Precise Estimation of Fetal Y Chromosome Concentration Based on XGBoost Regression and Entropy-Weighted TOPSIS Optimization

Haifeng Xu [1,*,#], Hao Sun [2,#], Yuqi Fu [2,#]

[1] School of Accounting, Capital University of Economics and Business, Beijing, China, 100070

[2] School of Finance, Capital University of Economics and Business, Beijing, China, 100070

* Corresponding Author Email: 32023030222@cueb.edu.com

[#]These authors contributed equally.

Abstract. Addressing challenges in fetal Y chromosome concentration estimation for non-invasive prenatal testing (NIPT) — including inconsistent measurement units, significant noise, weak monotonic correlations among variables, and strong potential nonlinearity—this study focuses on constructing and optimizing a Y chromosome concentration prediction model. First, bubble plots and correlation analysis revealed significant weak nonlinear relationships between fetal Y chromosome concentration and gestational age at testing, as well as maternal BMI. Building upon this, an XGBoost regression model was developed as the primary predictor and compared with Support Vector Machine Regression. Results demonstrated significantly improved model fit for XGBoost, achieving value of 0.958 and a low standardized root mean square error of 0.707, substantially outperforming the SVR model. To further enhance model robustness and interpretability, this study introduced the entropy weighting method to determine objective weights for key features. Combined with the Two-Operator Partial Indexed System of Selection method, this approach performed weighted optimization of model parameters. The optimized model successfully simplified complex nonlinear relationships into a clearer structure while significantly improving fitting accuracy for extreme and marginal samples. Ultimately, paired t-tests on the same sample error confirmed that the improvement of the optimized model over the baseline model was highly statistically significant.

Keywords: XGBoost Regression, Entropy-Based Method, TOPSIS Method.

1. Introduction

Non-invasive prenatal testing (NIPT) holds significant value in fetal genetic material detection. Its core challenge lies in accurately estimating fetal Y chromosome concentration across varying gestational weeks and maternal physical conditions, thereby informing optimal clinical testing timing and ensuring result reliability. Current research predominantly relies on linear statistical models or univariate empirical thresholds for concentration estimation [1-3]. However, these approaches face significant limitations when dealing with observed biological data, which are often multi-source and heterogeneous, exhibiting inconsistent units, high noise, and coexisting heteroscedasticity. Key variables (e.g., Y chromosome concentration, gestational age, BMI) frequently demonstrate weak monotonic correlations, alongside significant potential interactions and complex nonlinear relationships [4-5]. Consequently, existing discrimination strategies based on single-variable empirical rules or conventional linear models struggle to adequately capture the underlying data complexity, often leading to compromised predictive robustness and limited clinical reproducibility [6-7]. There is, therefore, an urgent need to establish a technical approach that not only achieves high predictive accuracy but also balances statistical significance with clinical interpretability [8-9].

To address these gaps, this paper proposes a comprehensive multivariate modeling and optimization framework. The research systematically investigates the construction and optimization of a Y-chromosome concentration prediction model, followed by decision-making for the optimal NIPT testing window stratified by maternal BMI. The main contributions and workflow of this study are as follows: Firstly, we explicitly identify and quantify the nonlinear relationships among key

variables through advanced visualization techniques (e.g., bubble charts) and nonparametric correlation analyses (Spearman, Kendall's tau-b), moving beyond linear assumptions. Secondly, an XGBoost regression model is developed as the primary predictor to handle the complex nonlinearities and interactions, demonstrating superior performance compared to Support Vector Machine Regression (SVR). Thirdly, to enhance model robustness and interpretability, we introduce an integrated optimization strategy combining the entropy weighting method to objectively determine feature importance and the TOPSIS method for comprehensive model evaluation and parameter weighting. Finally, the optimized model undergoes rigorous significance validation via paired t-tests to ensure the statistical reliability of the improvements [10]. This integrated methodology aims to provide a more precise, reliable, and clinically actionable tool for fetal Y chromosome concentration estimation.

2. Model Establishment and Solution

2.1. Data Preprocessing

The data is from https://www.mcm.edu.cn/. First, the attached data is preprocessed, including handling potential sequencing failures and multiple tests for the same pregnant woman, removing missing values, converting the gestational age in string format (e.g., 11w+3) into gestational days. Statistical models are established based on constructing histograms of individual data and line charts comparing the three types of data [6-7].

The visualization results are shown in the following figures:

As can be seen from Figure 1, the BMI of pregnant women, Y-chromosome concentration, and gestational days for testing approximately follow a normal distribution, which can be used for subsequent significance analysis.

