Zillow

To Predict Errors on Zillow’s estimated home value

Data

The data from Zillow through Kaggle.com. The data is found at :

https://www.kaggle.com/c/zillow-prize-1

1. **properties_2016.csv**: The full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016 including 59 features.
2. **train_2016.csv**: all the transactions before October 15, 2016

Data Cleaning

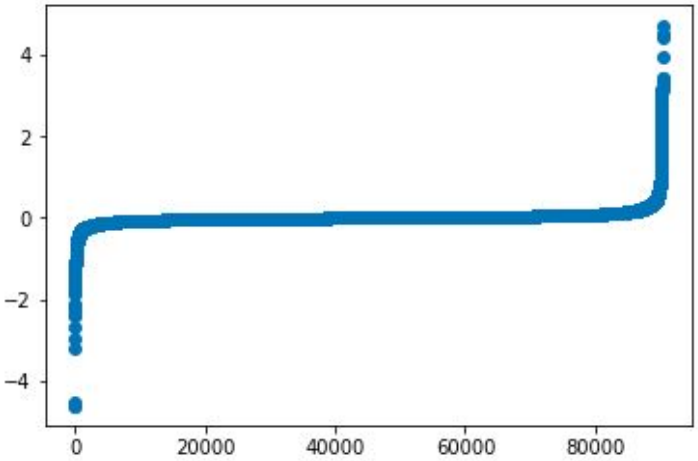
Duplication: 125 duplicated data, However, it meant they were transitioned for more than twice for a year.

Negative Values: Two columns, logerror and longitude, have negative values which are reasonable to have for them.

Unusual Object : 5 columns which are objective type. Each column does not have unusual values, for example “?” , “$”

Outliers

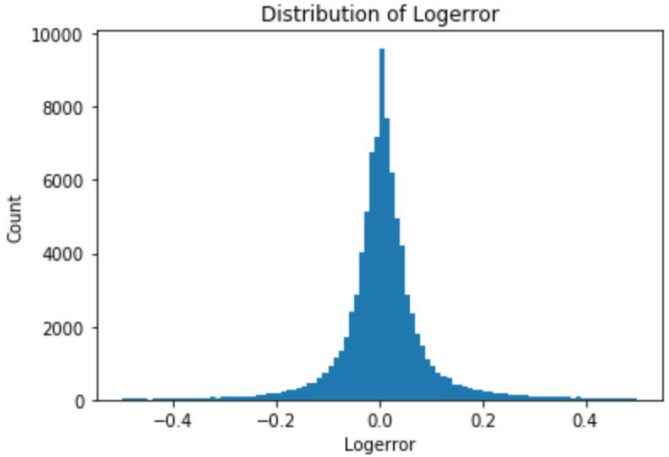
Let’s draw a scatter plot on "logerror". there are some outliers at the end of both sides. Our task in the project is to find where the zillow algorithm fails. These outliers mean where the zillow algorithm fails the most.

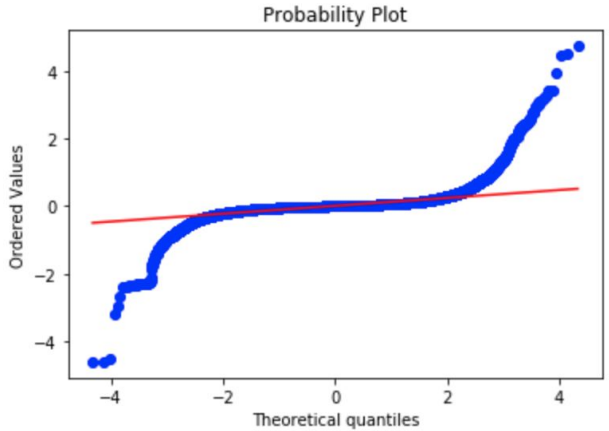


Data Storytelling

Distribution of Logerror

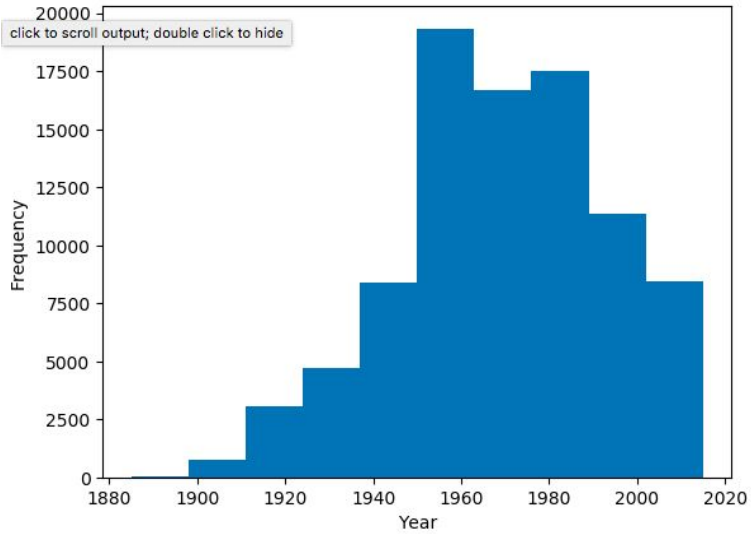
The distribution of logerror follows a normal distribution by checking QQ plot.

