Analysis of Zillow Data : A Spatial Random Forest Approach

Arkajyoti Saha

Abstract

In the manuscript I analyse the property data available through the Zillow competition in order to predict the log error of the House value estimate by Zillow. A suitable modelling of the log error of the "Zestimate" can improve the house value estimate of 7.5 Million properties across all of USA. Moreover, the spatial nature of the data, along with absence of any correlation between the covariates and response, makes it a challenging statistical problem independent of the competition objective. I take a random forest based approach to model the effect of predictors on the response, along with a nearest neighbor Gaussian process based spatial correction, to model the log-error of "Zestimate".

Keywords: Missing data; Imputation; random forest; Spatial data; Gaussian process; nearest neighbors; Bayesian inference.

1 Introduction

Zillow's Zestimate home valuation has been a noteworthy name in the U.S. real estate industry since it's arrival 11 years ago. The "Zestimate" provides consumers as much information as possible about homes and the housing market, based on 7.5 million statistical and machine learning models that analyze a number of data points on each property. "Zillow Prize" is an attempt to further improve upon the ever evolving error margin (from 14% at the onset to 5% today) through a competition. The development of methodology for prediction of log error associated with the home value estimate by Zillow is an interesting challenge from the perspective of a statistician, as analysis of such a huge dataset with high percentage of missing data along with presence of Spatial component, presents some interesting challenges in terms of methodology development and application. A better prediction of the error associated with the Zillow's estimate will in turn lead to a better model building mechanism for home value estimation which stand to impact the home values of 110M homes across the U.S.

The goal of the article is to find a statistical modelling of the log error rate of the home value estimate based on non parametric (justification provided afterwards) dependancy on the covariates along with spatial correction. The rest of the manuscript is described as follows: In section 2, the data under consideration along with exploratory data analysis is

discussed. Section 3 discusses in detail the methodology developed for the data analysis; section 4, discusses the obtained results. In section 5, conclusion is drawn with some limitations of the project and pointers toward future work.

2 Data

The data under consideration is available in the official competition website.

2.1 Description

The predictors of the dataset are the home features, where as the response is the log-error between "Zestimate" and the actual sale price. We are provided with the full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) in 2016 and 2017. It contains 58 variables, which includes number of bedrooms, bathrooms, total area, location of the property etc. The difference between the log values of the actual house price and the "Zestimate" corresponding to the sold properties along with the transaction date are also provided to us.

2.2 Objective and Evaluation

In the perspective of the competition, the goal is to predict the log-error at 6 time points for all properties: October 2016, November 2016, December 2016, October 2017, November 2017, and December 2017. As the evaluation depends on property sells that are to be made in coming months and cannot be performed till January, 2018 (that too only through a submission to Zillow competition, corresponding to publicly unavailable data), we propose a model of our own to evaluate the developed method. We begin by splitting the data corresponding to properties with available sales data in training and test dataset (in 2:1 ratio). We train our model with the training dataset and try to predict the log-error corresponding to the properties in the test dataset. The accuracy of the prediction performs the model evaluation. The Mean Squared Error is used as the measure of the accuracy of the prediction.

2.3 Exploratory Data Analysis

In the exploratory data analysis, we begin with exploration of the interesting features of the data under consideration. One of the biggest feature of the data under consideration is the missing pattern of the data. We plot the percentage of the missing data corresponding to each of the variables (Fig. 1). For the project under consideration, we decide to keep ourselves restricted to the variables, with less than 40% missing data. The missing ness pattern present in the data is not known a priori (hence safe to assume, they are missing not at random (Rubin, 1976)) and imputation of majority of the data may lead to presence of false signals in the data.

Next we try to explore the components of the response. We begin by exploring linear relationship of the responses with the features under consideration. The correlation plot

(Fig. 2) clearly demonstrates that the correlation of response with all the variables under consideration is negligible.

