Kaggle Project
Machine learning

Zillow's Home Value Prediction (Zestimate)

Team Alpha: Nathalie Cohen, Yiming Wu, Stefan Hainzer, Summer Sun

Structure

- Introduction
- EDA
- Data cleaning
- Feature engineering
- Models
- Conclusion

Introduction

- Zillow: A real estate database company

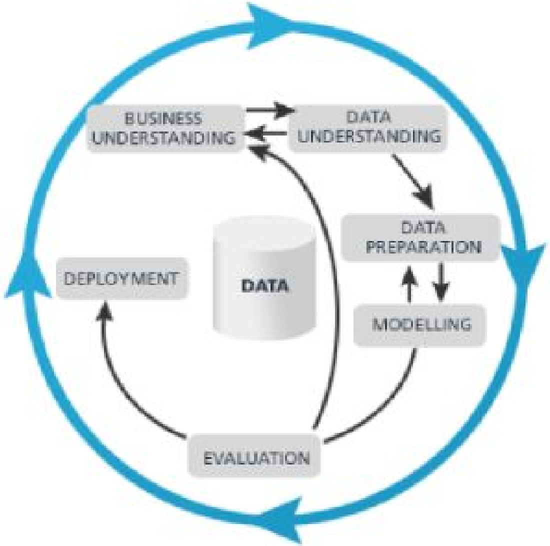
- Zestimate: Estimated home value based on 7.5 million statistical and machine learning models

- Improve the Zestimate residual error:

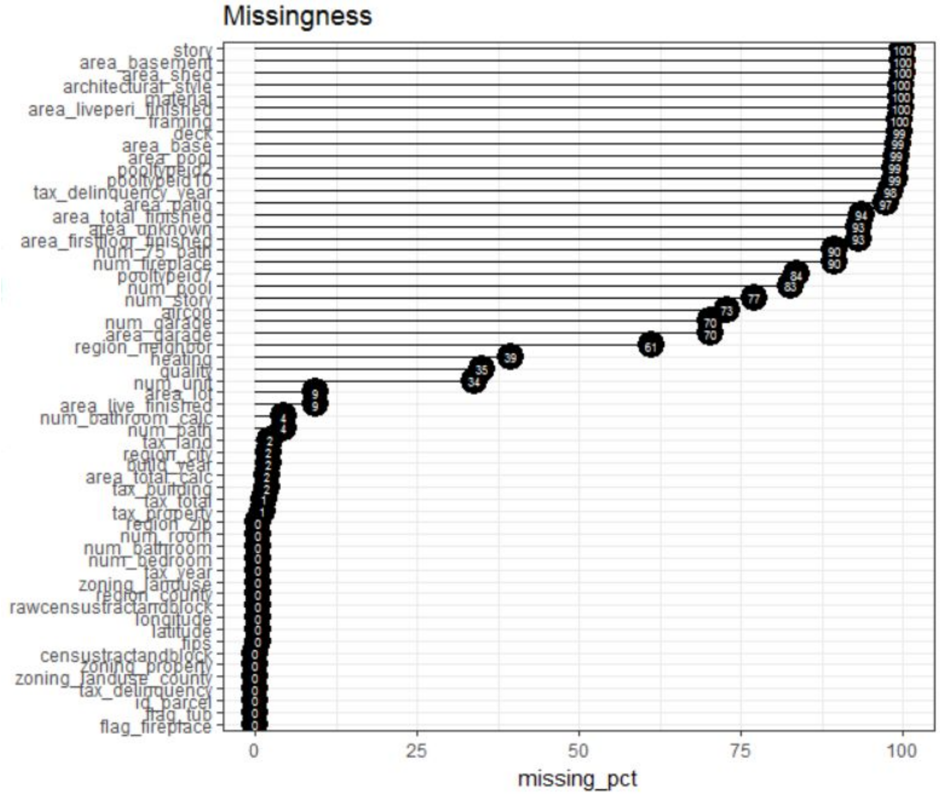
  $$logerror = log(Zestimate) - log(SalePrice)$$

