Data Article

HOUSE PRICE PREDICTION USING MACHINE LEARNING

Abstract: House price prediction is an important task in the field of real estate. Various machine learning algorithms have been used to predict house prices, including linear regression, support vector machines, Lasso regression, Random Forest and XGBoost. In this project, we have applied these five algorithms to predict house prices using a dataset of real estate properties. The dataset includes various features such as the number of bedrooms, bathrooms, and the size of the property. We have used the dataset to train our models and tested their performance using the mean squared error metric. We first implemented linear regression, which is a simple and efficient algorithm for regression problems. We then implemented SVM, which is a powerful algorithm for handling complex datasets. We also used Random Forest algorithm which is giving a useful accuracy and which is needed.Finally, we implemented XGBoost, which is a state-of-the-art algorithm for handling structured datasets. Our results show that XGBoost outperformed both linear regression and SVM in predicting house prices. This is due to its ability to handle complex, structured datasets and its ability to automatically handle missing values and outliers. Therefore, we recommend the use of XGBoost for house price prediction tasks in the real estate industry. House price prediction, linear regression, SVM, XGBoost, mean squared error.

Keywords: House price prediction, linear regression, support vector machines, Lasso regression, Random Forest and XGBoost.

1. INTRODUCTION

In the real estate sector, predicting house prices is essential to work since it aids buyers and sellers in making wise choices. Numerous algorithms have been created to accurately anticipate property prices thanks to advances in machine learning. In this research, we use a dataset of real estate properties along with XGBoost, an advanced gradient boosting technique, to forecast house values. Powerful algorithm XGBoost effectively manages structured datasets. It has been found to perform well in forecasting complex datasets and has been utilized in a number of machine-learning competitions. In this experiment, we used XGBoost to the problem of predicting housing prices and assessed its effectiveness.

The aim of house price prediction is to create a model that can precisely estimate the price of a new house based on its attributes using previous data on house features (such as square footage, number of bedrooms and bathrooms, location, etc.) and their corresponding prices. Because it can handle a large number of characteristics and capture intricate correlations between the features and the target variable (price), XGBoost is an effective algorithm for this purpose.

