Original Article

HOUSE PRICE ESTIMATION BY USING DEEP LEARNING: A CASE STUDY

Jagbeer Singh[1], Shubham Singh[1], Utkarsh Chuhan[1], Prabhakar Vats[1]

[1] Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, U.P., India.

DOI: 10.47750/pnr.2022.13.S07.476

Abstract

Every year, the cost of housing increases, which generates the needs of a system that predict future house prices? A developer can estimate the promotion cost of a residence with use of a house price forecast. It also assists clients in deciding on the apparent timing of a property purchase. The cost of a home is influenced by a number of factors, and it mainly depends on physical surroundings, type of home, dimensions, area or locality, etc. House prices vary by location and community. Real estate prices can be predicted in a number of ways. One of the efficient methods is to use regression techniques. Regression is considered a reliable method for determining variables that exert influence on topics of significance. Random forests are extremely reliable and accurate against over-fitting. The process of running regression ensures that you can determine the most important factors, negligible factors, and how much each of these things influences the other. The primary intention is to implement the most cutting-edge forecasting method.

Keywords: House-prices, Regression, Price-prediction, Lasso-regression.

Introduction

Among the businesses in which most people are interested in this era of globalization is investment. There are several products that are often used as investments, such as gold, bonds, stocks and real estate [1]. Real estate developers must thoughtfully consider and use the appropriate methodology for setting home prices. Due to the fact that the value of properties never stop rising and hardly ever decline in the short or long run.[2].

Predictive analytics is among one of several approaches that can be used to determine home prices. The challenge in this is to get as near as possible to the results based on the built model. Housing prices are mainly dependent on region, dimensions, housing type, city, country, tax system, business cycle, demographic movements, interest rates, and numerous other factors that can affect supply and demand. Local House Price Forecast has many useful regression algorithms that you can use. Regression analysis is a set of statistical operations that is used to estimate relationship between variables. It features numerous approaches for modeling and analyzing multiple variables, where the main aim is to establish a relation between the dependent variable and one or more independent variables.

Moreover, regression analysis is about understanding how the independent variables is altered while the other is kept constant, and how that affects the typical value of the dependent variable. The main edge of using regression-based forecasting techniques is the adoption of research and investigation to figure out what will unfold over the next three months or even in the future. For small-scale business owners, regression-based projections can provide insight into how tax increases will affect consumer spending and changes in local economies.

Regression and Prediction techniques give a scientific approach to small-scale business management by transforming enormous amounts of raw data into information that may be used. The dataset obtained has training

sets containing 79 attributes (i.e. traits, variables or predictors) and 1460 houses (i.e. observations) with the sales price of each house. Test set contains 1459 houses with the same 79 attributes, but no sales price. Because this is the target variable. In this research work, the proposed house price forecast is based on the technique Gradient Boosting regressor.

Literature Survey

Using Hedonic Pricing model a survey was conducted on real estate prices of Savannah city, Georgia[3].The papers' information includes2,888 single-own circle of homes for the period between 2000 and 2005. Residential log price included the number of restroom, bedrooms, fireside, and garage space, floors, and square feet homes. Additionally, the paper includes three dummy variables for the month of May, June, and July to take into consideration the seasonal influences on housing prices. For an instance if house was sold in the month of May, the value of variable May is set to 1, if not it is set to 0. The other variables June and July are also created in the similar manner. The study found that the logarithm of home sales prices was convincing and related in a good way with May and July, and was insignificant for June. This means that homes that are sold in May or July tend to have higher prices.

The social and economic impacts of housing development in the Scottish countryside are examined. Investing in housing finance has direct and indirect effects on the economy. employment, GDP, productivity and many other important factors Impact of housing finance investment. The study found that housing is a key indicator of increasing wealth in the country. It was then concluded that the purpose of Scottish Housing Policy was to improve housing quality standards and increase investment in the old housing sector.

