Research Reports on Computer Science
https://ojs.wiserpub.com/index.php/RRCS/

UNIVERSAL WISER PUBLISHER

Research Article in Special Issue: Selected Papers from the 4th International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (MIND-2022)

An Analysis of House Price Prediction Using Ensemble Learning Algorithms

Boyapati Sai Venkat[1], Maddirala Sai Karthik[1], Konakanchi Subrahmanyam[1], B Ramachandra Reddy[2*]

[1]Department of Computer Science and Engineering, SRM University, AP, India
[2]Department of Computer Science and Engineering, National Institute of Technology, Jamshedpur, Jharkhand, India
E-mail: saivenkat_boyapati@srmap.edu.in

Received: 8 March 2023; **Accepted:** 28 March 2023

Abstract: It is very important to understand the market drifts in the wake of booming civilization and ever-changing market requirements. The principal purpose of the study is the prediction of house prices based on current conditions. From historical data on property markets, literature attempts to draw useful insights. Business trends must be understood so that individuals may prepare their budgetary needs accordingly. A society that is ever-expanding is driven by the growing real estate industry. A lot of clients have been duped by agents setting up a fake market rate. As a result, the real estate industry has become less transparent in recent years. Due to decreased accuracy and overfitting of data, the previous model reduced efficiency, whereas the newly developed model resolves such issues and provides a rich user interface with a better model. An important part of this study is to develop an extensive model that is beneficial to both business societies and individuals. This is the main objective of this study. In order to simplify the client's fieldwork and free up his time and money, this software is intended to assist him. Machine learning algorithms enable models to be enlightened such as root mean square error, random forest, support vector machine, k-nearest neighbors, mean squared error, extreme gradient boost, mean absolute error, R-squared score, linear regression, AdaBoost, CatBoost.

Keywords: house price prediction, machine learning, ensemble learning, support vector machine, gradient boost, random forest, CatBoost

1. Introduction

In today's society, the real estate market is constantly a hot topic. Housing is critical to one's development. People are responsible when it comes to spending and economic approaches when currently looking for a new property. The real estate market is rapidly growing due to the growth of the economy. Because the price of housing is controlled by emerging markets and a variety of other factors. There are many factors that influence the price of homes. Buyers, on the other hand, are particularly sensitive to price fluctuations. Investors need to be aware of business trends to ensure that underwriting can be done correctly and that business output is augmented.

The main purpose of the project is to forecast home price volatility to determine the best methodology for

DOI: https://doi.org/10.37256/rrcs.2320232639

predicting house prices with less inappropriate data through the use of appropriate algorithms. This estimate is consistent with house prices for nonhomeowners based on their financial resources and objectives. Predicted prices will indeed be generated by analyzing the preceding merchandise, fare ranges, and forewarning developments. Developing a high-accuracy model is feasible for researchers who have confidence in the data set.

In this article, we use ensemble learning to forecast house prices. The ensemble learning technique is now thought to be an outstanding prediction tool algorithm. Ensemble learning reduces prediction error by reducing the various factors of the base classifiers. The stability of the basic classifiers determines the overall model's performance. It doesn't focus on any specific occurrences in the training data set because the likelihood of each sample being picked is the same. Ensemble learning performs exceptionally well in prediction, classification, and regression tasks. Ensemble learning beats the basic learner in terms of generalization and prediction accuracy, according to the findings. It also indicates that the proposed methodology is appropriate to property price forecasts and has some pragmatic and referential value.

The rest of the work is structured as follows. Section 2 summarizes some related work on software failure prediction using different machine learning models. Section 3 briefly describes the proposed deep neural network model for predicting software failures. Section 4 presents the experimental data and conclusions in Section 5.

