IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

Housing Price Prediction using Machine Learning Technique

Ranjith S[1] and Dr. Ganesh D[2]
PG Student, Department of MSc CS-IT[1]
Professor, School of CS & IT[2]
Jain (Deemed-to-be University), Bangalore, India
ranjithsgowda06@gmail.com

Abstract: *As house prices increase every year, so there is a need for implementing machine learning in a real estate to develop a model that predicts the current house prices. House prices prediction can help the developers to determine the estimated selling price of a house and can help the buyers to arrange the budget at the right time to purchase a house. There are many factors that influence the price of a house positively and some impact negatively such as the age of the house and include physical conditions, concept and location. This paper presents a model that can generate predictions for housing prices by training the system with a 2001 data set so that it can learn how to forecast the market price of the house based on certain factors which have already mentioned bellow. The connection between the houses costs and the economy of the nation is a critical factor for predicting house prices as housing prices trends are not only the concern of buyers and sellers, but it also plays major role in the current economic situation. Therefore, it is important to predict housing prices to help all the stack holders such as the buyers and the sellers to make their decisions accordingly. In order to select a prediction method, various regression methods were explored such as a multiple linear regression will be chosen for this research proposed work due to its Flexibility and probabilistic approach to learning and model prediction is very high.*
Factors affecting the housing prices Including house age, longitude, latitude, number of rooms, number of restrooms, number of schools around the house, carports, parks, shopping centers, condition of the house, total number of floors a house has and other distinctive.

Keywords: ML multiple linear regression, gradient descent, Kmeans, RMSE, MSE, R squared, flask

I. INTRODUCTION

Housing market plays a big role in shaping the economy of the country therefore there's a requirement of implementing machine learning within the field of land because the housing renovation and construction boost the economy by increasing the house sales rate, employment and expenditures For many people, buying a property is one of the most important decision and purchase in life. Other than the moderateness of a house, different factors, for example, area where the house is located, age and the investment additionally influence the basic decision making process. Therefore by analyzing all these this project came up with a hybrid machinelearning prediction system that can learn from the data set to teach itself in order to make data driven predictionsaccordingly. There are different machine learning algorithms to perform predictions but this project will use supervised Multiple Linear Regression with gradient descent and unsupervised K-Means clustering to predict houses prices very accurately and minimizes error rate, The motivation for choosing Multiple linear regressionalgorithm is that it can accurately predict the result when there are more than one factors that influence the priceof the house. This project focuses on predicting the selling cost of the house based on numerous parameters likeYear designed, square feet, number of bedrooms and bathrooms and many features. The performance of this machine learning model is measured by four parameters of accuracy, SSE, RMSE, MSE, R squared. The proposed system is a hybrid model which is based on both supervised learning algorithm (Multiple linear Regression) and unsupervised learning algorithm (Kmeans algorithm) to cluster the data into groups then MLR obtains the predicted price of the house based on the 12 factors including (floors, bathrooms, bedrooms, squarefeet of the house, square feet of living room, year built, condition, basement ….etc) and the statistical relationship between these features and price of the house also obtained using MLR. This work also uses manytechniques like backward elimination, PCA,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

II. REALATED WORK

A similar study was done by Robin A. Dubin in 1998, in which he attempted to predict housing prices using 22 input variables. However, he only had a total data set of 1493 observations, so he had 1000 observations for making a predictor and the remaining 493 were used for testing. Interestingly, he noted that observations couldbe more correlated the closer the houses were geographically located to each other.

More recently, In [2] essential algorithms, for example, linear regression can accomplish 0.113 prediction errorsutilizing both characteristic highlights of the house properties (living zone, number of rooms) location and so on.) and additional geographical features (socio demographical features such as average income, population density, etc..).

III. PROPOSED SYSTEM

The proposed system is a hybrid model which is based on both supervised learning algorithm (Multiple linear Regression) and unsupervised learning algorithm (Kmeans algorithm) to cluster the data into groups then MLR obtains the predicted price of the house based on the 12 factors including (floors, bathrooms, bedrooms, squarefeet of the house, square feet of living room, year built, condition, basement ….etc) and the statistical relationship between these features and price of the house also obtained using MLR. This work also uses manytechniques like backward elimination, PCA, Gradient descent to optimize the performance of the prediction model. The chosen algorithm for this work are Multiple Linear regression and Kmeans clustering.

