~~$350,000~~

$300,000

PREDICTING LOG ERROR IN ZILLOW REAL ESTATE PRICING PREDICTIONS

-PETER GIERKE

Background on Zillow

- *Zillow has data on 110 million homes across the United States, not just those homes currently for sale.*
- *Zillow determines an estimate ("Zestimate," pronounced "ZEST-imate") for a home price based on a range of publicly available information, including sales of comparable houses in a neighborhood.*
- *In 2007, The Wall Street Journal studied the accuracy of Zillow's estimates and found that they "often are very good, frequently within a few percentage points of the actual price paid. But when Zillow is bad, it can be terrible.*
- *Using data published on the Zillow website, the typical Zestimate error in the United States in July 2016 was $14,000.*
- *From launch until today Zillow has improved Zestimate median error from "14% at the onset to 5% today"*

https://en.wikipedia.org/wiki/Zillow

Background on competition

- *Zillow is looking to improve the accuracy of Zestimates*
- *Target for prediction is the log error as described:*
  - *logerror=log(Zestimate)−log(SalePrice)*
- *Goal is to predict logerror of 6 individual timepoints:*
  - *October, November and December of 2016 and October, November and December of 2017*
- *Submission Accuracy is evaluated on Mean Absolute Error*

Dataset -Data

- Data
  - 2,985,217 samples
  - 58 columns of data.
    - 1 of type int
    - 5 of type object
    - 52 of type float
  - Large portion of the columns missing more than 80% of the data.

parcelid	0.000000
logerror	0.000000
transactiondate	0.000000
airconditioningtypeid	68.306703
architecturalstyletypeid	99.712590
basementsqft	99.952649
bathroomcnt	0.590237
bedroomcnt	0.590237
buildingclasstypeid	99.982381
buildingqualitytypeid	36.831441
calculatedbathnbr	1.891841
decktypeid	99.275418
finishedfloor1squarefeet	92.450254
calculatedfinishedsquarefeet	1.318122
finishedsquarefeet12	5.742696
finishedsquarefeet13	99.963661
finishedsquarefeet15	96.075365
finishedsquarefeet50	92.450254
finishedsquarefeet6	99.536400
fips	0.590237
fireplacecnt	89.420885
fullbathcnt	1.891841
garagecarcnt	67.033729
garagetotalsqft	67.033729
hashottuborspa	97.395690
heatingorsystemtypeid	38.245367
latitude	0.590237
longitude	0.590237
lotsizesquarefeet	11.767297
poolcnt	80.287630
poolsizesum	98.932949
pooltypeid10	98.721521
pooltypeid2	98.674169
pooltypeid7	81.613461

propertycountylandusecode	0.591338
propertylandusetypeid	0.590237
propertyzoningdesc	35.786414
rawcensustractandblock	0.590237
regionidcity	2.575679
regionidcounty	0.590237
regionidneighborhood	60.344011
regionidzip	0.628778
roomcnt	0.590237
storytypeid	99.952649
threequarterbathnbr	86.775831
typeconstructiontypeid	99.670745
unitcnt	35.742366
yardbuildingsqft17	97.086256
yardbuildingsqft26	99.895387
yearbuilt	1.422735
numberofstories	77.348559
fireplaceflag	99.755536
structuretaxvaluedollarcnt	1.008688
taxvaluedollarcnt	0.591338
assessmentyear	0.590237
landtaxvaluedollarcnt	0.591338
taxamount	0.596844
taxdelinquencyflag	98.036581
taxdelinquencyyear	98.036581
censustractandblock	1.256456

Data -Target

- *3 columns*
  - *1 nominal (parcel ID)*
  - *1 continuous (logerror) *competition target*
  - *1 interval (transaction date)*
- *90811 samples*

	parcelid	logerror	transactiondate
0	11016594	0.0276	2016-01-01
1	14366692	-0.1684	2016-01-01
2	12098116	-0.0040	2016-01-01
3	12643413	0.0218	2016-01-02
4	14432541	-0.0050	2016-01-02

Data Cleaning and Feature Engineering

- *Merge data onto target dataset to start working with features for available target data.*
- *Drop data where more than 20% of the column is NaN. Leaves us with 26 columns.*

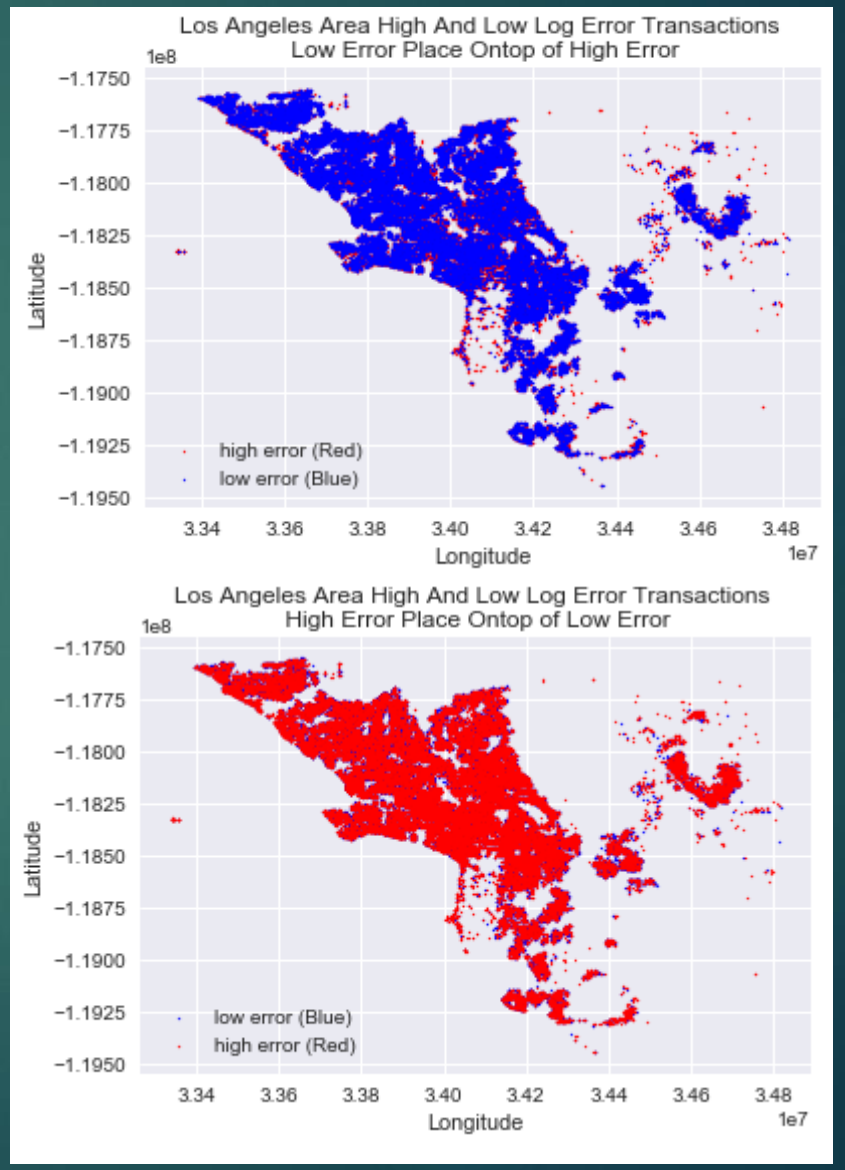
parcelid	0.000000
logerror	0.000000
transactiondate	0.000000
bathroomcnt	0.590237
bedroomcnt	0.590237
calculatedbathnbr	1.891841
calculatedfinishedsquarefeet	1.318122
finishedsquarefeet12	5.742696
fips	0.590237
fullbathcnt	1.891841
latitude	0.590237
longitude	0.590237
lotsizesquarefeet	11.767297
propertycountylandusecode	0.591338
propertylandusetypeid	0.590237
rawcensustractandblock	0.590237
regionidcity	2.575679
regionidcounty	0.590237
regionidzip	0.628778
roomcnt	0.590237
yearbuilt	1.422735
structuretaxvaluedollarcnt	1.008688
taxvaluedollarcnt	0.591338
assessmentyear	0.590237
landtaxvaluedollarcnt	0.591338
taxamount	0.596844
censustractandblock	1.256456

