Machine Learning based Real Estate Price Prediction

Anupam Mukherjee
Dept. of AIML, School of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
anupammukherjee2003@gmail.com
ORCID: 0009-0005-1632-2084

Rtwik Nambiar
Dept. of AIML, School of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
rtwik.nambiar@gmail.com
ORCID: 0009-0002-2202-1569

Deepjyoti Choudhury*
Dept. of AIML, School of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
deepjyotichoudhury05@gmail.com
ORCID: 0000-0001-7288-2207

***Abstract*— The dynamic nature of real estate markets gives rise to huge challenges for buyers and sellers in accurately predicting property prices. This paper presents a detailed study on real estate price predictions, focusing on using machine learning techniques to forecast property prices effectively. Using a dataset from Bengaluru, as a case study, it explored the intricacies of Real Estate markets and developed a predictive model to estimate the property prices. Our studies consist of data preprocessing, featuring engineering, model selection, and evaluations, giving insights into the factors that affect property prices and the performances of various regression algorithms. After performing multiple experiments, we showcase the efficiency of our approaches in predicting property prices. Existing techniques that have tried to predict house prices have often struggled to deal with nonlinear relationships present among the factors. Recent techniques such as Linear regression, Ridge and Lasso Regression and Support Vector Machines(SVM) have come up short of being perfect when it comes to dealing with computational intensity and these techniques require extensive hyperparameter tuning. However, our predictive model is able to deal with these challenges by leveraging the strengths of Random Forest regression. This is done with high accuracy and the implications for real estate stakeholders are discussed, subsequently. This paper also highlights the various avenues through which the performance of real estate estimation models can be improved in the future.**

Keywords— Real Estate Price Prediction, ML, Regression Analysis, Property Valuation, Feature Engineering, Market Dynamics.

I. Introduction

The real estate market operates within a dynamic framework which is affected by a number of factors such as economic conditions, trends in the demographic, urbanization, and governmental policies. Taking these points into consideration, predicting the price of properties becomes significantly important for many stakeholders, which includes homebuyers, real estate investors, policymakers, and developers. House price prediction can be done by various deep and learning machine learning models such as hybrid regression [1][2], Bayesian optimization [3], ML techniques in addition to sentiment analysis [4]. ML Techniques such as lasso, ridge, and linear regression have been used to make accurate precisions with varying success [5]. The innate complexity and uncertainty of the real estate market, however, brings up significant challenges in forecasting property prices with a high precision score [6]. We tackle this challenge through this paper by presenting an extensive study on real estate price prediction, while focusing on leveraging machine learning techniques to enhance the overall predictive accuracy. Ensemble learning techniques like bagging, boosting and stacking have given good results in increasing the robustness of the predictive models [7]. Some papers have illustrated how effective deep learning frameworks like TensorFlow can be in accurately forecasting the real estate prices[8][9]. However, machine learning techniques like KNN, bayesian and regression models are still the norm [10][11][12][13][14][15]. Studies have been done to utilize open data in an efficient manner which can be contribute improving the predictive powers of the R.E.P.P model[16]. Homebuyers can facilitate informed purchasing decisions, by accurately forecasting the prices and also ensuring that they secure properties at relatively fair market values. The real estate investors also rely on price predictions to identify profitable investment opportunities and in turn optimize portfolio returns. Policymakers take the help of price predictions to calculate the housing policies, attain progress in urban development and the regulation of market dynamics.

After analysing the dataset, we are able to understand the intricate relationship between property attributes and prices. Our study consists of a systematic methodology that includes data preprocessing, feature engineering, model selection and evaluation. Our aim is to uncover key insights into the factors that determine property valuation and also assess and compare the performance of various regression algorithms in accurately predicting real estate prices. With the help of various machine learning techniques, we are able to tackle the built-in uncertainties and intricacies that are present in the real estate market and subsequently, in real estate price prediction. This proves advantageous for making the way for forecasting models that are accurate and reliable. Our main goal is to provide the stakeholders with insights that eventually lead to value creation and help them in making well informed decisions across the real estate industry.

