Pratibodh
A Journal for Engineering
A free and Open Access Journal
Homepage: https://pratibodh.org

Data-Driven House Price Prediction Models

Khushi Sharma[1], Manu Garg[2], Ishita Goyal[3], Ms. Geerija Lawania[4]

Department of Artificial Intelligence & Data Science, Department of Artificial Intelligence & Data Science Jaipur Engineering College & Research Centre, Jaipur Engineering College & Research Centre

1kaushalyadav.ai24@jecrc.ac.in, 2geerijalavania.cse@jecrc.ac.in

Abstract

The real estate market is outstanding the most important thing is the price, which is always changing. that one of the main areas of application of mechanical ideas learn how to increase and predict high costs accuracy. The purpose of this work is market value of real estate. This system will help you find the property's starting price based on geographic variables. By breaking past markets patterns and value ranges and future advances expected future costs. The meaning of this test is predicting real estate prices using decision trees regressor. Helps customers invest resources in legacy without contacting a broker. The result research has shown that decision tree regressors give the following results: Accuracy 89%.

Article Status

Keywords: House Price Prediction, Machine Learning Model, Support Vector Regression, Artificial Neural Network

Available online :

1. Introduction

Any organization in today's real estate industry the company is doing well to remain competitive advantage over other competitors. There is a demand
simplify the process for the public bring you the best results. This article proposes a system predict real estate prices using regression engines if you want to sell a learning algorithmic, you need to be aware of what sticker price you need to put on it. Additionally, price calculations provide more accurate results.
Measuring instrument!

This regression model is not just built to:
Predict the price of a house for sale the same goes for houses under construction. Regression is a machine learning machine encourage raising expectations by assuming something current Measurable Information – Connections between target parameter and other parameters independent parameters. According to this definition, a house is the cost depends on the parameters, e.g., rooms, living areas, areas etc. You may be able to find a way to do that using a fake using these parameters, you can calculate the valuation of your home. Region of the specified country. The target feature of this proposed model is price property and independent equipment a: No. number of bedrooms, number of bathrooms, carpet area, Exclusive area, number of floors, building age, postal code, Latitude and longitude of the facility. Unlike those of the above functions that are usually required to predict real estate prices, we have included two more features - Air quality and crime rate. Features make a valuable contribution to forecasting the high value of these properties increases real estate prices this will lead to a decline in real estate prices.

All implementations are done using Python programming language. for the construction of Predictive model using decision tree regressor
"Scikit-learn" machine learning library. grid search
CV helps find the optimal maximum depth value for building decision trees. According to the trained model, once complete, it will be integrated into Flask's user interface.
(Python framework).

2. Methodology

This project used many machines learning algorithms, including: B. Linear Regression, Random Forest Regression, and Cat Boost
Regressors, SVR, KNN, XGB regressors, AdaBoost regressors for predicting real estate prices. 80% of information comes from known datasets
The remaining 20% of the information is used for testing purposes. This work contains several elements technologies such as transformation techniques, reduction techniques, and the search for new connections. Predicting real estate prices has a lot to explore and requires knowledge of machine learning. Housing prices are generally fixed; considering various variables. They call these elements concept, intensity, and placement. Also consider your physical condition
That includes no. the size of the room, the dimensions of the property, the age of the building, the size of the garage and kitchen. This project used many machine learning regression algorithms, including: B. Linear regression, decision trees, K-means, and random forest. Many factors influence real estate prices, including physical characteristics, location, and economics. factor. We consider RMSE as a performance matrix applied to

different datasets and these algorithms then determine the most accurate model that predicts better outcomes.

3. IMPLEMENTATION

3.1 Exploring your data

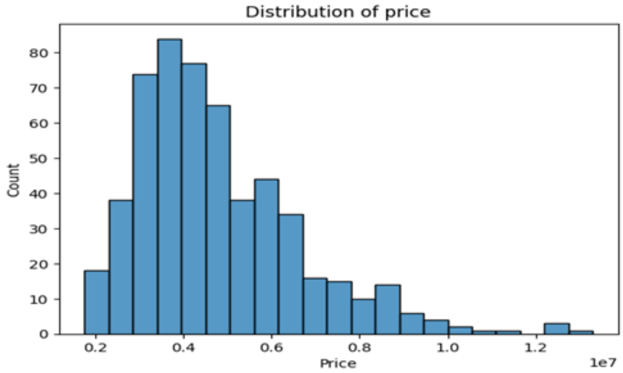
All tuples in the dataset define suburbs or cities of Boston. This data was collected by Boston SMSA (Standard). (Metropolitan Statistical Area) in the 1970s. Below are the attributes extracted from UCI MLR1.CRIM: Crime rate per capita by city