2. Related work

Real estate is not only a vital engine for economic growth, but it is also a major source of concern for the general public. People's attention to house prices continues to increase dramatically with increasing housing demand. The real estate sector has emerged as something of a competitive and opaque market [1]. Alternative approaches to forecasting property prices based on all intrinsic and extrinsic factors that contributed to the price without any fluctuations have been developed over time [2]. Although the data provided play a part in determining the best future prediction model. Our methodology examines a set of parameters specified by the customer in determining the most effective pricing for their needs and desires [3]. The price of a property might fluctuate by location, area, amenities, and other characteristics, and researchers have been trying to find the best possible predictive model to forecast that price over the past decade [4]. Various approaches to estimating the property price employing various models and a combination of models were identified in our literature review. As a result, providing accurate house price estimates is crucial [5]. Multiple factors influence housing prices, including time and space, house age, surrounding conditions, communities, transportation, and so on. Existing forecasting models are often single predictors, i.e., the prediction is made using only one forecasting model. When data sets are noisy, this model's prediction accuracy is not strong sufficient [6].

To find the best algorithms for identifying housing market movements, researchers analyzed previous rates of growth or house price indices, which are usually computed from a nationwide average house price [7]. Ensemble learning is now regarded to be an outstanding prediction tool algorithm. Ensemble learning's generalization performance can be enhanced by combining basic learners [8]. By reducing the dispersion of the base classifiers, ensemble learning improves prediction error. The stability of the basic classifiers determines the overall model's performance. Ensemble the learning accuracy of an instrument well in prediction, classification, and regression tasks [9].

To estimate the price of real estate, an ensemble learning algorithm model was used, and the results have been compared to a single decision tree classifier. Ensemble learning beats the basic learner in terms of adaptability and prediction accuracy, according to the findings. It also shows that the proposed model is applicable to predict property prices and that it has considerable practical and reference value [10].

3. Proposed method

Several machine learning methods are implemented in the proposed framework (Figure 1), which is then assembled into the voting classifier. Various machine learning classifications are analyzed and compared, and their results were reported.

