Housing Price Prediction Via Different Machine Learning Techniques

Aayush Agrawal
Computer Science
Dept., University Of
Central Missouri
Lees Summit, U.S.A.
aayushsanghai1994@gmail.com

Keyurkumar
Dharmeshkumar Patel
Computer Science
Dept., University Of
Central Missouri
Lees Summit, U.S.A.
keyurpatel3310@gmail.com

John Conley
Computer Science
Dept., University Of
Central Missouri
Lees Summit, U.S.A.
JCD57610@ucmo.edu

Karthik Reddy
Kothakapu
Computer Science
Dept., University Of
Central Missouri
Lees Summit, U.S.A.
kxk86810@ucmo.edu

***Abstract*- The Standard & Poor's Case-Shiller and Office of Federal Housing Enterprise Oversight (OFHEO) housing price indices are used to determine house sales [2]. Housing price fluctuations are frequently estimated using the House Price Index (HPI)[8] . HPI alone cannot forecast a person's housing price because housing prices are highly associated with other criteria including location, area, and population [1]. The researchers and many other interested parties have become interested in the phenomenon of growing or declining housing values. There have been a lot of publications that use typical machine learning techniques to successfully estimate house prices, but they rarely analyze the performance of different models and ignore the less well-known but more sophisticated models. The motivation of this project comes from the fact to improve the accuracy of house value predictions through a comparison between different models. As a result, this study will utilize advanced machine learning methodologies to investigate the differences between a number of advanced models in order to explore diverse influences of features on prediction methods [15]. We analyze different models in this paper and compare the accuracy of the models based on various accuracy predicting methods. The weighted mean of several different methodologies was used to calculate our results, which are the most accurate. We have even tried to reduce the number if attributes to see if that makes any difference in predicting the housing prices and also calculated our error percentage to check the accuracy. In light of this assertion, we then provide an enhanced housing price prediction model to help a home seller or real estate agent make more educated choices. We also suggest to use latest real-time neighborhood data and details provided by the state or central government's website or pull some latest datasets available in websites like Kaggle, scikit etc. to get better and updated results.**

Keywords- House Price Prediction, Machine Learning, Linear Regression, Random Forest, Decision Tree, Mean Absolute Error.

I. INTRODUCTION

The analysis of real estate value is considered essential for guiding decisions on urban planning. A risky nonlinear process is the land framework. To obtain the highest results, financial experts base their decisions on the patterns that are now in play. Designers are curious in the future trends in their fundamental leadership. We can all agree that a house's price is a selection of numbers, hence predicting a house's price is a regression job [5]. To anticipate the price of a property, one person will often look for comparable houses in their neighborhood and attempt to predict the price using the information gathered. All facts imply that house price prediction is an emerging study topic of regression which requires the understanding of machine learning. Machine learning can be useful in predicting the price of a property given the recent expansion of the real estate market. Few academics have, however, used machine learning algorithms to determine the selling price for real estate holdings. Real estate agents, buyers, and sellers are all significant participants in the real estate market. Homeowners might hire a real estate agent to represent them if they want to sell their townhouse. The agent enters the seller's townhouse's details into a Multiple Listing Service (MLS) [13]. The information will subsequently be available to other real estate brokers as an active listing. Townhouses are being sold by their owners at the asking amount. Buyers, on the other hand, try to make a deal by offering less than the asking price. As a result, the listing price that the seller initially anticipated and the closing price that the buyers pay may differ in price. From the seller's perspective, the sale is profitable if the closing price is more than or equal to the listed price. The seller can incur a loss if the final price is less than the listed price [2]. This has inspired us to work in this field. Machine learning entails the provision of valid datasets, and predictions are then based on them. The machine itself learns the potential significance of a given event for the system as a whole based on the data it has already

loaded, and it predicts the outcome appropriately. The list of contemporary uses for this approach is vast and includes forecasting stock prices, the likelihood of an earthquake, corporate sales, and many more. For our project, we are doing an **Empirical Evaluation** and we will consider California Housing Data as our primary location and will try to predict the housing prices for various localities based on the latitude and longitudes. We are planning to use various parameters like Average no. of rooms, Population, Average Occupancy, Median House Value etc. as given in the data set which we will be procuring through a website named scikit learn. We will try to use various algorithms listed below and the weightage of each algorithm will be determined by the accuracy percentage. After reviewing over a number of test runs, we come to the conclusion that a succession of algorithms, as opposed to a single method, produces superior outcomes.

