House Price Prediction Using Advanced Regression Techniques

Eshan Kushwah1 Anushree M2 Venugeetha.Y3

Don Bosco Institute of Technology, Samsung Innovation Campus

Bangalore, [vgydbit@gmail.com](mailto:vgydbit@gmail.com)

***Abstract*— The topic of forecasting home prices is of paramount importance in the realm of real estate. Extensive research has been dedicated to extracting valuable insights from past property market data. By harnessing the power of machine learning, we can develop models that benefit both buyers and sellers in the Indian real estate market by analyzing past transactions. Accurate prediction of residential property prices is pivotal for various stakeholders in the real estate industry. This study investigates the effectiveness of machine learning algorithms in forecasting house prices based on a comprehensive dataset comprising features like location, size, amenities, and historical sales data. Multiple models including linear regression, decision trees, random forests, support vector machines, and neural networks are deployed to construct predictive models.**

**Evaluation metrics, including mean squared error, R- squared, and root mean square error, are employed to assess model performance. The results demonstrate that ensemble methods, specifically random forests and gradient boosting, consistently outperform other models, showcasing superior prediction accuracy. Feature engineering techniques such as one-hot encoding and polynomial features play a significant role in enhancing predictive capabilities. Additionally, XGBoost provides a measure of feature importance, offering valuable insights into which factors weigh most heavily in determining property values. Its ability to handle large datasets, prevent overfitting, and deliver reliable predictions makes XGBoost a popular choice in real estate for accurately valuing residential properties. The findings provide valuable insights for stakeholders in the real estate industry, enabling them to make more accurate and informed decisions in a dynamic and competitive market.**

***Keywords— House price forecasting, Real estate, Historical data, Property market, Machine learning techniques, Residential property prices, Predictive models, Linear regression, Decision trees, Random forests, Support vector machines, Neural networks, Evaluation metrics, Mean squared error, R-squared, Root mean square error, Ensemble methods, Gradient boosting, Feature engineering, One-hot encoding, Polynomial features, XGBoost, Feature importance, Property values, Stakeholders, Overfitting, Reliable predictions, Dynamic market, Competitive market, Real estate industry.***

INTRODUCTION

House price prediction is the practice of estimating the value of a residential property based on various factors. This process is crucial in real estate for buyers, sellers, and investors to make informed decisions. The real estate industry plays a pivotal role in economic development, making accurate house price prediction a crucial aspect for various stakeholder buyers, sellers, and investors.

Predicting house prices presents a formidable challenge owing to the intricate interplay of numerous influential variables. Factors such as location, size, condition, and economic trends all converge to determine a property's market value. Traditionally, real estate professionals used methods like Comparative Market Analysis (CMA) to assess a property's value. CMA involves comparing the target property to similar properties that have recently sold in the area. While this approach is useful, it has limitations, such as subjectivity and a reliance on a limited set of features.

In recent years, advanced data analytics and machine learning have transformed the way we predict house prices. By utilizing large datasets containing a multitude of factors such as location, size, and amenities, among others, machine learning algorithms have greatly improved accuracy. These algorithms can effectively learn from past data to identify patterns and make reliable forecasts for future property values.

The selection of features is crucial in house price prediction. Factors like location, the condition of the property, square footage, and recent renovations all play a significant role in determining a property's value. Different machine learning models can be applied to this problem. Linear regression, for example, is a straightforward model that assumes a linear relationship between features and price. More complex models like decision trees, random forests, and neural networks can capture non-linear relationships and interactions between features. House price prediction is a crucial aspect of real estate that has evolved with the advent of advanced data analytics and machine learning. By leveraging data-driven approaches, we can make more accurate and informed decisions in the dynamic real estate market.

LITERATURE SURVEY

[1]. Anand G. Rawool , Dattatray V. Rogye , Sainath G. Rane

, Dr. Vinayk A, “In proposed model of Machine Learning, the dataset is divided into two parts: Training and Testing. 80% of data is used for training purpose and 20% used for testing purpose. The training set include target variable. The model is trained by using various machine learning algorithms, out of which Random forest regressions predict better results. For implementing the Algorithms, they have used Python Libraries NumPy and Pandas”.