Study [8] revealed that each of the five variables in the simple regression model (floor, rent, heating system, seismic zone and land price) had a significant result on the values of the dependent variables when the significance level was accepted as 0.05. We found that land and rental prices have the greatest impact on house prices whereas the remaining variables floor, heating system and seismic zone are found to be insignificant according to the study, but the sample size may vary accordingly. As the sample size increases, the regression model is again recommended for further study. The application of multiple regression analysis on the housing dataset provided a good example of the strategic application of mathematical tools to explain or model changes in housing prices, support analysis, and aid decision-making in real estate investment. House price volatility provides a good example of the strategic application of mathematical tools for support [5] (2010) SVM regression is used to estimate the 1993–2002 Estimates China's house prices during the year. A specific district of Tangshan City between 2000 and 2002. In this paper, the use of a genetic algorithm to adjust hyper-parameters of the Support Vector Machine model. His SVM regression model's error levels for both China and the one district he studied in Tangshan City are both under 4%.This indicates how accurately his SVM regression model predicts Chinese housing prices. For residential market in Singapore, a decision tree model (2006) was used to study the price impact of properties [6]The paper concluded that owners of two- to four-bedroom dwellings were more concerned with basic dwelling characteristics such include the type of model and owners' ages of dwellings with five or more bedrooms. I'm here. Additionally, Executive Home owners are more concerned with service characteristics such as neighborhood locations and recreational facilities than with basic residential characteristics.

In a 2014 study [7], relationships between different house characteristics and residential property asking prices were developed and analyzed with both Using ordinary least squares, we can do basic linear regression and multiple linear regression. It was finished. For simple linear regression, the living area served as the explanatory variable, and for multiple linear regression, the plot area, the number of bedrooms, the building year, and additional explanatory factors were included. The findings of multiple linear regression reveal the bias brought on by the 4484 Simple Linear Regression's exclusion of critical components. The size of the house, as opposed to the size of the garage, was shown to be the component that most significantly affected the price of a residential property.

Numerous prior studies have found legit evidences to support compelling correlations between house prices and several aspects of the economy, including the labor market, interest rates, building prices, and income [8] [9][10].

Method and Materials

Various types of regression techniques are available to make predictions [11] [20]. All these techniques are principally influenced by three indicators (how many independent variables are there, the type of dependent variable, and the regression line's shape) as shown in Figure 1.

Various Algorithms used for the purpose of predicting Housing-prices are listed below

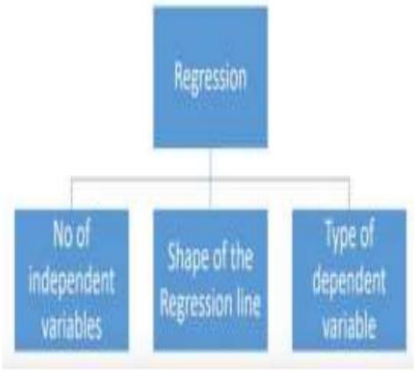


Fig. 1: metrics of regression

Hedonic Pricing Model

According to hedonic pricing theory, a commodity like a home may be thought of as a group of distinct parts or characteristics [12]. commonly used to gauge real estate costs. Elements inside a house (for instance, the number of bedrooms, baths, and other rooms) and external elements are combined in a hedonic price model (neighborhood walkability rating, public school rating, etc.). I respect his morals. Regression models may be used to create hedonistic pricing. For predicting the price, a regression model is displayed in Equation 1.

$y = a.\ x1 + b.\ x2 + \cdots + n.\ x1$

Where y be the anticipated price as well as x1, x2, and xi are characteristics of a house, respectively While a, b,... n denote the correlation coefficients of each factor used to calculate home prices. While utilising the hedonic methodology to account for attribute differences in a model used to determine home prices is allowed, doing so is typically impractical because housing markets vary greatly by region. Therefore, Incorporating geographic information or a location component into a model that takes into account variations in the level of real estate prices seems more logical.

Artificial Neural Network Model

Building hedonic pricing models is akin to using neural network models. However, a data set must first be used to train the neural network [13][21]]. The model generates an output (estimated house price) for a given input. Following that, the model contrasts the output values with the observed output (actual house prices). The total mean squared error serves as a measure of the values' correctness, and back-propagation is used to attempt to lower the forecast error. By changing the connection weight, this is accomplished. The network's performance could be affect by both the number of hidden-layers and the nodes in each one. [14]. Iterative testing is applied to discover the finest ANN model [22][23].

