Mathematical Modeling of Housing price in Beijing

Guannan Liu

Abstract

The purpose of this study is to develop several statistical models for predicting individual apartment and house prices regarding size, time, and detailed descriptive information, including the study multivariable regression, tree regression, generalized boosted regression model and cluster analysis. Each model will be evaluated with its r-square and the mean square error. The data for the metropolitan areas of Beijing through 2011 to 2017 are analyzed. The model is shown to have a better predictive ability in cluster analysis and being grouped with charactered subset. The result generally suggests the variable of size, constriction time/type and facilities having greater impacts to house and apartment price than other variables.

Introduction

Many parties are interested in the market value of the residential real estate. Buyers and sellers want to know market values for bidding and listing. Banks want to know the extent to know the value of their collaterals. Local governments use market values in part to set real estate taxes. Modeling the house price is also one of the hottest topics in machine learning

.

Modeling real estate prices presents a unique set of challenges. Each observation is distinctive, each has its own set of hedonic characteristics: number of bedrooms, square size, view, amenities, facilities and the most important variable, location. In the other hand, the fair market price of the real estate only recorded while transaction. Each transaction involves large and influential capital. However, the sales occur infrequently. (Wheaton 1999) As a result, during a certain period of time, out of the entire population of apartment and house, only a small percentage of the population are actually been sold and the price are recorded. In Yaman's analysis, the turnover rate also varies from one area to another area, which will also add variation to the dataset and difficulty to modeling. (Yaman 2014) From these information, this project's objective is to develop a practical model to predict prices from which we can construct a clear picture to identify the key factor which impacts the market price.

Data

As discussed in the project proposal in the October of 2020: the real estate dataset usually polarized. The macro-level studies and data usually published by the government, organization, and institutions. These macro data I found are complete and well organized. The problem is these data usually being lack of detail, which only contains the data of transaction time, price and regional

information. In the other hand, the detailed datasets obtained from industry, reginal realtors associations are not complete.

This project studies the dataset of Beijing housing price from 2011 to 2017, obtained from the Lianjia.com. Because of the monopoly power of Lianjia in china's real estate market, Lianjia's data could be considered as a comprehensive and reliable source to study with. This project study a dataset direct fetched from bj.Lianjia.com/chengjiao. The web crawler read all the key information from Lianjia.com and save it to the .csv file. In this dataset, it includes 318,851 observations in a 7-year range.

Lianjia formerly called Homelink, is a Chinese real-estate brokerage company founded in 2001. In the following 19 years, the company has shown tremendous growth, with a backing of multiple funding from giants like ByteDance, Vanke, Sunac China Capital, Tencent, Baidu, and Fosun Group. Today, it has over 8000 offline stores within the country and has enrolled over 150,000 brokers to carry out its deals on the native grounds. Started as a property agent, the company is now a full-stack real estate agent which provides multiple services like second-hand housing, new housing, renting, residential real estate, overseas real estate, internet platform, wealth management, post-real estate market etc. As of 2019, it had approximately 6,000 brokerage offices and more than 150,000 brokers. By 2018, its market value had reached US$6 billion. (Yanogya 2018)

Data cleaning

The raw data fetched form the internet contain many errors and missing variables. There are two steps to prepare the data.

The first step is to review the data type and checking the missing and "NA" variables, because handling missing data is important as many machine learning algorithms do not support data with missing values. checking the missing value and data type, I found that the most observations with missing value and error are the data before the year of 2015. I think this is understandable because the company may have different requirements over the year. 2016's total count is also the highest. After removing the data with missing values, there are still 91,242 data with 26 variables.

Second, it is necessary to change each variable to correct type, such as changing the binary variable and floor to numerical from string. The raw data is directly fetched form the Chinese website by neutral Chines language. It is needed to translate into numerical, such as 4b4b to 4 bedroom and 4 bathrooms.

Third, deleting variables by a test of collinearity. A collinearity is a special case when two or more variables are exactly correlated. This means the regression coefficients are not uniquely determined. In turn it hurts the interpretability of the model as then the regression coefficients are

not unique and have influences from other features. Based on what we observed in step 1 and human common sense, some variables have strong collinearity. For example, unit price lost its value when we know it equals the total price divided by the size of the apartment, as while as ladder ratio is also high related to the floor.

