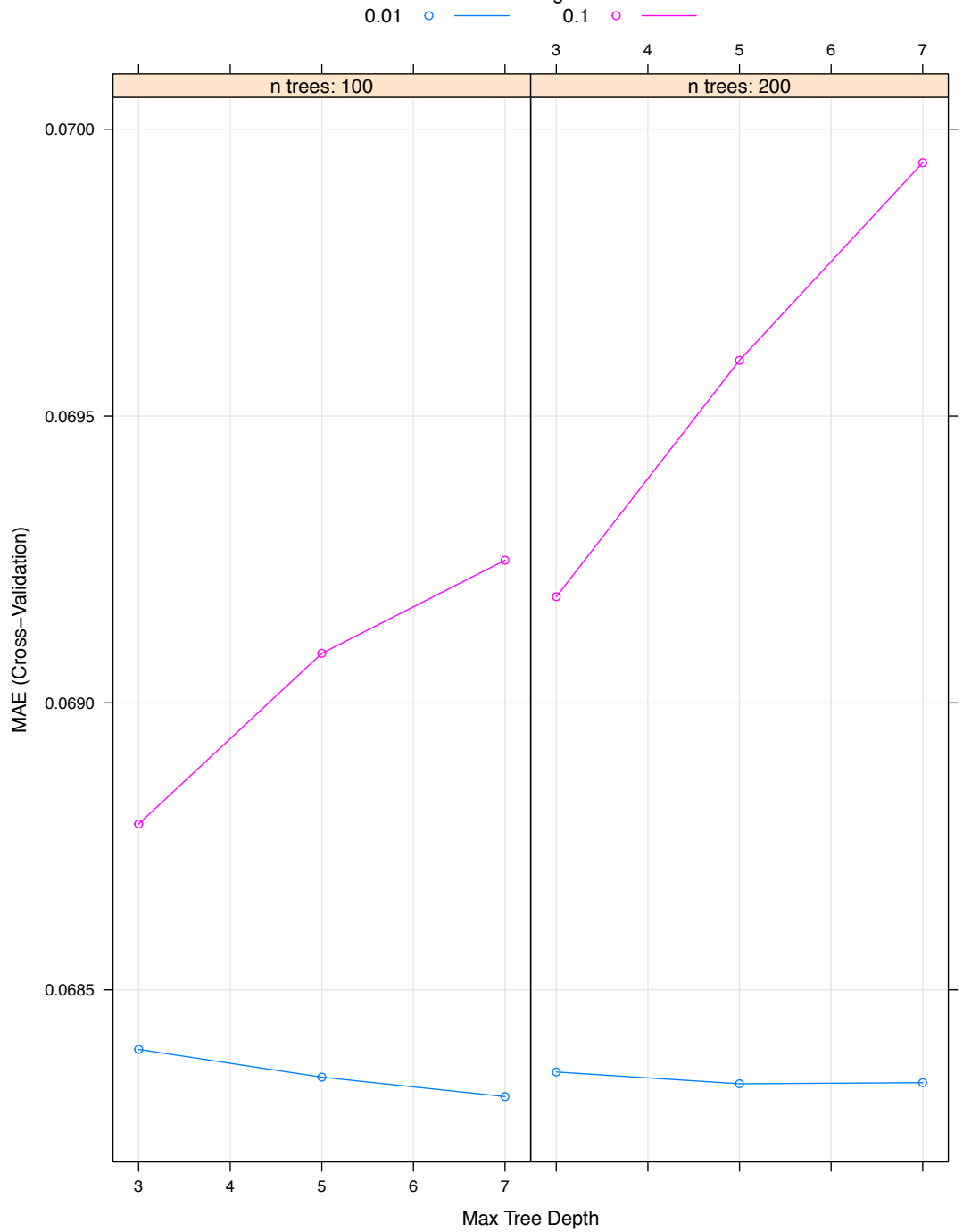


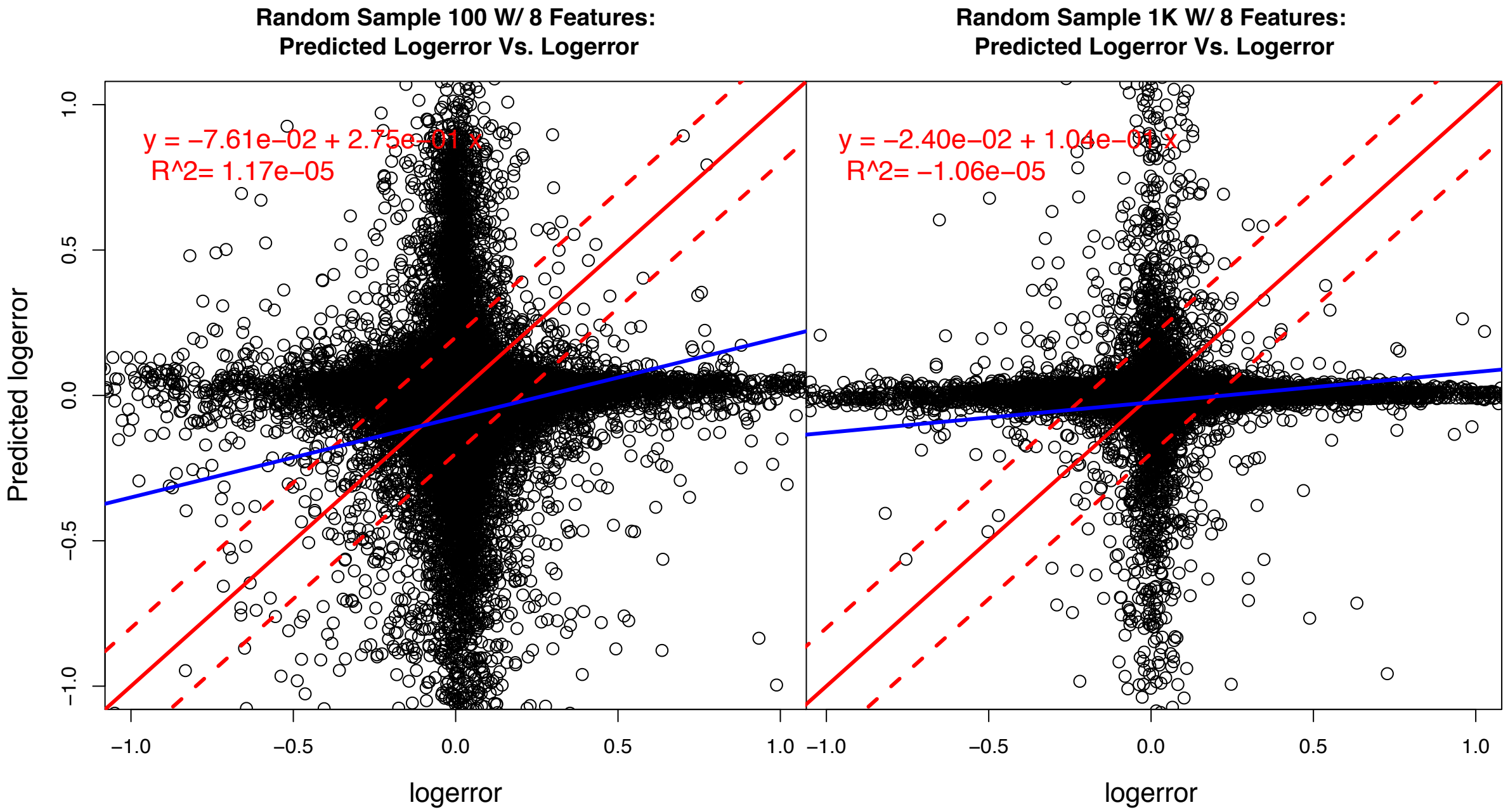