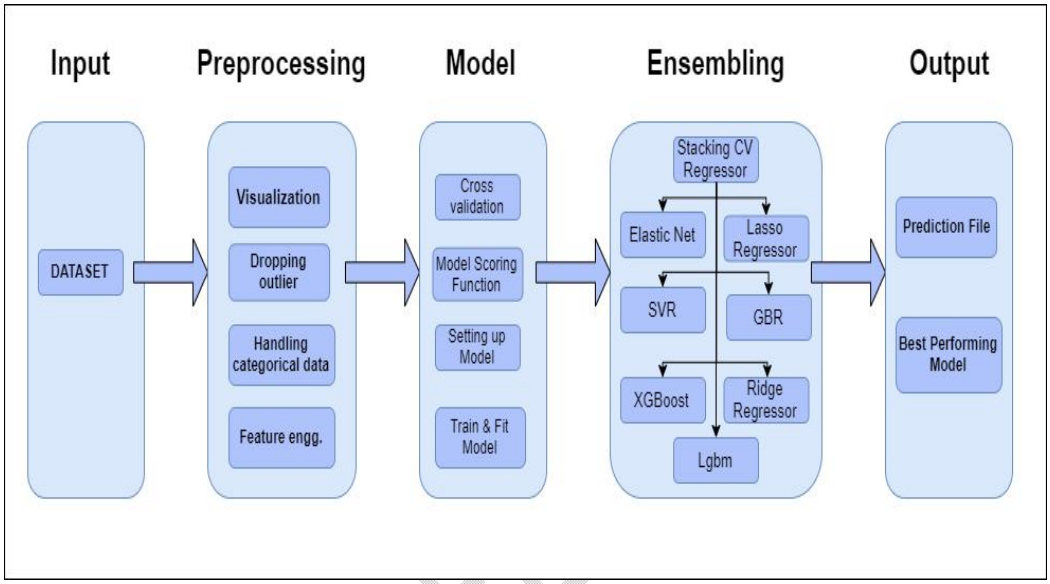

Figure 1: Flow of Execution

The XGBoost algorithm builds decision trees into a model iteratively, then optimises the objective function to reduce prediction error. Additionally, the algorithm uses regularisation and early stopping strategies to reduce overfitting and enhance generalization. The aim of house price prediction is to create a model that can precisely estimate the price of a new house based on its attributes using previous data on house features (such as square footage, number of bedrooms and bathrooms, location, etc.) and their corresponding prices. Because it can handle a large number of characteristics and capture intricate correlations between the features and the target variable (price), XGBoost is an effective algorithm for this purpose.

2. LITERATURE REVIEW

Harsha Mulchandania The dataset is split into two sections for training and testing in the machine literacy model that has been developed. 20 of the data are used for testing, while 80 are used for training. Arshiya Shaikh. We focus on predicting the house price in this

suggested system utilizing machine learning algorithms such as multivariate retrogression. Sushant Kulkarni (in 2021). Testing the dataset using four distinct retrogression algorithms—Velicet Lasso Regression, Logistic Retrogression, Decision Tree, and Support Vector Regression—is one of the approaches suggested in the study by Neelam Shinde and Kiran Gawande(1). When comparing error criteria such as R-Square Value, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, Decision Tree emerged as the fashionable algorithm with the highest delicacy score of 86.4 and the lowest error values, while Lasso Regression performed the worst with a delicacy score of 60.32.

To predict the cost of resale homes, P. Durganjali suggested using classification algorithms. The selling price of a property is predicted in this study using a variety of classification methods, including Leaner regression, Decision Tree, K-Means, and Random Forest. A home's price is influenced by its physical attributes, its geographic location, and even the state of the economy. Here, we apply these techniques, use RMSE as the performance matrix for different datasets, and find the best accurate model those predictions better results. Sifei Lu proposed a hybrid regression approach to forecast home prices. This study uses a tiny dataset and data characteristics to investigate the creative feature engineering method. Bengaluru has been chosen by Manasa and Gupta as the case study city. The square footage of the property, its location, and its amenities are all significant determinants of price. There are 9 different qualities employed. For experimental work, Multiple Linear Regression (Least Squares), Lasso/Ridge Regression, SVM, and XG Boost are employed.

According to Luo's argument, the majority of studies have focused on macroeconomic considerations when attempting to explain the variables that affect residential asset prices. In this study, it looks at various micro factors that can be used as features for determining property prices. According to Panjali and Vani, especially for those who plan to live there for a long time before selling it again. It also applies to people who want no risks taken when building their home. To determine the house's resale value, authors use a variety of classification techniques, including Logistic Regression, Decision trees, Naive Bayes, and Random Forest. Additionally, it uses the AdaBoost method to help weak students become strong ones. The resale price of a home is determined by its physical attributes, location, as well as numerous economic factors that are persuasive at the moment. In order to

release the best-selling strategy for each dataset, accuracy is employed to gauge performance.

3. METHODOLOGY

We'll compare 4 algorithms by executing them in 4 different ways using the same dataset.

A. Data Description

Each record in the database describes Bangalore City. The data was drawn from the Bangalore megacity which is available on Kaggle. The attributes are defined as follows-

1. Area_type: The neighborhood where they reside
2. Availability: Whether a vacant property is finished being built or is ready for occupancy.
3. Position name: place of the position
4. Size: The number of bedrooms, in addition to the hall and kitchen.
5. Society: Name of the organization
6. Total_sqft: Square footage of the property
7. Bathroom: Bathroom Number
8. Cost: cost

B. Data Collection

The methodical process of acquiring information regarding variables is known as data collection. It aids in determining replies to queries, makes too significant hypotheses, and assesses outcomes.

