

House Prices Prediction Using Machine Learning Techniques

#1 Dr Yamarthi Narasimha Rao, #2 Sravanthi Srinivas Addepalli

#1 Professor & HOD, #2 M.Tech., Scholar

Department of Computer Science and Technology, QIS College of Engineering and Technology, Ongole

Abstract:

In places like Bangalore, developing predictive models for home selling prices is still a difficult and time-consuming job. The price at which houses in major cities are being sold as in bangalore is reliant on a variety of interconnected variables. The area of the house where you live may have a significant impact on the pricing. Property, including the physical location as well as the features and services on the property. Due to the fact that there has been research done and an analysis done by taking into account the data set that is still available to the general public using a machine to depict the various housing options platform for hack athon there are nine different variables in this collection of information. According to the findings of this investigation, attempts have been made to develop a model that can predict assessing the pricing in light of the influencing variables. Some regression methods, such as are used in modeling experiments, such the least squares method of multivariate linear regression, the lasso, and the ridge models of regression, support vector regression, and methods for enhancing such as extreme gradient boost regression (xg) methods; boost). These models are used in the construction of prediction models and in the analysis of decide on the most effective design by doing a comparison an examination of the disparity in prediction error between the two models. It is hoped that by doing this, a predictive model may be built for assessing the pricing in light of the influencing variables.

1. Introduction

Machine learning algorithms are used in the modeling process. A computer learns from past data and applies that knowledge to make predictions about the future data. Predictive analysis is most often performed using a model called regression. So far, the suggested model has shown to be very accurate. In economics, it is useful to make predictions about the future. Business, finance, healthcare, e-commerce, and other related industries the arts, sports, and so on using a technique like this helps predict the future. There are many variables that influence home pricing [7]. In a big city potential homebuyers take into account places like bangalore there are a number of variables such as the location, size of the property, and distance to all of these types of amenities may be found in parks as well as most most significantly, the cost of a home. Multiple linear regression is a statistical technique that is widely used. The methods of statistical analysis used to determine the significance of target variable, as well as a number of independent variables. A lot of regression methods are used in the construction of a model for forecasting the price that takes into account various variables. According to the findings of this investigation, we've made an effort to develop a housing-price forecasting model.

Model for regression on data set that is still accessible public on the hackathon platform for machines. We had a discussion about it. There are five different prediction models: conventional least squares, models of regression using lasso and ridges, svr model, and xgboost a model that incorporates regression analysis. A side-by-side comparison with assessment measures are also important to consider. Once we've found something that works for both of us, we can go forward. A model for estimating the future monetary worth of a certain house property. Most of the time, property prices increase when it is necessary to compute the passage of time and the estimated monetary worth of that period. The value of this appraisal is needed when selling either while submitting an application for a loan or when availability of the property on the market. According to these evaluations, are decided by experts in the field. This technique, however, has the disadvantage that these appraisers may be skewed by their own interests while doing appraisals. Whether it's from buyers, sellers, or lenders as a result, we need a computer-based approach for making predictions that can forecast the value of the property without any prejudice. This first-time purchasers may benefit from an automated model. Consumers with less experience to determine if the affordability of real estate is exaggerated or exaggerated, depending on your perspective. Now, the price of a home is determined by a number of different factors. In the business world and in society as a whole. Nevertheless, in the past according to research, home prices are rising rapidly. Depending on the house's size and layout location [2], [3]. In addition, we examined a number of internal factors, including bedrooms, square footage of living space, and quality of construction for example, material) as well as external variables closeness to other developments in the area.) [4], [5]. After that, these parameter values were applied to two there are many algorithms for machine learning. As previously stated, we have a linear regression model and a support vector machine have been explored the price of the will be predicted using a home and assessed the differences in their results.

2. Literature Survey

Property value forecasting has grown in importance during the past two decades. The rise in property demand and the unpredictable nature of the economy have compelled academics to discover a method to accurately forecast real estate prices. As a result, finding all the minute variables that may influence property costs and creating a prediction model that takes them all into account is a difficult task for academics. A comprehensive understanding of real estate price appraisal is required when creating a predictive model. This issue has been studied by a large number of researchers, all of whom have published their findings.