The flow of the paper can be described in the following manner: Section I corresponds to the introduction; Section II includes the literature survey; Section III comprises of the design and implementation; Section IV discusses the results, and we conclude with Section V.

979-8-3503-8657-8/24/$31.00 ©2024 IEEE
DOI 10.1109/ICISC62624.2024.00067

II. LITERATURE REVIEW

TABLE 1: Comparative Analysis among existing literature

S no.	Author Name	Technology Used	Outcome
1	Quang Truong, Minh Nguyen, Hy Dang, Bo Mei (2019) [1]	Hybrid Regression	This study involves the usage of different types of hybrid ML methods for finding out correct solutions.
2	Phan TD (2018) [3]	Bayesian optimization method, deep neural network	This study uses the Bayesian and DNN methods that give high accuracy.
3	Rajesh Kumar, Praveen Ranjan Srivastava (2023) [4]	Hybrid Model (Machine Learning + Sentiment Analysis)	This study uses a hybrid model for predicting the real estate prices. The following combines machine learning techniques with sentiment analysis. It integrates the sentiment analysis of real estate market trends with traditional housing data to improve the accuracy of price predictions.
4	Fadhil M. Basysyar, Gifthera Dwilestari (2022) [5]	Linear ridge and lasso regression	This study uses lasso, ridge and linear regression, to aid in the improvement of accuracy of the model.
5	Xin Li, Shuai Zhang, and Xianfeng Tang (2019) [6]	Feature Selection Techniques	This study compares different feature selection techniques in relation to real estate house price prediction. It further evaluates methods such as filter, wrapper, and embedded approach to improve the accuracy of the model.
6	Sheng Chen, Jie Liu, Ying Shen (2018) [7]	Ensemble Learning Techniques (Bagging, Boosting, Stacking)	This study provides us with an overview of ensemble learning techniques used for real estate housing price prediction and their applications in improving the accuracy and robustness of predictive models for housing prices.
7	Fatemeh Mostofi, Vedat Togan, Hasan Basri Basanga(2022) [8]	DNN models, PCA	This study uses Deep neural networks and performs dimensionality reduction using PCA based on current real estate data to determine the market evaluation.
8	Xiangyang Ma, Pengfei Li, Hua Zhou (2021) [10]	KNN algorithm	This study calculates the usage of the KNN algorithm for precisely predicting real estate house prices.
9	Dhanush Gowda R (2024) [11]	Fuzzy Logic, Polynomial linear prediction model	This study accurately forecasts the real estate prices using a polynomial linear prediction model with the help of fuzzy logic. The following model compares and analyses the various new techniques that have been recently used in machine learning
10	Fatima Trindade Neves, Manuela Aparicio, Miguel de Castro Neto (2024) [16]	XGBoost, Employed SHAP	This study aims to utilize open data in an efficient and resourceful manner, which contributes to improving the predictive powers of the REPP model.

A. Motivation and requirement

This paper accurately predicts the housing prices of different regions according to multiple factors. House price prediction can be done by various deep and learning machine learning models. Hybrid regression is used to do the same. Three different types of machine learning models are compared and analyzed for optimal solutions. A fair amount of knowledge about the land pricing in different regions is important for all buyers. Machine learning is the primary approach for implementing house price prediction with supervised learning algorithms such as bayesian optimization and K-Nearest Neighbors. Machine learning and data analysis techniques are used to supply the buyer with all the necessary information about buying or selling a property. Research has been done to even apply sentiment analysis in collaboration with machine learning techniques to predict the house prices. Ensemble learning techniques like bagging, boosting and stacking have given good results in increasing the robustness of the predictive models. Reducing the uncertainty and providing a smoother experience for the buyer as well as the seller is a significant part of the process.

