IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

House Price Prediction Using Ensemble Learning

[1]Sushant Kulkarni, [2]Shefin Shajit, [3]Akshay Mohite, [4]Dr.Swati Sinha

[1]Student, [2]Student, [3]Student, [4]Professor
Department of Information Technology,
Vidyalankar Institute of Technology, Mumbai, India

Abstract: This paper presents a system that works on a set of data containing house prices of places in Mumbai along with the major parameters affecting the price such as area, location, swimming pool, etc. obtained from open web source Kaggle Inc. and predict the resale price under the parameters. The model implemented incorporates ensemble learning i.e. a combination of machine learning algorithms instead of relying on a single algorithm for improved predictions. The ensemble model incorporated in our system (weighted average of Decision Tree, Linear Regression, and K-Nearest Neighbor) brings an added advantage over using solo algorithms in the process of obtaining minimum error in prediction. The trained model demonstrated a Mean Absolute Percentage Error (MAPE) of 16.09%.

***Index Terms* - House Price Prediction, Linear Regression, Decision Tree, KNN, Ensemble Learning**

I. INTRODUCTION

The real estate sector is a major sector influencing India’s economy. In India, about 15 percent of the total jobs are generated by the real estate sector. In 2021, the sector must adopt innovative ways of dealing with the requirements. While houses will continue to be sold, they will now be done with creative disruption. The reinvention will include technology playing a lead role in meeting altered norms being considered by home buyers. It is important to have the best tools at disposal which would guide the buyers about where to put their money into. Since property prices rarely decrease rapidly, it is a major contender for investment. The property prices depend on various intrinsic and extrinsic factors which directly or indirectly affect the long-term price values. A fair share of India’s economic condition affects property prices in the long run. This scenario calls for technology to bring out the best ways to help out the customer’s investment decisions. A smart property investment decision about where to put the money depends mainly on three factors conditions, concept, and location. House price prediction can help the customer make a calculated risk in investment. The price point can vary per the square foot area, location, availability of swimming pool, lifts, etc. This paper explains a system that incorporates a dataset containing certain parameters that affect property prices in Mumbai, India. Further details about the dataset and the parameters incorporated are explained in detail in the dataset section.

II. LITERATURE SURVEY

Over the years there have been numerous approaches in predicting house prices per all the intrinsic and extrinsic factors affecting the price without any fluctuations. Although finding the best possible prediction model depends on the data available. The price of a property can differ based on its location, area, amenities, etc., and finding the best predictive model to predict that price has been a concern for researchers over the past decade. In our literature survey, we found various such approaches to find the house price using various models and a combination of models.

One of the methods proposed in the paper by Neelam Shinde and Kiran Gawande [1] includes testing the dataset with four different regression algorithms namely Lasso Regression, Logistic Regression, Decision Tree, and Support Vector Regression. On comparing the error metrics such as R-Squared Value, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, Decision Tree turned out to be the best algorithm giving a higher accuracy level of 86.4% and low error values whereas Lasso Regression performed the worst giving an accuracy level of 60.32%.

The majority of our work around ensemble learning for our predictive model has been widely accepted as a success in improving predictions [2]. Where ensemble models can vary depending on the needs of the data and the manner of predictions where even a small amount of improvement can have a big impact. The integration of two or more ensemble members depends on the type of integration the developer would see fit for the data i.e. Constant Weighting Functions and Non- Constant Weighting Functions.

A different approach has been taken in the paper [10]. The paper focused on linking various researches related to housing market prices analysis. Substantial focus is provided to hedonic price modeling and its application on the house price market and the possible submarket existence.

Decision Tree is used to make predictions in paper [5] after giving the highest accuracy in terms of prediction values among other algorithms tested namely Linear Regression, Multiple Linear Regression, Decision Tree Regressor, and KNN. Apart from parameters like no. of bedrooms, carpet area, built-up area, age of the property, zip code, no. of bathrooms, latitude, and longitude of the property, they have also included two other features – air quality and crime rate to better the prediction.

