УДК 005

ENHANCING HEART DISEASE PREDICTION USING GAUSSIAN PROCESS REGRESSION AND SEQUENTIAL ANALYSIS FOR OPTIMAL SAMPLING EFFICIENCY

He Runhai, Li Fusheng, Zhang Zhenxing, Zhang Shoujing
Belarusian State University
e-mail: fpm.he@bsu.by

Summary. *Heart disease poses a significant risk to global health, and accurate prediction of this risk is critical to public health. This study leverages the Kaggle heart disease dataset, employs machine learning models to predict heart disease risk, and introduces sequence analysis to minimize sample collection while maintaining accuracy. Of the four models, linear regression, Bayesian regression, random forest, and Gaussian process regression, Gaussian process regression proved to be the most accurate. Combined with sequential analysis, we found appropriate sampling stops to reduce acquisition costs while maintaining prediction accuracy.*

Heart disease is one of the leading causes of death worldwide, and accurate prediction of heart disease risk can help identify high-risk groups early, achieve timely intervention, and reduce the risk of disease. With the development of data-driven medical research, machine learning-based heart disease prediction models have become a research hotspot due to their efficient data processing and high-precision prediction. However, previous studies have mostly focused on model optimization, and less attention has been paid to model performance under small samples. In this study, 14 key attributes were selected to compare the prediction results of four common machine learning models using the cardiac dataset of the Kaggle platform, and the changes in prediction accuracy during the gradual increase of sample size were explored through sequential analysis to achieve a balance between model accuracy and sampling cost.

Linear Regression models the relationship between a target and predictors through a linear function [1]. Bayesian Ridge Regression incorporates Bayesian principles into linear regression, allowing for uncertainty estimation in the model parameter [2]; Random Forest Regression utilizes an ensemble of decision trees to improve predictive performance and control overfitting (Breiman, 2001) [3]; Gaussian Process Regression applies a non-parametric approach based on Gaussian processes, providing both predictions and uncertainty estimates over function space (Rasmussen & Williams,2006) [4]; Sequential Analysis offers a method for sampling that minimizes the required sample size to achieve the desired level of statistical precision (Wald, 1945) [5].

Four machine learning models, including Linear regression, Bayesian ridge regression, Random Forest regression, and Gaussian process regression, were used to compare their performance in heart disease prediction. The results of the prediction experiment are shown in the figure1, figure 2 and table 1 below.

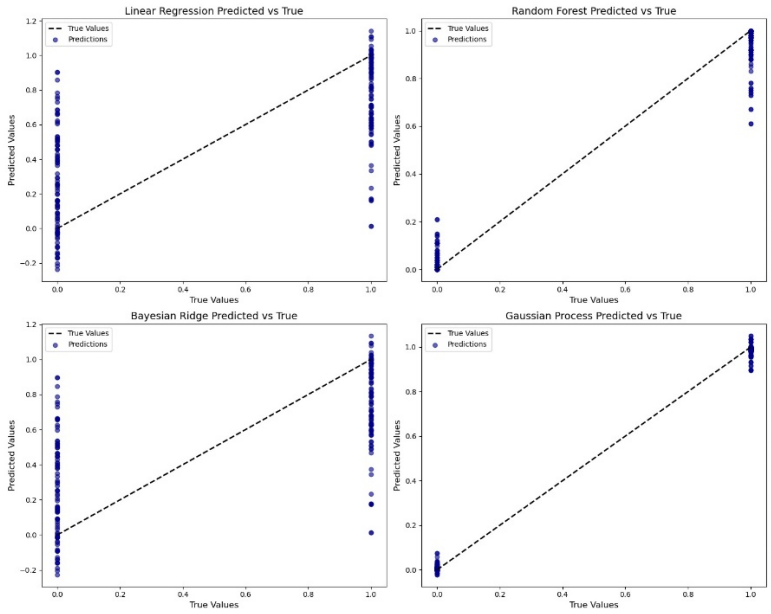

Figure 1 – Predicted results of each model with the actual values

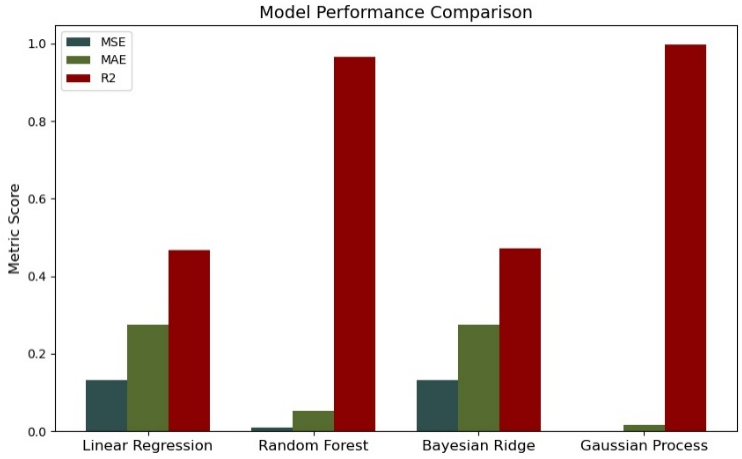

Figure 2 – Visualization of prediction performance comparison

Table 1 – Prediction performance comparison of each model

	Linear Regression	Random Forest	Bayesian Ridge	Gaussian Process
MSE	0.1324	0.0087	0.1314	0.0007
MAE	0.2751	0.0525	0.2747	0.0164
R^2	0.4676	0.9651	0.4715	0.9972

Since the experimental results above show that the Gaussian process regression model has the best effect, this paper further introduces sequential analysis to optimize the sampling process of the Gaussian process regression model to study when to stop sampling under the premise of ensuring the prediction accuracy, so as to reduce the sampling cost. The basic idea of sequential analysis is to automatically stop sampling when the prediction error reaches a certain

standard, thereby effectively reducing the amount and cost of data acquisition. The results are shown in the figure 3, table 2 below.

