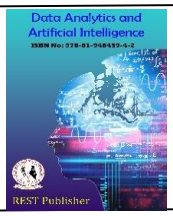
DOI: https://doi.org/10.46632/daai/2/4/2

Data Analytics and Artificial Intelligence
Vol: 2(4), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/



Enhancement in House Price Prediction using Advanced Regression Technique

*** K. Mahalakshmi, Ankam Usha Shree, Monica K, Madhumitha K**
Vel Tech Hightech Dr.Rangarajan Dr.Sakunthala Engineering College Chennai, Tamil Nadu, India.
*Corresponding author Email: mahalakshmi.k@velhightech.com

Abstract. House cost gauging is a significant subject of land. The writing endeavors to infer helpful information from authentic information of property markets. Computer based intelligence methods are applied to obvious property trades in India to track down important models for house buyers and vendors. Revealed is the high mistake between house costs in the most expensive additionally, most sensible rustic regions in the city of Mumbai. Additionally, tests show that the Numerous Direct Relapse that relies upon mean squared bumble assessment is a not kidding approach.
Keywords: House rent technique, Regression technique, House price prediction

1. Introduction

The land area is a significant industry with numerous partners going from administrative bodies to privately owned businesses and financial backers. Among these partners, there is popularity for a superior comprehension of the business functional component and driving variables. Today there is a lot of information accessible on significant measurements as well as on extra logical variables, and it is normal to attempt to utilize these to work on how we might interpret the business. Outstandingly, this has been done in Kaggle rivalries on lodging costs. Now and again, forward thinking factors have ended up being valuable indicators of land patterns. For instance, it is seen that Seattle lofts near specialty food stores, for example, Entire Food varieties encountered a higher expansion in esteem than normal. This venture can be considered as a further advance towards more proof based decision making to serve these partners. The point of our venture was to construct a prescient model for change in house costs in the year2021 in view of certain time and geology subordinate variable.

2. Literature Survey

Our House Value Expectation utilizing an AI Model: An Overview of Writing" December 2020 Worldwide Diary of Present Day Training and Computer Programming Information mining is at present ordinarily applied in the real estate market. Information mining's capacity to eliminate gigantic data from unpleasant information makes it particularly significant to anticipate house costs, key lodging credits, and some more. Research has bestowed that the qualifications in house costs are usually a concern for house owners and the real estate market. An assessment of making is done to seclude the reasonable properties and the most accommodating models to calculate the house costs. The disclosures of this assessment certified the utilization of the Fake Brain Organization, Backing Vector Re-slip by and XG Boost as the best models veered from others. Besides, our discoveries likewise recommend that location ascribes and underlying traits are unmistakable variables in foreseeing house costs. This study will be of enormous advantage, particularly to lodging designers and scientists, to find out the main ascribes to decide house costs and to recognize the best AI model to be utilized to direct a concentrate in this field. Expecting property costs with AI calculations" Winky K.O. Ho, &Siu Wai Wong Feb 2020, Acknowledged 01 Oct 2020, and Distributed on the web: 19 Oct 2020 this study utilizes three AI calculations including, support vector machine (SVM), inconsistent backcountry (RF) and propensity helping mother chine (GBM) in the appraisal of property costs. It applies these strategies to look at an information preliminary of around 40,000 lodging exchanges a time of north of 18 years in Hong Kong, and sometime later contemplates the results of these calculations. To the degree that insightful power, RF and GBM have accomplished better execution when stood apart from SVM. The three-execution assessments including mean squared mess up (MSE), root mean squared blunder (RMSE) and mean totally rate bungle (MAPE) related with these two computations likewise unambiguously beat those of SVM. In any case, our assessment has observed that SVM is now a strong assessment in information fitting since it can pass on sensibly careful guesses in-side a tight time need. Our decision is that AI offers a promising, elective procedure in property valuation and evaluation research particularly comparable to property cost figure.

