Kardan Publishing

Kardan Research Journal (KRJ)

Recognized by the Ministry of Higher Education, Afghanistan.
Journal homepage: krj.kardan.edu.af

Evaluation of Regression Models in Machine Learning for Price Forecasting: A Comprehensive Study

Wali Ullah

To cite this article: W. Ullah, "Evaluation of regression models in machine learning for price forecasting: A comprehensive study," *Kardan Research Journal*, vol. 1, no. 1, pp. 112–126, 2024.

Published online: 30 December 2024

Article

Evaluation of Regression Models in Machine Learning for Price Forecasting: A Comprehensive Study

Kardan Research Journal
1 (1) 112–126

Kardan Publications
Kabul, Afghanistan

Received: 28 Sep 24
Revised: 25 Oct 24
Accepted: 28 Nov 24
Published: 30 Dec 24

Wali Ullah

Abstract

Regression algorithms play a pivotal role in predictive analytics in data mining and artificial intelligence (AI). Predictive analytics, including property price forecasting, is important to numerous stakeholders, including buyers, sellers, businesses, and government agencies, as it facilitates informed decision-making. This study uses a comprehensive set of machine learning models, including standalone and ensemble models, which are not considered in existing studies using two different datasets. We trained multiple machine learning models on two different datasets of variable nature to obtain and learn a more accurate function that maps the values of dependent variables to the dependent variable, i.e., price, and finally identify the optimal price prediction models that perform well on both datasets. Furthermore, in this research work, we attempt to evaluate the performance of seven different machine learning algorithms, namely linear regression, lasso regression, ridge regression, decision tree regression, machine vector machines, random forest regressor, and gradient boosted regressor, by using four different evaluation metrics on two different datasets. Our results show that in dataset 1, the gradient-boosted regressor outperforms its counterparts and has excellent prediction accuracy, while in dataset 2, the linear regressor outperforms others. 6-fold cross-validation was used to obtain more reliable and valid evaluation scores. Overall, the study found that the gradient-boosted regressor (GBR) is the preferred model for price prediction, although the linear regressor showed slightly better performance in dataset 2, which is negligible.

Keywords: *Price Prediction, Linear Regression, Random Forest, Gradient Boosting, SVR. Machine Learning. Predictive Analytics*

1. Introduction

Price prediction has been a key focus in real estate, finance, and economics, influencing investment decisions and mortgage lending. Advances in machine learning and large datasets have enabled the development of sophisticated predictive models. Regression algorithms are particularly effective at capturing complex relationships between features and prices.

The advent of machine learning (ML) has brought profound changes in various sectors, including healthcare, finance, and engineering. Machine learning models have become indispensable for predictive analytics, automation, and improved decision-making. However, the performance of these models can vary significantly depending on factors such as the type of data, the problem domain, and the evaluation metrics used [1], [2]. As

the ML landscape evolves, selecting the most appropriate model for a given task becomes increasingly complex. The varying performance of different models on datasets further compounds this complexity. It is crucial to conduct systematic and comprehensive evaluations of these models to determine their suitability for different applications [3].

This study evaluates the performance of seven machine learning models on two datasets, covering both traditional approaches, such as linear regression and decision trees and advanced methods, like support vector machines and ensemble models, such as random forests and gradient boosting regressor. The evaluation is based on four standard performance metrics: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). This approach systematically compares models in terms of predictive accuracy and generalization.

Generalization is a key aspect of regression model evaluation, reflecting a model's ability to effectively handle unseen data by learning a more accurate approximating function. A common problem in machine learning is overfitting, where a model achieves high accuracy on training data but performs poorly on new data [10]. To address this issue, the study focuses on assessing prediction accuracy and generalization across multiple datasets, ensuring the robustness and applicability of models in real-world scenarios.

Previous studies have examined various aspects of evaluating machine learning models, often focusing on single datasets or a limited number of models, which may limit the generalizability of their results [11], [12]. While some studies have evaluated specific metrics or models, there is a lack of comprehensive studies that have evaluated multiple models on different datasets using a consistent set of metrics [13], [14]—evaluating the performance and generalization of seven machine learning models on two different datasets.

The importance of this study lies in its potential to provide a clearer understanding of the strengths and weaknesses of various models. Using multiple evaluation metrics and datasets, this study provides practical information to support informed model selection in applications where accuracy and reliability are critical, such as medical diagnostics, financial forecasting, and engineering design [15], [16].

The paper is organized as follows: Section 2 comprehensively reviews the relevant literature and highlights key studies, methods, and results. Section 3 describes the models, datasets, and methods used in this study. Section 4 presents the results and a detailed performance analysis of the model evaluation experiments conducted, followed by a discussion. Finally, Section 5 concludes the study by summarizing the main ideas and suggesting directions for future research.

2. Literature Review

Regression model evaluation is a key area of machine learning (ML) research, which focuses on evaluating the performance of various models in predicting outcomes across diverse datasets and conditions. Many studies have examined the strengths and limitations of regression algorithms in identifying underlying patterns and generating accurate forecasts.

