EasyChair Preprint

№ 7089

Machine Learning Model for House Price Prediction

V Impana

November 26, 2021

Machine Learning Model for House Price Prediction

Impana V,Mtech Student
Vanishree K,Assistant Professor
Information Science and Engineering
RV College of Engineering
Banglore,India

Abstract **- Real estate is one of the transparent industry in our ecosystem. Housing prices keeps on changing based on various factors. In the current system, house prices are calculated without the necessary prediction about the future price increase and market trends. Predicting the house prices based on the real factors is one of the main mission of the project. Project mainly involves functioning of a website which accepts customer specifications combining the applications of linear regression algorithm and the results obtained are not the solo determination of one technique, it entails the weighted mean of various techniques which yields minimum number of errors and maximum accuracy than the algorithm which is applied individually.**

Sale prices for the homes in Banglore are predicted by applying various machine learning techniques. The dataset consists of the information about number of floors, square feet, location, and rooms.Representation of the data is done with the help of Python libray, Regression techniques such as Support vector regression and multiple linear regression are used here to build a predictive model comparing them on the various metrics such as R-Squared Value , Mean Squared Error(MSE),Mean Absolute Error(MAE) , Root Mean Squared Error(RMSE) .Here, the main goal is to build a model that evaluates the house prices based on various factors.

Keywords: Houseprice, MultipleLinearregression, Support vector regression

IV. INTRODUCTION

Machine learning is nothing but providing a valid set of the data and making predictions based on the data provided.Here,data is always the heart of innovations.A Machine learns by itself that how important a particular event may be, based on the data which is pre-loaded and makes some predictions accordingly. One of the principle point of machine learning is that they learn things naturally without the assistance of people.

The dataset has been taken from Kaggle website. Kaggle allows the users to build and explore the models in a web based environments.The housing dataset consists of 36000 records with distinct set of features. Multiple linear regression is one of the widely used techniques for accessing the relationship between the target variables and several independent variables. Data analysis is carried out by one of the efficient tool called sklearn. The proposed model focuses on building self learning algorithms as to project the evaluated solutions based on the previous data with low mean absolute error and High accuracy.

V. Related Work

Many of the research people have worked on the models based on the predictions Nitis Monburinon. [1] who applied linear regression technique to predict the prices of the cars where the model is trained based on the data of the used cars collected from German e-commerce website, and as a result the multiple linear regression technique gives the best performance with the mean absolute error of MSE=3D 0.55 respectively. According to E. Sreehari [2] on analyzing the R-squared value and p-value the evaluation of MLR prediction is done while by considering the ratio of the total variations of these outcomes ,R-squared

value measures at what extent the prediction model replicates the values of outcomes. Considering the total variations of these outcomes, the values always lies between 0 and 1 whereas one indicates the variability of the response around the mean and zero indicates no variance .The independent variables and dependent variables established that are established using the regression models should be less than 0.05.

Another research conducted by T . D. Phan[3], where the historical property transactions were analyzed using machine learning techniques. Its main aim was to get the helpful information from the dataset of property markets in Melbourne city of Australia, to discover the models that would be helpful to anticipate the estimation for a given set of the attributes. The study showed the high difference between the costs in most moderate and most costly rural areas in the city of Melbourne. Different regression techniques were obtained in order to obtain results. The study suggests that one of the efficient approach is the blend of stepwise and SVM that is established on a mean squared error measurement.

VI. System Design and Architecture

The design mainly consists of three stages: Initial stage, middle stage and final stage. Intial stage comprises of Data Analysis and accumulation. The middle stage consists of different sub-stages like Feature selection, Regression Model training, SVM display and validation of the model. The last stage comprises of the visualization of the product or the solution obtained.

Phase 1: Collection of data

The process of gathering and measuring the information with the help of the software is known as collection of data.Machine learning heavily depends on the data. It is one of the crucial aspects that makes algorithms to possibly get trained. Before machine learning analysis takes place, collection of data is a must which comprises of categorizing and collecting the structured quantitative data .The dataset in the proposed model consists of nineattributes :price, place, built, sqft, total-floors, bhk,years-old, sale,floor. There is no point of analyzing the data without validation, hence the data validity should be checked before hand.

Data collection is the most important aspect in the machine learning process, predictive models are only as good as the data from which they are built, so collection of the good data practices are crucial for developing high performance models.

