Zillow's Home Value Prediction: Midway Report

Harish Pullagurla
Department of ECE
North Carolina State University
hpullag@ncsu.edu

Pooja Mehta
Department of ECE
North Carolina State University
pmehta@ncsu.edu

Ratika Kapoor
Department of Mathematics
North Carolina State University
rkapoor3@ncsu.edu

1 Introduction

A house is the largest and perhaps the most expensive purchase a person makes in her or her lifetime. The goal of this project is to provide an estimated prediction of the value of a house for future sale.

The accurate prediction of a house price is key to prospective homeowners, developers, investors and other real estate participants such as mortgage lenders and insurers. Local government use market values to set real estate taxes.

A property's appraised value also plays an important role in several real estate transactions such as sales, loans and its marketability. Conventional method for determining property prices involved professional appraisers. This method had the disadvantage of the appraiser having a vested interest from the seller, buyer, mortgage broker or lender leading to a biased estimate of the house price. This led to the creation of automated house prediction systems, independent of the bias and providing a more accurate and less biased estimate of the house price.

Zestimate is one such house price prediction model. It was created by Zillow to give consumers as much information as possible about homes and the housing market.It is computed using a proprietary formula and is a starting point in determining a home's value. It is based on public and user-submitted data and also takes into account special features, location, and market conditions.

To further improve the accuracy of prediction, Zillow introduced an open competition on Kaggle (1). In this competition, Zillow provided full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) from 2016 and 2017. The goal of this competition was to minimize the log error between their Zestimate and the actual sale price. Error modeling is considered to be a powerful method to find areas to improve in a good existing model like Zestimate.

This report is divided into the following sections. Section 2 provides the background on house price prediction and several data mining techniques that were used in this field. Section 3 provides information of the implementation strategy that was followed as part of this project. Section 4 talks about Experiments and Results, and Section 5 provides the conclusion based on the observations made.

2 Background

With a large amount of unstructured resources available, the Real Estate industry has become a highly competitive business. The data mining process in such an industry provides an advantage to the developers by processing this data, forecasting future selling prices and thus, helping people make knowledge-driven decisions.

In the area of house price prediction, many attempts were made using different methods.Park et.al (2) conducted an analysis of 5359 townhouses in FairFax County, VA and used machine learning to develop house prediction models. They concluded that RIPPER outperformed other house prediction models such as C4.5, Bayesian and AdaBoost in all the tests. In their conclusion, they underlined the fact that housing markets can be influenced by macro-economic variables and future research should

Submitted as a Project Midway Report for CSC522 Automated Learning and Data Analysis.

consider these variables along with environmental amenities variables as input to house prediction models.

Wang et.al (3) performed real estate forecasting based on SVM optimized by PSO. Support Vector Machine (SVM) is a type of learning machine that has helped in solving the problems of limited sample learning, nonlinear regression and in overcoming the "curse of dimensionality". To determine the parameters of SVM, which determine its learning and generalization; particle swarm optimization (PSO) was chosen. Their experimental results showed that PSO-SVM had a high forecasting accuracy compared to other models.

3 Method

For this competition, Zillow provided all the properties with their home features for 2016 and 2017 from 3 counties. Each home feature had a 58 dimensional vector talking about features like year built, number of rooms, location, tax paid etc. Training set was provided with a list of 160,000 values, each having the log error and the date on which the transaction took place. Log errors corresponding to 6 time points for each of the property are to be predicted at the end.

$$logerror = log(Zestimate) - log(SalePrice)$$

Steps towards building a regression model to improve Zestimate residual error are provided in the following subsections.

3.1 Data Exploration

As a first step, exploratory data analysis was performed. This helped in understanding the given data and its distribution. In addition, the importance of different features was analyzed before prepossessing the data.

Training set was sampled, by parcel id, from all the properties given. It was important to understand if the training set followed a similar trend as the complete properties data provided. Figure : 1 shows the density vs feature value plot for few features. It was observed that both of them follow a similar trend. Hence, deductions made on training data would be valid at other data.