L. Liu [11] used four regression techniques: linear regression, support vector machines (svm), and multiple regression. K-nearest neighbors (knn) and random number generator by combining forest regression with an ensemble approach prediction using knn and random forest techniques a home's asking price. According to the results of the ensemble analysis, applying pca on prices with a smallest inaccuracy of 0.0985 had no effect reduce the inaccuracy rate of the forecast. There have also been a number of investigations on the assembling of characteristics as well as methods for extracting them.

Different feature selection and feature extraction methods have been evaluated by wu, jiao yang [12]. Using support vector regression in addition to a few scientists neural networks have been created for the purpose of predicting home values linsombunchai made a comparison between hedonic pricing and predictive modeling for home values using an artificial neural network [13] the neural network model's r-squared value than the hedonic model and the root mean square error there was a significant decrease in the value of the neural network model. Hence according on their findings, artificial neural networks outperform traditional neural networks compares well to the hedonic model using the hedonic pricing model,

Cebula can accurately forecast the housing market. Savannah, georgia, costs are quite reasonable. The cost of logs, in dollars the value of homes has been shown to be significantly and positively the number of bathrooms and bedrooms is related to hearths, garages, levels, and a house's overall square footage the length of your feet liuguangyan uses a support vector machine chinese home prices will be predicted using support vector machines (svms). Between the years 1993 and 2002. It has been discovered that using the genetic algorithm, svm regression model hyper-parameters may be tuned. The the svm regression model's error scores were lower. Less than 4% [15].

A study conducted by tay and ho examined the accuracy of price predictions made by in predicting regression analysis and artificial neural network prices for an apartment are listed below. The researchers came to the conclusion that the neural when compared to a regression analysis model, the network model offers superior results with a 3.9 percent mean absolute error [16].

3. Methodology

Methodology is an explanation of how to accomplish something. A framework that has been established. It's made up of a variety of necessary milestones to meet in order to obtain the end goal we've collected a variety of information. Ideas for data mining and machine learning as follows figure 1 depicts a flowchart of the tasks shown in the diagram.

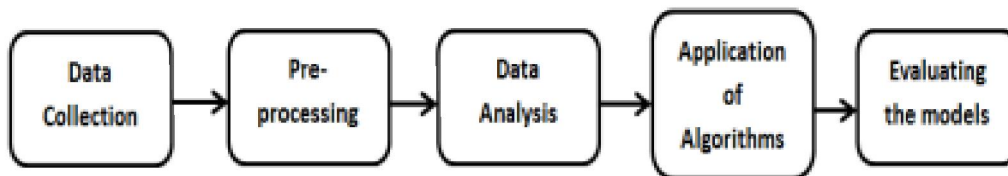


Figure 1: System Architecture

Collection of information

This project made use of an open-source dataset. Database provided by kaggleinc [22]. It has 3000 pieces. 80-parameter recordings with the potential of having an impact on the value of real estate however, just one of them stands out. Only 37 of the 80 possible parameters have been selected as being fixed. Can have an impact on housing costs measures like area are examples of overall quality measured in square meters as well as its interior and exterior finishes location, construction year, and the number of the number of bedrooms and bathrooms, the size of the garage, and the kind of swimming pool space, as well as vehicles that fit in garages the year when the property was first put on the market and the asking price sold. The selling price is a function of a number of different factors. Additional factors that aren't within your control. A few settings had been altered. Some of the numbers were scores,

while others were ratings. These in the end, the scores were quantified numerically. Table 1 provides a succinct overview of the most significant variables influencing the pricing of a product pertaining to the residence.

Parameters	Description	Datatype
OverallQual	Rates the overall material and finish of the house	Numerical
YearBuilt	Original construction date	Numerical
TotalBsmntSF	Total square feet of basement area	Numerical
GrLivArea	Above grade (ground) living area square feet	Numerical
FullBath	Full bathrooms above grade	Numerical
GarageCars	Size of garage in car capacity	Numerical
GarageArea	Size of garage in square feet	Numerical
WoodDeckSF	Wood deck area in square feet	Numerical
PoolArea	Pool area in square feet	Numerical
YrSold	Year Sold (YYYY)	Numerical
SalePrice (Dependent Variable)	Selling Price of the house	Numerical

Table 1
The Parameters

Preparation of data

It's a method for converting large amounts of unstructured data into something more usable. Systematized information that may be accessed and understood. It is concerned with the process of locating the missing and the dataset contains duplicate data. The whole collection of data is checked for nan and the result of the most recent experiment the value of nan will be removed from the input. As a result, homogeneity is achieved. In the collection of data. On the other hand, there was none in our study's data. Data were discovered to be missing, indicating that each record was formed the basis for the relevant feature values.

