International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

REAL ESTATE PRICE PREDICTION USING MACHINE LEARNING

[1]*Manudev Chhiller, CSE, SRM IST Ramapuram*
[2]*Shivam, CSE, SRM IST Ramapuram*
[3]*Rishav Kumar, CSE, SRM IST Ramapuram*

Abstract - The real estate industry's notorious lack of transparency and volatile housing prices have spurred research projects to enhance housing price predictions. These initiatives employ a combination of regression techniques, emphasizing the weighted average of these methods to yield more accurate outcomes. They also suggest the utilization of real-time neighborhood data from Google Maps to enhance real-world valuations. With the real estate market's competitiveness and its reliance on factors like physical condition, concepts, and location, these projects strive to forecast residential prices based on customer financial needs, studying historical market trends and forthcoming developments. Multiple regression techniques such as Linear Regression, Lasso Regression, Forest Regression aiming to optimize house price predictions and assist lower- and middle-class individuals with financial parameter-based price estimations. Overall, these projects leverage data science and machine learning to address real estate pricing challenges, seeking to enhance transparency and decision-making for both buyers and investors.

***Key Words*:** Real Estate, Linear Regression, Lasso Regression, Forest Regression, Decision-making.

1.INTRODUCTION

At the core of technological innovations lies the pivotal role of data, enabling the realization of various outcomes through predictive models. Machine learning is the predominant tool in this endeavor, functioning on the premise of providing a robust dataset, with subsequent predictions based on a machine's learning from this pre-loaded data. This methodology finds extensive application in diverse domains, such as forecasting stock prices, earthquake possibilities, and company sales. Our approach encompasses a comprehensive dataset, encompassing factors like square footage, bedroom and bathroom counts, flooring type, elevator and parking availability, and furnishing condition. A diverse dataset is employed to ensure accurate results for various scenarios. We utilize various algorithms, and after careful evaluation, we ascertain that a combination of algorithms, rather than a single one, yields superior results. This research is indispensable for informed urban planning and is crucial in addressing the unpredictability of property prices, particularly in India, where soaring urban property costs make standardization and transparency imperative. The need for data-driven insights is pronounced, given the lack of standard pricing mechanisms in the Indian real estate market, where media influence can skew perceptions. For both real estate professionals and laymen, comprehending market dynamics and their impact on property values is challenging. Thus, a tool that grasps these dynamics and parameter effects is essential. To build such a system, we need an accurate predictive model that minimizes errors, leveraging substantial historical data. With limited research focused on property prediction in India, this project aims to construct a robust system that considers various parameters affecting property values, employing machine learning techniques. Regression models, particularly multiple linear regression, are instrumental in predictive analysis and are commonly used to establish relationships between target variables and various independent factors. This research project utilizes data accessible through the Machine Hackathon platform, specifically examining three prediction models: Linear Reggression, Lasso and Ridge regression, Forest Regression with a comparative assessment using evaluation metrics. The paper comprises sections outlining prior research, data description and preprocessing, model development and evaluation, summarization of results, and future prospects. It underscores the role of data science in real estate, where thorough analysis offers users valuable insights into property cost, availability, and potentially prevents fraud.

2. EXISTING SYSTEM

MULTIPLE LINEAR REGRESSION

Multiple Linear Regression examines the correlation between a scalar response variable and two or more explanatory variables. It reveals how the value of the dependent variable is linked to the values of independent variables.

LIMITATIONS

The dependent variable, denoted as y, needs to be continuous. In contrast, the independent variables can take on any type. Typically, the dependent variable is influenced by the independent variables.

PROBLEM STATEMENT

Acquiring a house stands out as one of the most significant investments an individual makes in their lifetime. Consequently, it’s crucial to exercise extreme caution and ensure that the right amount of money is invested in the property.

In the upcoming document, we delve into various machine learning techniques and methodologies designed to forecast house prices. The dataset encompasses both training and testing sets. Our goal is to predict house utilizes prices by

considering the preferences and requirement of users. The model utilizes sample data to anticipate the price of a house.

3. PROPOSED SYSTEM

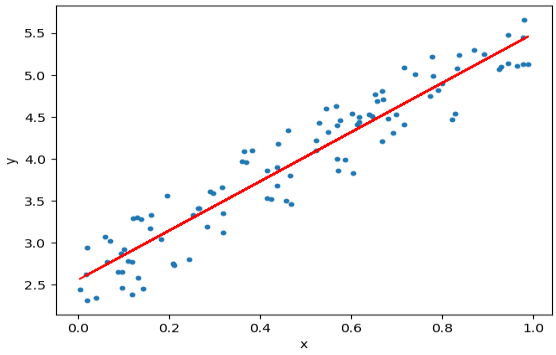
Linear Regression is a method employed to discern the connection between a dependent variable and an independent variable. The regression techniques applied in this context encompass linear regression, Lasso Regression and Forest Regression.

PROPOSED WORK

Linear Regression

The property rates database includes different categories like "quarter," "upper," "normal," and "lower." The "upper" category represents high-cost houses, while "normal" and "lower" encompass mid-range and low-cost houses, respectively. To employ linear regression, we assign the "quarter" attribute to the x-axis and property rate values to the y-axis. Linear regression is performed individually for each category.

In linear regression, we assume a connection between an independent variable vector and the dependent target variable. By using these independent parameters, we can predict the target variable. The independent data vector can consist of N parameters or properties, also known as regressors. Linear regression assumes a linear relationship between the dependent variable and the regressors. The difference between the predicted value and the observed value is referred to as the "error."

