Article

Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices

Marzieh Khosravi [1],*, Sadman Bin Arif [2], Ali Ghaseminejad [3], Hamed Tohidi [4], and Hanieh Shabanian [5]

[1] Independent researcher, PA, USA; msh.khosravi2@gmail.com
[2] Department of Electrical and Electronics Engineering, University of Asia Pacific, Dhaka 1205, Bangladesh; sadmanbinarif93@gmail.com
[3] Post-Doctoral Associate, Department of Civil, Construction and Environmental Engineering, The University of Alabama, Tuscaloosa, AL 35487, USA; aghaseminejad@ua.edu
[4] Assistant Professor, Department of Civil Engineering, University of Memphis, TN 38152, USA; htohidi@memphis.edu
[5] Assistant Professor, Department of Computer Science, Northern Kentucky University, KY 41099, USA; shabanianh1@nku.edu
* Correspondence: msh.khosravi2@gmail.com

Abstract: Real estate market analysis and place-based decision-making can both benefit from understanding house price development. Although considerable amounts of interest have been devoted to housing price modelling, the assessment of house price fluctuation still requires further comparing studies. Housing price prediction is challenging as contributing factors are quite dynamic and subject to a variety of regulating elements. Future insight into housing market trends not only increases customer investment trust potential but also makes it possible for financial support to proceed more realistically in the future. In this study, a comprehensive data-informed framework is developed to investigate and anticipate real estate house prices using historical data by combining explanatory features. We examined about 500 houses in the Boston area as a case study and discussed how the increase in housing prices could vary by each of the contributing components. Fourteen machine learning (ML) regressors imply to the dataset and lead to a comparative study of the accuracy of all the models. ML-based regressors forecast real estate home prices as a function of thirteen influencing factors. The most informative features were also selected by conducting the permutation feature importance technique on all the features The study provides a comprehensive tool for evaluating the robustness and efficiency of ML models for housing price predictions. The results highlighted random forest as the best model has an R^2 equals to 0.88 and voting regressor as the second highest rated model has R^2 equals to 0.87. Results of multivariate exploratory data analysis also implied that the average number of rooms and percentage of the lower status of the population have the most significant impact on the price range predictions.

Keywords: Real State, Regressors, Artificial Intelligence, Machine Learning, Data-informed, Boston

1. Introduction

The continued growth in residential houses all over the world as one of the main human beings required living amenities brought so many researchers' and merchandise' attention to this field. The housing market is a major market that strengthens the world economy, where housing significantly impact every household. The recent global experience has highlighted that price variations and the housing market's transient nature are risk factors that can jeopardize the performance of the entire financial sectors. Consumption expenditure is believed to be beneficial for households' welfare assessment since it is fundamentally rooted in the concept of money metric utility and compared to household income, overall household spending often exhibits less variability. Furthermore, the household average expenditure includes the value of the house as a significant component [1]. There may not be a linear relationship between all the house's features and its price where it is essential to understand the potentially non-linear relationship to prevent potentially incorrect implications for real estate development and urban planning [2]. Price fluctuations are crucial because they obviously affect people's choices of where to dwell, work, and invest in real estate. As a result, several expert researchers have given house price forecasting a lot of attention. Machine learning models and algorithms have recently proven to be effective and comprehensible solutions to a diverse variety of house prices forecasting problems, as well as stock prices, and other monetary variables [3,4].

Regression analysis is used to determine how significantly each feature contributes individually to house pricing. According to an analysis of the housing market and housing price assessment literature, two major research trends include the application of the hedonic-based regression approach called as hedonic pricing model (HPM) and artificial intelligence (AI) approaches for establishing house price forecasting model [5–7]. Several hedonic-based approaches were used to investigate the connection between house prices and their associated effective housing attributes [8–11]. Whereas the comparison results of previous studies indicated that the AI technique performed better at forecasting property values than the HPM approach. Machine learning (ML) as one of the most innovative methods in AI for discovering, evaluating, and analyzing extremely complex structures of data and connections enables substantial learning, and a systematic entry of more existing information to improve model predictions [12,13]. ML algorithms are utilized for house price forecasting across numerous nations and regions considering the wide range of possible impact of housing locations [14,15]. Among all the various forecasting algorithms in ML, linear regression (LR) model is one of the common ML algorithms which also includes Ridge regressions (RR) and Lasso regressions (LaR) [16,17]. Other commonly used ML approaches are decision tree (DT) [18–20], random forest (RF) [21–23], support vector machine (SVM) [24–28]. Couple of comparative studies investigated the performance of these algorithms. SVM founded to have a more satisfying results compared to LR, DT, RF [29].

gradient boosting machine (GBM) [30–33], k-nearest neighbors (KNN) [34–37], and artificial neural network (ANN) [38–42] were also investigated for further analysis on house pricing. GBM and extreme gradient boosting (XGB) outperformed compared to the SVM as the previously better model [43]. Ho et al. compared three ML methods of SVM, RF, and gradient boosting machine (GBM) on housing prices over a period of 18 years considering three error metrics of mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) [44–49]. Although successful ML models using out-of-sample data set over predicted housing pricing, including SVM, RF, RR, LaR, decision tree, bagging, boosting, and ensemble learning, have been discovered to be more efficient and realistic, a comprehensive comparative research still needed to identify the forecasting model with the best performance for only housing prices.[50–52].

AI algorithms can be a superior method for predicting home prices, according to various studies. Graczyk et al. conducted a study to examine various ensemble models for estimating the value of the real estate and properties and found that all ensemble classifiers utilizing additive regression significantly reduced error when compared to the initial models [53]. Liu et al. concluded that back propagation neural network algorithm can performed well when evaluating home prices [54,55]. SVR and RF are the second and third noteworthy models that were investigated by researchers, despite the fact that neural network field of knowledge appears to be more substantially used in the house price prediction model [56]. Other studies revealed the greater predictive potential of generalized housing price pattern by ML approaches than typically applying ordinary least squares (OLS) on a hedonic pricing [51]. Wang et al. [57] presented two AI methods; particle swarm optimization (PSO) and SVM as the great real estate price forecasting models compared to the grid and genetic algorithms. Park and Bae [58,59] develop a housing price forecasting model based on AI and ML algorithms such as naïve Bayesian, and AdaBoost while comparing them to the classification accuracy performance.

doi:10.20944/preprints202209.0341.v1

Steurer et al. performed a comparative study of seven metrics among all 48 possible various metrics to evaluate the automated valuation models (AVMs) based on ML algorithms for house pricing prediction [60]. Peng et al. provide a new method as the first life-long property valuation prediction model which is automated by using a long short-term memory (LSTM) network to model the temporal relationship for property transaction data across time after first using a graph convolutional network (GCN) to extract the geographical information [61–63]. Su et al. propose a system that uses a hybrid approach based on genetic algorithm optimized ML (GA-GBR) [64]. They make use of a comprehensive data interpretation, enhance data on property valuation, and automate particular property valuation processes that can eventually perform a more accurate property valuation estimation. Kang et al. introduce a data-fusion methodology to test the accuracy of multi-data source projections of probable housing price development by integrating multiple data sources such as house structural attributes, their actual photos, locational amenities, street view images, transportation accessibility, and socioeconomic attributes of neighborhoods [65]. Deriving the R^2 of 0.74 by GBM resulted in discovering the higher house appreciation potential of low house prices within small house areas. Which is the same case in Boston house's recent rapidly increased pricing for both single-family homes and condominiums.

This article proposed an integrated framework for comparing multiple ML regression models considering different error metrics to conclude the best-fitted model with the highest accuracy and facilitate the housing price prediction. Even with all the preceding comparative studies that have been conducted in the domain of property's value, a more thorough comparative investigation is still required to determine whether additional housing market prediction analysis should incorporate certain algorithms or not. From the following perspective, this study contributes to the current literature already dedicated to numerous participations in property valuation. Although the correlation between algorithms has been thoroughly examined in previous research, integrating fourteen ML regression models on a well-known dataset enhances each model's accuracy when compared to each regressor's housing price result. The paper is structured as follows. Section two introduces the data and feature information utilized in this study and introduces the background and current status of the Boston property valuation trend. A comprehensive exploratory data analysis (EDA) is conducted on the multivariate dataset to investigate the values distribution and bivariate correlation coefficients followed by methodology and introduction to all the ML regression models and error analysis methods. Section 3 details the results of regressors performance and comparisons on their accuracy as well as deriving the feature importance ranking. Section 4 concludes the article and discusses ML regressions algorithms comparisons.