Another proposed system used Lasso and Random Forest regression techniques and picked the best model for the data depending on error values [3]. The data was passed onto 6 stages including data pre-processing, test-train 50:50 split, training the data with Lasso and Random Forest models and testing with the test data, and picking the best model.

In the paper by Prof. Pradnya Patil, Darshil Shah, Harshad Rajput, and Jay Chheda [4], the proposed system incorporates the UiPath Studio Platform to develop the RPA Flowchart. The UiPath Studio provides data scraping capabilities with the assistance of scraping wizards. A bunch of machine learning algorithms are compared and implemented on the dataset. A comparison between boosting algorithms is done namely XGBoost, Light BGM, and CatBoost. Random Forest was found to do well with small amounts of data and doesn’t improve accuracy with more samples. CatBoost was termed the clear winner in comparisons. RPA provided a major improvement in efficiency in terms of fast extraction and less prone to errors.

A further step has been taken in the paper by P. Durganjali and M. Vani Pujitha [8]. It analyses different classification algorithms such as Decision Tree, Logistic Regression, Random Forest, AdaBoost, Naïve Bayes with an accuracy of 92%, 81.5%, 86.5%, 96%, and 88% respectively. AdaBoost and Decision Tree using C 5.0 were selected to predict values of the house and using rules, they predicted profit or loss.

Detailed study of different machine learning algorithms namely Multiple Linear Regression, Elastic Net Regression, Ridge Regression, Ada Boosting Regression, LASSO Regression, and Gradient Boosting has been done on a public output dataset of a specified region in the USA [9]. The attained scores of the algorithms were 0.73, 0.66, 0.73, 0.78, 0.73 and 0.97 respectively. Gradient Boosting turned out to be the best algorithm as it gave low error values.

III. Dataset

The dataset incorporated in the system is taken from a public dataset source Kaggle Inc. It has data for house prices and features from 413 unique locations of Mumbai, India. It consists of 6347 records with 17 parameters that have the possibility of affecting the property prices. However, out of these 17 parameters, only 7 were chosen (Area, No. of Bedrooms, New/Resale, Gymnasium, Lift Available, Car Parking, Swimming Pool) along with 2 added parameters (Location Id and Price Area) which are bound to have a major effect on housing prices. The area is the total built-up area in square feet. New/Resale specifies if the property is a resale property or a new property. Gymnasium, Lift, Car Parking, and Swimming Pool mention if the property happens to provide these amenities (Binary value i.e. 1s and 0s). Location id is a unique id to all the locations present in the dataset in ascending order of Price Area. Price Area is the average price per area of a location.

3.1 Data Pre-processing

Data Pre-processing is a major step in transforming the dataset into an efficient format. This includes removing the NaN values (Missing Values) and improper noisy data in the dataset to narrow down the training of the model to the appropriate level. Records containing NaN values were deleted in our dataset to make it fit for training the ensemble learning model.

3.2 Data Analysis

Every single parameter in the dataset is analyzed with every other parameter to check the dependence and correlation using an SNS heatmap. The correlation is measured in the range of -1 to +1 where a higher absolute score shows better correlation and the lower absolute score worse the relation. Below Fig.1 depicts the correlations among the 17 parameters which affect the house prices. We removed parameters that gave bad scores and kept in a total of 7 parameters that majorly affected house prices namely, Area, No. of Bedrooms, New/Resale, Gymnasium, Lift Available, Car Parking, Swimming Pool, Location Id, and Price Area.

We removed outliers from the dataset using Interquartile Rule. On multiplying 1.5 (a constant to discern outliers) to the Interquartile Range (IQR) we set a limit to values outside Q1 and Q3 (Quartile 1 and Quartile 3 respectively). Practically any value that exists outside the fence i.e., values greater than 1.5 * (IQR) below Q1 or greater than 1.5 * (IQR) above Q3 are termed as outliers. Removing outliers helps to make the predictions statistically significant.

