An Applied Econometric Approach To Predict Housing Value  
  
Introduction

Yifan Cheng, Jincen Jiang, Wenyu Li, Shuxian Mao, Feiyang Ni,Yirui Zhang

USC Econ 460 Final Project

Abstract

[The abstract should be one paragraph of between 150 and 250 words. It is not indented. Section titles, such as the word Abstract above, are not considered headings so they don’t use bold heading format. Instead, use the Section Title style. This style automatically starts your section on a new page, so you don’t have to add page breaks. Note that all of the styles for this template are available on the Home tab of the ribbon, in the Styles gallery.]

Keywords: [Click here to add keywords.]

**An Applied Econometric Approach To Predict Housing Value  
  
Introduction**

Real Estate Property is not only a human's essential requirement; it also serves as a symbol of a person's wealth and prestige in the modern era. Investing in real estate appears to be profitable in general, as property values do not depreciate rapidly. Changes in the real estate market can have a ripple effect on a variety of stakeholders, including household investors, lenders, and policymakers. Investment in real estate appears to be an appealing investment option. Thus, anticipating the value of real estate is a critical economic indicator.

Machine learning is a cutting-edge approach that enables the identification, interpretation, and analysis of enormously complex data structures and patterns (Ngiam & Khor, 2019). It enables consequential learning and improves model predictions by inputting more recent data in a methodical manner (Harrington, 2012). Machine learning has emerged as a critical predictive technique in recent years, owing to the growing trend toward Big Data. This paper provides a method that demonstrates the capabilities of the machine learning technique and evaluates the performance of several models in order to determine the most effective strategy for predicting housing prices.

# Literature review

In a relatively traditional industry like the real estate industry, some authors believe that the using of machine learning will enable the industry to develop better. In the case of Pai and Wang, they use machine learning models and actual transaction data to predict real estate prices (Pai &Wang 2020). They think real estate forecasts are crucial to the formulation or real estate policies and it also can help real estate owners and agents make wise decisions. In their article, they launched four machine learning models, least square support vector regression, classification and regression tree, general regression neural networks and backpropagation neural networks to predict real estate transaction prices with actual transaction data in Taichung, Taiwan. And the actual transaction data they used include the attributes and transaction prices of real estate which are also used as independent and dependent variables of the machine learning models. The empirical results revealed that all four machine learning models provides accurate predictions. In addition, least squares support vector regression is better than the other three prediction models, and in terms of MAPE, it has obtained more accurate results than some previous studies. Therefore, the author believes that the least squares support vector regression using genetic algorithms is a reliable machine learning technique for predicting real estate prices.

In the case of Park and Kwon Bae, they also relied on the use of machine learning techniques for predicting housing values (Park, 2015). Decision Tree, repeated incremental pruning to reduce errors (RIPPER), Naive Bayes, and AdaBoost are among the algorithms compared by the author. RIPPER which is a propositional rule learner with an average error of roughly 25%, produces the best results. They looked at housing statistics for 5,359 townhouses in Fairfax County, Virginia between 2004 and 2007. These assets have 76 attributes, of which 28 were chosen via a t-test and logistic regression filtering process. It also has three variables that refer to the ratings of the area's elementary, middle, and high schools. The remaining eight factors are interest rates on mortgage contracts, location, and construction and selling dates. As a result, rather than being a regression problem, the problem can be viewed as a classification problem that decides if an investment is beneficial.

Even though the hedonic pricing model based on ordinary least squares (OLS) linear regression has been utilized for many real estate appraisals, there are still issues with the model's stability and accuracy. In Hong, Choi and Kim’s paper, they investigate the features of a house price predictor based on the Random Forest method by comparing it with that of a conventional hedonic pricing model (Choi & Kim, 2020). They used apartment transaction data from 2006 to 2017 in the Gangnam district of South Korea, one of the most developed areas. Using a data set that represents 40% of all transactions in the sample area. The average percentage deviation between predicted and actual market price in the RF predictor was found to be around 5.5 percent, whereas it was nearly 20 percent in the OLS-based predictor. The probability of the predicted price being within 5% of the actual market price was 72% with the RF predictor, whereas only about 17.5 percent of the regression-based predictions fell within the same range. These findings indicate that, in the practice of mass appraisal, the random forest method may be a useful complement to hedonic models, as it better captures the complexity or non-linearity of actual housing markets.

Levantesi and Piscopo think machine learning techniques are better to explain which variables are more important in describing the evolution of the real estate market (Levantesi & Piscopo, 2020). They applied random forest algorithm on London real estate market and looked at the local factors that influence the interaction of housing demand, supply, and price. The variables were chosen with an urban viewpoint, as the interaction of local factors such as population growth, net migration, new development, and net supply is the key driver driving the market. Their numerical results reveal that random forests outperform traditional regression approaches based on GLM in terms of prediction improvements. Random forests are commonly used in tiny data sets by data scientists because the bootstrapping that random forests rely on allows the algorithm to perform well in any case. Random forests are simple to build and do not necessitate costly hyperparameter modifications.