Analyzing the data

As a prerequisite to running any model on our data, we must see what our dataset's properties are. As a result, we'll require to conduct an analysis of our data and look at the various factors as well as the connections among them parameters. If there are any outliers, we may additionally determine that in the data we collected. Outliers arise as a result of a mistake in the experimental design, and they must be eliminated from further consideration. As shown by the dataset. Our research has shown that there is one or at least two extreme cases. The overall trend in the sale price has shifted downwards. A variety of variables grlivarea and totalbsmtsf are two examples of such areas. Have a straight line relationship with 'saleprice,' it seems the the selling is boosted by the home's overall quality and location the house's price has risen in tandem with it. Having said that, the overall level of quality and the quantity of restrooms are not linked. They're not connected in any way. The total floor space of the basement there is a connection between the ground living area and each other. In each of the total graphs, there is a single outlier. Subterranean space.

4. Results

The data in the next part is broken down into different algorithms that are put to use. We've taken everything into consideration. Take into account several performance indicators, such as correctness, r-squared value, and rms rmse, and rmsd are all measures of calculation of the median (mse). By using these specifications,

Table 2 compares the four different designs.

	Accuracy	R-Square	RMSE	MAE	MSE
LR	72.81%	0.987	8922	6118	79604145
SVR	67.81%	0.968	14101	76429	1.99E+08
Lasso	60.32%	0.81	34275	21058	1.7E+08
DT	84.64%	0.99	217	5.68	47184.93

Table 2 Results

We can see from the above table that the decision tree provides a greater r-squared value and a lower accuracy values that indicate errors. When evaluating different designs, we discover that the decision tree with the greatest degree of precision works well. 84.64 percent of the time, and lasso has the lowest accuracy of any weapon tested. 60.32 percent of the population. With rmse, decision tree generates almost no errors. 217-point value, and lasso has the poorest results with that there is an rmse of 34245. This leads us to the conclusion that the decision tree is too fitted the correctness of our dataset, although this problems may arise as a result of taking this into consideration. Take into account the many sounds that may be heard all around. Also we've split our data in half, 50/50, to find out prevent our side from becoming too fitted.



Figure 1

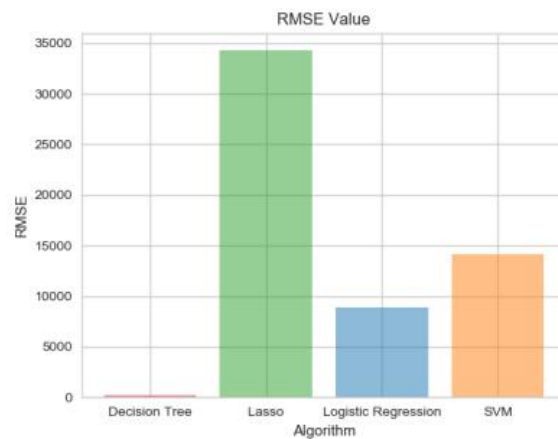


Figure 2

5. Conclusion

We utilized machine learning in this study. Use of computer algorithms to forecast changes in the value of homes as previously stated, we have outlined the step-by-step analysis process and establishing a link between the parameters. As a result, we are able to choose the variables that are not linked to one other in any way and are entirely separate in the course of nature. These capabilities were subsequently made available as input was fed through four different algorithms, and the result was a csv file. Include expected home price increases. As a result, we determined each model's performance by comparing it to the and compared them using various performance measures on the basis of these parameters. Decision tree was discovered to be useful in our research. Our data and provides the greatest degree of accuracy 84.64 percent of the population. When it comes to precision, lasso comes in last with a 60.32 percent accuracy rating. Support and logistic regression using a vector with 72.81 percent and 67.81 percent accuracy respectively as a result, we may say that the use of classifiers for the regression issue allows us to see how it is possible that classification may suit regression issue [21] effectively. We suggest that further effort be focused on a vast amount of data would provide a more accurate and complete picture. About the design of the model we've only taken on a handful of projects. Methods for machine learning that work classifiers, but there are a slew of new classifiers that need to be trained. And comprehend their prognostication for continual values are also an option. By reducing the number of mistakes this study's findings may aid in the creation of submissions for a number of different cities.