- Crime: Per capita crime rate of the city.
- ZN: Residential property portion designated for properties over 25,000 square feet.
- INDUS: Percentage of non-retail space per capita.
- CHAS: Also known as Charles River dummy variable. Its value is 1 if the area is adjacent to a river, 0 otherwise.
- NOX: The concentration of the toxic gas nitric oxide in the area. It is measured in parts in 10 million.
- RM: Apartment consisting of an average number of rooms.
- Age: Percentage of apartments built before 1940.
- DIS: How close is it to the 5th nearest employment office?
- RAD: Radial road accessibility index.
- Taxes: Total property tax rate per $10,000.
- PTRATIO: Ratio of teachers to city residents.
- B: Calculated using the formula: 1000(Bk−0.63)2. where Bk is the percentage of black people in each city.
- LSTAT: Proportion of population with low status individuals.
- Price: Average price of a home (in thousands of dollars)

We can clearly see that our attributes are fused with many entities

3.2 Data preprocessing

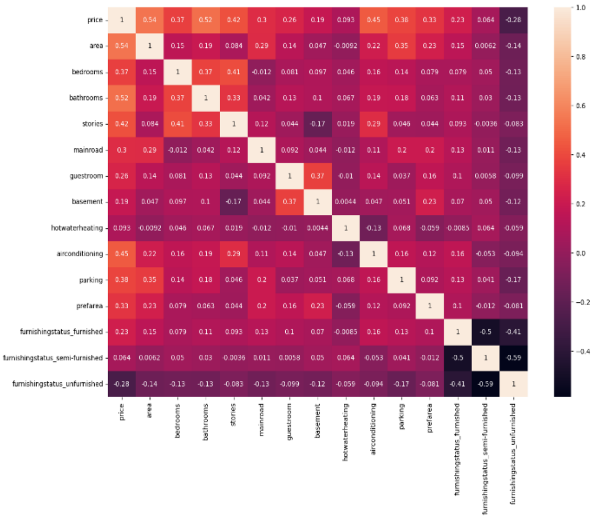
- Check for NULL values: Missing elements or NULL values are defined as data that is not saved or whose variables are missing. Dataset provided. There are several reasons why certain values are missing from the information. Some descriptions are outdated Improper maintenance can corrupt the data, and for some reason observations may not be recorded in a sufficiently accurate field. For some reason, errors may occur in recording values due to human error or the user not providing a value. Intentionally.
- Check if the data is normally distributed. Since the normal distribution is one of the important concepts in statistics, It can also be considered the backbone of data distribution using machine learning. Knowledgeable mathematicians must figure out the distribution after using a linear model, it should perform well when the information is normally distributed and there is a central limit theorems and exploratory data analysis are considered. If it's a normal distribution, the distribution may include that assumption. It can be given entirely by two parameters. These are means and therefore variances. For mean and variance you can access any date on the data curve if you know the value.



- Check for outliers: From a machine learning perspective, an outlier is defined as a point that is far away from all other points. The above statement indicates that it is an outlier, such as an outsider or someone far removed from the gang. a bit in statistics, an outlier is defined as something that has an underlying behavior that differs from the rest of the data. Alternatively, an outlier may be information that stands apart from all others. There's no reason to confuse this although this statement refers to unbalanced datasets, there may be little similarity within the definitions. I won't go into detail about this lots of details
- Check for correlations between attributes: Data correlation: This could be due to understanding connections or dependencies between many variables or attributes of a dataset. With the help of correlations, we can gain some insights. 1 or more an attribute depends on another attribute, or a reason for additional attributes. one or more related attributes remaining attributes.

a) Positive correlation: This means that when characteristic X decreases, characteristic Y also decreases or characteristic Y increases feature X also increases. Both features move correspondingly and there is a linear relationship between them.

b) Negative correlation: means that when feature X decreases, feature Y must increase and vice versa. c) No correlation: There is no relationship between these two attributes.