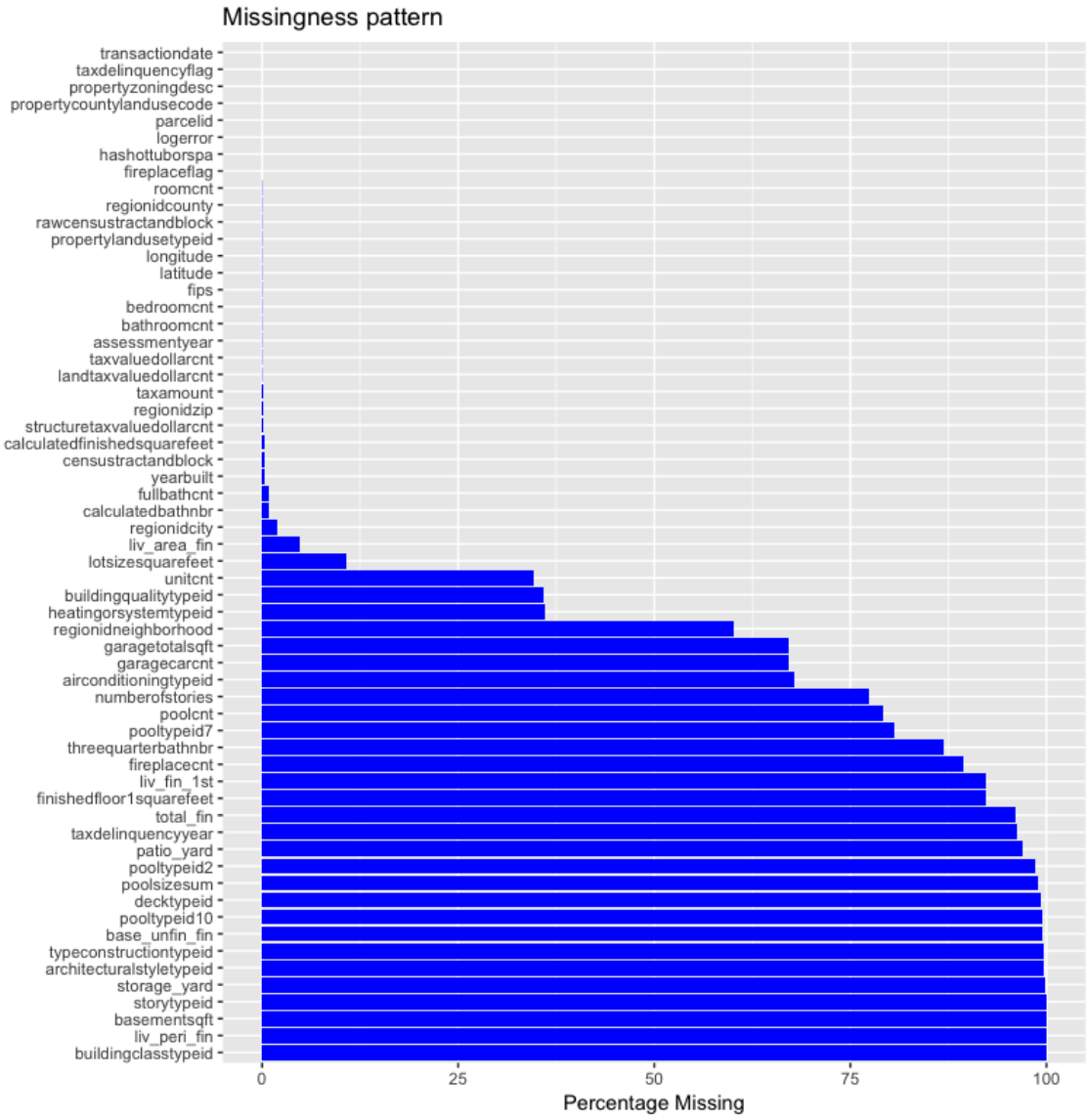

Figure 1: Missing Percentage corresponding to each of the variables

3 Model

The exploratory data analysis suggests that the response can be broken down into two separate components:

- The covariate based component : There is no linear relationship, so we opt for machine learning based nonparametric Random Forest (Breiman, 2001; Liaw et al., 2002)

approach. Random forest combines tree predictors in a way, such that each tree depends on an independent and identically distributed (across trees in the forest) sample/subsample (with/without replacement) of the original data. Using a random selection of features to split each node yields error rates comparable to that of Adaboost (Freund and Schapire, 1995) with additional robustness to noise. Additionally, Internal estimates can also be used to measure variable importance.

- Spatial Component : As the price of a house depends on the location of the house, the error associated with the estimation of the price of the house is also expected to be location dependant.