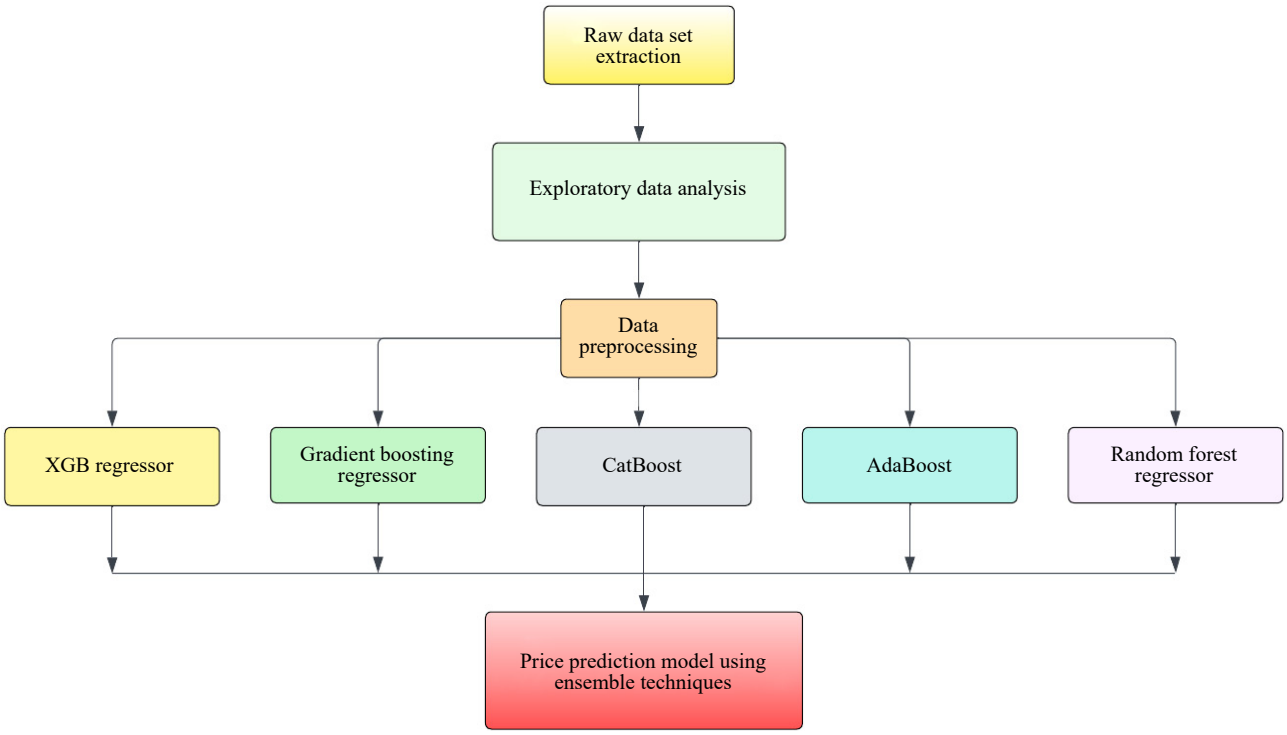

Figure 1. Proposed method

3.1 *Machine learning techniques*

K-Nearest Neighbor (KNN). The KNN is a regression and classifying predictive analysis machine learning method. It's also known as a lazy learner algorithm because it doesn't study the data it's trained on, instead, it classifies the new data it's tested on based on its similarities. KNN uses feature similarity to estimate house price values, which assigns a value to new data based on how closely it relates to the points in the training set.

$$\sqrt{\sum_{i-1}^{k}(x_i - y_i)^2}, \quad \sum_{i-1}^{k}|x_i - y_i| \tag{1}$$

where X is the new argument, Y denotes the existing point, and K indicates the K-factor (number of clusters evaluated by the algorithm before allocating a value).

It's a supervised learning algorithm that's easy to use, versatile, and implement. It is based on the idea that creating a bond is near in proximity. To express the concept of similarity, it calculates the distance between two points on a graph. The numerical parameter '*k*' in the algorithm shows how many data points should be taken into account when voting. We encircle a new point with K number of data points and assign it to the group with the most points within the circle to classify it. The best technique to figure out the value of K is to try out a few different values before settling on one, which decreases the error while maintaining the prediction's accuracy.

Linear Regression. This model uses statistical methods to determine the relationship between two or more features present in the data set. The relation between the dependent and independent is computed using the function containing the least squares.

As a general rule, linear regression equations are as follows:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{2}$$

The beta represents the estimated parameter for x_i as independent variables, y as the dependent variable.

Random Forest. The random forest techniques are a decision tree ensemble method that is difficult to understand.

It employs a method known as tree bagging. By designing the decision tree, the problem of decision tree instability and excessive variety can be addressed. Because they are built using random sampling methods, these decision trees are referred to as random forests (Figure 2).

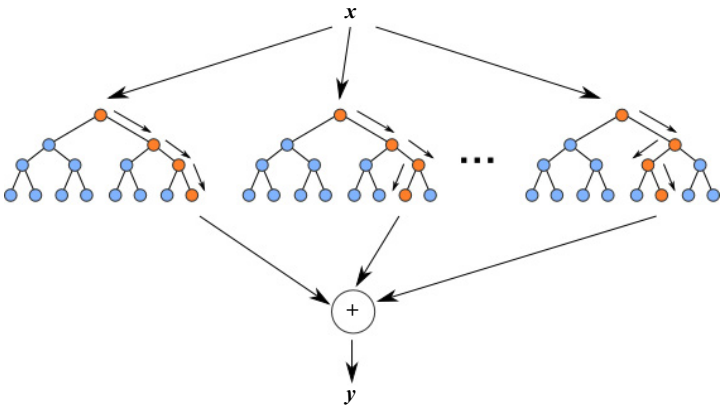

Figure 2. Random forest

It selects a random subset of the total features and factors to train the weak learners. A model averaging strategy is used to integrate the forecasts of each tree. It chooses a set of features and factors at random and creates its own type of decision tree.