III. DESIGN AND IMPLEMENTATION

The workflow of the entire implementation process has been represented in fig.1.

A. Dataset Collection and Loading

The process of actively accumulating data from any reliable source is called Data Collection. The foundation of our analysis is completely dependent on the data that we select. The dataset should consist of different features such as size of the property, location of the property, number of bathrooms in the property and other similar features. These features serve as the building block of an accurate model. We source our data from a credible database and then further analyze the scope of the data. After we get a reliable dataset, the next step comprises of loading it into the computation environment. Python libraries such as numPy and Pandas are used to speed up the data loading process. Multiple mathematical operations can be performed by using the numPy library. Whereas, Pandas library provides us with tools that can be used to analyze and manipulate data. This in turn enables us to manage large datasets. For performing data preprocessing and exploration, we load the dataset into Pandas DataFrame giving us access to large sets of functionalities.

B. Dataset Preprocessing

Data preprocessing is performed to avoid loss of quality and integrity of a given data. Firstly, we make improvements to the fitness of the models while simultaneously considering the problems in the data. Handling the missing values is a primary step in data preprocessing. Different kinds of attribution techniques ensure that no critical information is lost due to the occurrence of missing data. Outlier detection is used to eliminate the impact of irregularities in the performance of the model. We then have to get rid of the data inconsistencies in our dataset. The presence of duplicate records, null values, and wrong information are the different types of errors that come up. Identifying and providing a solution to the problems makes sure that the data integrity is preserved. Feature selection and principal component analysis are the methods used for reducing dimensions. We get a proper understanding of our dataset which is useful for further processes. They also help in increasing the prediction powers of the model. Different techniques to encode, such as target encoding and ordinal encoding are used. All these steps in conversion to a numerical format. This becomes the best case for any machine learning algorithm.

C. Dataset Transformation

This process includes preparing the dataset for model training where we use different types of transformation techniques. These techniques are used to enhance the predictive power of the model. One of the most important aspects of this process is Feature Engineering. We create new features and edit existing ones to get valuable knowledge in this process. To further explain with an example, we make the feature: "price_per_square_feet" to regulate pricing across different properties. This also helps in giving us a better view of pricing trends. One-hot encoding and standard scaling are the techniques that are used to ensure a certain uniformity across contrasting scales. This finally leads us to notice significant progress in the model's power to interpret.

D. Model Selection

The implementation phase comes after thoroughly completing the preprocessing and transforming steps. This phase includes evaluating different types of predictive models. It consists of several significant steps such as model selection and training, evaluation of the model, and finally deploying the model. Model selection is the first step that consists of identifying the best suitable machine learning algorithms. Use of different types of models such as linear, ridge and lasso regression models, support vector machine model, xgboost regression model and random forest regression. All the algorithms are calculated on the basis of some metrics. The metrics that we used are as follows: accuracy, precision, mean squared error (MSE), mean absolute error (MAE), recall and F1-score.