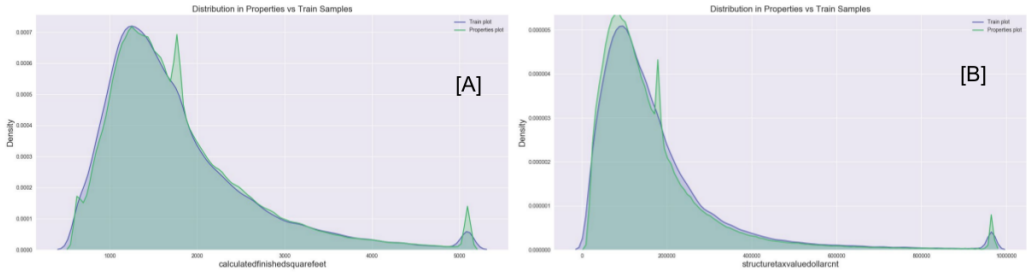

Figure 1: Comparison of Distribution between Training and Properties data for 2 different features (A,B)

Figure :2 A, displays the trends in log error values on the training data set. It was seen to follow a normal distribution centered at zero, implying that Zestimate was able to accurately predict most of the values. Also Figure :2 B, validated the information that the training set consisted of transactions that occurred before October 15, 2016 plus some transactions that took place after October 15, 2016 as said.

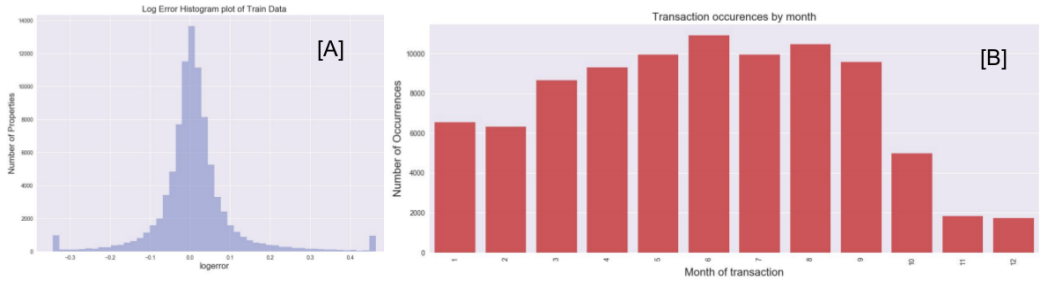

Figure 2: A. Histogram plot of log error values from training data , B. Month wise transaction plot of training data

Random Forests is one of the most common methods used to find the importance of different features. The tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This means a decrease in impurity over all trees (called gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, a subset of the most important features can be created. Figure : 3 shows the relative importance of each of the 58 dimensional vectors provided, which acted as a initial queue for selection of parameters for the machine learning model towards final log error prediction.