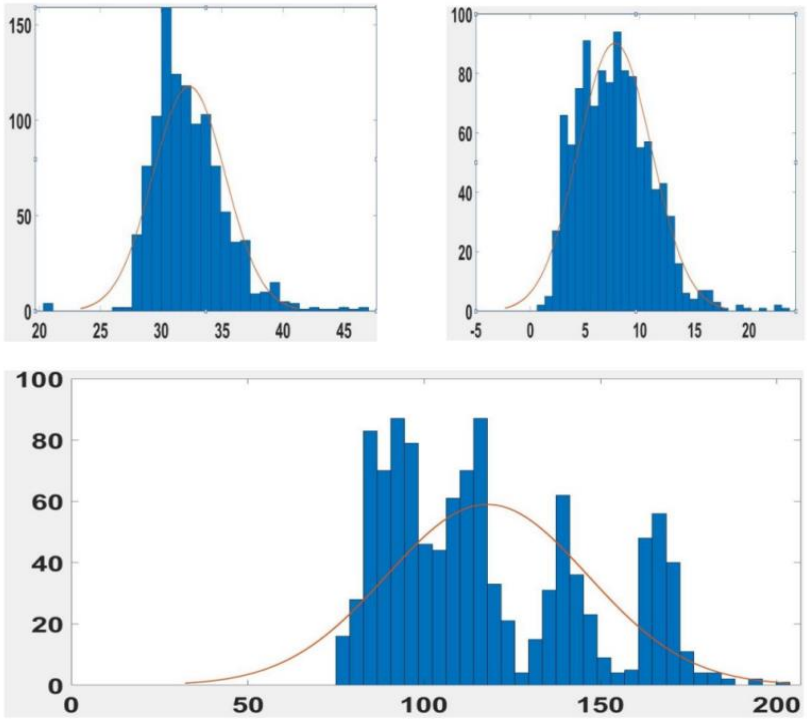

Figure 1. Histogram of Pregnant Women's BMI, Y-Chromosome Concentration, and Gestational Days for Testing

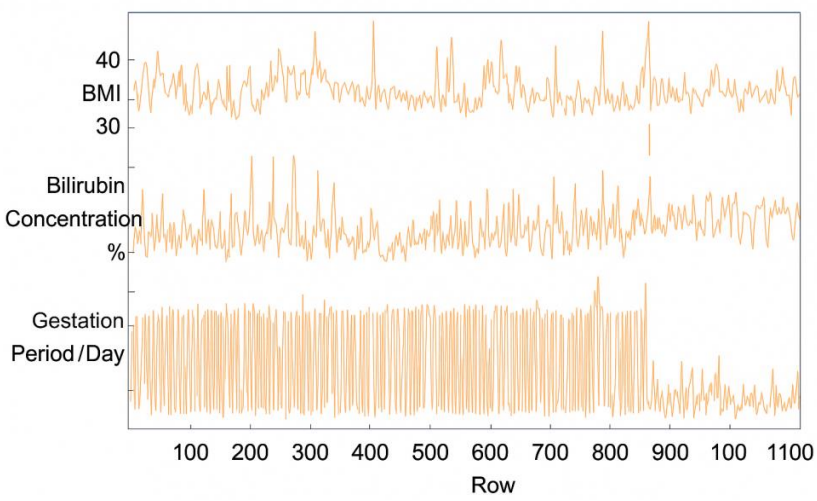

Figure 2. Line Chart Comparison of Three Types of Data

As can be seen from Figure 2, the fluctuation of Y-chromosome concentration has a low linear correlation with the linear changes of gestational weeks and BMI. By further comparing the volatility and differences of the three types of key data, it is convenient for us to initially clarify the relationship between them [8-9]. Bubble Chart of BMI vs. Y-Chromosome Concentration is shown in figure 3.