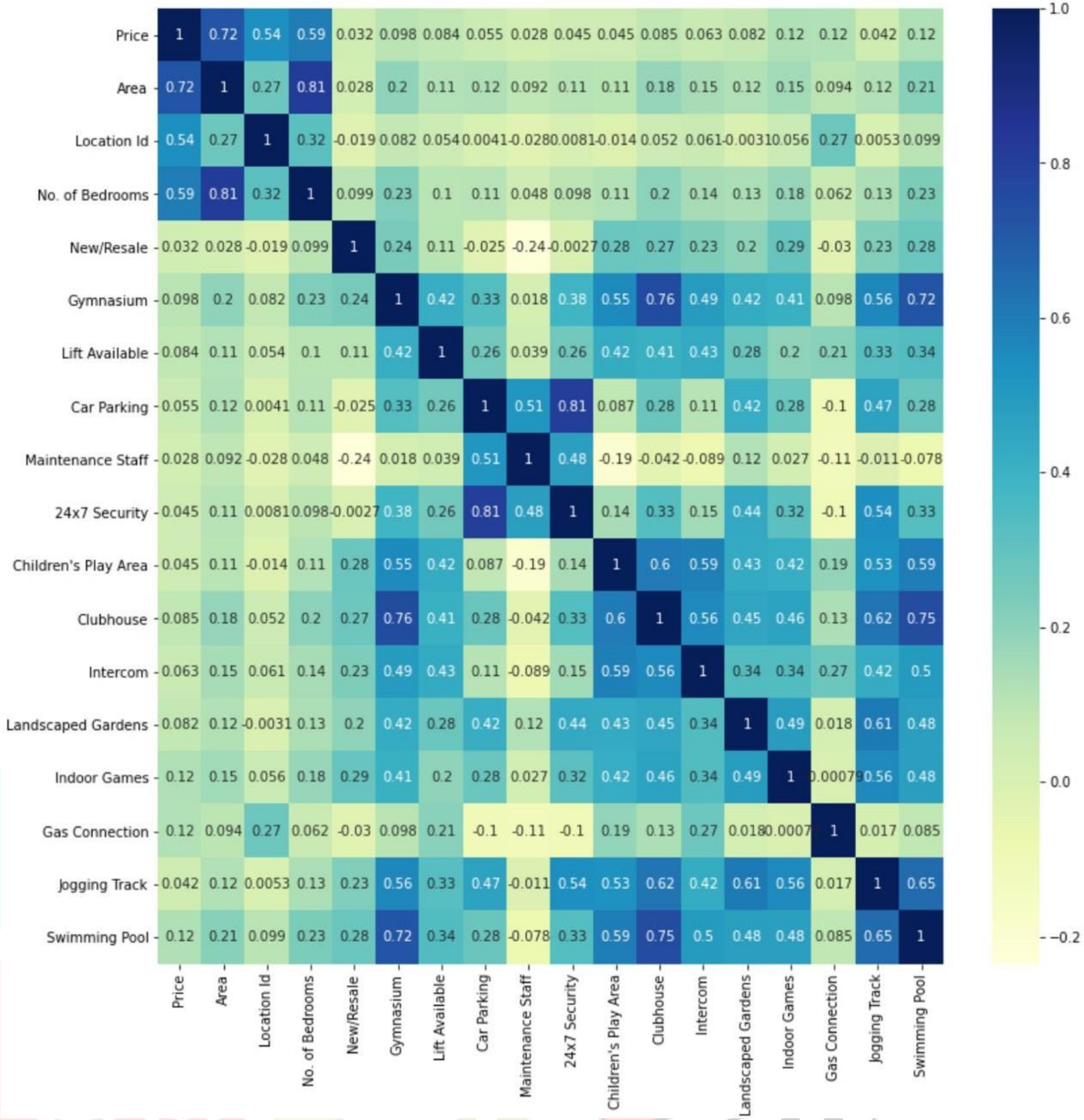

Fig.1. Correlation Matrix Heatmap

IV. METHODOLOGY

After analyzing and visualizing the data the next step is to train the model to help us in predicting the house prices. It involves dividing the dataset into training and testing sets. In the process of developing the model for training the data, we found from our experimental results that a combination of algorithms or models worked better than a single algorithm on the data i.e., gave lower error values. Various regression algorithms were tested including Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K- Nearest Neighbors (KNN), Bagging, and Boosting where Bagging, Boosting, and Random Forest were ensemble models in themselves. Ensemble Learning includes a combination of algorithms or models to provide results.

The ways of combining models depend exclusively on the features of the dataset. Simple averaging of results of ensemble members (models used in the combination) implies equal contribution to the final prediction. Weighted averaging is an extension to simple averaging which deals with the limitation of simple averaging when some models areknown to perform better or much worse than the others. A weighted ensemble is a model where the weight contribution of each model participating (ensemble members) is computed according to the efficiency of predictions. Below given Fig.2 provides an insight into the flow of the system.