II. MOTIVATION

Real estate value analysis is considered essential to urban planning decisions. A dangerous non- linear process is the floor frame. For best results, financial professionals make decisions based on current models. Designers are interested in future trends in core management [14].

To predict the price of a property, people often look for similar homes nearby and try to predict the price based on the information they gather. All facts indicate that house price prediction is an emerging topic of regression research that requires an understanding of machine learning. It inspired me to work in this field. Machine learning requires us to provide valid datasets and make predictions based on them. Based on the data already loaded, the machine itself learns the importance of certain events to the system as a whole and predicts outcomes accordingly. Predicting house prices is crucial for maximizing real estate productivity. As before, house prices were calculated by adding together the prices for buying and selling in a certain area. In order to close the information gap and increase real estate efficiency, the house price forecast model is very crucial. With the help of this model, we could forecast prices.

It also helps both the consumer and the developer by assisting in the selection of the appropriate time to sell a home. The physical state, design, and location of a house are the three variables that affect its cost.

III. MAIN CONTRIBUTION AND OBJECTIVES

We are doing an empirical valuation, considering California as the primary location, and trying to predict property prices for different locations based on latitude and longitude. As described in a data set obtained from a website called scikit learning, using various parameters such as average number of rooms, population, average occupancy, and average house price. We trying to use different algorithms, each weighted by a percentage of accuracy.

We are using three models Linear Regression, Decision Tree, Random Forest. To evaluate the accuracy of each model we use Mean Absolute Error and Mean square Error (MSE).

- Linear regression: linear regression is a supervised machine learning model where the model finds a linear line of best fit between the independent and dependent variable. It finds the relationship between the dependent and independent variable. This model makes approximation process simple.
- Decision Tree: A decision tree is a supervised learning algorithm used for both classification and regression tasks. Our problem is also a type of regression and supervised task. This model required less data preparations for implementation.
- Random forest: random forest is also used for classification and regression problems. It builds decision tree for different samples and take the average for regression problem. This model solves the problem of overfitting as output is based on majority.
- Mean absolute error: absolute error means difference between predicted and original value. MAE refers to the average value of all the absolute error.
- Mean square error: MSE is the average of the square of the difference between prediction and original value.

We are going to implement these three models and by using two error measure we can decide which method is more accurate and useful. Then assert our predictions based on the calculations and accuracy of the different models implemented.

IV. RELATED WORK

After going through various research papers on the related topic, we see two major challenges that researchers face while predicting the housing prices viz, the number of features that we need to take into account that will help us accurately predict the house prices and secondly, to figure out that which machine learning technique will be most effective in meticulously predicting the housing prices [4].

Roy E. Lowrance developed a linear model of residential real estate prices for 2003 through 2009 in Los Angeles County [6]. He tested designs for linear models to determine the best form for the model as well as the training period, features, and regularize that produce the lowest errors. And he compared the best of linear models to random forests, the result showed that the random forests model, with minimal design work, outperformed the carefully designed local linear model [6]. Hu et al. built multivariate regression models of home prices using a dataset composed of 81 homes [6]. They applied the maximum information coefficient MIC) statistics to observed home values and predicted ones as the evaluation method, and found high strength of the relationship between observer and predictor [6]. Vapnik introduced three kinds of a support vector machine (SVM) such as a hard margin SVM (H-SVM), a soft margin SVM (S-SVM) and a kernel SVM. Statistical users accept SVM examined by the real data [6]. Due to high nonlinearity of the house value data, Mu et al. used support vector machine (SVM), least squares support vector machine (LSSVM) and partial least squares (PLS) to forecast the values of Boston suburb houses [6]. They found SVM and LSSVM are superior to PLS on dealing with the problem of nonlinearity, and the best performance can be achieved by SVM because of solving quadratic programming problem [6]. All these previous work clearly indicates that non-linear models would be much better than linear regression for house value prediction [6].

