MERGING LANES: WHERE E-LEARNING DIVERSITY MEETS FUTURE TRENDS

VOLUME 11, 2026
e-ISBN : 978-629-98755-9-8

eISBN 978-629-98755-9-8
9 786299 875598

SIG CS@e-Learning
Unit Penerbitan
Jabatan Sains Komputer & Matematik
Kolej Pengajian Pengkomputeran, Informatik & Matematik
Universiti Teknologi MARA Cawangan Pulau Pinang

MERGING LANES: WHERE E-LEARNING DIVERSITY MEETS FUTURE TRENDS

Advisor
Dr. Nor Hanim Abd Rahman,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Chief Editor
Ts. Jamal Othman,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Editors
Ts. Dr. Syarifah Adilah Mohamed Yusoff,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Dr Arifah Fasha Rosmani,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Mohd Saifulnizam Abu Bakar,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Published by:
Unit Penerbitan Jabatan Sains Komputer & Matematik (JSKM)
Bahagian Hal Ehwal Akademik (BHEA)
Universiti Teknologi MARA
Cawangan Pulau Pinang
13500 Permatang Pauh
Pulau Pinang
Malaysia

e ISBN : 978-629-98755-9-8

e-ISBN : 978-629-98755-9-8

SIG : e-Learning@CS
https://appspenang.uitm.edu.my/sigcs/
Publication Date : 15 – Apr - 2026

APPLICATIONS AND ADVANTAGES OF BOOSTED REGRESSION TREES IN STATISTICAL MODELING

*Wan Nur Shaziayani Wan Mohd Rosly[1], Norshuhada Samsudin[2], Sharifah Sarimah Syed Abdullah[3] and Fuziatul Norsyiha Ahmad Shukri[4]
**shaziayani@uitm.edu.my[1], norsh111@uitm.edu.my[3]*
sh.sarimah@uitm.edu.my[2], fuziatul@uitm.edu.my[4]

[1,2,3,4]Jabatan Sains Komputer & Matematik (JSKM),
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

**Corresponding author*

ABSTRACT

Boosted Regression Trees (BRT) have become an important technique in statistical modelling due to their strong predictive performance and flexibility in handling complex data structures. This paper presents an overview of the concept, methodology, advantages, and applications of BRT in modern data-driven analysis. The BRT approach combines regression trees with boosting algorithms, where multiple simple models are sequentially developed to improve prediction accuracy by minimizing errors from previous iterations. One of the key strengths of BRT is its ability to capture nonlinear relationships and interactions among variables without requiring strict assumptions about data distribution. In addition, BRT is robust to outliers, capable of handling missing values, and suitable for analysing different types of data, including continuous and categorical variables. The paper also reviews various applications of BRT across multiple domains, such as environmental modelling, epidemiology, and economic analysis. Overall, BRT provides a powerful and efficient framework for both predictive and explanatory modelling, making it a valuable tool for researchers and practitioners in statistical and machine learning fields.

Keywords: Boosted Regression Tress, Statistical Modeling, Machine Learning Methods, Multiple Linear Regression, Predictive Modeling

Introduction

In recent years, machine learning techniques have become essential tools for statistical modelling and data-driven decision making. Among these, ensemble methods have proven to be particularly effective in improving prediction performance by combining multiple models. Boosted Regression Trees (BRT), also known as Gradient Boosting Machines (GBM), represent one such approach that merges the strengths of regression trees and boosting algorithms. The BRT technique builds multiple small regression trees sequentially, with each tree correcting the errors of the previous ones, ultimately resulting in a strong and accurate predictive model.

BRT has several advantages that make it superior to many traditional regression and classification techniques. Firstly, it is highly flexible and capable of capturing nonlinear relationships and complex interactions among predictor variables without requiring prior specification. Secondly, BRT is robust to outliers and missing data, which makes it suitable for real-world datasets that are often noisy or incomplete. Thirdly, it can handle mixed data types, such as continuous, categorical, and ordinal variables, within the same model framework. Another major advantage is its ability to provide variable importance measures and partial dependence plots, allowing researchers to interpret the influence of each predictor on the response variable. Additionally, by tuning parameters such as the learning rate, number of trees, and tree depth, BRT can balance between accuracy and generalization, minimizing overfitting while maintaining predictive strength.