3. System Analysis

Existing System: The work that exists in the past exploration directed by Gharehchopogh utilizing straight relapse approach get 0,929 blunders with the genuine cost. In straight relapse, deciding coefficients for the most part utilizing the un square strategy, however it requires a long investment to get the best recipe. They utilized four lose the faith approaches to be express Direct Regression, Support Vector Machine, K-Nearest Neighbours (KNN) and Random Forest Regression

and a party approach by joining KNN and Random Forest Technique for anticipating the property's expense regard. The social occasion approach expected the costs with least blunder of 0.0985 and applying PCA didn't further cultivate the guess screw up.

Disadvantages of Existing System: In current framework if any alteration is to be made in increments manual work and are blunder inclined. As the framework is overseen and kept up with by labourers, blunders are a portion of the conceivable outcomes.

Proposed System: The proposed system helps with inspecting and finished by considering the enlightening assortment that stays open to general society by addressing the available housing properties in machine hackathon stage. An undertaking has been made to fabricate a farsighted model for surveying the expense taking into account the factors that impact the expense. Demonstrating investigations apply some relapse strategies like different straight relapse (Least Squares), Displaying assessments apply some lose the faith frameworks like different straight apostatize (Least Squares). Lasso and Ridge backslide models. Such models are utilized to make a shrewd model, and to pick the best performing model by playing out an in general assessment on the farsighted botches got between these models.

Advantages of Proposed System: In this framework, we notice the assessment measurements acquired for cutting edge relapse models, we can say both act likewise. We can pick one for house cost assumption appeared differently in relation to major model. On the off chance that present, we can kill exemptions and really check out at the model execution for improvement. We can collect models through state of an art technique to be explicit unpredictable woods, cerebrum associations, and atom swarm headway to deal with the accuracy of assumptions.

4. Architecture Diagram

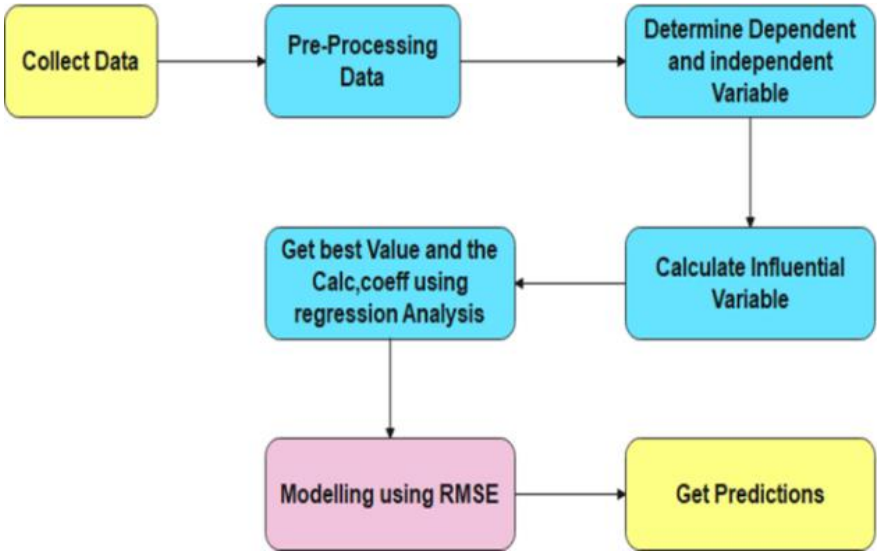


FIGURE 1 Proposed system architecture

The above architecture represents the flow of methods which is used to predict house prices. The assessment is done to pre-process the information and overview the suspicion accuracy of the models. The fundamental has various stages that should stop by the guess results. These stages can be portrayed as: Pre-handling: both datasets will be checked and pre-dealt with using the strategies. These methods have various ways to deal with dealing with data. Hence, the pre-taking care of is done on various accentuations where each time the precision will be evaluated with the used blend. Information parting: assigning the dataset into two regions is fundamental for train the model with one moreover; utilize the other in the assessment. The dataset will be isolated 75% for arranging and 25%for testing. Assessment: the precision of both datasets will be evaluated by assessing the R2 and RMSE rate while setting up the model nearby an appraisal of the genuine expenses on the test dataset with the costs that are being expected by the model. Performance: nearby the appraisal estimations, the vital opportunity to set up the model will be assessed to show the computation shift with respect to time. Connection: connection between the accessible elements and house cost will be reviewed utilizing the Pearson Coefficient Correlation to perceive whether the highlights have a negative, positive or no relationship with the house cost.