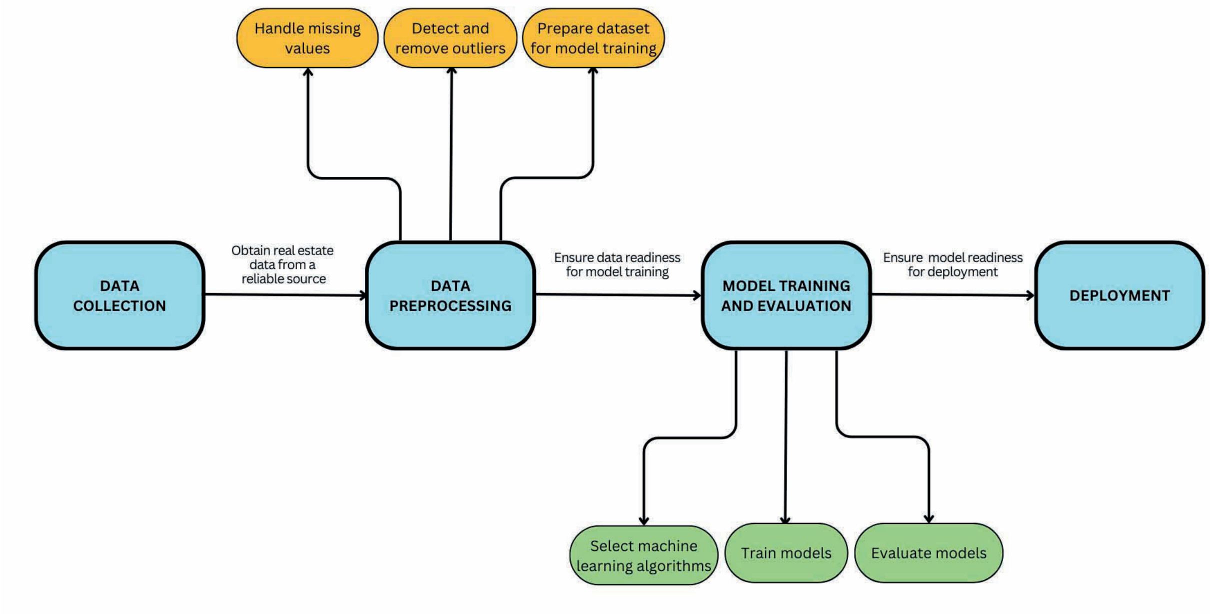

Handle missing values
Detect and remove outliers
Prepare dataset for model training
DATA COLLECTION
Obtain real estate data from a reliable source
DATA PREPROCESSING
Ensure data readiness for model training
MODEL TRAINING AND EVALUATION
Ensure model readiness for deployment
DEPLOYMENT
Select machine learning algorithms
Train models
Evaluate models

Fig. 1: Workflow Diagram

IV. PERFORMANCE ANALYSIS

A. Model Training and Evaluation

After choosing the models, we train them using our pre-processed dataset. We then split the entire dataset into training and testing data. Training phase makes use of techniques like cross-validation that avoids overfitting of the model. Models are then evaluated using different metrics to get to know their capabilities in prediction. Metrics such as accuracy, precision, MSE, MAE, recall, and F1 score are used to measure the performance of models across different evaluation criteria.

B. Metrics

1) Accuracy: The accuracy is computed as the ratio of true predictions to total predictions. It is calculated using the formula shown in equation (1):

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{1}$$

These evaluation metrics are useful to calculate the accuracy.

2) Precision: The precision is calculated as the ratio of true positive predictions to the total predicted positives, as shown in Equation (2):

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Here, TP and FP represent true positive and false positive predictions, respectively.

3) Recall: The recall metric measures the ratio of true positive predictions to the total actual positives, as given in Equation (3):

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Here, TP and FN represent true positive and false negative predictions, respectively.

4) F1-Score: The mean of precision and recall is called F1 score, computed using the fomula shown in Equation (4):

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Here, precision and recall are the precision and recall values, respectively.

5) Mean Absolute Average (MAE): measures the average absolute difference between the predicted and the actual values. Lower MAE indicates better model performance. It is computed using the formula shown in equation (5):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | Yi - \widehat{Yi} | \tag{5}$$

6) Mean Squared Error (MSE): It calculates the average of the differences that are squared between predicted values and actual values.

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2 \qquad (6)$$

V. RESULTS AND DISCUSSION

The proposed model is implemented using the following tools and technologies: M1 Macbook Air (8GB RAM, 3.2 GHz CPU) along with Python (Jupyter Notebook)

