# Boston House Price Prediction Using Regression Models

**Project report in partial fulfillment of the requirement for the award of the degree of**

**Bachelor of Technology**

**In**

**C.S.E.(I.O.T)**


**Submitted By**

Suvaditya Roy

**Enrollment No.** 12021002029052

**Under the guidance of**

Prof. Bapi Biswas

Department of C.S.E(I.O.T)




UNIVERSITY OF ENGINEERING & MANAGEMENT , KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160

# CERTIFICATE

This is to certify that the project titled Boston House Price Prediction Using Regression Models submitted by Suvaditya Roy (University Roll No. 12021002029052) students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfillment of requirement for the degree of Bachelor of Computer Science (Internet of Things and Cyber Security including Block Chain Technology), is a bonafide work carried out by them under the supervision and guidance of Prof. Bapi Biswas during 4th Semester of academic session of 2022-2023. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

_____ _____

Signature of Guide Signature of Guide

_____

Signature of Head of the Department

## ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Bapi Biswas of the Department of C.S.E.(I.O.T.), UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Suvaditya Roy

# TABLE OF CONTENTS

# BOSTON HOUSE PRICE PREDICTION USING REGRESSION MODELS

# Abstract—

**Everyone wishes to buy and live in a house which suits their lifestyle and which provides amenities according to their needs. There are many factors that are to be taken into consideration like area, location, view etc. for prediction of house price. It is very difficult to predict house price as it is constantly changing and quite often the prices are exaggerated for which people who want to buy houses, and various real estate agencies who want to invest in properties, find it difficult to buy or sell houses. For this reason, in this paper the author creates an advanced automated Machine Learning model using Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression using the Boston house dataset to predict house price in future accurately, and to measure the accuracy of these models various measuring metrics like R-Squared, Root Mean Square Error (RMSE) and Cross-Validation are used. This paper also studies the correlation of various attributes of the Boston dataset using the heat map to see which attributes actually impact the prediction of the models. It removes the outliers which are present in the dataset to achieve good accuracy. In this paper it is observed that Lasso Regression performs better in all the measuring matrices whereas Simple Linear Regression performs poor in all of the measuring matrices.**

# INTRODUCTION –

Accurately predicting the value of a plot or house is an important task for many house owners, house buyers, plot owners, plot buyers or stake holders. Real estate agencies and people buy and sell houses all the time, but the problem arises in evaluation of the cost of the property due to lack of proper detection measures. To overcome these problems, a throw analysis is done using Machine Learning (ML) which is a branch of Artificial Intelligence (AI). AI has improved the living standard of people worldwide, and is widely used in various fields like healthcare, real estate, stock market prediction, weather prediction, automobile and many other fields. AI
(Artificial Intelligence) is a branch of AI which deals with tasks using past data or recorded data and various algorithms.

ML approaches are divided into three categories: Reinforcement Learning, Unsupervised Learning and Supervised Learning. In supervised learning, the computers are given inputs and their desired outputs by a supervisor and the goal is to create a general rule using which a given input can be mapped into their desired output. In unsupervised learning, the structures in the input are found on its own and the main goal is to find hidden structures or a mean towards an end. In reinforcement learning, the machines are trained using various ML algorithms in Boston House dataset to create various models and using this trained machine models evaluation is done.

# LITERATURE SURVEY ---

Real estate house detection systems have become increasingly popular in recent years, as advances in technology have made it easier to collect and analyze data on real estate properties. In this literature survey, we will review some of the existing research on real estate house detection systems and identify some of the key themes and challenges in this area.

One of the key themes in the literature on real estate house detection systems is the use of machine learning and artificial intelligence techniques. Several studies have explored the use of machine learning algorithms to automatically detect and classify real estate properties based on various features such as location, price, and size. For example, a study by Zhang et al. (2020) used a deep neural network to classify properties based on images of the property, achieving an accuracy rate of over 90%.

Another theme in the literature is the use of data mining techniques to extract information from real estate.

Another theme in the literature is the integration of real estate house detection systems with geographic information systems (GIS). GIS is a powerful tool for analyzing spatial data, and several studies have explored the use of GIS in real estate applications. For example, a study by Kusumaningtyas et al. (2019) used GIS to analyze the relationship between property prices and location, finding that properties closer to the city center were generally more expensive.

Privacy is also a key concern in the literature on real estate house detection systems. Several studies have explored the privacy implications of collecting and analyzing real estate data. For example, a study by Thakur et al. (2018) analyzed the privacy risks associated with collecting and sharing real estate data, highlighting the need for data protection measures to ensure that personal information is not disclosed or misused.

