HOUSE PRICE FORECASTING USING MACHINE LEARNING TECHINQUES

Dr. Saravanan T M
Associate Professor
Department of Computer Applications
Kongu Engineering College
Erode, India
???

Rithika B M
PG Research Scholar
Master of Computer Applications
Kongu Engineering College
Erode, India
rithikamoorthy2001@gmail.com

Karan G
PG Research Scholar
Master of Computer Applications
Kongu Engineering College
Erode, India
Karansekar2233@gmail.com

Tamil Selvan R C
PG Research Scholar
Master of Computer Applications
Kongu Engineering College
Erode, India
tamilselvanrc@gmail.com

Abstract: Housing has become one of the most important form of investments. Everyone now wants to invest their money and buy houses. Due to this increased interest towards investments in housing, the price of the houses are set to change accordingly .This change can cause difficulty in predicting the price of the houses in a particular area as it could've either increased or decreased over a period of time. Due to this uncertainty, the people who want to buy a house may not be able to plan their finance accordingly. They could've planned to buy a house which satisfies their own personal requirements for a lesser capital but in fact it could be either higher or even lesser than they thought it would be. This not only affects the buyers but also the people who own houses and decide to use it for renting as they are not aware of the change in prices in a particular area and may decide to let people in for a lower rent. This problem is solved by us using Machine Learning the best machine learning algorithms are chosen and our dataset is trained using Linear regression and K-Means Alogrithm .The best model can be used to predict the prices of the houses and can help both the buyers and the renters.

*Keywords:*Linear regression,house price forecasting, K-Means

1.INTRODUCTION

In this ever changing world, one of the most important and constant desire of human beings is to have a nice shelter. This shelter in form of houses are also considered to be a good form of investment as they can be passed down to generations. Normally a person either buys or rents houses. The requirements for the house changes with respect the number of people in a family as more people require more rooms and more space. Some people also expect some other necessary requirements. They may want their houses to be in a colder or hotter place, near oceans and even in an area where the number of houses are less as they want less disturbances. Even the age of an individual may influence the data as young bachelors may not require bigger houses with many rooms when compared to older married people with children. Also one of the most important crucial factor would be the income of the person as people with higher incomes tend to go for houses with high prices. Even a land in particular popular area can cost more than that of a land in rural areas.Since there are numerous factors deciding the price of the houses we can't exclude any of them as the price is dependent on all of them. Since we have multiple variables we have to use the linear regression algorithm. But we can't get accurate results just from a single algorithm. A different algorithm may even perform better than multiple linear regression. So we also train our dataset with K-Means algorithms.

2. LITERATURE REVIEW

In [1] Dr. M. Thamarai et al. (2020) surveyed about the system of House Price Prediction Modeling Using Machine Learning. House prices are increasing every year which has necessitated the modeling of house price prediction. These models constructed, help the customers to purchase a house suitable for their need. Proposed work makes use of the attributes or features of the houses such as number of bedrooms available in the house, age of the house, travelling facility from the location, school facility available nearby the houses and Shopping malls available nearby the house location. House availability based on desired features of the house and house price prediction are modeled in the proposed work and the model is constructed for a small town in West Godavari district of Andhrapradesh. The work involves decision tree classification, decision tree regression and multiple linear regression and is implemented using Scikit-Learn Machine Learning Tool.

In [2] Darshil Shah et al. (2020) surveyed about the system of House Price Predication Using Machine Learning and RPA. Robotic Process Automation for real-time data extraction. Robotic Process Automation involves the use of software robots to automate the tasks of data extraction while machine learning algorithm is used to predict house prices with respect to the dataset. The system uses RPA to extract the data and also makes optimal use of machine learning algorithms which satisfies the customer by providing accurate output and preventing the risk of investing in the wrong house.

In [3] Swarali M. Pathak et al. (2021) surveyed about the system of Comparing of Machine Learning Algorithms for House Price Prediction using Real Time Data. The project tends to use Regression technique for Machine learning as we are dealing with continuous outcome variable. The most effective model to resolve given problem statement. The goal of this research project is to create an effective machine learning model that is able to accurately estimate the price of the house based on given features and deploy the machine learning model in the form of a website to reach out individuals.

