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

Elizabeth Vincent

October 25, 2017

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 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. While Zillow firmly states that their Zestimate is exactly what it sounds like – an *estimate* – they recently faced a lawsuit over the accuracy and real-world implications of the Zestimate. In May 2017 a homeowner in Illinois sued Zillow claiming that the Zestimate under-valued the home and impeded the sale of the home at its true market value, and that because the Zestimate is "promoted as a tool for potential buyers to use in assessing [the] market value of a given property," it meets the legal definition of an appraisal under Illinois state law[2].

The lawsuit against Zillow was dismissed in August 2017[3], but not before Zillow launched a competition on Kaggle, a cite that hosts data science competitions with an emphasis on machine learning methods, to crowdsource an improved method for calculating the Zestimate. 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 towards more expensive homes[4]. 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 land parcels in Los Angeles, Orange, and Ventura counties in California. There are 58 features across all parcels, but not each parcel contains information on each feature. The competitor must predict the log error of the Zestimate for each of the parcels at the time of their sale. The accuracy of the predicted log error for the transaction 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. The data contain many missing values and many non-numeric features. Non-numeric features were converted to numeric where possible, and missing values were imputed as the median.

As the Zestimate formula is proprietary, the competitor starts with a blank slate to model the log error values. The models tested on these data were: predicting the mean, predicting the median, generalized linear models (GLM), and random forest models (RF). Predicting the mean and predicting the median were used as baseline models; any acceptable model must outperform these naïve models. The RF performed better than both the naïve models and the GLM, and including information on missing values improved the GLM but not the RF. According to Zillow, several factors clearly effect the accuracy of the Zestimate, two of which are the number of homes on the market in the area and the amount of information about the property that is available to them[1]. With this in mind, models were trained with and without information on missingness to determine if missing data was indeed an important factor in the accuracy of the Zestimate.

2 Methods

2.1 Data Cleaning

The data were obtained from the Zillow competition on Kaggle and consist of 58 features for 2,985,217 parcels that were split by Zillow into a training group and a test group, for which the true log error is provided. As the log error for the test data is not released, test error metrics can only be computed through submission to the Kaggle competition. Therefore, for the purpose of this report, only the 167,888 parcels for which the log error is available were used in the analysis. Not all 58 features are independent, nor is every feature available for every property. Prior to any imputation or removal of missing values, the degree of missing data (missingness) was calculated for each feature as well as for each property. The result is a binary matrix of missingness (see code section 2.1), where each cell in the missingness matrix corresponds to a single cell in the original data. The value in each cell of the binary matrix is 1 if the cell is empty in the original dataset, and 0 otherwise.

As many models are incapable of dealing with missing or non-numeric features, the data required extensive cleaning prior to any analysis. Several features only contained one entry type, such as "yes" or "true", and the remaining entries were empty. In these cases, it was assumed that an empty entry was a negative response. The "yes" or "true" entries were converted to 1 and any empty entries were converted to 0, creating a binary, numeric feature. This conversion was performed for the following features: whether or not a parcel was flagged for tax delinquency, if a parcel has a fireplace, if the parcel has a pool, and if the parcel has a hot tub or spa. In a similar manner, other features that contained three or fewer possible entries were one-hot-encoded: a binary feature was made for each possible entry. This conversion was performed for the type of pool and county ID. See code section 2.2 for one-hot-encoding and imputation methods.

After minimal imputation and one-hot-encoding, the missingness of each feature was recalculated (see code section 2.3) as described previously in this section. For model building, features that were missing in more than 95% of cases after imputation were removed, as well as features that were redundant (see code section 2.5). 17 features were removed due to redundancy or a high percentage of missing data.

2.2 Correlation Analysis

The binary missingness matrix described in section 2.1 was used to calculate Pierson's correlation coefficients between the missingness of each feature and the log error. Additionally, Pearson's correlation coefficients were calculated for the correlation of numeric variables with the log error. P values were calculated for the correlation coefficients[6].

