New Trends in Mathematical Sciences
http://dx.doi.org/10.20852/ntmsci.2018.327

A new machine learning approach to house price estimation

Changchun Wang[1] *and Hui Wu*[2]

[1]International Monetary Fund, Washington, DC 20431, USA
[2]Department of Mathematics, Clark Atlanta University, Atlanta, GA 30314, USA

Received: 18 September 2018, Accepted: 4 November 2018
Published online: 29 December 2018.

Abstract: In this paper, we propose using Random forests, a machine learning approach, to build house price estimation model. Compared to benchmark linear regression model, random forests model can better capture hidden nonlinear relations between the price and features of houses and gives an overall better estimation. Numerical experiments have been done on North Virginia house price data, which can strongly demonstrate our findings.

Keywords: Machine learning, random forest, house price estimation.

1 Introduction

House price estimation has been broadly studied a lot as described in [1,2]. Establishing a housing price estimating model can greatly help the formulation of housing prices and improve the accuracy of estimation of future real estate policies [3,4]. A lot of models to estimate house price has been implemented a lot, as in[5,6,7,8]. People have used some machine learning methods, like self- organizing map (SOM) [9] to model and predict house price [10]. The main possible influencing factors of housing prices can be divided into the following categories: Location (Urban, Midtown, Average/Price, etc.): House prices are positively correlated with urban development levels and negatively correlated with distances from developed regions. Transportation (subway, bus, road rating, etc.): House prices are positively correlated with traffic convenience. Housing conditions (type, construction year, renovation level, floor, area, etc.): In most cases, the better the housing conditions, the higher the price. Supporting facilities (school district level, public facilities include parks, hospitals, shopping malls, etc.): The overall quality and convenience of the package play a positive role in housing prices. The quality of house can be easily measured and compared. However, the convenience is often hard to measure and sometimes missed in recorded data. Features defines convenience are of the with nonlinear relations. Our motivation is the use machine learning method to capture that convenience information. We will compare random forest and linear regression on house price data without specific feature specify defining convenience. We think zip code or coordinates of house contains the convenience information, and by training machine learning model on these feature, that information will be reflected in the model and improve model accuracy.

In this paper, we mainly focus on a machine learning technique called Random Forests to build a house price model. Unlike the traditional method, it can capture hidden non-linear relations between the price and features of houses and gives overall better estimations. We use the classical multiple linear regression model as benchmark. By including features such as zip code, longitude and latitude, which are not linearly related to house price, we found that random forest model performs much better and captures the hidden information in those features.

* Corresponding author e-mail: ccwang81@gmail.com

Random Forests (RF), an ensemble method in machine learning based on the week learner decision trees, has now been widely used in prediction. It was created and presented by Tin Ham Ho [11,12]. Leo Brakeman developed this method in [13]by using internal estimates to monitor error, strength, and correlation and to measure variable importance. RF has been applied in medical area, for example, using data to find the cluster of patients [14], RF diversity is very useful because it can handle mixed variables well, the dissimilarity of RF is useful for detecting tumor sample clusters [15]. RF can easily deal with data in ecosystem and geology, because it handles multi-purpose data sets and generate high classification accuracy graph [16].

The rest of this paper is structured as follows: In the second section, we introduce the Random Forests method. In the third section, we provide a short review of the classical linear regression model. In the fourth section, we explain the house price data and provide some analysis of the attributes in the model. In the fifth section, some numerical results are provided. Finally, in the last section, we present the conclusion.

2 Random forests method

Random forests model is an ensemble method of decision trees, which is a weak learner and can easily be over-fitting. By ensembling the decision tree, the problem of instability and high variance of decision trees can be overcome. Because these decision trees are generated using random sampling methods, they are called random forests. The logic of the method is to randomly select a subset of explanatory variables and train with weak learners separately. It builds a prediction model by independently training weak prediction models, typically decision trees. Predictions of each tree are combined using some model averaging technique. As shown in Fig. 1, the random forests method is based on randomly selecting a bag of features, and build a decision tree separately. These decision trees in random forests are independent to each other. A majority voting can be used on such kind of trees.