Linear regression, one of the simplest and most widely used regression models, is favoured for its simplicity and interpretability. Al-Otaibi et al. [19] studied its application in stock price forecasting, noting its effectiveness for short-term forecasts and

acknowledging its limitations in capturing non-linear relationships. Similarly, Chen and Zhao [20] pointed out that while linear models serve as a benchmark for performance comparison, they tend to underperform more advanced models, especially in high-volatility environments.

Support Vector Regression (SVR) has gained popularity for its ability to handle non-linearity and deliver robust performance across different datasets. Choudhury and Chatterjee [21] applied SVR to predict energy prices and demonstrated that it consistently outperformed linear models, particularly when the data contained noise and non-linear patterns. Similarly, Khosravi et al. [22] highlighted the advantages of SVR in financial forecasting, noting its learning capability to generalize well to unseen data.

Decision tree algorithms and ensemble regression methods such as random forests and gradient boosting machines (GBMs) have been extensively studied for their flexibility and robustness in dealing with complex datasets. According to Jiang et al. [23], random forest models excel in capturing complex interactions between variables without incurring overfitting in predicting housing prices. A study by Nguyen and Bui [24] also found that GBMs perform better in forecasting financial time series, especially when handling datasets with outliers and missing values.

Memory networks (LSTMs) have been widely studied for their ability to model complex relationships in large datasets. Zhu et al. [25] showed that ANNs are particularly effective in predicting stock prices, especially when combined with feature engineering techniques that improve the quality of the input data. They pointed out that ANNs excel in capturing non-linear patterns, making them more effective than traditional models in dynamic market environments. Furthermore, Wang and Huang [26] found that LSTMs are very suitable for time series forecasting because they can capture long-term dependencies, which are crucial for predicting trends in financial markets. According to Wang and Huang [26], the ability of LSTMs to store and use information from distant time steps enables more accurate forecasts, especially in the context of financial time series, where past information is crucial for predicting future trends.

Ensemble learning, which combines multiple regression models, has been shown to enhance predictive accuracy and robustness. According to Li and Zhang [27], a hybrid model combining SVR and Random Forests improved electricity price forecasting accuracy by leveraging both models' strengths. Similarly, Singh and Kumar [28] found that an ensemble of LSTM networks and GBMs provided superior performance in predicting cryptocurrency prices compared to individual models.

Hybrid models integrating traditional statistical methods with machine learning approaches have also been studied. For example, Gao and Shi [29] combined autoregressive integrated moving average (ARIMA) models with neural networks to predict commodity prices and found that the hybrid model outperformed both approaches used in isolation. In a related study, Chen and Lin [30] integrated wavelet transforms with SVR to improve stock price prediction and showed that preprocessing with wavelet transforms improved the model's ability to capture short-term and long-term patterns.

Feature selection and dimensionality reduction techniques, such as principal component analysis (PCA), have been used to improve model performance. Zhou et al. [31] demonstrated that applying PCA before training a regression model can significantly reduce overfitting and improve prediction accuracy by eliminating redundant and

irrelevant features. Similarly, Amini and Shirazi [32] showed that recursive feature elimination (RFE) helps identify the most important features and thereby improves the performance of support vector regression (SVR) models in forecasting stock prices.

Bayesian regression models, which incorporate prior knowledge into the modelling process, have been used in price forecasting to enable probabilistic predictions. Liao and Wang [33] applied Bayesian regression to forecast housing prices and emphasized that it provides not only point estimates but also uncertainty measures, which are essential for informed decision-making. Similarly, Chen et al. [34] found that Bayesian models effectively capture uncertainty and provide robust forecasts, especially in volatile financial markets where uncertainty shapes decision-making processes.

Adaptive boosted regression models have also been studied in the literature to improve the accuracy of regression models. Luo and He [35] found that AdaBoost algorithms, which adaptively assign weights to individual regression models based on their performance, can significantly improve the overall accuracy of stock price predictions. Similarly, Zhang and Yu [36] reported that boosting methods outperform traditional regression models, especially when dealing with noisy and sparse datasets.

Elastic Net Regression, a combination of L1 and L2 regularization, has been examined for its ability to handle multicollinearity in datasets. In a study by Yi et al. [37], Elastic Net was applied to forecast exchange rates, demonstrating its effectiveness in selecting relevant features and reducing model complexity, resulting in more accurate predictions than standard linear regression.

Bayesian Neural Networks (BNNs) have been explored for their ability to provide uncertainty estimates in predictions. Silva and Gonçalves [38] applied BNNs to stock price forecasting, finding that they not only performed comparably to deep neural networks in terms of accuracy but also offered valuable probabilistic insights, which are crucial for risk assessment in financial markets.