RANDOM FOREST
XGBOOST

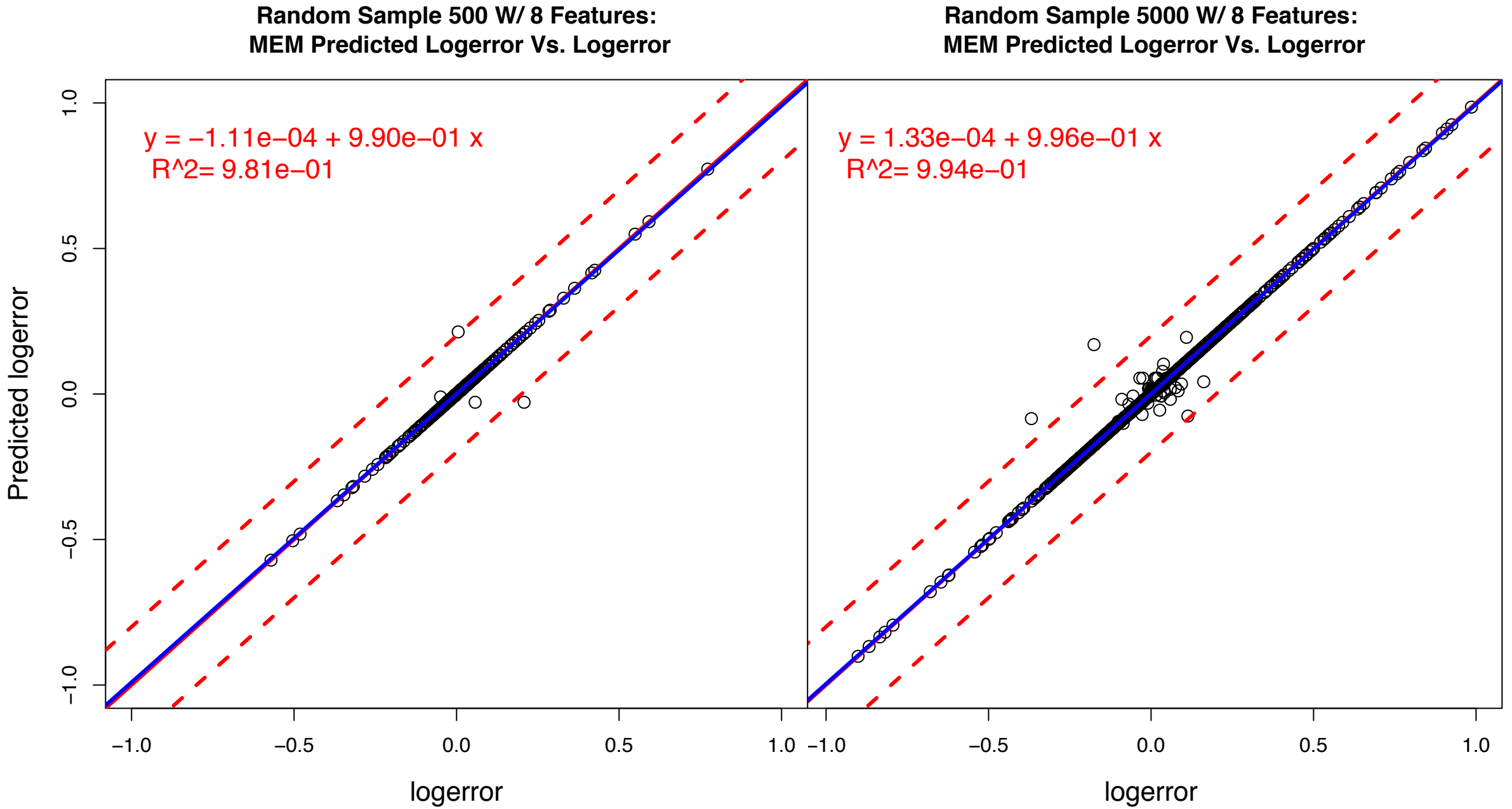
LM &
Linear Mixed