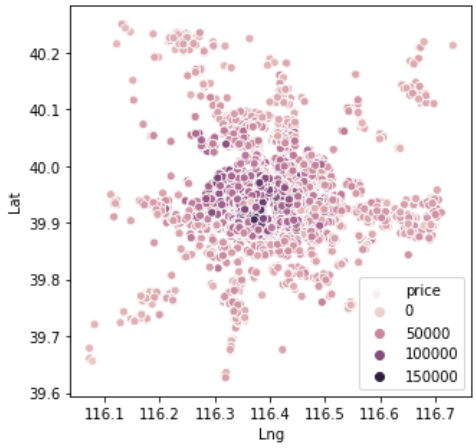
Variable Name	**Meaning**	**Variable Name**	**Meaning**	**Variable Name**	**Meaning**
URL	*Link to the source*	Lng,Lat	*GPS information*	TradeTime	*Time of the transaction*
Followers	*How many followers on website*	DOM	*Days of listing*	TotalPrice	*Total Price*
Price	*Unit Price*	Square	*Size in m^2*	DrewingRoom	*Numbers of living room*
LivingRoom	*Number of Bedroom*	Kitchen	*Number of kitchen*	BuildingType	*Range of 1-4,$ is the best*
Ladder ratio	*How many ladder need to take by the total floor of the building*	Floor	*Floor*	Elevator	*True/false to have an elevator*
Fiveyears	*True/false variable to A Chinese reginal tax policy*	Subway	*True of false to have a subway station close by*	District	*Area of the city*
Renovation condition	*Level of the decoration*	Construction Time	*Year to build*		

Table 1. Data dictionary

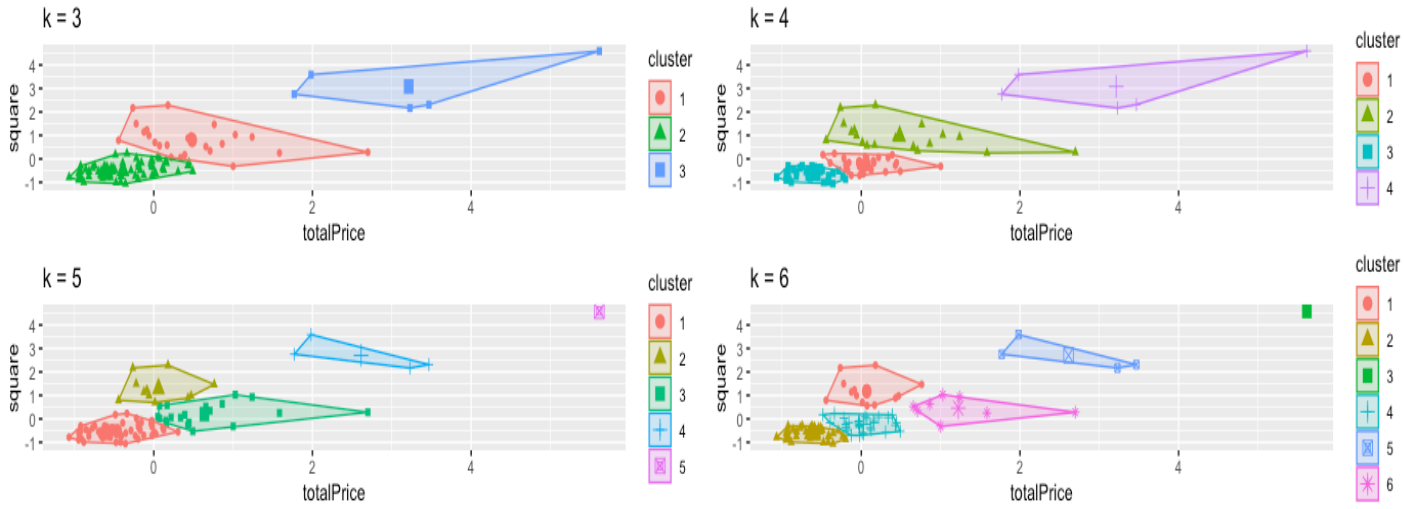
Methodology

In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms function by making data-driven predictions or decisions, through building a mathematical model from input data. This project will develop the models from the randomly select training set and check the prediction with the valid set. Each subset has 45,621 observations, which is large enough to conduct the proposed statistical model analysis with 26 variables.