Predicting real estate costs is a challenging issue, as seen by the multiple research projects that authors have undertaken utilizing a variety of machine learning algorithms. The study displays a range of findings from each publication, but one important element is missing—it does not project the costs of the client-specified homes into the future. As a result, the risk of interest in a condo or a zone increases significantly. Clients frequently hire brokers to help them avoid making this error, which raises the procedure's cost. This forces the current framework to be modified and enhanced.

The present work is unique from all of the works mentioned above as we are trying to look at the problem from various perspectives instead of just one. We are trying to predict the price of the housing market by implementing different set of models and calculating the accuracy of each model by predicting the difference between the predicted and original values.

V. PROPOSED FRAMEWORK

The entire work process can be broken into the three sections below. Data preprocessing, data analysis, the use of machine learning techniques, and performance monitoring processes are some aspects.

A. Data Preprocessing

We saw that the data contains 21000 instances and almost 8 attributes. We converted the data into pandas for smooth import and checked out each of the fields. Then we tried to get an overview of how our data looked like with the help of various pandas functions.

We also checked if our data had any null or disparaging values which could affect our models or predictions.

B. Data Analysis

Prior to creating a model, exploratory data analysis is crucial. We can identify implicit patterns in the data in this way, which will helps us select the best machine learning techniques. We can use various inbuilt libraries to perform data analysis and check the distribution of data. Some of the analysis libraries we used was :

- Heatmap: It help to show the correlation between the attributes in our data. The correlation is between -1 and +1. There is no linear trend between the two variables if the values are closer to zero. The closer the correlation is to 1, the more positively associated they are; that is, as one rises, the other does as well, and the stronger this relationship is, the closer to 1 the correlation is. Through this analysis, we could see a significant relation between median income and the median house value. The heatmap gave a strong and compelling insight on what are the attributes we could use to our

advantage in selecting the machine learning models.

```
In [21]: #The Heatmap give us an insight on the correlation between different attributes in the dataset
         plt.figure(figsize=(12, 6))
         sns.heatmap(data.frame.corr(),
                     cmap = 'YlGnBu',
                     linewidths = 0.5,
                     annot = True)
Out[21]: <AxesSubplot: >
```

Figure 1: Heatmap showing correlation

- Scatterplot: a scatter plot matrix is used to show bivariate correlations between sets of variables. Numerous associations can be investigated in a single chart thanks to the scatter plots in the matrix, which each show the link between a pair of variables. Scatterplot was one of the most important exploratory analysis for our project as we could plot the median house value based on the latitude and longitude. This formed a map of California giving us a clear picture of the accuracy of the data and how the housing prices differ with respect to terrain, county etc.

```
In [6]: #the x- and y-axis is the latitude and longitude and the circle size and
        #color is linked with the house value in the district.
        sns.scatterplot(data=data.frame, x="Longitude", y="Latitude",size="MedHouseVal",hue="MedHouseVal")
Out[6]: <AxesSubplot: xlabel='Longitude', ylabel='Latitude'>
```

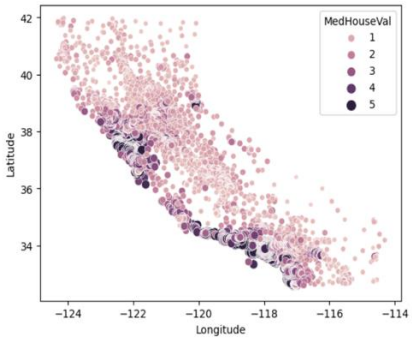

Figure 2: Scatterplot

- Pair Plot: Pair plots are used to determine the most distinct clusters or the best combination of features to describe a connection between two variables. By creating some straightforward linear separations or basic lines in our data set, it also helped to create some straightforward classification models.