IV. METHODOLOGY

K-Means Algorithm

K-Means clustering is an unsupervised machine learning algorithm that is widely used method for cluster analysis to group the given data set into k clusters and returns the clustered data. In this proposed work both multiple linear regression algorithm (supervised algorithm) and K-Means algorithm (unsupervised algorithm) are combined to build a hybrid housing price prediction model that outperforms the existing models.

Before applying Multiple linear regression algorithm on housing data set, The data are pre-processed, standardized and PCA is applied on the and then given to K-Means algorithm to cluster the 2000 observations into three clusters labeled as cluster 0, cluster 1, cluster 2 hence optimal number of clusters is 3 which is obtainedby using elbow curve method.

Multiple Linear Regression

This proposed work uses Multiple linear regression for multivariate predictive analysis, therefore the value of dependent variable like housing price depends on many factors like longitude latitude, square feet of the house, number of bathrooms and goes no, we don't throw all the independent variables as a input to Linear regression algorithm instead we use various technique such as Feature selection and backward elimination formula. The final outcome of MLR is real valued price of the house in terms of Dollars

Multiple Linear Equation

y= a + (B1 * X1) + (B2 * X2).... + (Bp * Xp).

Let's test this out with an example!

Price = a+(B1*no_of_bathrooms) +(B2 * floors) + (B3 * longitude) + (B3* sqf_of_living_room)

Like this w have 13 features and each feature is multiplied with slop and then all results result are summed up to get the predicted price

Linear regression evaluation metrics used in the work are:-

- R Square (Coefficient of Determination) - $R^2 = 1 - SSE/SSTO$
- Adjusted R^2-
- MSE - MSE is the average of the squared error that is used as the loss function for linear regression
- MAE - This is mean absolute error

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

Gardinet descent algorithm : Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). Sum of Squared Errors (SSE)In order to fit the best intercept line between the points in the scatter plots, we use a metric called "Sum of Squared Errors" (SSE) and compare the lines to find out the best fit by reducing errors. The errors are sum difference between actual value and predicted value To minimize the residual errors OLS and Gradient descent has been used for optimizing the model performance of the housing price prediction model.

$$h_\theta = \theta_0 + \theta_1 x$$

$$\theta_0: \quad \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

Where

$$\theta_1: \quad \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)}$$

Fig 1.1 Gradient descent hypothesis function

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_0} \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Gradient hypothesis function for linear regression.

V. RESULT AND DISCUSSION

K-Means

This work combines both supervised machine learning algorithm (means algorithm) and unsupervised machine learning algorithms (Multiple linear regression, Gradient Descent) to build a powerful prediction model.

Means algorithm took 2001 housing observations and clustered into three cluster as shown in the fig 1.2 , assignment of the data point to the cluster is done based on the minimum Euclidean distance between the data point and the centroid. Number of cluster is obtained by using Elbow curve method, K=3

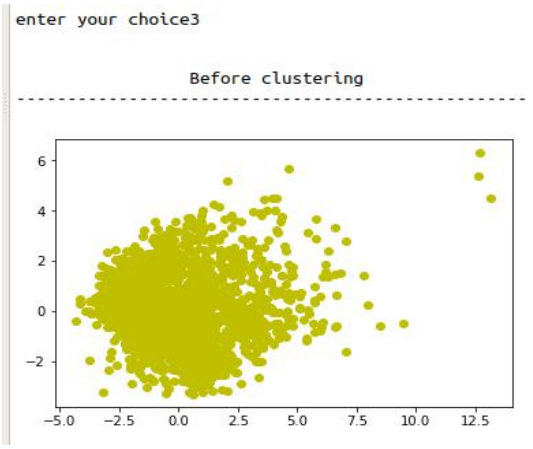
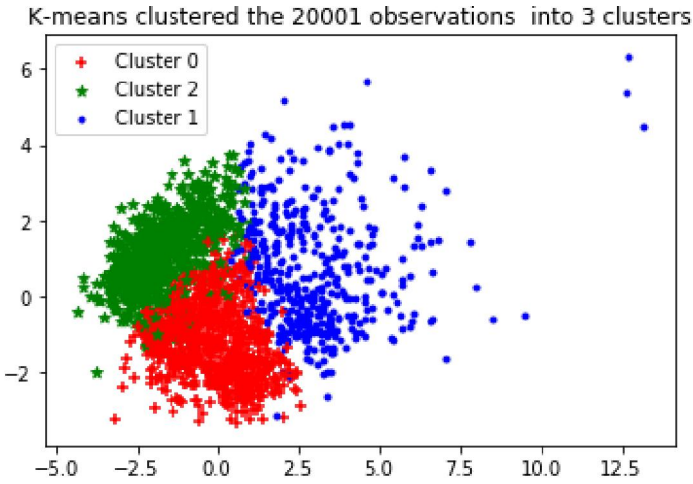