	A	B	C	D	E	F	G	H	I	J	K
1	price	place	built	sqft	sale	yearsOld	floor	totalFloor	bhk		
2	6300000	BTM Layo	Super built-up	Area	1450	Resale	5	1	4	1	
3	11500000	Yelahanka	Super built-up	Area	2190	Resale	5	3	5	3	
4	3800000	Whitefiel	Super built-up	Area	1019	Resale	1	2	5	2	
5	10500000	Ambalipu	Super built-up	Area	1857	Resale	15	4	5	4	
6	11500000	Yelahanka	Super built-up	Area	2190	Resale	5	3	5	3	
7	15000000	Devarabe	Super built-up	Area	1672	Resale	10	6	10	6	
8	7350000	Yelachena	Super built-up	Area	1330	Resale	10	1	4	1	
9	15000000	Devarabe	Super built-up	Area	1672	Resale	10	6	10	6	
10	10500000	Ambalipu	Super built-up	Area	1857	Resale	15	4	5	4	
11	4900000	KR Puram	Super built-up	Area	1100	Resale	5	3	4	3	
12	3800000	Whitefiel	Super built-up	Area	1019	Resale	1	2	5	2	
13	3800000	Whitefiel	Super built-up	Area	1019	Resale	1	2	5	2	
14	15000000	Devarabe	Super built-up	Area	1672	Resale	10	6	10	6	
15	4025000	Electronic	Super built-up	Area	1543	Resale	10	3	4	3	
16	4025000	Electronic	Super built-up	Area	1543	Resale	10	3	4	3	
17	4800000	Abbaiah R	Built-up	Area	1200	Resale	10	1	4	1	
18	10500000	Ambalipu	Super built-up	Area	1857	Resale	15	4	5	4	
19	7900000	Subraman	Super built-up	Area	1784	Resale	5	18	18	18	
20	4025000	Electronic	Super built-up	Area	1543	Resale	10	3	4	3	
21	7350000	Yelachena	Super built-up	Area	1330	Resale	10	1	4	1	
22	11500000	Yelahanka	Super built-up	Area	2190	Resale	5	3	5	3	
23	10500000	Ambalipu	Super built-up	Area	1857	Resale	15	4	5	4	
24	4900000	KR Puram	Super built-up	Area	1100	Resale	5	3	4	3	
25	15000000	Devarabe	Super built-up	Area	1672	Resale	10	6	10	6	

fig 1 : Dataset collected from kaggle website

Phase 2: Data cleaning and loading

The process of measuring and gathering the information with the help of a software is known as Data cleansing. The training data containing various errors or the garbage values in the data set are common, and these errors can be removed by checking whether any missing values are present in data or not and also the value needs to be present in a particular range. If a variable has many values missing values we can drop these values. Data cleaning does not necessarily improve the model accuracy, performing model selection can atleast reduce any negative effects. Normalizing the data is also necessary before applying algorithm to it because every parameter has different units and the output will not be normalized.

Phase 3: Feature Selection

Feature selection is one of the core concepts in machine learning that hugely impacts the performance of the model. Partially relevant or irrelevant features can negatively impact model performance. It is the process which selects the features automatically or manually that contributes the most to prediction variable or the outputs. Machine learning follows the rule of garbage in-garbage out. Hence, the model should be fed with the data carefully.

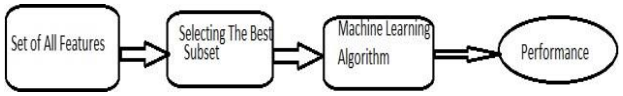

fig 2 : Data flow model for feature selection.

Phase 4: Train LR model

Training a model means learning good values from all the weights from labeled examples. The data will be split into two modules: Test set and Train set. Initially the data will be trained .In the proposed model 80% of the data will be considered as Train set and the Test set consists of 20% of the data. Thus linear regression tries to curve the model according to the given dataset with minimum errors. The training data should contain correct answers, which is known as learning algorithm and Target attribute finds the patterns in the input data attributes of the Training set to the target, and finally outputs the ML model that captures these patterns.

Phase 5: Validation of the model

Model validation is generally carried out after training the model. It is the process of checking whether the applied algorithm is the given dataset or not where the trained model is evaluated with a testing set .It compares the outputs from the system that is under consideration to the outputs that are obtained from the model. The values that are obtained from the model are recorded. It basically provides the generalization ability of the trained model. With the model Training, model validation aims to find the optimal model with the best performance.

Algorithm used: Multiple Linear Regression

Multiple Linear Regression is used to predict the value of a variable based on the value of two or more other variables. The variable that is to be predicted is known as dependent variable, while the variables that is used to predict the value of the dependent variable are known as independent or explanatory variables. The formula that is used for prediction in Multiple Linear Regression is :

$$yi = \beta0+\beta1xi1+\beta2xi2+\ldots\ldots+\beta pxip+€$$

yi is the predicted or the dependent variable,β0 is the y-intercept that is the value of y when xi and xi2 are 0, β1 and β2 are the regression coefficients, βp is the slope coefficient for each independent variable and € is the random error of the model.

There is a connection between needy target variable and autonomous variable vector. By utilizing the free parameters the objective variable can be anticipated. The autonomous information vector can be a vector of N properties or parameters. It accepts that the connection between subordinate variable and regressors is direct.

Regression residual technique ascertains the distinction between the vertical separation from the best fitting line and watched information.The MSE(Mean Squared Error) is a quality measure for the estimator by partitioning RSS by add up to watched information focuses. Linear regression will predict the exact numerical target which can only classify the output .It plays a strong role in predicting the price value of real estate property.

VII. Methodology

Predicting the house prices requires large number of factors such as sqft sale yearsOld, floor, total Floor, bhk, thus there are various factors which decide the price of the house. Therefore, it becomes difficult to use numerous variables which are dependent,target value is predicted using :Linear Regression Model.

