Capstone project proposal: Estimate error from Zillow's home value prediction

Yun Zhang

June 3, 2017

1 Introduction

Home value prediction is an efficient way to represent the general information for homes, which is important for both buyers and sellers to understand the housing market. The online real estate database company Zillow provides its own home value prediction (Zestimate) to help consumers predict home value at no cost. The 'Zestimate' estimates home values according to 7.5 million statistical and machine learning models that analyze hundreds of data points on each property, and its current median margin of error compared to the real sale price is about 5%. Although the prediction has a relatively high accuracy in percentage (95%), due to the high total price of each property, the error in price is still large enough to affect its reliability especially for buyers who need to apply for financial loan based on accurate home value.

In this project, different machine learning techniques will be applied to build a predictive model that estimates the error between the 'Zestimate' and real sale price according to the information of each property. The predictive model can quantitatively analyze how the performance of 'Zestimate' varies under different scenarios for different properties, determines the key factors or features which induce errors due to the possible inadequate representation in the current 'Zestimate', and helps improving the design of a new algorithm for home value prediction.

2 Problem Statement

In this project, different features of a home will be applied to estimate the log error ($\xi_{\log}$) between 'Zestimate' and real sale price. Here the log error is defined by

$$\xi_{\log} = \log(P_{zest}) - \log(P_s) \ , \tag{1}$$

where P_{zest} is the price provided by 'Zestimate' while P_s is the real sale price for a property. The problem is clearly a regression problem, which predicts the log error by the basic information for each house. The input may include multiple features like house location, home size, tax information and facilities. This problem may be solved by many regression techniques including linear regression, k nearest neighbors and regression by decision tree and different boosting techniques.

3 Datasets

The dataset includes two portions, a full list of real estate properties data and the completed transactions in three counties (Los Angeles, Orange and Ventura, California) during 2016. It is provided by Zillow for the Zillow prize competition held on the Kaggle platform (https://www.kaggle.com/c/zillow-prize-1). The first portion of the dataset provides the basic information for nearly 3 million houses. For each house, this portion provides information of 58 numeric and categorical features on property location, home size, tax information and facilities. All the features will be treated as the input for the predictive model to estimate $\xi_{\log}$. On the other hand, the second portion of the dataset provides the sale date and $\xi_{\log}$ of over 90000 transactions for all houses sold within year 2016. The dataset is split into training and test portions. The training data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. The rest of data forms the test dataset.

4 Solution Statement

Since the $\xi_{\log}$ prediction is a regression problem, the solution basically includes two steps which are data preprocessing and the optimization of regression model. Possible effective models may be linear regression, k nearest

neighbors and tree-based regression. The final decision of the chosen model will be based on the comparison of the evaluation metrics (Section 6) of different models. For data preprocessing, more details will be discussed in Section 7.

5 Benchmark Model

The benchmark model is a predictive model that uniformly estimates $\xi_{\log}$ as the median value of $\xi_{\log}$ data in the training set. This model is equivalent to a linear regression model with only one constant coefficient.

6 Evaluation Metrics

In this project, the mean absolute error (MAE) between the predicted and actual $\xi_{\log}$ (defined by Eqn. 1) is applied as the evaluation metrics, which is defined as

$$M = \frac{1}{N} \sum_{i=1}^{N} \left| \xi_{\log,pred} - \xi_{\log,real} \right| \ , \tag{2}$$

where N is the number of test data. The lower MAE will indicate a better prediction by the predictive model.

7 Project Design

The workflow of the proposed project can be divided into five steps, which are given by

- Data preprocessing: This step includes the basic process of data cleaning like filling missing values, reasonably normalizing numeric features and transferring multi-type categorical features into dummy vector. The basic information of size and quality will be extracted by this step.

- Exploratory analysis: This step aims to understand the relation between different features and $\xi_{\log}$ of the training dataset, and will effectively help the the next step for feature selection.

- Feature selection and creation: Each existing feature is chosen as input of the predictive model based on its quality (percentage of missing values) and correlation with $\xi_{\log}$ from exploratory analysis. Several new features may be created according to the principal component analysis of the training dataset.

- Model comparison and optimization: Different regression models will be applied to predict $\xi_{\log}$, and the model with the best performance (lowest MAE) will be selected and optimized.

- Model analysis: This step aims to analyze and explain the optimized model and extract more insights on how $\xi_{\log}$ is induced and varies under different scenarios. For example, if the tree-based regression is selected, the production of $\xi_{\log}$ can be directly explained by the multiple splits of the decision tree. All the insights will help Zillow understand the shortcomings of their 'Zestimate' model.

8 Summary

In this project, the dataset on the 2016 real estate properties data and the completed transactions of three counties (Los Angeles, Orange and Ventura) in California will be analyzed and applied to predict the log error between real sale price and home value prediction ('Zestimate') provided by Zillow. The final predictive model will help Zillow understand the performance of 'Zestimate' under different scenarios and provide more insight on future improvement. This project requires various statistical and machine learning techniques, and will be a good capstone project for the completion of Udacity machine learning nano degree.