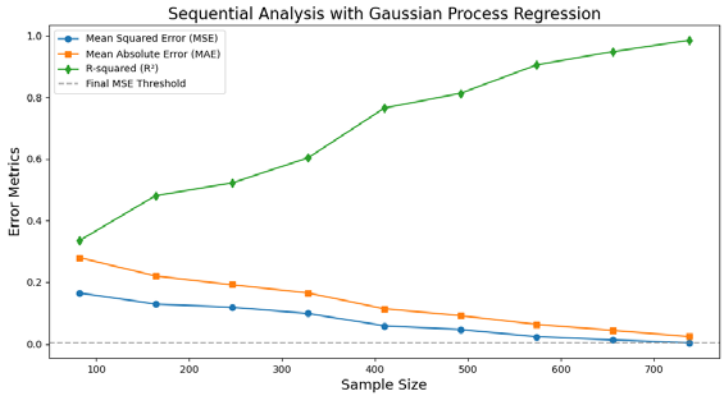

Figure 3 – Dynamic visualization of MSE convergence and evaluation metrics

Table 2 – Dynamic numerical results during sequential analysis

Batch	Sample Size	MSE	MAE	R^2
1	82	0.1652	0.2796	0.3355
2	164	0.1290	0.2200	0.4812
3	246	0.1188	0.1919	0.5222
4	328	0.0985	0.1652	0.6038
5	410	0.0584	0.1139	0.7652
6	492	0.0465	0.0916	0.8131
7	574	0.0235	0.0628	0.9054
8	656	0.0129	0.0438	0.9483
9	738	0.0037	0.0239	0.9851

Based on the experimental results, the initial sample size was 1025. However, only 738 samples were required to achieve an error convergence of less than 0.01, with excellent metrics in Mean Squared Error, Mean Absolute Error, and R^2. This indicates that sequential analysis significantly reduces the necessary sample size while maintaining high model prediction accuracy. This approach effectively lowers data collection costs and provides an efficient data sampling strategy for future heart disease risk prediction studies. Furthermore, it underscores the potential of small sample sets to achieve high prediction accuracy, particularly in medical applications where data access is limited. This finding underscores the value of sequential analysis in data-driven health predictions.

References

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

2. Shi, Q., Abdel-Aty, M., & Lee, J. (2016). A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. – Accident Analysis & Prevention. – 88. – P. 124–137.

3. Breiman, L. (2001). Random forests. – Machine learning. – 45. – P. 5–32.

4. Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning

(Vol. 2, No. 3, p. 4). – Cambridge, MA: MIT press.

5. Wald, A. (1992). Sequential tests of statistical hypotheses. In Breakthroughs in statisti cs: Foundations and basic theory. – New York, NY: Springer New York. – P. 256–298.

УДК 004.021

COMPARATIVE STUDY OF BOSTON HOUSE PRICE PREDICTION MODELS BASED ON GAN DATA AUGMENTATION

He Runhai
Belarusian State University
e-mail: fpm.he@bsu.by

Summary. *This study aims to compare the performance of four machine learning models in the Boston housing price prediction task, and explore whether GAN data augmentation can improve the prediction performance of the model. Experimental results indicate that on the original data set, The CatBoost model demonstrated superior predictive performance, with an MSE of 4.76, RMSE of 2.18, MAE of 1.65, and R^2 of 0.94. However, no significant performance improvement was observed when GAN data augmentation was applied to these models.*

The real estate market is an important part of the national economy, and house price prediction has always been a hot topic of concern in academia and industry. The Boston house price dataset is a recognized classic benchmark dataset for house price prediction. The dataset comes from a real estate study in the 1970s [1]. In recent years, with the rapid development of machine learning technology, various advanced models have been widely used in house price prediction tasks and have achieved good results.

This study aims to compare the performance of four mainstream machine learning models (random forest, XGBoost, gradient boosting, CatBoost) on the Boston house price prediction task, and explore whether the introduction of GAN data augmentation technology on this basis can further improve the prediction accuracy of the model [2, 3, 4, 5]. GAN-based data augmentation generates synthetic samples that mimic the statistical characteristics of the original dataset, aiming to enhance model generalization capabilities [6].

The Boston house price dataset contains 506 samples, 13 features, and the target variable is the house price. During the model training phase, a grid search optimization is performed on the hyperparameters of four basic models and then, compared them with the version with GAN data augmentation.

The Figure 1 below shows the comparison of the predicted values and actual values of each basic model on the test set. As can be seen from the figure, the predicted curve of the CatBoost model is closest to the actual value curve and has the best prediction performance.