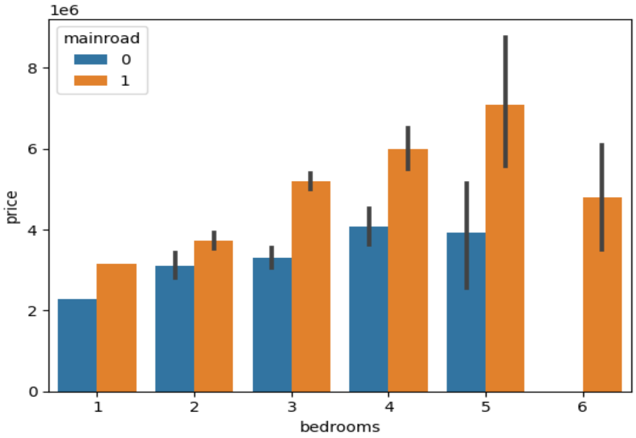
Figure

3.3 Pre-processing method

- Robust Scaler: Features are scaled using mathematical statistics so that outliers are felt to be robust. Robust scaler technology removes the median and shift the data consistently to the displacement range. IQR stands for interquartile range between the 1st and 3rd quartile. Scaling and centering are done independently of each other ability to calculate statistics on sample data in the training set. The median and interquartile range are stored as follows this transformation technique can be used with future data.
- Quantile Transformer: Transform features using quantiles information. Quantile transformation technique transforms the attributes to act in according to a similar distribution. So, for a given feature, this transformer tends to show up the foremost and recent values. It also minimizes the effect of the outliers: and so, can be therefore a steady preprocessing method. Independently, for each feature the transformation is to be applied. At first, some approximated value of the function distributed cumulatively of a feature is selected to map the first values to some similar distribution. Then the values obtained are mapped to the output required distribution by the associated quantile method.
- Yeo - Johnson Transformer: Apply an influence transform feature wise makes the information more like Gaussian. The group of parametric, monotonic transformations only include the power transformer techniques which are put together to form data more like Gaussian. This is more kind of useful while modeling the issues and problems associated with hetero-secede elasticity or other situations where desired output should be distributed normally. At present, the power transformation techniques assist the box cox transformation and hence the Yeo-Johnson transformer. The parameter which is optimal for making the variance steady and skewness has to be reduced is estimated through likelihood in most of the cases

3.4 Comparison of 3 transformation techniques

After contrast of three transformation strategies skew distribution facts. We see that the Yeo-Johnson transformer yields skew distribution is near 0. With this we are able to verify that the Yeo-Johnson transformer top-rated answer among the three transformation strategies and the facts converted via this transformer drives to higher predictions.

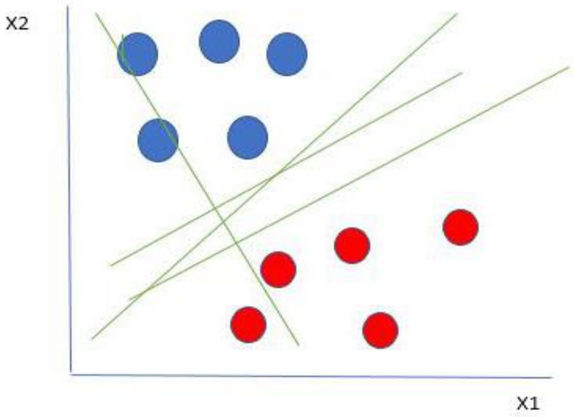


Figure

3.5 Explore different ML models.

- Linear Regression: Supervised type of machine learning is supported by linear regression algorithm. it accomplishes that regression task. Regression techniques model target/output predictions based on independent variables. This is the best an algorithm for identifying relationships between various attributes and predictions. Many regression models differ it provides good results because it supports the nature of the relationship between dependent variable and the required independent variable. Number of independent variables used. This regression algorithm takes on the task of predicting variable values. (v2) supported specified variables (v*). Therefore, in this model we find a linear relationship between v* (input) and v(output). Therefore, it is also called statistical regression.
- K-Nearest Neighbors (KNN): K-Nearest Neighbors algorithm is one of the most popular machine learning algorithms. Algorithms that support supervised learning categories. This algorithm assumes that the data is current and that the data is homogeneous. Review the available data and assign the most recent data to the category that is most similar to one of the available categories. This algorithm stores all prefetched data and consistently separates new information. It helps if it's fresh when displaying data, it is often easy to categorize the data into compatible classes.
- Support Vector Regression (SVR): Support Vector Regression is one of the supervised learning algorithms. Ordinary discrete value prediction. This model uses the same principles as the SVM principle. essential idea behind this SVR is the search for the most

effective adjustment line. In SVR, the most effective fitting line is the existing hyperplane. Best score. Machine learning regression models attempt to reduce the error between predicted values. Importantly, the model attempts to fit the simplest line around the threshold. The edge value is the distance from. Hyperplane up to the boundary. The fitting time complexity of this model depends approximately quadratically on the amount of sample acquired. This makes it difficult to scale to available datasets containing more than 10,000 sample pairs



Random Forest Regressor (RFR): Decision trees can be used for both regression and classification tasks. you visually its name comes from the tree-like river. Regression begins at the base of the tree and continues down the branches. Variable results up to leaf nodes are also supported and displayed. Forest can estimate this assign, average, and use decision trees with different labels for many subsamples of a given data. Prediction accuracy and overfitting control. The Maximum Samples parameter controls the size of each subsample. Bootstrap is True (default), otherwise the entire data is used to build each tree