2.3 Model Building

The first two models, predicting the mean and predicting the median, were used as a benchmark. To get accurate MAE values, the MAE was averaged over 10-fold cross-validation. A random sample of 80% of the

data for the test set, and the other 20% as the test set. The mean was calculated from the train subset and the test MAE was calculated by predicting the mean of the log error in the training subset mean as the log error for the test set. To cross-validate, this was repeated with nine new random splits of the data.

The models were each trained on two separate datasets. The first contained only the numeric features with missingness percentages below 5. These features are: lot sqft., total finished living area sqft., number of baths, year built, number of bedrooms, total number of rooms, latitude coordinates, and longitude coordinates. The second dataset contained all the features in the first data set, with the addition of the binary matrix of missingness (see report section 2.1). Prior to model building, any missing values in the data to be used for the models were imputed as the median for their given feature across all observations. After median imputation, the data were centered and scaled. By definition, there is no missing data in the missingness matrix, and therefore nothing was imputed in the missingness matrix. The missingness matrix was neither centered nor scaled.

All models were trained using 10-fold cross-validation. Only features that are numeric and which are missing values for less than 5% of properties were used for training the models. GLM[7] were trained[8] with a variety of values for the tuning parameters alpha and lambda to optimize the MAE. The alpha tuning parameter is a numeric value between [0,1] that represents the mixing of lasso (0) versus ridge (1) GLM. Lasso models penalize the total number of covariates with nonzero beta values, whereas ridge models penalize the magnitude of the beta values. Lambda is any numeric, and represents the degree that a model should be penalized based on the alpha.[7]

Similarly, RF[9] were trained[8] with a variety of values for the mtry parameter to optimize the MAE. Mtry is the number of random parameters to select on for each branch in the tree. For all GLM and RF, the expected MAE was taken as the reported MAE from the model, while the training MAE was calculated by predicting values on the whole dataset using the model and calculated the MAE of those predictions.

3 Results

Data show a high degree of missingness, but this missingness is not well correlated with the log error. All the correlations values shown in Figure 1B are significant (Pearson's correlation, $p<0.05$), but the magnitude of the correlations are all below 0.1 (see Figure 1). Regardless, the binary matrix of missingness was included in the models to evaluate the effect of missing data on the accuracy of the Zestimate, given Zillow claims the availability of data does in fact affect it[4]. The correlation between numeric features in the original data and the log error were also calculated, but again, for significant correlations ($p<0.05$), the magnitude of the correlation does not exceed 0.12 (see code section 2.4).

The median is immune to the skewing effects of outliers, unlike the mean. It is therefore unsurprising that predicting the median performed better than predicting the mean (see Table 1). The data are not normally distributed, 91.6% of observations fall within *one* standard deviation of the mean (see code section 2). Perhaps it is not surprising then that none of the models performed better than either of the naïve models, as determined by the expected MAE on test data, because predicting the mean will be within one standard deviation of the true value in over 90% of the data.

For both the GLM and the RF, models were fit on numeric data from the original dataset, as well as on a dataset that included information on missing values. While the GLM performed similarly on the training and test data. The model that does not include information on missing data resulted in MAE of 0.0625 for the training data and 0.0626 for the test data, and the model that does include information on missing data resulted in MAE of 0.06918 for the training data and 0.06919 for the test data (see Table 1). For the GLM that does not include information on missing data, one scaled and centered unit of total finished living area results in an increase in log error of 4.60×10^{-3}, one more scaled and centered dollar of property tax results in a decrease in log error of 1.64×10^{-3}, and for every scaled, centered, added bedroom the log error increases by 6.71×10^{-4}. The same reasoning extends to the beta and covariate values for the GLM that does include information on missing data.