2. Data and Methods

This project's dataset was taken from the University of California Irvine (UCI) machine learning repository and the online data science platform Kaggle that both have a collection of databases, domain theories, and data generators that are used by the machine learning community [66,67]. Each of the 506 entries in data set used for this study, contains aggregate information about 14 characteristics of residences in different Boston areas. Boston home prices increased 9.6% over the previous year in July 2022, with a median sale price of $800K for all home types. In comparison, the median price for single-family houses and condominiums is $885K and $730K, respectively. Home prices in Boston are rapidly increasing because of a serious supply and demand imbalance in the housing market. 604 properties were sold in July 2022 compared to 856 in July of the previous year, a decline of 19, 45, and 30 percent in the number of single-family homes, townhouses, and condominiums sold, respectively. The housing price growth might be the cause of less tendency and less capability of buyers in townhouses compared to single-family houses [68,69].

The summary of the original dataset that includes the initial features is represented in Table 1. The information is divided into features and the target variable to create a model that can estimate the worth of homes and keep them in the variables for features and prices, accordingly. The features mentioned above provided quantifiable data on each data point. The BLACK variable is a representative of the proportion of blacks by town (Bk), where the BLACK is the equivalent if $1000(Bk - 0.63)^2$. LSTAT is the proportion of the population that a lower status of them is equivalent to half of the proportion of adults without, some high school education and the proportion of male workers classified as laborers [70]. The variable we aim to forecast is the target variable, "MEDV" which represents the median value of owner-occupied homes in $1000s.

Table 1. Full Description of the features and target variable

Features	ID	Description
CRIM	x_1	Per capita crime rate by town
ZN	x_2	Proportion of residential land zoned for lots larger than 25,000 sq.ft.
INDUS	x_3	Proportion of non-retail business acres per town.
CHAS	x_4	Charles River dummy variable (equals to 1 if tract bounds river, otherwise it is 0)
NOX	x_5	Nitric oxides concentration parts per 10 million
RM	x_6	Average number of rooms per dwelling
AGE	x_7	Proportion of owner-occupied units built prior to 1940
DIS	x_8	Weighted distances to five Boston employment centers
RAD	x_9	Index of accessibility to radial highways
TAX	x_{10}	Full-value property-tax rate per $10,000
PTRATIO	x_{11}	Pupil-teacher ratio by town
BLACK	x_{12}	Related to the proportion of blacks by town
LSTAT	x_{13}	Lower status of the population (percent).
MEDV	y_1	Median value of owner-occupied homes in $1000s

2.1. Multivariate Exploratory Data Analysis (EDA)

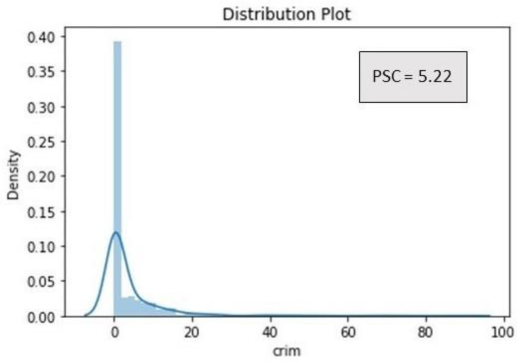
To understand the properties and characteristics of the multivariate dataset, a thorough EDA is carried out in **Error! Reference source not found.**. The crucial stage in performing EDA on data is to get the ML model to operate successfully. All of the real estate variables' internal temporal distribution is examined using a variety of visual techniques and numerical indexes. EDA is the process of conducting an initial inquiry into the real estate variables to identify any hidden patterns in the variables' distribution and further broken down into a variety of activities. They are known as the normality check, outliers/extreme values identification, and descriptive statistics. Descriptive statistics provide a brilliant way to depict the distribution of the variables' values by employing the variables' range, percentiles, inter-quartile range, number of data points, mean, and standard deviation. **Error! Reference source not found.** displays complete multivariate descriptive statistics. Histograms with density plots are used as a visual representation to show the normality of the variables, and Pearson's coefficient of skewness (PCS) is used as a numerical measure of skewness. By substituting the nearest neighbors of the data points for missing values, numerical imputation is used to make the dataset consistent. **Error! Reference source not found.** illustrates the basic descriptive statics of the input features as well as the target variable. It involves the number of data points considered to train/test the ML models, central tendency (mean), measure of spread (standard deviation), range of the values (minimum and maximum), and the percentiles to provide insights into the distribution of the variable with numerical indicators.

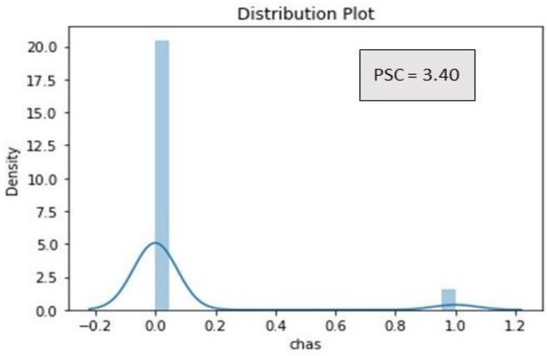
Table 2. Descriptive Statistics of the real estate variables

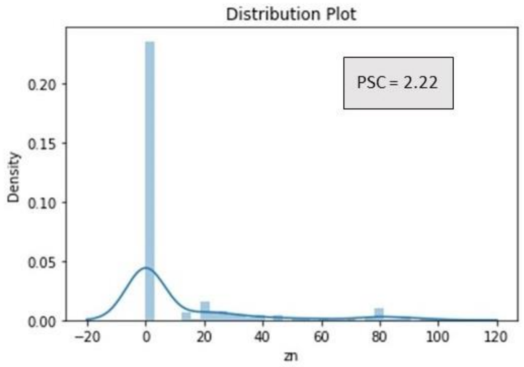
	Count	Mean	Std	Minimum	25%	50%	75%	Maximum

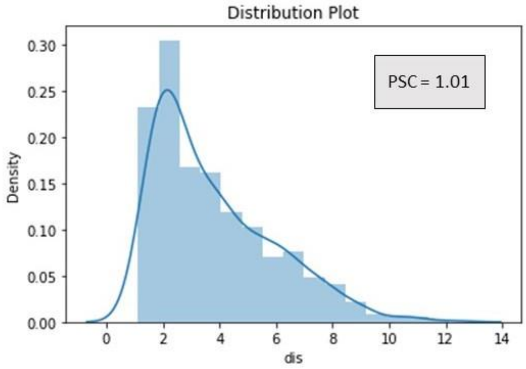
CRIM	506	3.61	8.60	0.00	0.08	0.25	3.67	88.97
ZN	506	11.36	23.32	0.00	0.00	0.00	12.50	100.00
INDUS	506	11.13	6.86	0.46	5.19	9.69	18.10	27.74
CHAS	506	0.06	0.25	0.00	0.00	0.00	0.00	1.00
NOX	506	0.55	0.11	0.38	0.44	0.53	0.62	0.87
RM	506	6.28	0.70	3.56	5.88	6.20	6.62	8.78
AGE	506	68.57	28.14	2.90	45.02	77.50	94.07	100.00
DIS	506	3.79	2.10	1.12	2.10	3.20	5.18	12.12
RAD	506	9.54	8.70	1.00	4.00	5.00	24.00	24.00
TAX	506	408.23	168.53	187.00	279.0	330.00	666.00	711.00
PTRATIO	506	18.45	2.16	12.60	17.40	19.05	20.20	22.00
BLACK	506	356.67	91.29	0.32	375.3	391.44	396.22	396.90
LSTAT	506	12.65	7.14	1.73	6.95	11.36	16.95	37.97
MEDV	506	22.53	9.19	5.00	17.02	21.20	25.00	50.00

Error! Reference source not found.Error! Reference source not found. graphic represents the distribution of the variables and demonstrates the significant level of overall non-normality for the top five features with the highest values and one feature with the negative significant level of overall non-normality. When compared all the variables exhibited more non-normality. The values of 5.22 (CRIM), 3.40 (CHAS), 2.22 (ZN), 1.10 (MEDV), 1.01 (DIS), 1.00 (RAD), 0.90 (LSTAT), 0.72 (NOX), 0.66 (TAX), 0.40 (RM), 0.29 (INDUS), -0.59 (AGE), -0.80 (PTRATIO), -2.89 (BLACK) are PCS, a numerical indicator of non-normalcy/skewness, are likewise greater than the PCS values of other real estate variables, indicating considerably less normality.