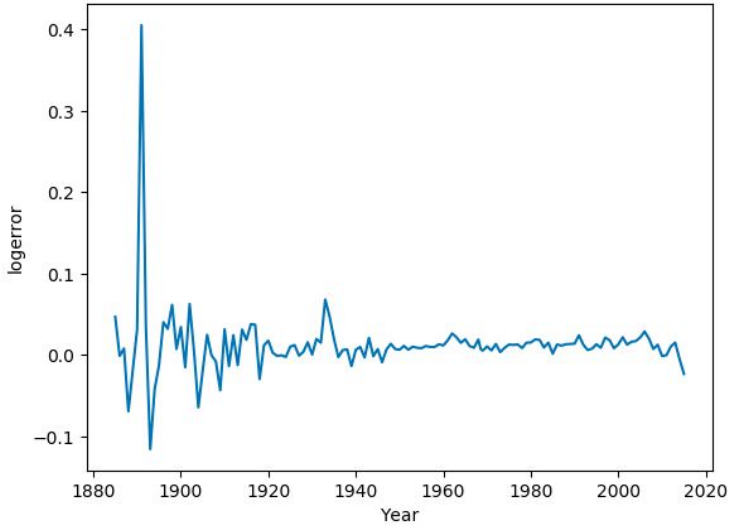




Built year

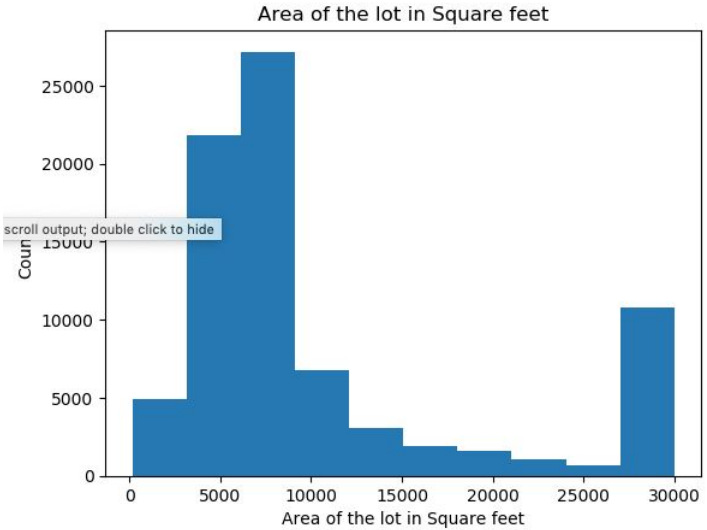
After observing plot for density of built year, we can find that most houses, 59.37%, are built between 1950 and 1990. Logerror is getting smaller with newer houses. Zestimate predicts home value better with newer homes.





Size of lot

"lotsizesquarefeet" is area of the lot in square feet. The mean of lot size is 29110. 25% of the data is between 5962 and 7570. The maximum is 6971010.