5. Algorithm

Linear regression: Direct relapse is one of the most notable calculations in insights and AI. The target of a direct relapse model is to track down a connection between at least one highlight (free/logical/indicator factors) and a nonstop objective variable (subordinate/reaction) variable. Assuming that there is just a single component, the model is basic straight relapse furthermore, assuming there are different highlights, the model is numerous straight relapse. It includes 3 stages: (1) Examining the relationship and directionality of the information, (2) Estimating the model i.e., fitting the line, and (3) Evaluating the authenticity and support of the model. Ridge regression: Edge fall away from the confidence model

is a regularization model, where a functional a variable (tuning limit) is added and progressed to promotion dress the effect of various parts in straight apostatize which is by and large avoided as unsettling influence in quantifiable setting. In mathematical turn of events, the model can be conceded as

$$y= xb +e$$

Here, y is the dependent variable x gathers features in network plan and b suggests break confidence coefficients and e keeps an eye out for residuals. Taking into account this, the elements are normalized by deducting the different factors and disseminating by their standard deviations. The tuning work showed as ƛ is then shown as a piece of regularization in the edge fall away from the confidence model. If ƛ‟ s regard is monster, then the squares ' remaining by firm has each of the stores of being zero. Expecting that it isn't exactly the plans. Conform to least square system. ƛ is sorted out using a system called cross - support. Edge lose the faith reduces the coefficients to considering no prominent extreme objective terrible characteristics in any case not to Zero. We will similarly perform network search cross-support to tune the regularization hyper limit λ. We have picked wide reach for hyper-cut off and considered to be 0.001 as best worth. The model got has an assortment of coefficients: Show ([2.86025752e-02, 4.16108845e-05 and 1.83397566e01, 5.65203565e-02, -3.85346357e-02, 4.89806025e-01])**.**

Lasso Regression: Performs L1 regularization, for instance adds discipline indistinguishable from out and out worth of the size of coefficients Minimization objective = LS Obj + α * (measure of inside and out worth of coefficients) Note that here LS Obj‟ insinuates „least squares objective‟, for instance the straight backslide objective without regularization Tether recommends least endlessly out shrinkage, and the choice executive is a LR technique that likewise regularizes use-fulness .It is indistinct from edge backslide; of course, really it varies in the potential gains of regularization. The all-around possible increases of how much lose the faith coefficients are taken into thought. It even sets the coefficients to nothing so it totally reduces the goofs. So determination of parts is occurred by tie break faith. In the really insinuated edge condition, the part ' e ' has completely credits rather than squared values. It is to be seen that computationally Lasso apostatize procedure is totally more raised than Ridge lose the confidence system. We have performed structure search cross-guaranteeing to tune the regularization hyper limit λ. We have picked wide reach for hyper confines and thought about 0.001 as best worth. The model acquired has an assortment of coefficients: Array ([2.26872602e-02, 3.84815301e05,1.88127460e-01, 6.35927974e-02, - 0.00000000e+00, 4.65240926e-01, - 5.40347504e-02,2.12552087e-01, 1.13374140e-01, 3.49669654e-07]).

Root mean square error: It is a root mean squared goof of the log ascertaining changed expected and log changed certifiable qualities. The blunder term can be detonated to an especially high worth if peculiarities are open if of RMSE, but if there should be an occasion of RMLSE the irregularities are beyond a shadow of a doubt.