The cluster analysis suggests the price of each house have a strong tendency to be grouped. Each house group which shares similar features have a significant small variation. The plot below shows Beijing's house price by location. (Lu 2019) It is clear that the location of the community may a have a big impact to the house price. In reality, the housing price also won't change too much in one community, because they have similar facilities, build year and type. Therefore, this project creates some subset according to their features, for example the newly built house close to subway and the house over 10 million and have more than 2 bathrooms.



Picture 1. Beijing housing price map



Picture 2. Cluster analysis result

The project will evaluate each model with its R square and MSE, RMSE. The R-squared represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is. The MSE represent

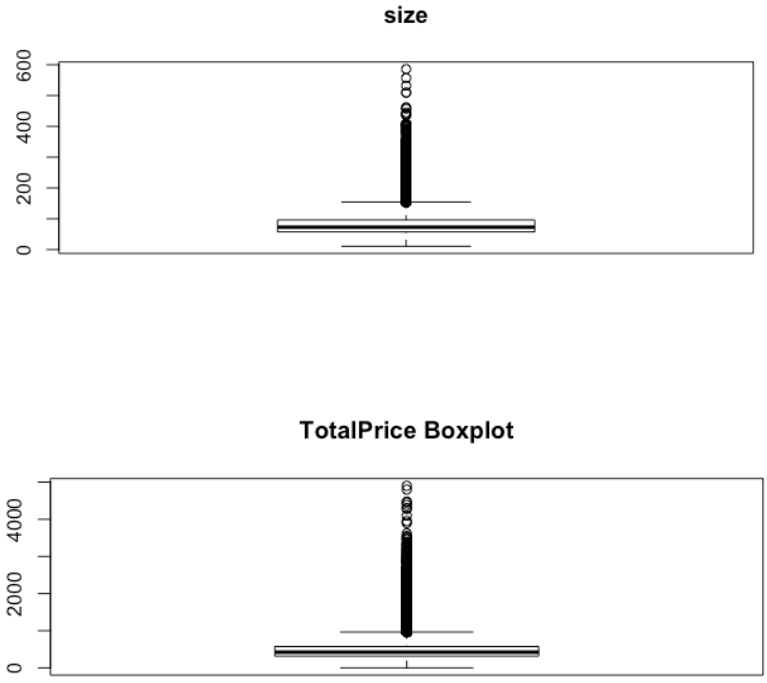
represents the difference between the original and predicted values extracted by squared the average difference over the data set. A lower MSE is better.

Model building

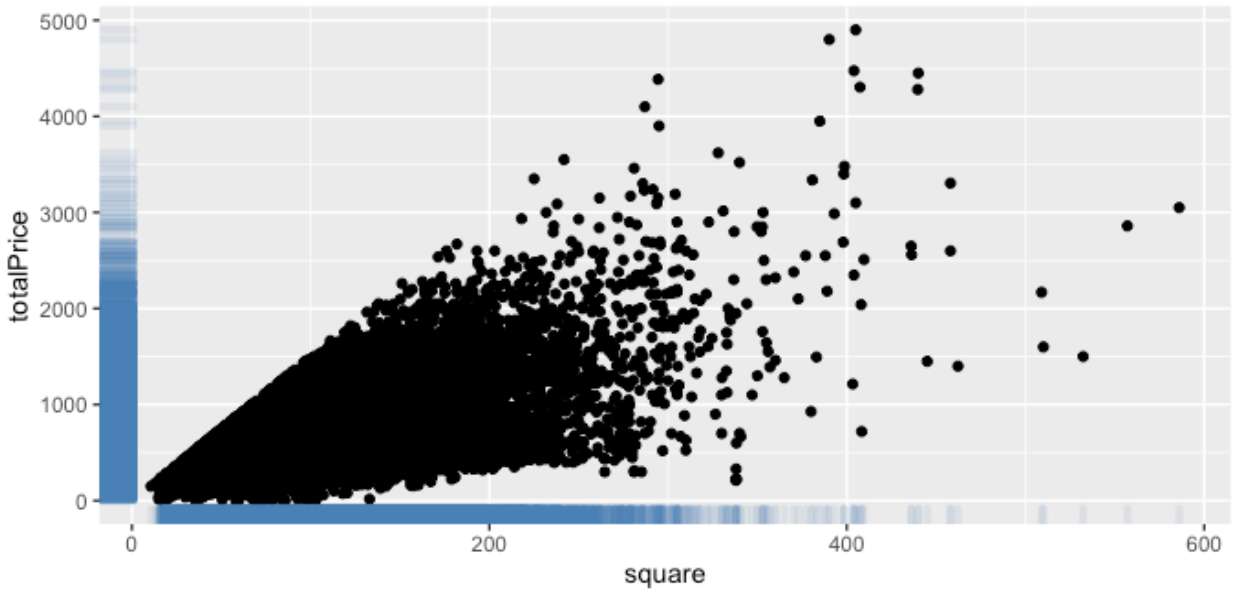
There are two main parts for model building. The first is to explore the dataset. It is important to get some knowledge of the data before conducting any model building. In this part, we will have a better understanding of the data by querying, data visualization, and reporting techniques.