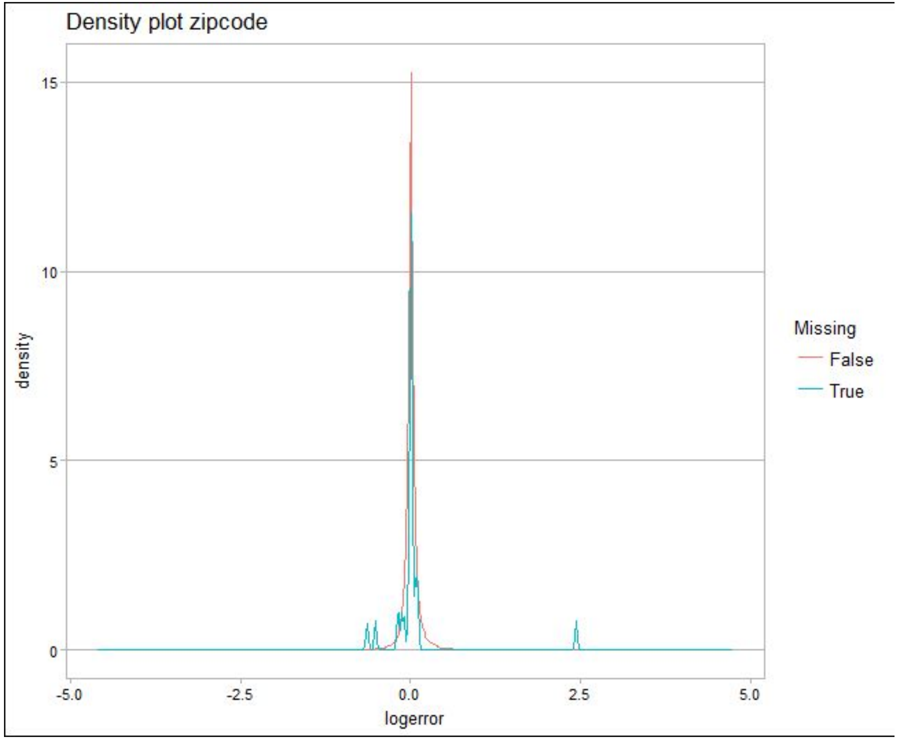
- Importance?
  - Median error rate of 4.3 percent → costs
  - Lawsuits

