Model Prediction Study of Boston Dataset Based on Random Forest Fusion Algorithm

Peiyi Gao

School of Chemistry and Life Sciences, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023

Abstract: *This study is based on the random forest fusion algorithm for model prediction of the Boston dataset, aiming to explore the application of statistics in biomedical research. Through data visualization and a variety of statistical methods, this study provides an in-depth analysis of the characteristics of the dataset and variable relationships. The study first introduces the importance of statistics in the biomedical field, including the application of descriptive statistics, inferential statistics, Bayesian statistics, probability theory, regression analysis, multivariate analysis, and survival analysis. Subsequently, this study elaborates on the Random Forest algorithm and constructs a hybrid model to improve the prediction accuracy by fusing the Gradient Boosted Tree (GBDT) model. Experimental results show that the fusion model performs well in reducing prediction errors and improving model stability, especially in the house price prediction task, where the fusion model outperforms the single model in terms of mean square error (MSE) and coefficient of determination (R^2). By analyzing the Boston house price dataset, this study finds that variables such as the average number of rooms per dwelling (RM), the percentage of low-income people (LSTAT), and air quality (NO_X) have particularly significant effects on house prices. This study not only verifies the validity of the fusion model, but also provides a scientific basis for urban planning and policy making. Future research can further optimize the model, combine deep learning techniques, expand the dataset and variables, and deepen interdisciplinary applications to enhance the universality and practicality of the research results.*

Keywords: *Random Forest Algorithm, Gradient Boosting Tree, Data Visualization, House Price Prediction, Model Fusion*

1. Introduction

Statistics is a science that focuses on the collection, analysis, interpretation and presentation of data, and is widely used and important in the biomedical field[1]. Descriptive statistics summarize patient population characteristics, such as age and weight, through metrics such as mean and median, while inferential statistics draw conclusions from sample data and generalize to the population for assessing drug effectiveness and disease risk[1]. Bayesian statistics combine prior knowledge with new data to support drug development and clinical trial design. Probability theory explains random events to help understand genetic variation and the probability of disease occurrence; regression analysis explores relationships between variables and is used to predict disease risk and assess treatment efficacy. Multivariate analysis explores the complex relationship between biomarkers and disease and identifies disease subtypes; survival analysis assesses treatment effects and predicts probability of survival. Statistics are indispensable in clinical trial design, bioinformatics, and epidemiology, and support genomics, proteomics data analysis, and public health policy development.

In the field of statistics and machine learning, the Random Forest algorithm is widely used in classification and regression tasks due to its strong predictive power and robustness. Breiman (2001) first proposed the Random Forest algorithm and demonstrated its effectiveness in classification and regression tasks[2]. However, despite the many advantages of random forests, such as high accuracy and robustness to high-dimensional data, it also has some limitations.

First, the computational cost of random forests is high, especially when dealing with large-scale datasets. Since multiple decision trees need to be trained, the complexity and training time of the algorithm increases with the number and depth of trees. In addition, the model complexity of random forests is high and difficult to interpret. The “black-box” nature of Random Forests limits them in scenarios that require model interpretability compared to single decision trees. Finally, although Random Forests reduce the risk of overfitting through integrated learning, they may still face bias and variance problems in some cases if the parameters are not properly tuned.

To address these limitations, subsequent studies have improved and optimized random forests. For example, Liaw and Wiener (2002) experimentally verified the superiority of random forests in dealing with high-dimensional data[3], while Antipov and Pokryshevskaya (2012) applied it to house price prediction in Moscow and achieved high prediction accuracy[4]. In addition, Gradient Boosting Tree (GBDT), as an integrated learning algorithm, significantly improves the predictive ability of the model by gradually optimizing the loss function[5]. In summary, although the Random Forest algorithm performs well in the prediction task, it still has disadvantages such as high computational complexity and poor model interpretability. The existence of these drawbacks is mainly due to its integrated learning-based nature, which leads to an increase in model complexity and training cost. Therefore, subsequent studies have further enhanced the performance and applicability of the model by improving the algorithm or fusing it with other algorithms.