Effective Model

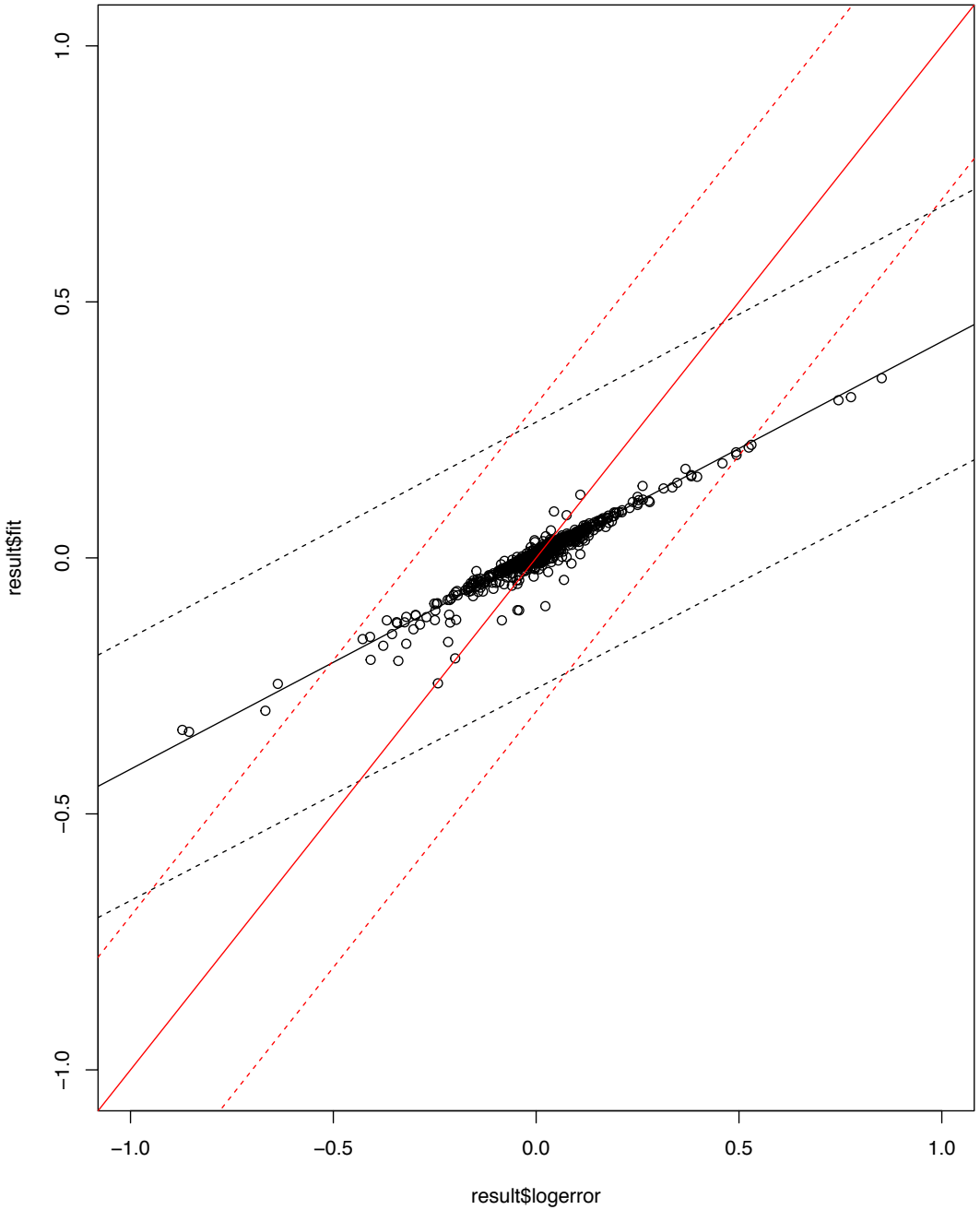

2016–01–01
2016–02–01
2016–03–01
2016–04–01
2016–05–01
2016–06–01
2016–07–01
2016–08–01
density
7.5
5.0
2.5
0.0