In [4] Bindu Sivasankar et al. (2020) surveyed about the system of House Price Predication. The goal of the project is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. From our analysis we set the threshold value of RMSE as 0.12 and integrate those algorithms (Ridge regression, Lasso regression, XG Boost regression) with RMSE value less than 0.12. This definitely increases the accuracy. In future this paper may help in the upcoming development of these areas.

In [5] Anand G. Rawool et al .(2021) surveyed about the system of House Price Predication. the main aim of our project is to predict accurate price of house without any loss. There are many factors that have to be taken into consideration for predicting house price and try to predict efficient house pricing for customers with respect to their budget as well as also according to their priorities. Linear Regression, Decision Tree Regression, K-Means Regression and Random Forest Regression. This model will help people to put resources into a bequest without moving towards a broker. The result of this research provide that the Random Forest Regression gives maximum accuracy.

3. METHODOLOGY

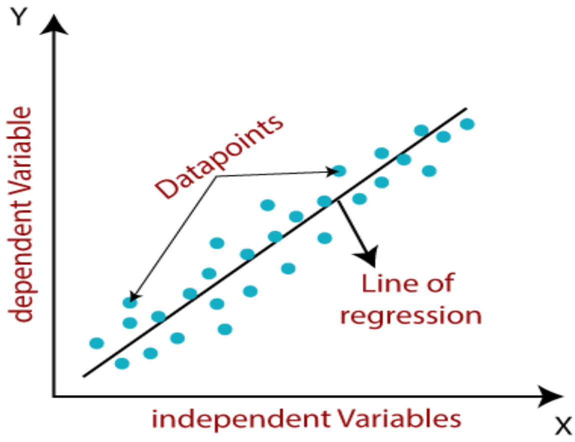
A. Regression Techniques

Machine Learning comes with two approaches of learning namely Supervised and Unsupervised Learning. Supervised Learning used needs that the data which is being used to the train the algorithm, has already some samples of correct outcomes. This is the most commonly used method as it increases the chances of getting much accurate results. We have performed supervised learning approach for this project. Supervised Learning is further divided into two groups Regression and Classification. The difference between the two is that the dependent attribute or outcome (predicted) variable is Numerical for regression which has numeric continuous value and Categorical for classification which has results in the form of categories.

Regression analysis is a type of predictive modeling technique that analyses the relation between the target or dependent variable and independent variables in a dataset. It involves determining the best fitting line that passes through all the data points in such a way that distance of the line from each data point is minimal. For most accurate Predictions we are trying different Regression techniques on given problem statement to find out best fitting model. This includes linear regression.

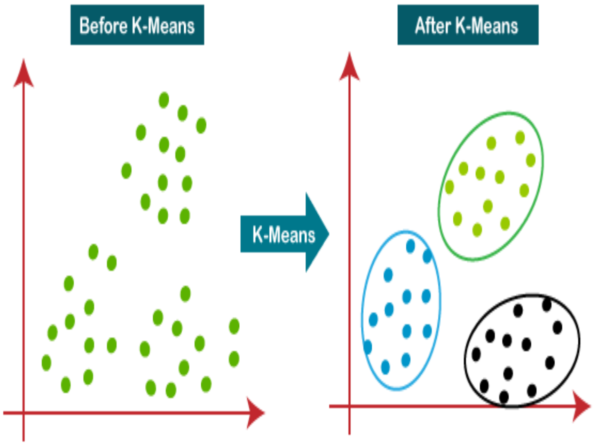
1.Linear Regression

The main aim of Linear Regression model is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is defined as the difference between the actual value and Predicted value. The goal is to reduce this error or difference. Linear Regression is of two types based on number of independent variables: Simple and Multiple. Simple Linear Regression contains only one independent variable and the model has to find the linear relationship between this and the dependent variable. Whereas, Multiple Linear Regression contains more than one Independent variable for the model to find the relationship with the dependent variable an Unsupervised Learning algorithm.



B. K-Means

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in K-Means clustering is the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. algorithm mainly performs two tasks : Determines the best value for K center points or centroids by an iterative process. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.