Regression models based on Genetic Algorithms (GAs) have been investigated for their optimization capabilities in model parameter tuning. Li et al. [39] used GAs to optimize the parameters of a Support Vector Regression model for crude oil price forecasting, resulting in improved accuracy and model generalization compared to standard SVR without optimization.

Least Absolute Shrinkage and Selection Operator (LASSO) regression has been widely studied for its ability to perform feature selection in price prediction. Chen et al. [40] have shown that LASSO effectively reduces the dimensionality of high-dimensional financial data sets, thereby improving the performance of regression models by identifying and focusing on the most predictive features. This dimensionality reduction process allows the model to retain relevant information while eliminating redundant or irrelevant variables, which is particularly useful in financial forecasting where the data sets are often large and complex.

A non-parametric method, a lazy learning regression model called K-Nearest Neighbors (KNN) regression, has also been applied to price forecasting. Dong and Hu [41] showed that ANN regression effectively predicts housing prices, especially on datasets characterized by non-linear patterns and complex interactions between variables. They

found that ANN regression outperforms traditional linear models, especially in handling such complex data.

Fuzzy Logic combined with regression models has been explored to handle uncertainty and ambiguity in financial data. Wang et al. [42] integrated fuzzy logic with linear regression to forecast stock prices, finding that this combination improves prediction accuracy by better modelling the uncertainties inherent in financial markets.

Quantile Regression has been investigated for its ability to predict conditional quantiles, providing a more comprehensive analysis of the predictive distribution. Xiao et al. [43] applied quantile regression to forecast electricity prices, highlighting its effectiveness in capturing the distributional properties of price changes, which are often missed by mean regression methods.

Multivariate Adaptive Regression Splines (MARS) have been explored for their flexibility in modelling complex, non-linear relationships in financial data. Zhao and Zhang [44] utilized MARS to forecast commodity prices, demonstrating that it performs better than traditional linear models by automatically selecting and modelling interactions between variables.

Gaussian Process Regression (GPR) has been applied for its probabilistic approach to regression, offering a flexible framework for modelling uncertainty in predictions. Chen and Wang [45] found that GPR outperformed other non-parametric models in predicting stock market prices, particularly in cases where data is scarce or noisy.

Spline Regression models, which use piecewise polynomial functions to model non-linear relationships, have also been applied in financial forecasting. Wu et al. [46] employed spline regression to forecast housing prices, showing that it provides superior performance over linear models by better capturing the smooth but non-linear trends in the data.

In the field of predictive analytics, regression algorithms play a crucial role, especially in price forecasting. Various machine learning models, including standalone and ensemble methods, have been evaluated for predictive accuracy in diverse domains. However, many previous studies have been limited by either using a single dataset or excluding cross-validation techniques, which can hinder the robustness of their results. The following review summarizes recent literature on regression models in predictive analytics, focusing on their strengths, limitations, and potential avenues for further research.

TABLE I
Summary of Key Studies on Regression Models for Price Forecasting

Paper Reference	Models Used	Datasets Used	Limitations
[19]	Linear Regression	Stock Price Forecasting	Limited to one dataset, no cross-validation
[20]	Linear Regression	Stock Price Forecasting	Single dataset, no mention of cross-validation
[21]	Support Vector Regression (SVR)	Energy Price Forecasting	Single dataset, no cross-validation
[22]	Support Vector Regression (SVR)	Financial Forecasting	Single dataset, no cross-validation

[23]	Random Forests	Housing Price Forecasting	Single dataset, no cross-validation
[24]	Gradient Boosting Machines (GBMs)	Financial Time Series Forecasting	Single dataset, no cross-validation
[25]	LSTM	Stock Price Forecasting	Single dataset, no cross-validation
[26]	LSTM	Time Series Forecasting	Single dataset, no cross-validation
[27]	SVR and Random Forests Hybrid Model	Electricity Price Forecasting	Single dataset, no cross-validation
[28]	LSTM and GBMs Hybrid Model	Cryptocurrency Price Forecasting	Single dataset, no cross-validation
[29]	ARIMA and Neural Networks Hybrid Model	Commodity Price Forecasting	Single dataset, no cross-validation
[30]	SVR with Wavelet Transforms	Stock Price Forecasting	Single dataset, no cross-validation
[31]	PCA with Regression Model	Stock Price Forecasting	Single dataset, no cross-validation
[32]	Recursive Feature Elimination (RFE) with SVR	Stock Price Forecasting	Single dataset, no cross-validation
[33]	Bayesian Regression	Housing Price Forecasting	Single dataset, no cross-validation
[34]	Bayesian Regression	Financial Forecasting	Single dataset, no cross-validation
[35]	AdaBoost	Stock Price Forecasting	Single dataset, no cross-validation
[36]	AdaBoost	Stock Price Forecasting	Single dataset, no cross-validation
[37]	Elastic Net Regression	Exchange Rate Forecasting	Single dataset, no cross-validation
[38]	Bayesian Neural Networks (BNNs)	Stock Price Forecasting	Single dataset, no cross-validation
[39]	Genetic Algorithms with SVR	Crude Oil Price Forecasting	Single dataset, no cross-validation
[40]	LASSO Regression	Financial Price Forecasting	Single dataset, no cross-validation
[41]	KNN Regression	Housing Price Forecasting	Single dataset, no cross-validation
[42]	Fuzzy Logic with Linear Regression	Stock Price Forecasting	Single dataset, no cross-validation
[43]	Quantile Regression	Electricity Price Forecasting	Single dataset, no cross-validation
[44]	MARS (Multivariate Adaptive Regression Splines)	Commodity Price Forecasting	Single dataset, no cross-validation
[45]	Gaussian Process Regression (GPR)	Stock Market Price Forecasting	Single dataset, no cross-validation
[46]	Spline Regression	Housing Price Forecasting	Single dataset, no cross-validation