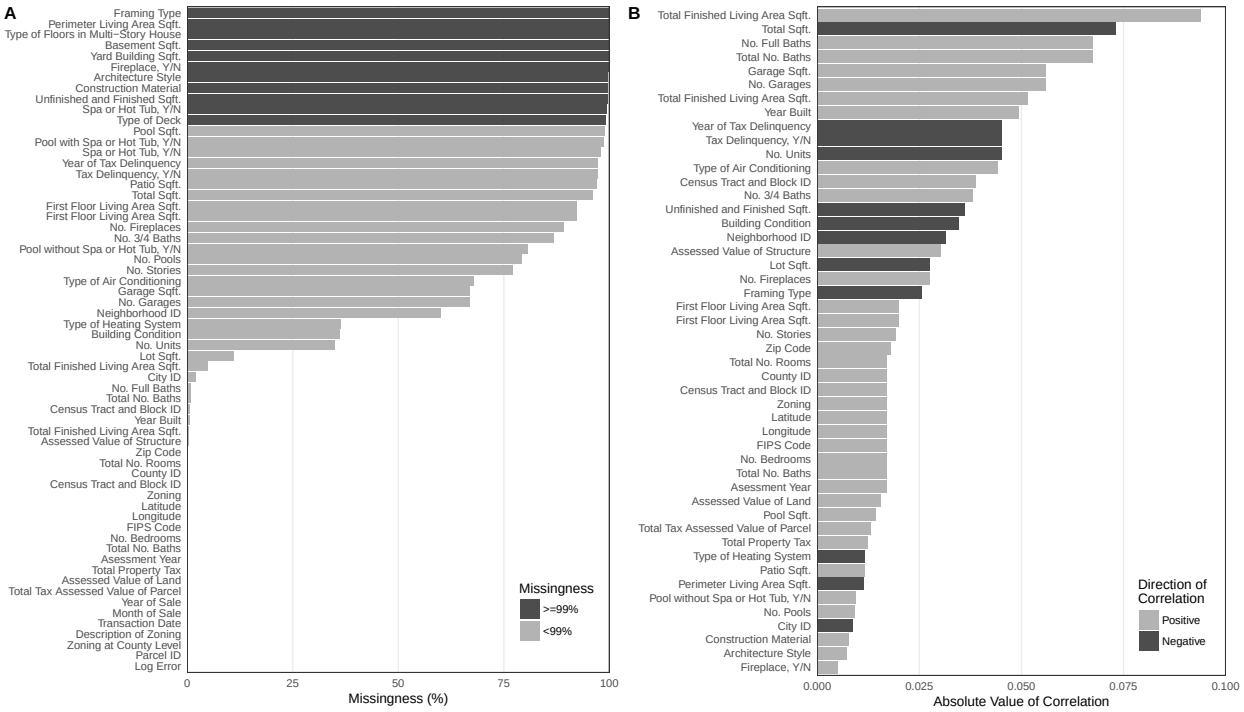

Figure 1. (A) Missingness of Data. The bars represent the percentage of observations for which there is no data for a given value. Values for which at least 99% of observations have no data are shown in dark gray, values for which less than 99% of observations have no data are shown in light gray. **(B) 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. Note the range of the x-axis is 0 to 0.100. Some features appear more than once due to redundancy in the data.

RF models had a much greater difference in performance. The model that did not include information on missing data resulted in MAE of 0.05601 for the training data and 0.07067 for the test data, and the model that did include information on missing data resulted in MAE of 0.06791 for the training data and 0.06923 for the test data (see Table 1). In both cases, the RF showed the worst expected performance on test data of the three model types. This indicates that the models were likely overfit to the test data.

4 Discussion

The Zestimate is already a highly accurate algorithm with a median margin of error of 5%[5]. It is therefore unsurprising that the log error of the Zestimate does not correlate well with missingness in the data or with individual numeric features, 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 not expect GLM to substantially outperform predicting the mean, as is the case. The RF performed significantly better on the training data than both the naïve models and the GLM, but performed worse than the naïve models on the test data. This is a sure sign of over-fitting the data, a common issue with random forests. In addition to overfitting, random forests and other machine learning methods are substantially less interpretable than GLM. GLM return coefficients that directly relate the input variables to the output variable, but RF do not.