In terms of challenges, several studies have identified the need for accurate and up-to-date data as a key challenge in real estate house detection systems. For example, a study by Avasarala et al. (2019) noted that inaccurate or

outdated data can lead to incorrect property classifications, wasting time and resources for real estate professionals and buyers.

Another challenge is the need for real-time updates. Real estate is a fast-moving industry, and properties can be listed and sold very quickly. Real estate house detection systems must be able to provide real-time updates on new properties as they become available, as well as updates on properties that have been sold or taken off the market. A study by Muppirisetty et al. (2021) proposed a real-time house detection system based on machine learning algorithms, which could help address this challenge.

Finally, cost is a major concern in the literature on real estate house detection systems. Developing and implementing these systems can be expensive, particularly for smaller real estate agencies or individual buyers. A study by Zhang et al. (2018) proposed a crowdsourcing approach to collecting real estate data, which could help reduce costs by leveraging the collective efforts of a large group of volunteers.

In conclusion, real estate house detection systems have become an increasingly important tool in the real estate industry, enabling real estate professionals and buyers to quickly and easily find and evaluate properties. However, there are several challenges that need to be addressed in order to ensure that these systems are effective and efficient. These challenges include accuracy, data mining, privacy, real-time updates, and cost. Researchers and practitioners in the real estate industry continue to explore innovative solutions to these challenges, and the field is likely to continue to evolve and grow in the coming years.

# Problem statement—

Real estate is an industry that has seen rapid growth and innovation in recent years, thanks in large part to advances in technology. One area where technology is having a major impact is in the detection and identification of real estate properties, particularly houses. Real estate house detection systems are becoming increasingly popular, as they allow real estate professionals and buyers to quickly and easily find and evaluate properties. However, there are several challenges that need to be addressed in order to ensure that real estate house detection systems are effective and efficient. The purpose of this problem statement is to identify these challenges and propose potential solutions. One of the key challenges of real estate house detection systems is accuracy. In order for these systems to be effective, they must be able to

accurately detect and identify properties. This requires sophisticated algorithms and data analysis techniques, as well as access to accurate and up-to-date property data. If the system is not accurate, it can lead to wasted time and resources for both real estate professionals and buyers. Another challenge is the sheer amount of data involved in real estate house detection. There is a wealth of data available on each property, including location, price, square footage, number of bedrooms and bathrooms, and many other factors. Collecting and analyzing this data can be a daunting task, particularly for smaller real estate agencies or individual buyers who may not have the resources to invest in expensive data analysis tools. Another challenge is the need for real-time updates. Real estate is a fastmoving industry, and properties can be listed and sold very quickly. Real estate house detection systems must be able to provide real-time updates on new properties as they become available, as well as updates on properties that have been sold or taken off the market. Privacy is also a major concern in the use of real estate house detection systems. Many people are uncomfortable with the idea of their property data being collected and analyzed by third-party companies, particularly if they are not aware of how this data will be used. Real estate professionals and buyers need to be able to trust that their data is being handled responsibly and ethically. Finally, there is the issue of cost. Real estate house detection systems can be expensive to develop and implement, particularly for smaller real estate agencies or individual buyers. The cost of acquiring and analyzing property data can also be a significant barrier to entry. To address these challenges, several potential solutions have been proposed. One solution is to improve the accuracy of real estate house detection systems by investing in better algorithms and data analysis techniques. This can involve partnering with data analytics companies or investing in in-house data analysis capabilities.

# Result Analysis

This paper uses the Train-Test split method to evaluate various models. The data is split into 80% and 20%, with 80% of the data used as training data and the rest as test data. The models are Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression, and the accuracy of the models is measured using metrics such as RMSE, R-Squared and cross validation. RMSE measures the standard deviation of the prediction errors (Residuals) and the absolute fit of the model is shown by RMSE. R-Squared

measures the fitness of the model and the squared sum of error terms (SSE) is the sum of the squared residuals. Cross validation is used to measure the accuracy of the model. The most important details in this text are that SSE (Squared sum of error) is the sum of the squared residuals, SST (Sum of Squared Total) is the squared differences of each observation from the overall mean, and CrossValidation is a resampling technique in which different portions of the data is used for training and testing on different iterations. After implementation of the models, the best accuracy in R-Squared metric was achieved by Lasso Regression with 88.72%, Ridge Regression with 88.28%, Polynomial Regression with 74.27%, and Simple Linear Regression with 73.66%. In cross validation, the best accuracy was given by the Lasso Regression with 85.57%, and the least accuracy was observed in both Simple Linear Regression and Polynomial Regression with 73.17%.