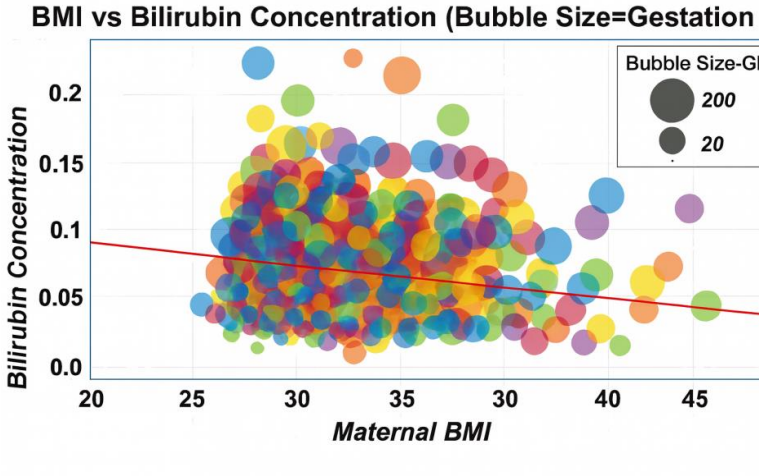

Figure 3. Bubble Chart of BMI vs. Y-Chromosome Concentration

2.2. Establishment of the Spearman Model

Spearman's correlation coefficient is a non-parametric measure of correlation, usually used for correlation analysis of continuous variables. This paper mainly uses this coefficient to evaluate the pairwise correlation between gestational days for testing, pregnant women's BMI, and Y-chromosome concentration. Let the two variables be X and Y, with a sample size of n. Let the m-th observation be X_m and Y_m, then the sample rank difference can be expressed as $d_m = \text{rank}(X_m) - \text{rank}(Y_m)$.

The formula for calculating Spearman's correlation coefficient r is as follows [10]:

$$r = 1 - \frac{6\sum d_m^2}{n(n^2-1)} \quad (1)$$

The calculation results are shown in the following table 1:

Table 1. Calculation Results of the Spearman Model

	Pregnant Women's BMI	Gestational Weeks for Testing	Y-Chromosome Concentration (%)
Pregnant Women's BMI	1 (0.000***)	0.138 (0.000***)	-0.154 (0.000***)
Gestational Weeks for Testing	0.138 (0.000***)	1 (0.000***)	0.085 (0.005***)
Y-Chromosome Concentration (%)	-0.154 (0.000***)	0.085 (0.005***)	1 (0.000***)

Note: The values in parentheses are p-values; *** indicates $p < 0.001$, indicating extremely significant correlation.

The Spearman three-variable correlation heatmap drawn using MATLAB is shown in the following figure 4:

As can be seen from Figure 4, the correlation coefficient between gestational days for testing and BMI is 0.14, showing a weak positive correlation, and the correlation coefficient with Y-chromosome concentration is 0.08, also showing a weak positive correlation. The correlation coefficient between BMI and Y-chromosome concentration is -0.15, showing a weak negative correlation. The pairwise correlations among the three variables are all weak, indicating that there is no obvious monotonic relationship between Y-chromosome concentration and gestational days for testing or BMI, which further confirms the previous conclusion that the relationship among the three is non-linear. Therefore, it is more suitable to construct a multivariate or non-linear model to solve the problem.

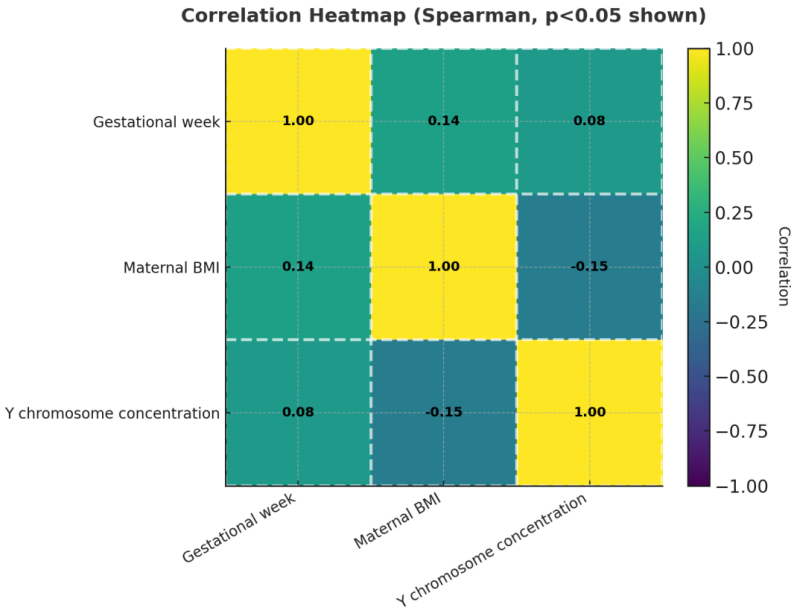

Figure 4. Spearman Three-Variable Correlation Heatmap

Note: The color depth represents the magnitude of the correlation coefficient, with values ranging from -1 to 1. Only correlations with $p < 0.05$ are displayed. Correlation coefficient is shown in table 2.