Data Cleaning and Feature Engineering

- *7 of the remaining 26 columns are categorical and can be excluded from modeling*
  - *fips*
  - *Propertycountylanduse*
  - *Regioncityid*
  - *Regioncityzip*
  - *Regionidcounty*
  - *Censustractandblock*
  - *rawcensustractandblock*

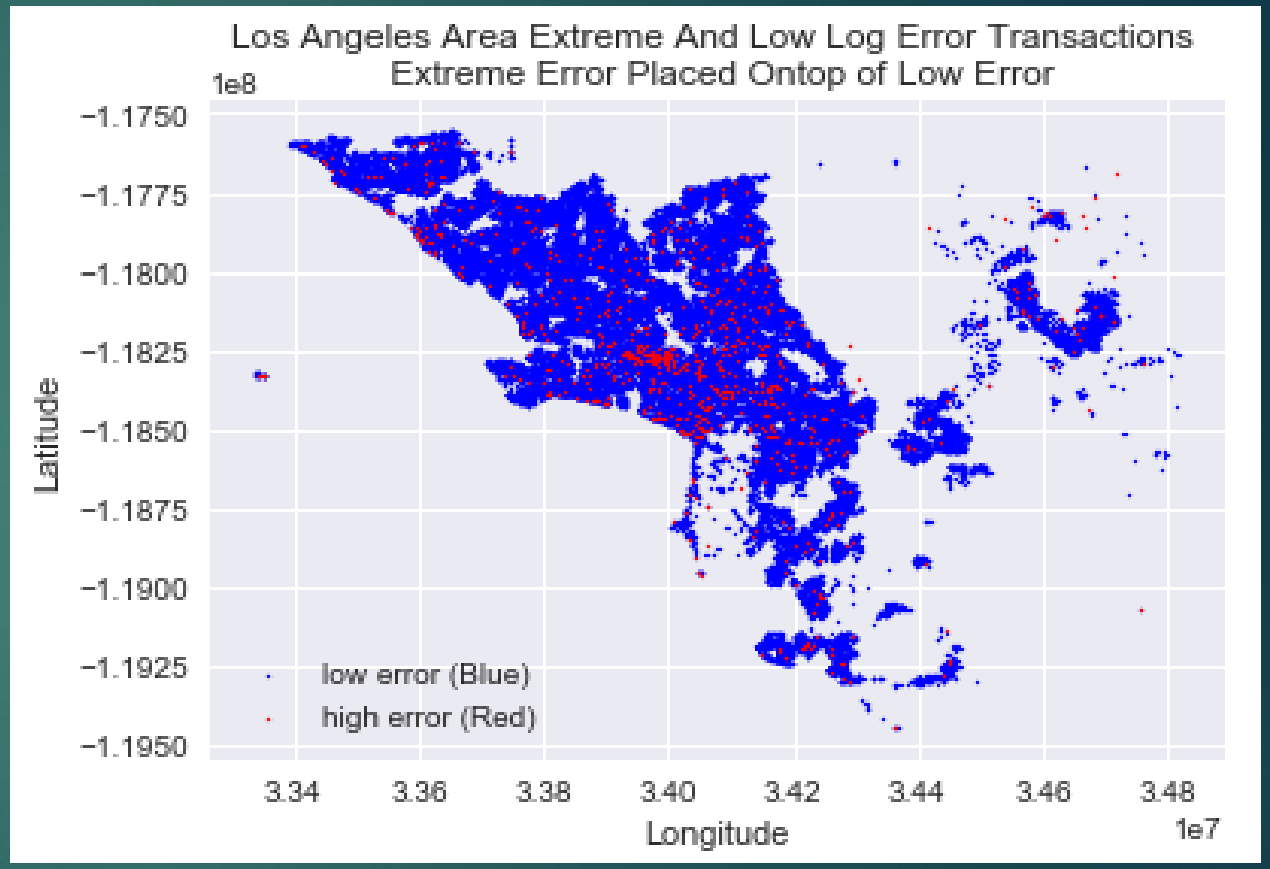
Data Cleaning and Feature Engineering

- *2 columns are continuous and refer to location*
  - *Latitude*
  - *Longitude*
- *Exploration of data shows no clear patterns based on geography*
  - *Low Error defined as between -0.0253 and 0.0392.*
  - *High error defined as the remaining data.*
  - *No clear geographical pattern for high vs low error predictions.*
- *Usefulness of data questionable*



Data Cleaning and Feature Engineering

- *Further adjusting definition of high error to less than -0.6300 or greater than 0.6534 show that there is some clustering of very high log error.*