2. Methods

2.1 Random forests based on decision trees

Random Forest is an integrated learning algorithm based on Bagging, which is widely used in classification and regression tasks. The specific structure is shown in Figure 1, which improves the accuracy and robustness of the model by constructing multiple decision trees and synthesizing their predictions. The core idea of Random Forest is to utilize the advantage of “integrated learning” to combine multiple weak learners (decision trees) into a single model by “majority voting” (classification task) or “averaging” (regression task). (decision tree) into one strong learner by means of “majority voting” (classification task) or “averaging” (regression task) [6].

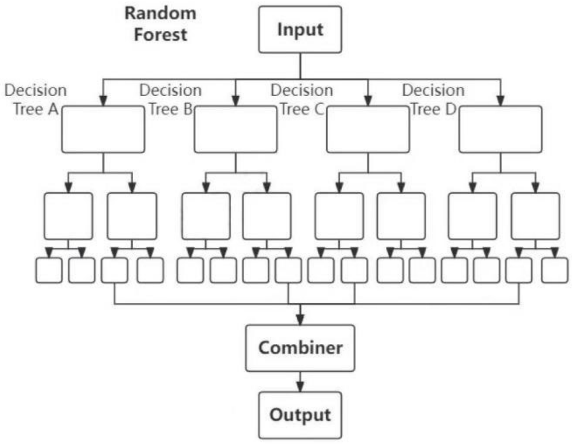


Figure 1 RF random forest structure

Random forests are based on decision trees, which are simple models that predict a target variable through a series of “questions” (features) and “answers” (decision nodes)[7]. However, decision trees have some drawbacks, such as being prone to overfitting (performing well on training data but poorly on new data) and being sensitive to small changes in the data. Random forest solves these problems by constructing multiple decision trees and synthesizing their results[8]. The prediction formula for the Random Forest regression model is given in Equation (1).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{1}$$

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

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

The mean square error (MSE) (Eq. (2)) and the coefficient of determination R^2 (Eq. (3)) used by Random Forest to evaluate the model performance.

2.2 The Fusion model

Assume that the prediction result of the random forest is $\hat{y}_{\mathrm{RF}}$, the prediction result of the gradient boosting tree is $\hat{y}_{GBDT}$, and the prediction result of the fusion model is $\hat{y}_{Ensemble}$ [9]. The fusion model can be realized by weighted average as shown in equation (4):

$$\hat{y}_{Ensemble}=\alpha\cdot\hat{y}_{RF}+(1-\alpha)\cdot\hat{y}_{GBDT} \tag{4}$$

Where α is a weight parameter, usually determined by cross-validation.

$$MSE_{Ensemble}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_{Ensemble,i})^2 \tag{5}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_{Ensemble,i})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{6}$$

Correspondingly, the mean square error (MSE) (Equation (5)) and the coefficient of determination R^2 (Equation (6)) used in the fusion model are shown above, where, y$_i$ is the true value, $\hat{y}_{Ensemble,i}$ is the predicted value of the fusion model, $\overline{y}$ is the mean of the true value, and n is the sample size. The numerator represents the sum of squared prediction errors of the fusion model, which reflects the deviation between the predicted and true values of the model [10-12]. The denominator represents the total variance between the true value and the mean, reflecting the variability of the data itself. The closer the R^2 value is to 1, the better the model explains the data; the closer the value is to 0, the model has little or no explanatory power.

3. Experiments

3.1 Dataset exploration and visualization

In order to gain a deeper understanding of the relationships between the features and variables of the Boston dataset, this study explores a variety of visualization methods, including correlation heatmaps, 3D scatter plots, radar diagrams, network diagrams, chordal plots, and 3D surface plots. These methods reveal the structure of the dataset and the complex relationships among variables from different perspectives.

The correlation heatmap visualizes the strength of correlation between the variables as shown in Figure 2(1). The results show that the proportion of low-income strata (LSTAT) is significantly negatively correlated with the median house price (MEDV), while the average number of rooms per dwelling (RM) is significantly positively correlated with the MEDV.

The 3D scatter plot demonstrated the relationship between LSTAT, RM and MEDV through a three-dimensional view, as shown in Figure 2(2). The figure shows that an increase in LSTAT is associated with a decrease in MEDV, while an increase in RM is associated with an increase in MEDV, revealing an interaction between the variables.