References

- [1] r. J. Shiller, "understanding recent trends in house prices and home ownership," national bureau of economic research, working paper 13553, oct. 2007. Doi: 10.3386/w13553. [online]. Available: <http://www.nber.org/papers/w13553>.
- [2] d. Belsley, e. Kuh, and r. Welsch, regression diagnostics: identifying influential data and source of collinearity. New york: john wiley, 1980.
- [3] j. R. Quinlan, "combining instance-based and model-based learning," morgan kaufmann, 1993, pp. 236–243.
- [4] s. C. Bourassa, e. Cantoni, and m. Hoesli, "predicting house prices with spatial dependence: a comparison of alternative methods," journal of real estate research, vol. 32, no. 2, pp. 139–160, 2010. [online] available:<http://econpapers.repec.org/repec:jre:issued:v:32:n:2:2010:p:139-160>.
- [5] s. C. Bourassa, e. Cantoni, and m. E. Hoesli, "spatial dependence, housing submarkets and house price prediction," eng, 330; 332/658, 2007, id: unige:5737.[online]. Available: [http:// archive - ouverte. Unige. Ch/unige:5737](http://archive-ouverte.unige.ch/unige:5737).
- [6] pow, nissan, emil janulewicz, and l. Liu. "applied machine learning project 4 prediction of real estate property prices in montréal." (2014).
- [7] limsombunchai, visit. "house price prediction: hedonic price model vs. Artificial neural network."new zealand agricultural and resource economics society conference. 2004.
- [8] park, byeonghwa, and jae kwon bae. "using machine learning algorithms for housing price prediction: the case of fairfax county, virginia housing data."expert systems with applications 42.6 (2015): 2928-2934.
- [9] bhuriya, dinesh, et al. "stock market predication using a linear regression." electronics, communication and aerospace technology (iceca), 2017 international conference of.vol. 2.ieee, 2017.
- [10] majumder, manna, and md anwar hussian. "forecasting of indian stock market index using artificial neural network."information science (2007): 98-105.
- [11] li, li, and kai-hsuan chu. "prediction of real estate price variation based on economic parameters." applied system innovation (icasi), 2017 international conference on.ieee, 2017.
- [12] hromada, eduard. "mapping of real estate prices using data mining techniques." procedia engineering 123 (2015): 233- 240.
- [13] razi, muhammad a., and kuriakoseathappilly. "a comparative predictive analysis of neural networks (nns), nonlinear regression and classification and regression tree (cart) models." expert systems with applications 29.1 (2005): 65-74.
- [14] wu, jiao yang. "housing price prediction using support vector regression." (2017).
- [15] pedregosa, fabian, et al. "scikit-learn: machine learning in python." journal of machine learning research 12.oct (2011): 2825-2830.
- [16] manning, richard l. "logit regressions with continuous dependent variables measured with error." applied economics letters 3.3 (1996): 183-184.
- [17] smola, alex j., and bernhard schölkopf. "a tutorial on support vector regression." statistics and computing 14.3 (2004): 199-222.
- [18] jaen, ruben d. "data mining: an empirical application in real estate valuation." flairs conference. 2002.

- [19] lim, wan teng, et al. "housing price prediction using neural networks." natural computation, fuzzy systems and knowledge discovery (inc-fskd), 2016 12th international conference on. iee, 2016.
- [20] manning, richard l. "logit regressions with continuous dependent variables measured with error." applied economics letters 3.3 (1996): 183-184.
- [21] torgo, luis, and joao gama. "regression using classification algorithms." intelligent data analysis 1.4 (1997): 275-292.
- [22] <https://www.kaggle.com/ohmets/feature-selection-forregression/data>