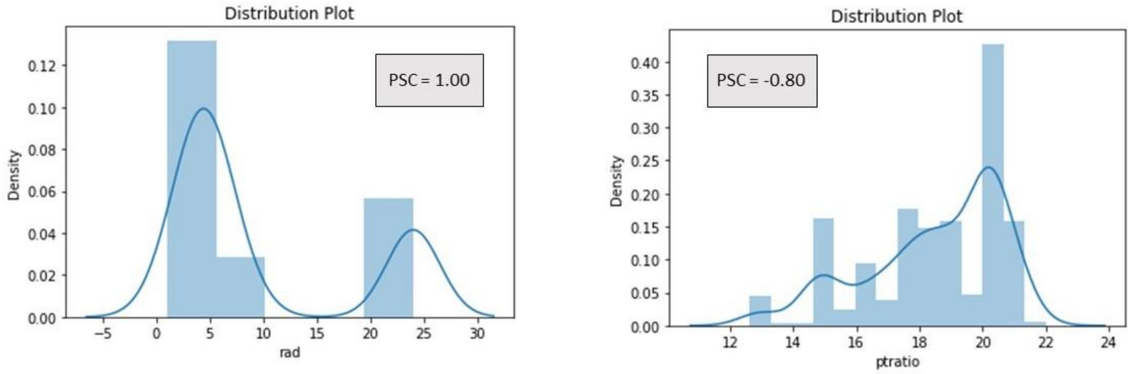
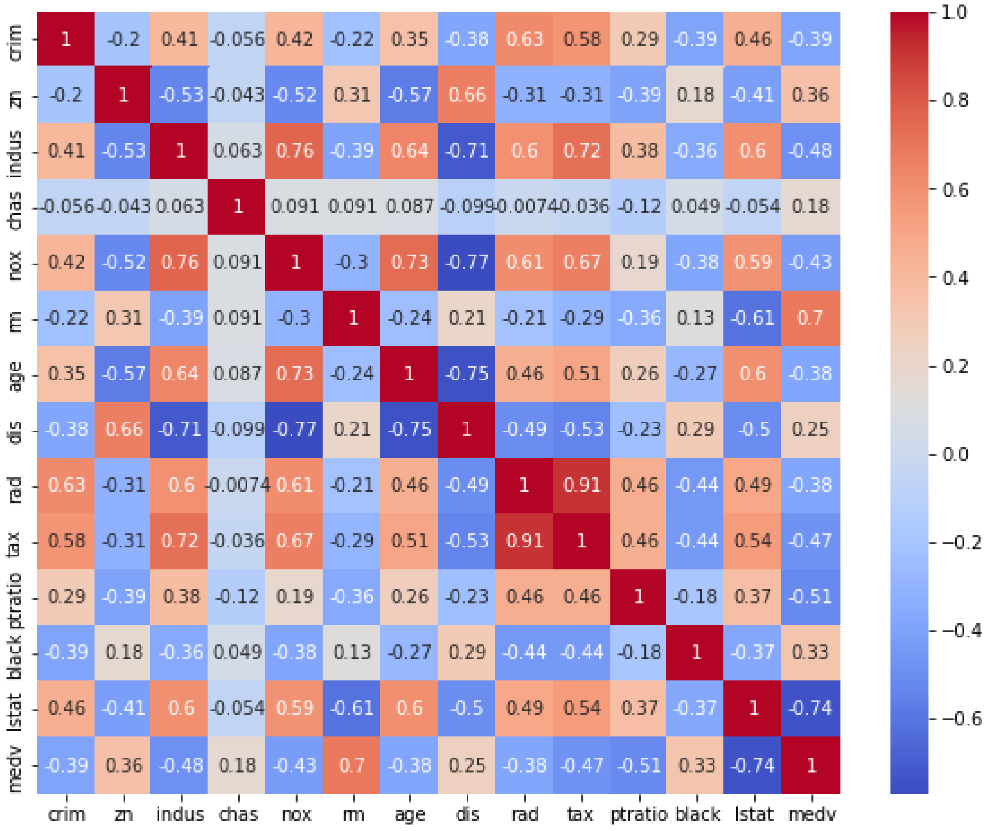


Figure 1. Distribution of the values of housing variables, 5.22 (CRIM), 3.40 (CHAS), 2.22 (ZN), 1.10 (MEDV), 1.01 (DIS), -0.80 (PTRATIO) using histogram

The linear relationship between two of the housing variables is shown in **Error! Reference source not found.Error! Reference source not found.**. Low linear coefficient values indicate a high level of overall nonlinearity among a few variables. It is discovered that linear relationships might have either a positive or negative direction. Several factors, including LSTAT, RM, PTRATIO, INDUS, and TAX have moderate to high linear relationships with the target variable, MEDV.



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	1	-0.2	0.41	-0.056	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39
zn	-0.2	1	-0.53	-0.043	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41	0.36
indus	0.41	-0.53	1	0.063	0.76	-0.39	0.64	-0.71	0.6	0.72	0.38	-0.36	0.6	-0.48
chas	-0.056	-0.043	0.063	1	0.091	0.091	0.087	-0.099	-0.0074	-0.036	-0.12	0.049	-0.054	0.18
nox	0.42	-0.52	0.76	0.091	1	-0.3	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
rm	-0.22	0.31	-0.39	0.091	-0.3	1	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.7
age	0.35	-0.57	0.64	0.087	0.73	-0.24	1	-0.75	0.46	0.51	0.26	-0.27	0.6	-0.38
dis	-0.38	0.66	-0.71	-0.099	-0.77	0.21	-0.75	1	-0.49	-0.53	-0.23	0.29	-0.5	0.25
rad	0.63	-0.31	0.6	-0.0074	0.61	-0.21	0.46	-0.49	1	0.91	0.46	-0.44	0.49	-0.38
tax	0.58	-0.31	0.72	-0.036	0.67	-0.29	0.51	-0.53	0.91	1	0.46	-0.44	0.54	-0.47
ptratio	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1	-0.18	0.37	-0.51
black	-0.39	0.18	-0.36	0.049	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1	-0.37	0.33
lstat	0.46	-0.41	0.6	-0.054	0.59	-0.61	0.6	-0.5	0.49	0.54	0.37	-0.37	1	-0.74
medv	-0.39	0.36	-0.48	0.18	-0.43	0.7	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1

Figure 2. Bivariate correlation coefficients among the housing variables represented by the correlation heatmap.

2.2. Feature Engineering (FE)

Following a successful initial EDA analysis of the dataset, FE is performed. The ML procedure might not provide an accurate performance that is satisfactory in the absence of a successful FE. Iterative gradient descent cannot provide effective optimization without a thorough analysis of the dataset. Therefore, a thorough FE process is used to change the variables into those that are best suited for the ML algorithms. Imputation, data transformation, data standardization, and the division of the dataset into training, testing, and validation sets are all used in this study's FE. Imputation is used to fill in the null values and make the dataset consistent. Every series contained null values or observations as a result of sensor errors. The values of the nearest neighbors to the blank cell are imputed to these dataset cells. Following a successful imputation using the variable median values, the distribution of the variable series is examined visually and quantitatively to verify its normalcy. A measure of the normality of the variables is the PCS. The neural network regression algorithms do not produce sufficient results with good optimization because the distribution of the discharge and water level variables is significantly non-normal and heavily skewed to the left. The values of a variable are rescaled as part of the data standardization process so that the variable has a mean of 0 and a variance of 1, which is the same as the bell-shaped normal distribution curve (also known as Z-score normalization). The ML recurrent neural network model employs the gradient descent method, where the step size of the method is influenced by the feature value. In gradient descent, smooth progress towards minima necessitates updating the steps for all feature values at the same pace. In the gradient descent procedure, reaching the lowest requires a standardized variable. Equation 1 displays the variable series normalization formula.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X stands for the relevant variable, while the subscripts norm, max, and min signify the normalized variable's maximum and minimum values. The range of the series is divided by the difference between the variable of interest and the minimum of the entire series to provide normalized data, which is then used in the ML's training and testing phases. The full set of normalized variables is divided into two parts: a training set for training the model and a testing set for testing and evaluating the model. Twenty (20%) of the dataset is used for testing, and the remaining eighty (80%) is used for training. In a nutshell, FE and EDA are crucial phases for the successful operation of the ML model.

2.3. Machine Learning Regressors

We will develop the methods and tools required for a model to make a forecast in this project's second portion. The confidence in the forecasts is substantially increased by being able to evaluate each model's performance accurately using these tools and methodologies. To evaluate the accuracy of a particular model, performance on the training and testing sets must be quantified. Generally, a performance metric in some form is employed for this, whether it is through the computation of an error, the goodness of fit, or another useful measurement. We will divide the Boston housing dataset into training and testing subsets for this part. To eliminate bias in the ordering of the dataset, the data are frequently also shuffled into a random order when the training and testing subsets are created. Once trained, it is useful for evaluating our model. We want to determine whether it was correctly learned from a training split of the data. There are three possible circumstances. Underfitting is the term for when a model has learned poorly from the data and is unable to predict even the results of the training set. This is brought on by a high bias. Overfitting is when a model learns the training data so well that it memorizes it and cannot generalize to new data. High variance is what causes overfitting. The model learned well, was able to accurately predict results on new data, and had precisely the proper amount of bias and variation. The model used in this study as a comparison study of the data is represented in **Error! Reference source not found.**.