The second part is building the model with multivariable regression, tree regression, and generalized boost regression.

The dependent variable, the total price of the house varies from 1 to 4,900 the median price of 422(thousand of RMB). The second variable people usually pay attention to is the size of the house, which is "square" as its column name in our dataset. This variable ranges from 10.7 to 586 square meters with a median of 72.65. Based on the observation, there is a positive relation between the total sale price of the house and the size of the house, which is following people's common sense.

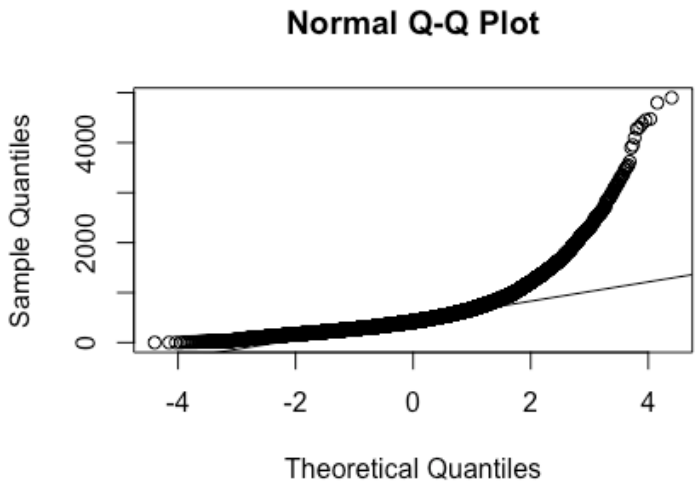


Picture 2. Boxplot of total price and the size



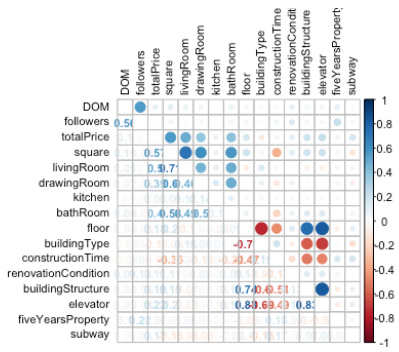
Picture 3. Relation between Price and size

At the Q-Q Plot the sample quartiles and the theoretical quantiles are plotted against each other. If they both came from the same distribution, we should see the points forming a line that is roughly straight. However, for the total price that are far away from the average price, the plot deviates heavily from the qq line. Especially for high prices a steep slope with a following sharp flattening out can be observed.



Picture 4. QQ-plot to check normality

The picture below shows the correlations between each variables in our dataset.



Picture 4. Correlation plot

Multivariable regression

Multivariable linear regression establishes the relationship between a dependent variable (i.e., The total price) and more than other independent variables.

$$Total\ price = b0 + b1*x1+b2*x2...+bn*xn$$

It is one of the most popular mathematic models for analyzing the real-world problem. In this particular project, the multivariable regression will help us to (i) identify each factor associated with the total price of the house, (ii) measure the marginal effect of each variable to the total price, and (iii) interpretation the regression and the forecasting to the public.