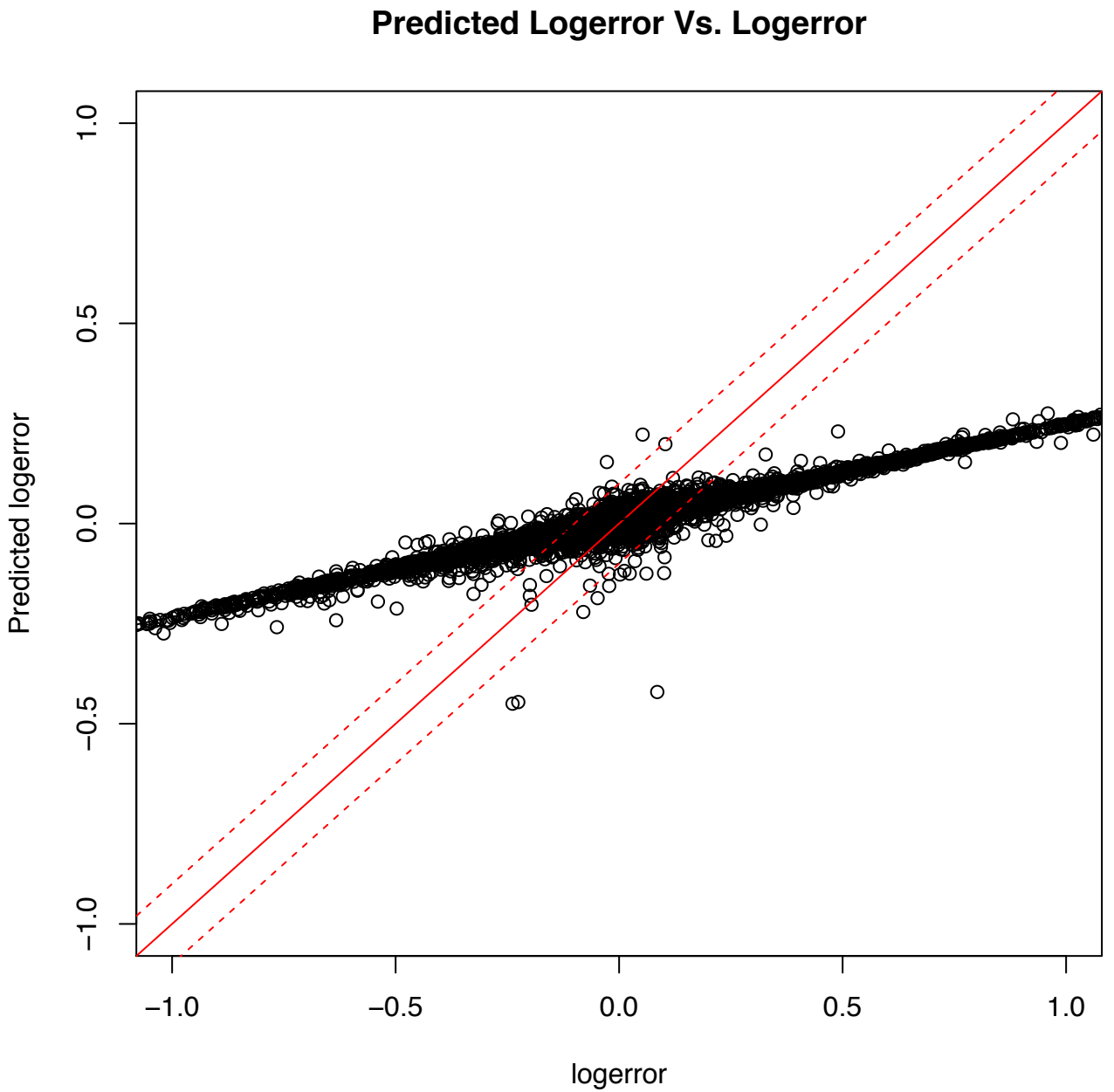

Boosting Iterations
3245
3508
4644
interaction depth: 7
n minobsinnode: 5
interaction depth: 7
n minobsinnode: 16
interaction depth: 4
n minobsinnode: 5
interaction depth: 4
n minobsinnode: 16
MAE (Cross−Validation)
Shrinkage
0.20
0.25
0.30
0.35
0.40
0.45
0.095
0.100
0.105


0.01
0.1
n trees: 100
n trees: 200
MAE (Cross−Validation)
Max Tree Depth
3
4
5
6
7
0.0685
0.0690
0.0695
0.0700


Random Sample 100 W/ 8 Features:
Predicted Logerror Vs. Logerror
y = −7.61e−02 + 2.75e−01 x
R^2= 1.17e−05
Random Sample 1K W/ 8 Features:
Predicted Logerror Vs. Logerror
y = −2.40e−02 + 1.04e−01 x
R^2= −1.06e−05
Predicted logerror
logerror
logerror
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0


Random Sample 500 W/ 8 Features:
MEM Predicted Logerror Vs. Logerror
y = −1.11e−04 + 9.90e−01 x
R^2= 9.81e−01
Random Sample 5000 W/ 8 Features:
MEM Predicted Logerror Vs. Logerror
y = 1.33e−04 + 9.96e−01 x
R^2= 9.94e−01
Predicted logerror
logerror
logerror
−1.0
−0.5
0.0
0.5
1.0


result$fit
result$logerror
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0


Predicted Logerror Vs. Logerror
Predicted logerror
logerror
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