# Data

This paper collects the data from the American Housing Survey (AHS) 2013 Metropolitan dataset. The survey has been the most comprehensive national housing survey in the United States, which provides current information in various aspects related to house price, such as its size, composition and quality. The raw data amounts to 70044 observations and 2167 variables, including 145 census regions. This high dimensional data makes statistical computations challenging and this paper implements the following steps to handle missing observations and select the variables: Firstly, we consider all owner-occupied units (TENURE = 1, VACANCY = −6) with non-missing and positive value (VALUE > 0), which keep 35852 observations. Secondly, we conduct manual screening to select 173 out of 2167 variables that may be correlated with house prices, where we merge the dummy variables in the categories of Appliance, Heating & Cooling, Housing Problems and Disaster Planning into four aggregate variables. Thirdly, we remove the variables that have over 80% of missing values or non-applicable values, with 74 variables left. Lastly, we remove the observations that have missing values in any of the variables and reach our processed dataset with 20415 observations and 74 variables.

The table 1 shows the composition of the remaining variables and the table 2 shows top 20 census regions that have the most observations, which accounts for more than 80% of total observations.

|  |  |  |
| --- | --- | --- |
| Categories | Example of variables description | Count |
| Accessibility | Household has a disabled person | 1 |
| Admin | Central city / suburban status | 7 |
| Demographics | Educational level of householder | 4 |
| Disaster planning | Disaster planning | 1 |
| Equipment and Appliances | Flag indicating emergency food | 3 |
| Healthy Homes | Water safe for drinking and cooking | 2 |
| Home Improvement | Annual cost for routine maintenance | 1 |
| Housing Costs | Annual real estate tax payment | 11 |
| Housing Problems | Water leak in roof | 2 |
| Income | Household income | 4 |
| Neighborhood Features | Rating of neighborhood as place to live | 2 |
| Occupancy and Tenure | Owner or renter status of unit | 2 |
| Structural | Number of bedrooms in unit | 34 |
| **Total** |  | **74** |

*Table 1*

*Table 2*

|  |  |
| --- | --- |
| **Census Region** | **Number of Observations** |
| 9999: Non-metro or suppressed | 10325 |
| 2160: Detroit, MI | 1276 |
| 6160: Philadelphia, PA-NJ | 1207 |
| 9991: Chicago Areas | 745 |
| 9993: Northern New Jersey Areas | 739 |
| 4480: Los Angeles-Long Beach, CA | 297 |
| 1600: Chicago, IL | 270 |
| 9992: New York Areas | 247 |
| 8840: Washington, DC-MD-VA | 173 |
| 5600: New York City | 157 |
| 6200: Phoenix, AZ | 151 |
| 5380: Nassau-Suffolk, NY | 146 |
| 5120: Minneapolis-St. Paul, MN | 142 |
| 3360: Houston, TX | 141 |
| 7320: San Diego, CA | 122 |
| 0360: Anaheim-Santa Ana (Orange County), CA | 118 |
| 1920: Dallas, TX | 118 |
| 7600: Seattle, WA | 109 |
| 1120: Boston, MA | 106 |
| 1680: Cleveland, OH | 106 |
| Rest | 3720 |
| **Total** | **20415** |

# Methodology

# Result

**Linear regression:**

We use 310 features, such as AMTE, DENS, DINING, FAMRM…, as independent variable, use VALUE as dependent variable. We found that out of sample mean square error is 1601.103, and root mean square error of linear regression is 40.01378.

**Lasso:**

图表

描述已自动生成

Lambda between -6 and -4.

We use the same variables and found that out of sample root mean square error of linear regression is 40.1186.

**Regression Tree:**

图示

描述已自动生成

full size

Full tree

图表

描述已自动生成

Best size

图示

描述已自动生成

Pruning tree

The best size = 6

Out of sample root mean square error of linear regression is 43.32149

**Random forest:**

图片包含 文本

描述已自动生成

Feature importance

Out of sample root mean square error of linear regression is 39.9963

**Comparation:**

Box plot drawn based on the results of 10 random samples.

图表, 箱线图

描述已自动生成

According to the figure of root-mse comparation, the minimum mean root-mse of these 10 random samples is from the Random Forest model, so, Random Forest is the best model in this data set. (Although the above results show that OLS performs better than LASSO, the mean of 10 times random pick-up samples shows that LASSO performs better than OLS).

**Predict result:**

表格

描述已自动生成

We used 1/5 of the sample as the prediction dataset and used the best model-random forest model to predict the value of this sample. We define the accuracy of the prediction model using the formula:

accuracy = 1 – abs (predict value – true value)/true value.

Although with some extreme predicted results which impact the mean of the accuracy of predict sample, the accuracy of 3/4 of the sample is higher than 0.6, and the accuracy of half of the sample is higher than 0.8. We believe that our model has a good performance in the prediction dataset.

References

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet. Oncology, 20(5), e262–e273

Harrington, Peter. *Machine Learning in Action*. Manning Publications Co., 2012.

Feggella, D. (2019). What is machine learning? emeRJ.

Pai, Ping-Feng, and Wen-Chang Wang. 2020. "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices" Applied Sciences 10, no. 17: 5832. https://doi.org/10.3390/app10175832

Park, B.; Kwon Bae, J. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. Expert Syst. Appl. 2015, 42, 2928–2934. [CrossRef]

Hong, Jengei, Heeyoul Choi and Woo-sung Kim. “A HOUSE PRICE VALUATION BASED ON THE RANDOM FOREST APPROACH: THE MASS APPRAISAL OF RESIDENTIAL PROPERTY IN SOUTH KOREA.” International Journal of Strategic Property Management 24

Levantesi, Susanna, and Gabriella Piscopo. 2020. "The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach" Risks 8, no. 4: 112. <https://doi.org/10.3390/risks8040112>