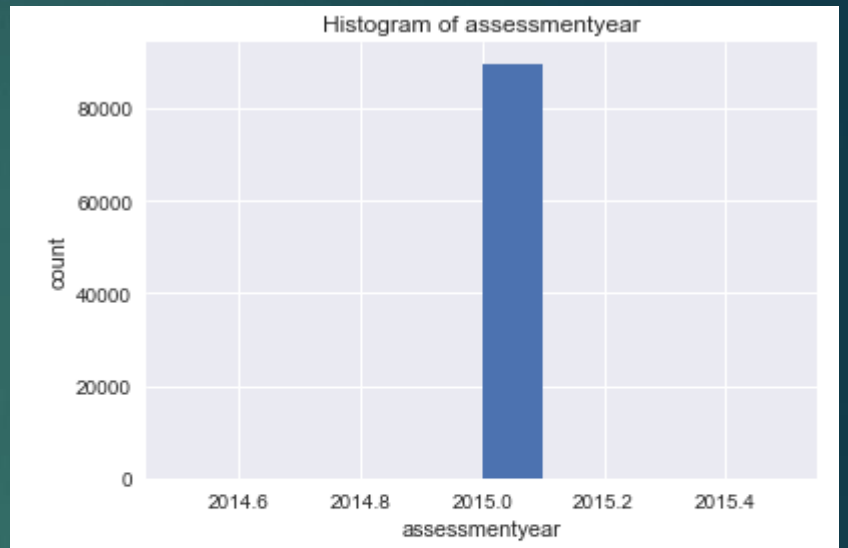
Data Cleaning and Feature Engineering

- *3 columns are useful for other purposes:*
  - *Parcelid*
    - *Index value*
  - *Logerror*
    - *Target value used for evaluation*
  - *Transaction Date*
    - *Potentially useful for further evaluation and feature generation*

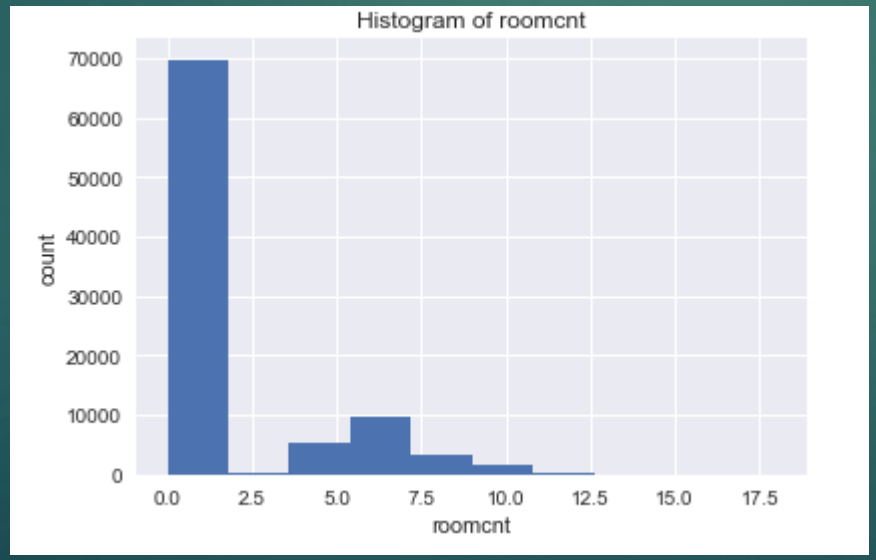
	parcelid	logerror	transactiondate
0	11016594	0.0276	2016-01-01
2	12098116	-0.0040	2016-01-01
3	12643413	0.0218	2016-01-02
4	14432541	-0.0050	2016-01-02
5	11509835	-0.2705	2016-01-02

Data Cleaning and Feature Engineering

- *2 variables contain bad data*
  - *Roomcnt*
    - *Majority of values set to zero, structures with zero rooms make no sense*
  - *Assessment Year*
    - *All values set to 2015, meaningless.*

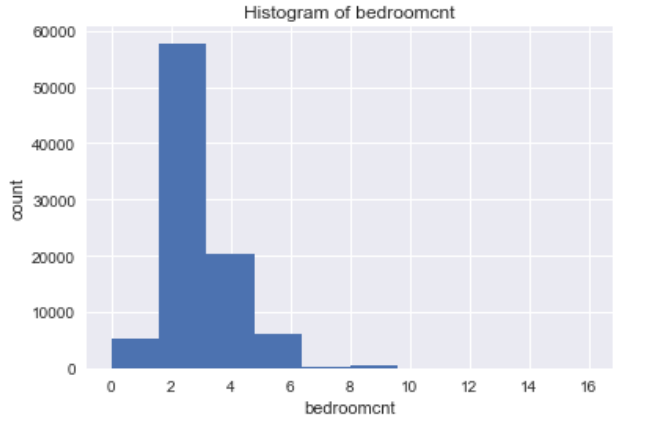



```
% Nan:0.0
Unique values:[ 2015.]
No. of unique values:1
Mean:2015.0
Median:2015.0
Mode:0    2015.0
dtype: float64
```

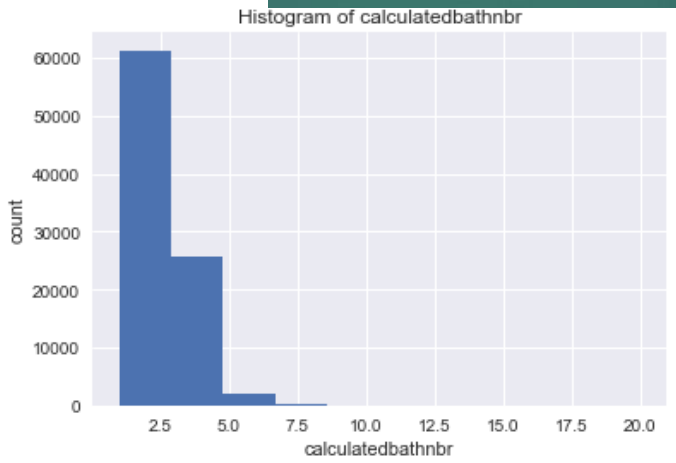


Data Cleaning and Feature Engineering

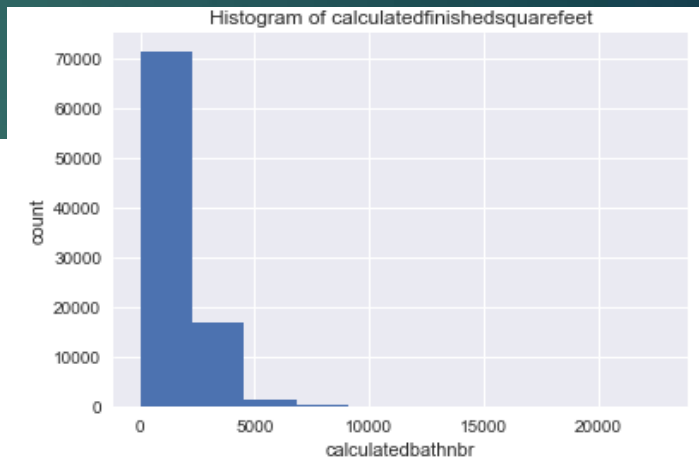
- *Remaining 12 columns potentially useful in model building.*




```
[ 3.  4.  2.  5.  1.  6.  7.  0. 12. 11. na
  14. 13. 15.]
% Nan:0.005902368655779586
Mean:3.0318692882857934
Median:3.0
Mode:0    3.0
dtype: float64
```




```
[ 2.   3.5  3.   2.5  4.   1.   5.   5.5  1.5  8.   na
  9.   7.   6.   10.   6.5  7.5  12.  11.  20.   8.5  15.
% Nan:0.018918412967592033
Mean:2.309216212272569
Median:2.0
Mode:0    2.0
dtype: float64
```