The decision trees of random forests are independent of one another. Majority voting can be used on these types of trees.

Support Vector Machine (SVM). The SVM is a supervised machine learning technique used for prediction purposes. To find a hyperplane for data segmentation, the SVM technique turns the initial set into a higher dimensional space. Support vectors, also known as "essential training tuples," define the hyperplane. SVM is more accurate than other algorithms because of its ability to accept nonlinear boundaries.

3.2 *Ensemble learning techniques*

An ensemble is basically a combination of decisions from various models to make decisions that help in improving the accuracy of predicting the results.

The major divisions/methods in ensemble learning, namely Bagging, Boosting, and Stacking methods. This paper mainly focuses on Bagging and Boosting techniques in order to achieve high accuracy scores.

Gradient Boost. We're utilizing gradient boost for regression when we use it to predict a continuous value. This is distinct from the use of linear regression. By translating data into a tree representation, the decision tree tackles the problem of machine learning. It consists of internal nodes for each attribute and the leaf node is represented as the class level. In most cases, the loss function is the squared error.

XGBoost. A gradient boost framework is used in XGBoost, which is one of the ensemble techniques. It's a gradient-boosting method that minimizes over-fittings and bias by combining parallel processing, tree pruning, missing value management, and regularization. XGBoost uses ensemble learning to forecast a single value, which takes into account a set of models known as base learners. Because not all base learners are expected to make poor predictions, the poor guesses are balanced out by the outstanding predictions when they are combined together.

AdaBoost. The AdaBoost is a regression approach that uses both "simple" and "weak" classifiers to create a "strong" classifier. The algorithm fits a regression to the initial data set, then fits multiple copies of the regression to the alike data set, weighing each instance according to the current prediction's error.

CatBoost. In boosting, multiple weak models are sequentially integrated and a strong competitive predictor is constructed using greedy search. Additions of new functions are added until and unless the chosen loss function can no longer be minimized.

3.3 *Data set description*

Ames Housing Data Set (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data). The current house price data set was extracted from Kaggle. The data set chosen has test and train data in which the training data set consists of 80 independent variables and 1 target variable with 1,460 rows, and the test data consists of 80 dependent variables with 1,459 rows.

The target variable from the train data has been removed and assigned to another variable in order to combine train and test data.

The data set consists of some redundant attributes which are removed in the next phase to avoid data disturbance. Important parameters that hugely affect the price of a house, GrLivArea, OverallQual, YearBuilt, YearRemodAdd, and GarageArea are thoroughly examined in the data set.

King County Data Set (https://www.kaggle.com/datasets/harlfoxem/housesalesprediction). The second data set based on King County Housing is a whole single data set. The data set consists of 21 columns, some of which are redundant and some of which are unnecessary.

The target feature is the price column which is also termed the dependent variable. All the other independent features will be used to predict house prices.

3.4 *Data preprocessing and analysis*

Data preprocessing is a critical step in transforming a large-scale data set into a usable format. Narrow down the training of the model to the right level, this encompasses eliminating the NaN values (missing values) and inappropriate noisy data in the data set. To make our data set acceptable for training the ensemble learning model, we removed records that contained NaN values.

Ames Housing Data Set. The paper uses a house price training data set. On the basis of the analysis, it is observed that the data set contains numerical as well as categorical attributes. On further analysis, it is found that the data set contains null and missing values.

During the analysis, some of the attributes were found to contain outliers and therefore skewness. The data set contains the dependent attributes (features), i.e., SalePrice, and independent attributes that included some redundancies. Correlation analysis was performed between different features in the data set to see how well the features are correlated with each other. We then dropped the unnecessary features (columns) from the data set which when present can affect the prediction model resulting in a prediction that is less accurate.

The next process is to handle the missing and incorrect values. For this process, we first plotted a heatmap (Figure 3) to check which columns (features) have missing values. We then replaced the incorrect values with “NAN” which can be visualized in Figure 4.