A. Comparison with Models

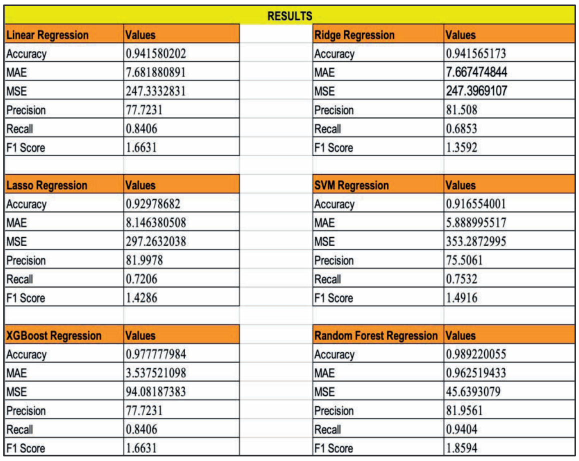
All of the three linear models that are Linear, Ridge and Lasso Regression, give similar results when it comes to calculating accuracy, mean squared error, and mean absolute error. Linear and Ridge Regression have MSE, MAE, and accuracy scores that are almost the same. Lasso Regression on the other hand has higher MSE and MAE scores and lower accuracy when compared to the other two. The only thing that is similar is the Precision, recall, and F1 score for all three models. Lasso Regression shows a little better results in terms of precision and recall. Due to the presence of a linear structure, these models function well but in most cases they may not find complex relationships present in the dataset.

SVM Regression calculates larger MSE and MAE values when compared to the linear models. This shows during error calculation, it displays poor performance. The accuracy score found out by SVM is higher in value but when it is compared with other linear models, it is seen to be much smaller. Values for metrics like Precision, recall, and F1 score are calculated. This regression model is useful to deal with unseen data in a dataset. SVM Regression is usually more useful for non-linear relationships in any dataset.

XGBoost Regression and Random Forest Regression perform way better than the linear models and SVM Regression. These ensemble methods are better in performance in terms of MSE, MAE, and accuracy. They have very low values of MSE and MAE when compared with all the models. This shows having a greater performance in prediction accuracy by these models. They also have the highest accuracy values. These methods also can find out complex relationships that are present in the dataset. These high and precise values confirm their usefulness in predicting real estate prices. Figures 3, 4, 5, 6, 7 and 8 represent the actual value vs predicted value plots of each of the models. Figures 9, 10, 11 and 12 are comparison graphs plotting the values of the evaluation metrics (Accuracy, F1 Score, Recall and Precision) for each of the models.

B. Discussion

On comparison with the existing techniques in the market, our Random Forest model comes out on top due to its ability to handle nonlinearity and interaction between the features. This results in our model being able to achieve higher accuracy and precision scores as shown in figure 2. After comparing all the regression models we find out that the ensemble methods, like Random Forest Regression, perform better than the linear models and SVM Regression in housing price prediction. Ensemble methods are usually known to identify non-linear relationships present in the data which may be missed by some linear models. Linear Regression, Ridge Regression, and Lasso Regression give a fairly good performance, with results that can be compared in terms of accuracy and error. The linear nature of these models does not allow them to understand the important details of the dataset. SVM Regression has the ability to find out the non-linear relationships by using different types of kernel functions. Even after that it shows a worse performance when it is compared to the other ensemble methods. The linear kernel that is used in the calculations, does not clearly identify the underlying patterns present in the data. Because of having a great performance in accuracy and being versatile in finding out complex relationships ensemble methods such as Random Forest Regression is recommended.



RESULTS				
Linear Regression	Values		Ridge Regression	Values
Accuracy	0.941580202		Accuracy	0.941565173
MAE	7.681880891		MAE	7.667474844
MSE	247.3332831		MSE	247.3969107
Precision	77.7231		Precision	81.508
Recall	0.8406		Recall	0.6853
F1 Score	1.6631		F1 Score	1.3592
Lasso Regression	Values		SVM Regression	Values
Accuracy	0.92978682		Accuracy	0.916554001
MAE	8.146380508		MAE	5.888995517
MSE	297.2632038		MSE	353.2872995
Precision	81.9978		Precision	75.5061
Recall	0.7206		Recall	0.7532
F1 Score	1.4286		F1 Score	1.4916
XGBoost Regression	Values		Random Forest Regression	Values
Accuracy	0.977777984		Accuracy	0.989220055
MAE	3.537521098		MAE	0.962519433
MSE	94.08187383		MSE	45.6393079
Precision	77.7231		Precision	81.9561
Recall	0.8406		Recall	0.9404
F1 Score	1.6631		F1 Score	1.8594