C. Gradient boosting

Gradient boosting is a machine Taking in strategy to relapse Also arrangement problems, that produces a prediction model in the structure of an group from claiming powerless prediction models. The exactness of a predictive model might be helped to two ways:. Possibly by grasping characteristic building alternately. Toward applying boosting calculations straight

far. There are a significant number boosting calculations in.

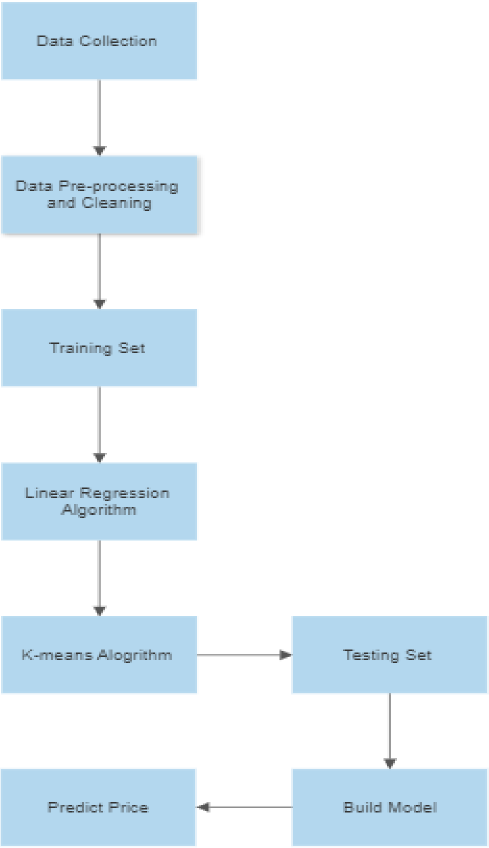
- Gradient Boosting
- XGBoost
- AdaBoost
- Gentle Boost etc.

Each boosting algorithm need its own underlying math. Also, a slight variety may be watched same time applying them. Boosting calculation will be a standout among those The greater part capable Taking in thoughts acquainted in the final one twenty A long time. It might have been intended to order problems, yet all the it can be developed should relapse too. The inspiration to gradient boosting might have been An technique. That combines those outputs about large portions “weak” classifiers to process An capable “committee. ” a powerless classifier (e. G. Choice tree) will be person whose slip rate is main superior to irregular guessing.

4.PROPOSED METHODOLOGY

The proposed framework has an implementation of several machine learning algorithms, which are then ensembled into the voting classifier. Comparison and analysis of various machine learning classifiers are done, and their outcomes are noted.

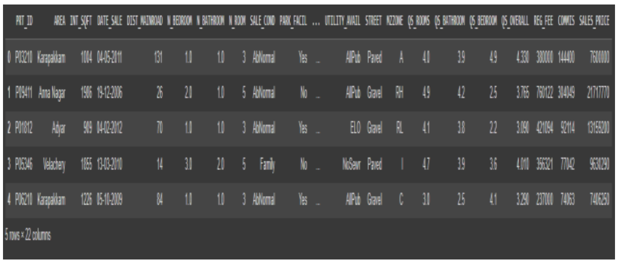
4.1 EVALUATING MODELS



4.2 DATA PRE-PROCESSING

After the manual collection of data through web scraping, there could be some mistakes in the collected entries, some null or blank values, human errors or some impractical values which we call as outliers. So to overcome these inaccuracies, we need to Pre-process and clean the data from these clutter values. There is a high need of Data Pre - processing because if the Data that we are providing to our model is accurate and faultless, then only the model will be able to give precise estimations which are very close the actual value. In Data Pre-processing and Cleaning, we remove the null values, take an overview of the dataset and also removal of unnecessary data columns (independent attributes) is done for the sake of accuracy and over fitting of the model.