Fig Unclustered 2001 housing data

In fig Before applying Kmeans to the housing data set all the 2001 observation are scattered to form onecluster.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024




```
Kmeans clustered 2001 Real estate data

in 13.40812087059021 ms
```

Fig after applying kmeans to housing data

In figure Kmenas algorithm clustered the data into three clusters red, blue, green and labeled as 0,1,2 respectively. scatter graph in the above figure shows 3 clusters all the data points in one cluster have similar features and internally it creates cluster column which is obtained by using labels_ attribute of kmeans object.

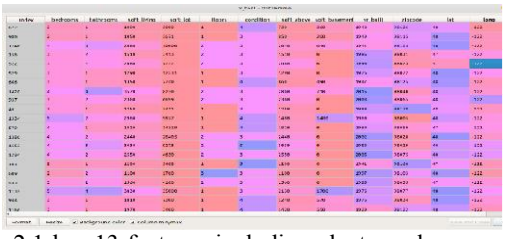
- All the data in cluster 0 are coloured with red.
- All the data in cluster 1 are coloured with blue.
- All the data in cluster 2 are coloured with green.

And the time taken by Kmeans algorithm to cluster 2001 data **is 13.5** millisecond

Finally the cluster column is obtained from kmeans algorithm and added to the original data set which will be used for training and testing our regression model.

Multiple Linear regression Result

As MLR is a supervised learning algorithm the price is removed from the data set and assigned to y pandas data frame object, and all the dependent variables are assigned to x pandas data frame object Splitting ratio used in thiswork is **80.20** which means 80% (1600) of the observations are used for training the model and 20%(4001) of theobservations are used for testing how well the model can predict the prices of the new houses which have not seenby the model in training phase



Original data which is shown in the fig 2.1 has 13 features including cluster column so step wise regression has been used to find the those features which have high significant on the price of the house in fig 7.4. To do this P value is compared with 0.05 if its greater than 0.5 then that feature will be dropped and the model is trained withthe remaining data this process is repeated until p is less than 0.05. And after OLS we train the model with thosefeatures which have p < 0.05

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

OLS Regression Results

Dep. Variable:	price	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	1420.
Date:	Tue, 16 Oct 2018	Prob (F-statistic):	0.00
Time:	19:27:53	Log-Likelihood:	-27619.
No. Observations:	2001	AIC:	5.526e+04
Df Residuals:	1992	BIC:	5.531e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bedrooms	-6.652e+04	7478.009	-8.896	0.000	-8.12e+04	-5.19e+04
bathrooms	7.271e+04	1.1e+04	6.633	0.000	5.12e+04	9.42e+04
sqft_living	193.3882	6.535	29.591	0.000	180.571	206.205
condition	2.749e+04	8421.893	3.264	0.001	1.1e+04	4.4e+04
sqft_above	113.7430	6.562	17.333	0.000	100.874	126.612
sqft_basement	79.6452	8.871	8.978	0.000	62.247	97.043
yr_built	-2302.9700	231.754	-9.937	0.000	-2757.475	-1848.465
zipcode	-540.9143	68.883	-7.853	0.000	-676.004	-405.825
lat	6.191e+05	4.06e+04	15.237	0.000	5.39e+05	6.99e+05
long	-2.295e+05	5.06e+04	-4.535	0.000	-3.29e+05	-1.3e+05