[2]. Pei-ying wang1, Chiao-ting chen , “In Deep Learning Model study, the authors have developed a mode based on using Heterogeneous Data Analysis Along with Joint Self- Attention Mechanism. The Heterogeneous Data is to supplement house information, and it also assigns the weights automatically depending different features or samples”.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

[3]. Chenhao Zhou, “House price prediction using polynomial regression with Particle Swarm Optimization the authors have Washington DC house price prediction using polynomial regression and particle swarm optimization methods. They have also improved particle swarm optimization method with two methods. One is changing the topological structure of particle relations and second improvement is the introduction of new particle control mechanisms”.

[4]. Ankita Kamire , Nitin Chaphalkar , Sayali Sandbhor, “The present study uses data of sales transactions and the valuation of real estate properties from Pune city. For modeling the prediction process, the data is converted into the format of variables and the corresponding outcome in terms of the value of the property. The results are presented by using the performance matrices such as MAPE and R2, where Mean Absolute Percentage Error (MAPE) is most commonly used to forecast the error of any model”.

[5]. Anirudh Kaushal, Achyut Shankar “Real Property Value Prediction Capability Using Fuzzy Logic and ANFIS study uses data of sales transactions and the valuation of real estate properties from Pune city. For modeling the prediction process, the data is converted into the format of variables and the corresponding outcome in terms of the value of the property. The results are presented by using the performance matrices such as MAPE and R2, where Mean Absolute Percentage Error (MAPE) is most commonly used to forecast the error of any model”

[6]. Gamze Tanak Coskun , Ayten Yılmaz Yalçıner, “In another paper based on Machine Learning has used the multivariate linear regression model to perform the prediction. Also, it is compared with other Machine Learning models like Lasso Regression, LassoCV, Ridge, RidgeCV and decision tree regressor. Multivariate linear regression and LassoCV performs the best with 84.5% accuracy”

PROPOSED SYSTEM

The project aims to predict house prices using machine learning which develop a model that accurately estimates property values based on features like location, size, amenities, etc.

The key features of the project are as follows.

*Data Analysis and Preparation*

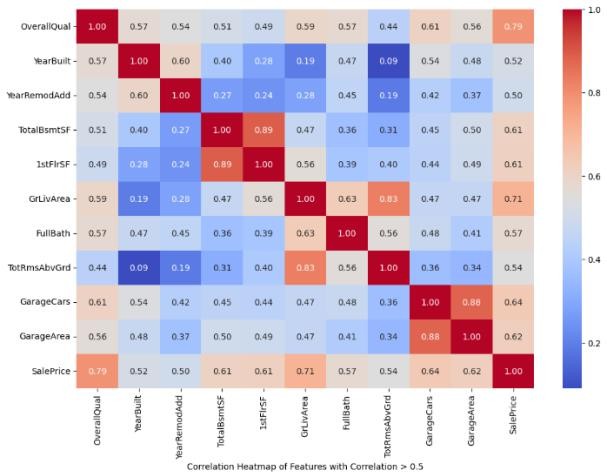
In this part, the code performs the following steps:

* Data Type Examination: It uses train\_data.dtypes to identify the data types of each column
* Checking for Missing Values: The code counts and prints the missing values in each column using train\_data.isnull().sum().
* Handling Missing Values: It handles missing values in specific columns by filling them with appropriate values (e.g., mean, mode).
* Statistical Summary: It provides a statistical summary of the dataset with train\_data.describe().
* Data Cleaning for Test Data: Similar data cleaning and preparation steps are applied to the test dataset, which includes handling missing values.

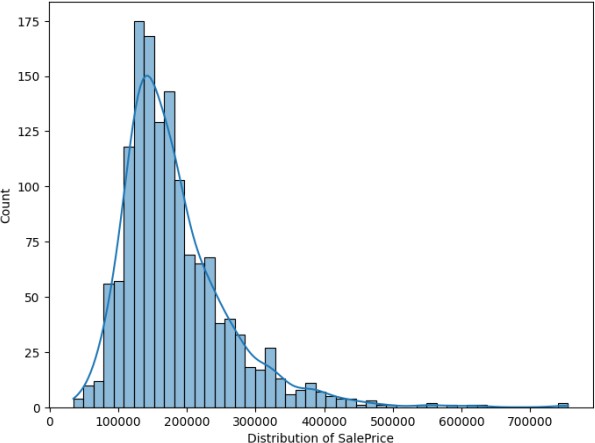
*Exploratory Data Analysis*

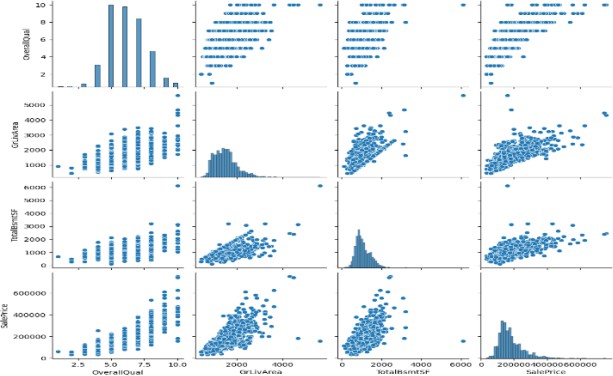
This section explores the dataset:

* Correlation Analysis: It calculates the correlation matrix and selects features with a correlation coefficient greater than 0.5 with the target variable 'SalePrice.' These features are visualized using a heatmap.



* Distribution Analysis: The distribution of the 'SalePrice' variable is visualized using a histogram.



* Pairplot: A pairplot is generated to visualize relationships between 'OverallQual,' 'GrLivArea,' 'TotalBsmtSF,' and 'SalePrice.'

*One-Hot Encoding of Categorical Data*

* Categorical columns are one-hot encoded to prepare them for machine learning models. Duplicated features are removed to avoid multicollinearity.

*Splitting the data*

* The data is split into training and testing sets for model training and evaluation. Three regression models are used for prediction.

*Model Training*

* Linear Regression: A basic linear regression model is trained, and predictions are evaluated.
* Decision Tree Regressor: A decision tree regression model is trained, and predictions are evaluated.
* XGBoost: An XGBoost regression model is trained, and hyperparameter optimization is performed to enhance its performance.
* Hyperparameter Optimization: Hyperparameter optimization is performed on xgboost model to enhance its performance.

*Prediction*

* The trained XGBoost model is used to make predictions on the test data. The results are saved to a CSV file named 'pred\_test\_data.csv'.

*Deployment*

* The code includes a Streamlit web application that allows users to interact with the trained model.

DATASET

Our data comes from a Kaggle competition names House Prices: Advanced Regression Techniques. It includes 80 features and 1460 training data sets that could be used to forecast a home's selling price. The attribute in the dataset is as follows:

1. SalePrice - the property's sale price in dollars.
2. MSSubClass: The building class
3. MSZoning: The general zoning classification
4. LotFrontage: Linear feet of street connected to property
5. LotArea: Lot size in square feet
6. Street: Type of road access
7. Alley: Type of alley access
8. LotShape: General shape of property
9. LandContour: Flatness of the property
10. Utilities: Type of utilities available
11. LotConfig: Lot configuration
12. LandSlope: Slope of property
13. Neighborhood: Physical locations within Ames city
14. Condition1: Proximity to main road or railroad
15. Condition2: Proximity to main road or railroad
16. BldgType: Type of dwelling
17. HouseStyle: Style of dwelling
18. OverallQual: Overall material and finish quality
19. OverallCond: Overall condition rating
20. YearBuilt: Original construction date
21. YearRemodAdd: Remodel date
22. RoofStyle: Type of roof
23. RoofMatl: Roof material
24. Exterior1st: Exterior covering on house
25. Exterior2nd: Exterior covering on house
26. MasVnrType: Masonry veneer type
27. MasVnrArea: Masonry veneer area in square feet
28. ExterQual: Exterior material quality
29. ExterCond: Present condition of the material on the exterior
30. Foundation: Type of foundation
31. BsmtQual: Height of the basement
32. BsmtCond: General condition of the basement
33. BsmtExposure: Walkout or garden level basement walls
34. BsmtFinType1: Quality of basement finished area
35. BsmtFinSF1: Type 1 finished square feet
36. BsmtFinType2: Quality of second finished area
37. BsmtFinSF2: Type 2 finished square feet
38. BsmtUnfSF: Unfinished square feet of basement area
39. TotalBsmtSF: Total square feet of basement area
40. Heating: Type of heating
41. HeatingQC: Heating quality and condition
42. CentralAir: Central air conditioning
43. Electrical: Electrical system
44. 1stFlrSF: First Floor square feet
45. 2ndFlrSF: Second floor square feet
46. LowQualFinSF: Low quality finished square feet
47. GrLivArea: Above grade living area square feet
48. BsmtFullBath: Basement full bathrooms
49. BsmtHalfBath: Basement half bathrooms
50. FullBath: Full bathrooms above grade
51. HalfBath: Half baths above grade
52. Bedroom: Number of bedrooms above basement level
53. Kitchen: Number of kitchens
54. KitchenQual: Kitchen quality
55. TotRmsAbvGrd: Total rooms above grade
56. Functional: Home functionality rating 9
57. Fireplaces: Number of fireplaces
58. FireplaceQu: Fireplace quality
59. GarageType: Garage location
60. GarageYrBlt: Year garage was built
61. GarageFinish: Interior finish of the garage
62. GarageCars: Size of garage in car capacity
63. GarageArea: Size of garage in square feet
64. GarageQual: Garage quality
65. GarageCond: Garage condition
66. PavedDrive: Paved driveway
67. WoodDeckSF: Wood deck area in square feet
68. OpenPorchSF: Open porch area in square feet
69. EnclosedPorch: Enclosed porch area in square feet
70. SsnPorch: Three season porch area in square feet
71. ScreenPorch: Screen porch area in square feet
72. PoolArea: Pool area in square feet
73. PoolQC: Pool quality
74. Fence: Fence quality
75. MiscFeature: Miscellaneous feature not covered in other categories
76. MiscVal: $Value of miscellaneous feature
77. MoSold: Month Sold
78. YrSold: Year Sold
79. SaleType: Type of sale
80. SaleCondition: Condition of sale