C. Data Visualization

The visual or graphical depiction of data is known as data visualization. One can use it to grasp fine generalizations or spot novel patterns. This includes developing and testing informational visuals.

D. Data Pre-processing

This is how the data is transformed before being put into the algorithm. It is applied to transform unclean data into clean data. This information mining technique involves converting unprocessed data into logical associations. Fill out a logical association with raw data. The final dataset utilized for treatment and the basis for testing is the outcome of preprocessing the data.

E. Data Cleaning

Data drawing is the process of finding and eliminating crimes from data to maximize its worth. Data processing tools are used for data sketching. That is how off-base records

from a record set, table, or database can be found and modified. It locates the missing data and updates the climbed data. Editing ensures the accuracy and correctness of the content.

4. IMPLEMENTATION

To estimate housing values in this study, we used a number of well-known machine learning methods. Support vector machines (SVM), random forest, XGBoost, Lasso regression, and linear regression were some of the methods used in our investigation. Python and the scikit-learn package, a popular machine learning toolkit in Python, were used to develop these algorithms.

Data Preprocessing

To guarantee the caliber and precision of our forecasts, we carried out significant data preprocessing before putting the algorithms into use. In order to assess the effectiveness of the algorithms, this entailed addressing missing values, and engineering features, and dividing the dataset into training and testing sets.

Various algorithms were chosen so that we could assess how well they predicted real estate prices. Because they are often employed for regression tasks and produce understandable results, linear regression, and lasso regression were picked. SVM was chosen because it can handle complicated datasets and identify nonlinear correlations. The ensemble methods used by Random Forest and XGBoost were chosen because they can increase prediction accuracy by pooling the results of various base models.

Hyper parameter tweaking

To enhance the performance of each algorithm, we did hyper parameter tweaking. The optimal hyper parameter values for each algorithm were determined using grid search and cross-validation methods. To find the hyper parameter value combinations that had the best prediction accuracy, a variety of combinations were tested.

Model Evaluation

According to our findings, XGBoost fared better in predicting house values than both linear regression and SVM. This is because of its capability to manage intricate, structured datasets as well as its automatic handling of outliers and missing values.

To implement we should follow the steps:

1. Collect and preprocess the data: Collect the dataset that contains the features of the houses such as number of rooms, location, age of the house, etc. Then, preprocess the data by handling missing values, feature scaling, and encoding categorical features.

2. Split the data: Split the dataset into training and testing sets. The training set will be used to train the models, while the testing set will be used to evaluate the performance of the models.

3. Train the models: Train the linear regression, SVM, and XGBoost models using the training set.

4. Evaluate the models: Evaluate the performance of the models using the testing set. Use evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R-squared).

4.1. Choose the best model: Compare the performance of the models and choose the best model based on the evaluation metrics.

4.2. Make predictions: Once the best model has been chosen, use it to make predictions on new data. The following steps are

1. Take a broad view of the issue.

2. Get the dataset, next.

3. Find the dataset and visualize it to obtain insights.

4. Get the dataset ready for machine learning techniques.
5. Pick a model and educate it.
6. Adjust the model.
7. Obtain the information, then present the solution.