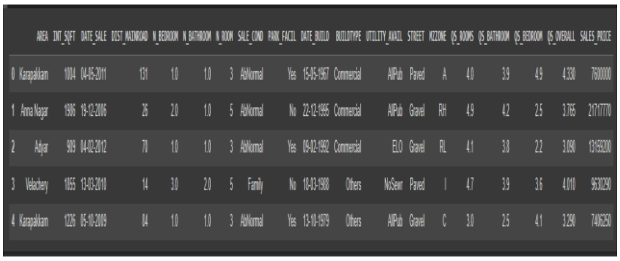
- Data cleaning - Process of removing incorrect values, duplicate data in the dataset.
- Data integration - Process of combining multiple data to create a unique set of information.
- Data reduction – Process of reducing the number of data records by eliminating the invalid data.
- Data transformation – Process of converting the raw data into structured format without affecting the original content



	PRT_ID	AREA	INT_SQFT	DATE_SALE	DIST_MAINROAD	N_BEDROOM	N_BATHROOM	N_ROOM	SALE_COND	PARK_FACIL	...	UTILITY_AVAIL	STREET	MZZONE	QS_ROOMS	QS_BATHROOM	QS_BEDROOM	QS_OVERALL	REG_FEE	COMMIS	SALES_PRICE
0	P03210	Karapakkam	1004	04-05-2011	131	1.0	1.0	3	AbNormal	Yes	...	AllPub	Paved	A	4.0	3.9	4.9	4.330	380000	144400	7600000
1	P09411	Anna Nagar	1986	19-12-2006	26	2.0	1.0	5	AbNormal	No	...	AllPub	Gravel	RH	4.9	4.2	2.5	3.765	760122	304049	21717770
2	P01812	Adyar	909	04-02-2012	70	1.0	1.0	3	AbNormal	Yes	...	ELO	Gravel	RL	4.1	3.8	2.2	3.090	421094	92114	13159200
3	P05346	Velachery	1855	13-03-2010	14	3.0	2.0	5	Family	No	...	NoSewr	Paved	I	4.7	3.9	3.6	4.010	356321	77042	9630290
4	P06210	Karapakkam	1226	05-10-2009	84	1.0	1.0	3	AbNormal	Yes	...	AllPub	Gravel	C	3.0	2.5	4.1	3.290	237000	74063	7406250

5 rows × 22 columns

Fig 1 Before pre-processing



	AREA	INT_SQFT	DATE_SALE	DIST_MAINROAD	N_BEDROOM	N_BATHROOM	N_ROOM	SALE_COND	PARK_FACIL	DATE_BUILD	BUILDTYPE	UTILITY_AVAIL	STREET	MZZONE	QS_ROOMS	QS_BATHROOM	QS_BEDROOM	QS_OVERALL	SALES_PRICE
0	Karapakkam	1004	04-05-2011	131	1.0	1.0	3	AbNormal	Yes	15-05-1967	Commercial	AllPub	Paved	A	4.0	3.9	4.9	4.330	7600000
1	Anna Nagar	1986	19-12-2006	26	2.0	1.0	5	AbNormal	No	22-12-1995	Commercial	AllPub	Gravel	RH	4.9	4.2	2.5	3.765	21717770
2	Adyar	909	04-02-2012	70	1.0	1.0	3	AbNormal	Yes	09-02-1992	Commercial	ELO	Gravel	RL	4.1	3.8	2.2	3.090	13159200
3	Velachery	1855	13-03-2010	14	3.0	2.0	5	Family	No	18-03-1988	Others	NoSewr	Paved	I	4.7	3.9	3.6	4.010	9630290
4	Karapakkam	1226	05-10-2009	84	1.0	1.0	3	AbNormal	Yes	13-10-1979	Others	AllPub	Gravel	C	3.0	2.5	4.1	3.290	7406250

Fig 2 After pre-processing

4.3 DATA VISUALIZATION

Data Visualization Visualization of data is an analytical skill exercised in machine learning. Data visualization helps to get a detailed understanding of various attributes of datasets. Several machine learning algorithms are perceptive to the range and distribution of the characteristics. It can help identify corrupt data, outliers, and patterns. It also aids to investigate the market drifts and assist the investor tomaximize this throughput. Here in our design, we have

deployed several graphs to classify the key relationships.