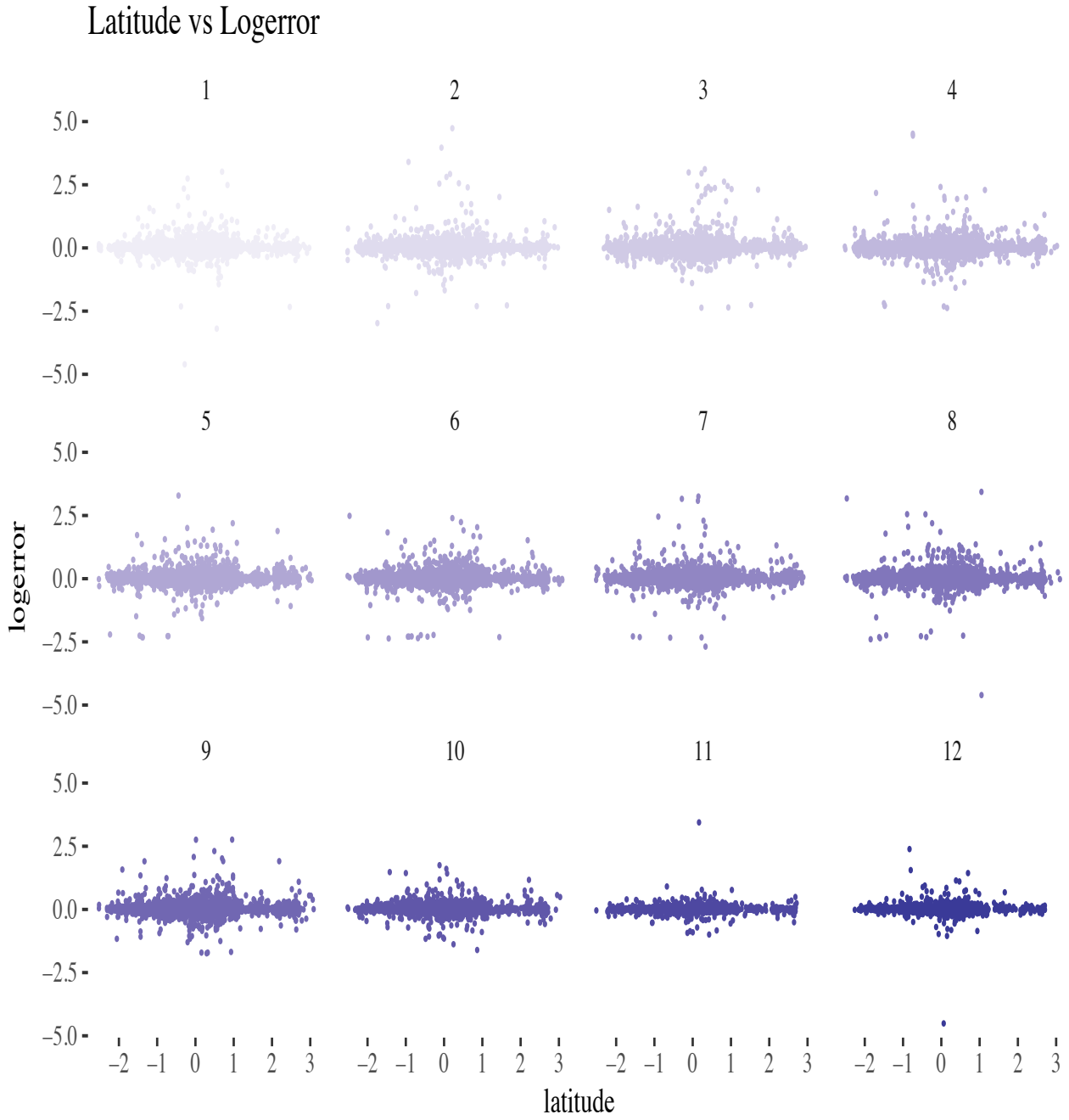
Random Forest

Region features matter.

Mean Absolute Error =

0.07071

	feature	importance
1	structuretaxvaluedollarcnt	0.085605
4	taxamount	0.083288
5	calculatedfinishedsquarefeet	0.082269
2	taxvaluedollarcnt	0.081811
3	landtaxvaluedollarcnt	0.080706
8	latitude	0.078395
0	lotsizesquarefeet	0.077873
9	longitude	0.077280
15	yearbuilt	0.063726
13	regionidzip	0.046730
17	rawcensustractandblock	0.045288
19	censustractandblock	0.045158
6	month	0.043462
11	regionidcity	0.029804
20	bedroomcnt	0.023638
18	bathroomcnt	0.022217
10	propertylandusetypeid	0.013478