Workflow - CRISP



EDA - Understanding missingness

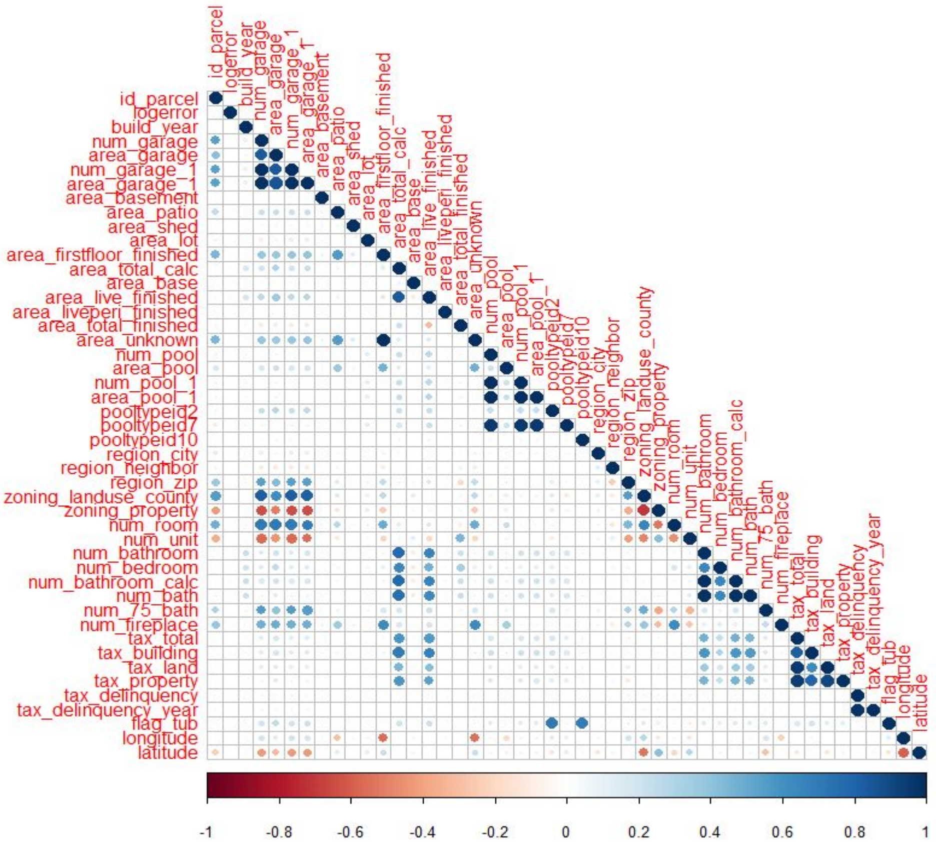




EDA - Understanding the variables

- Low correlations between logerror and the various variables

- Missing features/ranges of features where improvements are to be made?



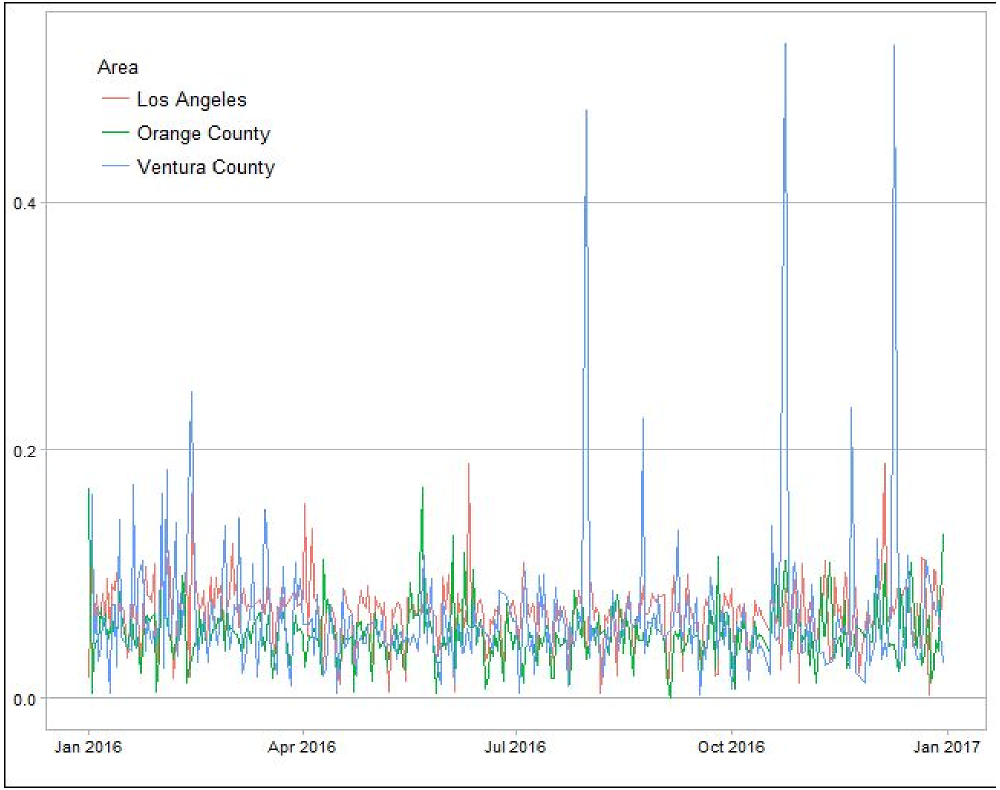
EDA - Understanding the market

- Houses geographically located in Los Angeles, Orange County, Ventura County

Housing market in 2016:

- “Median home prices in *Orange County* have surpassed their bubble-era height in mid 2016”

- “Huge demand for purchasing property and not many homes from which to choose”
  → Does market sentiment result in inaccurate Zestimate?