We did a 1:3 split on the dataset to make the test and train sets. After training and building the ensemble model with the training set of data, it was tested with the test set of data to make the final predictions. The model was then implemented with a web UI to input parameter values from the user and predict house prices based on those values. After working on various models as a combination for ensemble learning, we found the best fit of three algorithms namely Linear Regression, K-Nearest Neighbor (KNN), and Decision Tree. A weighted average of predictions fromthese three algorithms gave us the least error range compared to other combinations.

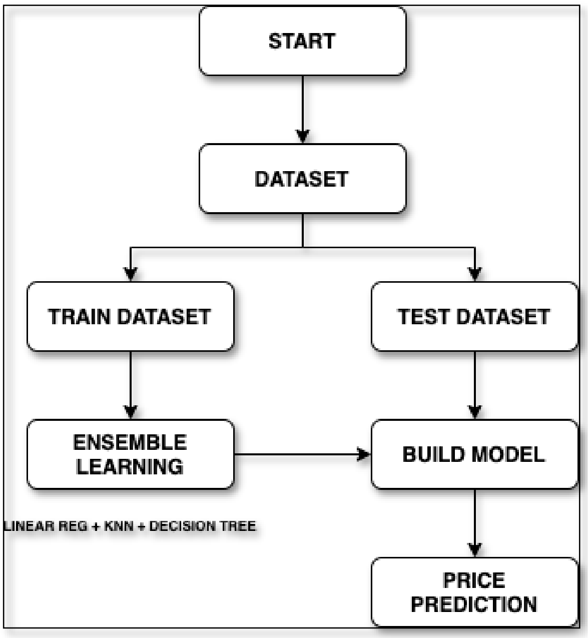

Fig.2. Methodology Flow Chart

4.1 Linear Regression

Simple linear regression algorithm analyses the relation between two entities where one is dependent and the other is independent. Change in the independent entity reflects the change in the dependent entity. This algorithm does not calculate the dependency instead only the association between the two entities or variables. The equation of the line of linear regression is as follows:

$$y = A + Bx \qquad (4.1)$$

Where X is the independent variable and Y is the dependent variable. A refers to the intercept whereas B refers to the slope of the line. Here, we have trained the linear regression model using the training dataset and then tested it on the testing dataset to make predictions. Below given Fig.3 is a scatter plot between the original house price value and the predicted house price value by the model.

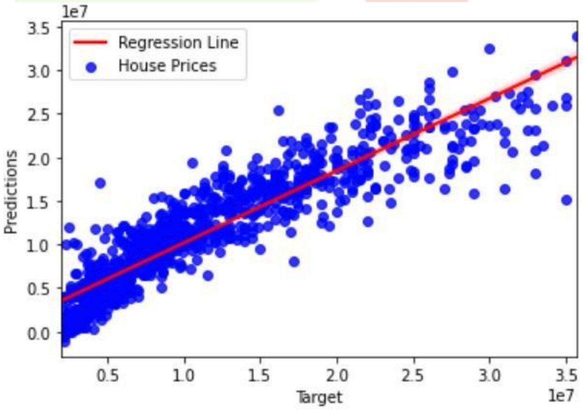

Fig.3. Linear Regression Scatter Plot

Below given Table.1 mention the error metrics of the model: Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE).