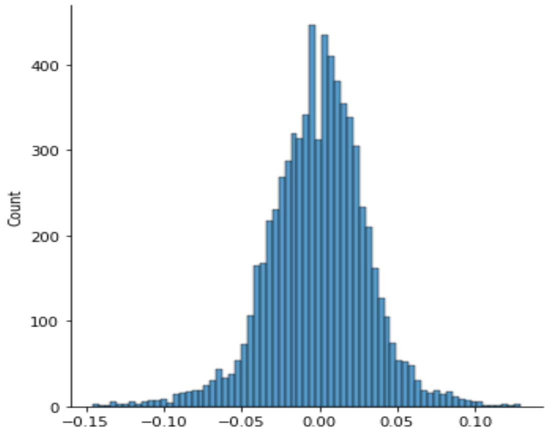


Fig 1 Data pre-processed graph

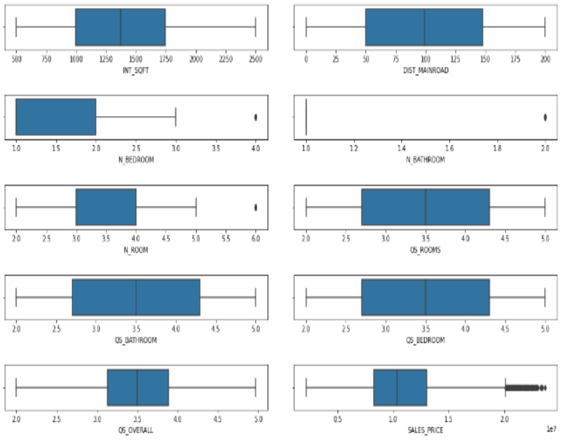


Fig 2 Outlier detection

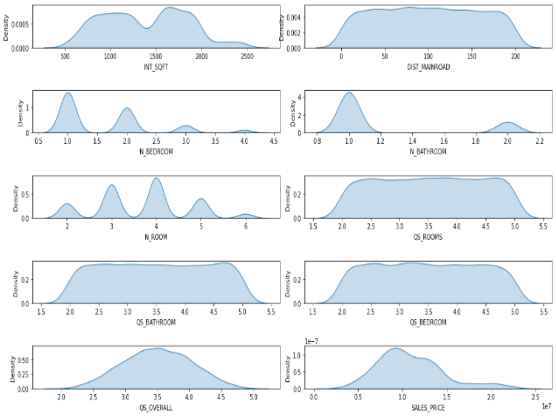


Fig 3 Distribution Overview

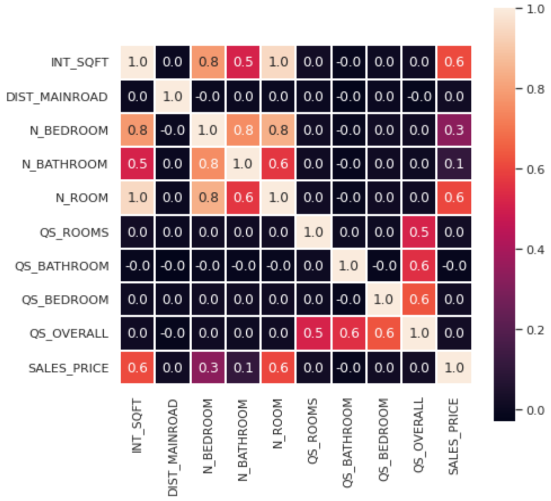


Fig 4 Correlation of numerical features

4.4 TRAINING & TEST SET

The given dataset should be split into two for training and testing our machine learning model. Normally 50% - 80% of the dataset is used as training set and the

rest is used for testing. The results from testing set is used for calculation of accuracy and scores. Stratified sampling is used to remove any bias while splitting the dataset into training and testing set.

4.5 CREATION OF MODELS

The training data set is fit into each models(Linear Regression,K-Means). Including all the independent variables to our models will decrease its score as it may contain least significant variables. So to remove the insignificant variables, backward elimination is used.

4.6 EXECUTION AND OUTPUTS

When the code gets executed first we get outputs plots and then prediction takes place. These plots help us to understand the correlation between target variable (price) and different predictor variables.