Table 3. The machine learning regressors models

Linear Regression (LR)	Extreme Gradient Boosting (XGB)
Ridge Regression (RR)	Stochastic Gradient Descent (SGD)
Lasso Regression (LaR)	Gaussian Process Regression (GPR)

Support Vector Machine (SVR)	AdaBoost Regression (ABR)
K-Nearest Neighbors (KNN)	Histogram-based Gradient Boosting (HGBRT)
Decision Tree (DT)	Voting Regressor (VR)
Random Forest (RF)	Stacking Regressor (SR)

Linear regression as one of the ordinary least squares regressors fits a linear model to all the features with the coefficient β where the coefficients are not raised to any power and do not combine in any term to minimize the residual sum of squares between the observed house price (y) and the predicted ones by the model [17] (Eq. 2). Ridge regression leverages L2 regularization to penalize the magnitude of the coefficients to resolve some of the concerns with ordinary least squares [71]. Lasso regression is defined by L1 regularization and panelized terms based on the summation of the coefficient absolute values [72]. The α controls the penalty strength represented in Equations 3 and 4 for both regularized methods.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

$$\text{Penalty term: } P_{RR} = \alpha \sum_{i=1}^{n} {\beta_i}^2 \tag{3}$$

$$\text{Penalty term: } P_{LaR} = \alpha \sum_{i=1}^{n} |\beta_i| \tag{4}$$

Support vector machine basically considers the data points that are within the decision boundary line and the hyperplane that is capable of including a maximum number of data points can be selected as the best-fitted line [73]. For each training point $x_i (i \leq n)$ ($n =$ total number of data points). the objective is to identify target ($\hat{y}$) variable that deviates from observed values by a value no more than ε (Eq. 5) [74,75]. K-nearest neighbors provide the input in the k closest training sets that the predicted output is the average of the values of KNN where the weights to the contributions of the nearest neighbors are greater than the further sets. This approach uses an inverse distance weighted average of the KNN by the following distance functions; Euclidean, Manhattan, and Minkowski (Eq. 6).

$$|y_i - \hat{y_i}| \leq \varepsilon \tag{5}$$

$$\text{Euclidean distance function: } \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

A decision tree is a non-parametric supervised learning technique with an objective to derive decision rules from the features to predict the target variable. To build a model that predicts the value of a target variable, the objective is to comprehend simple decision rules generated from the data features. DT learns regression by dividing the training examples in a manner that minimizes the Sum of Squared Residuals (RSS) and generates a forecast for the output value by averaging all the cases ($\bar{y}_i$) (Eq. 7). Random forest is a stochastic predictor that leverages averaging to increase predicted accuracy and reduce overfitting after fitting various DT to diverse dataset subsamples. RF takes advantage of utilizing the bootstrapping method with a statistical resampling technique that involves random sampling with the replacement of the data set.

$$RSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{7}$$

Extreme gradient boosting provides how to perform the gradient boosting algorithm efficiently for regression predictive modeling. The gradient boosting algorithm minimizes the loss gradient while maximizing the model accuracy [76]. Stochastic gradient descent selects a few data points randomly from the entire data set for each iteration to drastically reduce the computational effort. SGD fits a linear model by minimizing a regularized empirical loss [77].

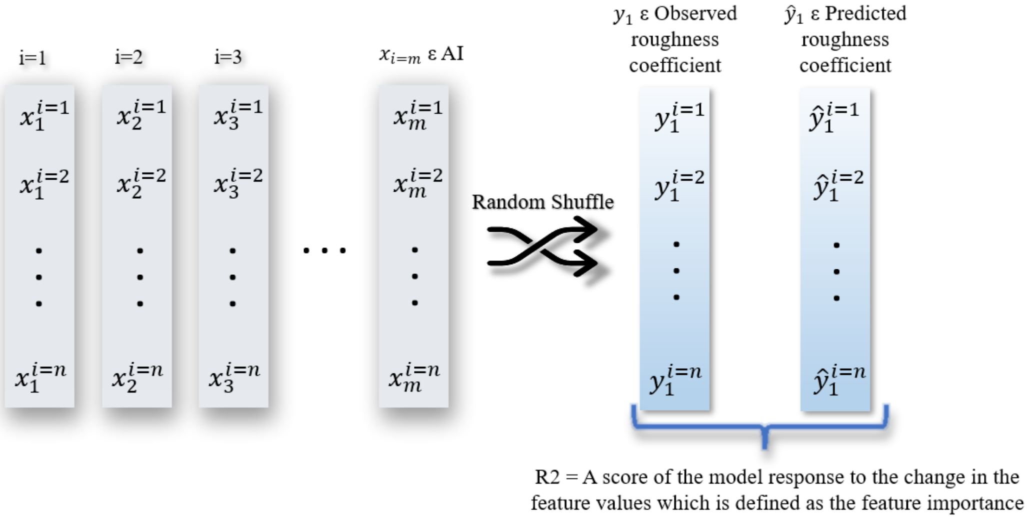
Gaussian process regression as one of the supervised learning methods is nonparametric kernel-based probabilistic model that implements the Gaussian processes [78,79]. AdaBoost regression is a meta-estimator that originates by fitting one regressor on the original dataset and fitting subsequent versions of the regressor on the same dataset with having the weights updated in compliance with the error of the recent prediction [80,81]. Histogram-based gradient boosting is one of the ensembles of DT algorithms that resolve the problem of training on a large dataset. Histogram-based gradient boosting (HGBRT) extracts the continuous input variables to a specific segment of a few hundred distinct values to significantly expedite the training process by this input variable transformation [82].

Voting regressor is an ensemble meta-estimator that fits several base regressors on the entire dataset and the ultimate prediction is then created by averaging the individual predictions from each model [83,84]. This algorithm performs well in balancing individual regression vulnerabilities. Stacking regressor is an ensemble learning technique to combine multiple regression models via a meta-regressor [85–89]. SR is capable of stacking the output of each individual estimator and computing the final prediction by using each estimator's output as an input for a final estimator. Hence it enables the utilization of each estimator's major strengths in the resultant prediction.

2.4. Hyperparameters Optimization

Hyper-parameters are not directly learned from the ML models. In scikit-learn, they are passed as a set of arguments to the constructor of the ML models. It is highly recommended to search the hyper-parameter space for the best model performance. Searching framework involves an algorithm, a parametric space, a method for searching or sampling the candidates of hyperparameters, and a scoring function. In this study, two approaches to parameter search are considered e.g., the grid search cross-validation method (GSCV) and randomized search CV (RSCV). GSCV exhaustively takes all the hyperparameter combinations, while the RSCV samples a few hyperparameter candidates from a parametric space with a given distribution.

The relative feature importance of the predictors is studied by analyzing the permutation feature importance (PFI) technique in the computational domain in **Error! Reference source not found.**. In PFI, the impact of shuffling the values of a feature over the target variable, housing price is quantified to observe the response in output variables due to the change in input variables. The score of the error matrix (R^2) derived from the observed and predicted values of the housing price as a result of the shuffle in the predictors provides the score of the feature importance. In the PFI technique, ML models are run with the values of a specific feature permuted/shuffled keeping the other features constantly and the change in the RMSE values are recorded. Only the output from the ML model i.e., housing price is used as a target variable in estimating the feature importance of other input variables as the output from ML regressors are already used as an input variable in the ML model.



R2 = A score of the model response to the change in the feature values which is defined as the feature importance

Figure 3. Permutation feature importance

2.5. Error Analysis

Several evaluation measures are outlined in the ML literature for various approaches to compare the accuracy of predictions made using regressors. The estimates' error is a measure of the discrepancy between the measured values and predicted values for the data points, as measured by various methods. In this research, various settings for model accuracy and reliability were employed. One error metric cannot accurately capture the value of the housing price predictions among various regressions. Essence, correlation coefficient (r), RMSE, and mean absolute error (MAE) were the top three approaches used to examine the derive regressors' prediction conclusions and evaluate the model accuracies. "Norms" refer to the numerous multi-dimensional error measure designs. The norm normalizations create a relative metric with no dimensions and reduce the error measurements' dependency on the data frame's dimensions.