Omnibus:	1784.293	Durbin-Watson:	0.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	131478.018
Skew:	3.848	Prob(JB):	0.00
Kurtosis:	41.958	Cond. No.	5.16e+17

Fig OLS after removing the variables with high p value

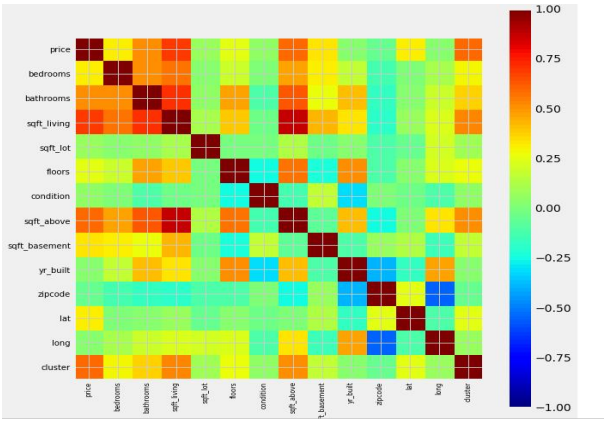


Fig Correlation matrix.

The above figure shows a Pearson's correlation matrix which shows the relation between the price of the house and other variables, large positive and negative value indicates a high correlation. Diagonals are colored with maron color which indicates high correlation between the variables but they are useless as it's a self correlation.

Rules for translating the Correlation coefficient:

- 0 shows no straight relationship
- +1 indicates a positive relationship –as one variable increases other variable decreases
- −1 indicates a negative relationship as one variable increases the other decreases

Fig 1.1 states the bedrooms, bathrooms, sqft_living, sqft_lot,cluster …etc have high impact on the price of the houses.
Features correlation plot

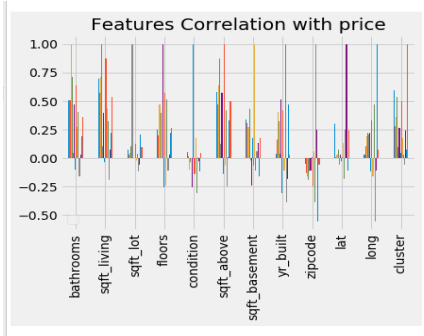


Fig Feature correlation with price factor.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

Feature correlation plot shows the correlation between the price of the house and other factors unlike fig 7.7 variables having positive one or close to +ve one are having high correlation with the price of the price.
In the event that we pass so much related variables to the model we may make the model to consider unnecessary features and we may have revile of high dimensionality issue so its important to visualize the features correlationplot to find variables which are having high impact on predicting the price of the house.

Accuracy and Variance

```
Database is  succesafuy closed

End of Reading the data from sqlite3 house table
                Linear Regression RMSE:  291028.7639



----------Evaluation of MLR model--------
          RMS 291028.76393904694
          r2 0.5954968043010864
```

Fig MLR score

The accuracy obtained from Multiple Linear Regression (MLR) before implementing gradient descent algorithmis 59% and RMSE is 290128 which is not as expected so we implemented Gradient descent to reduce RMSE andincrease the r squared value/explained variance

Gradient Descent with MLR result

```
gradient descent
                    RMS 136581.45452180353
                    r2 0.9109090491042903
                    MSE 18654493719.291504
Variance score: 0.9109090491042903
```

Fig Gradient Descent with MLR result

Gradient descent algorithm has increased the score from 59% to 91% and reduce RMSE to 136581.

Deployment of the model

To productionise the housing price prediction model so the clients can consume the prediction by providing their housing data to the model via web page. Bellow HTML page is used to accept user's data and we have created Flask restful API to make the communication possible between the client and housing price prediction model.
Request from the client is handled by the API and extracts the data from json object and pickled machine learningmodel in the server is unpickled and then the received data are passed as parameters to the model to make prediction and finally it returns a price of the house as a result which is serialized into json object and sent to theclient via the API.