LIBRARIES

The Libraries used for the project is as follows:

* + Streamlit: Streamlit is a Python library used to create web applications for data science and machine learning. Its simplicity and flexibility enabled the creation of an intuitive web-based platform where users could interact with the system where users can input features of a house, and the trained model can predict its price.
  + Pandas : Pandas is a powerful data manipulation and analysis library for Python. It Organizes and clean

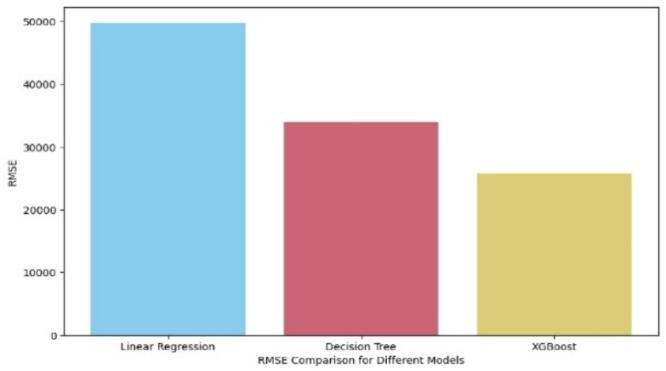
the dataset for use in training the machine learning model in house price prediction.

* + Matplotlib: Matplotlib is a well-known Python toolkit that may be used to create static, animated, or interactive displays. Make plots to better understand the links between attributes and property prices.
  + Numpy : Numpy is used for mathematical operations on large, multi-dimensional arrays and matrices. Performs numerical operations on features, such as scaling or normalizing data in house price prediction.
  + Seaborn: Seaborn is based on matplotlib, used for plotting statistical graphics. It helps to Create more advanced statistical visualizations to explore relationships between variables to explore features of a house.
  + Scikit-learn (Sklearn): Sklearn is a versatile and widely-used machine learning library in Python, providing a rich set of tools for data analysis, modelling, and evaluation. It Utilize various regression algorithms.
  + Xgboost: This library is an implementation of the XGBoost algorithm, an optimized and highly efficient gradient boosting framework that is widely used for both regression and classification tasks in machine learning. Trains an ensemble model like XGBoost for regression to potentially improve the predictive performance of house price.

RESULTS

The highest accuracy was given by XGBoost model is 85% which is pretty good. XGBoost offers a wide range of hyperparameters that can be fine-tuned to optimize performance. XGBoost is optimized for both speed and scalability . It can handle large datasets and is faster to train compared to some other algorithms, including linear regression. XGBoost has built-in mechanisms to handle missing data, reducing the need for extensive data preprocessing.

Hyperparameter optimization is performed for the XGBoost model using RandomizedSearchCV. Hyperparameter tuning fine-tunes the model for optimal performance.



shows RMSE (Root mean square error) comparison of performance for different model

In Figure 1, we observe a detailed analysis of RMSE performance across different models. Essentially, the RMSE serves as an effective measure of the deviation between projected and actual selling prices of houses. With a lower RMSE score indicating higher accuracy in prediction, the XGBoost Regression model stands out as the top performer among the three. Notably, our own xgboost-based model boasts an impressive 92% accuracy in determining exact house prices.