$E(Y / X) = f(X, B)$

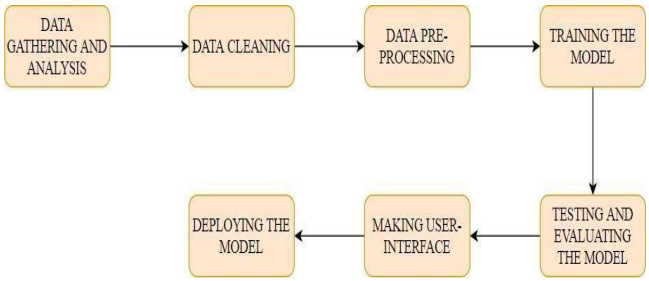

Fig 3 :Data Flow model

Machine learning is largely implemented to only take data that is in a numeric format as the input. Many of the data set are categorical and take a limited set of values. One example of nominal variable is "place" .Such categorical values cannot be interpreted by conventional machine learning algorithm without preprocessing them to numerical format.

The dataset is used in two ways: First to train the algorithm,and then to test it,and for these intents we have split the set in two. Using Five-fold cross validation is common practice, meaning 20% of the dataset is used for testing data and altered for five runs and 80% is used as training data. And at the final stage regression techniques is used to find and predict the prices based on various parameters.

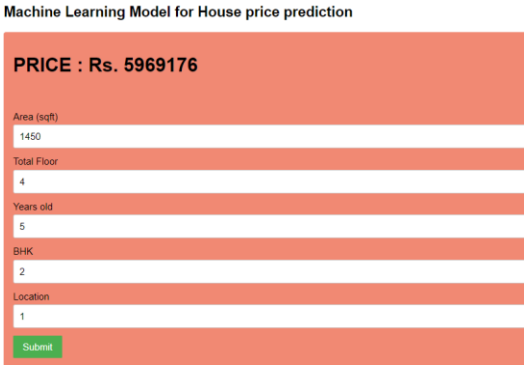

Fig 4: A Front end page for predicting house prices.

III. Conclusion

A system that provides the accurate pricing of the house prices has been developed .The system makes optimal use of Multiple Linear Regression . it has turned out to be difficult to store huge amount of information and concentrate them for one's own prerequisite. Likewise, the separated information ought to be helpful. The framework makes ideal utilization of the Linear Regression Algorithm. Additional features can be added to the system without disturbing it's core functionality for the customers benifit. The system will satisfy customers by preventing the risk of investing in the wrong house providing and by providing the accurate output .

IV. Future Work

A learning system can be created which will gather users feedback and history so that the system can display the most suitable results to the user according to his preferences . More factors like subsidence that influence the house costs should be included.Top to bottom subtle elements of each property could be added to give plentiful points of interest of a coveted domain.This will help the framework to keep running on a bigger level. Several more cites can be included in the system if the size and computational power increases of the system.

VI. References

[1] Eduard Hromada, "Mapping of real estate prices using data Mining techniques" ,Czech Technical University, Czech Republic , 2015.

[2] Pallav Ranka and Prof.Kripa Shanker, "Stock Market Prediction using Artificial Neural Networks".Indian Institute of Technology , Kanpur,2016.

[3] Adyan Nur Alfiyatin and Ruth Ema Febrita, "Modeling House Price Prediction using-Regression and Particle Swarm Optimization". International Journal of advanced Computer Science ,2017.

[4] Li and Kai-Hsuan chu, "Prediction of Real Estate Price Variation Based on Economic Parameters",Nankai University,Department of Financial Management,2017.

[5] Nissan Pow, Emil Janulewicz and liu Dave, "Applied Machine Learning Project for Prediction of real estate property prices in Montreal",2016.

[6] Dr. Swapna Borde, Aniket Rane, Gautam Shende and Sampath Shetty, "Real Estate Investment Advising Using Machine Learning," IRJET, 2017.

[7] Chen, J.-H. et al. Forecasting spatial dynamics of housing market using Support Vector Machine. International Journal of Property Management,2017.

[8] Risse M. & Kern M. Forecasting house-price growth in the Euro area with dynamic model averaging. North American Journal of Economics and Finance,2016.

[9] Ahmad I., Hussain M., Alghamdi, & Alelaiwi A.. Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components. Neural computing and applications(2019).

[10] Yu, H., Chen, R., & Zhang, G. A SVM stock selection model within PCA. Procedia computer science (2018).

[11] Jing, C., & Hou, J. (2015). SVM and PCA based fault classification approaches for complicated industrial process.

[12] Bork L. & Moller S., 2015. Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. International Journal of Forecasting, 31(1), pp.63 –78.

[13] Balcilar, M., Gupta, R., & Miller, S. M. (2015). The out -ofsample forecasting performance of nonlinear models of regional housing prices in the US. Applied Economics. [14] Park & Bae J. K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications (2015).

[15] Plakandaras, Gupta, & Papadimitriou, T.Forecasting the US real house price index. (2015).