Table 2. Correlation Coefficient Table

	Pregnant Women's BMI	Y-Chromosome Concentration (%)	Gestational Weeks for Testing
Pregnant Women's BMI	1 (0.000***)	-0.105 (0.000***)	0.094 (0.000***)
Y-Chromosome Concentration (%)	-0.105 (0.000***)	1 (0.000***)	0.06 (0.004***)
Gestational Weeks for Testing	0.094 (0.000***)	0.06 (0.004***)	1 (0.000***)

Note: The values in parentheses are p-values; *** indicates $p < 0.001$, indicating extremely significant correlation.

2.3. Robustness Test

To verify the robustness of the Spearman model, this paper constructs the Kendall's tau-b model to conduct a double test on the relationship among the three variables.

Kendall's correlation coefficient is roughly similar to Spearman's rank correlation coefficient, but it mainly measures whether there is a monotonic relationship between two variables by comparing the consistency of the pairwise order of samples, and draws conclusions from another perspective. Its main calculation formula is as follows:

$$\text{tau}_b = \frac{C-D}{\sqrt{(C+D+T_X)(C+D+T_Y)}} \quad (2)$$

Where C represents the number of concordant pairs, D represents the number of discordant pairs, and T_X and T_Y are the tied pairs on X and Y respectively.

The three-variable correlation heatmap is shown in the following figure:

As can be seen from Figure 5, the correlation coefficient between gestational days for testing and BMI is 0.09, showing a weak positive correlation, and the correlation coefficient with Y-chromosome concentration is 0.06, also showing a weak positive correlation. The correlation coefficient between BMI and Y-chromosome concentration is -0.11, showing a weak negative correlation. The conclusion is the same as the previous result of the Spearman model, which can confirm the robustness of the Spearman model in testing the correlation of the three variables.

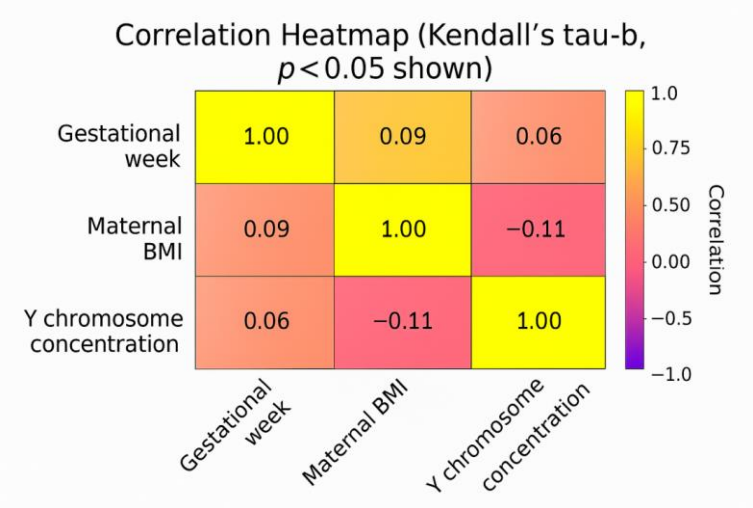

Figure 5. Kendall's tau-b Three-Variable Correlation Heatmap

Note: The color depth represents the magnitude of the correlation coefficient, with values ranging from -1 to 1. Only correlations with $p < 0.05$ are displayed.

2.4. XGBoost Regression Model

This paper first uses Support Vector Regression (SVR) for regression fitting of the data, but the obtained R^2 is too small. On this basis, we adopt the XGBoost regression model for evaluation and obtain a more ideal R^2 value. Therefore, this paper compares the test set data calculated by the two methods, and the comparison results are shown in the following table 3:

Table 3. Calculation Results of XGBoost Regression and SVR

Model	MSE	RMSE	R^2
XGBoost	0.5	0.707	0.958
SVR	11.301	3.362	0.031

Compared with SVR, XGBoost has a significantly higher R^2 and lower RMSE, and the overall fitting effect is significantly improved. Therefore, this paper will use the XGBoost regression model as the main model for regression fitting.