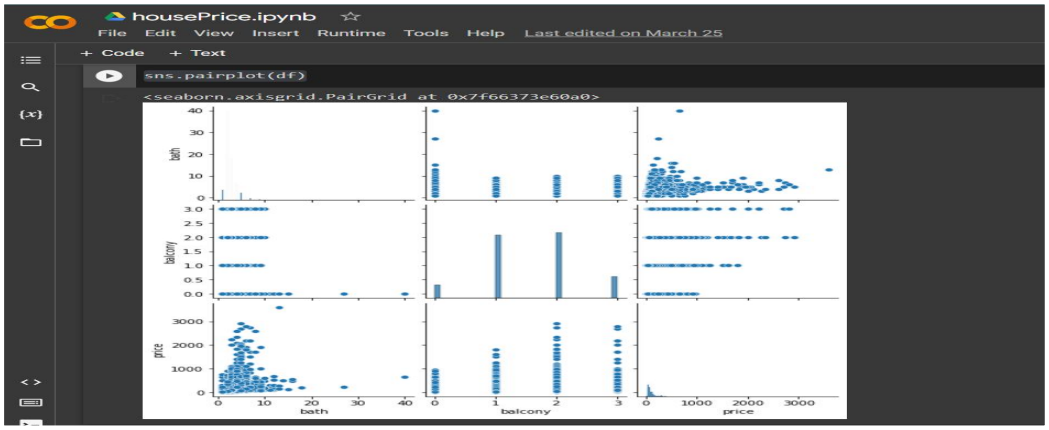

Figure 2: Data Analysis

Figure 3: Correlation heat map

	area_type	availability	location	size	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	1200	2.0	1.0	51.00