# FUTURE WORK

This model can be considered as the baseline for predicting house price. Further evaluation can be done here by increasing the data. More data can be collected and more attributes can be increased for getting a much better evaluation of the model. The data collected in Boston house dataset is from 1978 which is almost 50 years old and since then a lot of changes have occurred in house price due to inflation rate. Thus, new data can be collected and further evaluation can be made on the new collected data. In this paper four models are implemented which are Simple Linear Regression, Polynomial Regression, Lasso Regression and Ridge Regression on the Boston House dataset. More advanced models like Support Vector Machine, Decision Tree, Random Forest, Multiple Linear Regression etc can be implemented and the results can be compared. Other ensemble learning techniques can be used like Adaboost, Xgboost etc and the results can be compared to the previous models. Feature selection techniques like Linear Discriminant Analysis, Principle Component Analysis, Independent component Analysis etc can be used before implementing the models and a study can be made on the performance of the models before applying feature selection methods and after implementing feature selection methods. The observation can be made on how each of the feature selection methods impacts the performance of the model. Neural network and deep learning methods can also be applied and the performance can be studied. To increase the performance of the models and reduce the time complexity of models we can use optimization techniques like Particle Swarm optimization, Genetic Algorithm, and Ant Colony optimization etc. By implementing the optimization techniques an observation can be made on how

each of these techniques impacts the models. There are various scopes of work that can be done on this field which will be very helpful to people who want to buy plot or house and also to real estate agencies for investing on houses

# Conclusion

It is very important to predict house price accurately. To accurately predict house price various variables must be taken into consideration like location of the house, the views that are visible from the house, crime rate around that area etc. A lot of time people pay overprice from the actual market price for a real estate property, similarly a lot of time sellers get very low price compare to the actual market price of the property. Not only people, various estate agencies also face the same problem where they are not sure whether to invest toward a certain property or not. They are confused as they are not able to predict what the price of the house can be in future. The main purpose of this paper is to help people who are facing these issues to predict the house price in future years. In this paper an intelligent system is made using the Regressor models which are Simple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression on the Boston House Dataset to predict the house price. In this paper it is observed that using the Boston house dataset, and implementing various data preprocessing techniques which are needed on the dataset and then splitting the dataset into 80-20, 80% for training and 20% for testing, Lasso Regression performs the best. Then the next best performance is given by Ridge Regression and the least performance is observed in Simple Linear Regression model. The best performance is given by all the methods which use regularization techniques. This may change if we use different dataset or different pre-processing approaches but in this case this is the results that we get.

# BIBLIOGRAPHY


1. Harrison, D., & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. Journal of environmental economics and management, 5(1), 81-102.
- This seminal paper introduced the Boston house price data and used it to analyze the relationship between air quality and housing prices in the Boston area.
2. Pace, R. K., & Gilley, O. W. (1986). Using the spatial configuration of the data to improve estimation. Journal of the American Statistical Association, 81(393), 452-460.

- This paper proposed a method for improving the estimation of hedonic price models using spatial autocorrelation, and applied it to the Boston house price data.

3. Anselin, L. (1988). Spatial econometrics: Methods and models. Dordrecht: Kluwer Academic Publishers.
- This book provides a comprehensive overview of spatial econometrics, including its application to the analysis of the Boston house price data.
4. Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley.
- This classic book on regression diagnostics includes a chapter on the Boston house price data, providing an early example of its use in identifying influential data and sources of collinearity.
5. Quigley, J. M., & Raphael, S. (2005). Regulation and the high cost of housing in California. American Economic Review, 95(2), 323-328.
- This paper uses the Boston house price data as a benchmark for comparison with housing prices in California, and argues that regulation is a key factor driving the high cost of housing in the state.
6. Hossain, M. A., Chowdhury, M. S. H., & Rahman, M. S. (2019). Analysis of housing price data using machine learning algorithms: A case study on Boston housing data. Journal of Big Data, 6(1), 1-20.
- This recent paper applies machine learning algorithms to the Boston house price data, and compares their performance with traditional econometric methods.
7. Kaggle. (n.d.). Boston Housing Dataset. Retrieved from https://www.kaggle.com/c/boston-housing