The RF that included information on missingness performed worse on the train data but better on the test data than the RF that did not include this information. Additionally, the GLM that included information on missingness performed better on both the training and the test data than the GLM that did not. While this alone would support Zillow's assertion that the availability of data affects the accuracy of the Zestimate, none of the GLM or RF models perform better on the test data than either of the naïve models. It is probable that missing data for some features is informative while others create noise, and that removing the missing data features that introduce noise may boost the performance of the models above that of the naïve

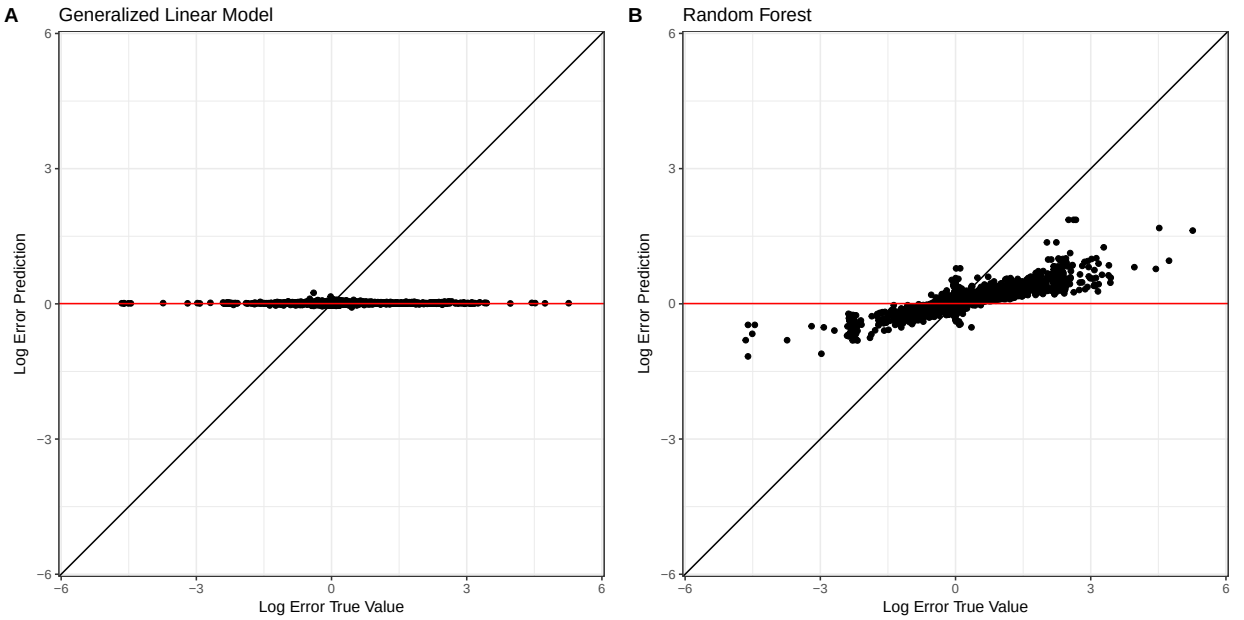

Figure 2. (A) Generalized Linear Model Predictions. The values of the log error predicted by the generalized linear model that includes information on missing data are plotted against the true log error values. The black line represents the line of unity, x=y. The red line shows the median value of the true log error for all properties. Mean absolute error = 0.06918. **(B) Random Forest Model Predictions**. The values of the log error predicted by the random forest model that does not include information on missing data are plotted against the true log error values. The black line represents the line of unity, x=y. The red line shows the median value of the true log error for all properties. Mean absolute error = 0.05601

models, the GLM training algorithm[7][8] should have been able to parse these out. It is therefore unclear if the availability of data is truly influencing the accuracy of the Zestimate.

Overall, none of these models performed better than the naïve models, but the performance of models that included information on missing data does indicate that, even though missingness of data does not correlate with the log error, it may still be influencing the accuracy of the Zestimate in less obvious ways. For that, Zillow will have to rely on the second round of the competition, in which competitors attempt to reengineer the Zestimate algorithm.