A common error metric for the model's accuracy and its competency to the values of the data points is the coefficient of determination (R^2) which is derived from the correlation coefficient. A model is a better fit for the data by having a greater R^2 value. The correlation coefficient is the square root of the coefficient of determination calculated by Equation 8.

$$r = \frac{\sum_{i=1}^{N}\left(Q_{i(com)} - \bar{Q}_{(com)}\right)\left(Q_{i(obs)} - \bar{Q}_{(obs)}\right)}{\sqrt{\left[\sum_{i=1}^{N}\left(Q_{i(com)} - \bar{Q}_{(com)}\right)^2\right]\left[\sum_{i=1}^{N}\left(Q_{i(obs)} - \bar{Q}_{(obs)}\right)^2\right]}} \tag{8}$$

Where $Q_{i(com)}$ is the median value of owner-occupied home computed, and $Q_{i(obs)}$is the MEDV observed, for the i^{th} data point. $\bar{Q}_{(com)}$ and $\bar{Q}_{(obs)}$ are the average calculated and observed values of MEDV, respectively. Furthermore, the total number of observations represented by N, is equal to 506 for this study. Between observed and computed values, the R^2 scale ranges from 0 to 1, with 0 indicating a lack of association and 1 denoting perfect correlation.

The RMSE is one of the most used evaluation statistics, which is more vulnerable to significant abnormalities due to the exponential multiplication of major errors by the squared term. RMSE is the average of the errors' absolute values, normalized by the total number of data points, and typically the lowest RMSE score correlates with the highest predicted accuracy (Eq. 9).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|Q_{i(obs)} - Q_{i(com)}\right|^2} \tag{9}$$

MAE is the average absolute error between actual and expected values that assists the understanding of model performance over the entire dataset by discarding the negative differences as all the differences have equal weight. This made the MAE a better interpretable error metric as it utilizes the same predicting variable scale (Eq. 10).

$$MAE = \frac{\left|\left(Q_{i(obs)} - Q_{i(com)}\right)\right|}{N} \tag{10}$$

3. Results

3.1. Predicted and Observed Data

The ML models are deployed to forecast the median house price for testing after a successful training using the obtained median house price data. The best ML-based regression performance is visually represented in **Error! Reference source not found.** by a scatterplot that compares predicted median house price values to observed median house price values. The chosen hyperparameters are used to run the best ML model, RF regressor (number of trees = 100, minimum number of samples needed to divide an internal node = 10, minimum number of samples needed to be at a leaf node = 1). The maximum tree depth is 50, the number of features to be considered while determining the appropriate split is "sqrt," Whether bootstrap samples are used while creating trees is set to True, and the randomness of bootstrapping samples after optimization is set to 0. The chosen hyperparameters for MLP are the size of the hidden layers (100, 50), the activation function (ReLU, Sigmoid), the solver (Adam), the alpha (0.0001), and the learning rate (constant). The R^2 values for the top models are 0.88, 0.87, 0.86 and 0.84 respectively (**Error! Reference source not found.**). In **Error! Reference source not found.**, the fitted regression line's (black dotted lines) statistical distance from the expected median house price values is measured statistically by R^2. where the values of observed and expected median house prices are the same.

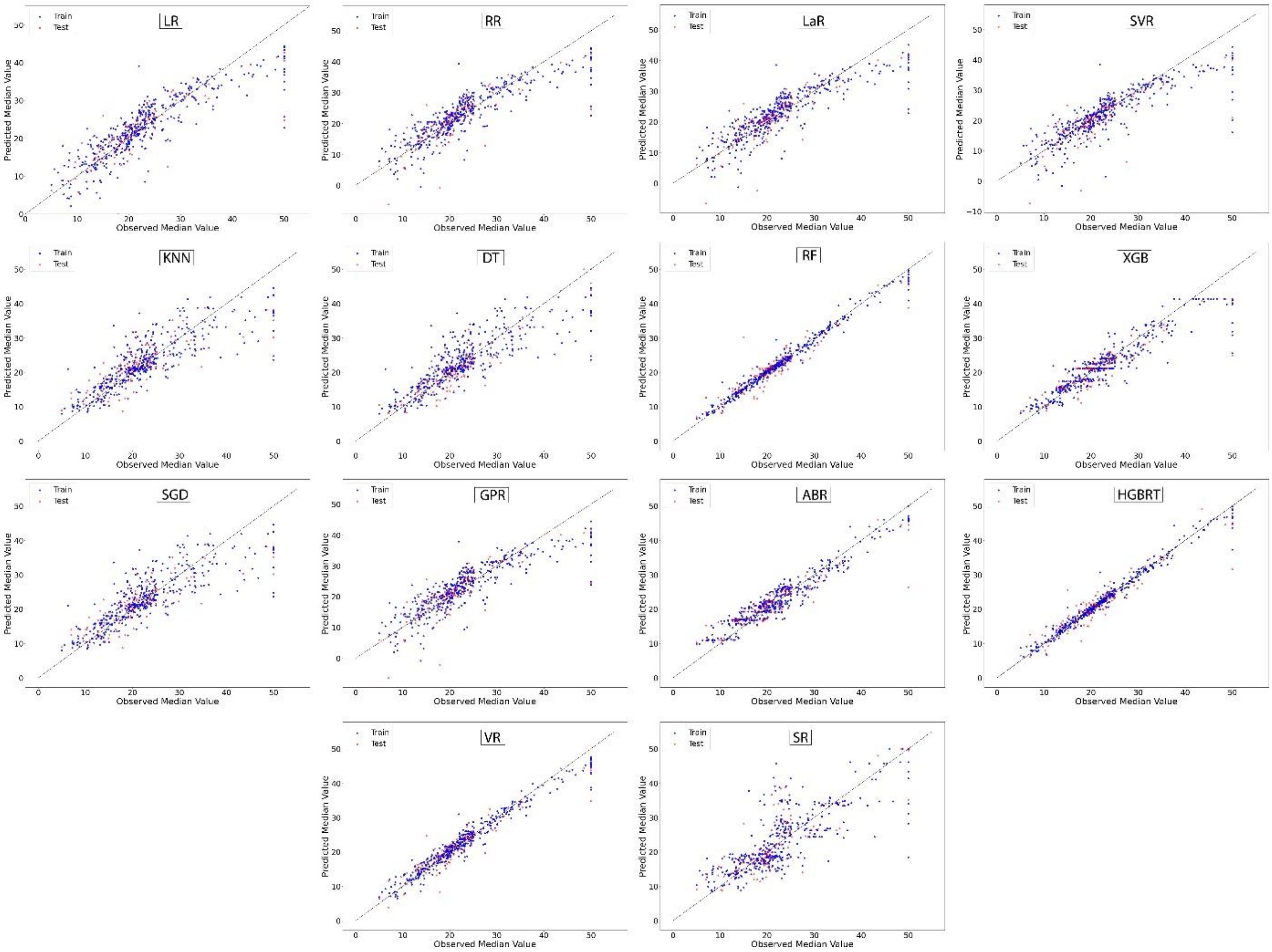


Figure 4. Scatterplots of the observed and predicted median house price from the ML models.

4. Discussion

4.1. Model Evaluation and Improvement

Results of a comparative analysis of the model's efficiency employing all the fourteen regressors are tabulated in **Error! Reference source not found.**. R^2 values range from %0.66 to %0.88 whereas the RMSE ranges from 2.93 to 4.92, for the highest and lowest R^2, respectively. Furthermore, the resultant ranges for MAE consist of the range 2.06 to 2.45. With an R^2 of 0.88, the RD method surpasses all the other thirteen models in forecasting the Boston housing prices depending on the dataset used for this study. VR ranked as the second fitted model based on accuracy scores of 0.87, while the RMSE increases by %3.07 and MAE remains the same compared to RF. On the other hand, HGBRT obtained the third-ranked score of 0.86 with %2.27 decrease in R^2 and %8.53, %4.37 increase in the RMSE and MAE compared to the RF, respectively. RF regressor is run with the selected hyperparameters (number of trees equals to 100, the minimum number of samples required to split an internal node equals to 10, the minimum number of samples required to be at a leaf node equals to 1, number of features to consider when looking for the best split equals to 'sqrt', maximum depth of the tree equals to 50, Whether bootstrap samples are used when building trees equals to “True”, the randomness of the bootstrapping of the samples equals to 0) after optimization. With an R^2 score of 0.698, the tuned RF method surpasses VR, HGBRT, and DT in forecasting median house price, while their respective scores are 0.87, 0.86, and 0.84. The models' RMSE and MAE scores also demonstrated satisfactory performance.