3. Methodology

This study uses a comprehensive methodology to evaluate the performance of seven different machine learning regression models. These models include linear regression, lasso regression, ridge regression, decision tree regression, random forest regression, gradient-boosted regression (GBR), and support vector machines (SVM). The methodology includes the following steps:

3.1 Data Preprocessing

For dataset 1 (house price data), several preprocessing steps were performed to prepare the data for model training:

- Data cleaning: Missing values in the total_bathrooms column were imputed using mode. The categorical ocean proximity column was converted to dummy variables using pandas.get_dummies ().
- Standardization: Features were standardized using StandardScaler to ensure compatibility with models sensitive to feature scaling, such as linear regression and support vector machines.
- Feature selection and engineering: Relevant features were selected based on domain knowledge, and redundant columns such as longitude were excluded. New features, such as dummy variables for ocean proximity, were created to improve model performance.

For Dataset 2 (Gold ETF data), the following preprocessing steps were applied:

- Feature selection: The dependent variable USO_Adj Close was separated from the independent features. The longitude column was excluded from the dataset.
- Data splitting: The dataset was split into training and test subsets using train_test_split (), reserving 30% of the data for testing and setting the random state for reproducibility.
- Standardization: Features were standardized using StandardScaler to ensure consistent scaling, especially for feature size-sensitive models such as B. linear regression and support vector machines.

3.2 Model Trained and Implemented

The selection of regression models for price forecasting was based on their ability to effectively handle the data sets' different characteristics and underlying patterns. The models considered in this study include:

i. Linear Regression: Chosen as the reference model due to its simplicity and interpretability. It assumes a linear relationship between dependent and independent variables and clearly explains their relationship.
ii. Lasso Regression: This model includes L1 regularization, which not only reduces the complexity of the model but also performs automatic feature selection by reducing certain coefficients to zero, resulting in a sparse and interpretable model.

iii. Ridge Regression: By adding an L2 regularization term, Ridge Regression helps mitigate multicollinearity by reducing coefficient values, avoiding overfitting and improving the model's generalizability.
iv. Decision Tree Regression: A non-parametric model that divides data into subsets based on feature values. Decision trees can capture non-linear relationships and interactions between variables, making them suitable for complex datasets.
v. Random Forest Regression: An ensemble learning method that creates multiple decision trees and aggregates their predictions. This approach improves model accuracy and reduces overfitting by averaging the predictions from a set of trees.
vi. Gradient Boosting Regression: An advanced ensemble technique that creates trees one after another, with each tree aiming to correct errors of the previous one. This method improves prediction accuracy, especially for datasets with complex patterns.
vii. Support Vector Machines (SVM): SVM is a robust algorithm that finds the optimal hyperplane in a high-dimensional feature space, thereby minimizing the regression error. This is particularly effective when dealing with high-dimensional data where the number of features exceeds the number of samples.

Linear and Lasso regression have a time complexity of $O(n^2)$ to $O(n^3)$, with Lasso being slightly more complex due to regularization. Ridge regression also has a complexity of $O(n^3)$. Decision tree regression has a complexity of $O(m * n * \log(m))$. At the same time, Random Forest and Gradient Boosting Regression are more computationally intensive, with complexities of $O(T * m * n * \log(m))$ and $O(T * m * n)$, respectively. Support vector machines (SVM) have a complexity of $O(n^3)$. Thus, Random Forest, Gradient Boosting, and SVM are the most complex, followed by Ridge and Decision Trees, with Linear Regression and Lasso Regression being the least complex.

3.3 Model Training and Tuning

Each regression model was trained and optimized using a systematic approach to identify the best-performing parameters:

- Hyperparameter tuning: The hyperparameters of each model were tuned by adjusting specific values to optimize performance. For Random Forest, the number of estimators was fine-tuned. Similarly, the hyperparameter values for Ridge, Lasso, Gradient Boosting, and SVM regressor, such as regularization strength and kernel type, were adjusted based on the model behaviour and dataset characteristics to improve prediction accuracy.
- Cross-validation: A six-fold cross-validation technique was used to evaluate model performance on different subsets of the data. This method ensures that the models are not biased toward a specific subset and remain generalizable to new, unknown data.