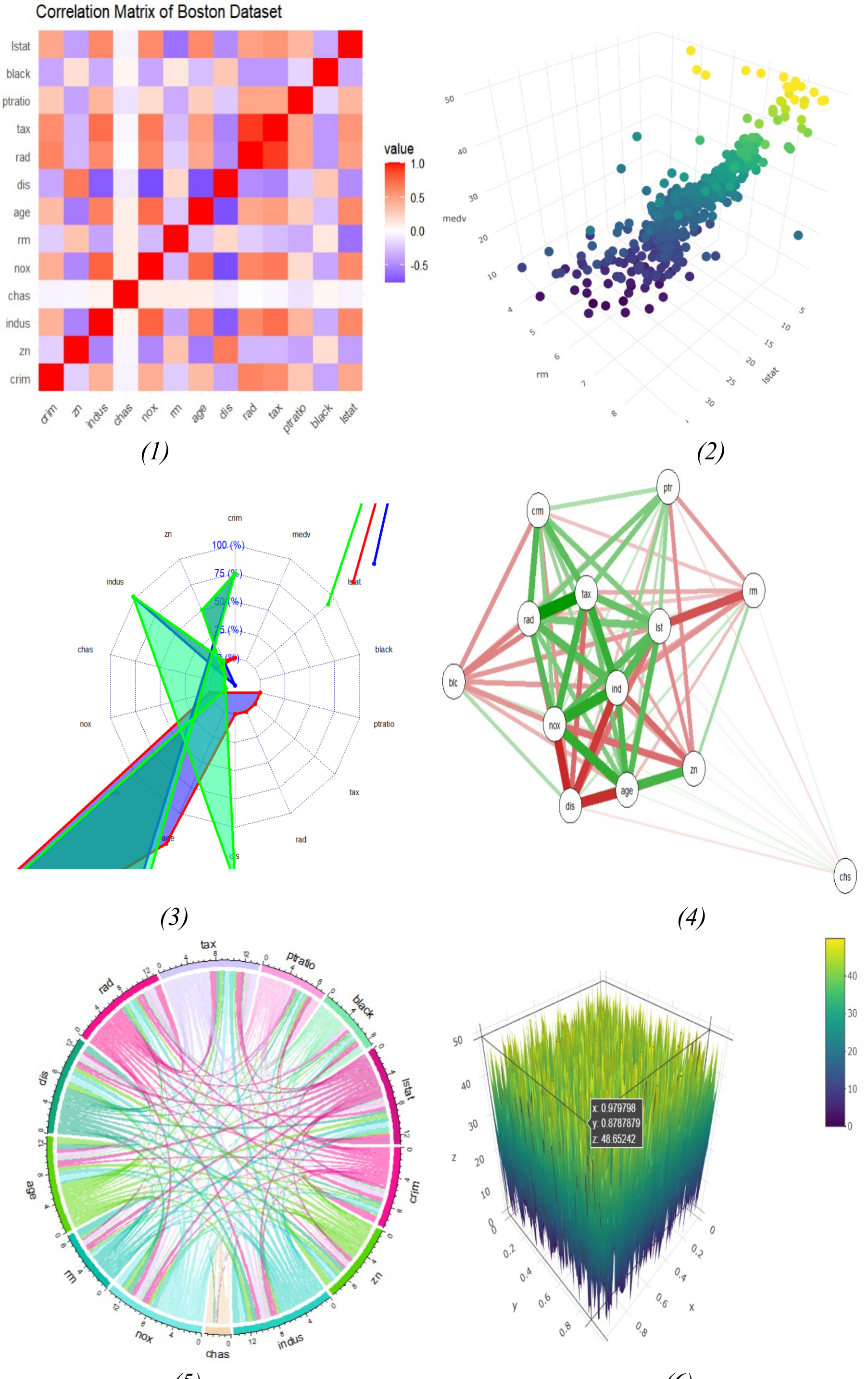
The radar plot shows the characteristics of the top 5 observations in the dataset on multiple variable dimensions as shown in Figure 2(3). By comparing the performance of different observations on each variable, the radar plot helps to identify outliers and patterns in the data.