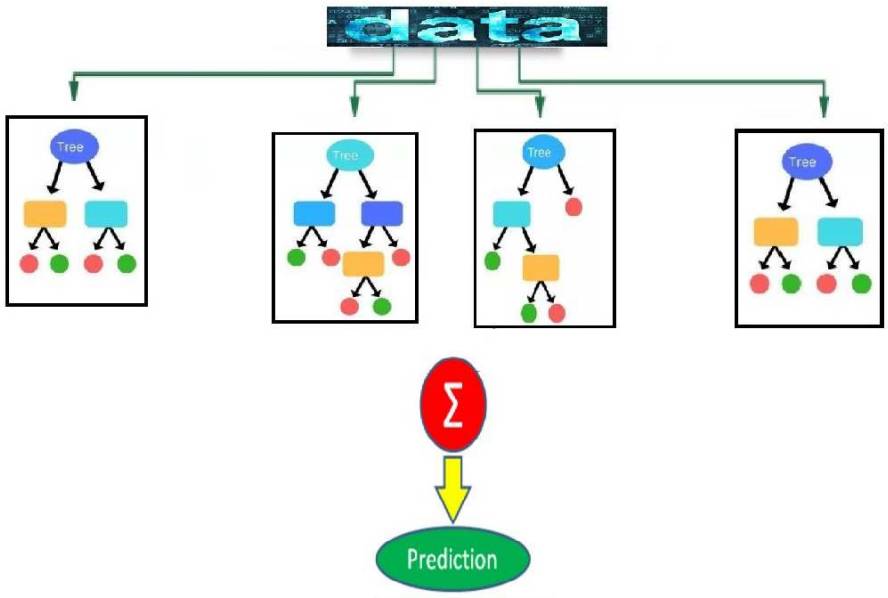

Fig. 1: Random forest model.

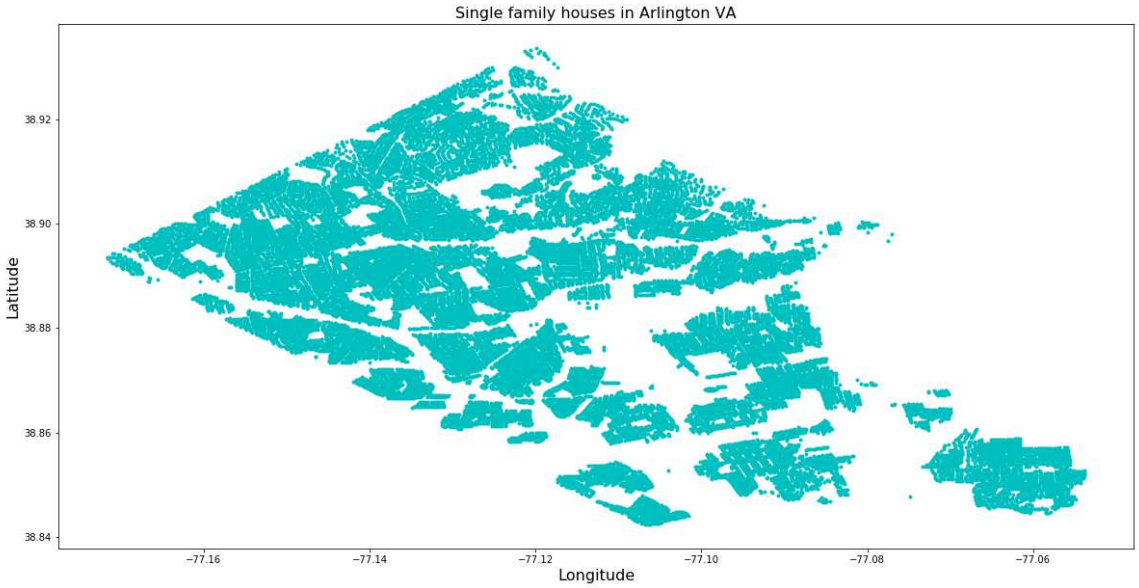

Fig. 2: Single family houses in Arlington county.

3 Linear regression model

We use a linear regression model as a benchmark to predict the house price. Generally, a linear regression model is a statistical analysis method that uses the regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. It models the relationship between one or more independent variables and dependent variables using a least-squares function called a linear regression equation. This function is a linear combination of one or more model parameters called regression coefficients. If there is only one independent variable, it is called simple regression, and if more than one independent variable involved, it is called multiple regression. Linear regression has many practical applications.

Generally, Linear regression is a model that predicts the proportional relationship between a dependent variable and a predictor and fit the mapping between data input and output. The standard equation for linear regression is given as follows:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{1}$$

In the formula, the beta is the estimated parameters for independent variables x_i, y is the dependent variables, and ε if the error term. The objective of the linear regression model is to estimate the parameters β_i by minimizing the sum of squared errors. In our model, the y is the estimated house price, and x_i are variate features of estimating the house price, such as area, zip code.

4 Data and methodology

The data we use is the single-family house assessment price of 2015 of Arlington country, Virginia USA. There are totally 27649 data points (houses). We have features such as the lot size and year built clearly defines house quality but we don't have any specific feature that obviously defines house convenience.
Besides the house price, we also have some basic features of the house such as the lot size of the house, the year the house has been built, zip code of the house and the location of the house.