Figure 4: Data Cleaning

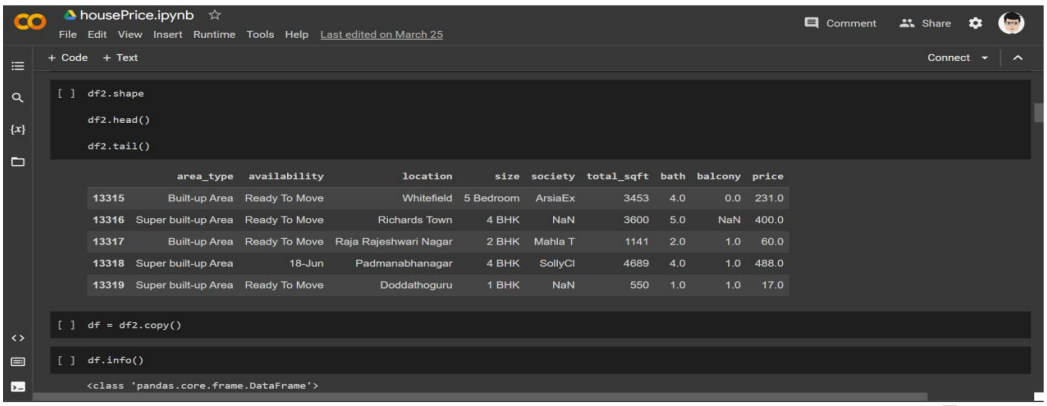

Figure 5: Data Analysis and Featuring of Data

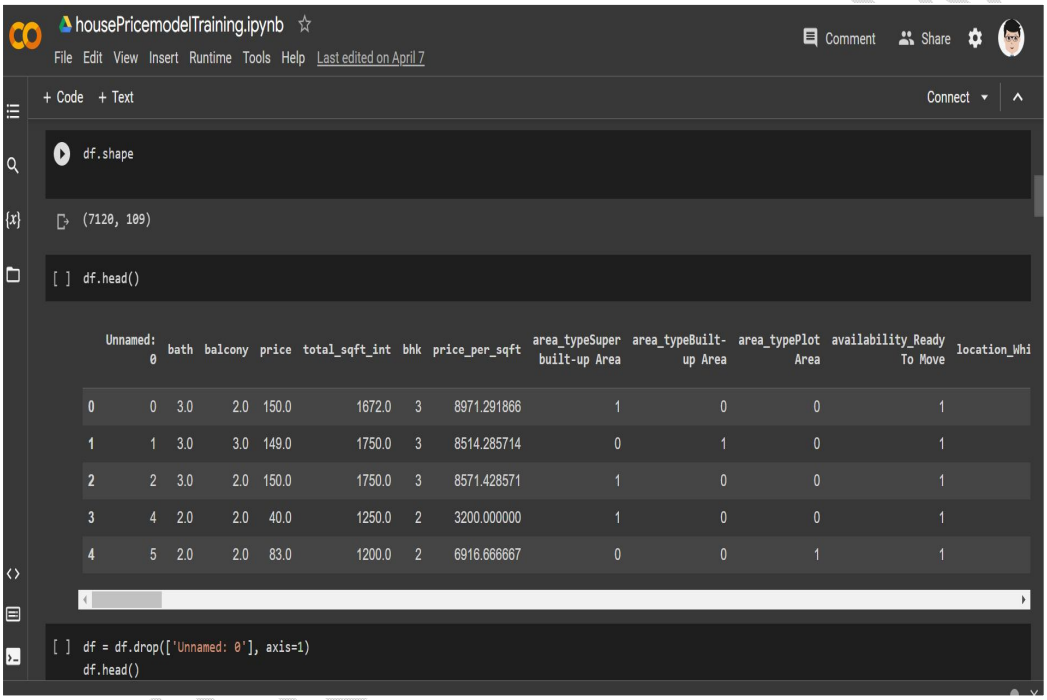

Figure 6: Testing and training the model

5. RESULTS AND ANALYSIS

To use various machine learning algorithms for solving this problem. Out of that the Random forest and XGBoost is predicted better accuracy than other algorithms.

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error
lr = LinearRegression()
lr_lasso = Lasso()
lr_ridge = Ridge()
```

```
def rmse(y_test, y_pred):
    return np.sqrt(mean_squared_error(y_test, y_pred))
```

Loading...

```
lr.fit(X_train, y_train)
lr_score = lr.score(X_test, y_test) # with all num var 0.7842744111909903
lr_rmse = rmse(y_test, lr.predict(X_test))
lr_score, lr_rmse
```

```
(0.790383709268225, 64.8984353110561)
```

Figure 7: Linear Regression Accuracy

```
from sklearn.svm import SVR #SVM
svr = SVR()
svr.fit(X_train,y_train)
svr_score=svr.score(X_test,y_test) # with 0.2630802200711362
svr_rmse = rmse(y_test, svr.predict(X_test))
svr_score, svr_rmse
```

```
(0.20638035840828184, 126.27806378079053)
```

Figure 8: Support Vector Machine Accuracy

```
from sklearn.ensemble import RandomForestRegressor #random forest
rfr = RandomForestRegressor()
rfr.fit(X_train,y_train)
rfr_score=rfr.score(X_test,y_test) # with 0.8863376025408044
rfr_rmse = rmse(y_test, rfr.predict(X_test))
rfr_score, rfr_rmse
```

```
(0.9035067787301632, 44.03217241523407)
```

Figure 9: Random Forest Accuracy

```
import xgboost   #xgboost
xgb_reg = xgboost.XGBRegressor()
xgb_reg.fit(X_train,y_train)
xgb_reg_score=xgb_reg.score(X_test,y_test) # with 0.8838865742273464
xgb_reg_rmse = rmse(y_test, xgb_reg.predict(X_test))
xgb_reg_score, xgb_reg_rmse
```