The prediction results of the XGBoost regression model are shown in the following table 4:

Table 4. XGBoost Regression Prediction Results

Predicted Y (Test Set)	Y-Chromosome Concentration (%)	Pregnant Women's BMI	Gestational Weeks for Testing (Days)
6.706212043762207	10.4313885	35.1384194021275	128
11.959758758544922	8.8884151	32.4255843278135	94
11.902836799621582	8.8995704	32.4296307575877	101
8.45714282989502	9.6715264	32.7096476852238	108
8.312301635742188	10.6577778	33.0673974561598	115
7.910775184631348	5.8193093	30.6293529774186	88
8.064091682434082	6.4847597	30.870736391688	95
6.598770618438721	7.3592653	31.187951821331	102
7.046818733215332	8.169636	31.4819031897525	109
5.84748649597168	11.9299526	27.8813428504966	89
6.291099548339844	12.529372	28.0987743540234	96
6.33254861831665	13.1490703	28.3235617918159	103
9.000640869140625	13.8775448	28.5878063879599	110
8.152375221252441	7.9291459	35.5478156953067	89
10.507523536682129	7.93142	35.5486405759471	96

Figure 6. Validation Residual Chart

The horizontal axis represents the true response value (figure 6), and the vertical axis represents the residual (predicted value - true value). The residuals are generally distributed around 0, with the most concentration in the middle interval, reflecting the robustness of the model in the main body.

2.5. Model Optimization

This paper intends to use the entropy weight method combined with the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) for weighting and comprehensive evaluation of the model. The formula used in the entropy weight method is as follows:

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{3}$$

Where w_j reflects the information effectiveness of the j-th indicator. The greater the dispersion, the higher the corresponding weight. $d_j = 1 - e_j$, and e_j is the information entropy of the j-th indicator.

The calculation results are as follows table 5:

Table 5. Calculation Results of the Entropy Weight Method

Item	Information Entropy e	Information Utility Value d	Weight (%)
Gestational Weeks for Testing / Day	0.964	0.036	66.453
Height / cm	0.993	0.007	13.58
Weight / kg	0.993	0.007	13.672
Pregnant Women's BMI	0.997	0.003	6.296

The weight of gestational days for testing is 66.453%, the weight of height is 13.58%, the weight of weight is 13.672%, and the weight of pregnant women's BMI is 6.296%. Among them, the maximum indicator weight is gestational days for testing, and the minimum is pregnant women's BMI. Entropy Weight Method Weight Chart is shown in figure 7.

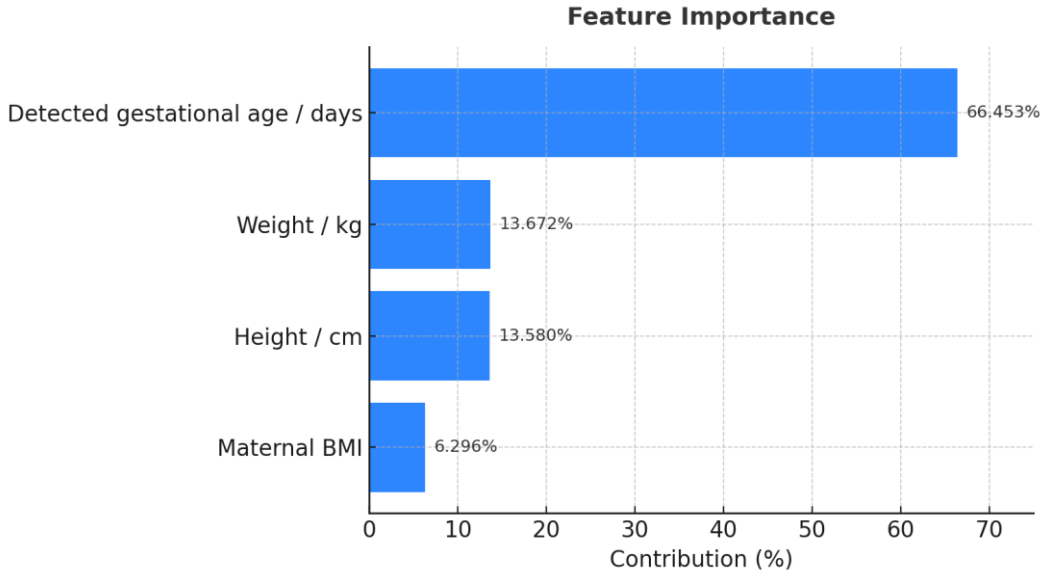

Figure 7. Entropy Weight Method Weight Chart

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) measures the distance between multiple indicators and the ideal value. On the basis of the weights of each variable obtained by the entropy weight method, the comprehensive score of each variable is calculated to better optimize the model. The calculation formulas are as follows:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} w_j (x_{ij} - x_j^+)^2} \tag{4}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} w_j (x_{ij} - x_j^-)^2} \tag{5}$$

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{6}$$

Where D_i^+ represents the distance to the positive ideal solution, and D_i^- represents the distance to the negative ideal solution.