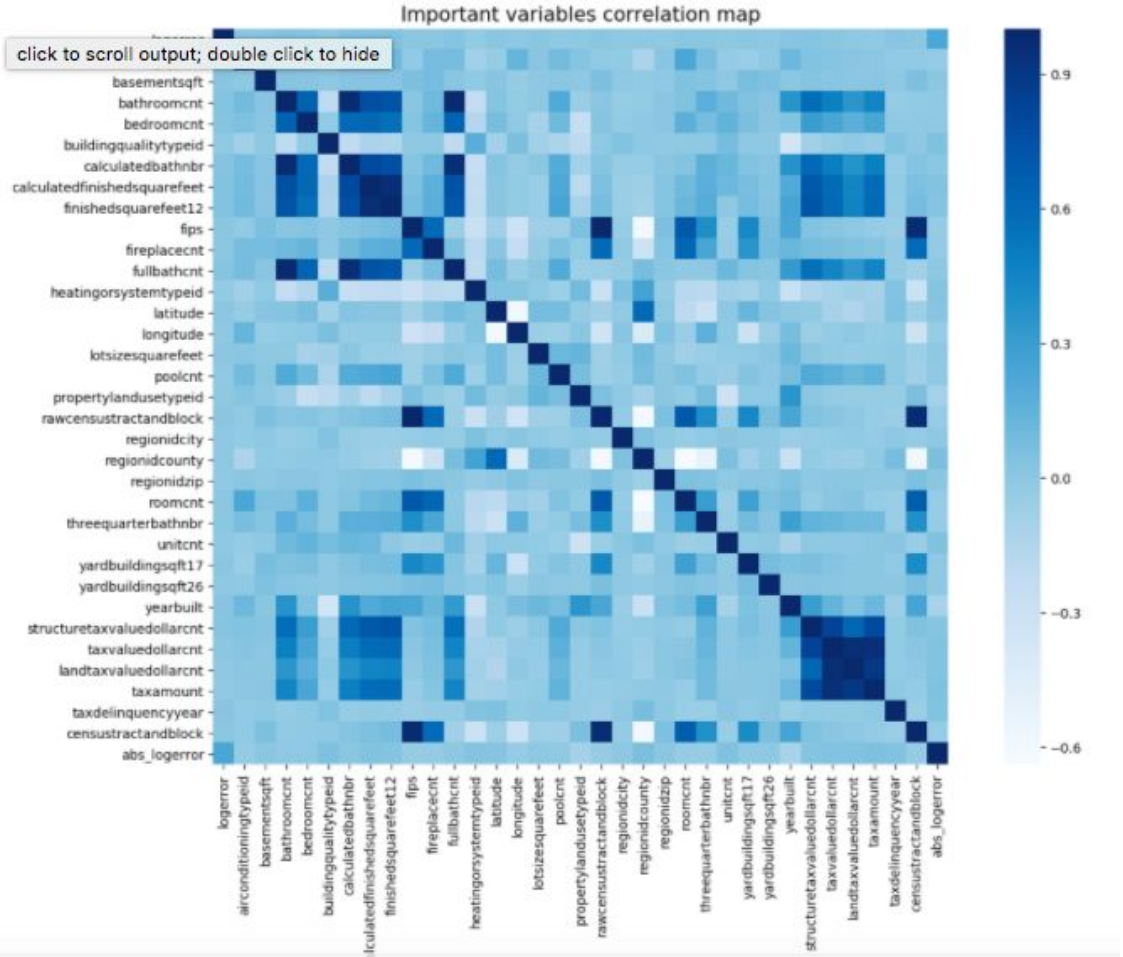


Inferential Statistics

not variables which can be particularly significant in terms of predicting logerror based on correlation.

Also, there are strong multicollinearity between dependent variables.

Therefore, a linear regression is not suitable for the model because of multicollinearity.



Machine Learning

Linear Regression

Training data was randomly chosen from 75% of entire data 50 times to get mean squared error. The mean squared error was 0.02603 which were the average of 50 mean squared errors. R squared is 0.0052 The top 10 properties which have large impacts on logerrors were like the table below.

	name	coeff
102	propertylandusetypeid_263.0	0.123889
103	propertylandusetypeid_264.0	0.046871
88	propertycountylandusecode_1	0.040308
108	propertylandusetypeid_275.0	0.039954
92	propertycountylandusecode_122	0.037242
93	propertycountylandusecode_34	0.037155
71	regionidcity_47568.0	0.025231
4	calculatedbathnbr	0.018124
43	heatingorsystemtypeid_18.0	0.017372
34	heatingorsystemtypeid_1.0	0.016581

Random Forest

To find the best fitted random forest model, grid search is used. possible combination of options were applied to find the better model. From grid search, the model with max_depth of 5, min_samples_spli of 20 and n_estimators of 30 was selected.

The MSE for the model was 0.02580.

name	description	coeff
taxamount	tax amount	0.158339
longitude	longitude	0.128765
finishedsquarefeet12	finished living area	0.119519
calculatedfinishedsquarefeet	total finished living room area	0.112954
taxvaluedollarcnt	total tax assessed value of the parcel	0.092970

Gradient Boosting Regressor

To find the best parameters for Gradient Boosting Regressor, Grid search was used. After the best parameter chosen.

The mse is 0.02488.

name	coeff	coeff
taxamount	taxamount	0.126420
structuretaxvaluedollarcnt	The assessed value of the built structure on the parcel	0.113069
taxvaluedollarcnt	total tax assessed value of the parcel	0.104929
calculatedfinishedsquarefeet	total finished living room area	0.085786
finishedsquarefeet12	finished living area	0.07742

Lasso

To find the alpha for Lasso try many possible variables, 0,0.0001,0.001, 0.01,0.1,0.5,1,2,3,4 for alpha, then choose the best variables.

The best alpa which gives lowest mse is 0.001 with mse of 0.02896

name	description	coeff
taxvaluedollarcnt	The total tax assessed value of the parcel	1.010697e-07
lotsizesquarefeet	Area of the lot in square feet	3.656718e-09
latitude	latitude	2.463666e-09
longitude	longitude	1.907122e-11
structuretaxvaluedollarcnt	The assessed value of the built structure on the parcel	-4.335424e-08
landtaxvaluedollarcnt	The assessed value of the land area of the parcel	-8.562764e-08
taxamount	The total property tax assessed for that assessment year	-2.273446e-06

Conclusion

As we see on the above, the Gradient Boosting Regressor has the lowest MSE. MSE is the measure of how predicted values are different from the actual values, the smaller MSE the better.

So Gradient Boosting Regressor is the model I chose. The variables which have larger impact on logerror are related to tax, living area and total assessment value.

We need to improve Zillow's estimated home value, Zestimate in those areas.

Model	MSE
Linear Regression	0.02603
Random Forest	0.02580
Gradient BoostingRegressor	0.02488
Lasso	0.02896