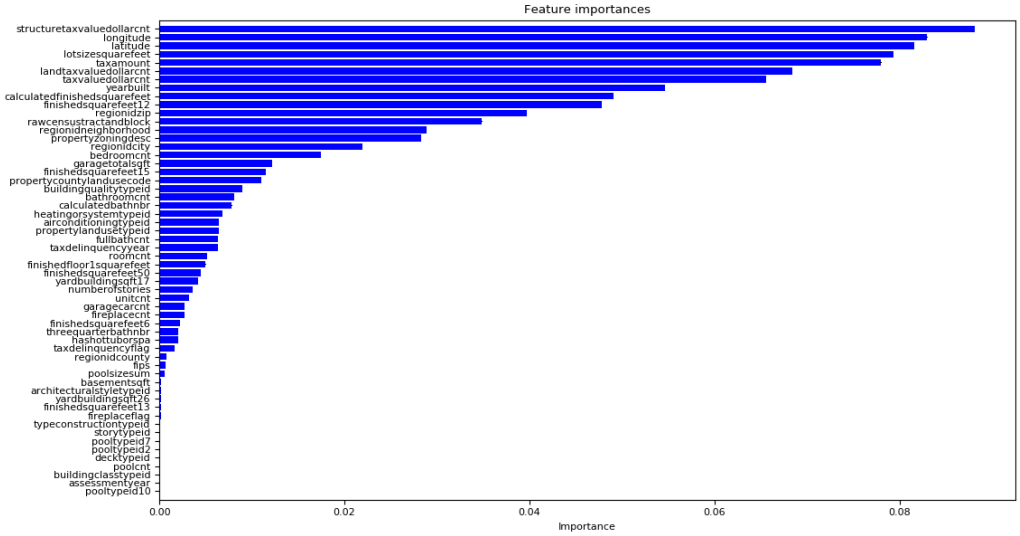

Figure 3: Feature Importance Scores using Random Forests

3.2 Data Pre-processing

Each house featured in the properties had a 58 dimensional feature vector associated with it. A large proportions of these 58 dimensional vectors were missing. Figure : 4 shows the percentages of missing values for each of the features. To impute extreme missing values would indirectly mean creating artificial data, hence imputation for these extreme cases was not considered to be appropriate. Thus, properties with missing values greater than 90 percent were dropped from further processing. Out of the initial 58 features provided, 20 of them had extreme missing values in this data.