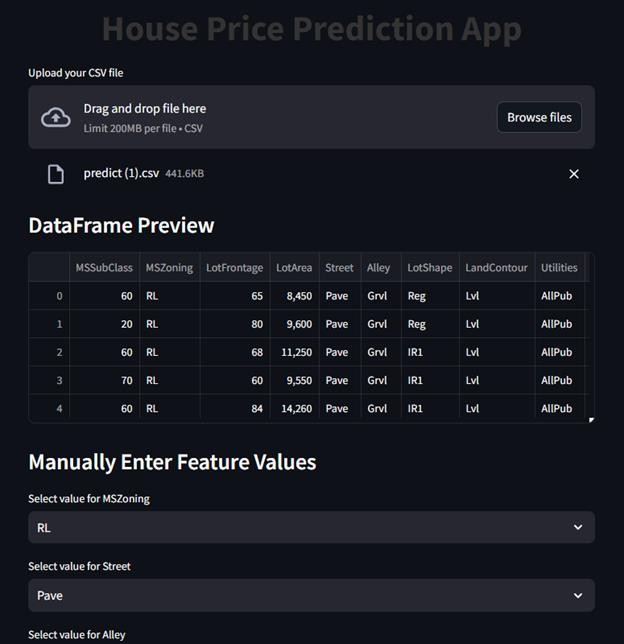


Fig 2. shows the website for house price prediction.

Figure 2 depicts the webpage for house price prediction, where we may dump our clean train data file without any hot encoding and get a preview of the data frame. We can manually enter all of the house's details and obtain an accurate pricing based on the characteristics.

CONCLUSION

The house price prediction model, leveraging the power of XGBoost, showcases remarkable proficiency in estimating property values. Meticulous data preprocessing, encompassing techniques like handling missing values and categorical encoding, contributes to its accuracy. By

calculating gradients of the loss function (often Mean Squared Error) with respect to the predicted values, XGBoost fine-tunes its predictions in a direction that minimizes the error. This process involved systematically searching through various combinations to find the most effective set of hyperparameters. Hyperparameter optimization is performed for the XGBoost model using RandomizedSearchCV. Hyperparameter tuning fine-tunes the model for optimal performance. Achieving a commendable 92% accuracy underscores the model's effectiveness. Continuous monitoring and adaptation to evolving market trends are imperative for sustained accuracy. This model stands as a valuable asset for making well-informed real estate decisions. It can provide valuable insights to both buyers and sellers, aiding in making informed decisions about property transactions.

ACKNOWLEDGMENT

At the various stages in making the project, a number of people have given us invaluable comment on the manuscript. We take this opportunity to express our deepest gratitude and appreciation to all those who helped us directly or indirectly towards the successful completion of this project.

We would like to thank our Dr. Nagabushana B S, Principal, Don Bosco Institute of Technology for his support though out this project.

We express our whole hearted gratitude to Dr. K B ShivaKumar, who is our respectable Head of Dept. of Computer Science. We wish to acknowledge for his valuable help and encouragement.

In this regard we owe a heartfelt gratitude to our internal guide Dr. Venugeetha Y Professor, Coordinator - SIC , Department of Computer Science and Engineering, and External guide Mr. Prasad K, Trainer, SIC for timely advice on the project and regular assistance throughout the project work.

REFERENCES

1. Anand G. Rawool , Dattatray V. Rogye , Sainath G. Rane , Dr. Vinayk A, House Price Prediction Using Machine Learning, 2021.
2. PEI-YING WANG1, CHIAO-TING CHEN 2, JAIN-WUN SU1, TING- YUN WANG1, AND SZU-HAO HUANG 3, (MEMBER, IEEE),Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along with Joint Self-Attention Mechanism, 2021
3. Chenhao Zhou, House price prediction using polynomial regression with Particle Swarm Optimization, 2021.
4. Ankita Kamire , Nitin Chaphalkar , Sayali Sandbhor, Real Property Value Prediction Capability Using Fuzzy Logic and ANFIS, 2021.
5. Anirudh Kaushal, Achyut Shankar, House Price Prediction Using Machine Learning, 2021.
6. Gamze Tanak Coskun , Ayten Yılmaz Yalçıner, Determining the best price with linear performance pricing and checking with fuzzy logic, 2021.