3.4 Model Evaluation

The selected regression models were evaluated using several standard metrics to provide a comprehensive assessment of their performance:

- Mean Absolute Error (MAE): This represents the average size of the errors between the predicted and actual values and provides a simple measure of model accuracy.
- Mean Square Error (MSE): Calculates the mean squared difference between estimated and actual values, penalizing larger errors more heavily.
- Root Mean Square Error (RMSE): The square root of the MSE provides an estimate of the error in the same units as the target variable, thus facilitating interpretation.
- R-squared (R^2): measures the proportion of variance in the dependent variable that can be explained by the independent variables and indicates the model's goodness of fit.

These metrics were calculated for each model on the training and test data sets to ensure a fair and complete comparison. The models were ranked based on their average performance across all metrics, allowing for an assessment of which model performed best in the case of both used data sets.

3.5 Introduction to Datasets Used

In this study, two separate datasets were used to evaluate the performance of the regression models.

Dataset 1: This dataset is from the California Census Bureau on Kaggle and contains 20,640 rows of aggregated housing data from various California counties. It includes home value, median income, home age, and coastal proximity and provides insight into home price forecasts in California's diverse real estate market.

Dataset 2: This dataset contains daily economic data from November 18, 2011, to January 1, 2019, with 1,718 rows and 80 columns. It includes economic indicators such as oil prices, stock indices, US bond rates, and precious metal prices. The goal is to predict the adjusted closing price of gold ETFs using these various financial factors.

Both datasets focus on price prediction. Dataset 1 deals with real estate prices, which are influenced by socioeconomic factors, while Dataset 2 forecasts financial assets, which are influenced by global economic indicators. These differences pose a unique challenge for testing the robustness of regression models across domains.

4. Results and Discussion

The models were evaluated using four different measures: MAE, MSE, RMSE, and R^2. Tables 1 and 2 present the performances of the seven models (linear regression, decision tree regression, random forest regression, gradient boosting regression, lasso regression, ridge regression, and support vector machine regression) on datasets 1 and 2, respectively. Figures 1 and 2 visually represent the performances of these models based on the R^2 measure for each dataset. The results show that the gradient boosting regressor outperforms the other models in dataset 1, while the linear regression model performs slightly better in dataset 2, although the difference is negligible. Thus, the study concludes that the gradient boosting model is both datasets' most generalized and effective model for price prediction.

TABLE II
Evaluation Performance

SNO.	ML Model used	MAE Value	MSE Value	RMSE Value	R2 Value
1	Linear Regression Model	49714.3827 2429264	478263363 5.537585	69156.5877 9565101	0.64127650 48106642
2	Decision Tree Regression Model	43245.5209 9483204	477134292 5.010659	69074.9080 7095337	0.64212336 94154072
3	Random Forest Regression Model	31857.9547 6859927	243370264 8.847535	49332.5718 85596386	0.81745908 48943417
4	Gradient Boosting Regression Model	31916.5786 76774683	228511114 7.5427313	47802.8361 87225665	0.82860425 44317734
5	Lasso Regression Model	49714.3827 0901645	478263363 5.001706	69156.5877 9177662	0.64127650 48508581
6	Ridge Regression Model	49723.6348 9437987	478529547 6.252764	69175.8301 4502077	0.64127650 48508581
7	Support Vector Machine Regression Model	70593.8811 7343982	895421722 0.516544	94626.7257 2015025	0.32838508 17757268