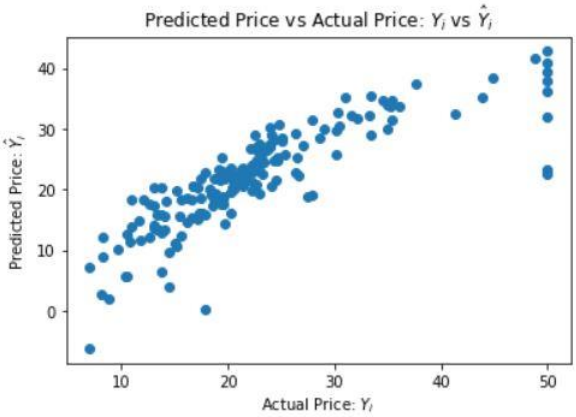


The next step is to determine the best-fitting relationship (line) between these variables. The most common method used is the Residual Sum of Squares (RSS). This method calculates the discrepancy between observed data (actual value) and its vertical distance from the proposed best-fitting line (predicted value). It squares each difference and sums them up. The Mean Squared Error (MSE) is a measure of the quality of the estimator, obtained by dividing RSS by the total observed data points. It is always a non-negative number, with values closer to zero indicating a smaller error. The Root Mean Squared Error (RMSE) is the square root of MSE and represents the average deviation of predictions from observed values. This is easier to interpret compared to MSE, which can be a large number.

Linear regression excels in predicting precise numerical target values, unlike other models that primarily classify outputs. As a result, it plays a vital role in predicting the price of real estate properties.

Lasso Regression

LASSO, or "Least Absolute Shrinkage and Selection Operator," is a linear regression technique that also incorporates regularization. While it shares similarities with ridge regression, it differs in how it applies regularization. LASSO considers the absolute values of the sum of regression coefficients and has the unique ability to reduce some coefficients to zero, effectively eliminating them. This feature makes LASSO regression a valuable tool for feature selection. In contrast to ridge regression, where the regularization term involves squared values, LASSO employs the absolute values of the coefficients, which can be thought of as the magnitude of their influence on the model.



It's worth noting that from a computational standpoint, LASSO regression can be more computationally intensive than ridge regression. To find the optimal regularization hyperparameter (λ), a grid-search cross-validation approach is often used. This entails testing a wide range of hyperparameter values, and in your case, the best value was determined to be 0.001.

Forest Regression

Random Forest regression relies on a technique called "Bagging of trees" to improve its predictive accuracy. The central idea behind this method is to create a multitude of decision trees that are uncorrelated, effectively reducing variance and enhancing the robustness of predictions. The process involves generating a significant number of decision trees through the random forest training algorithm, which applies the principle of "bootstrap aggregating" or "bagging" to the tree learners. This process begins with a training dataset (X) and the corresponding response data (Y). Iteratively, random samples of training examples, with replacement, are selected, forming Xb and Yb. Subsequently, a classification or regression tree (fb) is trained on these newly created Xb and Yb samples. By combining the outputs of these diverse decision trees, Random Forest significantly improves its

predictive performance, making it a powerful tool for regression tasks.

4. REQUIRED SYSTEM

HARDWARE REQUIREMENTS

The standard set of requirements outlined by any operating system or software application pertains to the physical computer resources, commonly referred to as hardware. A hardware requirements list is typically accompanied by a hardware compatibility list, particularly in the context of operating systems. The minimum hardware specifications include:

1. PROCESSOR: PENTIUM IV
2. RAM: 8GB
3. PROCESSOR SPEED: 2.4 GHz
4. MAIN MEMORY: 8GB RAM
5. PROCESSING SPEED: 600 MHz
6. HARD DISK DRIVE: 1TB
7. KEYBOARD: 104 KEYS

SOFTWARE REQUIREMENTS

Similarly, software requirements involve specifying the resource prerequisites and installations necessary on a computer to facilitate the functioning of an application. These requirements typically require separate installations. The minimal software requirements include:

1. FRONT END: PYTHON
2. IDE: ANACONDA
3. OPERATING SYSTEM: WINDOWS 10

5. CONCLUSION

A sophisticated housing price prediction system has been developed, combining Linear Regression, Lasso Regression, Forest Regression to ensure accuracy. This algorithm effectively minimizes the risk of investing in the wrong property and offers additional features to enhance customer satisfaction. Future updates may include expanding the database to encompass larger cities, enabling users to explore a broader range of houses for more precise decision-making. Challenges arise when dealing with vast and diverse real estate data, and the system optimally employs Linear Regression to enhance decision accuracy. Further enhancements could involve incorporating additional factors like economic recessions and in-depth property details for a more comprehensive operation on a larger scale.

6. REFERENCE

1. Housing Price Prediction Through Machine Learning Algorithm: A Case Study in Melbourne City, Australia, by Danh Phan.
2. Sales Price Prediction of Houses using Regression Methods in Machine Learning authored by Parasich Andrey Viktorovich, Parasich Viktor Aleksandrovich, Kaftannikov Igor Leopoldovich, and Parasich Irina Vasilevna.
3. Real Estate Value Prediction Using Linear Regression, written by Nehal N Ghosalkar and Sudhir N Dhage.
4. Prediction of Housing Market Trends Using Twitter Data, authored by Marlon Velthorst and Cicek Guven.
5. House Price Prediction Using Machine Learning and Neural Networks, by Ayush Verma, Abhijit Sharma, Sagar Doshi and Rohini Nair.