- AdaBoost Regression Algorithm (ABR): The AdaBoost regression algorithm can be used as an estimator starting with appropriate adjustments. Since the regressor is based on the initial data, additional copies of the regressor are fitted to the same dataset, but with all the weighting, the instances are then adjusted according to the error in this prediction. As a result, subsequent regression models focus on: even more in more difficult cases. The decision tree is then boosted using the Ada Boost algorithm*. This algorithm works on 1D a sine wave data set with some Gaussian noise. 299 boosts to 300 decision trees are analyzed.
- XGBoost Regressor (XGBR): The XG of the XGBoost stands for Extreme Gradient which is a free and open-source library that produces an effective implementation of the gradient boosting algorithm. Soon after the development and its first release, this algorithm became the go to technique and infrequently the important aspect to win the solutions for a variety of tasks in machine learning contests. Prediction based regression modeling tasks involve the prediction a numeric value like an amount or a distance. This algorithm is often used very directly for prediction-based regression modelling. The gradient boosting points to a class of ML algorithms related to ensemble learning which can be used for either of classification or regression tasks.
- CatBoost Regressor (CBR): Cat Boost is based on a decision tree algorithm and gradient boosting inference algorithm. The idea of boosting is to add a large number of weakly constructed models so that a robust model can be built through greedy search techniques. Competitive models for prediction. The learned tree learns because gradient boosting adjusts the selected trees one by one. There are fewer mistakes because there are no more mistakes. This allows you to add new functionality to existing functionality this methodology is valid until the selected loss method no longer reduces

The support vector regression method gives the best accuracy of over 89%, while the CatBoost algorithm
Accuracy is greater than 88%. This is approximately the same accuracy as support vector regression.

Conclusion

Throughout the project, we built several machine learning regression models from scratch and gained extensive knowledge additionally, some insights were gained about regression models, power transformers, and their development. we looked into it many algorithms are used to improve the accuracy of real estate price predictions, including: Examples: support vector regressor, linear regression, nearest neighbor, random forest regressor, AdaBoost regressor, CatBoost regressor, XGBoost regressor, etc. I compared all of these the algorithm mentioned in the previous statement concluded that it is the CatBoost regressor and SVR. This provides the highest accuracy of approximately 90%. Improved prediction accuracy by up to 15% compared to existing models a detailed comparison of the performance of all algorithms used in this project is also presented graphically.

Acknowledgments

We would like to thank all the faculty members of ECM Department, Sreenidhi University of Science and Technology for their support continued assistance, collaboration and support to complete this project.

We are Drs. Many thanks to Mr. D. Mohan, Dean, ECM Department, Sreenidhi University of Science and Technology, Ghatkesar you took the initiative in this project and provided valuable and timely suggestions regarding our project. We thank you for your cooperation project completion.

We would like to thank Dr.T.Ch. Very heartfelt. Mr. Siva Reddy, Director, Srinidhi Institute and Mr. Chakkarakal Tomy, Managing Director, Srinidhi Institute Doctor of Science and Technology, Ghatkesar for providing the resources to complete this project. Finally, we would like to express our gratitude to you almighty Ones, all our friends, teachers and non-teachers who have helped us directly or indirectly in this endeavor.

References and notes

1.Garriga C., Hedlund A., Tang Y., Wang P, "Regional Science and Urban Economics, Rural-Urban Migration and Real Estate Prices in China", Region Science and Urban Economy (2020), p. 103613, March 2020

2.Wang 101715, May 2018.

3.G. Naga Satish, Ch. V. Raghavendran, M. D. Sugnana Rao, Ch. Srinivasulu "Housing Price Prediction Using Machine Learning". IJITEE, 2019.

4. Bharatiya, Dinesh, et al. "Stock Market Prediction Using Linear Regression," Electronics, Communications, and Aerospace Engineering (ICECA), 2017. international conference. Vol. 2. IEEE、2017.

5.Anand G. Rawool1、Dattatray V. Rogye、Sainath G. Rane, Dr. Vinayk A. Bharadi, "House price predition using Machine Learning, IRE Journals, May 2021.

6.E.Laxmi Lydia, Gogineni Hima Bindu, Aswadhati Sirisham, Pasam Prudhvi Kiran, "Electronic Governance of Housing Price using Boston Dataset Implementing through Deep Learning Mechanism", IJRTE, Volume-7 Issue-682, April-2019.

7.Li Yu, Chenlu Jiao, Hongrun Xin, Yan Wang, Kaiyang Wang, "Prediction on Housing Price Based on Deep Learning", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol.12, No.:2, 2018.