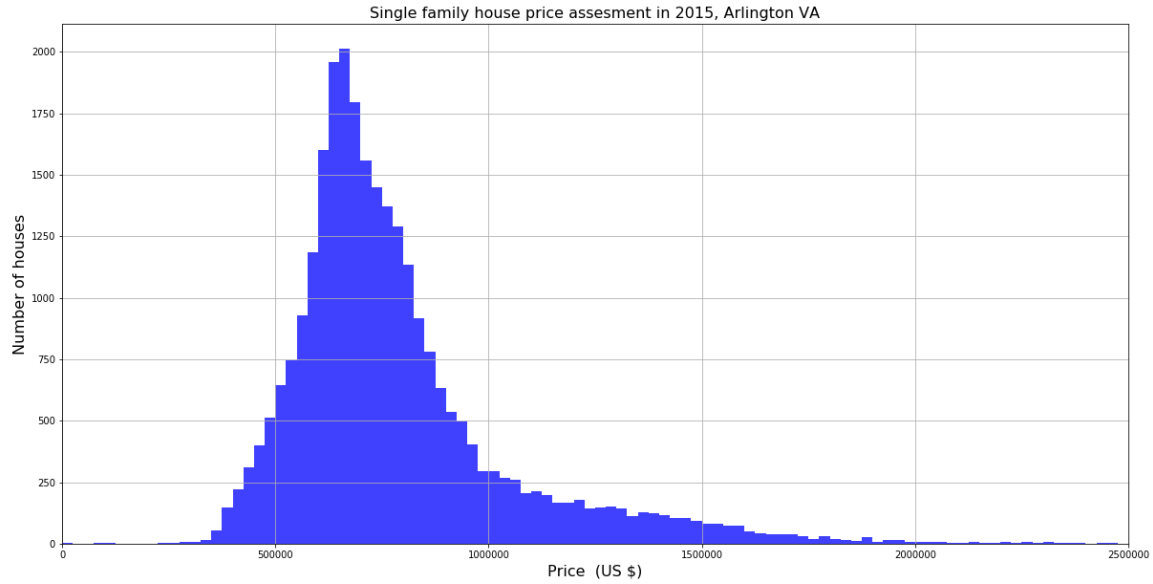

Fig. 3: 2015 House price assessment histogram, Arlington county, VA.

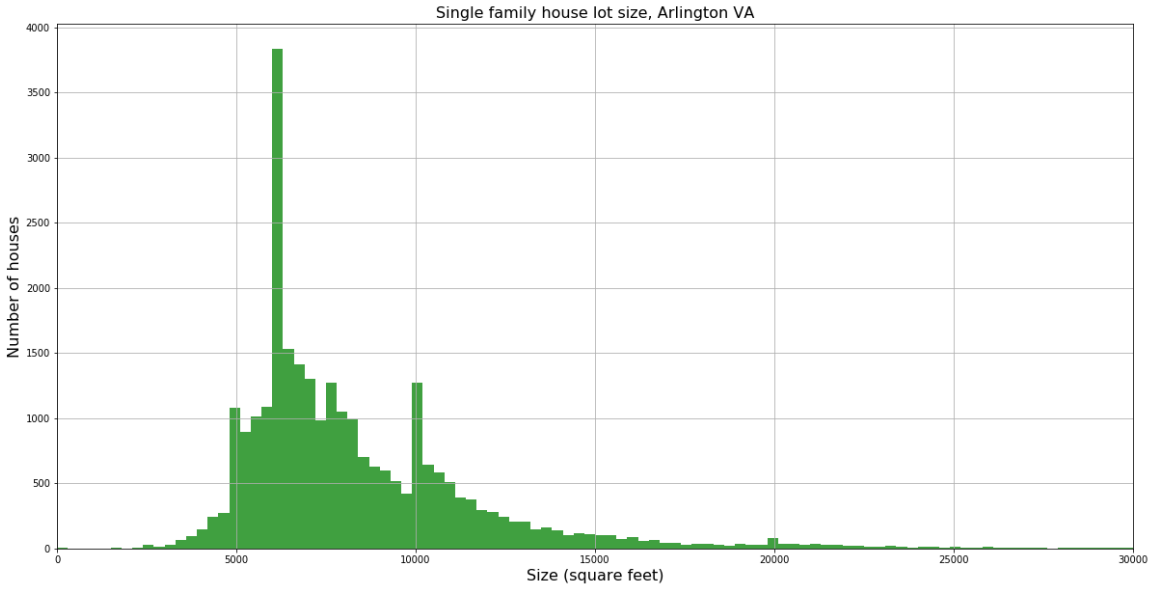

Fig. 4: House lot size histogram, Arlington county, VA.