BRT is also considered a branch of machine learning, which is a subset of artificial intelligence (AI). As an AI-based algorithm, BRT belongs to the family of ensemble learning methods that combine the outputs of multiple weak learners to achieve higher predictive performance. Other well-known machine learning techniques that share similar objectives include Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbors (k-NN). Each of these methods has its own strengths and limitations, but BRT stands out for its balance between accuracy, interpretability, and computational efficiency, making it suitable for both predictive modelling and explanatory analysis.

Because of these advantages, BRT has gained wide adoption in various research domains. It is now considered one of the most reliable ensembles learning methods for predictive analytics, especially when the relationships between variables are complex and not easily captured by linear models. The popularity of BRT continues to grow with the development of efficient computational tools and open-source software such as R and Python, which make implementation easier for researchers and practitioners. This paper aims to introduce the concept, methodology, advantages, and applications of Boosted Regression Trees (BRT) in a simple and structured manner suitable for academic understanding. To achieve this objective, the paper provides a comprehensive review of existing literature, explains the underlying algorithm and mathematical formulation of BRT, and presents a comparative analysis with traditional methods such as Multiple Linear Regression (MLR) to highlight its strengths and practical relevance.

Literature Review

The foundation of BRT lies in the development of decision tree algorithms and the concept of boosting. The Classification and Regression Tree (CART) methodology introduced by Breiman et al. (1984) laid the groundwork for tree-based modeling. Later, the concept of boosting was introduced by Freund and Schapire (1996) through the AdaBoost algorithm, which improved weak learners by iteratively emphasizing misclassified observations. Friedman (2001) then extended the boosting concept to regression problems through the Gradient Boosting Machine (GBM) framework, which uses the gradient of the loss function to guide model improvement. Elith, Leathwick, and Hastie (2008) popularized the application of BRT in ecological and environmental modeling due to its ability to model complex nonlinear relationships. Since then, BRT has been applied in various disciplines such as finance, epidemiology, and social sciences.

In Malaysia, the application of Boosted Regression Trees (BRT) has gained considerable attention in recent years across diverse research areas. For instance, Cheong et al. (2014) employed BRT to analyze the relationship between land use factors and the incidence of dengue cases in Selangor. Their study demonstrated that BRT effectively identified important predictors such as built-up areas, water bodies, and vegetation cover, showing high accuracy in mapping dengue risk areas. Similarly, McCluskey et al. (2014) applied BRT in the field of real estate to perform mass appraisal of residential property values in Malaysia. Their findings revealed that BRT produced more reliable valuation results compared to conventional regression models, mainly due to its ability to capture nonlinear and complex interactions among property characteristics.

In environmental studies, Shaziayani et al. (2021) utilized an enhanced hybrid model integrating Support Vector Machines (SVM) with BRT and quantile regression (QR) to predict PM_{10} concentration levels across several Malaysian cities, including Alor Setar, Klang, and Kuching. Their model, termed SVM_BRT-QR, outperformed traditional predictive methods in forecasting air quality levels three days ahead. Additionally, Shaziayani et al. (2020) explored the integration of quantile regression into the BRT framework for emission modeling and air pollution forecasting in Malaysia. The study highlighted the robustness of BRT-based models in handling environmental datasets characterized by high variability and uncertainty. These studies collectively demonstrate that BRT is an adaptable and accurate modeling approach widely applied in Malaysia for health, environmental, and economic prediction problems.