Figure 2. Results Obtained

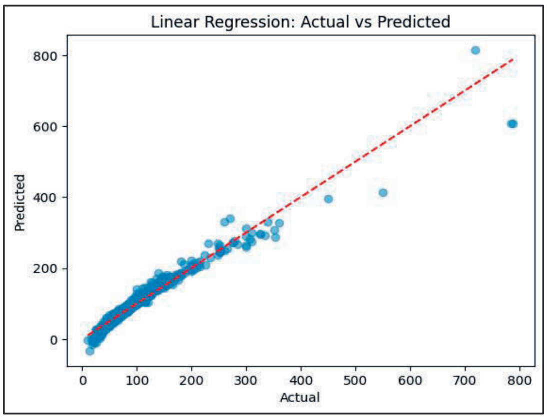


Fig. 3 Actual vs Predicted Plots of Linear Regression Model

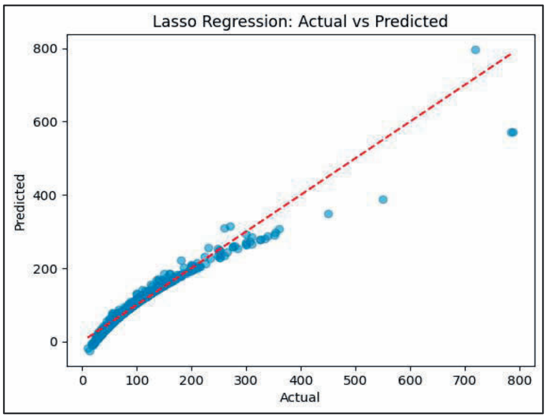


Fig. 4 Actual vs Predicted Plots of Lasso Regression Model

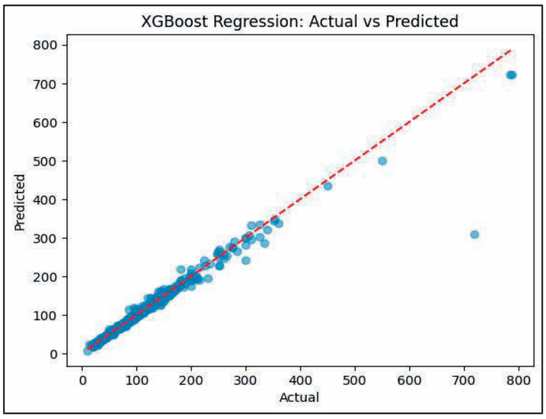


Fig. 5 Actual vs Predicted Plots of XGBoost Regression Model

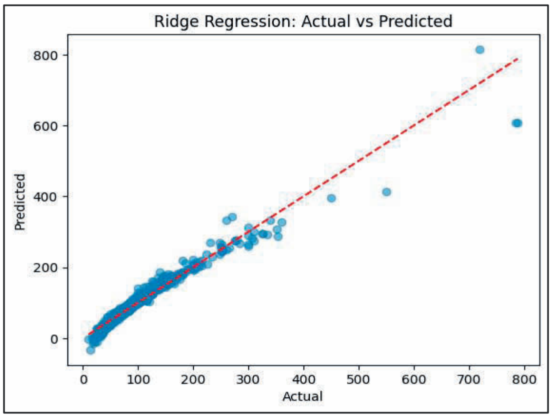


Fig. 6 Actual vs Predicted Plots of Ridge Regression Model

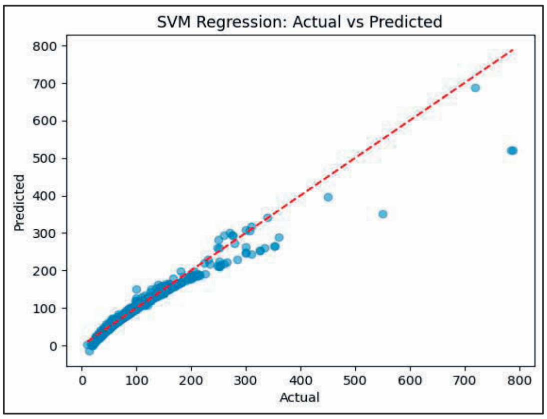


Fig. 7 Actual vs Predicted Plots of SVM Regression Model

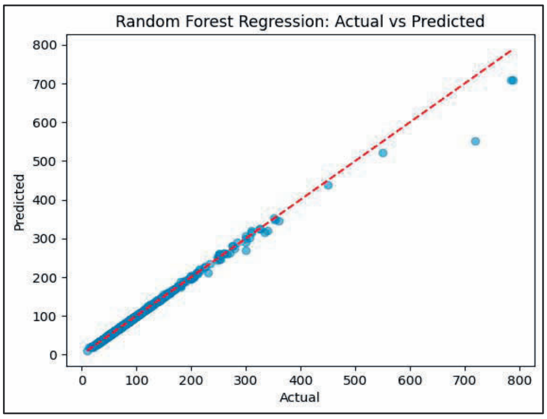


Fig. 8 Actual vs Predicted Plots of Random Forest Model

The red dotted line denotes the regression line.