We use linear regression model as our bench mark and compare it with Random Forests model. We will repeat the test on different sets of features. Some sets only contain features related to the quality of houses and some sets contains features with information about convenience of houses. By comparison, we can show how Random Forests can take advantage of those information and give overall better estimation.

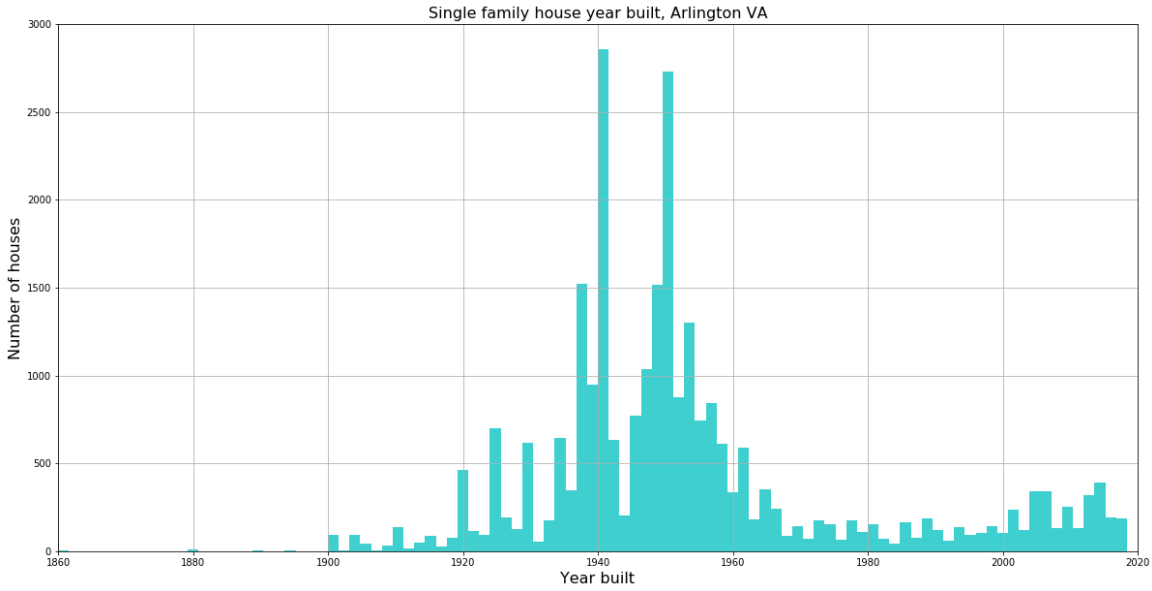

Fig. 5: House year built histogram, Arlington county, VA.

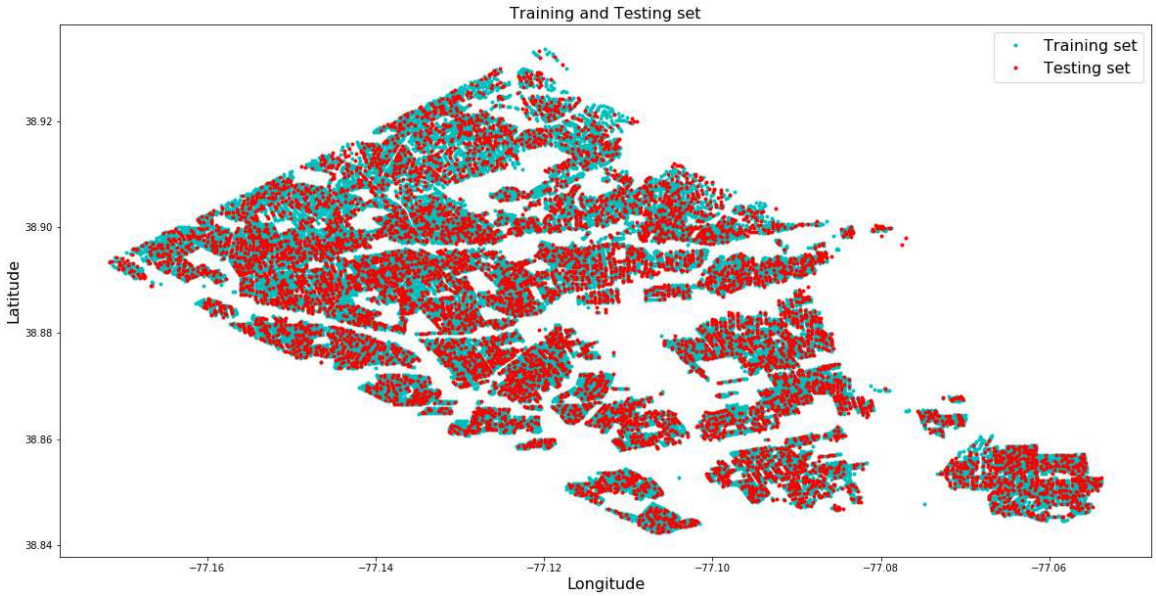

Fig. 6: Training and testing data distribution.