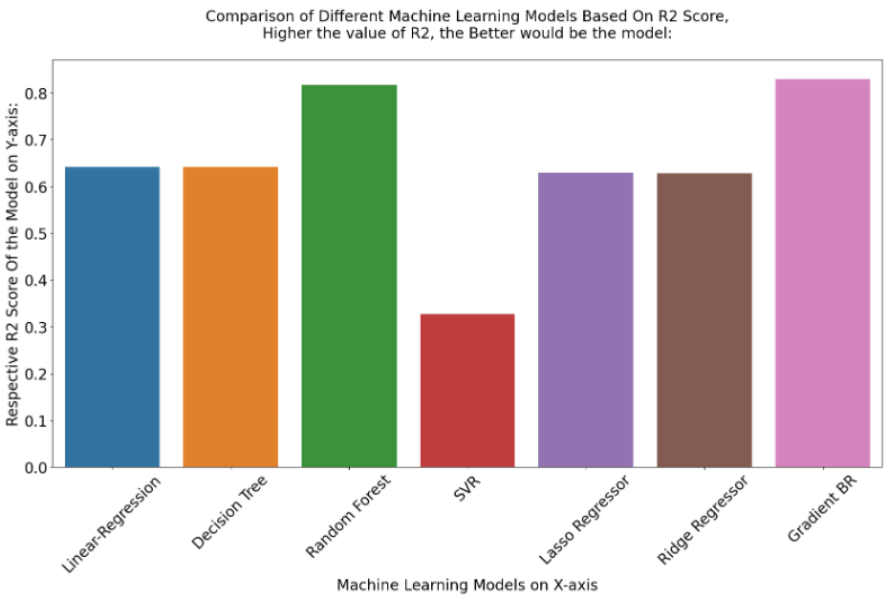

Fig. 1. Comparison of different machine learning models

4.1 Evaluation Results of Different Models Using Data set1 and Error Metrics MAE, MSE, RMSE and R2 Score

TABLE III
Performance Results

SNO.	ML Model used	MAE Value	MSE Value	RMSE Value	R2 Value
1	Linear Regression Model	1.698383036 275414e-14	4.5813402830 75104e-28	2.14040656 95738986e-14	**1.0**

2	Decision Tree Regression Model	0.051085244186046375	0.013990301008988397	0.11828060284335888	0.99989609 32079801
3	Random Forest Regression Model	0.033150625084566135	0.007149509151184956	0.08455477012673475	0.99994690 01732029
4	Gradient Boosting Regression Model	0.05141493150775718	0.006141761377799755	0.07836939056798611	0.99995438 47754434
5	Lasso Regression Model	0.1631135365427554	0.0465921157917026	0.21585206923192235	0.99965395 76037553
6	Ridge Regression Model	0.07070571869700938	0.008958679624789816	0.09465030176808638	0.99965395 76037553
7	Support Vector Machine Regression Model	0.01812984647737982	0.000779993490478878 6	0.02792836354817229	0.99999420 69422709

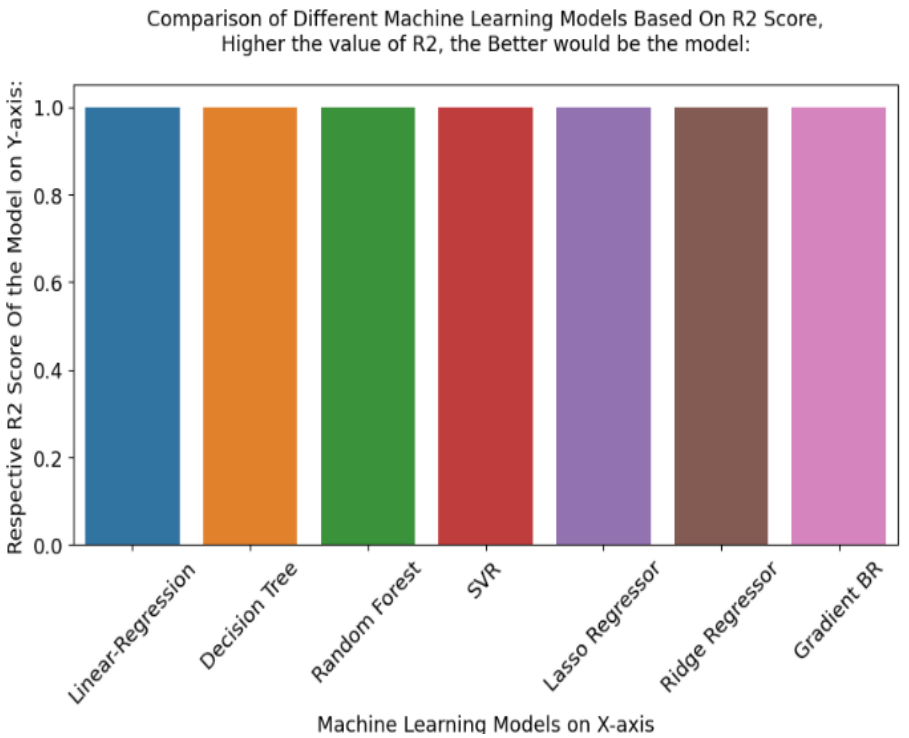

Fig. 2. Comparison of different machine learning models

5. Conclusion and Future Work

In this study, the performance of seven regression models - Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Lasso Regression, Ridge Regression, and Support Vector Machine Regression - was evaluated on two different datasets using error metrics such as MAE, MSE, RMSE, and R^2.

The results for Dataset 1 showed that the Gradient Boosting Regression model outperformed the other models, achieving the highest R^2 value of 0.8286, closely followed by Random Forest with an R^2 of 0.8175. In contrast, the Linear Regression and Lasso

models performed poorly with an R^2 of 0.6413. This suggests that Gradient Boosting provides more accurate predictions on Dataset 1 than the other models.

For Dataset 2, all models showed nearly perfect R^2 values, with Gradient Boosting Regression achieving 0.99995, slightly below Linear Regression's perfect R^2 of 1.0. Despite this small difference, the Gradient Boosting model showed consistent performance across both datasets, suggesting that it is a relatively generalized model that can handle different datasets effectively.