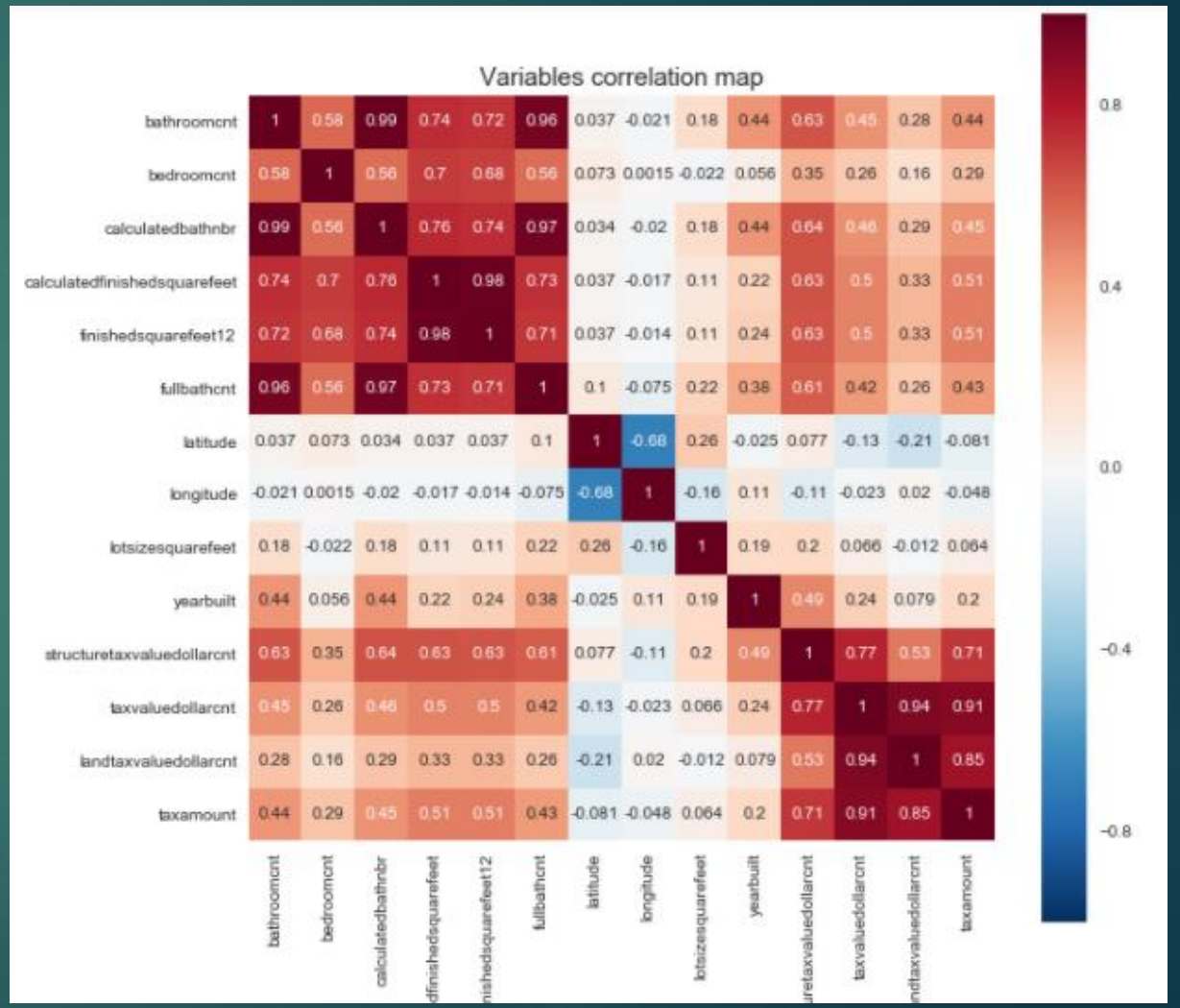



```
% Nan:0.01318122253911971
Unique values:[  1684.   2263.   2217. ...,   5937.  10681.  10014.]
No. of unique values:5102
Mean:1773.1859865645993
Median:1540.0
Mode:0    1200.0
dtype: float64
```

- *Each variable is individually assessed for missing values and filled accordingly.*

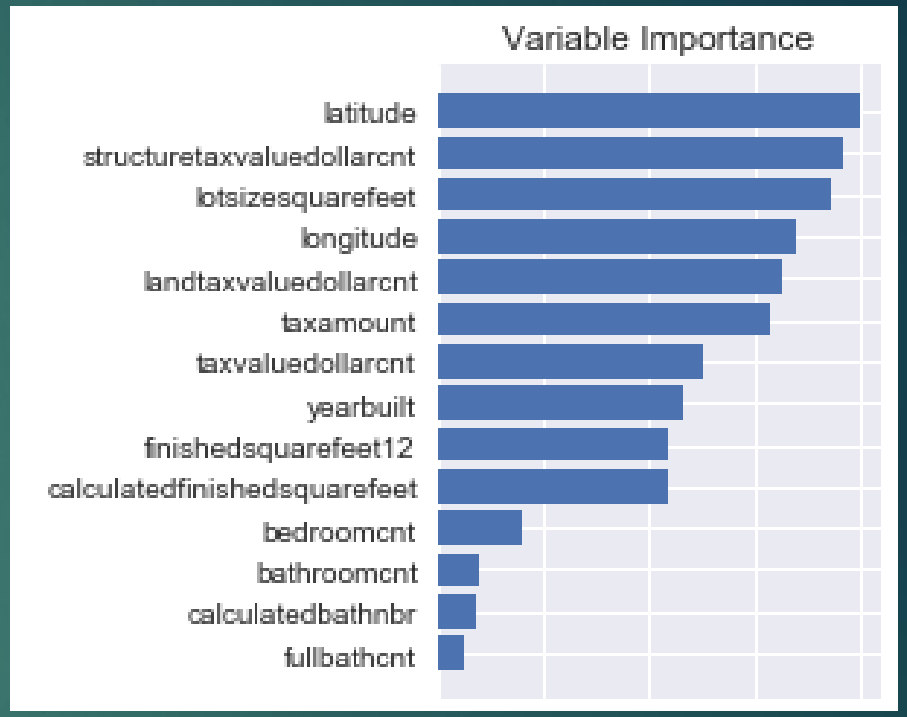
Correlation Analysis

- *Heat map shows heavily correlation in 2 blocks:*
- *bedroomcnt, bathroomcnt, calculatedbathbr, calculatedfinishedsquarefeet, fullbathcnt are all highly correlated.*
- *Structuretaxdollarvaluecnt, taxvaluedollarcnt, landtax valuedollarcnt, and tax amount all highly correlated.*
- *Yearbuilt and lotsizesquarefeet stand alone*
- *Latitude and Longitude highly negatively correlated*