Proposed Methodology

There are two different data sets, a test data set and a training data set. Both contain numerous variables in the form of features that describe the home. The training data set contains 1460 observations for which house sales

prices are provided. A predictive model is created based on this data. The test dataset contains 1459 observations that predict a selling price of. The quality and quantity of the property's numerous physical attributes are the focus of 80 variables. The majority of the variables are simply the kind of details that a regular homebuyer needs to be aware of while considering a property[24][25].

This study is based on house price data from the Ames Housing dataset. The dataset includes several factors that are not inversely correlated with home prices. B. The functions "Date", "Long", and "Latitude"—which, respectively, stand for the date the home was sold and the longitude and latitude of the house—should be altered or omitted. Start by figuring out the building's age using the "Date" (the day the home was sold) and the "Year Built" (the year the house was built). Create a new binary feature to indicate if the home has been renovated using the Year Renovated feature (which indicates the year the property was renovated). Although zip codes do not directly correlate to prices, they may nonetheless provide valuable information regarding home values. As a result, it is handled as a category characteristic. Then the following features are removed: 'id', 'date', 'year built', 'latitude', 'long', 'date year', and 'year renamed'**.**

Gradient Boosting Algorithm

Gradient boosting one of the most widely used machine learning techniques for tabular datasets. It is enough capable of finding non-linear relationships between model targets and features, and simple to use to deal with missing values,[26].

Gradient-Boosting Algorithm

1. Initialize model with a constant value:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

2. for $m = 1\ to\ M$:

2.1 Compute residuals

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} for\ i = 1, \ldots, n$$

2.2 Train regression tree with features against x and r create terminal node: R_{jm} for $j = 1, \ldots, J_m$

2.3 Compute $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)\ for\ j = 1, \ldots, J_m$

2.4 Update the model: $F_m(x) = F_{m-1}(x) + v \sum^{J_m} \gamma_{jm} 1(x \in R_{jm})$

outliers, and high-cardinality categorical values in features without special handling. Some popular libraries, such as XGBoost and LIGHT GBM, allow you to construct bare-bones gradient-boosted trees without knowing the details of the algorithm, while others such as hyper-parameter optimization and loss-function fitting can be used. Before beginning, you'll need to understand how it functions. Improve model quality.

Random Forest

Classification plays a big role in machine learning. Which class, also known as group, does an observation fall under? In a wide range of commercial applications, the capacity to classify observations properly is incredibly useful. To predict whether a certain customer will purchase a goods or not, or whether a particular credit will default. Numerous classification techniques, including Logistic Regression, Support Vector Machines, Naive

Bayes Classifiers, and Decision Trees, are available in data science. A random forest classifier, on the other hand, sits atop the classification hierarchy.

House Price Affecting Factors

Real estate values are influenced by a number of things. According to research [16], there are three key categories of elements that influence home prices: location, concept, and physical condition. Physical qualities are those aspects of a home that can be felt by a person's senses. These physical qualities can be the dimensions of the residence whole number of bed-rooms, whether the area kitchen would be normal or modular, if garage is there or not, if a garden is present or not, etc. Developers can entice potential purchasers with ideas like sites and structures, the age of housing [17], and concepts. An elite setting, a healthy and green environment, or a minimalist house are a few examples. When establishing a home's pricing, location is crucial. Since a property's location often affects how much it costs [18].Moreover, the Place is identified convenient if it has easy access to public amenities such as schools, grocery shops, banks, hospitals and health facilities, as well as family recreational places like shopping centers, food courts, and also locations with stunning sight [19], [20].