6. Modules

Cleaning information: Information cleaning is the most common way of fixing or eliminating inaccurate, defiled, mistakenly designed, copy, or deficient information inside a dataset. Feature Engineering: Incorporate planning is the strategy associated with changing unrefined data into features that better locate the central issue to the farsighted models, achieving unrivalled model precision on subtle data. Variable Selection: It empowers the AI calculation to prepare quick, It reduces the assorted thought of a model and simplifies it to interpret, It deals with the exactness of a model expecting the right subset is picked, It decreases over fitting. Splitting the data and implementing cross validation: Divide your information into preparing and testing (80/20 is for sure a decent beginning stage), Part the preparation information into preparing and approval (once more, 80/20 is a fair parted), Subsample irregular choices of your preparation information, train the classifier with this, and record the presentation on the approval set. Regression Models: Apostatize models anticipate a value of the Y variable given known possible increases of the X parts. Assumption in-side the level of values in the dataset used for model-fitting is intimated nonchalantly as improvement. Measure outside this level of the data is known as Extrapolation.

7. Result

This examination shows that element designing can generally affect the model exhibition, and assuming the information are adequately huge, get approval ought to be liked over train-test-split to build model assessment. In my case, even though the predictors have high multi co linearity, their coefficients were not shrunk by the Lasso model, and it is shown that regularization does not always make big improvement on a given model. In the end, the Lasso regression has the highest R2 when predicting on the test set, and categories of car model appear to be the most important features to predict houses prices.

	INT_SQFT	DIST_MAINROAD	N_BEDROOM	N_BATHROOM	N_ROOM	QS_ROOMS	QS_BATHROOM	QS_BEDROOM	QS_OVERALL
count	7109.000000	7109.000000	7108.000000	7104.000000	7109.000000	7109.000000	7109.000000	7109.000000	7061.000000
mean	1382.073006	99.603179	1.637029	1.213260	3.688704	3.517471	3.507244	3.485300	3.503254
std	457.410902	57.403110	0.802902	0.409639	1.019099	0.891972	0.897834	0.887266	0.527223
min	500.000000	0.000000	1.000000	1.000000	2.000000	2.000000	2.000000	2.000000	2.000000
25%	993.000000	50.000000	1.000000	1.000000	3.000000	2.700000	2.700000	2.700000	3.130000
50%	1373.000000	99.000000	1.000000	1.000000	4.000000	3.500000	3.500000	3.500000	3.500000
75%	1744.000000	148.000000	2.000000	1.000000	4.000000	4.300000	4.300000	4.300000	3.890000
max	2500.000000	200.000000	4.000000	2.000000	6.000000	5.000000	5.000000	5.000000	4.970000

