Can we predict house prices using known features of each house and a supervised learning approach?

Florence Galliers

18/10/2020

Contents

1 Background: **1**
1.1 Objectives: 2

2 Data: **2**
2.1 Data Description 2
2.2 Data Preparation 2

3 Methods: **2**

4 Results: **2**
4.1 Conclusions: 2

References **2**

1 Background:

House prices are an important part of the economy and usually reflect trends in it. They can be influenced by the physical condition of the house and by other attributes such as location (Bin 2004). Prices are important for homeowners, prospective buyers and estate agents, prediction methods could help lead to more informed decisions to each of these stakeholders. Gao et al. (2019) suggest that prediction models may be useful for a few reasons, firstly in narrowing down the range of available houses for prospective buyers. People looking to put their house on the market could use prediction models to look for the optimal time to do so. Prediction accuracy is important.

In most countries there is some form of house price index that measures changes in prices (Lu et al. 2017). This contains a summary of all transactions that take place but not the individual features of each house sold, therefore it cannot be used to make predictions of house price.

Something else to take into account is that it may be difficult for prospective buyers to visualise how square footage measurements of a house are calculated or how this measurement translates into physical size if they have not visited the house themselves. Buyers therefore rely on factors such as the number of bedrooms, bathrooms or house age to get an idea of the value of the house. This analysis will focus on which features of a house have the largest influence on the prediction of house prices. This report does not look at the effect of time on house prices. It is already a well known fact that house prices increase every year (Alfiyatin et al. 2017).

Many house price prediction models have been created using machine learning methods. The hedonic price model is the most extensively researched and uses regression methods (Gao et al. 2019). Machine learning methods use data in a ‘training set’ to build a model, this model is then used to make predictions on an unseen ‘test set’ of data. The accuracy of models can be calculated by taking the predicted values from the actual values.

1.1 Objectives:

- Understand which attributes of houses given in the data set can be used to effectively construct a prediction model for house price (dependent variable).
- Minimize the differences between predicted and actual house prices by using model selection to choose the most accurate model.

2 Data:

2.1 Data Description

The dataset chosen for this analysis is from houses sold in 2014 in Washington, USA. It contains the sale price (US dollars) along with attributes of each house such as number of bedrooms, number of bathrooms, etc. There were 4600 observations with 17 variables in the original data set downloaded from [Kaggle] (https://www.kaggle.com/shree1992/housedata). Although the dataset is from 2014, it was a particularly interesting data set because it contains a large amount of information and interesting selection of variables. A more recent data set, or one from the UK could not be found, and so this analysis will go ahead with this data set.

2.2 Data Preparation

The first task was preparing and cleaning the dataset. This served two purposes, firstly to get to know the different variables in the data and existing patterns or correlations between them and secondly to carry out feature selection. Missing values were searched for and removed and any observations in which price equalled zero were removed. The cleaned dataset was exported ready for use in the main analysis. This cleaned data set contained 4522 observations and 12 variables (Table 1).

The remaining variables and their descriptions are shown in Table 1:

3 Methods:

```r
set.seed(2)
n = nrow(data) #number of rows
train_index = sample(1:n, size = round(0.8*n), replace=FALSE)
train = data[train_index ,] #takes 80% of the data for training set
test = data[-train_index ,] #remaining 20% for the test set
```

4 Results:

4.1 Conclusions:

References

Alfiyatin, Adyan Nur, Ruth Ema Febrita, Hilman Taufiq, and Wayan Firdaus Mahmudy. 2017. “Modeling House Price Prediction Using Regression Analysis and Particle Swarm Optimization.” *International Journal of Advanced Computer Science and Applications* 8.

Bin, Okmyung. 2004. “A Prediction Comparison of Housing Sales Prices by Parametric Versus Semi-Parametric Regressions.” *Journal of Housing Economics* 13 (1): 68–84.

Gao, Guangliang, Zhifeng Bao, Jie Cao, A Kai Qin, Timos Sellis, Zhiang Wu, and others. 2019. "Location-Centered House Price Prediction: A Multi-Task Learning Approach." *arXiv Preprint arXiv:1901.01774*.

Lu, Sifei, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh. 2017. "A Hybrid Regression Technique for House Prices Prediction." In *2017 Ieee International Conference on Industrial Engineering and Engineering Management (Ieem)*, 319–23. IEEE.