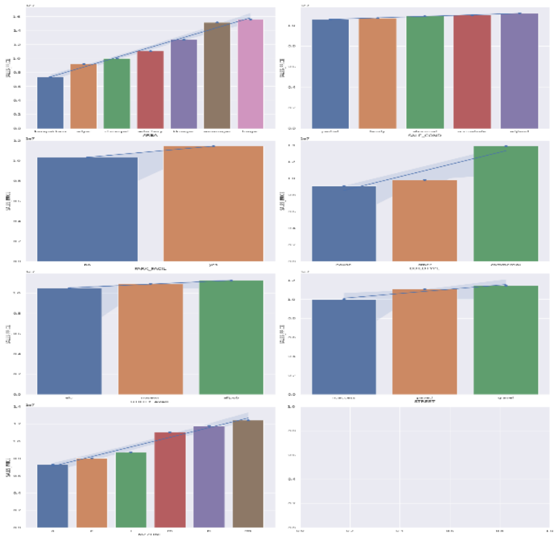
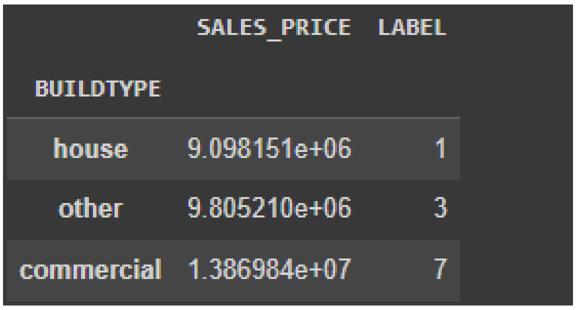


Fig 1 Bar plots for categorical features



```
                SALES_PRICE  LABEL
BUILDTYPE
house           9.098151e+06     1
other           9.805210e+06     3
commercial      1.386984e+07     7
```