The network graph demonstrates the complex pattern of correlation between variables through the connectivity of nodes and edges as in Fig. 2(4). The graph clearly shows the negative correlation of LSTAT with several variables, revealing direct and indirect relationships between variables.

The chord plot demonstrates the strength of the bi-directional correlation between the variables in a circular layout, as in Figure 2(5). The negative correlation between LSTAT and MEDV is particularly significant in the plot, indicating that this relationship is more prominent in the data.

The 3D surface plot visualizes the nonlinear relationship between LSTAT, RM and MEDV through the shape of the surface, as shown in Figure 2(6). The figure shows that MEDV peaks when LSTAT is lower and RM is higher, revealing a complex interaction between the variables.

Through the above visualization analyses, this study reveals the characteristics of the Boston house price dataset and the relationships among variables from multiple perspectives. These analyses provide an important basis for subsequent model construction and help identify the key factors and their interactions that affect house price.



(1) *(2)*

(3) *(4)*

(5) *(6)*

Figure 2 Visualization results graph presentation

3.2 Fusion model prediction datasets

3.2.1 Assessment model indicator selection

The dynamics of housing prices have always been affected by a variety of factors, with more or less fluctuations and ups and downs. In the case of the Boston area reflected in the Boston dataset, housing prices are affected by a variety of factors, such as the proportion of students and teachers, the proportion of low-income people, the distance of the property to the city center, and the quality of air, etc. The complexity of these factors not only adds to the difficulty of urban planning and construction, but also brings to the residents of the housing choice uncertainty. In view of this, this study chooses the mean square error (MSE) and the coefficient of determination (R^2) as the model evaluation indexes, as shown in Table 1.

Since errors are squared when making model predictions, larger prediction errors are magnified, which makes the model focus more on reducing those predictions that deviate significantly from the true value. MSE addresses exactly this problem, a property that is particularly important in house price prediction, where over- or under-estimation of house prices can lead to large economic impacts. In addition, MSE has good mathematical properties, especially the derivability, which makes it perform well in gradient descent-based optimization algorithms and helps the model learn and tune the parameters more efficiently on the Boston dataset.

The coefficient of determination (R^2), on the other hand, reflects the model's ability to capture the relationship between house prices and characteristics (e.g., education, occupation, socio-economic structure).R^2 is a relative indicator that not only measures the model's forecasting accuracy, but also reflects the extent to which the model improves relative to a simple mean forecast. A high R^2 value indicates that the model is better able to explain the variation in house prices, thus providing a scientific basis for policy making. Through R^2, the degree of influence of various factors (e.g., teacher-student ratio, proportion of low-income people) on house prices can be visualized, thus providing data support for urban planning and policy adjustment.

Table 1 Selection of Indicators for the Boston House Price Forecasting Model

Indicator Name	Indicator Content
MSE	Mean Square Error
R2	Coefficient of determination

3.2.2 Construction of the fusion model

Calle et al. used random forests in genomics research for disease prediction via the AUC-RF strategy, demonstrating its potential in biomedical data analysis. Inspired by this, this study attempts to apply the Random Forest algorithm to the analysis of the Boston house price dataset, aiming to explore its performance in the task of house price prediction. However, during the experimental process, the limitations of Random Forest when used alone gradually appeared, mainly in the form of a greater risk of overfitting and the need to improve the prediction accuracy. In view of this, this study introduces the gradient boosted tree model (GBDT), which has significant advantages in solving the overfitting problem and improving the prediction accuracy. In order to fully utilize the advantages of the two models, this study constructs a hybrid model integrating Random Forest and GBDT with the help of R language and applies it to the prediction task of Boston house price dataset. By comparing and analyzing the prediction results with the single model, the hybrid model achieves a more significant improvement in prediction performance, which verifies the effectiveness and superiority of the hybrid model in the field of house price prediction.

3.2.3 Fusion modeling for house price forecasting

In order to visualize the key factors affecting house prices, this study calculates the feature importance with the help of the random forest model and combines it with the gradient boosting tree model for correction. As shown in Figure 3, the results of the analysis indicate that among the 13 features included in the dataset, the average number of rooms per dwelling (RM), the proportion of low-income strata (LSTAT), and the environmental indicator (NO_X) have a significant impact on house prices and are almost dominant. In addition, distance to the five Boston employment centers (DIS), crime rate (CRIM), and the percentage of students who are teachers in the town (PTRATIO) also have a significant impact on house prices.