In the pair plot we dropped the attribute latitude and longitude because they separately did not affect the target attribute. From the above pair plot, we could also see that as the average rooms and average bedrooms increases, the price of the house also increases. Also, average occupancy is the least corelated with the other attributes.

C. Model Selection

For implementation, we used three machine learning models viz,

- Linear Regression: The simplest prediction technique is linear regression. The predictor variable and the variable that comes first in importance, whether the predictor variable and su, are the two items it employs as variables. The link between one dependent variable and one or more independent variables is explained using these estimations To train the model for Linear Regression, first we imported the required libraries from sklearn and then then created the model using Linear Regression() [12]. We used X_train and Y_train to train the model. To test the model, we used X_valid and assign the predicted output to LR_Pred.

- Random Forest: A large number of decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees choose is the output of the random forest for classification problems.
Representation of it is as follows:

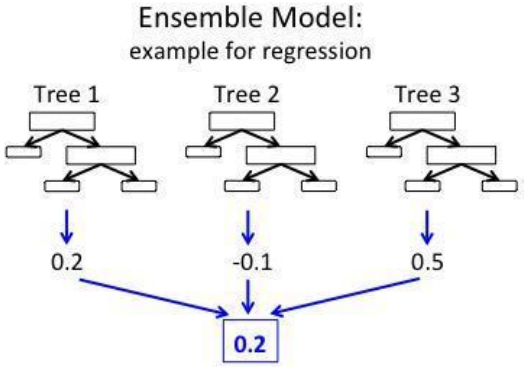

Figure 3: Random Forest regression model [7]

It builds decision tree for different samples and take the average for regression problem. This model solves the problem of overfitting as output is based on majority.

We have use Random Forest algorithm in our project and helped us train our model and predict the Housing prices. This one gave one of the most promising results as compared to other models with an accuracy of almost 80%.

- Decision Tree: The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes. In decision tree, we select attributes based upon which attribute has highest information gain [9].

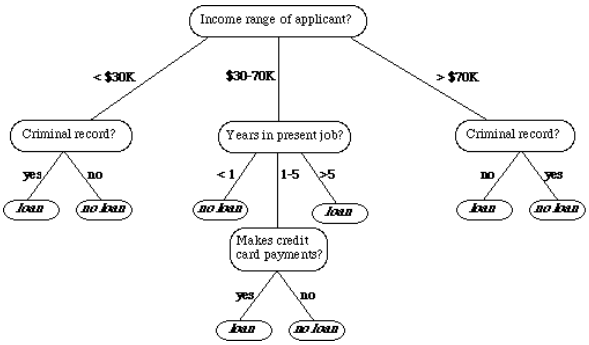

Figure 4: Decision Tree

When do we end growing our tree must be a question you are asking yourself. Real-world datasets typically contain a lot of features, which leads to a lot of splits, which produces a massive tree. These trees take long to construct and may result in overfitting. In other words, the tree will provide extremely accurate results on the training dataset but inaccurate results on the test dataset.

Since the Random Forest performed really well with our data, it was a plausible step for us to evaluate our model using Decision Tree [12] .

VI. DATA DESCRIPTION

The dataset utilized in the current study comes from a competition that was organized by the website statlib and uses data from that website. This work utilizes feature selection techniques such as loading the dataset from the sklearn library, searching for null values or outliers and we have also used Variable Influence Factor (VIF).
A measure of a variable's association with other variables is called the Variance Inflation Factor (VIF). We try to maintain a set of variables such that the Variance Inflation Factors (VIFs) of all the variables are less than those that the classifier predicts 1 when the target value is 1, true negatives are those values in which the classifier predicts 0 when the target value is 0, and false positives are those values in which the classifier predicts 1 when the target value is 1. As a general rule, if the correlation between the variables is high, the Variance Inflation Factor is 1.

For our project we are using dataset “California housing”. Dataset is obtained from statlib repository. The task is to predict the housing price based on many features. This dataset has 20640 instances and 8 features and a target attribute. The target column is MedHouseVal which is median value of house for each district. All the features have float datatype and there are no null values in the dataset.