Table 4. Performance comparison of ML regressors

Model	R^2	RMSE	MAE
Random Forest (RF)	0.88	2.93	2.06
Voting Regressor (VR)	0.87	3.02	2.06
Histogram-based Gradient Boosting (HGBRT)	0.86	3.18	2.15
Decision Tree (DT)	0.84	3.37	2.53
Ada-Boost Regression (ABR)	0.82	3.6	2.45
XGBoost (XGB)	0.79	3.83	2.52
Linear Regression (LR)	0.66	4.92	3.18
Ridge Regression (RR)	0.66	4.93	3.14
Gaussian Process Regression (GPR)	0.66	4.93	3.13
Lasso Regression (LaR)	0.65	5.01	3.14
K-Nearest Neighbors (KNN)	0.64	5.08	3.66
Stochastic Gradient Descent (SGD)	0.64	5.08	3.66
Support Vector Machine (SVR)	0.59	5.42	3.14
Stacking Regressor (SR)	0.59	5.44	3.9

4.2. Feature Importance

Based on the change in the R^2 value as a numerical indicator, the PFI technique is used to assess the influence of the features on the ML-based regression. According to PFI analysis, compared to the other characteristics, LSTAT (lower status of the population) has the most significant in predicting median house prices. To demonstrate the strong reaction of median house price predicted from the ML-based regression model due to the change in the predictors, the significance scores are displayed in **Error! Reference source not found.**. RM (Average number of rooms per dwelling), NOX (Nitric oxides concentration), INDUS (Proportion of non-retail business acres per town), PTRATIO (Pupil-teacher ratio by town) and DIS (Weighted distances to five Boston employment centers are the top five most significant features.

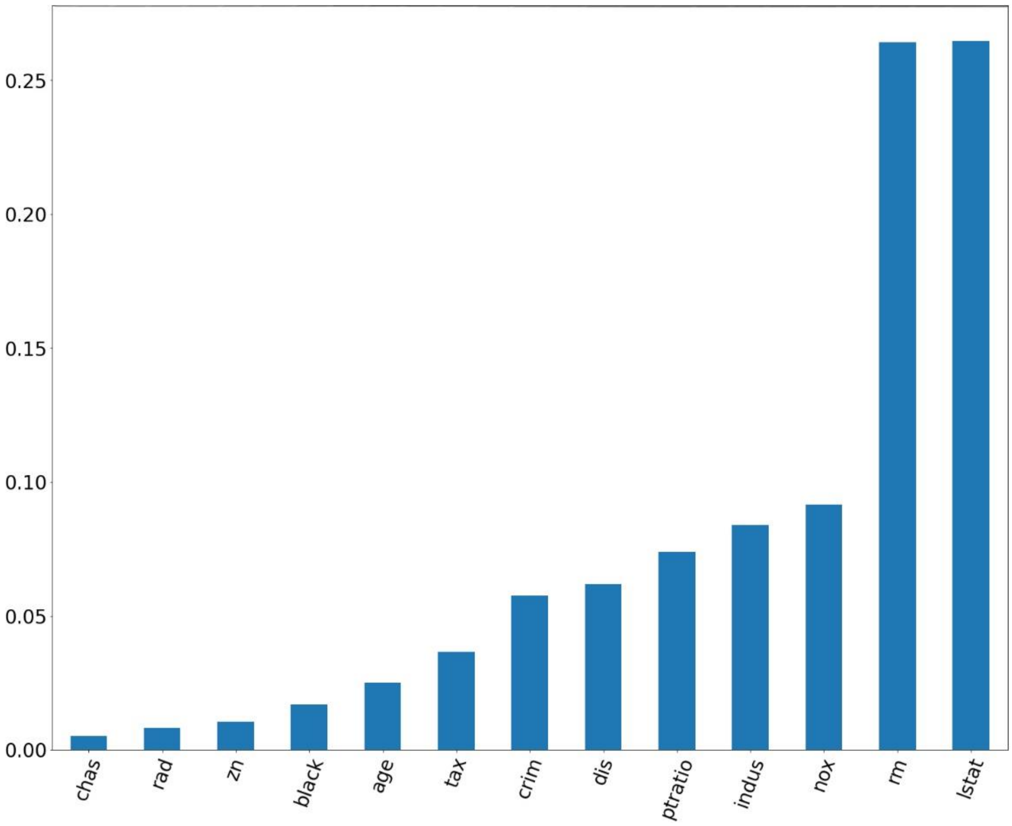


Figure 5. Rank of influencing factors according to their Feature Importance.

5. Conclusions

Data-informed ML approaches provide a path to circumvent the complexities of real estate influencing factors and the computational burden of traditional predictive models. In addition, estimating housing price with model calibration is highly expensive and inefficient computationally at a large scale. A novel ML-based predictive framework is developed to predict housing price in the Boston area in this study. Several features are retrieved and combined from an open-source database. A set of supervised regressors e.g., RF, HGBRT, VR, etc. showed satisfactory performance. Additionally, the feature importance task employed in this study indicates the influencing factors for housing price in chronological order However, data scarcity may result in poor model performance in several regions. By introducing more parameters with greater resolutions, the models' performance can be improved.. Model validation with previously calibrated predictive models at the study locations may enhance the model's applicability and robustness. Further, other ML/DL regressors such as Bayesian regression Multi-layer Perceptron can be explored for higher model performance. The proposed framework illustrates a promising ground and potential to deploy the predicted house price to aid the decision-makers in financial sectors.

Author Contributions: Conceptualization, M.K, S.B.A.; methodology, M.K., S.B.A. and A.G.; Software, M.K., A.G., and H..T.; validation, M.K. A.G., and H.S.; formal analysis, A.G. H.T.; investigation, M.K. H.S.; resources, M.K., A.G.; data curation, H.T. H.S.; writing—original draft preparation, M.K., S.B.A., A.G.; writing—review and editing, M.K., A.G., H.T., and H.S.; visualization, A.G., H.T.; supervision, M.K., H.T.; project administration, H.T., H.S.; funding acquisition, H.S.. All authors have read and agreed to the published version of the manuscript

Funding: This research was supported in part by the Computer Science program in the School of Computing and Analytics at Northern Kentucky University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data collected for the study can be made available upon request from the corresponding author.

Acknowledgments: This study was supported by the Computer Science program in the School of Computing and Analytics at Northern Kentucky University. The authors would also like to thank Dr. Hanieh Shabanian for her support and guidance immensely.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deaton, A. Household Surveys, Consumption, and the Measurement of Poverty. *Econ. Syst. Res.* **2003**, *15*, 135–159, doi:10.1080/0953531032000091144.
2. Olszewski, K.; Waszczuk, J.; Widłak, M. Spatial and Hedonic Analysis of House Price Dynamics in Warsaw, Poland. *J. Urban Plan. Dev.* **2017**, *143*, 04017009, doi:10.1061/(ASCE)UP.1943-5444.0000394.
3. Gu, J.; Zhu, M.; Jiang, L. Housing Price Forecasting Based on Genetic Algorithm and Support Vector Machine. *Expert Syst. Appl.* **2011**, *38*, 3383–3386, doi:10.1016/j.eswa.2010.08.123.
4. Kang, J.; Lee, H.J.; Jeong, S.H.; Lee, H.S.; Oh, K.J. Developing a Forecasting Model for Real Estate Auction Prices Using Artificial Intelligence. *Sustainability* **2020**, *12*, 2899, doi:10.3390/su12072899.
5. Li, S.; Jiang, Y.; Ke, S.; Nie, K.; Wu, C. Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM). *Land* **2021**, *10*, 533, doi:10.3390/land10050533.
6. Pérez-Rave, J.I.; Correa-Morales, J.C.; González-Echavarría, F. A Machine Learning Approach to Big Data Regression Analysis of Real Estate Prices for Inferential and Predictive Purposes. *J. Prop. Res.* **2019**, *36*, 59–96, doi:10.1080/09599916.2019.1587489.
7. Abidoye, R.B.; Chan, A.P.C. Improving Property Valuation Accuracy: A Comparison of Hedonic Pricing Model and Artificial Neural Network. *Pac. Rim Prop. Res. J.* **2018**, *24*, 71–83, doi:10.1080/14445921.2018.1436306.
8. Adair, A.; Berry, J.; McGreal, W. Hedonic Modelling, Housing Submarkets and Residential Valuation. *J. Prop. Res.* **1996**, *13*, 67–83.
9. Selim, H. Determinants of House Prices in Turkey: Hedonic Regression versus Artificial Neural Network. *Expert Syst. Appl.* **2009**, *36*, 2843–2852, doi:10.1016/j.eswa.2008.01.044.
10. Meese, R.; Wallace, N. House Price Dynamics and Market Fundamentals: The Parisian Housing Market. *Urban Stud.* **2003**, *40*, 1027–1045.
11. Stevenson, S. New Empirical Evidence on Heteroscedasticity in Hedonic Housing Models. *J. Hous. Econ.* **2004**, *13*, 136–153, doi:10.1016/j.jhe.2004.04.004.
12. Ngiam, K.Y.; Khor, I.W. Big Data and Machine Learning Algorithms for Health-Care Delivery. *Lancet Oncol.* **2019**, *20*, e262–e273, doi:10.1016/S1470-2045(19)30149-4.
13. Hastie, T.; Rosset, S.; Tibshirani, R.; Zhu, J. The Entire Regularization Path for the Support Vector Machine. 25.
14. Liu, L.; Wu, L. Predicting Housing Prices in China Based on Modified Holt's Exponential Smoothing Incorporating Whale Optimization Algorithm. *Socioecon. Plann. Sci.* **2020**, *72*, 100916, doi:10.1016/j.seps.2020.100916.
15. Ge, C.; Wang, Y.; Xie, X.; Liu, H.; Zhou, Z. An Integrated Model for Urban Subregion House Price Forecasting: A Multi-Source Data Perspective. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM); November 2019; pp. 1054–1059.
16. Mullainathan, S.; Spiess, J. Machine Learning: An Applied Econometric Approach. *J. Econ. Perspect.* **2017**, *31*, 87–106, doi:10.1257/jep.31.2.87.
17. Sklearn.Linear_model.LinearRegression Available online: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed on 5 September 2022).