Data cleaning

Two approaches:

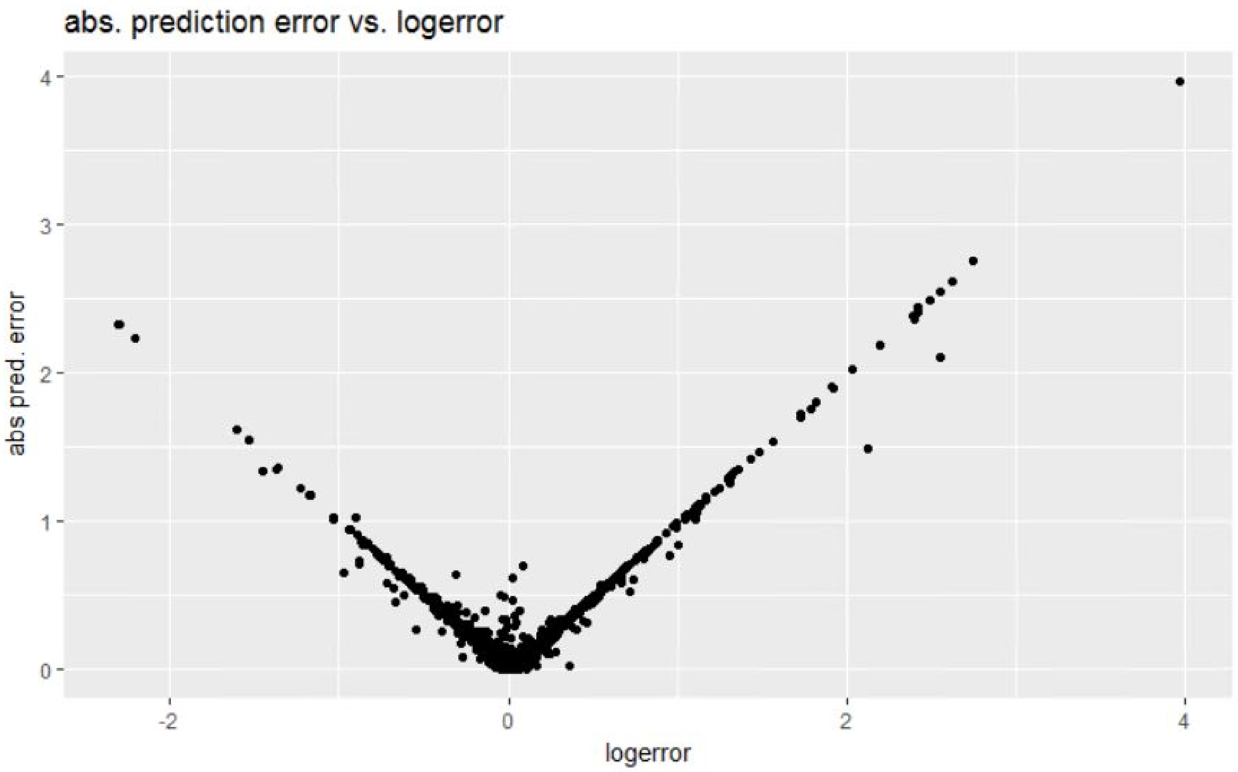
1. Assuming the logerror from Zillow was introduced by NA values
    - Preserved all columns, made some reasonable imputation;
      e.g. area / num_garage, area / num_pool, etc.
    - set NAs to zero;
    - Shrunk number of levels to fit in different methods, *i.e.* Rpart, RandomForest.