There is total 8 features:
MedInc - it is the median income in the group
HouseAge – median house age of the district
AveRooms- Average number of rooms
aveBedrms – Average number of bedrooms
Population – populations of the district
AveOccup – Average number of members in the household
Latitude and Longitude – latitude and longitude of the district

This is the target attribute of the dataset. Which is median house value of the district. Here we have to predict the numeric value, so this is a regression problem. This dataset has labels, so this is a classification problem.

VII. RESULTS/EXPEREMENRATION & COMPARISON ANALYSIS

Different models have been implemented in this study. The models are Linear Regression, Decision Tree, Random Forest. All the models are supervised models. To implement these models’ prior knowledge is required. To measure the accuracy and performance of these model we use two different measures Mean absolute error and mean square error. Also, we calculated the accuracy score of each model for the given dataset.

- Mean absolute error: One of the measures for describing and evaluating a machine learning model's performance is MAE (mean absolute error). We find this error by subtracting predicted value from actual value.

  Error = Actual value – Predicted value. (1)

  This is done for every tuple then all Error are converted into positive value by applying absolute function.

  Absolute error = |Error| (2)

  After that mean is calculated for all absolute.

  MAE = (1/n) $\sum_1^n |O - E|$(3)

  Where O is known value, E is predicted value and n is the total number of records. The mean absolute error for any dataset is means the mean of the absolute values of each predicted error. Predicted error means difference between original and predicted value. Mean Absolute Error (MAE) is a statistical term that describes the outcomes of measuring the difference between two continuous variables [10].
- Mean square error: Mean square error computed by doing square of difference between actual and predicted value then taking mean of each values computed.

  MSE = (1/n) * Σ(actual – Predicted)2 (4)

  This error evaluate how close are the original and predicted values. Large error suggests that predicted values are very far from the actual value. The mean squared error (MSE); a risk function used in analytics which calculates the square of errors. To ensure that the mean squared error is always greater than or equal to zero, the differences are squared, which eliminates any negative values for the differences. Squaring will increase the influence of larger errors and it affect large errors more than small ones. MSE is used mostly when the data is normally distributed, and you want large errors to be charged more than small ones while performing regression.
- Accuracy score: every model that is implemented with the help of scikit learn has a function name score. Which takes train dataset as input to compute the accuracy of the model. This method does not need the predicted values to measure the accuracy of the model. This method makes prediction using test dataset on its own and use this result to compute accuracy of the model. This method is used to calculate accuracy instead of other methods which involves many steps. There is only one condition before running this method that is it needs test dataset ready.

In this study some of the prediction made by the three models are very close to actual values while some values are very far from the actual value. For each model individual result is described below.

- Linear regression: After training the model with X_train and Y_train. We predicted the result using X_valid. The predicted result is the median house price which is LR_Pred. The predicted results are [2.4435 0.9813 2.1672 ... 0.9648 1.8676 2.0355]
- Random Forest: The result of the Random Forest model is stored in the RF_Pred after training the model with same dataset used for linear regression. And the predicted output is [2.4435 0.9813 2.1672 ... 0.9648 1.8676 2.0355].
- Decision Tree: DT_Pred is the predicted result for decision tree model. The values are [2.911 1.231 1.397 ... 1.036 2.063 2.362]. Same dataset is used to train all three models.

To compare each model's performance and accuracy for the given dataset MAE, MSE and accuracy score is used. The comparison between these values is shown in the given table.

