EDA

Yuman Lin

Pay attention!

Goal of the Zillow competition:

predict the log-error between their Zestimate and the actual sale price, given all the features of a home.

We need to predict:



Metrics in the competition: MAE

The data we have: You are provided with a full list of real estate properties in three counties **(Los Angeles, Orange and Ventura, California)** data in 2016.

**General process:**

1. Merge property csv with train csv, remove all column which does not have logerror column – the shape new dataframe: 167888 \* 60
2. Drop some useless columns (such as parcelid)
3. 21 category columns and 32 numeric columns

**Missing values and outliers:**

Numeric columns:

1. Remove columns which missing more than 90% of the values
2. Replace nan with mean values

Category columns:

1. Remove some columns (see analysis process in code)
2. Replace null value with False for binary column
3. Replace null value with mode/ or the most frequent type (according to the column definition in excel) for other category columns

Data analysis results:

1. None of the features have strong relationship with logerror. (PCA and LDA won’t work, feature selection ❌)
2. Some of the features suffer from multicollinearity, such as longtitude and latitude, bathroomcnt and calculatedbathroomnbr.

Suggested model:

1. XGBoost
2. LightGBM