2. Using MICE imputation, random select ('sample')
    - Deleted columns with more than 75% missingness;
    - Removed duplicated, highly correlated columns, which may cause collinearity;
    - Scaled the geographical information;
    - Removed all NA property observations

Tested Models

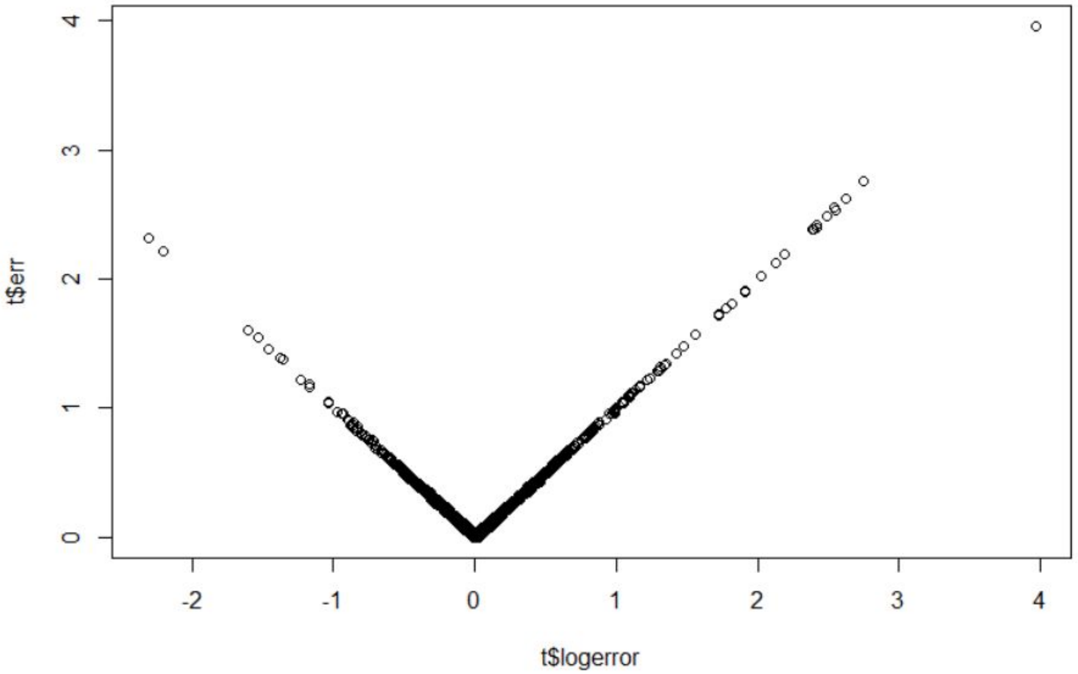
Regression	Classification
Ridge Regression	Decision Tree
Lasso Regression	Logistic Regression
Decision Tree	Random Forest
Random Forest	Gradient Boosting Machine
Gradient Boosting Machine	
XGBoost	