Methodology

The Boosted Regression Tree (BRT) method, originally proposed by Friedman (2001), combines two fundamental ideas: regression trees and boosting. Regression trees partition the predictor space into regions based on decision rules, where each terminal node represents a prediction value. However, a single regression tree often suffers from instability and overfitting. Boosting addresses these issues by iteratively adding weak learners (shallow trees) to minimize the residual errors of the ensemble model, resulting in a more accurate and robust prediction model. The general algorithm of BRT can be described as Figure 1. Figure 1 shows the general algorithm of the Boosted Regression Tree (BRT) model, which begins by initializing the model with a constant value, typically the mean of the response variable. The process then iteratively computes the residuals between the observed and predicted values, fits a regression tree to these residuals, and updates the model by adding a fraction of the new tree's predictions, controlled by a learning rate (λ). This iterative process continues for a fixed number of iterations or until the model converges, meaning further improvements in prediction accuracy become minimal. Mathematically, the model can be expressed as:

$$F_m(x) = F_{m-1}(x) + \lambda h_m(x) \qquad (1)$$

where

$F_m(x)$ is the updated model at iteration m

$h_m(x)$ is the fitted tree to the residuals

λ ($0 < \lambda \leq 1$) is the learning rate controlling the contribution of each tree.

This stepwise procedure (1) enables BRT to build a strong predictive model by sequentially improving upon the weaknesses of previous iterations.

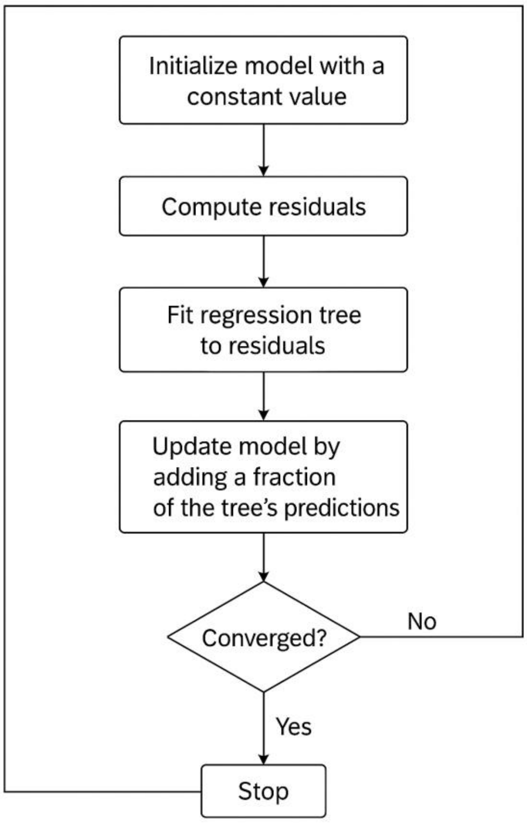

Figure 1. Iterative Framework of the Boosted Regression Tree (BRT) Algorithm

Comparison of Multiple Linear Regression and Boosted Regression Trees

Multiple Linear Regression (MLR) has been widely applied in statistical modelling, particularly for predicting environmental variables such as air pollutants. It is a well-established technique that models the relationship between a dependent variable and multiple independent variables under the assumption of linearity. MLR is often preferred due to its simplicity, ease of interpretation, and straightforward implementation. However, despite its popularity, the method is built on several strict assumptions, including linearity, normality of residuals, homoscedasticity, and independence of errors. In real-world applications, especially in environmental and atmospheric studies, these assumptions are frequently violated, which can lead to biased parameter estimates and reduced predictive performance.

In addition, previous studies have highlighted that MLR may perform poorly when dealing with complex systems. Shahraiyni et al. (2016) reported that although MLR is commonly used in air quality modelling, its effectiveness is limited in the presence of multicollinearity and nonlinear relationships among variables. Similarly, Dormann et al. (2013) emphasized that violations of model assumptions can significantly reduce the reliability of linear regression models, particularly when predictors are highly correlated or exhibit nonlinear interactions. As environmental datasets are often dynamic and complex, traditional linear approaches may not be sufficient to capture underlying patterns accurately.

In contrast, Boosted Regression Trees (BRT) offer a more flexible and powerful alternative for modelling complex relationships. BRT combines regression trees with boosting techniques, where multiple decision trees are built sequentially and each new tree is trained to correct the errors of the previous ones. This iterative learning process enables BRT to capture nonlinear relationships and interactions among variables without requiring strict statistical assumptions about the data distribution. As a result, BRT is particularly suitable for high-dimensional and complex datasets commonly found in environmental studies.