FIGURE 2 Collected dataset description

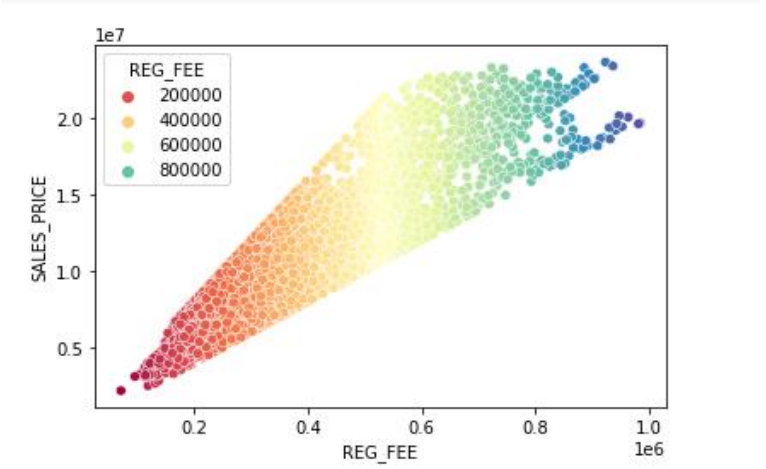


FIGURE 3.Sales_Price vs Reg_Fee

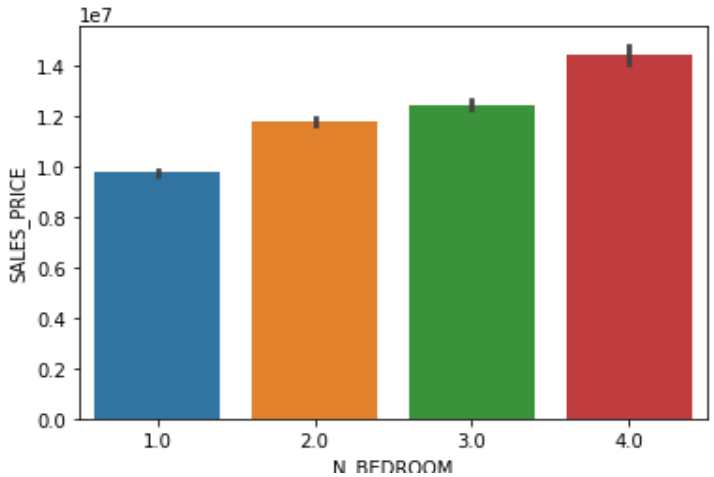


FIGURE 4. Sales_Price vs N bedroom bar plot

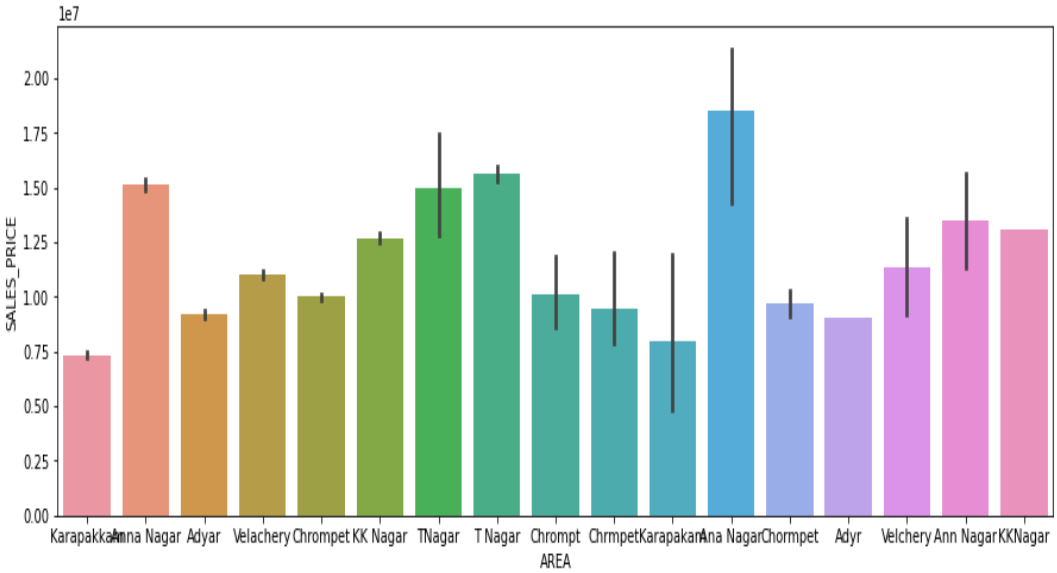


FIGURE 5.Barplots of various areas

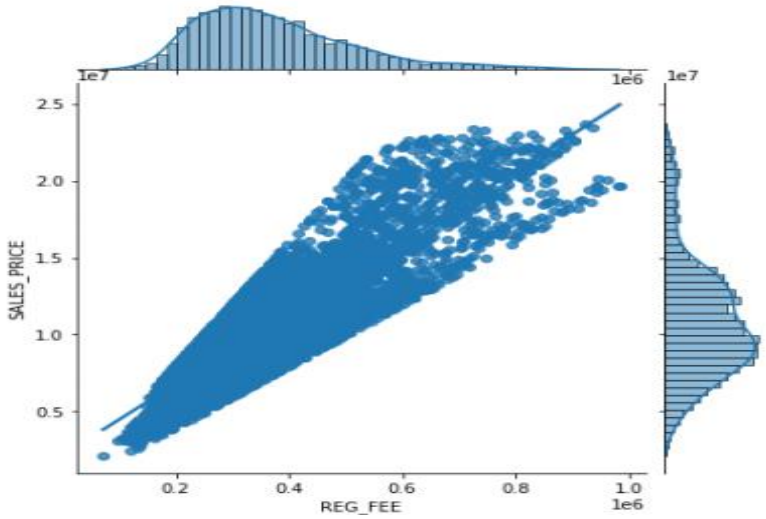


FIGURE 6 Jointplot of Sales_price vs Reg_fee

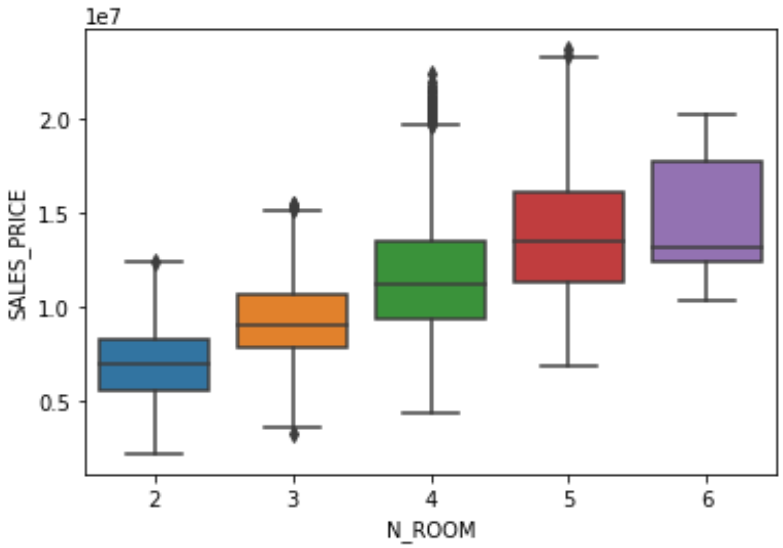


FIGURE 7 Boxplot of Sales price vs N Room

```
[ ]
    labels = ['Linear', 'Ridge', 'Lasso']

    metrics = {'MAE': mae_vals, 'MSE': mse_vals, 'RMSE': rmse_vals, 'R^2': r2_vals}
    metrics_df = pd.DataFrame(metrics, index=labels)

    metrics_df
```

	MAE	MSE	RMSE	R^2
Linear	0.126815	0.028447	0.168661	0.801900
Ridge	0.078689	0.012423	0.111459	0.916872
Lasso	0.136466	0.033465	0.182934	0.776076

FIGURE 8. comparison of various algorithms

8. Conclusion

This framework helps an ideal model doesn't be guaranteed to address a strong model. A model that every now and again utilize a learning calculation that isn't reasonable for the given information structure. Now and again the confirmed data might be unreasonably obviously or it could contain took a couple of tests to engage a model to definitively get the objective variable which recommends that the model additional parts fit. These models are used to make a shrewd model, and to pick the best performing model by playing out an overall assessment on the farsighted missteps got between these models. Future Scope: In future the framework will