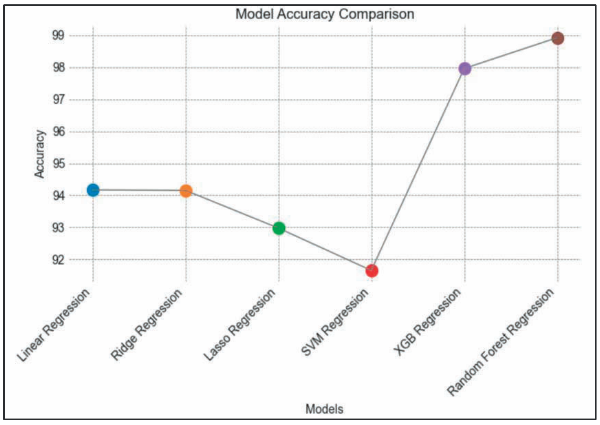


Fig. 9 Accuracy Comparison

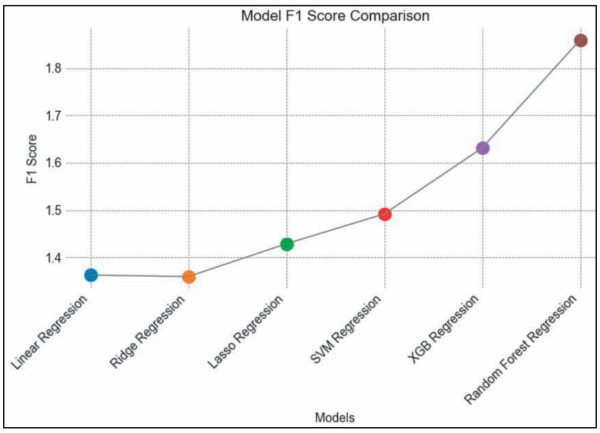


Fig. 10 F1-Score Comparison

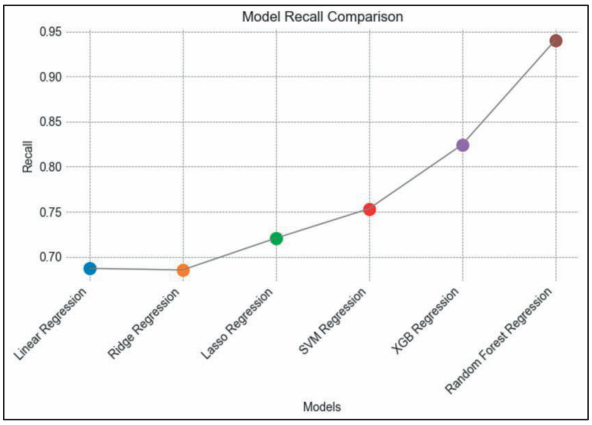


Fig. 11 Recall Comparison

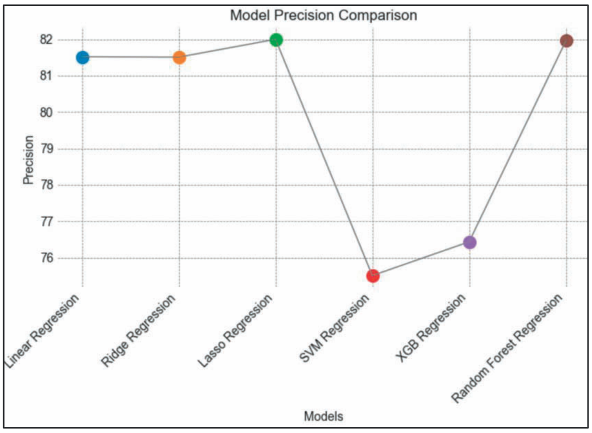


Fig. 12 Precision Comparison

VI. CONCLUSION AND FUTURE WORK

From the analysis, it is evident that Random Forest Regression outperformed other methods in terms of functionality and accuracy for price prediction. Despite testing the linear models and SVM Regression, Random Forest Regression was preferred due to its ability to handle complex relationships within the dataset. Moving forward, several enhancements can be done to further refine the model's accuracy and reliability. These include optimizing model parameters, exploring additional predictive methods, and employing feature engineering to create more informative features. Ensuring data quality and conducting various types of analysis will also be critical. Additionally, studying historical trends can provide valuable insights. These improvements will contribute to building a more robust model, ultimately aiding buyers and sellers in making informed decisions in the real estate market.

REFERENCES

[1] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, “Housing Price Prediction via Improved Machine Learning Techniques”, In: 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)

[2] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 2017. doi:10.1109/ieem.2017.8289904.

[3] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018. doi:10.1109/icmlde.2018.00017