Evaluation of Prediction Error



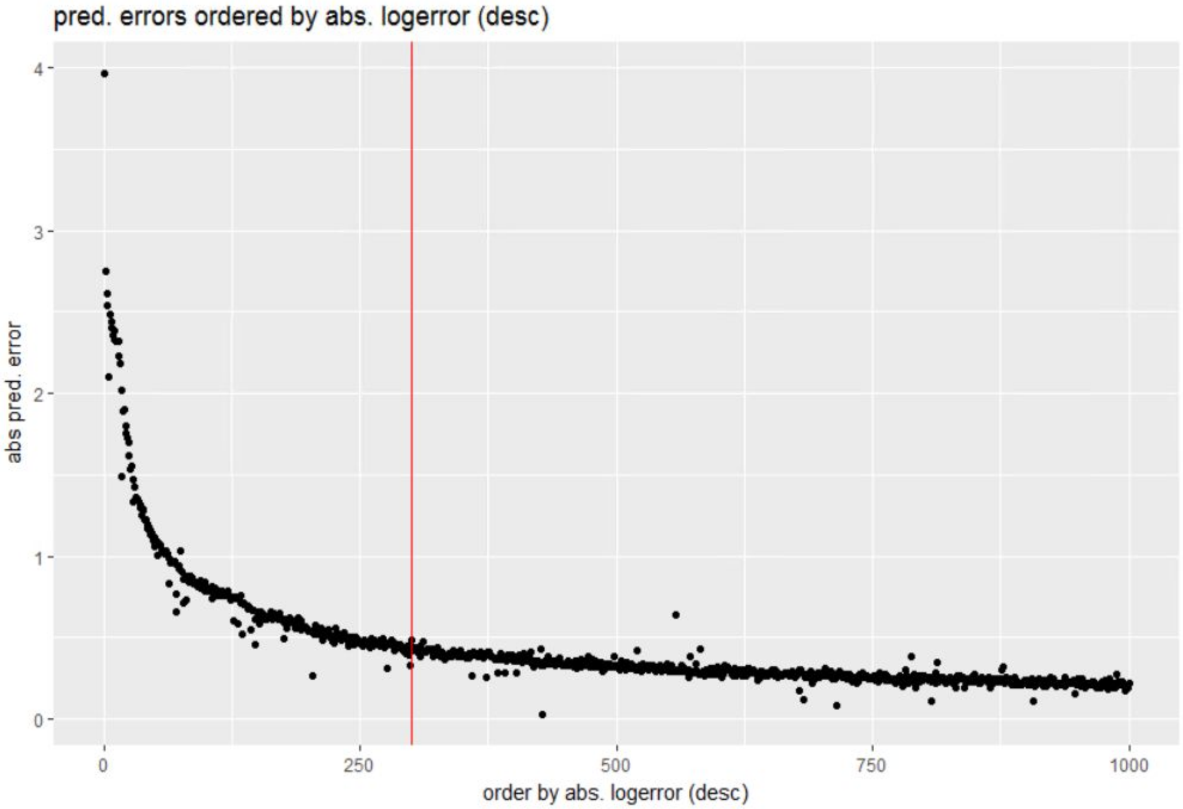
- Prediction error depends linearly on logerror.
- Low predictive power.
- Example is from a random forest..

Evaluation of Prediction Error



- Prediction error depends linearly on logerror.
- Low predictive power.
- Example is from a gbm..

