Evaluating Predictive Models for the Log Error Values of Zillow's Sale Price Zestimate

Elizabeth Vincent

October 11, 2017

Abstract

Abstract placeholder.

1 Introduction

the Zillow Prize competition on Kaggle

Data for approximately 3 million real estate properties in Los Angeles, Orange, and Ventura counties in California

Six time points to predict for the 3 million homes: October, November, and December of 2016, and October, November, and December of 2017. Data on when or if the properties were sold is not available except for the ”training” data sets, which are properties for which the log error is provided, indicating they were sold. In theory we could check if the others were sold and use the log error of those homes to further evaluate the model, but it would not help the actual standings because 2016 data is only used for the public leaderboard and does not actually factor in to standings. The competition closes on October 15 as the predictions will be evaluated on real estate sold between October 15 and December 31, 2017.

The goal of the competition is to accurately predict the log error of the Zestimate, where the log error is defined as

$$log(error) = log(Zestimate) - log(SalePrice) \tag{1}$$

Submissions to the competition will be evaluated based on the mean absolute error (MAE) of the predicted vs. the actual logerror of the Zestimate, where the MAE is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where y_i is the actual log error and $\hat{y}_i$ is the predicted log error.

Using MAE instead of RMSE because data is skewed and heteroscedastic (check paper cited on the kaggle competition for more details as to why)

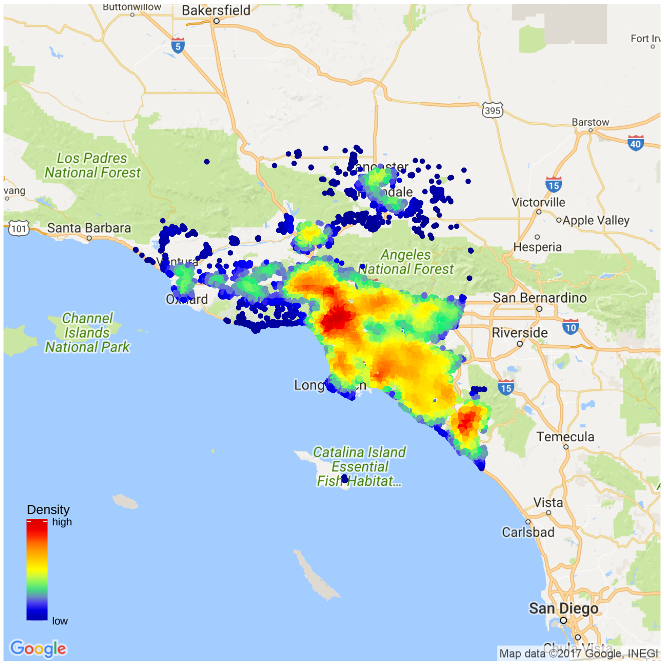

Figure 1. Property Locations: The locations of the parcels from the Zillow data are shown above. Red indicates a higher density of properties and blue indicates a lower density. The parcels are located in the Los Angeles, Orange, and Ventura counties, CA.

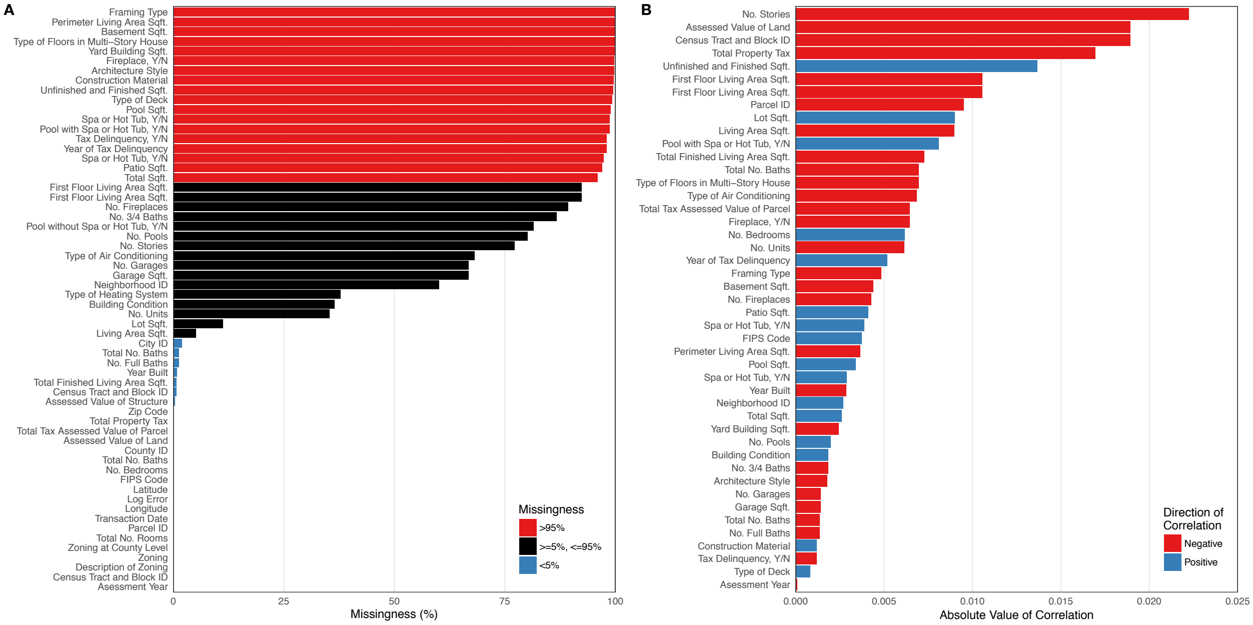

Figure 2. Missingness of Data: Missingness (degree of missing data) was calculated for each variable by counting the number of rows missing data for a given variable in the training data. Some variables share the same name because several variables in the data are redundant. (A) The bars represent the percentage of observations for which there is no data for each variable in the training data. Variables for which data is missing in more than 95% of observations are shown in red, those for which data is missing in at least 5% but no more than 95% are shown in black, and those for which data is missing in less than 5% are shown in blue. (B) The bars represent the absolute value of the correlation of missing data for a given variable with the log error. Red bars represent positive correlation, blue bars represent negative correlation.

2 Methods

1. Data gathering
2. Data cleaning
3. Missingness
4. Imputation
    - City and zip based on unique city-zip or zip-city combinations, and based on long/lat coordinates
5. EDA
    - Correlation of missingness between vars
    - Correlation of logerror with missingness
    - Correlation of logerror with numeric vars
6. Analysis
    - Predict the mean
    - Linear model
    - Random forest

3 Results

RMSE and R^2 for each method. Compare methods, which gave lowest RMSE and highest R^2

4 Discussion

Interpretation of what factors influence the logerror, which method was best.