Tree regression

Tree regression is also called as the decision tree. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value. (Wu 2018) One of its advances is this model is very easy to interpretate. Audiences could find out the answer by following the tree's branch.

Generalized boosted regression

Generalized boosted regression is realized by the r package of gbm. It combines Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. Boosting is the process of iteratively adding basis functions in a greedy fashion so that each additional basis function further reduces the selected loss function. This implementation closely follows Friedman's Gradient Boosting Machine (Friedman 2001).

Result

Multivariable regression:

The multivariable regression suggests the variable of DOM (days on market), followers (The number of people followed the house on website), floor and five-year-porperty are **not** significant

to the model with a large p value. This result generally follows people's common sense because the DOM and follows only reflect people's attention, which may only have effects to how fast the house will be sold. The variable of Five-Year property reflects a local law enforced by Beijing Government, which requires buyer and seller pay more transaction tax and capital gain tax if the property's last recorded transaction happened with five years ago. This variable is also not significant to the total price of the house.

Based on the result, the model is:

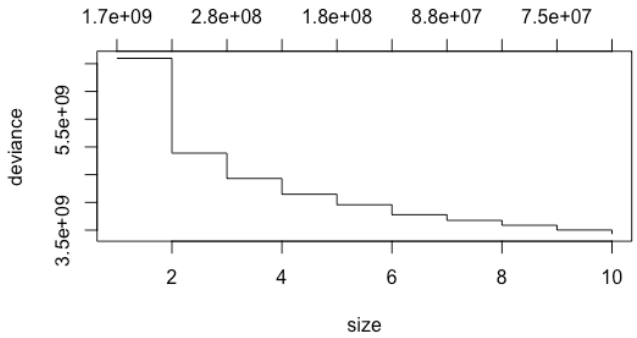
Total Price=
*-300.61+4.177*size+21.61*bedroom+10.87*livingroom+87.32*kitchen+33.54*bathroom+12.01* buildtype+3.242Time+6.843condition+9.074Structure+66.18elevator+92.58subway*

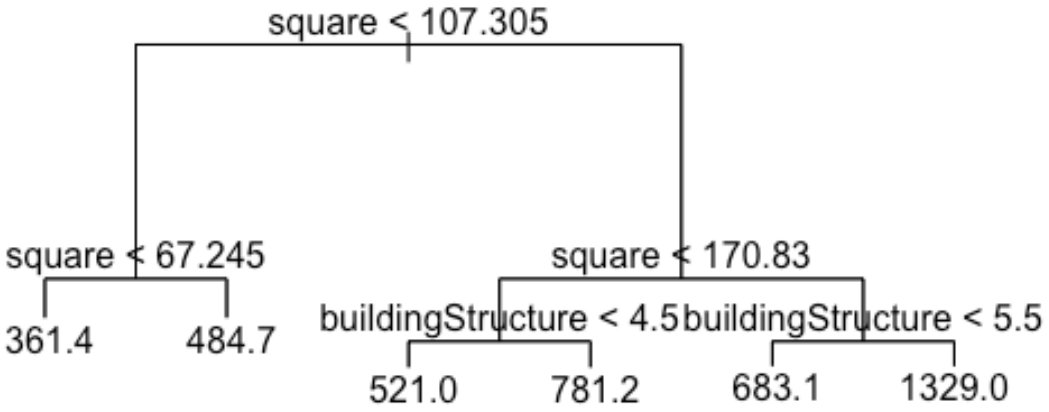
This regression includes all variables results an acceptable adjust r-square of 49.46, and mse of 36539.85. It is a good tool to make interpretation to the decision maker in real world, such as the predicted housing price will be 10.87(thousand) more if the house have one more living room with other parameters holds same or the price will increase 66.18 (thousand) is the apartment has an elevator if holding other parameters same.