Working Model

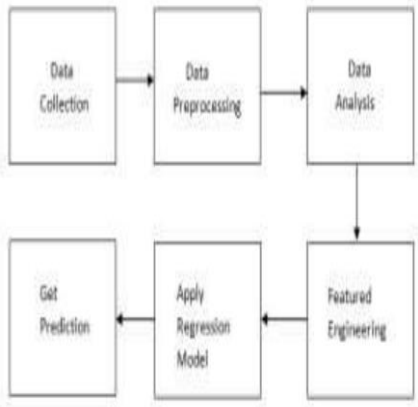


Fig. 2: Steps involved for prediction

a) Reading data

At this stage the data is read. Next, we need to concatenate the training data with the test data. This happens mainly due to the presence of text variables. These will later be replaced with dummy variables. If you treat the training and test sets separately, each has a different number of dummy variables, which can break the predictions.

b) Data Preprocessing

It is the process of turning unstructured, complicated data into knowledge that can be systematically understood. Finding duplicate or missing data in datasets is a part of this procedure. The whole dataset is examined for the presence of Na, as well as any further observations that include Na are eliminated. Therefore, the dataset now has consistency. Last but not least, we must separate the data into training and test data.

c) Data Analysis

We must identify the characteristics of a dataset before we can apply the model to it. As a result, it is essential to evaluate the dataset and investigate different factors as well as their correlations. Additionally, we can identify outliers in the dataset. Outliers should be removed from the dataset since they are the result of experimental error.

d) Feature Engineering

Manipulating features (variables or predictors) is one of the most important steps in creating a model. It only becomes apparent when the function is manipulated in some way. Below are just a few examples of the 's capabilities. If House was built in a different year than it was renovated, renovation could increase the value of the property.

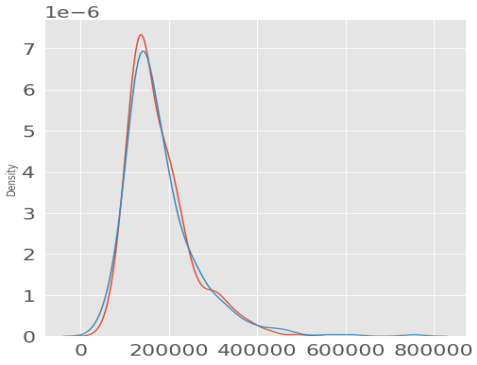
Seasonality (Category): Combined month sold and year sold. More homes were sold during the summer, but this is likely to change over the years, especially during the period when his 4,444 homes were sold, coinciding with the 4,444home crash. I was doing it.

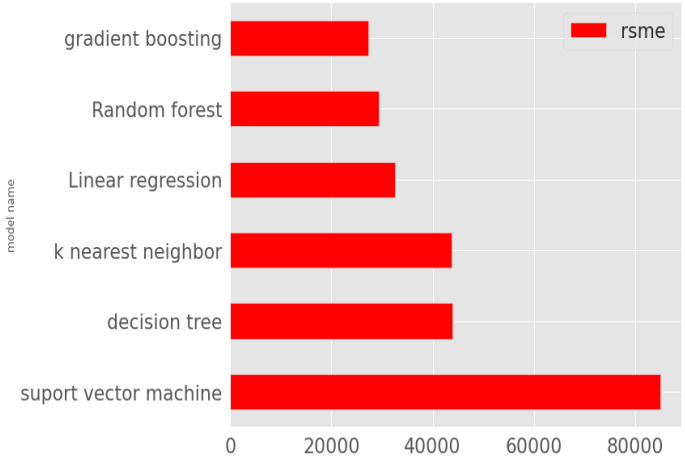
New Homes (Category):Yes or No if sale year is same as construction year. If the residence was sold in the same year as its construction, it would likely be in high demand and the would sell for a higher price.

Gross Area (Continuous): The sum of all variables representing the area of different parts of the building. There are many variables related to the area of different sides of each house. A total area of square meters can be expected to have a consequential result on the selling price.

e) Modeling

The process of choosing a collection of statistical models is known as model selection, and it involves fusing current data with previous knowledge. Covariates are either included or excluded when developing a model, and the choice of how to incorporate covariates in the design matrix of any model is dependent on both prior hypotheses and data. GRADIENT BOOSTING REGRESSOR is a regression analysis technique that does both variable selection and regularization to increase predicted accuracy and interpretability of the final statistical model.





Conclusion

In this article, we implemented the GRADIENT BOOSTING REGRESSOR regression method to predict the price of homes. A step-by-step procedure for analyzing datasets and finding correlations between parameters is described. Thus, essentially uncorrelated and independent parameters could be selected and this set of features was given as input 4484. To improve the prediction accurate we can perform variable selection and regularization both together.

