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

Elizabeth Vincent

October 18, 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]. Zestimates are calculated for all homes, not just those currently on the market, and are offer a means of assessing the current market. The Zestimate is useful for those who own a home and want to know the current value of their investment, and for tracking the market to determine the best times to buy or sell a property. For these reasons as well as others, it is desirable to have the most accurate Zestimate possible. According to Zillow, several factors clearly effect the accuracy of the Zestimate: the number of homes on the market and the amount of information about the property that is available are two of them [1]. To improve the accuracy of the Zestimate, it is necessary to know which parameters have the greatest effect on the accuracy and therefore which parameters should be optimized to best improve the accuracy of the Zestimate.

With the goal of improving the accuracy of the Zestimate, Zillow launched a competition on Kaggle, a cite that hosts data science competitions with an emphasis on machine learning methods. In the first round of the competition competitors must predict the accuracy of the Zestimate for a given property at a given time point. For this competition, the accuracy of the Zestimate is measured by the log error, 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 therefore a relative error metric, the log ratio error, is used to estimate the accuracy of the Zestimate, as opposed to an absolute error metric, such as root mean square error (RMSE), which would bias model evaluation for more expensive homes [2]. The second round of the competition involves improving the accuracy of the Zestimate itself, but will not be covered in this report.

Zillow has provided data for approximately 3 million properties in Los Angeles, Orange, and Ventura counties in California. The competitor must predict the log error of the Zestimate for each of the properties for three transaction periods: October, November, and December of 2017. The first round of the competition closed 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.

2 Methods

2.1 Data Cleaning

The data were obtained from the Zillow competition on Kaggle and consist of 58 variables for 2,985,217 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 variables are independent, nor is every variable available for every property. The degree of missing data (missingness), was calculated for each variable as well as for each property. A binary matrix of missingness of the same dimensions as the original data (roughly 3 million observations of 58 variables) was created for later correlation analysis (see section 2.2).

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.

2.2 Correlation Analysis

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.

2.3 Model Building

Analysis

- Predict the mean
- Linear model
- Random forest

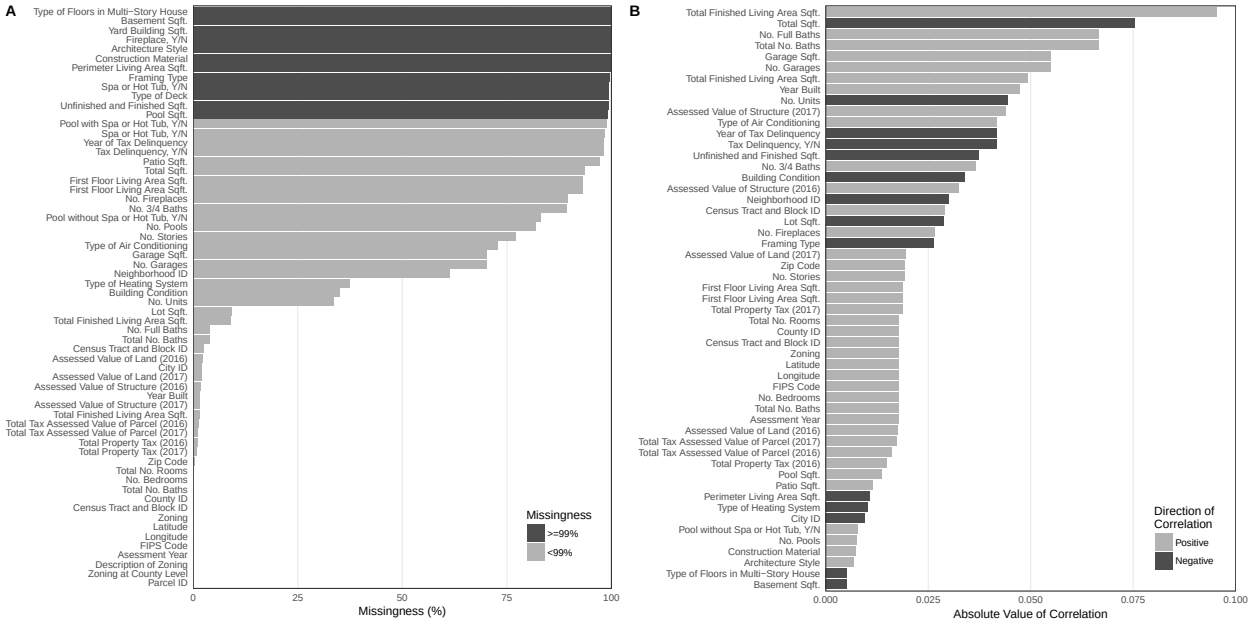

Figure 1. (A) Missingness of Data. The bars represent the percentage of observations for which there is no data for a given value. Values for which at least 99% of observations have no data are shown in dark gray, values for which less than 99% of observations have no data are shown in light gray. **(B) Correlation of Missing Data with the Magnitude of the Log Error**. The bars represent the absolute value of the correlation coefficient (Pearson's correlation, $p<0.05$) of missing data for a given value with the magnitude of the log error. Light gray represents positive correlation, dark gray represents negative correlation. Some variables share the same name due to redundancy in the data.

3 Results

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

Method	Model	Train MAE	Test MAE	R^2
Predict the mean	= mean(log error)	0.0694	0.0651	0
GLM	~x1 + x2 + x3	###	###	###

4 Discussion

The Zestimate is already a highly accurate algorithm with a median margin of error of 5% [3]. It is therefore unsurprising that the log error of the Zestimate does not correlate well with missingness in the data or with individual variables, as these correlations would be easy to detect and correct. As the correlation between the log error of the Zestimate and any one factor is weak, one would expect generalized linear models to perform only slightly better than predicting the mean, which is indeed the case.

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

[3] Kaggle: *Zillow Prize: Zillow's Home Value Prediction (Zestimate).*

https://www.kaggle.com/c/zillow-prize-1#description