One of the main distinctions between MLR and BRT lies in their predictive capability and robustness. While MLR performs adequately when the relationship between variables is linear and assumptions are satisfied, its performance declines in the presence of nonlinearity, interactions, and noisy data. On the other hand, BRT generally demonstrates higher predictive accuracy and greater robustness to outliers and missing data, making it more reliable in practical applications.

However, despite its advantages, BRT is often considered less interpretable than MLR. MLR provides explicit coefficient estimates that are easy to interpret and explain, whereas BRT requires additional tools such as variable importance measures and partial dependence plots to understand the influence of predictors. Therefore, the choice between MLR and BRT depends on the objective of the study: MLR is preferred when interpretability is essential, while BRT is more appropriate when the focus is on prediction accuracy and modelling complex nonlinear relationships.

Conclusion

In conclusion, this paper has presented a comprehensive overview of Boosted Regression Trees (BRT), including its conceptual foundation, methodological framework, and key advantages in statistical modelling. BRT is a powerful ensemble learning technique that integrates regression trees with boosting algorithms to iteratively improve predictive performance. Unlike traditional methods, it does not rely on strict statistical assumptions, making it highly adaptable to complex and real-world datasets.

The comparative discussion between BRT and Multiple Linear Regression (MLR) highlights the limitations of conventional linear approaches when dealing with nonlinear relationships, multicollinearity, and high-dimensional data. While MLR remains useful due to its simplicity and interpretability, its performance is often constrained in complex systems. In contrast, BRT demonstrates superior predictive accuracy, flexibility, and robustness, particularly in handling interactions among variables and irregular data structures commonly found in applied research.

Overall, the findings suggest that BRT is a highly effective alternative for modern statistical modelling, especially in fields involving complex and nonlinear data patterns. Although it is less interpretable compared to MLR, its strong predictive capability makes it a valuable tool for researchers and practitioners. Future studies may further explore hybrid approaches or model interpretation techniques to enhance the usability and interpretability of BRT in various applied domains.

References:

Breiman L, Friedman J. H., Olshen R. A., & Stone C. J. (1984). Classification and regression trees. *Wadsworth International Group, Belmont.*

Cheong, Y. L., Leitao, P. J., & Lakes, T. (2014). Assessment of Land Use Factors Associated With Dengue Cases in Malaysia Using Boosted Regression Trees. *Spatial and Spatio-Temporal Epidemiology, 10,* 75–84.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., & Carre, G., et al., (2013). Collinearity: a review of methods to deal with it and a simulation study eval- uating their performance. *Ecography, 36*(1), 27-46.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference* (pp. 148-156).

McCluskey, W. J., Daud, D., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction*. *19*(2).

Shahraiyni, H. T., & Sodoudi, S. (2016). Statistical Modeling Approaches for PM_{10} Prediction in Urban Areas; A Review of 21st-Century Studies. *Atmosphere, 7(2),* 15.

Shaziayani, W. N., Ul-Saufie, A. Z., Ahmat, H. & Al-Jumeily O. D. (2021). Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM10 concentration. Air Quality, *Atmosphere & Health. 14*(10).

Shaziayani, W. N., Ahmat, H., Razak, T. R., Zainan Abidin, A.W., Warris, S. N., Asmat, A., Noor, N. M., & Ul-Saufie, A. Z. (2022). A Novel Hybrid Model Combining the Support Vector Machine (SVM) and Boosted Regression Trees (BRT) Technique in Predicting PM10 Concentration. *Atmosphere, 7(*12).

eISBN 978-629-98755-9-8

SIG CS@e-Learning
Unit Penerbitan
Jabatan Sains Komputer & Matematik
Universiti Teknologi MARA Cawangan Pulau Pinang

e-ISBN : 978-629-98755-x-x

Design of the cover powered by
https://www.free-powerpoint-templates-design.com/