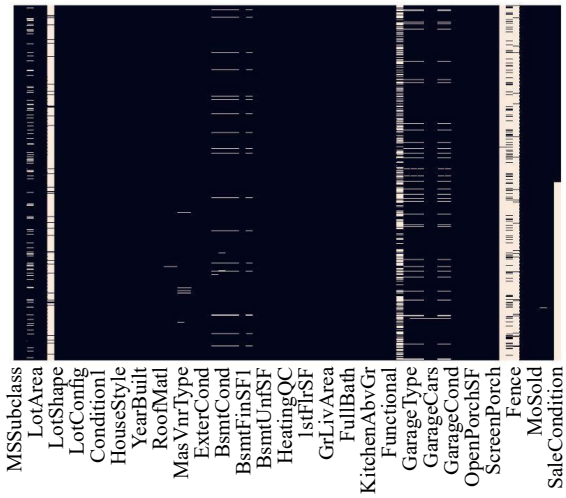

Figure 3. Before handling missing values

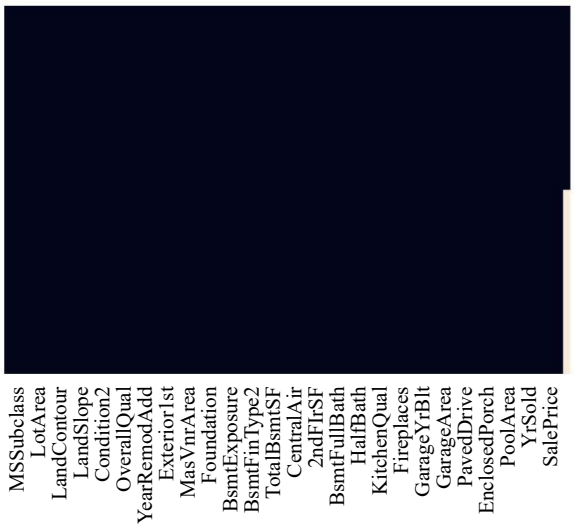

Figure 4. After handling missing values

The columns having a percentage of a missing value greater than 75% were dropped. Later, the entire data set is divided into numerical and categorical features respectively. The numerical features (columns) are then transformed using the data transformation techniques. In this case, we've used log transformation techniques to lessen the skewness of the features. These categorical features are then converted into numerical features that the machines can understand. This conversion is achieved through the process of "encoding".