Model	MAE	MSE	Accuracy
Linear Regression	0.3131	0.5332	60.47
Random Forest	0.1911	0.2820	79.09
Decision Tree	0.2455	0.5238	61.17

Table 1: Error and accuracy score

From the above table we can see the comparison between all the three models. Linear regression has the maximum of mean square error and mean absolute error and have the least accuracy for the given dataset which is 60.47%. while random forest has the minimum mean square error and mean absolute error and have the maximum accuracy

which is 79.09%. So, for this study random forest is the best choice for the prediction of the house price and linear regression is not a good choice for the given dataset. Decision tree model perform better than Linear regression but not better than Random Forest.

```
In [10]: #printing the original and predicted data by all three models.
predictdataset = pd.DataFrame()
predictdataset['original'] = Y_valid
predictdataset['LR_Predicted'] = LR_Pred
predictdataset['RF_Predicted'] = RF_Pred
predictdataset['DT_Predicted'] = DT_Pred
predictdataset
```

Out[10]:

	original	LR_Predicted	RF_Predicted	DT_Predicted
9288	4.359	2.970384	2.4435	2.911
1878	1.105	0.438590	0.9813	1.231
20439	2.641	2.438721	2.1672	1.397
10957	1.872	1.738766	1.7104	1.584
10316	3.049	2.591785	2.5844	4.222
...	...	...	...	...
13044	2.034	1.239351	1.5935	1.821
11419	2.904	2.830357	2.9547	3.440
1804	0.903	1.685755	0.9648	1.036
7136	1.894	1.881380	1.8676	2.063
16633	1.839	2.463608	2.0355	2.362

4128 rows × 4 columns

Figure 5: predicted result of each model

Above is the image which contains the predicted result of each model and actual value of the dataset. first column named original contains the actual value of the dataset. LR_Predicted is the predicted result from linear regression. RF_Predicted is the predicted result of the random forest model and DT_Predicted contains the predicted result of the model decision tree. From this result we can see that there are some values that are very far from the actual value and some predicted values are very close to the actual value. Also, we can see that values predicted from the Random Forest model are close compared to value predicted from other two models. So, random forest model is best choice for this study.

VIII. REFERENCES

[1]-Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, Housing Price Prediction via Improved Machine Learning Techniques, Procedia Computer Science, Volume 174, 2020, https://doi.org/10.1016/j.procs.2020.06.111.

[2]-Byeonghwa Park, Jae Kwon Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, 2015, https://doi.org/10.1016/j.eswa.2014.11.040.

[3]-A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, https://ieeexplore.ieee.org/abstract/document/8473231

[4]-D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, https://ieeexplore.ieee.org/abstract/document/8392275

[5]-N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, https://ieeexplore.ieee.org/abstract/document/8697639

[6]- Huang, Y. (2019). Predicting Home Value in California, United States via Machine Learning Modeling. *Statistics, Optimization & Information Computing* https://doi.org/10.19139/soic.v7i1.435

[7] The elements of statistical learning, Trevor Hastie - Random Forest Generation

[8] Bork, M., Moller, V.S.: House price forecast ability: a factor analysis. Real Estate Economics. Heidelberg (2016).

[9] , Alphaydin, Introduction to Machine Learning

[10] Andreas C. Müller & Sarah Guido, Introduction To Machine Learning With Python

[11] Andreas C. Müller & Sarah Guido, Machine Learning

[12] Christopher M. Bishop, Pattern Recognition and Machine Learning

[13] T. Kauko, P. Hooimeijer, J. Hakfoort, Capturing housing market segmentation: An alternative approach based on neural network modeling, Housing Studies, 17 (2002) 875-894.

[14] Lowrance, E.R.: Predicting the market value of single-family residential real estate. 1st edn. PhD diss., New York University, (2015).

[15] HousePriceIndex.FederalHousingFinanceAgency.https://www.fhfa.gov/

[16] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 2017

[17] Ke G, Meng Q, Finley T, Wang T, Chen W ,Ma W, et al. Light GBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 2017

[18] Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM

optimized by PSO. Optik-International Journal for Light and Electron Optics, 125(3), 1439–1443.
[19] Mansurul Bhuiyan and Mohammad Al Hasan, “Waiting to be sold: Pre- diction of Time-Dependent house selling probability,” IEEE International Conference on Data Science and Advanced Analytics, 2016.
[20] Li Li and Kai-Hsuan Chu, “Prediction of Real Estate Price Variation Based on Economic Parameters,” Department of Financial Management, Business School, Nankai University, 2017.