Table 1 Sales price -label

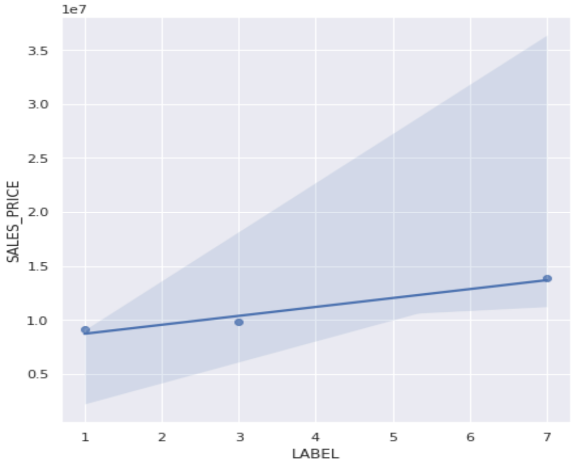


Fig 2 Sales price -label

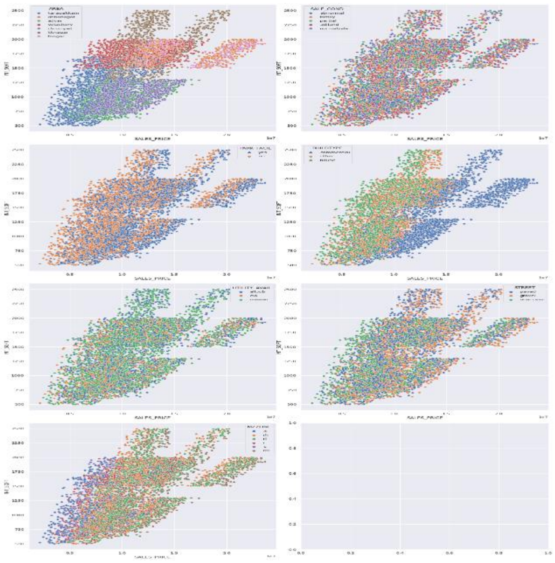


Fig 2 distribution of INT_SQFT and SALES_PRICE

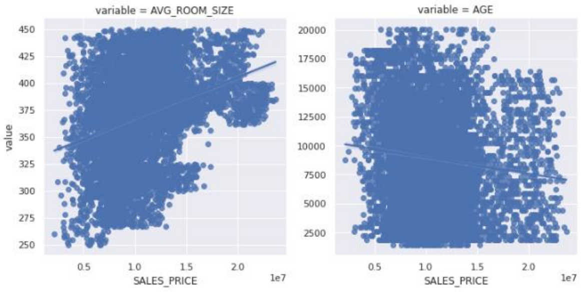


Fig 3 Liner corelation with Sales price

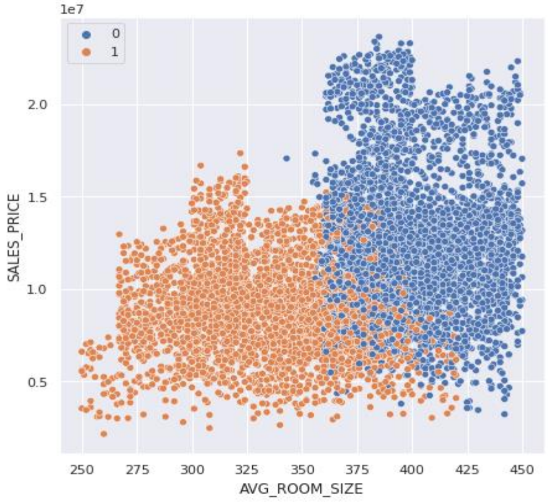
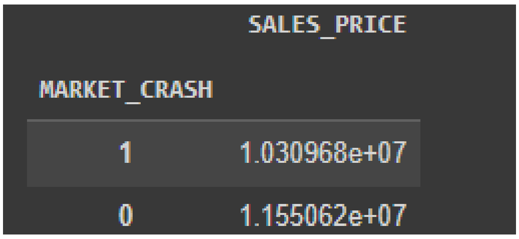


Fig 4 Average room size with sales price



	SALES_PRICE
MARKET_CRASH	
1	1.030968e+07
0	1.155062e+07

Table 2 Sales price -market crash

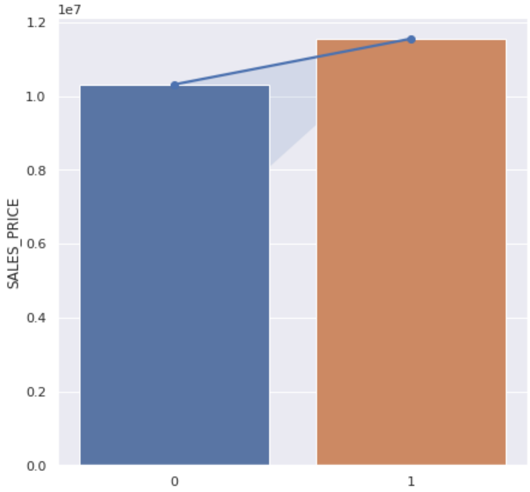


Fig 5 Sales price -market crash

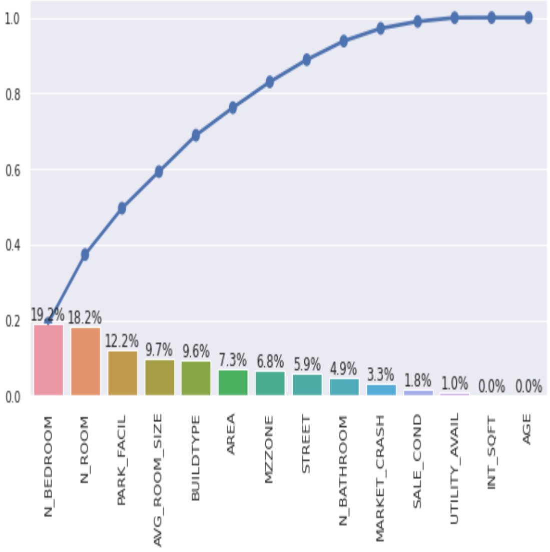


Fig 6 Feature Importance-Liner regression

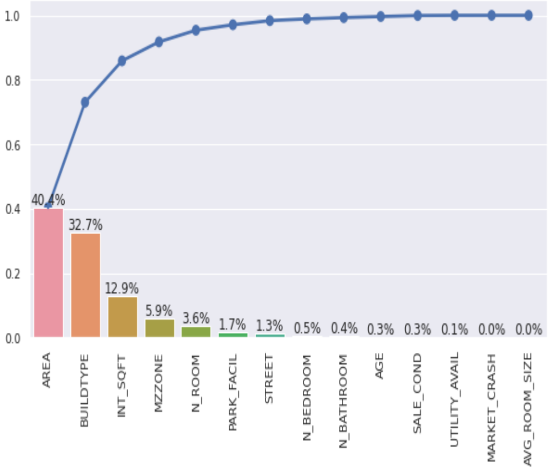
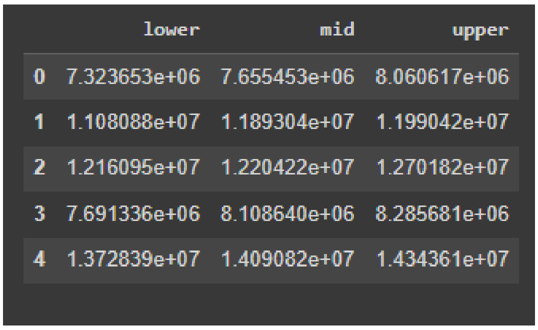


