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.

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.

Data Wrangling

Data Cleaning

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.

Missing Values

Let's check how many missing value each column has. Even though there are many missing values, I will just leave like this for now as we are not sure which columns would be needed

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.