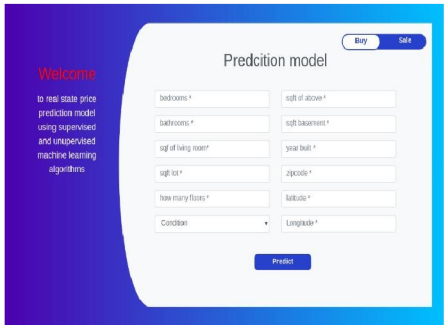


Fig Html page to that client uses to consume the prediction

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

VI. FUTURE ENHANCEMENT

The future work that we can contribute towards this work is to propose a method that uses time series data to obtain smaller error prediction values and using more data to get the expected result regardless of the location where the house is located.

For further analysis the Neural Network algorithms will be used in this work to improve the prediction power

VII. CONCLUSION

In this work, several machine learning techniques are combined to develop a prediction model for housing prices. Initially we clustered the housing data into three clusters to improve the performance of the MLR, and gradient descent algorithm was used for parameter tuning. The prediction power of the MLR determined by measuring how accurately a technique can predict whether the predicted price is close to the actual price. With all these techniques this problem overcomes the problems in some of the previous studies pertinent to housing price predictions have focused on hedonic-based methods which are conventional statistical approaches having some limitations of assumptions and estimations. However, with a combination of these approaches we were able to come up with an ensembling method that beat every hedonic based model by combining regression, clustered regression, Gradient descent regression, the accuracy obtained from this hybrid model is 91%.

Throughout this process we gained a solid understanding of the housing market as well as features that were most closely correlated with price, which could help us in developing a more generic forecasting model for predicting housing prices in different parts of the world in the future. Finally gained very deep understanding of various method how they work and how we might use them for future projects.

REFERENCES

[1] Validimir Bajic wive, 1985 ,“Housing- Market Segmentation and Demand for housing Attribute”, American Real Estate and Urban Economics Association, vol. 13(1), pages 58-75.

[2] John Quigley and Bradford case (1988),”Dynamic of real estate price prediction”, Vol. 73, No. 1 (Feb., 1991), pp. 50-58

[3] Hui Chen,Ting-Tzu, Chao Rung and John Francis ,(2014) , “tGray Relational Analysis (GRA) and Artificial Neural Network (ANN)”, ISBN: 978-9-8975-8149-6

[4] Clapp, John M. and Kim, Hyon-Jung and Gelfand, Alan E., Spatial Prediction of House Prices Using Lpr and Bayesian Smoothing (August 3, 2001). Available at SSRN: https://ssrn.com/abstract=288353 or http://dx.doi.org/10.2139/ssrn.288353

[5] Wu, Lynn and Brynjolfsson, Erik, (2013). The future of prediction: How Google searches foreshadow housing prices and sales. Available at SSRN 2022293, ICIS 2009 Proceedings. 147

[6] E. M´arquez-Chamorro , Gualberto Asencio-Cort´es , FedericoDivinaand Jes´us S. Aguilar-Ruiz, (2017) ,” Imbalanced classification techniques for monsoon forecasting based on a new climatic time series”

[7] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax county, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934

[8] Eman H, Ahmed and Mohamed Mustafa (2016), ”Price Estimation from Visual and Textual Features” , Proceedings of the 8th International Joint Conference on Computational Intelligence (IJCCI 2016) - Volume 3:
NCTA , pages 62-68ISBN: 978-989-758-201-1

[9] Khamis, A.B., Kamarudin, N. (2009). Comparative Study On Estimate House Price Using Statistical and Neural Network Model. International Journal of Scientific & Technology Research, vol. 19, no. 4, pp. 573-582, 2009. Available:[10] Raymond Y. C. Tse (2002), “Estimating Neighbourhood Effects in House Prices:
Towards a New Hedonic Model Approach”, Vol. 39, No. 7, 1165– 1180, 2002

[11] Charles A. Calhoun ,2003, “Property Valuation Models and House Price Indexes for the Provinces of Thailand” 1992–2000

[12] Ting Xu, (2008) "Heterogeneity in housing attribute prices: A study of the interaction behaviour between property specifics, location coordinates and buyers’ characteristics", International Journal of Housing Markets and Analysis, Vol. 1 Issue: 2, pp.166-181, https://doi.org/10.1108/17538270810877781

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

[13] Englund, Quigley and Redfearn , (1999)," Do Housing Transactions Provide Misleading Evidence About the Course of Housing Values?"