Let the set of locations corresponding to the n points in the transformed training data be denoted by $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n)$. Let $\mathbf{x}_i$, y_i be the feature and the response (respectively) corresponding to the i^{th} data point, then the model under consideration can be summarized as follows:

$$y_i = f(\mathbf{x}_i) + w(\mathbf{s}_i) + \epsilon_i \tag{1}$$

where, $f(\mathbf{x}_i)$, $w(\mathbf{s}_i)$ and ϵ_i denote the covariate based component, Spatial component and an error component respectively (ϵ_i are i.i.d. $N(0, \tau^2)$). A popular approach to model $w(\mathbf{s})$ is via Gaussian Processes (Rasmussen and Williams, 2005). The customary specification (with spatial parameters $\boldsymbol{\theta}$) $w(\mathbf{s}) \sim GP(\mu(\mathbf{s}), C(\cdot, \cdot|\boldsymbol{\theta}))$ using mean function $\mu(\mathbf{s})$ and covariance function $C(\cdot, \cdot|\boldsymbol{\theta})$ endows any finite collection $\mathbf{w} = (w(\mathbf{s}_1), w(\mathbf{s}_2), \ldots, w(\mathbf{s}_n))'$ with a multivariate Gaussian distribution with mean $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \mu(\mathbf{s}_2), \ldots, \mu(\mathbf{s}_n))'$ and covariance matrix $\mathbf{C}(\boldsymbol{\theta}) = (C(\mathbf{s}_i, \mathbf{s}_j|\boldsymbol{\theta}))$. The mean function $\mu(\mathbf{s})$ is usually chosen to be zero and the regression component is already modeling the mean, while a popular choice for $C(\cdot, \cdot|\boldsymbol{\theta})$ is the Matérn Covariance function, specified as:

$$C(\mathbf{s}_i, \mathbf{s}_j|\boldsymbol{\theta} = \left(\sigma^2, \phi, \nu\right)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\|\mathbf{s}_i - \mathbf{s}_j\|\phi\right)^{\nu} \mathcal{K}_{\nu} \left(\|\mathbf{s}_i - \mathbf{s}_j\|\phi\right); \phi > 0, \nu > 0, \tag{2}$$

where σ^2 controls the variance of the spatial component, ϕ denotes the decay in spatial correlation, ν controls the process smoothness and $\mathcal{K}_{\nu}$ denotes the Bessel function of second kind with order ν. If $\mathbf{y}$ denote the vector of observations and $\mathbf{X}$ is corresponding covariate matrix, then marginalizing out $\mathbf{w}$, the model for the observed data is given by $\mathbf{y} \sim N(f(\mathbf{X}), \mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I})$.

4 Data Analysis

As we hardly have any idea regarding the functional form of the relationship between the variables (the highly correlated variables are often missing simultaneously, another indicator of the fact that they are probably not missing at random), we use the "randomForest" package (Liaw and Wiener, 2002) for imputation purpose. This performs random forest based imputation of the data, based on the nature of the variable (i.e numeric or categorical), which provides it with an edge over the other model based approaches.

The algorithm starts by filling up the missing values by median/mode (corresponding to nature of the variable, numeric/factor). Then random forest is implemented with the completed data. The proximity matrix from the random forest is used to update the imputation of the missing values. For continuous predictors, the imputed value is the weighted average of the non-missing observation, with proximities as weights. Whereas, for categorical predictors, the imputed value is determined to be the category with the largest average proximity.

We begin by fitting a random forest model on the training data with the selected house features (as random forest is unable to handle variables with more than 53 features, we remove two features from the data. Additionally instead of the actual date of transaction, we are more interested in the monthly contribution in the log-error rate, hence we extract the month from the date and use it as a categorical variable while modelling the random forest and we get rid of the locations as they will be utilised while modelling the spatial component) with the "ranger' package (felicitates a fast implementation of random forests) (Wright and Ziegler, 2017). Let the random forest based estimation corresponding to the i^{th} data point be denoted by $\hat{y}_i$. The corresponding residual r_i, given by $y_i - \hat{y}_i$, where y_i is the original observation, captures the spatial and the nugget component of the model.

As far as the spatial components are concerned, analysis of geostatistical data through spatial process model necessitates computation and storage that becomes prohibitive with the increment in number of spatial locations. The massive size of the data under consideration precludes the conventional Gaussian Process fitting, hence we opt for Bayesian NNGP (Datta et al., 2016). By inducing sparsity in precision matrix, this algorithm drastically reduces the floating point operations (flops) (per iteration of the Markov Chain Monte Carlo (Gilks et al., 1995)) to linear in the number of spatial locations compared to cubic in conventional Gaussian process models. We implemented the NNGP model with the "spNNGP" package (Finley et al., 2017). We started with $\nu = 0.5$ (fixed, hence essentially reducing the covariance model to an exponential model) the initial value of $\sigma^2 = 5, \tau^2 = 1$ and $\phi = 6$, with $\texttt{uniform}(3, 300)$ prior on ϕ and inverse-Gamma prior on σ^2 and τ^2 with shape parameters fixed equal to 2 for both of them and scale parameters to be 5 and 1 respectively.The number of nearest neighbors (Datta et al., 2016) (m) was fixed to be 10. We ran the MCMC chain for $15,000$ iterations and drew inference on last $10,000$ samples with a burn in period of $5,000$.