Table 1. Linear Regression Error Metrics

MAE	MSE	RMSE	R2
2229371.20	9247442528422.93	3040960.80	0.83

4.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a machine learning algorithm that does regressive as well as classification predictive analysis. It is also called a lazy learner algorithm since it does not analyze the data it is trained with instead the algorithm only classifies the new data it is tested with by its similarity. Here, KNN uses feature similarity to predict the house price values, i.e., it assigns a value to the new data based on how closely it relates (using distance functions for continuous variables such as Euclidean (2) and Manhattan (3) distance functions) to the points in the training set.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (4.2)$$

$$\sum_{i=1}^{k}|x_i - y_i| \qquad (4.3)$$

Where X refers to the new point, Y refers to the existing point and K is the K-Factor (no. of neighbors the algorithm looks at before assigning a value). Below given Fig.4 is a scatter plot between the original house price value and the predicted house price value by the model.

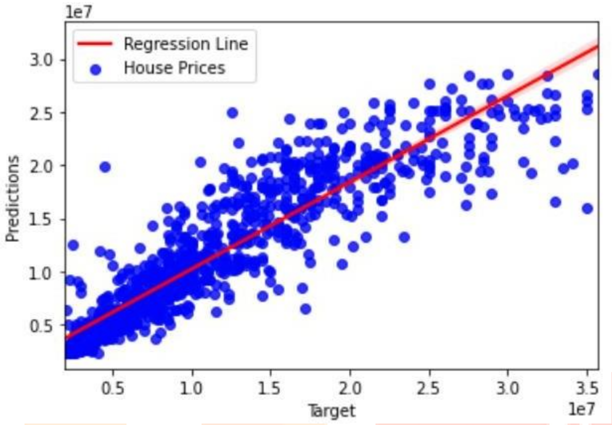

Fig.4. KNN Scatter Plot

Below given Table.2 mentions the error metrics of the model: Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE).

Table 2. KNN Error Metrics

MAE	MSE	RMSE	R2
1948884.96	8927720068103	2987929.06	0.84

4.3 Decision Tree

A decision tree algorithm builds the model in a tree structure with decision nodes and leaf nodes. It breaks down the dataset into smaller sets with similar values and the highest node is known as the root node. The tree is made of only conditional control statements with each decision node testing an attribute. Below given Fig.5 is a scatter plot between the original house price value and the predicted house price value by the model.

Below given Table.3 mentions the error metrics of the model: Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score, and Root Mean Squared Error (RMSE).

Table 3. Decision Tree Error Metrics

MAE	MSE	RMSE	R2
1851282.60	10135014746386	3183553.80	0.81

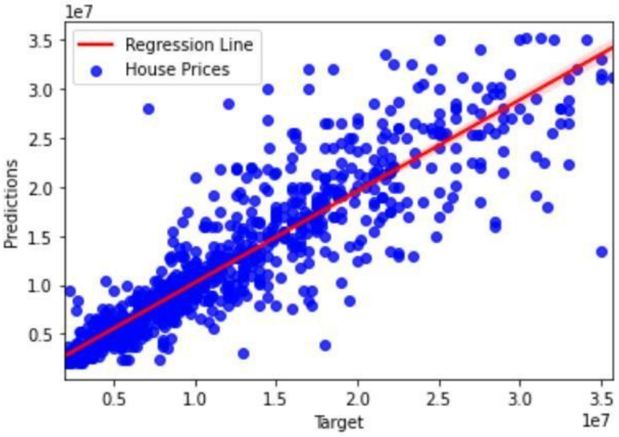

Fig 5. KNN Scatter Plot

4.4 Ensemble Learning Model

As discussed earlier, the ensemble learning model is a combination of algorithms or models which help to improve predictions depending on the features of the dataset. We could conclude from our experimental results that a weighted average of predictions from Linear Regression, KNN, and Decision Tree models provided the lowest error values compared to predictions from individual algorithms from Fig.6 and Table.4 below. The weights assigned to the predictions of models are based on its performance and features of our dataset exclusively.