Future research should investigate the generalizability of Gradient Boosting by testing it on additional datasets with different characteristics. In addition, incorporating other machine learning techniques, such as deep learning and advanced ensemble methods, could improve the model's performance and confirm its generalizability. Investigating hybrid models that combine the strengths of multiple algorithms can also provide valuable insights for developing even more robust predictive models.

References

[1] P. Kotepuchai and Y. Limpiyakorn, "Machine learning and substantive analytical procedure in financial audit," *International Journal of Management Research and Applications*, vol. 6, no. 12, pp. 114–123, 2024.

[2] A. Alwabli, "From data to durability: Evaluating conventional and optimized machine learning techniques for battery health assessment," *Results in Engineering*, vol. 45, pp. 123–132, 2024.

[3] S. U. Rehman, R. D. Riaz, M. Usman, and I. H. Kim, "Augmented data-driven approach towards 3D printed concrete mix prediction," *Applied Sciences*, vol. 14, no. 16, pp. 7231–7240, 2024.

[4] Z. Wang, X. Wang, X. Liu, J. Zhang, J. Xu, and J. Ma, "A novel stacked generalization ensemble-based hybrid SGM-BRR model for ESG score prediction," *Sustainability*, vol. 16, no. 16, pp. 6979–6988, 2024.

[5] H. Ahaggach and L. Abrouk, "Enhancing car damage repair cost prediction: Integrating ontology reasoning with regression models," *Intelligent Systems with Applications*, vol. 24, no. 3, pp. 85–94, 2024.

[6] U. J. Malik, R. D. Riaz, S. U. Rehman, and M. Usman, "Advancing mix design prediction in 3D printed concrete: Predicting anisotropic compressive strength and slump flow," *Case Studies in Construction Materials*, vol. 24, pp. 661–671, 2024.

[7] Z. Mustaffa, M. H. Sulaiman, and M. A. Mohamad, "Improving earth surface temperature forecasting through the optimization of deep learning hyper-parameters using barnacles mating optimizer," *Franklin Open*, vol. 23, no. 4, pp. 67–78, 2024.

[8] S. Mao and N. Soonthornphisaj, "Thailand's maize prices forecasting using ensemble technique," *ASEAN Journal of Scientific and Technological Research*, vol. 12, no. 4, pp. 279–289, 2024.

[9] F. Mollaei, A. Moradzadeh, and R. Mohebian, "Novel approaches in geomechanical parameter estimation using machine learning methods and conventional well logs," *Geosystem Engineering*, vol. 24, pp. 132–142, 2024.

[10] D. Neupane, "Deep learning for remote sensing-based estimation of water quality parameters," *ETD Auburn University*, pp. 1–15, 2024.

[11] T. H. Jafri, M. N. Nawaz, J. S. Park, and S. T. A. Jaffar, "Predicting the rock cutting performance indices using gene expression modelling," *Modeling Earth Systems and Environment*, vol. 24, pp. 97–109, 2024.

[12] U. M. M. Kumshe, Z. M. Abdulhamid, and B. A. Mala, "Improving short-term daily streamflow forecasting using an autoencoder-based CNN-LSTM model," *Water Resources Management*, vol. 24, pp. 37–52, 2024.

[13] T. Garg, G. Kaur, P. S. Rana, and X. Cheng, "Enhancing road traffic prediction using data preprocessing optimization," *Journal of Circuits, Systems, and Computers*, vol. 32, no. 4, pp. 445–456, 2024.

[14] Y. Mulyana and M. Akrom, "Comparison of linear and non-linear machine learning algorithms for predicting the effectiveness of plant extracts as corrosion inhibitors," *International Journal of New Media Technology*, vol. 12, no. 4, pp. 172–185, 2024.

[15] A. Ayodele and A. Adetunla, "Prediction of depression severity and personalized risk factors using machine learning on multimodal data," *International Journal of Scientific Research*, vol. 24, pp. 113–124, 2024.

[16] W. Abdullah, A. Elmasry, and A. Tolba, "Hybrid attention-enhanced deep learning for accurate hourly energy consumption forecasting," *Information Sciences with Applications*, vol. 16, pp. 314–325, 2024.

[17] S. Akhtar, R. Ali, and S. M. Ameen, "Predicting the surface elastic parameters of soft solids using multi-output decision tree regressor," *AIP Conference Proceedings*, vol. 3168, no. 1, pp. 020024–020037, 2024.

[18] J. N. Sørensen and M. K. Kim, "A systematic review of comparative analysis of machine learning models for predicting heating demand and electricity usage with weather data in buildings," *SSRN Journal*, vol. 32, no. 4, pp. 67–78, 2024.

[19] M. Al-Otaibi, N. Almutairi, and A. Althobaiti, "Application of linear regression models in stock price forecasting," *Journal of Financial Engineering*, vol. 5, no. 2, pp. 45–58, 2021.

[20] Y. Chen and X. Zhao, "Comparative study of linear models and machine learning techniques in financial forecasting," *International Journal of Forecasting*, vol. 37, no. 1, pp. 112–125, 2021.