According to the result, the top 4 significant variables are size, time and subway. The multivariable regression contains these 4 variables results a r-square of 48.59, which slightly lower that the regression of all variables

Tree regression

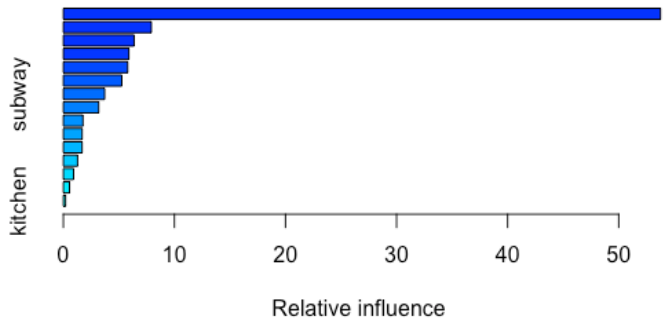
The picture below is the result of the tree regression. It shows the marginal deviance decrease is stable and small after the size of 6 branches. This indicate a decision tree regression model with 6 branches is good enough. The MSE of tree regression to all variable is 37556.88 while the six branch is 41410.15. The performance of tree regression is worse than the multivariable regression.





Picture 5. Decision tree

Third, the generalized boosted regression model results the best performance by output the MSE of 26528.66, which is 27.11% lower than the result from the multivariable regression. According to the result, the most important variable is the size of the house and the construction time of the house with relative influence rate of 5376 and 7.92. Interestingly, the third important variable is the days on the market, which shows a large p value in multivariable linear regression. The picture below is the result of the boosted regression model.



```
> summary(boost)
                                  var    rel.inf
square                         square 53.7613685
constructionTime       constructionTime  7.9226452
DOM                             DOM  6.3671151
floor                         floor  5.8993886
followers                 followers  5.7917883
buildingStructure     buildingStructure  5.2602985
subway                       subway  3.7170966
livingRoom                 livingRoom  3.1823050
elevator                     elevator  1.7610810
bathRoom                     bathRoom  1.6942129
renovationCondition renovationCondition  1.6899943
buildingType             buildingType  1.2787098
drawingRoom               drawingRoom  0.9191914
fiveYearsProperty     fiveYearsProperty  0.5570367
kitchen                       kitchen  0.1977683
```

Picture 6. GBM result

As aforementioned, the house price didn't follow the normal distribution, especially for the house in the high-end market. The house with similar features and characteristics also shows a strong tendency to be a group. The multivariable regression's performance is much better than the above three. The last model is to repeat the multivariable regression with different subgroup:

Condition1	**Condition2**	**R^2**	**MSE**	**Description**
constructionTime <10	Livingroom>4	0.5496	190392.3	Newly build large house
constructionTime <10	kitchen>1	0.7559	93091.97	New single family house(two kitchen)
constructionTime <10	Subway==1	0.6874	22135.3	New house with subway closeby
livingRoom ==1	constructionTime<5	0.9652	36.04376	Newly build One bedroom apartment
constructionTime<10	elevator==1	0.6588	18460.17	Newly build with elevator(tall new buildings)
DOM<100	Followers>100	0.5213	9797.021	Hot listing

Table 2. subset's performance

Conclusion

In conclusion, there are many factors could impact the house price in Beijing. The size of each house is the most straight forward variable. It is also could be observed that the categorical features with the largest effect on housing price predictive power are subway, building structure and type, and whether having the elevator.

Examining our final model coefficients, I observed that the categorical features with the largest effect on housing price predictive power are subway, elevator and building type and conditions.

References

Özbaş, Birnur, Özgün, Onur and Barlas, Yaman (2014) 'Modeling and Simulation of the Endogenous Dynamics of Housing Market Cycles' Journal of Artificial Societies and Social Simulation 17 (1) 19 <http://jasss.soc.surrey.ac.uk/17/1/19.html>. doi: 10.18564/jasss.2353

WHEATON, W. C. (1999). Real estate "cycles": some fundamentals. Real Estate Economics, 27(2), 209–230. [doi:10.1111/1540-6229.00772]

Lu, Guanpeng. “A Brief Analysis for BJ Housing Price & Prediction.” Kaggle, Kaggle, 5 Jan. 2020, www.kaggle.com/guanpeng/a-brief-analysis-for-bj-housing-price-prediction.

Sharma, Yanogya (10 August 2018). "How Lianjia leveraged Internet to evolve as $6 Bn online real estate company in China"

Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008-01-01). "Top 10 algorithms in data mining". Knowledge and Information Systems. 14 (1): 1–37. doi:10.1007/s10115-007-0114-2. ISSN 0219-3116. S2CID 2367747.

J.H. Friedman (2001). “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics 29(5):1189-1232.