Capstone project proposal: Zillow's Home Value Prediction

Shankar Adhikari

December 22, 2017

1 Introduction

Purchasing a house is one of the most important decision and also an expensive purchase for the person in his or her lifetime. This requires many years of planning and a lot of financial efforts. The house price is changing all the time also and it does not go as planned. Various factors are responsible changing house price; eg. tax plan of government, economic situation and so on. Predicting house price is a challenging task and in the meantime is very important for buyer or seller to understand the market price of the houses.

The online real estate database company Zillow provide free service for the consumer to predict house price, is called Zestimate. The estimated house prices (Zestimates) from Zillow are using 7.5 million statistical and machine learning models that analyze hundreds of data points on each property [1]. The Zillow continuously improve median margin error of prediction from 14 % to 5 % and now it is one of the largest and trusted marketplaces for real estate information in the U.S. Although the prediction has high accuracy 95%, the error is still a large amount when comes to million dollars investment.

2 Problem Statement

Zillow's home value prediction is done as stated in the Kaggle competition [1]. The goal is to estimate the residual error given by the Zestimates and the sale prices defined as;

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

where P_{zest} is the price provided by 'Zestimate' while P_s is the real sale price for a property.

This project here will implement multiple machine learning algorithm as stated by problem (regression the problem), through the use of input features such as house location, home size, tax information and others provided by the data.

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 [1]. The first portion contains properties information about 3M houses (properties_2016.csv), and second portion contains residual errors for about 90K houses (train_2016_v2.csv).

The properties file contain 58 numerical and categorical features of houses information those sold within the year 2016. Based on label data ($\xi_{\log}$) available, the whole data is divided into training and test set. The training data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016.

4 Solution Statement

This project is a supervised regression problem. Our goal is to predict residual error of Zestimates using provided features and labels for each house through the machine learning models. The entire task can be categorized as; data preprocessing and the optimization of models. The possible regression model 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

This project is ongoing Kaggle competition, where leaderboard shows the scores that can be used as the benchmark. Also, this project is about mini-

mizing the difference between predictions of our new model and the existing Zestimate model. If the difference is minimum, our model compares well with Zestimate and if it becomes large it goes another way. So, Zestimate itself is our benchmark model.

6 Evaluation Metrics

In this project, the mean absolute error (MAE) between the predicted and actual, is the evaluation metrics. This is defined as;

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

where N is the number of test data, $\xi_{\log,pred}$ is the predicted values from model and $\xi_{\log,real}$ is the value from label data.

7 Project Design

The project design will be as follow:

- Exploratory Analysis of Data: This step includes the graphical exploration of data features, missing value analysis, univariant and bivariant analysis of data, correlation analysis between features.

- Data preprocessing: In this step, we will implement basic data processing such as filling missing values, normalizing numeric features, feature transformation, removing outlier and so on.

- Feature selection: This step is for the determination of the best features that input to the model based on quality and correlation with label data. Also, new features will be created using Principle Component Analysis (PCA).

- Model Selection: Different regression models will be applied to predict output label, and we are selecting the best performing model. Also, we will use grid search method to choose best performance hyperparameters.

References

[1] https://www.kaggle.com/c/zillow-prize-1