We observe that the MSE of the error obtained from fitting the random forest and the spatial model in out of sample (test) data is **0.0312**. The medians of the post burn in period was reported as estimates. The estimates of σ^2, τ^2 and ϕ was respectively 0.01, 0.02 and 169.76 respectively. The 95% confidence interval corresponding to ϕ was given by $(14.139, 239.6)$, where as σ^2 and τ^2 had numerically degenerate distribution. This demonstrates that after accounting for the contribution by covariate through random forest, there is hardly any spatial trend left in the data. This is also indication of a great existing modelling scheme by the Zillow.

5 Conclusion

In this article we propose a random forest based spatial modelling in the perspective of the ZIllow data. Due to space and time constraint, there are a few drawbacks of the short

manuscript. A few noteworthy of them as follows:

- Exclusion of covariates with more than 40% missing data.
- Exclusion of variables with more than 53 categories due to the package based implementation of random forest.
- Absence of comparative discussion on performance of nonparametric approaches other than random forest (e.g. Neural Net, SuperLearner etc.)

As far as future progress is concerned, study of the effect of incorporation of different starting points and priors in the spatial model may lead to some interesting discoveries. I also plan to extend this approach for multivariate spatial data, which is also widely encountered in many fields but unfortunately lacks efficient implementation that can enable inference in popular statistical platforms. Finally, as theoretical properties of the spatial random forest remains elusive, I will seek to establish some theoretical justification of the algorithm in the future.

References

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* **111,** 800–812.

Finley, A., Datta, A., and Banerjee, S. (2017). *spNNGP: Spatial Regression Models for Large Datasets using Nearest Neighbor Gaussian Processes*. R package version 0.1.1.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News* **2,** 18–22.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news* **2,** 18–22.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, first edition.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77,** 1–17.

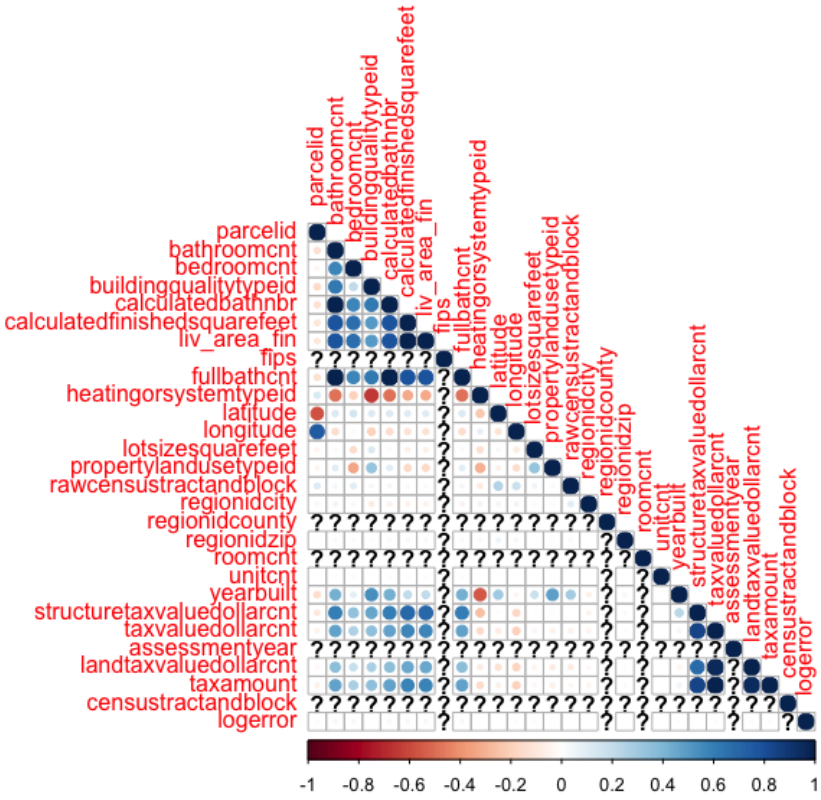

Figure 2: Correlation with the response ("logerror", last row of the triangle); no correlation proves absence of linear relationship