[4] Rajesh Kumar, Praveen Ranjan Srivastava, “Hybrid Model for House Price Prediction Using Machine Learning and Sentiment Analysis”, in:2021

[5] Fadhil M. Basysyar, Gifthera Dwilestari, “House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection”, in: 2022.

[6] Xin Li, Shuai Zhang, and Xianfeng Tang, "Feature Selection Techniques for House Price Prediction: A Comparative Study", in: 2019.

[7] Sheng Chen, Jie Liu, Ying Shen, “Ensemble Learning for Housing Price Prediction: A Review”, in:2018.

[8] Fatemeh Mostofi, Vedat Togan, Hasan Basri Basanga, “Real estate price prediction with deep neural network and PCA”, (2022).

[9] Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow. 2nd ed. Birmingham: Packt Publishing; 2017.

[10] Xiangyang Ma, Pengfei Li, Hua Zhou, “Predicting House Prices Using K-Nearest Neighbor Algorithm”, in:2021.

[11] Dhanush Gowda R, “House Price prediction using machine learning”,in: 2024.

[12] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/ (accessed May 4, 2024)

[13] Mu J, Wu F, Zhang A. Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis 2014;2014:1–7. doi:10.1155/2014/648047.K. Elissa, “Title of paper if known,” unpublished.

[14] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing-ICMLC 2018. doi:10.1145/3195106.3195133.

[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825–30.

[16] Fatima Trindade Neves, Manuela Aparicio, Miguel de Castro Neto. “The impacts of open data and explainable AI on real estate predictions in smart cities”, in: 2024