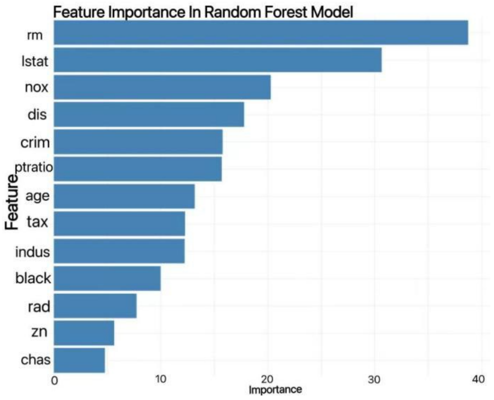


Figure 3 Bar chart of feature importance

The results reflected in Figure 3 are inextricably linked to the socio-economic logic of Boston's urban development. In Boston's urban planning, housing prices are influenced by a combination of socioeconomic and environmental factors, and there are complex interactions and recursive relationships between these factors. First, areas with a higher average number of rooms are usually newly built or remodeled residential areas, where developers may lower the threshold for occupancy in order to quickly attract tenants, attracting groups with relatively weak economic power and thus pushing up the proportion of low-income people in the area. This phenomenon further affects the pattern of functional zoning in cities: areas with a higher proportion of low-income people tend to have significant differences in land use and infrastructure development from other areas. Due to economic conditions, commercial development and residential construction in these areas are constrained, resulting in a weaker correlation with variables reflecting the city's economic vitality (e.g., the proportion of commercial land use, the proportion of teachers and students), thus providing a unique perspective for studying the interaction between the city's economy and other functional segments[4].

In addition, the concentrated distribution of industrial land is closely related to environmental pollution. Industrial areas usually concentrate a large number of factories with frequent production activities, leading to a significant increase in NO_X concentrations. This environmental problem not only affects the health of residents, but also may further aggravate the concentration of low-income groups, creating a vicious circle. Meanwhile, the proximity to the city center directly determines the accessibility and living comfort of the area. Areas closer to the city center have easier access to employment opportunities, commercial facilities and cultural and recreational resources, which attracts more residents and pushes up property prices, while areas farther away from the city center have lower demand for properties and lower property prices due to inconvenient transportation and lack of amenities.

Crime rates, as an important indicator of community safety levels, have a particularly significant impact on housing prices. High crime rates not only reduce residents' sense of safety and well-being, but may also trigger the loss of educational resources, teachers' reluctance to take up employment, students' choice to transfer to other schools, which in turn affects the teacher-student ratio (PTRATIO), creating a vicious cycle. The distribution of educational resources also has a profound impact on housing prices. Lower PTRs usually mean tighter educational resources, which may reduce the quality of education and thus the willingness of families to move in, reducing demand for real estate, while higher-quality educational resources can attract families to move in and push up the price of housing in school districts.

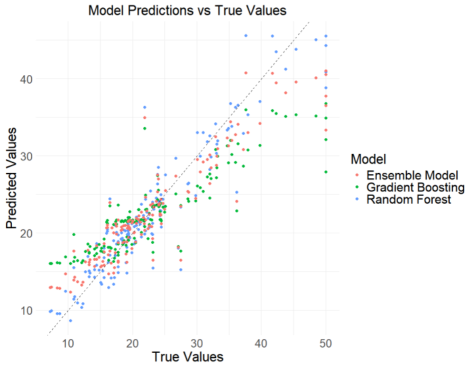


Figure 4 Comparison of model prediction performance

In order to verify the prediction accuracy of the constructed model, this study conducted a detailed comparative analysis of the model's prediction results with the real value of Boston house prices. In the comparative analysis, a scatter plot is drawn with the real value of Boston house prices as the horizontal coordinate (x-axis) and the predicted value of the model as the vertical coordinate (y-axis). The black dotted line in the graph indicates the ideal prediction state, i.e., the case where the true value is exactly the same as the predicted value. The closer the scatter points are to this dotted line, the closer the predictive effect of the model is to the ideal state and the higher the prediction accuracy. As shown in Figure 4, the prediction results of the fusion model present high accuracy and strong robustness, and most of the scatters are closely distributed near the dotted line, indicating that the model can effectively and accurately predict house prices in Boston.

