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

Elizabeth Vincent

October 14, 2017

Abstract

Abstract placeholder.

1 Introduction

Zillow is an online real estate marketplace founded in 2006 that, in addition to providing real estate listings, provides value estimates for properties. The value estimate of a property, termed "Zestimate", is calculated from public and user-submitted data using a proprietary formula that takes into account property features, location, and current market conditions [1]. With the goal of improving the accuracy of the Zestimate, Zillow launched a competition on Kaggle, a cite that hosts data science- and machine learning-based competitions. The goal of the Zillow competition is to predict the accuracy of the Zestimate for a given property at a given time point. The competition requires prediction of the log error of the Zestimate, where the log error is defined as follows:

$$log(error) = log\left(\frac{Zestimate}{SalePrice}\right) \tag{1}$$

Property prices are right-skewed and heteroscedastic, so a relative error metric, the log ratio error, is used as opposed to an absolute error metric, such as root mean square error (RMSE), to avoid biasing evaluation models towards more expensive homes [2].

Zillow has provided data for approximately 3 million properties in Los Angeles, Orange, and Ventura counties in California, shown in Figure 1. The competitor must predict what the log error of the Zestimate for each of the properties will be when the property is sold. The competitor must make predictions for all 3 million homes for three transaction periods: October, November, and December of 2017. The competition closes on October 16, 2017 and predictions will be evaluated based on the sale prices of properties that sell between October 17, 2017 and December 15, 2017. The accuracy of the predicted log error for the transaction period in which a property is sold will be evaluated based on the mean absolute error (MAE), where the MAE is defined as follows:

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

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

Paragraph about the different models (predict the mean, glm, ML), the purpose each model serves, what its strengths and weaknesses are.

2 Methods

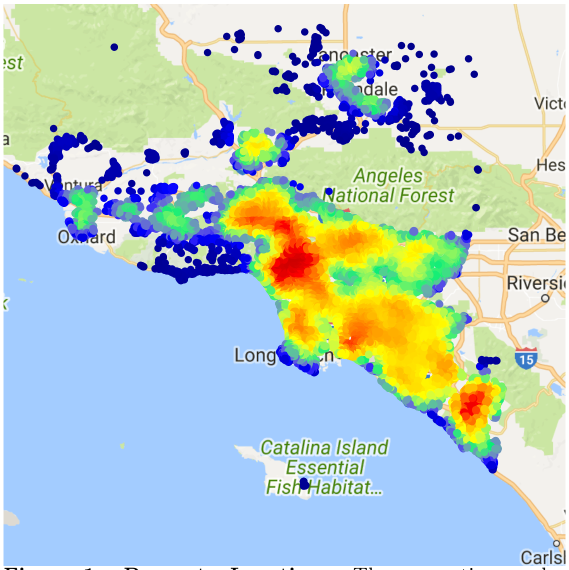

Figure 1. Property Locations: The properties are located in Los Angeles, Orange, and Ventura counties, CA. Red indicates high density of properties and blue indicates low density.

The data were obtained from the ZIllow competition on Kaggle and consist of 58 observations values for 2985217 properties that were split by Zillow into a training group and a test group, for which the true log error is provided. Not all 58 values are independent, nor is every value available for every property. A binary matrix of the same dimensions as the data was created to record for each property-value pair whether the datum is available (0) or not available (1). To ascertain the degree of missing data, termed missingness, the percentage of properties missing data for a given value was calculated as the column missingness, and the total number of missing values for a given property was calculated as the row missingness. Correlation for missingness was calculated for the following: the pattern of missingness for each column, or value, with the pattern of missingness for all other values; the pattern of missingness for each column with the log error; and the total number of missing entries for a row, or property, with the log error.

To decrease missing data, some values with low missingness were imputed based on other dependent values in the data. Latitude, longitude, city ID and zip codes are not independent because all are highly correlated with each other based on property location. For data that contained values for both city ID and zip code, dictionaries were made of all cities that are uniquely associated with one zip code, and all zip codes that are uniquely associated with one city ID. Not all city IDs and zip codes were uniquely paired. For properties containing either a zip code or a city ID, the missing value was imputed using the city ID and zip code dictionaries. To further impute city and zip codes, especially in properties that did not contain values for either, a distance matrix was calculated for all the properties that were missing data for either the city ID or zip code, as well as a random sample of 10,000 properties that contained data for both. The euclidean distance between two homes was calculated using the longitude and latitude of each home. City IDs and zip codes were imputed as the most frequent value from a home's 11 nearest neighbors.