Below given Table.4 mentions the error metrics of the model: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 Score, and Weight Assigned (W).

Table 4. Error Metrics Comparison

MODEL	MAE	MSE	RMSE	R2	W
Linear Reg.	2229371.20	9.2E+10	3040960.80	0.83	0.1
KNN	1948884.96	8.9E+!0	2987929.06	0.84	0.3
Decision Tree	1851282.60	1.0E+09	3183553.80	0.81	0.6
Ensemble	1658701.38	7.3E+10	2694486.53	0.87	-

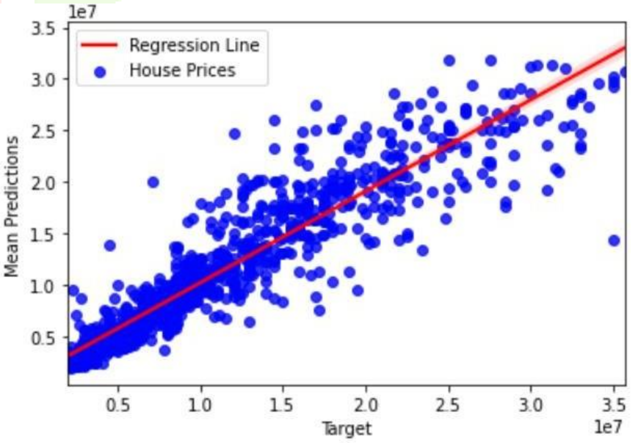

Fig 6. Ensemble Model Scatter Plot

V. EXPERIMENTAL RESULTS

The weighted average ensemble model performs substantially better than the individual models. The trained model attains an accuracy of 84%. The comparison of Mean Absolute Percentage Errors (MAPE) of the models used and the ensemble model is given below in the form of a bar chart Fig. 7. On comparing the various models, we find that the ensemble model works best with the lowest Mean Absolute Percentage Error (MAPE).

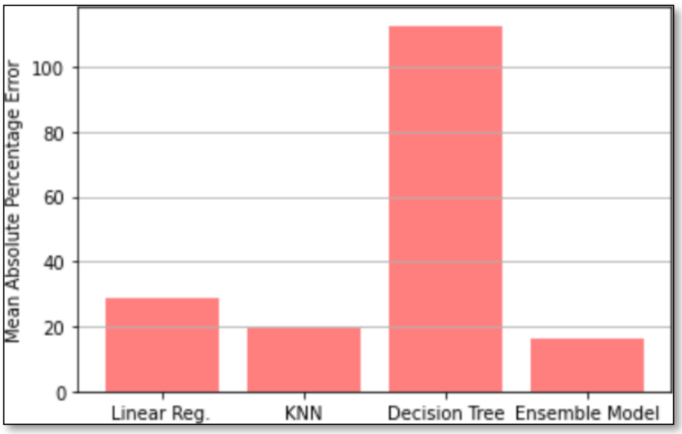

Fig 7. MAPE Comparison

VI. CONCLUSION

The price of housing units depends on a large number of factors. Therefore, price estimation strategy must consider multiple intrinsic and indicative parameters. The efficacy checking of the algorithms is generally done by testing it on certain data sets. Sample space of the data sets also affects the prediction performance. Work on many prediction methodologies is available in the literature. Most of them adapt the idea of comparing individual algorithms and selecting the best performing algorithms as the model for prediction. This paper presents a novel method of housing price prediction based on ensemble learning instead of using individual algorithms. The performance of the method was tested on 6347 records from the dataset incorporated from Kaggle Inc. It showed improved performance of 84% accuracy and MAPE of 16.09%. The innovation presented in this uses the weighted average ensemble model of Linear Regression, KNN, and Decision Tree. The performance can further be improved by optimizing the parameters. The key takeaway from this paper would be the significant improvement in the house price predictions using the ensemble learning model.

