Overview

The Project is to build a model to improve the Zestimate residual error.

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

“Zestimates” are Zillow's estimated home values. The model is to predict the difference between the Zillow's estimated home value, Zestimate, and the actual sale price.

Client

The client could be Zillow. Zillow can improve its algorithm with the model which would predict where zestimates will do good or bad. When we want to improve existing model, modeling errors can be a good way to find areas to improve the existing model.

Data

The data used in the project has been provided from Zillow through Kaggle.com.

The following files were used in the project. We will use data in 2016 as a train data and 2017 data as a test data. Data have 60 columns which have features for homes.

1. **properties_2016.csv**: a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016.
2. **train_2016.csv**: all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. It contains parcel ID , transaction date and calculated log error .
3. **properties_2017.csv**: a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2017.
4. **train_2017.csv**: all the transactions from Jan 1, 2017 to Sep 25, 2017. It can be used as a test dataset.

Approach to solving this problem

First of all, find out any variables which relate to the target variable, logerror. Then, apply regression model or any statistical model which can be applicable.

Deliverables

This would include code and a report.

Data Wrangling

Data Cleaning

- Duplication : I explored training data. 125 duplicated data for 2016 and 199 duplicated data for 2017 data were found. However, it meant they were trasacted for more than twice for a year. So, I didn’t delete any duplication.

- Negative values: Also, I checked if there were any negative numbers for each column. Two columns, logerror and longitude, have negative values which are reasonable to have for them.

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

Missing Values

Let's check how many missing value each column has. I found that 47 columns out of 60 columns have missing values and 18 columns among them have more than 95% of missing values.

Let's explore the top five columns which have the most missing values for now.

column	description	# missing value	% missing value
buildingclasstypeid	The building framing type	16	99.98%
finishedsquarefeet13	Perimeter living area	33	99.96%
basementsqft	Finished living area below or partially below ground level	43	99.95%
storytypeid	Type of floors in a multi-story house	43	99.95%
yardbuildingsqft26	Storage shed/building in yard	95	99.89%

- buildingclasstypeid: It represents the building framing type such as steel frame, wood frame or concrete/brick. Only 16 cells out of 90275 cells have the same value that is '4' and rest of them have all missing values. The '4' means that buildings have wood or wood and steel frames. It is not reasonable to replace missing values with other statics such as mean or median etc since most of them are missing. So I will delete the column.

- finishedsquarefeet13 : It is perimeter living area has 33 cells which are not missing. Every building has different sizes of living area so it is hard to decide what value to fill out for missing values. So I will delete the column as well

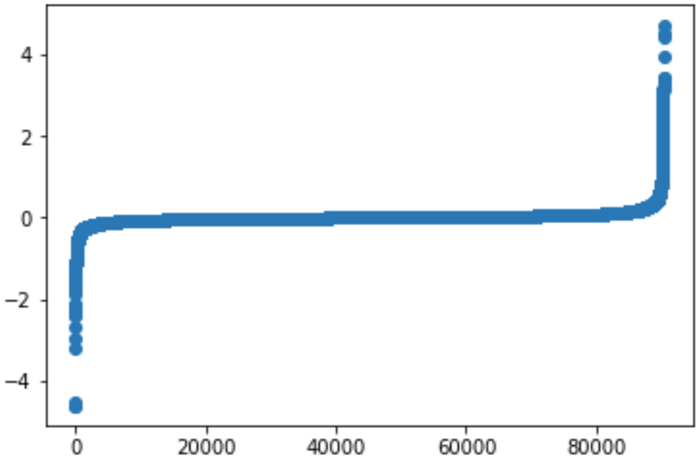
- basementsqft : It has 43 cells out of 90275 which are not missing. 'basementsqft' represents finished living area below or partially below ground level. There fore missing value might mean that it has not finished living room below ground leve. So I will filled missing value with 0.

- storytypeid: It is the type of floors in a multi-story house. It has 43 cells which are not missing and all have the same value of 7. Values range from 1 to 35. Every building has floor , so it should always have the value. Therefore, there is not any meaning for missing values. so I would delete the column

- yardbuildingsqft26: it is storage shed/building in yard. It has 95 cells which are not missing. Missing value might mean that it has no storage in yard. So I will fill missing values with 0.

Outliers

Let's draw a scatter plot on "logerror", then we can find that 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 means where the zillow algorithm fails the most. Thus, I will leave outliers just like that.


4
2
0
−2
−4
0
20000
40000
60000
80000