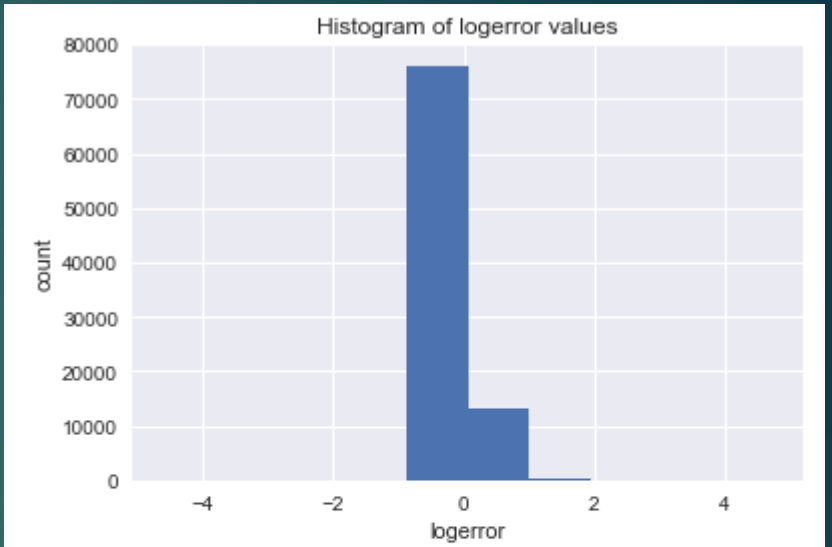
Variable Importance

- *Running a quick random forests and extracting variable importance to prediction of log error shows that the highly correlated group that relates to taxes are all very influential.*
- *Our stand alone variables of lot size square feet and year built also seem to hold some weight.*
- *The correlated group that related to structure size such as bedroom and bathroom count don't seem to show strong influence.*
- *Surprisingly Latitude and Longitude Appear to be very useful to modeling*



Target Data Analysis

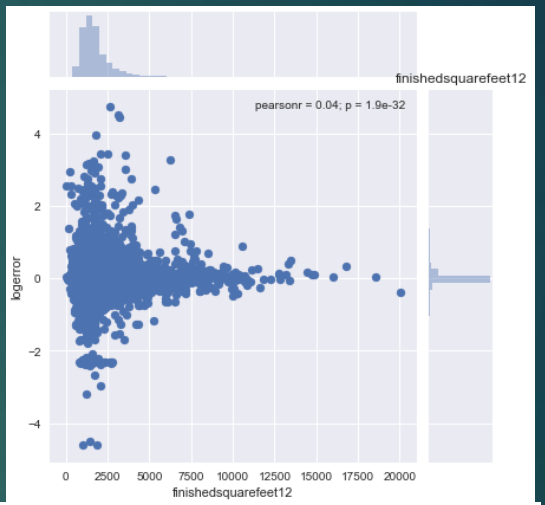
- *Logerror*
- *While there is a range of -4.6 to 4.6 more than half the data falls between -0.025 and 0.039. That's a very tight range for most of the data to fall in.*

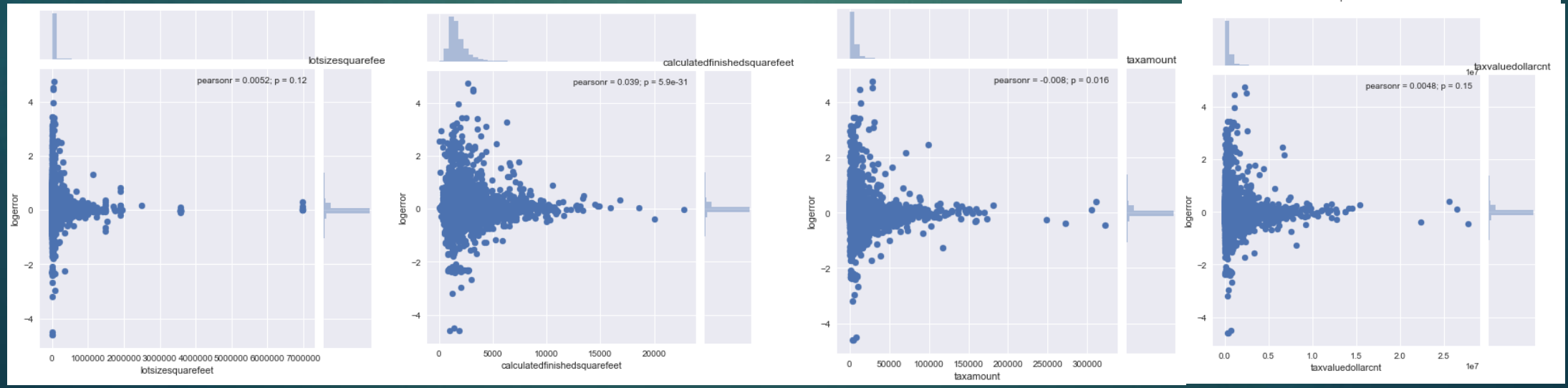



```
count    89669.000000
mean         0.011682
std          0.160438
min         -4.605000
25%         -0.025300
50%          0.006000
75%          0.039200
max          4.737000
Name: logerror, dtype: float64
```