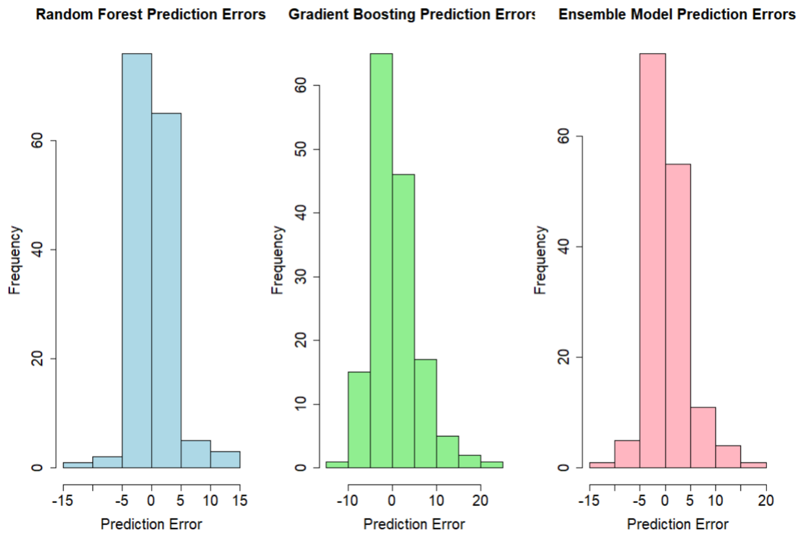


Figure 5 Prediction error

Inevitably, all three models have some degree of prediction error. By visualizing the prediction errors of the three models, the fusion model still shows significant advantages. As shown in Figure 5, the error distribution of the fusion model is the most concentrated and closest to the normal distribution, indicating that its prediction results have higher stability and accuracy.

The residual plot can visualize the difference between the model's predicted value and the real value, as shown in Fig. 6, the fusion model performs best in the Boston house price prediction task, with the most uniform distribution of the residuals and most of the residuals are close to zero, which indicates that the model's predicted value has the smallest deviation from the actual value.

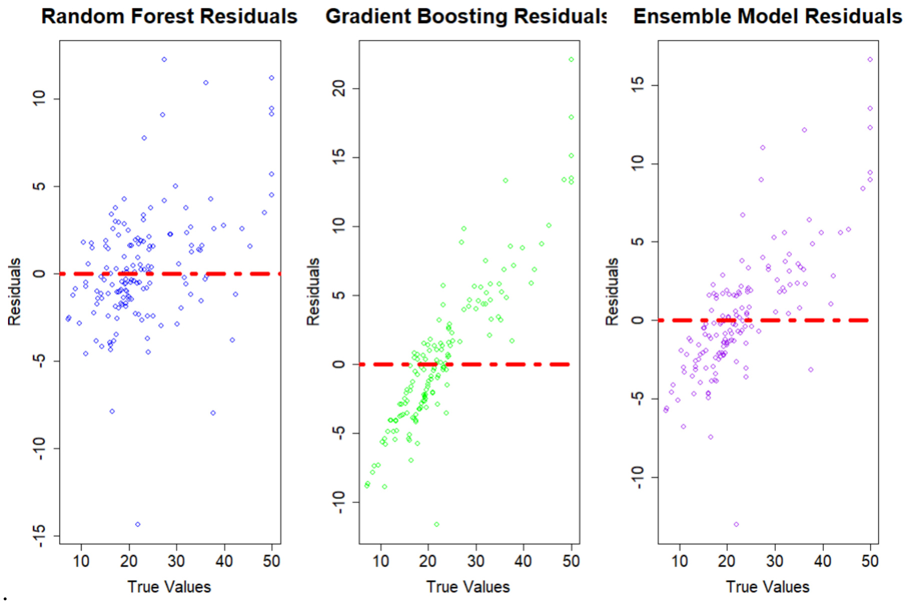


Figure 6 Residual plot

Finally, this study made a quantitative analysis of the assessment of the model's predictive ability by calculating the MSE and R^2 (to four decimal places), as shown in Table 2.