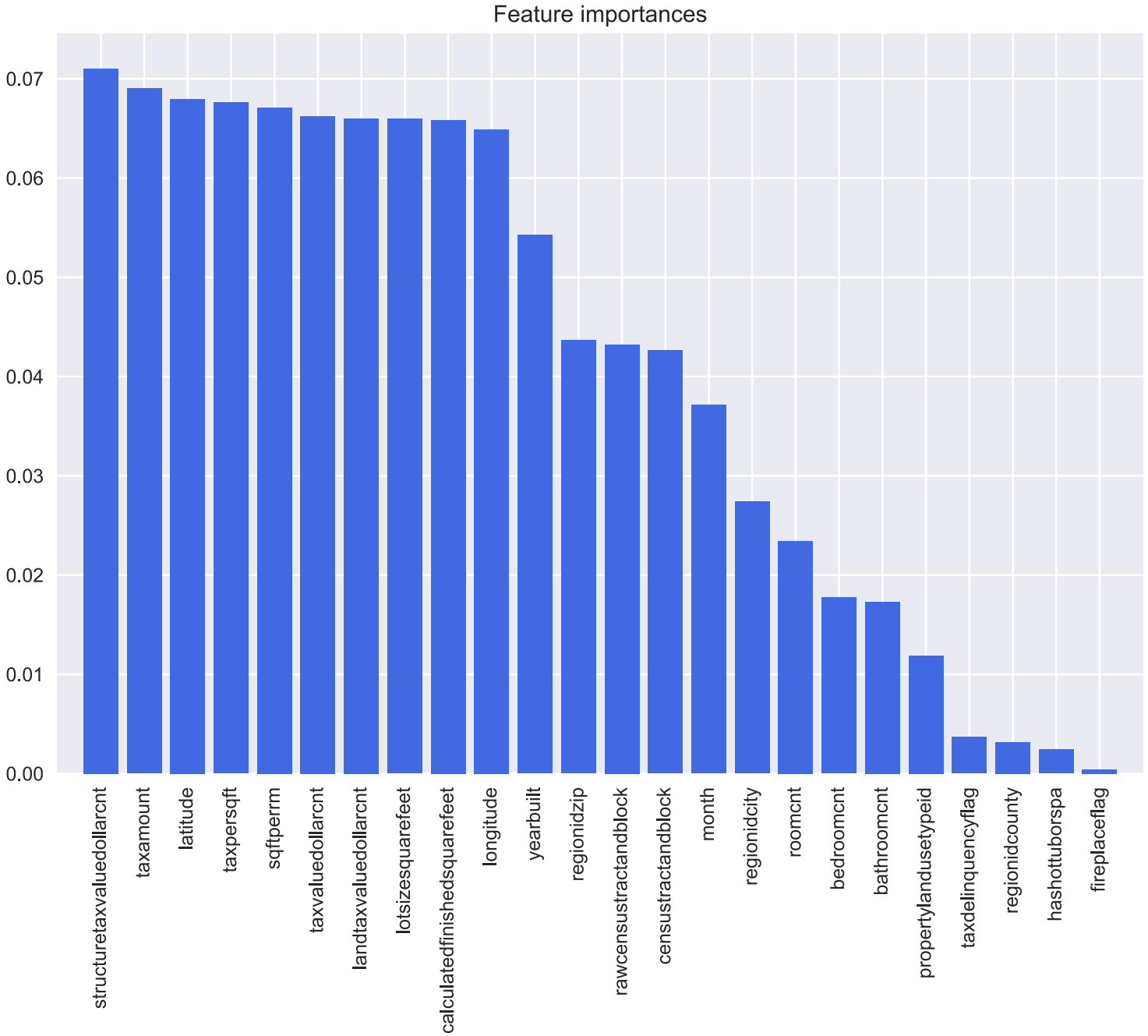
How about add more features?

We added some numeric features that might be linear related with logerror and also sensitive to regions.

‘Average Square Feet per Room’ = ‘calculatedfinishedsquarefeet’ / ‘roomcnt’

‘Tax per Square Feet’ = ‘taxvaluedollarcnt’ / ‘calculatedfinishedsquarefeet’