Examination of Influential Features compared to logerror

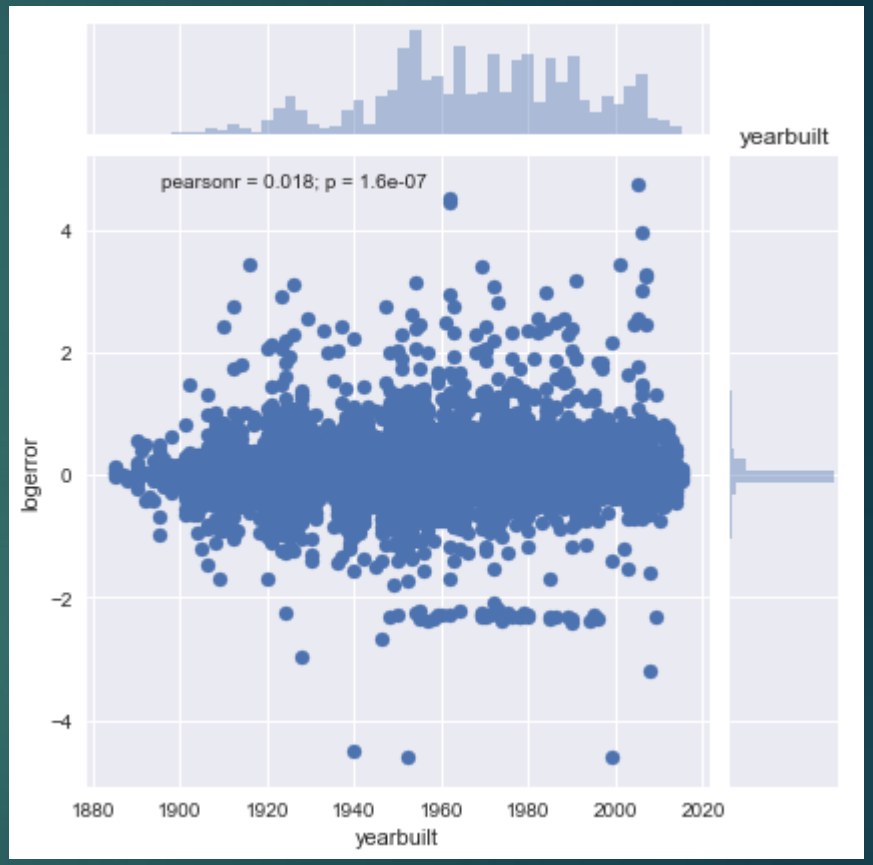
- *Calculatedfinishedsquarefeet, Lotsizesquarefeet, Taxamount, Taxvaluedollarcnt, Structuretaxvaluedollarcnt, finishedsquarefeet12 all share similar distributions, which makes sense considering how many were highly correlated.*





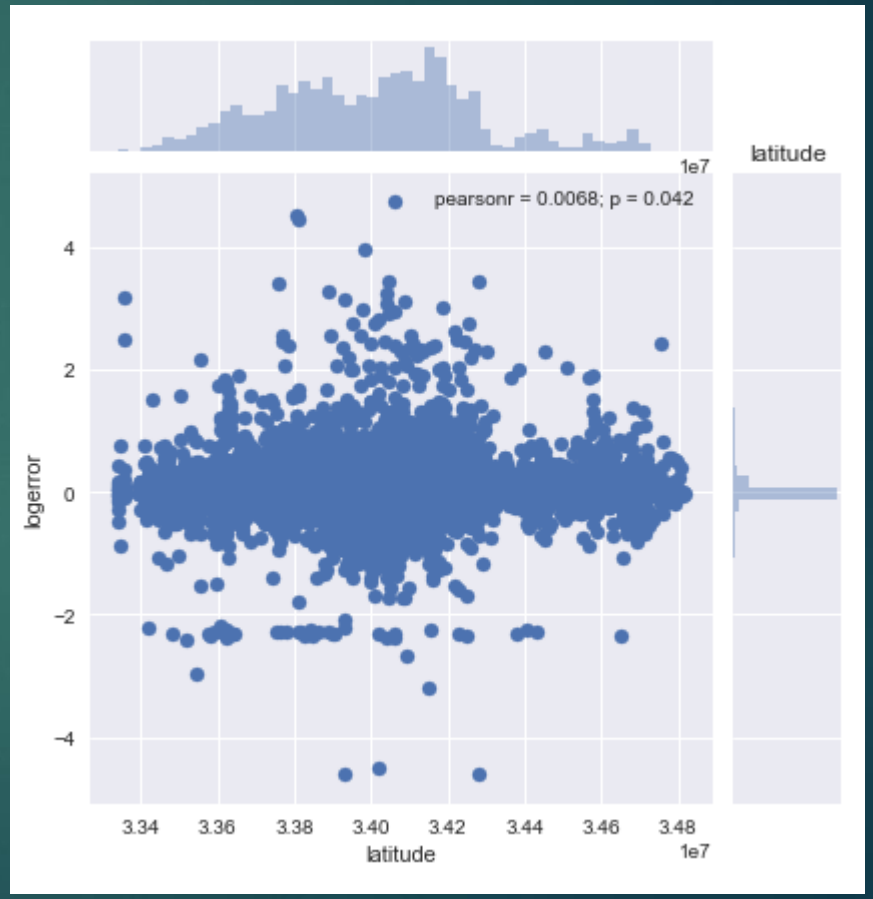
Examination of Influential Features compared to logerror

- *Yearbuilt*
- *Log error seems pretty evenly dispersed along year built.*



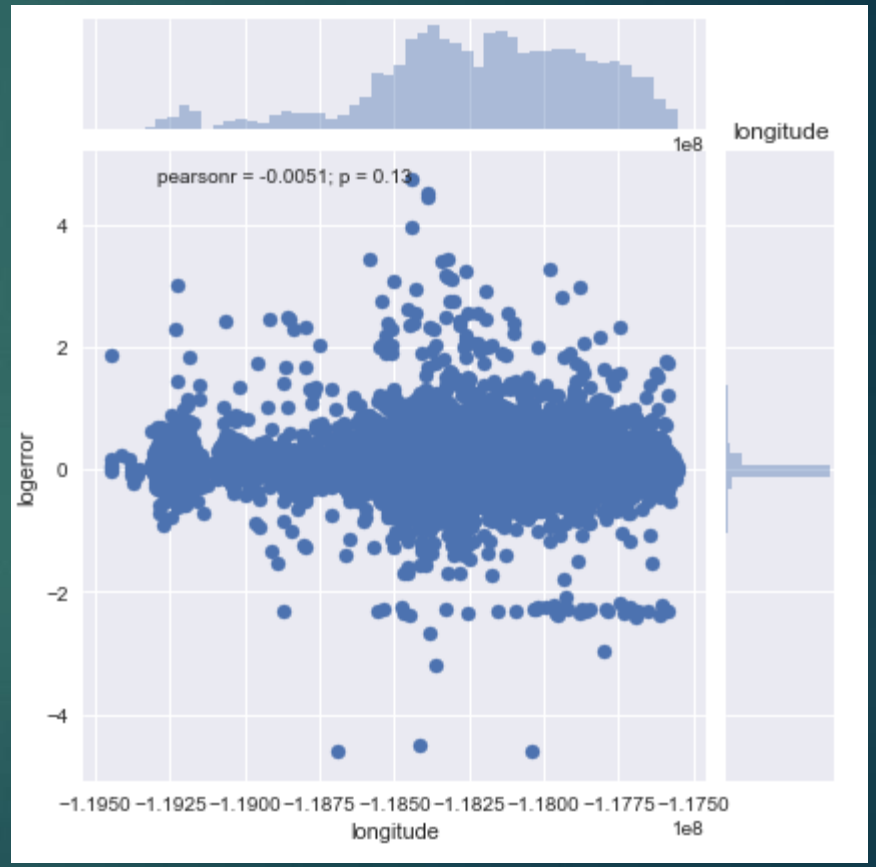
Examination of Influential Features compared to logerror