Fig 7 Feature Importance



	lower	mid	upper
0	7.323653e+06	7.655453e+06	8.060617e+06
1	1.108088e+07	1.189304e+07	1.199042e+07
2	1.216095e+07	1.220422e+07	1.270182e+07
3	7.691336e+06	8.108640e+06	8.285681e+06
4	1.372839e+07	1.409082e+07	1.434361e+07

Table 3 Final prediction

CONCLUSION AND FUTURE WORK

This paper investigates different models for house price forecasting.The most fundamental machine learning algorithm like Linear regression K-Means with Gradient Boosting.Work is implemented using Google colab machine learning tool.This work helps the user to forecast the price of the houses like the lowest,mid(average) and highest price of the particular house for 2 month duration .

In future the dataset can be prepared with more features and advance machine learning techinques,also increase the forecast duration period and even maintain the time complexity can be for constructing the house price forecasting model.

REFERENCES

[1] Udit., et al. “House Price Predication using Machine Learning.” *2021 Experimental Findings*. ResearchGate, 2021 .

[2] Elseviver B. V., et al. “House Price Predication Using Improved Machine Learning.” *International Journal of Advanced Computer Science and Applications 10.6* (2020).

[3] Robart Konwar, Bhagyashree Das, Angshuman Kakati. “House Price Predication Using Machine Learning.” *2021 International Journal of All Research Education and Scientific Methods*. IJARESM, 2021.

[4] Ozancan Ozdemir., et al. “House Price Predication Using Machine Learning : A Case in Lowa.” ResearchGate, 2022.

[5] Shuzlina Abdul Rahman., et al. “ House Price Predication Using A Machine Learning Model : A Survey Of Literature.” *I.J. Modern Education and Computer Science*, 2020.

[6] Anand G Rawool, Dattatray V Rogye, Sainath G Rane. “House Price Predication Using Machine Learning.” IRE Journals, 2021.

[7] Quang Truong, Minhnguyen, Hydang, BO MEI. “House Price Predication Via Improved Machine Learning Techniques.” *2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)*.ELSEVIER,2019.

[8] Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik. “House Price Forecasting Using Machine Learning.” *Proceedings of the 3rd International Conference on Advances in Science & Technology(ICST)*.SSRN,2020.

[9] Sumanth Mysore, Abhinay Muthineni, Vashnavi Nandikandi, Sudersan Behera. “Predication Of House Prices Using Machine Learning.” *International Journal for Research in Applied Science & Engineering Technology* IJRASET , 2022.

[10] G., Naga Satish, V., Raghavendran, M. D., Sugnana Rao. “House Price Predication Using Machine Learning.” *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), 2019.

[11] Bindu Sicasanka R, Arun P Ashok, Gouri Madhu, Fousiya S. “ House Price Predication.” *International Journal of Computer Sciencecs and Engineering*.JCSE,2020.

[12] Swarali Archana, Chaudha R. “Comparison of Machine Learning Algorithms For House Price Predication Using Real Time Data.” *International Journal of Engineering Research & Technology (IJERT)*,2021.

[13] Thamarai, M., and Malarvizhi S P. "House Price Prediction Modelling Using Machine Learning." *International Journal of Information Engineering & Electronic Business* 12.2 (2020).

[14] Patil, Pradnya, et al. "House Price Prediction Using Machine Learning and RPA." *International Research Journal of Engineering and Technology (IRJET), eISSN* (2020): 2395-0056.

[15] Modi, Maharshi, Ayush Sharma, and Dr P. Madhavan. "Applied research on house price prediction using diverse machine learning techniques." *International Journal of Scientific & Technology Research* 9.04 (2020).