Table 2 Quantitative analysis of assessment indicators

Model	MSE	R2
Random forest model	11.4696	0.8720
Gradient boosted tree model	28.2663	0.8054
Fusion model	16.4449	0.8568

The MSE of the fusion model is lower than that of the gradient boosted tree model, but slightly higher than that of the random forest model. This suggests that the fusion model is more effective than the gradient boosted tree in reducing prediction errors, even though it does not reach the lowest MSE value of the random forest. However, this slight increase in MSE may be acceptable considering the advantages that the fusion model may have in other aspects, such as better stability and generalization ability.

The R^2 of the fusion model is higher than that of the gradient boosted tree and close to that of the random forest. This means that the fusion model is better able to account for variability in the data, i.e., it is able to capture more accurately the relationship between house prices and features.

The fusion model combines the advantages of Random Forest and Gradient Boosted Tree, and reduces the bias and variance of individual models through model fusion, thus improving the overall prediction performance.

4. Conclusions

This study demonstrates the application of statistics in biomedical research using the Boston dataset, analyzing variable relationships through various statistical methods and data visualization techniques. For instance, correlation heatmaps revealed that the tax rate positively correlates with highway accessibility, while the percentage of lower-status individuals negatively correlates with the average number of rooms. For model prediction, a fusion model combining Random Forest (RF) and Gradient Boosted Tree (GBDT) was developed, outperforming single models by reducing prediction error (MSE) and improving explanatory power (R^2). Feature importance analysis identified key factors influencing house prices, including the average number of rooms (RM), low-income population proportion (LSTAT), air quality (NO_X), crime rate (CRIM), student-teacher ratio (PTRATIO), and distance to employment centers (DIS).

Based on these findings, the study proposes policy recommendations to promote real estate market development. These include balancing educational resources to attract families, improving living conditions for low-income groups, and enhancing air quality, safety, and transportation infrastructure to boost housing prices and support sustainable urban growth. Future research could integrate deep learning techniques to refine the model, expand datasets with macroeconomic and community factors, and adopt interdisciplinary approaches combining statistics with urban planning and sociology. Such efforts would provide deeper insights into housing price dynamics and inform comprehensive urban governance strategies.

References

[1] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[2] Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[3] Liaw A, Wiener M. Classification and regression by random Forest[J]. R news, 2002, 2(3): 18-22.

[4] Antipov E A, Pokryshevskaya E B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics[J]. Expert systems with applications, 2012, 39(2): 1772-1778.

[5] Bamford T, Easter C, Montgomery S, et al. A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos[J]. Human reproduction, 2023, 38(4): 569-581.

[6] Yan X, Li J, Smith A R, et al. Evaluation of machine learning methods and multi-source remote sensing data combinations to construct forest above-ground biomass models[J]. International Journal of Digital Earth, 2023, 16(2): 4471-4491.

[7] Louppe G. Understanding random forests: From theory to practice[D]. Universite de Liege (Belgium), 2014.

[8] Finkelshtein B, Baskin C, Maron H, et al. A simple and universal rotation equivariant point-cloud network[C]//Topological, Algebraic and Geometric Learning Workshops 2022. PMLR, 2022: 107-115.

[9] Theodoridis G, Tsadiras A. Retail Demand Forecasting: A Multivariate Approach and Comparison of Boosting and Deep Learning Methods[J]. International Journal on Artificial Intelligence Tools, 2024, 33(04): 2450001.

[10] Zhou Z H. Ensemble methods: foundations and algorithms[M]. CRC press, 2025.

[11] Merodio Gómez P, Juarez Carrillo O J, Kuffer M, et al. Earth observations and statistics: Unlocking sociodemographic knowledge through the power of satellite images[J]. Sustainability, 2021, 13(22): 12640.

[12] Kim M, Kim D, Jin D, et al. Application of explainable artificial intelligence (XAI) in urban growth modeling: A case study of Seoul metropolitan area, Korea[J]. Land, 2023, 12(2): 420.