- *Latitude*

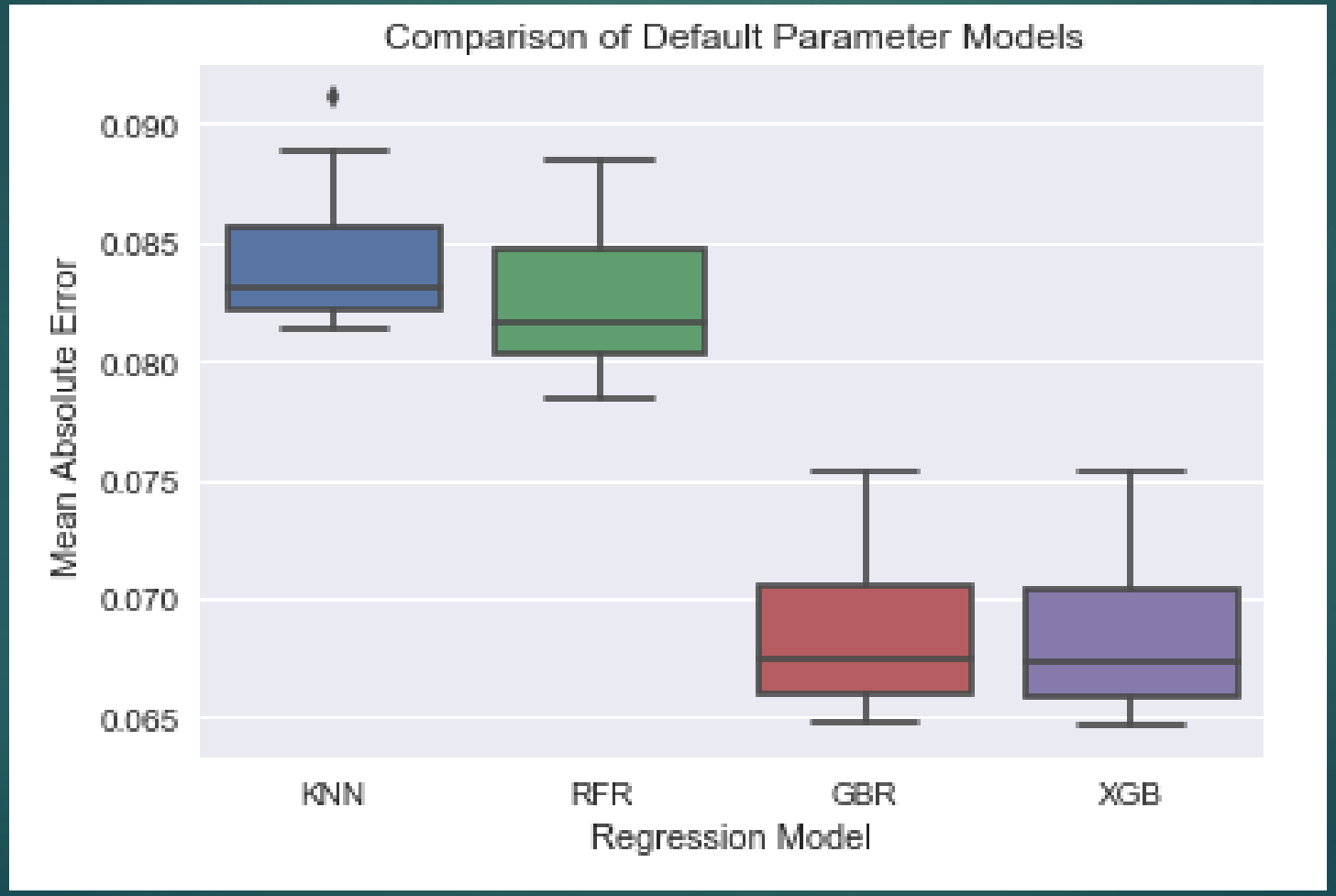


Examination of Influential Features compared to logerror

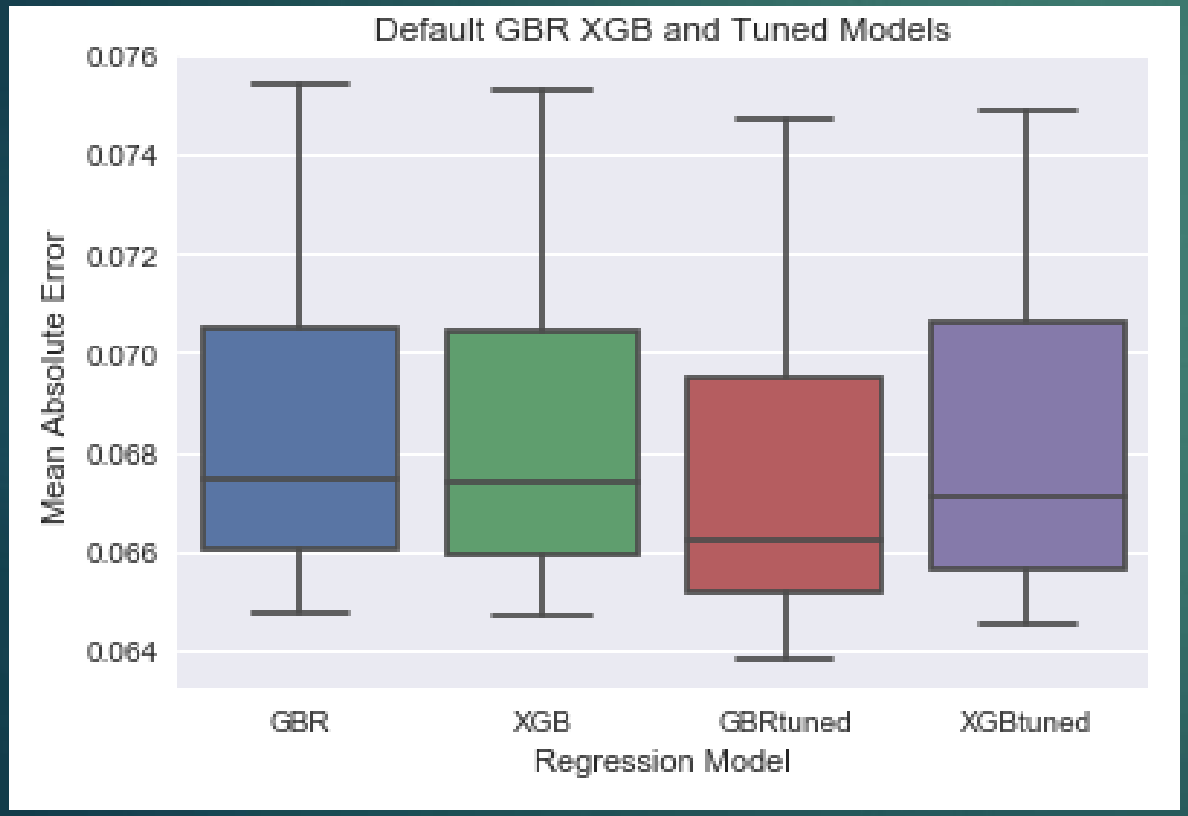
- *Longitude*

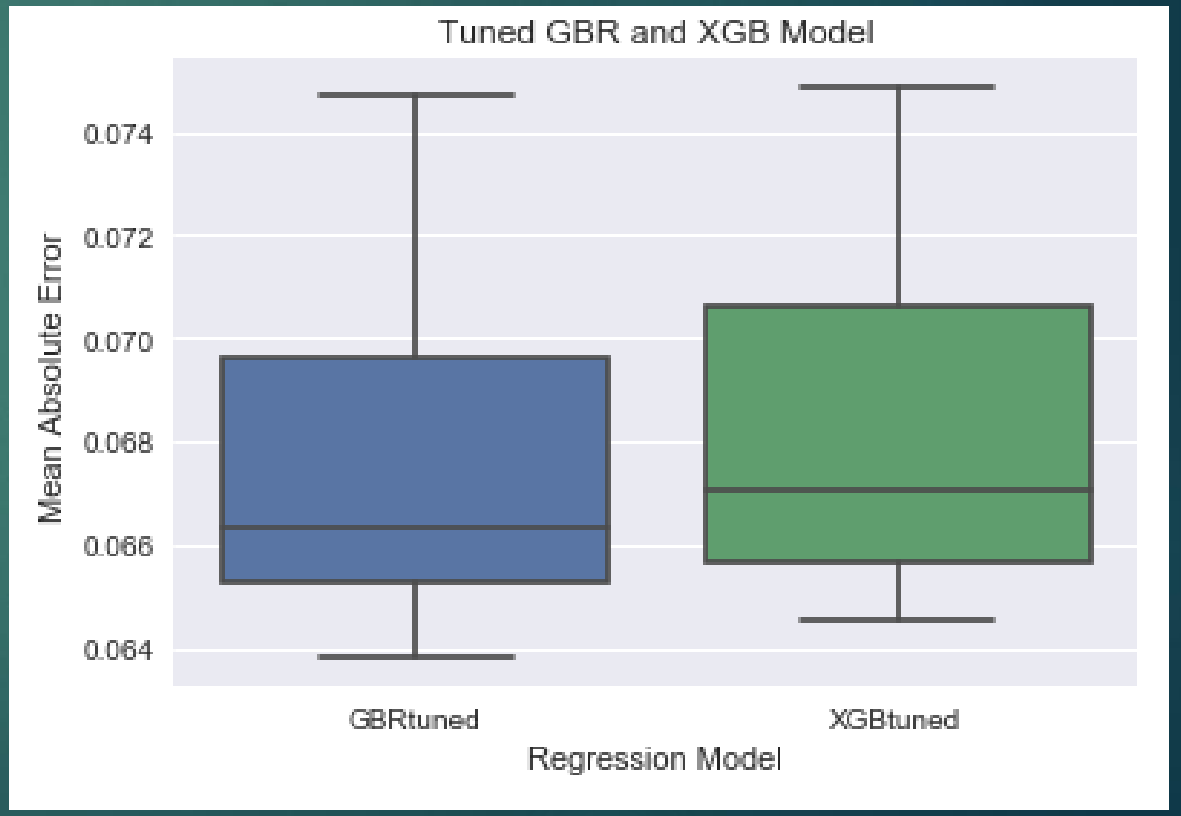


Model Building



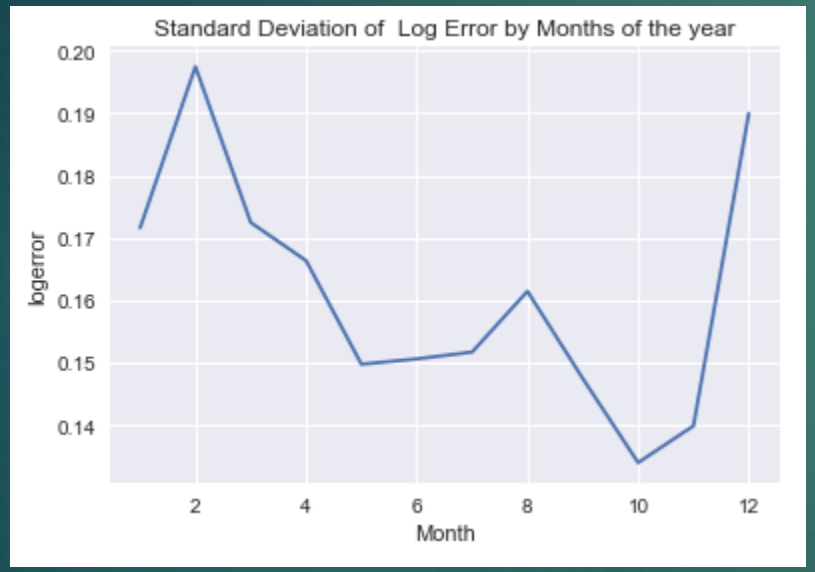
Hyper parameter Tuning





Predicting logerror by month

- *Standard Deviation and Mean of log error grouped by month*



Results

- *Baseline MAE (All predicted values as zero): 0.0663010*
- *Including all 14 continuous features and predicting MAE by fitting data only for corresponding months performed only slightly above the baseline*
  - *Fit and Predict by Month: 0.0660542*
- *Including all 14 continuous features and both scaling for average log error and scaling for standard deviations were roughly similar.*
  - *Standard Dev: 0.0649893*
  - *Average Log: 0.0650761*
- *Predicting all 6 targets (six months) identically and varying feature count improved results.*
  - *Drop Correlated (keep 6 features): 0.0650178*
  - *All features (14 features): 0.0649703*
- *The best final MAE resulted from having all 6 data points (six different months) reading identically and first 10 most important features.*
  - *Final Score: 0.0649469*

Conclusion

- *Zillow's Zestimates seem to be very accurate already, and further improvements of significant value seem difficult. The existing model seems very well designed.*
- *Despite high correlations discovered between many variables the very tight margin for the competition mean that highly correlated features can still contribute to predictions.*

Future Improvements

- *Further gains could be achieved potentially by lowering the threshold on other variables with high rates of Nan, bringing more data into the model.*
- *Further experimentation on relation of log error as it relates to month of sale could result in further improvement rather than predicting all six points as the same value.*
- *Analysis of categorical variables for feature generation and inclusion in modeling.*
- *Using multiple weighted models to predict high and low error separately*