notice the assessment measurements got for cutting edge relapse models, we can say both act along these lines. We can pick one for house cost suspicion seemed vary ently according to basic model. With the assistance of box plots, we can check for exceptional cases. If present, we can dispense with extraordinary cases and really look at the model's show for progress. We can manufacture models through state-of-the-art strategies to be explicit unpredictable woods, mind associations, and atom swarm improvement to deal with the precision of assumptions.

Reference

[1]. Nor HamizahZulkifley, Ismail Ibrahim, House Price Prediction utilizing a Machine Learning Model: A Survey of Literature, received: 15 July 2020; Accepted: 22 October 2020; Published: 08 December 2020.

[2]. Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong, predicting property costs with AI assessments To suggest this article: Ho, Bo-Sin Tang and Siu Wai Wong (2021) Predicting property costs with AI assessments, Journal of Property Research, 38:1, 48-70, DOI: 10.1080/09599916.2020.1832558.

[3]. LuSifie et al, A blend break confidence technique at house costs speculation. In procedures of IEEE meeting on Industrial Endlessly arranging Management: 2017.

[4]. R. Victor, Machine learning project: Predicting Boston house costs with break confidence in towards data science.

[5]. G. Kiran, Valuation of house costs utilizing perceptive procedures, Internal Journal of Advances in Electronics and Computer Sciences:2018,vol 5,issue-6