18. Khare, S.; Gourisaria, M.K.; Harshvardhan, G.M.; Joardar, S.; Singh, V. Real Estate Cost Estimation Through Data Mining Techniques. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1099*, 012053, doi:10.1088/1757-899X/1099/1/012053.
19. Louati, A.; Lahyani, R.; Aldaej, A.; Aldumaykhi, A.; Otai, S. Price Forecasting for Real Estate Using Machine Learning: A Case Study on Riyadh City. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6748, doi:10.1002/cpe.6748.
20. Sklearn.Tree.DecisionTreeRegressor Available online: https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html (accessed on 5 September 2022).
21. Wu, H.; Wang, C. A New Machine Learning Approach to House Price Estimation. *New Trends Math. Sci.* **2018**, *4*, 165–171, doi:10.20852/ntmsci.2018.327.
22. Varian, H.R. Big Data: New Tricks for Econometrics. *J. Econ. Perspect.* **2014**, *28*, 3–28, doi:10.1257/jep.28.2.3.
23. Sklearn.Ensemble.RandomForestRegressor Available online: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 5 September 2022).
24. Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes. *BMC Med. Inform. Decis. Mak.* **2010**, *10*, 16, doi:10.1186/1472-6947-10-16.
25. Karimi, M.; Khosravi, M.; Fathollahi, R.; Khandakar, A.; Vaferi, B. Determination of the Heat Capacity of Cellulosic Biosamples Employing Diverse Machine Learning Approaches. *Energy Sci. Eng.* **2022**, *10*, 1925–1939, doi:10.1002/ese3.1155.
26. Abdollahzadeh, M.; Khosravi, M.; Hajipour Khire Masjidi, B.; Samimi Behbahan, A.; Bagherzadeh, A.; Shahkar, A.; Tat Shahdost, F. Estimating the Density of Deep Eutectic Solvents Applying Supervised Machine Learning Techniques. *Sci. Rep.* **2022**, *12*, 4954, doi:10.1038/s41598-022-08842-5.
27. Sklearn.Svm.SVR Available online: https://scikit-learn/stable/modules/generated/sklearn.svm.SVR.html (accessed on 5 September 2022).
28. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27, doi:10.1145/1961189.1961199.
29. Huang, Y. Predicting Home Value in California, United States via Machine Learning Modeling. *Stat. Optim. Inf. Comput.* **2019**, *7*, 66–74, doi:10.19139/soic.v7i1.435.
30. Zhang, Y.; Haghani, A. A Gradient Boosting Method to Improve Travel Time Prediction. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 308–324, doi:10.1016/j.trc.2015.02.019.
31. Python API Reference — Xgboost 1.6.2 Documentation Available online: https://xgboost.readthedocs.io/en/stable/python/python_api.html (accessed on 5 September 2022).
32. Alkhammash, E.H. An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production. *Energies* **2022**, *15*, 6416, doi:10.3390/en15176416.
33. Lindgren, J. General Equilibrium with Price Adjustments—A Dynamic Programming Approach. *Analytics* **2022**, *1*, 27–34, doi:10.3390/analytics1010003.
34. Mehedi, M.A.A.; Yazdan, M.M.S.; Ahad, M.T.; Akatu, W.; Kumar, R.; Rahman, A. Quantifying Small-Scale Hyporheic Streamlines and Resident Time under Gravel-Sand Streambed Using a Coupled HEC-RAS and MIN3P Model. *Eng* **2022**, *3*, 276–300, doi:10.3390/eng3020021.
35. Beretta, L.; Santaniello, A. Nearest Neighbor Imputation Algorithms: A Critical Evaluation. *BMC Med. Inform. Decis. Mak.* **2016**, *16*, 74, doi:10.1186/s12911-016-0318-z.
36. Kang, S. K-Nearest Neighbor Learning with Graph Neural Networks. *Mathematics* **2021**, *9*, 830, doi:10.3390/math9080830.
37. Kück, M.; Freitag, M. Forecasting of Customer Demands for Production Planning by Local K-Nearest Neighbor Models. *Int. J. Prod. Econ.* **2021**, *231*, 107837, doi:10.1016/j.ijpe.2020.107837.

38. Ahmad, M.; Al Mehedi, M.A.; Yazdan, M.M.S.; Kumar, R. Development of Machine Learning Flood Model Using Artificial Neural Network (ANN) at Var River. *Liquids* **2022**, *2*, 147–160, doi:10.3390/liquids2030010.

39. Zurada, J.; Levitan, A.; Guan, J. A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *J. Real Estate Res.* **2011**, *33*, 349–388, doi:10.1080/10835547.2011.12091311.

40. Fior, J.; Cagliero, L.; Garza, P. Leveraging Explainable AI to Support Cryptocurrency Investors. *Future Internet* **2022**, *14*, 251, doi:10.3390/fi14090251.

41. Anand, C. Comparison of Stock Price Prediction Models Using Pre-Trained Neural Networks. *J. Ubiquitous Comput. Commun. Technol.* **2021**, *3*, 122–134, doi:10.36548/jucct.2021.2.005.

42. Kumar, R.; Yazdan, M.M.S.; Mehedi, M.A.A. Demystifying the Preventive Measures for Flooding from Groundwater Triggered by the Rise in Adjacent River Stage 2022.

43. Yan, Z.; Zong, L. Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms. In Proceedings of the Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence; Association for Computing Machinery: New York, NY, USA, August 25 2020; pp. 64–71.

44. Mehedi, M.A.A.; Yazdan, M.M.S. Automated Particle Tracing & Sensitivity Analysis for Residence Time in a Saturated Subsurface Media. *Liquids* **2022**, *2*, 72–84, doi:10.3390/liquids2030006.

45. Abdullah Al Mehedi, M.; Reichert, N.; Molkenthin, F. Sensitivity Analysis of Hyporheic Exchange to Small Scale Changes In Gravel-Sand Flumebed Using A Coupled Groundwater-Surface Water Model. **2020**, 20319, doi:10.5194/egusphere-egu2020-20319.

46. Zhu, X.; Khosravi, M.; Vaferi, B.; Nait Amar, M.; Ghriga, M.A.; Mohammed, A.H. Application of Machine Learning Methods for Estimating and Comparing the Sulfur Dioxide Absorption Capacity of a Variety of Deep Eutectic Solvents. *J. Clean. Prod.* **2022**, *363*, 132465, doi:10.1016/j.jclepro.2022.132465.

47. Liu, J.; Wang, B.; Xiao, L. Non-Linear Associations between Built Environment and Active Travel for Working and Shopping: An Extreme Gradient Boosting Approach. *J. Transp. Geogr.* **2021**, *92*, 103034, doi:10.1016/j.jtrangeo.2021.103034.

48. Wang, P.-Y.; Chen, C.-T.; Su, J.-W.; Wang, T.-Y.; Huang, S.-H. Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism. *IEEE Access* **2021**, *9*, 55244–55259, doi:10.1109/ACCESS.2021.3071306.

49. Ho, W.K.O.; Tang, B.-S.; Wong, S.W. Predicting Property Prices with Machine Learning Algorithms. *J. Prop. Res.* **2021**, *38*, 48–70, doi:10.1080/09599916.2020.1832558.