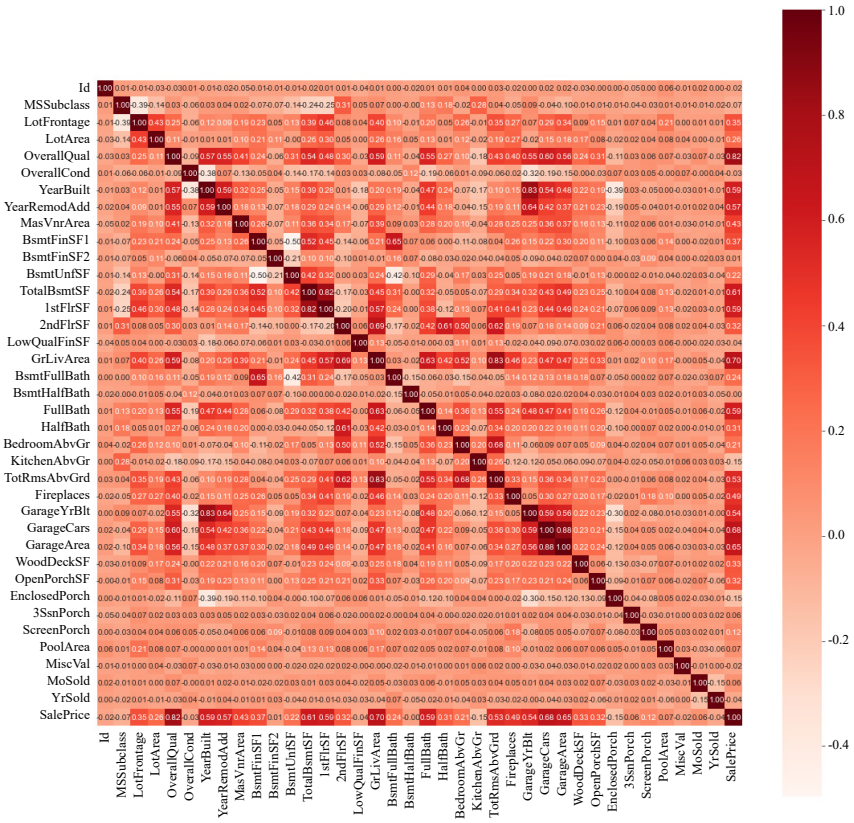

Figure 5. Correlation analysis between all the features in the Ames housing data set

King County Data Set. The King County data set is observed to have zero missing values on initial observation. Some unnecessary column features like "id" are then dropped from the data set. It is found that there are some redundant columns present in the data set. We first expanded the date feature into 3 separate columns, namely day, month, and year. The year is the only column that is sufficient for our analysis. Therefore, the other two columns are then dropped from the data set.

The next thing is to remove the redundant column features. So, we used the columns "yr_renovated", "yr_built" and "year" to add a new column "Age" that determines the age of the house. After this computation, all three features describing the year renovated, year built, and year are then dropped.

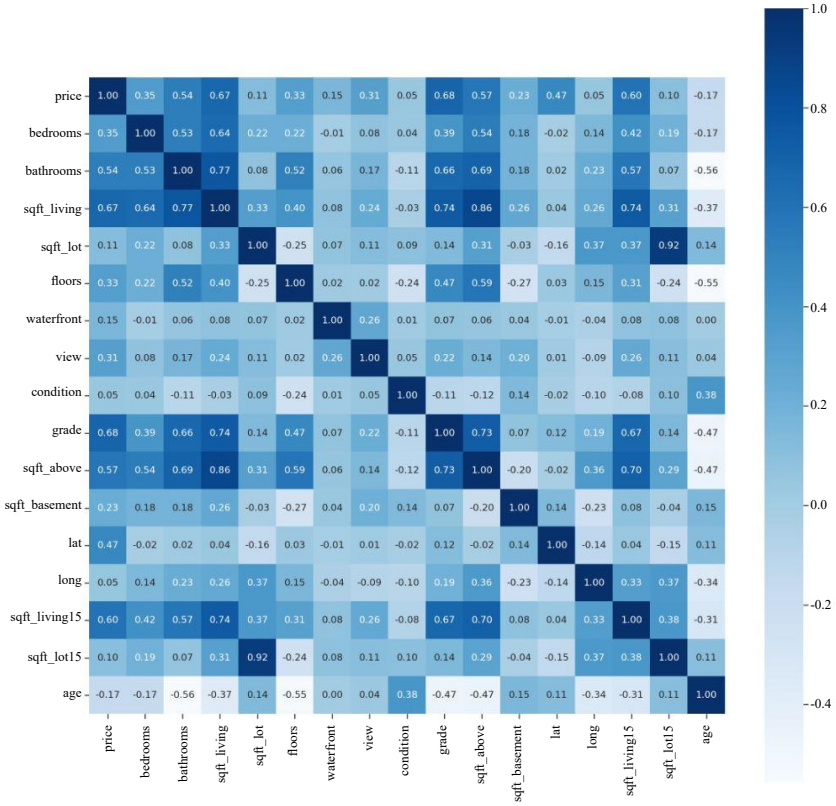

Figure 6. Correlation analysis between all the features in the King County data set

3.5 *Performance measures*

Root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), and R-squared (R^2) are the four factors that are considered in determining the accuracy of each of the models.

RMSE. The RMSE is used to compute expected performance by taking into account each data's prediction error.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Here, y_i is the actual value and $\hat{y}_i$ is the predicted value of the house price.