‘

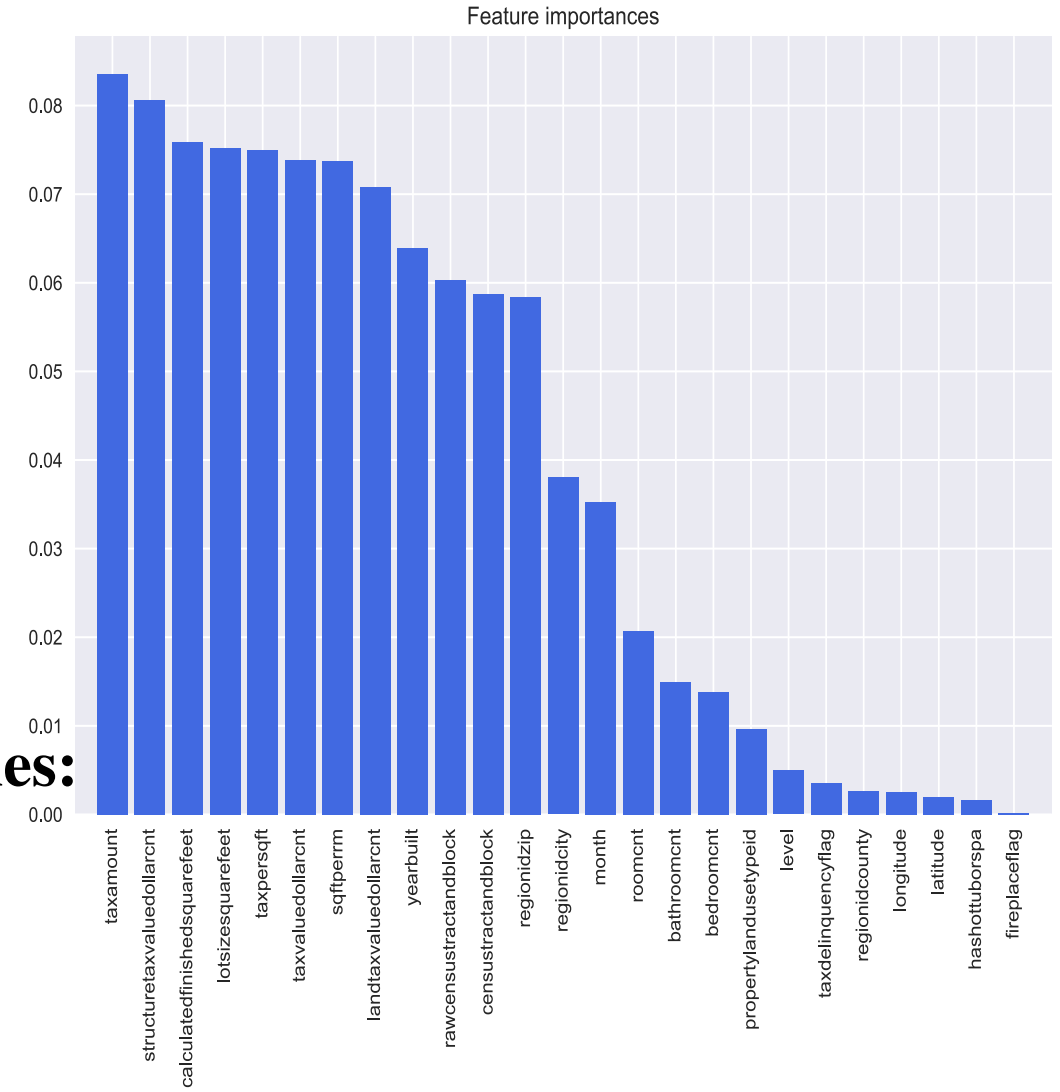


Mean Absolute Error =

0.06838

	feature	importance
2	structuretaxvaluedollarcnt	0.071002
5	taxamount	0.069018
9	latitude	0.067988
23	taxpersqft	0.067604
22	sqftperrm	0.067081
3	taxvaluedollarcnt	0.066242
4	landtaxvaluedollarcnt	0.066013
1	lotsizesquarefeet	0.065959
0	calculatedfinishedsquarefeet	0.065840
10	longitude	0.064885
15	yearbuilt	0.054256
14	regionidzip	0.043652

Kmeans Clustering on Regions



	feature	importance
5	taxamount	0.083614
2	structuretaxvaluedollarcnt	0.080609
0	calculatedfinishedsquarefeet	0.075894
1	lotsizesquarefeet	0.075170
23	taxpersqft	0.075031
3	taxvaluedollarcnt	0.073908
22	sqftperrm	0.073741
4	landtaxvaluedollarcnt	0.070805
15	yearbuilt	0.063928
18	rawcensustractandblock	0.060269
21	censustractandblock	0.058722
14	regionidzip	0.058399
12	regionidcity	0.038025
6	month	0.035238
20	roomcnt	0.020680
19	bathroomcnt	0.014978
7	bedroomcnt	0.013817
11	propertylandusetypeid	0.009644
24	level	0.004962
17	taxdelinquencyflag	0.003561
13	regionidcounty	0.002687
10	longitude	0.002540
9	latitude	0.001938
8	hashottuborspa	0.001659
16	fireplaceflag	0.000181

4 Region related categories:

- House Conditions
- Locations
- Size
- Tax-related

Mean Absolute Error = 0.06846

Results

In [272]:
```
from sklearn.metrics import mean_absolute_error
mean_absolute_error(y_test,rf_cv.best_estimator_.predict(X_test))
```

```
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.1s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:    0.6s
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed:    1.3s
[Parallel(n_jobs=4)]: Done 792 tasks      | elapsed:    2.3s
[Parallel(n_jobs=4)]: Done 1200 out of 1200 | elapsed:    3.3s finished
```

Out[272]: 0.06838219940869472 2

In [373]:
```
mean_absolute_error(y_test,rf_cv.best_estimator_.predict(X_test))
```

```
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.1s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:    0.5s
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed:    1.2s
[Parallel(n_jobs=4)]: Done 792 tasks      | elapsed:    2.0s
[Parallel(n_jobs=4)]: Done 800 out of 800 | elapsed:    2.0s finished
```

Out[373]: 0.0684677626995386 78

Future Work?

Concentrate of EDA

Mix Effective Model ….

Start Early!!!