50. Rico-Juan, J.R.; Taltavull de La Paz, P. Machine Learning with Explainability or Spatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. *Expert Syst. Appl.* **2021**, *171*, 114590, doi:10.1016/j.eswa.2021.114590.

51. Embaye, W.T.; Zereyesus, Y.A.; Chen, B. Predicting the Rental Value of Houses in Household Surveys in Tanzania, Uganda and Malawi: Evaluations of Hedonic Pricing and Machine Learning Approaches. *PLOS ONE* **2021**, *16*, e0244953, doi:10.1371/journal.pone.0244953.

52. Shahhosseini, M.; Hu, G.; Pham, H. Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction. In Proceedings of the Smart Service Systems, Operations Management, and Analytics; Yang, H., Qiu, R., Chen, W., Eds.; Springer International Publishing: Cham, 2020; pp. 87–97.

53. Graczyk, M.; Lasota, T.; Trawiński, B.; Trawiński, K. Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In Proceedings of the Intelligent Information and Database Systems; Nguyen, N.T., Le, M.T., Świątek, J., Eds.; Springer: Berlin, Heidelberg, 2010; pp. 340–350.

54. Liu, X.; Deng, Z.; Wang, T. Real Estate Appraisal System Based on GIS and BP Neural Network. *Trans. Nonferrous Met. Soc. China* **2011**, *21*, s626–s630, doi:10.1016/S1003-6326(12)61652-5.

55. Li, Z.; Piao, W.; Wang, L.; Wang, X.; Fu, R.; Fang, Y. China Coastal Bulk (Coal) Freight Index Forecasting Based on an Integrated Model Combining ARMA, GM and BP Model Optimized by GA. *Electronics* **2022**, *11*, 2732, doi:10.3390/electronics11172732.
56. Xu, X.; Zhang, Y. House Price Forecasting with Neural Networks. *Intell. Syst. Appl.* **2021**, *12*, 200052, doi:10.1016/j.iswa.2021.200052.
57. Wang, X.; Wen, J.; Zhang, Y.; Wang, Y. Real Estate Price Forecasting Based on SVM Optimized by PSO. *Optik* **2014**, *125*, 1439–1443, doi:10.1016/j.ijleo.2013.09.017.
58. Park, B.; Bae, J.K. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. *Expert Syst. Appl.* **2015**, *42*, 2928–2934, doi:10.1016/j.eswa.2014.11.040.
59. Bertoli, W.; Oliveira, R.P.; Achcar, J.A. A New Semiparametric Regression Framework for Analyzing Non-Linear Data. *Analytics* **2022**, *1*, 15–26, doi:10.3390/analytics1010002.
60. Steurer, M.; Hill, R.J.; Pfeifer, N. Metrics for Evaluating the Performance of Machine Learning Based Automated Valuation Models. *J. Prop. Res.* **2021**, *38*, 99–129, doi:10.1080/09599916.2020.1858937.
61. Peng, H.; Li, J.; Wang, Z.; Yang, R.; Liu, M.; Zhang, M.; Yu, P.; He, L. Lifelong Property Price Prediction: A Case Study for the Toronto Real Estate Market. *IEEE Trans. Knowl. Data Eng.* **2021**, 1–1, doi:10.1109/TKDE.2021.3112749.
62. Mehedi, M.A.A.; Khosravi, M.; Yazdan, M.M.S.; Shabanian, H. Exploring Temporal Dynamics of River Discharge Using Univariate Long Short-Term Memory (LSTM) Recurrent Neural Network at East Branch of Delaware River 2022.
63. Yazdan, M.M.S.; Khosravia, M.; Saki, S.; Mehedi, M.A.A. Forecasting Energy Consumption Time Series Using Recurrent Neural Network in Tensorflow Tensorflow. Preprints. **2022**. 2022090404, doi: 10.20944/preprints202209.0404.v1.
64. Su, T.; Li, H.; An, Y. A BIM and Machine Learning Integration Framework for Automated Property Valuation. *J. Build. Eng.* **2021**, *44*, 102636, doi:10.1016/j.jobe.2021.102636.
65. Kang, Y.; Zhang, F.; Peng, W.; Gao, S.; Rao, J.; Duarte, F.; Ratti, C. Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning. *Land Use Policy* **2021**, *111*, 104919, doi:10.1016/j.landusepol.2020.104919.
66. UCI Machine Learning Repository Available online: http://archive.ics.uci.edu/ml/index.php (accessed on 5 September 2022).
67. Boston Housing Available online: https://www.kaggle.com/datasets/schirmerchad/bostonhoustingmlnd (accessed on 24 October 2022).
68. Redfin Boston Housing Market: House Prices & Trends | Redfin Available online: https://www.redfin.com/city/1826/MA/Boston/housing-market (accessed on 5 September 2022).
69. Santarelli, M. Boston Housing Market: Prices | Trends | Forecasts 2022 Available online: https://www.noradarealestate.com/blog/boston-real-estate-market/ (accessed on 5 September 2022).
70. Boston Dataset Available online: https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html (accessed on 5 September 2022).
71. Sklearn.Linear_model.Ridge Available online: https://scikit-learn/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed on 5 September 2022).
72. Sklearn.Linear_model.Lasso Available online: https://scikit-learn/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed on 5 September 2022).
73. Khosravi, M.; Tabasi, S.; Hossam Eldien, H.; Motahari, M.R.; Alizadeh, S.M. Evaluation and Prediction of the Rock Static and Dynamic Parameters. *J. Appl. Geophys.* **2022**, *199*, 104581, doi:10.1016/j.jappgeo.2022.104581.
74. Understanding Support Vector Machine Regression - MATLAB & Simulink Available online: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html (accessed on 5 September 2022).

75. Platt, J.C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Proceedings of the Advances in Large Margin Classifiers; MIT Press, 1999; pp. 61–74.

76. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System 2016.

77. Sklearn.Linear_model.SGDRegressor Available online: https://scikit-learn/stable/modules/generated/sklearn.linear_model.SGDRegressor.html (accessed on 5 September 2022).

78. 1.7. Gaussian Processes Available online: https://scikit-learn/stable/modules/gaussian_process.html (accessed on 5 September 2022).

79. Gaussian Process Regression Models - MATLAB & Simulink Available online: https://www.mathworks.com/help/stats/gaussian-process-regression-models.html (accessed on 5 September 2022).

80. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139, doi:10.1006/jcss.1997.1504.

81. Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-Class AdaBoost. *Stat. Interface* **2009**, *2*, 349–360, doi:10.4310/SII.2009.v2.n3.a8.

82. Cai, Y.; Hang, H.; Yang, H.; Lin, Z. Boosted Histogram Transform for Regression. In Proceedings of the Proceedings of the 37th International Conference on Machine Learning; PMLR, November 21 2020; pp. 1251–1261.

83. Erdebilli, B.; Devrim-İçtenbaş, B. Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. *Mathematics* **2022**, *10*, 2466, doi:10.3390/math10142466.

84. Sklearn.Ensemble.VotingRegressor Available online: https://scikit-learn/stable/modules/generated/sklearn.ensemble.VotingRegressor.html (accessed on 5 September 2022).

85. Sklearn.Ensemble.StackingRegressor Available online: https://scikit-learn/stable/modules/generated/sklearn.ensemble.StackingRegressor.html (accessed on 5 September 2022).

86. Brownlee, J. Stacking Ensemble Machine Learning With Python. *Mach. Learn. Mastery* 2020.

87. Akter, M.S.; Shahriar, H.; Chowdhury, R.; Mahdy, M.R.C. Forecasting the Risk Factor of Frontier Markets: A Novel Stacking Ensemble of Neural Network Approach. *Future Internet* **2022**, *14*, 252, doi:10.3390/fi14090252.

88. Abdellatif, A.; Mubarak, H.; Ahmad, S.; Ahmed, T.; Shafiullah, G.M.; Hammoudeh, A.; Abdellatef, H.; Rahman, M.M.; Gheni, H.M. Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model. *Sustainability* **2022**, *14*, 11083, doi:10.3390/su141711083.

89. Khosravi, M.; Mehedi, M.A.A.; Baghalian, S.; Burns, M.; Welker, A.L.; Golub, M. Using Machine Learning to Improve Performance of a Low-Cost Real-Time Stormwater Control Measure. Preprints. **2022**. 2022110519. doi: 10.20944/preprints202211.0519.v1.

Abdellatif, A.; Mubarak, H.; Ahmad, S.; Ahmed, T.; Shafiullah, G.M.; Hammoudeh, A.; Abdellatef, H.; Rahman, M.M.; Gheni, H.M. Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model. *Sustainability* **2022**, *14*, 11083, doi:10.3390/su141711083.