Other values could be imputed by interpreting missing data as a negative response. The data contain information on whether taxes on a property were past due. All properties that were not flagged for tax delinquency were missing data for tax delinquency status. In this case, responses were imputed such that any missing data were interpreted as a negative response.

Redundant values did not always contain identical data, in which case the more informative value was kept and imputed based on the redundant value, which was then removed. For example, the data contain both a value for the type of floors a multi-story house contains, such as an attic or basement, and a value for whether a house has a basement. The only type of story recorded is basement, therefore the type of story was redundant with whether or not the house had a basement. Properties that were missing data for whether or not they have a basement but were recorded as having a basement type of floor were imputed as having a basement. The type of floor value was then removed from the data set.

Analysis

- Predict the mean
- Linear model
- Random forest

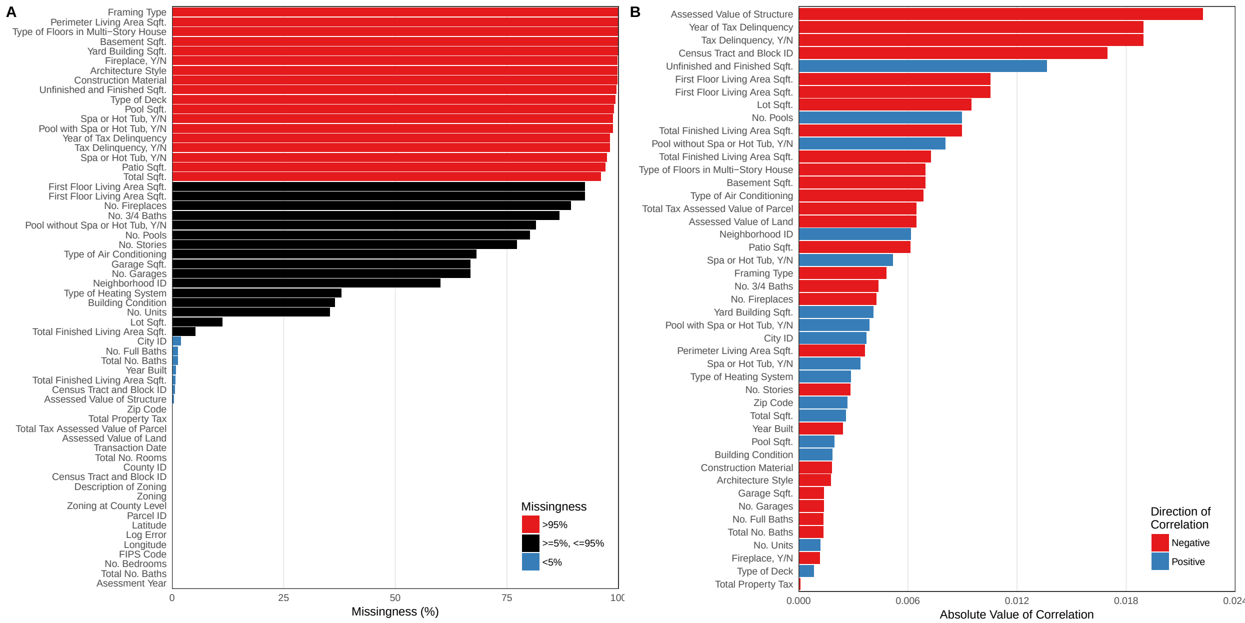

Figure 2. Missingness of Data: Missingness (degree of missing data) was calculated for each value by counting the number of rows missing data for a given value in the training data. Some variables share the same name because several values in the data are redundant. (A) The bars represent the percentage of observations missing a value. Values 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 value with the log error. Red bars represent positive correlation, blue bars represent negative correlation. Correlation cannot be calculated for values that have no missing data; these values are excluded in (B).

3 Results

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

4 Discussion

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

References

[1] Zillow: *What is a Zestimate? Zillow's Zestimate Accuracy*, https://www.zillow.com/zestimate/

[2] Tofallis, Chris. *A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation.* Journal of the Operational Research Society (2015) 66, 1352-1362