5 Numerical experiments

We split the data into training and testing sets. 30 of the total data are randomly picked as testing set, the rest are training set. We train our model only on the training set and compare out of sample result on test set. We compared the models on different feature sets. Random forest general performs better than bench mark linear regression model in terms of R2 and RMSE. But the much more important is that, when we include location related features, random forest model get significant improvement. This means the model can somehow capture the hidden nonlinear relationship between house price and house location, which also reflects the convenience components in determining house price.

Features/ Methods	R-square	RMSE
Features: Year built, lot size		
Random forests	0.639135706989	367.054972867
Linear regression	0.344215758946	407.940568982
Features: Zip code, latitude, longitude, year built, lot size		
Random forests	0.68614045456	357.59049678
Linear regression	413.775211328	413.775211328
Features: latitude, longitude, year built, lot size		
Random forests	0.701680132695	352.892749026
Linear regression	0.407037969505	389.06342124
Features: Zip code, latitude, longitude, year built, lot size		
Random forests	0.701310346391	352.065553406
Linear regression	0.539887986037	381.280477804

Table 1: Numerical results.

6 Conclusion

In this paper, we estimated the house price in the county of Arlington in Virginia, both by a linear regression model and a machine learning technique Random Forests. We found that the Random Forests can capture the nonlinear hidden relationship between house price and house location and give an overall better estimation than bench mark linear regression. This simple model can be scaled up for larger data with more features and captures the nonlinear information traditional models used to neglect.

Acknowledgments

We thank Arlington County Open Data, https://data.arlingtonva.us/home to provide the house price data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors have contributed to all parts of the article. All authors read and approved the final manuscript.

References

[1] The housing bubble and the GDP: A correlation perspective, R. M. Valadez, Journal of Case research in Business and Economics, 2010

[2] The relationship between Interest Rates, Income, GDP Growth and House Prices, Ting Xu, Research in Economics and Management ,Vol2, No. 1 2017

[3] Real Estate Prices and Economic Cycles, J. M. Quigley, International Real Estate Review, 1999 Vol. 2 No. 1, pp 1-20

[4] House Prices, Economic Output, and Inflation Interactions in Iran, Ali A. Naji Meidani, Research in Applied Economics, 2011, Vol 3, No. 1: E2

[5] Forecasting the US. Real house Index, Plakandaras V, July 2017

[6] Using machine learning algorithms for housing price prediction: The case of Fair fax Country, Virginia Housing Data, Byeonghwa Park, Jae Kwon Bae, Expert Systems with Applications. Vol 42, Issue 6, 2015.

[7] Housing Price Forecasting based on Genetic Algorithm and support vector Machine. Gu Ji Rong, Expert Systems with Applications, Vol 38, Issue 4, 2011
[8] Real Estate Price Forecasting Based on SVM optimized by PSO, Xibin Wang, Optik-International Journal for light and Electron Optics. Vol 125, Issue 3, 2014
[9] Customer Segmentation of Credit Card Default by Self Organizing Map, H. Wu and C.Wang
[10] Real Estate Investment Appraisal of Buildings using SOM, A Tulkki, 1998, Springer Finance
[11] Random Decision Forests, Ho, Tin Kam, Proceedings of the 3rd International conference on Document Analysis and Recognition, Montreal, QC 1.4-1.6 August 1995, pp 278-282.
[12] The Random Subspace Method for Constructing Decision Forests, Ho, Tin Kam, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8):832-844.
[13] Random Forests, Leo Breiman, Journal Machine Learning, Vol 45, Issue 1, 2001
[14] Tumor Classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Shi T. Seligson D, Modern Pathology, 2005, 1.8(4) 547-57
[15] Unsupervised Learning with Random Forest Predictors, Shi T., Horvath S. 2006, Journal of Computational and Graphical Statistics, 1.5(1) 1.1.8-1.3.8
[16] Decision Tree, Rule-Based, and Random Forest Classification of High-Resolution Multispectral Imagery for wetland mapping and inventory, Hongqing Liu, Semote Sens, Vol. 10, Issue 4 2018