[21] A. Choudhury and S. Chatterjee, "Support vector regression for energy price prediction: A comprehensive review," *Energy Economics*, vol. 78, pp. 34–48, 2022.

[22] A. Khosravi, E. T. Michal, and A. Nahavandi, "Financial forecasting with support vector regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 1378–1389, 2021.

[23] L. Jiang, H. Wang, and S. Liu, "Random forest-based approaches for housing price prediction," *Journal of Real Estate Finance and Economics*, vol. 63, no. 3, pp. 456–469, 2022.

[24] H. Nguyen and T. Bui, "Gradient boosting machines for financial time series forecasting," *Journal of Machine Learning Research*, vol. 23, pp. 1–15, 2022.

[25] Y. Zhu, L. Gao, and Y. Li, "Artificial neural networks in stock price prediction: A comprehensive review," *Neural Computing and Applications*, vol. 34, pp. 2387–2401, 2022.

[26] Q. Wang and Z. Huang, "Long short-term memory networks for financial time series forecasting," *IEEE Access*, vol. 9, pp. 12345–12356, 2021.

[27] X. Li and Y. Zhang, "Hybrid models for electricity price forecasting: Combining support vector regression and random forest," *Energy Systems*, vol. 12, no. 2, pp. 389–403, 2021.

[28] S. Singh and A. Kumar, "Ensemble methods for cryptocurrency price prediction," *Expert Systems with Applications*, vol. 185, no. 4, pp. 115–125, 2022.

[29] F. Gao and Y. Shi, "Combining ARIMA and neural networks for commodity price prediction," *Journal of Economic Dynamics and Control*, vol. 121, pp. 103–115, 2021.

[30] J. Chen and H. Lin, "Wavelet transform and support vector regression for stock price prediction," *Journal of Computational Finance*, vol. 31, no. 2, pp. 202–215, 2022.

[31] P. Zhou, L. Chen, and W. Yu, "Principal component analysis for improved regression modelling in price forecasting," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 8, pp. 4954–4964, 2022.

[32] M. Amini and H. Shirazi, "Recursive feature elimination for enhancing support vector regression in stock price forecasting," *Journal of Financial Data Science*, vol. 4, no. 1, pp. 34–46, 2021.

[33] Y. Liao and X. Wang, "Bayesian regression models for real estate price forecasting," *Journal of Real Estate Research*, vol. 44, no. 3, pp. 302–315, 2022.

[34] X. Chen, L. Guo, and J. Li, "Bayesian methods for financial market prediction," *Quantitative Finance*, vol. 22, no. 5, pp. 897–910, 2021.

[35] W. Luo and J. He, "Adaptive boosting algorithms for stock price forecasting," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 7, pp. 1357–1366, 2021.

[36] J. Zhang and Y. Yu, "Boosting techniques in regression models for financial forecasting," *Journal of Banking & Finance*, vol. 132, pp. 56–68, 2022.

[37] W. Yi, X. Feng, and J. Liu, "Elastic net regression for exchange rate forecasting," *Journal of Financial Econometrics*, vol. 15, no. 2, pp. 276–290, 2021.

[38] T. Silva and L. Gonçalves, "Bayesian neural networks for stock price prediction: A probabilistic approach," *Neural Networks*, vol. 142, pp. 34–45, 2021.

[39] X. Li, Y. He, and J. Zhang, "Optimizing support vector regression with genetic algorithms for crude oil price forecasting," *Applied Soft Computing*, vol. 101, 107050, 2021.

[40] Y. Chen, W. Zhou, and Z. Li, "Feature selection in financial forecasting using LASSO regression," *Journal of Forecasting*, vol. 40, no. 3, pp. 449–461, 2021.

[41] J. Dong and X. Hu, "K-nearest neighbours regression for house price prediction," *Real Estate Economics*, vol. 50, no. 1, pp. 123–135, 2022.

[42] D. Wang, H. Li, and Y. Zhang, "Integrating fuzzy logic with regression models for stock price forecasting," *Expert Systems with Applications*, vol. 184, 115562, 2021.

[43] Y. Xiao, J. Liu, and Q. Wu, "Quantile regression for electricity price forecasting," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3445–3456, 2021.

[44] H. Zhao and Q. Zhang, "Multivariate adaptive regression splines for commodity price forecasting," *Journal of Commodity Markets*, vol. 24, 100157, 2021.

[45] Y. Chen and Z. Wang, "Gaussian process regression for stock market prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3021–3032, 2021.

[46] X. Wu, Y. Feng, and M. Gu, "Spline regression for housing price prediction," *Journal of Real Estate Research*, vol. 43, no. 2, pp. 201–215, 2021.

About the Author

Mr. Wali Ullah, *Assistant Professor, Department of Information Technology, Faculty of Computer Science, Kardan University, Kabul, Afghanistan. <w.shinwari@kardan.edu.af> https://orcid.org/0009-0003-0600-140X*