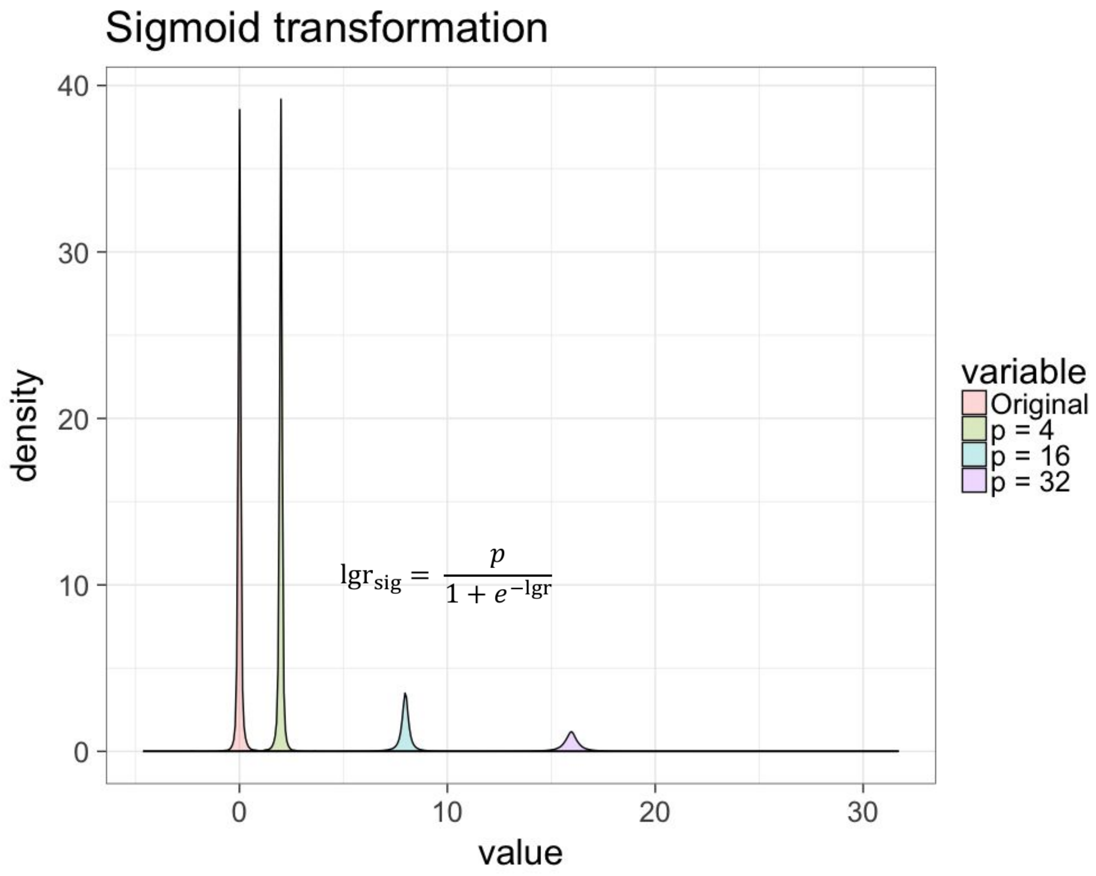
Identify MAE Drivers



- Overall:
  MAE = 0.06594
  (18055 obs.)
- Without highest 300 logerros:
  MAE = 0.05428
- Ability to predicting high logerrors improve MAE
- **There are only around 1000 observations with high logerrors in transaction data.**

Mastering high logerrors is a key success factor.