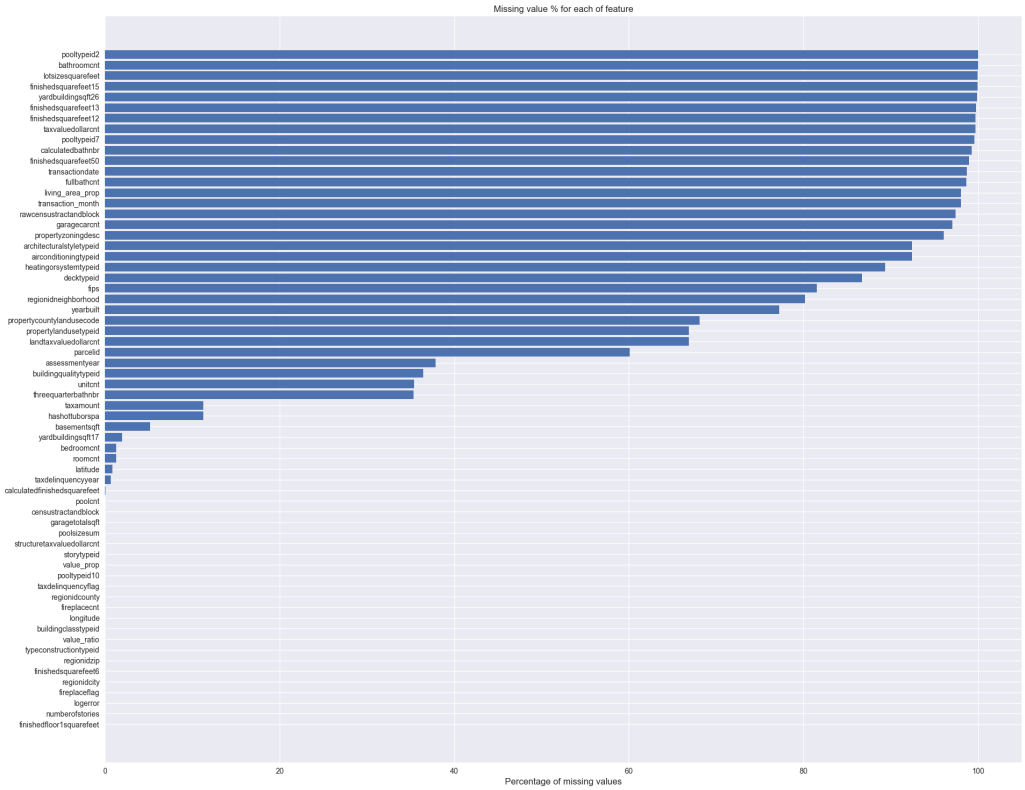

Figure 4: Missing value Percentages

Few features were dropped based on duplication or conveying the same information. For example, features such as 'FIPS (Federal Information Processing standard code)' and 'region zip' captured the same information and hence one was kept while the other removed. 3 other such duplicate entries were removed bringing the feature set size to 35. In addition, features with near zero variance like 'assessment year' were also removed as they wouldn't add any information to the machine learning models.

The 34 features available, were understood in detail to come up with methods to impute. These features could be broadly divided into multiple categories based on the information being conveyed. Features giving number count describing the house like number of bedrooms, bathrooms etc constituted one third of the feature set. This category of features were imputed based on mode representing the most frequently spotted count, with exceptions for a few. Other major categories included area of different portions of the house and tax amount payed in different fractions. These were imputed with the mean values computed ignoring the outliers.Features giving information about the location of the house constituted another major category. Features from this category didn't have many missing values.

Additional experiments would be done to identify better methods for imputation and for feature engineering.

3.3 Machine Learning Models

Most of the work until now was focused on data exploration and pre-processing. The remaining time of the project will be spent on model selection and parameter tuning along with some more data pre-processing that is required.

Different models that will be experimented for this given application are: XGBoost, lasso regression, linear regression, to name a few. Feature engineering would be done corresponding to the machine learning model selected to come up with better accuracy.

4 Experiment and Results

Experimentation on data cleaning and data reduction would be done in the coming days. This will help in drilling down to the most important features that will be helpful in better estimation/prediction of house prices.

- Analysis was done on the different properties and decision was taken to remove the attributes with greater than or equal to 90 percent missing values. This was based on the understanding that properties that have these many missing values will not give any useful information for prediction purposes.
- Different imputation techniques were tried, to fill missing values for the left over properties. For example, it was decided to impute with 0's where there could only be a binary outcome, imputation with the mode assuming it would be the most common value for the missing values for certain kind of properties like the kind of heating system for the house, the total number of bathroom.
- Random forests technique was used them to identify important features of a house for value prediction. This helped us rank down the features in the order of most important to the least important ones. It was observed that 'structuretaxvaluedollarcnt' was the most important feature. Location of the house in terms of 'latitude' and 'longitude' also played a crucial role in determining the price of the house.
- Ranking of features helped in dropping features that provided no useful purpose in estimating the value of a house.

Experimentation results from different machine learning models would be updated.

5 Conclusion

Considering the large amount of data provided by Zillow for this project, it was concluded that data exploration and pre-processing was an important step towards exploring any machine learning algorithm for house price prediction.

Imputation of missing values with the right technique also plays an important role as too much of missing data can affect the performance of any model. It was also observed that feature selection for understanding their relative importance and contribution for predicting a house price will be key to any model's forecast.

Further research on this project will focus on exploring the different algorithms and methodologies and their comparison to come up with the one that has the highest prediction accuracy.

6 References

[1] Kaggle.com. (2017). Zillow Prize: Zillow's Home Value Prediction (Zestimate) [online] Available at: https://www.kaggle.com/c/zillow-prize-1 [Accessed 07 Nov 2017].

[2] Park.B & Jae.K, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,(2016) *Expert Systems with Applications*

[3] Wang.X, Wen.J, Zhang.Y & Wang.Y, Real estate price forecasting based on SVM optimized by PSO, *Optik*

[4] Gan.V, Agarwal.B & Kim.B, Data Mining Analysis and Prediction of Real Estate Prices, *Issues in Information Systems*

[5] Jirong.G, Mingcang.Z & Liuguangyan.J, Housing price forecasting based on genetic algorithm and support vector machine, *Expert Systems with Applications*