MSE. Most people are familiar with and regularly use MSE, which is frequently taught in machine learning courses. MSE can't contain a negative value as the sum is applied to a squared number. The MSE can be mathematically defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

MAE. The average or mean of the absolute error is termed MAE.

$$\text{MAE} = \frac{1}{n}\sum_{i-1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

R^2 Score. In order to analyze a model's performance, the R^2 score is calculated. As a result, it is used as a tool to determine how well a model is trained to produce accurate results.

$$R^2 = \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \tag{6}$$

Here, y_i is the actual value and $\hat{y}_i$ is the predicted value of the house price and $\bar{y}$ is the mean value of the house price.

4. Experimental setup and result analysis

The purpose of this paper is to analyze and predict house prices using various machine learning and ensemble learning techniques and compare the results to select the best-performing model. Each model's performance is evaluated based on four major factors, namely, RMSE, MSE, MAE, and R^2 scores.

4.1 *Ames housing data set*

After applying the machine learning techniques (linear regression, KNN, SVM) and ensemble learning techniques (XGB regressor, AdaBoost, gradient boost regressor, random forest, CatBoost) to the Ames data set, the ensemble learning technique is giving the best results than machine learning techniques. In ensemble learning techniques, CatBoost is giving the best result as shown in Table 1 and Figure 7.

Table 1. Performance of the Ames data set

Algorithm	RMSE	MSE	MAE	R^2
LR	0.070	0.005	0.039	0.795
KNN	0.094	0.009	0.070	0.635
SVM	0.082	0.007	0.063	0.720
XGB	0.048	0.002	0.032	0.902
ADA	0.068	0.005	0.052	0.806
GRAD	0.050	0.003	0.035	0.895
RF	0.054	0.003	0.037	0.880
CAT	0.048	0.002	0.032	0.910

Note: LR = linear regression, XGB = XGBoost, ADA = AdaBoost, GRAD = gradient boost, RF = random forest, CAT = CatBoost

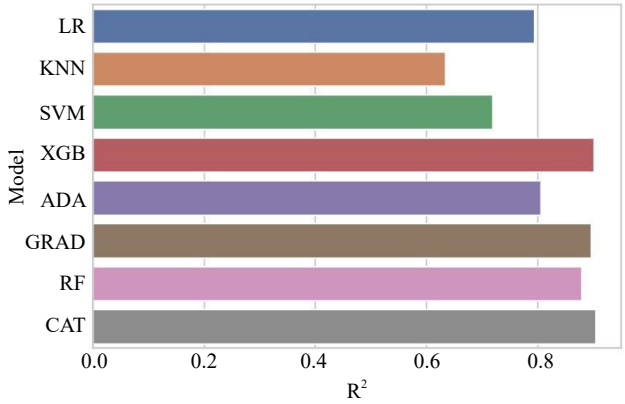

Figure 7. Result analysis for Ames data set

4.2 *King County data set*

After applying the machine learning techniques (linear regression, KNN, SVM) and ensemble learning techniques (XGB regressor, AdaBoost, gradient boost regressor, random forest, CatBoost) on the King County data set, the ensemble learning technique is giving the best results than machine learning techniques. In ensemble learning techniques, CatBoost is giving the best result as shown in Table 2 and Figure 8.

Table 2. Performance of the King County data set

Algorithm	RMSE	MSE	MAE	R^2
LR	0.5037	0.2537	0.3873	0.7463
KNN	0.4182	0.1749	0.3004	0.8251
SVM	0.3826	0.1464	0.2712	0.8536
XGB	0.3443	0.1186	0.2362	0.8814
ADA	0.4901	0.2402	0.375	0.7598
GRAD	0.3323	0.1105	0.2315	0.8895
RF	0.3499	0.1225	0.2414	0.8775
CAT	0.3237	0.1048	0.2221	0.8952