xgboost



- Sigmoid transformation on logerror

  slope around 0: $p/4$

- 10-fold cross validation

  For each fold, 100 iterations for best parameter selection

Public score: 0.06492

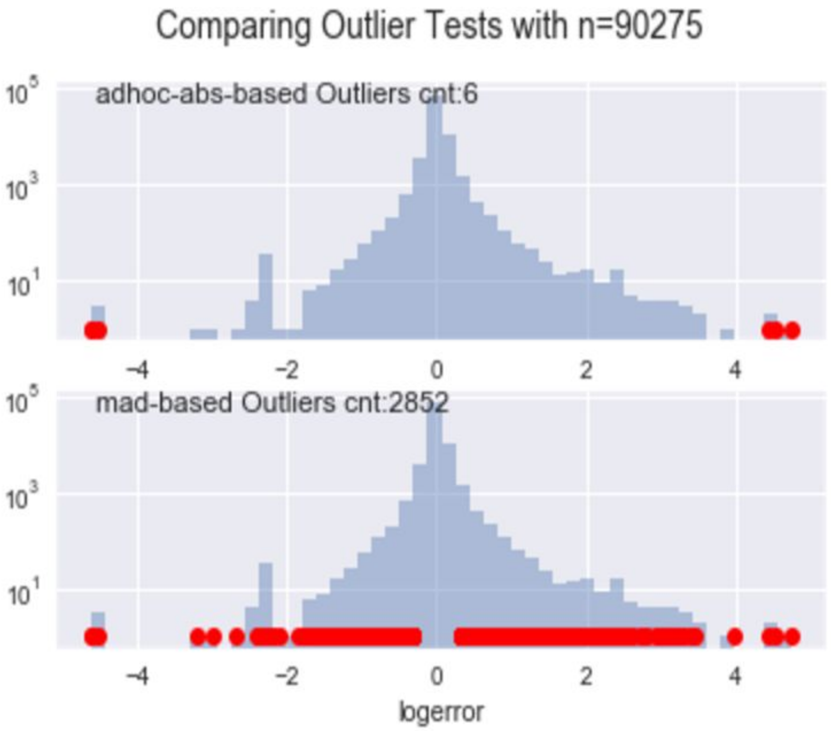
Feature engineering

1. Outliers
   - MAD methods
   - Percentile methods
   - Absolute Value methods

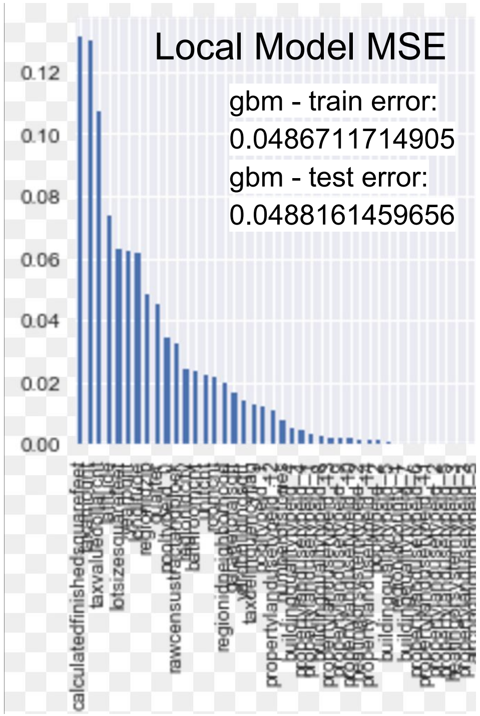
2. Combine Columns
   - Weighted average

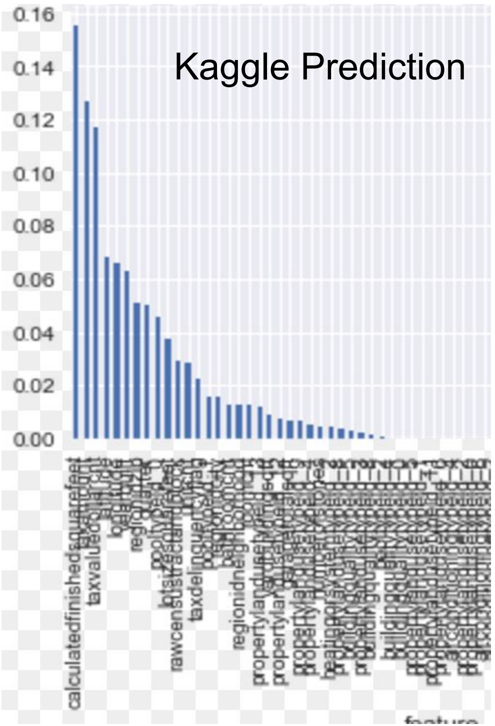
3. Observations
   - Binned Category
   - Dummies

4. Cross validations
   - Linear/Random Forest/GBM
   - Hyperparameters



GBM


Local Model MSE
gbm - train error:
0.0486711714905
gbm - test error:
0.0488161459656
0.12
0.10
0.08
0.06
0.04
0.02
0.00
calculatedfinishedsquarefeet
taxvaluedollarcnt
lotsizesquarefeet
rawcensustractandblock
regionidneighborhood
propertylandusetypeid


Kaggle Prediction
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
calculatedfinishedsquarefeet
taxvaluedollarcnt
rawcensustractandblock
taxdelinquencyflag
regionidneighborhood

Conclusion

- Predicting a predictor is a hard job
- Small portion of the data will make the difference
- Feature engineering is key
- XGboost is fast and practical (cross-validation)
- Zillow, keep your logerror!