References

[1] What is a Zestimate? Zillow's Zestimate Accuracy. *Zillow*. https://www.zillow.com/zestimate/ Accessed 10-20-2017.

[2] Harney, Kenneth R. Zillow faces lawsuit over 'Zestimate' tool that calculates a house's worth. *The Washington Post*, published online 05-10-2017. https://www.washingtonpost.com/realestate/zillow-faces-lawsuit-over-zestimate-tool-that-calculates-a-houses-worth/2017/05/09/b22d0318-3410-11e7-b4ee-434b6d506b37_story.html?utm_term=.e64be2e01192. Accessed 10-20-2017.

[3] MarksJarvis, Gail. Judge dismisses lawsuit that challenged Zillow's home price estimates. *Chicago Tribune*, published online 08-24-2017. http://www.chicagotribune.com/business/ct-judge-dismisses-zillow-zestimate-case-0825-biz-20170824-story.html. Accessed 10-20-2017.

[4] 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.

[5] Zillow Prize: Zillow's Home Value Prediction (Zestimate). *Kaggle*. https://www.kaggle.com/c/zillow-prize-1#description Accessed 10-20-2017.

[6] Alboukadel Kassambara (2016). ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package version 0.1.1. https://CRAN.R-project.org/package=ggcorrplot

[7] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22. http://www.jstatsoft.org/v33/i01/.

[8] Kuhn, Max *et al* (2017). caret: Classification and Regression Training. R package version 6.0-77. https://CRAN.R-project.org/package=caret

[9] Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17. doi:10.18637/jss.v077.i01

Method	Beta	Covariate	Tuning Parameters	Train MAE	Expected MAE
Predict the mean	1	mean	NA	0.06944	0.06900
Predict the median	1	median	NA	0.06886	0.06840
GLM	4.60×10^{-3}	total finished living area sqft.	alpha = 0.03 lambda = 0.04	0.06925	0.06926
	-1.64×10^{-3}	total property tax			
	6.71×10^{-4}	no. of bedrooms			
	1.39×10^{-2}	intercept			
GLM: missing value information included	-2.68×10^{-2}	assessed value of structure (missing)	alpha = 0.275 lambda = 0.005	0.06918	0.06919
	-1.48×10^{-2}	total no. baths (missing)			
	-1.04×10^{-2}	tax delinquency (y/n) (missing)			
	6.68×10^{-3}	total finished living area sqft.			
	-3.30×10^{-3}	assessed value of land (missing)			
	-2.87×10^{-3}	no. bedrooms (missing)			
	2.87×10^{-3}	no. pools (missing)			
	-2.62×10^{-3}	total property tax			
	5.99×10^{-4}	no. full baths (missing)			
	-5.62×10^{-4}	latitude (missing)			
	-3.81×10^{-4}	FIPS code (missing)			
	-3.53×10^{-5}	longitude (missing)			
	-5.44×10^{-6}	zoning (missing)			
	-6.75×10^{-7}	census tract and block ID (missing)			
	-7.98×10^{-8}	county ID (missing)			
	-8.14×10^{-9}	total no. rooms (missing)			
RF	NA	NA	mtry = 3 split rule = extra trees	0.05601	0.07067
RF: missing value information included	NA	NA	mtry = 5 split rule = extra trees	0.06791	0.06923

Table 1. Evaluation of predictive models. Comparison of the beta values, covariates, mean absolute error (MAE) on the training set, and expected MAE on the test set for each model. The method indicates the type of model fit: predicting the mean of the log error, predicting the median, generalized linear models (GLM) on two different data sets, and random forest models (RF) on two different data sets. The first GLM and RF include only numeric features with less than 5% of values missing, while the second GLM and RF include information on missing values. Beta and covariates are not provided in RF values, predicting the mean and predicting the median do not have any tuning parameters.