References

[1] R. M. A. van der Schaar, Analysis of Indonesian Property Market; Overview and Foreign Ownership,‖ Investment Indonesian. 2015.

[2] Y. Feng and K. Jones, Comparing multilevel modelling and artificial neural networks in house price prediction,‖ 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr. Knowl. Serv., pp. 108–114, 2015.

[3] Rochard J. Cebula. "The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District". In: The Review of Regional Studies 39.1 (2009), pp. 9–22.

[4] [Gang-Zhi Fan, SeowEng Ong, and HianChye Koh. "Determinants of House Price: A Decision Tree Approach". In: Urban Studies 43.12 (2006)

[5] Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan. "Housing price based on genetic algorithm and support vector machine". In: Expert Systems with Applications 38 (2011), pp. 3383–3386.

[6] Eric Slone, Haitian Sun, Po-Hsiang Wang, (2014), "Market Prices of Houses in Atlanta", from https://smartech.gatech.edu/bitstream/handle/1853/51632/Market%20Prices%20of%20Houses%20in%20Atlanta.pdf

[7] P. Linneman, An empirical test of the efficiency of the housing market‖. Journal of Urban Economics 20(1986): 140-154, 1986.

[8] J.M. Quigley, Real estate prices and economic cycles‖. International Real Estate Reviews 2: 1-20. 1999.

[9] K.Tsatasaronis, & H. Zhu, What drives housing price dynamics: Cross-country evidence?‖ BIS Quarterly Review of March.

[10]Torgo, Luis, and Joao Gama. "Regression using classification algorithms." Intelligent Data Analysis 1.4 (1997): 275-2.

[11] EzgiCandas, Seda BagdatliKalkan and Tahsin Yomralioglu, (2015), "Determining the Factors Affecting Housing Prices", FIG Working Week 2015, Sofia, Bulgaria, 17 - 21 May 2015.

[12] 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.

[13]Lenk M. M., Worzala E. M. and A. Silva, 1997, "Hightech Valuation: Should Artificial Neural Networks Bypass The Human Valuer?", Journal of Property Valuation & Investment, 15(1): 8 – 26. [14] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

[15] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, Factors influencing the price of housing in Indonesia,‖ Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.

[16]V.Limsombunchai, House price prediction: Hedonic price model vs. artificial neural network,‖ Am. J. …, 2004.

[17]D. X. Zhu and K. L. Wei, The Land Prices and Housing Prices Empirical Research Based on Panel Data of eleven Provinces as well as municipality in Eastern China,‖ Int. Conf. Manag. Science. Engineering. number 2009, pp. 2118–2123, 2013.

[18]S.Kisilevich, D. Keim, and L. Rokach, ―A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers' context,‖ Decis. Support Syst., vol. 54, no. 2, pp. 1119– 1133, 2013.

[19]C. Y. Jim and W. Y. Chen, ―Value of scenic views: Hedonic assessment of private housing in Hong Kong,‖ Landsc. Urban Plan., vol. 91, no. 4, pp. 226–234, 2009.

[20]Narayan, Vipul, and A. K. Daniel. "CHOP: Maximum Coverage Optimization and Resolve Hole Healing Problem using Sleep and Wake-up Technique for WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.2 (2022): 159-178.

[21]Narayan, Vipul, and A. K. Daniel. "Design Consideration and Issues in Wireless Sensor Network Deployment."." Invertis Journal of Science & Technology (2020): 101.

[22] Awasthi, Shashank, et al. "A Comparative Study of Various CAPTCHA Methods for Securing Web Pages." 2019 International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019.

[23]Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." (2022).

[24Narayan, Vipul, et al. "E-Commerce recommendation method based on collaborative filtering technology." International Journal of Current Engineering and Technology 7.3 (2017): 974-982.

[25] Pramanik, Sabyasachi, et al. "A Novel Approach Using Steganography and Cryptography in Business Intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.

[26] Irfan, Daniyal, et al. "Prediction of Quality Food Sale in Mart Using the AI-Based TOR Method." Journal of Food Quality 2022 (2022).