VII. ACKNOWLEDGMENT

A sincere thanks to Vidyalankar Institute of Technology for providing a platform and necessary resources to develop this project. Our deepest gratitude to each and every one who contributed substantially towards the completion of the project work.

VIII. REFERENCES

[1] Neelam Shinde, Kiran Gawande, “Valuation of House Prices using Predictive Techniques”, International Journal of Advances in Electronics and Computer Science – 2018.

[2] Joao Mendes Moreira, Alipio Mario Jorge, Carlos Soares, Jorge Freire de Sousa, “Ensemble Approaches for Regression: A Survey”, ACM Computing Surveys – 2012.

[3] Yashraj Garud, Hemanshu Vispute, Nayan Bisai, and Prof. Madhu Nashipudimath, “Housing Price Prediction using Machine Learning”, International Research Journal of Engineering and Technology (IRJET) – 2020.

[4] Prof. Pradnya Patil, Darshil Shah, Harshad Rajput, Jay Chheda, “House Price Prediction Using Machine Learning and RPA”, International Research Journal of Engineering and Technology (IRJET) – 2020.K. Elissa, “Title of paper if known,” unpublished.

[5] Alisha Kuvalekar, Sidhika Mahadik, Shivani Manchewar, Shila Jawale, “House Price Forecasting using Machine Learning”, 3rd International Conference on Advances in Science & Technology (ICAST) – 2020.

[6] Zhongyuan Han, Jiaming Gao, Huilin Sun, Ruifeng Liu, Chengzhe Huang, Leilei Kong, Haoliang Qi, “An Ensemble Learning-based model for Classification of Insincere Question”, FIRE – 2019.

[7] Ayush Varma, Sagar Doshi, Abhijit Sarma, Rohini Nair, “House Price Prediction Using Machine Learning and Neural Networks “, Second International Conference on Inventive Communication and Computational Technologies (ICICCT) – 2018.

[8] P. Durganjali; M. Vani Pujitha, “House Resale Price Prediction Using Classification Algorithms”, International Conference on Smart Structures and Systems (ICSSS) – 2019.

[9] CH. Raga Madhuri; G. Anuradha; M. Vani Pujitha, “House Price Prediction Using Regression Techniques: A Comparative Study”, International Conference on Smart Structures and Systems (ICSSS) – 2019.

[10] A. Adair, J. Berry, W. McGreal, “Hedonic modeling, housing submarkets and residential valuation”, Journal of Property Research – 1996.

[11] O. Bin,” A prediction comparison of housing sales prices by parametric versus semi-parametric regressions”, Journal of Housing Economics – 2004.

[12] T. Kauko, P. Hooimeijer, J. Hakfoort, “Capturing housing market segmentation: An alternative approach based on neural network modeling”, Housing Studies – 2002.

[13] Li Li, Kai-Hsuan Chu, “Prediction of Real Estate Price Variation Based on Economic Parameters “, IEEE International Conference on Applied System Innovation – 2017.

[14] G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch. Srinivasulu, “House Price Prediction Using Machine Learning “, International Journal of Innovative Technology and Exploring Engineering (IJITEE) – 2019.

[15] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization”, (IJACSA) International Journal of Advanced Computer Science and Applications.

[16] Ayşe SOY TEMÜR, Melek AKGÜN2, Günay TEMÜR, “Predicting housing sales in Turkey using ARIMA, LSTM, and Hybrid models”, Journal of Business Economics and Management ISSN – 2019.