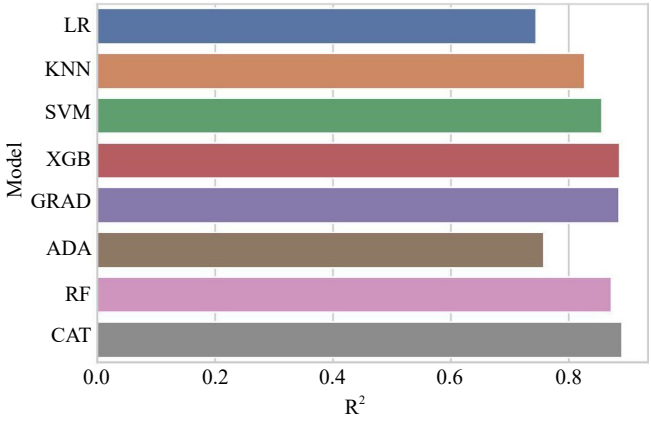

Figure 8. Result analysis for King County data set

5. Conclusion

From the results obtained from experimenting with different models. Using ensemble techniques like CatBoost and XGBoost has improved the accuracy of the model drastically when compared to simple regression techniques. These algorithms can be used in developing a web application where we could deploy these models for real-life analysis.

This process includes extracting relevant features from the data set. The merits of ensemble learning techniques are that it improves the accuracy compared to the traditional models by solving the problem of overfitting and being robust to outliers.

After applying the machine learning techniques (linear regression, KNN, SVM) and ensemble techniques (XGB regressor, AdaBoost, gradient boost regressor, random forest, CatBoost) ensemble learning techniques are giving the best result for both Ames and King County data sets. In ensemble learning techniques, CatBoost is giving the best result.

RMSE, MSE, MAE, and R^2 are the four factors that are considered in determining the accuracy of each of the models. The data set used here is a publicly available data set based on Ames housing. The same can be extended to predict house prices and rental prices across different locations.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Bork L, Møller SV. Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection. *International Journal of Forecasting*. 2015; 31(1): 63-78. https://doi.org/10.1016/j.ijforecast.2014.05.005

[2] Park B, Bae JK. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*. 2015; 42(6): 2928-2934. https://doi.org/10.1016/j.eswa.2014.11.040

[3] Plakandaras V, Gupta R, Gogas P, Papadimitriou T. Forecasting the US real house price index. *Economic Modelling*. 2015; 45: 259-267. https://doi.org/10.1016/j.econmod.2014.10.050

[4] She B, Shen H. An intelligent sensor ensemble learning method based on Bagging-SVM. *Sensor and Micro System*. 2016; 35(2): 26-28.

[5] Jui JJ, Imran Molla MM, Bari BS, Rashid M, Hasan MJ. Flat price prediction using linear and random forest regression based on machine learning techniques. In: Mohd Razman M, Mat Jizat J, Mat Yahya N, Myung H, Zainal Abidin A, Abdul Karim M. (eds.) *Embracing Industry 4.0*. Lecture Notes in Electrical Engineering, vol 678. Singapore: Springer; 2020. p.205-217. https://doi.org/10.1007/978-981-15-6025-5_19

[6] Sun S. Real estate price prediction based on data mining. *Modern Electronic Technique*. 2017; 40(5): 126-129.

[7] Yuan X, Zheng B, Jiao W. Commercial housing price prediction in Shanghai based on SVM. *Gansu Science Journal*. 2016; 28(1): 25-28.

[8] Cao B, Yang B. Research on ensemble learning-based housing price prediction model. *Big Geospatial Data and Data Science*. 2018; 1(1): 1-8. https://dx.doi.org/10.23977/bgdds.2018.11001

[9] Glaeser EL, Nathanson CG. An extrapolative model of house price dynamics. *Journal of Financial Economics*. 2017; 126(1): 147-170. https://doi.org/10.1016/j.jfineco.2017.06.012

[10] Rajan U, Seru A, Vig V. The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics*. 2015; 115(2): 237-260. https://doi.org/10.1016/j.jfineco.2014.09.012