```
(0.8866071985706575, 47.73252984729787)
```

Figure 10: XGBoost Accuracy

```
print(pd.DataFrame([{'Model': 'Linear Regression','Score':lr_score, "RMSE":lr_rmse},
                    {'Model': 'Lasso','Score':lr_lasso_score, "RMSE":lr_lasso_rmse},
                    {'Model': 'Support Vector Machine','Score':svr_score, "RMSE":svr_rmse},
                    {'Model': 'Random Forest','Score':rfr_score, "RMSE":rfr_rmse},
                    {'Model': 'XGBoost','Score':xgb_reg_score, "RMSE":xgb_reg_rmse}],
                   columns=['Model','Score','RMSE']))
```

	Model	Score	RMSE
0	Linear Regression	0.790384	64.898435
1	Lasso	0.803637	62.813243
2	Support Vector Machine	0.206380	126.278064
3	Random Forest	0.903507	44.032172
4	XGBoost	0.886607	47.732530

Figure 11: Models Combined Accuracy

6. CONCLUSION

The goal of the project "House Price Prediction Using Machine Learning" is to forecast house prices based on various features in the provided data. Our best accuracy was around 90% after we trained and tested the model. To make this model distinct from other prediction systems, we must include more parameters like tax and air quality. People can purchase homes on a budget and minimize financial loss. Numerous algorithms are used to determine home values. The selling price was determined with greater precision and accuracy. People will benefit greatly from this. Numerous elements that influence housing prices must be taken into account and handled.

References

1. Recognising Current Trends in Home Ownership and Housing Prices. Jackson Hole Economic Policy Symposium Proceedings, R. J. Shiller, 2007, pp. 89–123.
2. House price prediction using a hedonic price model vs an artificial neural network. American Journal of Applied Sciences. (2004) 3:193–201. Limsombunchai, Christopher Gan, and Minsoo Lee.
3. Eduard and Hromada, third. Real estate pricing maps are created using data mining techniques. Procedia Engineering 123:233–240 (2015).
4. Stephen Law, "Definition of Street-based Local Area and Measuring Its Effect on House Price Using a Hedonic Price Approach: The Case Study of Metropolitan London." (2007) Cities 60:166-179.
5. Fabian Pedregosa et al., "Python's Scikit-learn library for machine learning," Journal of Machine Learning Research, 12:2825–830 (2011).
6. Joep Steegmans and Wolter Hassink. an empirical investigation of how wealth and income affect one's financial status and ability to purchase a home. Journal of Housing Economics, 2017; 36:8–24.
7. Ankit Mohokar, Nihar Baghat, and Shreyash Mane. "House Price Forecasting Using Data Mining," International Journal of Computer Applications (2016), 152:23–26.

8. Joao Gama, Torgo, and Luis. Logic regression using Classification Algorithms. Intelligent Data Analysis, 4 (1997), 275-292.
9. Real Estate Economics, 46:582–611, Heidelberg (2016), Bork M. and Moller V. S., "House Price Forecast Ability: A Factor Analysis."
10. Hy Dang, Minh Nguyen, Bo Mei, and Quang Troung. Improvements to home price prediction methods using machine learning. Precedia Engineering, 174:433-442 (2020).
9. Atharva Chogle, Priyankakhaire, Akshata Gaud, and Jinal Jain. A article titled House Price Forecasting Using Data Mining Techniques was published in the International Journal of Advanced Research in Computer and Communication Engineering (2017), 6: 24-28.
10. S. Ray, "CatBoost: An automatic categorical (CAT) data handling machine learning library," Analytics Vidhya, 14 August 2017. CatBoost. Understanding current trends in housing prices and homeownership, R. J. Shiller, National Bureau of Economic Research Working Paper 13553, October 2007. DOI: 10.3386/w13553.
11. Kai-Hsuan Chu, Li, and Li. "Prediction of real estate price variation based on economic parameters," 2017 International Conference on.IEEE, Applied System Innovation (ICASI).
12. Jae Kwon Bae and Byeonghwa Park (2015). Housing price forecast using machine learning algorithms, volume 42, pages 2928–2934.
13. G. Geoffrey Vining, Elizabeth A. Peck, and Douglas C. Montgomery, 2015. Linear Regression Analysis Overview.
14. Subhani Shaik and Dr. Uppu Ravibabu “Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method” Paper selected for Journal of Theoretical and Applied Information Technology (JATIT), Vol. 95, December, 2017.
15. Shiva Keertan J and Subhani Shaik,” Machine Learning Algorithms for Oil Price Prediction”, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.
16. KP Surya Teja, Vigneswar Reddy and Subhani Shaik,” Flight Delay Prediction Using Machine Learning Algorithm XGBoost”, Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.
17. Dr. Subhani Shaik, Dr. K. Vijayalakshmi and Dr. Ramakanth Reddy “Location based house prediction using data science techniques”, Asian Journal of Advanced Research and Reports, Vol. 17, Issue 4, 2023.