Analysis of Zillow Data

Arkajyoti Saha

10/11/2017

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. This million-dollar competition consists of two stages:

- The qualifying (public) round, where the challenge is to model the residual error of the "Zestimate", based on the data already provided in the competition website (no use of external data is permitted).

- The final round (only open to top 100 teams from the first round), where the challenge is to build home valuation algorithm from scratch (no restriction on inclusion of external data).

For the project under consideration, we will only be concerned with the data and method corresponding to the first round of the competition.

2 Data

The data for the first round is posted 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 (**properties_2016.csv** and **properties_2017.csv** respectively). The train data includes: all transactions before October 15, 2016, plus some of the transactions after October 15, 2016 (**train_2016.csv**) and all the transactions from 1/1/2017 to 9/15/2017 (**train_2017.csv**).

2.2 Objective and Evaluation

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.

In order to judge the accuracy of the prediction, there is a public leaderboard test data, which has the transactions between October 15 and December 31, 2016. The rest of the test data, which will used for calculating the private leaderboard, is the response corresponding to all the sold (unsold ones will be ignored, in case of multiple sells within a month, the first reasonable response will be taken as the true response) properties in October 15, 2017, to December 15, 2017.

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 **properties_2016.csv** is the missing data. We plot the percentage of the missing data corresponding to each of

the variables (Fig. 1a). For the project under consideration, we decide to keep ourselves restricted to the variables, with less than 50% missing data. Model based data imputation is not very effective in Not missing at random scenario, 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. 1b) clearly demonstrates that the correlation of response with all the variables under consideration is negligible.

Traditionally house prices are dependant on the location, hence it is expected that the data will have a spatial component. Additionally, as we are supposed to predict the response in 6 different months, we also explore if monthly trends are present in the data, so that a monthly component can be incorporated in the model under consideration. From (Fig 1.c) it seems that there are some monthly trends in the data.

4 Model

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

- The covariate based component : There is no linear relationship, so we opt for machine learning based "SuperLearner" approach. SuperLearner as an algorithm, uses cross-validation to estimate the performance of multiple machine learning models (in some cases, same model with different parameter setups), and then creates an optimal weighted average of those models, based on the test data performance. This approach has been proven to of the same accuracy (asymptotically) of the best possible prediction algorithm that is tested.

- 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.

- Monthly Component : In the exploratory data analysis, we have seen that the response has a clear trend over the months, hence we are going to make that adjustment in the prediction by incorporating a monthly component into the model.

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 $\mathbf{x}_i$, t_i, y_i be the feature, the time of transaction (only in months) and the response (respectively) corresponding to the i^{th} response, then the model under consideration can be summarized as follows:

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

where, $f(\mathbf{x}_i)$, $w(\mathbf{s}_i)$, $u(t_i)$ and ϵ_i denote the covariate based component, Spatial component, Monthly component and an error component respectively (ϵ_i are i.i.d. $N(0, \tau^2)$).

5 Prediction

The prediction process consists of four stages, each of which are described in detail as follows:

5.1 Imputation

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 "missForest" package 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.

5.2 Monthly Component

In order to account for the monthly component we properly location shift mean response corresponding to each of the months to centre them around 0. Next we subtract the monthly contribution from the responses in order to empirically adjust for the contribution for the month (as we are not provided with the monthly trend corresponding to the months where prediction is to be made, monthly adjustment can only be done in the model building phase, not in the prediction phase, hence we centre the monthly trend to avoid any directional bias in the transformed data),

5.3 Covariate based Component

As mentioned earlier we use the package "SuperLearner" for handling the component $f(\mathbf{x}_i)$ in (1). The algorithms that were included in the SuperLearner library were "gam", "randomForest", "gbm", "nnet", "bart", and "polymars". As we have explored through the EDA that the response is hardly correlated with any of the predictors, we do not include any linear model in the SuperLearner library.

5.4 Spatial Component

Let the estimates from the Covariate based components corresponding to the i^{th} datapoint be denoted by $\hat{y}_i$. Next we fit a Bayesian conjugate spatial regression model with the Matérn covariance function (package "spNNGP") on the residual vector $\mathbf{y} - \hat{\mathbf{y}}$. The shape and scale parameter for the inverse-Gamma prior on σ^2 in the model is selected through a grid search and k.fold cross validation is performed to obtain the optimized predictions. By incorporating the nuggest in this model, we also take care of the i.i.d error component in (1).

N.B. The codes are available in the following link, I will be uploading the remaining codes by the deadline

https://www.kaggle.com/arkajyotisaha/analysis-of-zillow-data