The input results are shown in the following table 6:

Table 6. TOPSIS Calculation Results

Index Value	Distance to Positive Ideal Solution (D^+)	Distance to Negative Ideal Solution (D^-)	Comprehensive Score Index	Ranking
2.593584	0.85036833	0.27661499	0.24544728	920
3.4886856	0.68758211	0.35038297	0.3375672	536
6.6171003	0.52857154	0.49581149	0.48400986	248
6.1191623	0.44051674	0.60180815	0.57737098	141
5.9230127	0.81189509	0.21236799	0.20733734	1002
4.2401417	0.70420934	0.30781547	0.30415803	680
4.7824846	0.6036633	0.42869872	0.41526006	334
4.2674699	0.49676496	0.59323043	0.5442504	175
4.7362214	0.49676496	0.59323043	0.5442504	175
5.471245	0.79623571	0.28325129	0.26239435	855
5.6711486	0.79035252	0.28538599	0.26529308	843
6.5184837	0.78447856	0.28765038	0.26829831	831
6.3249404	0.64577025	0.37255583	0.36585121	442
5.2252975	0.51048561	0.49847437	0.49404771	234
6.6498894	0.41768362	0.61158402	0.59419338	117

According to the data in this table, when the comprehensive score is greater than or equal to 0.54, its ranking is less than or equal to 175, and the scheme should be given priority. When the comprehensive score is less than or equal to 0.27, the samples should be excluded to support the robust optimization of subsequent models.

2.6. Significance Test

This paper adopts the paired t-test to test its significance. In the same test set, the per-sample absolute error is constructed, and its calculation formula is as follows:

$$e_i^{(base)} = |y_i - \hat{y}_i^{(base)}| \; e_i^{(xgb)} = |y_i - \hat{y}_i^{(xgb)}| \tag{7}$$

Where $d_i = e_i^{(base)} - e_i^{(xgb)}$ represents the paired difference.

Table 7. Paired t-Test Results of Y-Chromosome Concentration vs. Gestational Weeks for Testing

Paired Variables	Mean ± Standard Deviation			t	Standard Deviation of Difference	p-Value	Effect Size of Difference
	Pair 1	Pair 2	Paired Difference				
Y-Chromosome Concentration vs. Gestational Weeks for Testing	0.077±0.034	117.896±28.543	-117.819±28.51	-135.797	1081	0.000***	4.128

Table 8. Paired t-Test Results of Y-Chromosome Concentration vs. Pregnant Women's BMI

Paired Variables	Mean ± Standard Deviation			t	Standard Deviation	p-Value	Effect Size of Difference
	Pair 1	Pair 2	Paired Difference				
Y-Chromosome Concentration vs. Pregnant Women's BMI	0.077±0.034	32.289±2.972	-32.212±2.939	-355.833	1081	0.000***	10.818

As shown in the above table 7 and 8, $d_i > 0$ and the t-value is far exceeding the critical value, with $p < 0.001$, and the result is significant. The above evidence shows that the average error of XGBoost is significantly smaller than that of the baseline model. The improvement is not accidental, but a statistically stable, sustainable, and reproducible improvement.

3. Conclusions

In this study, the TOPSIS method and the entropy value method complement each other, forming a progressive analysis framework from "objective weighting" to "comprehensive evaluation". By calculating the information entropy of variables, the entropy value method can objectively reflect the relative importance of each indicator in the overall evaluation, avoid the deviation of artificial subjective weighting, and is the premise and foundation for carrying out subsequent multi-indicator comprehensive evaluation. On the basis of the weight system determined by the entropy value method, the TOPSIS method realizes the comprehensive ranking and significance test of different individuals by constructing positive and negative ideal solutions and calculating the distance between each research object and the ideal solution.

The necessity of combining the two is mainly reflected in the following points: first, the entropy value method provides a scientific and reasonable weight basis for TOPSIS, ensuring the objectivity of the evaluation results; second, the introduction of TOPSIS transforms the simple weight analysis into a quantitative description of the overall performance of the object, thereby improving the interpretability and operability of the method; finally, through the combination of the two, it can not only reveal the weight differences between indicators but also verify the significance and reliability of the prediction results, providing a more solid support for the research conclusions.

Therefore, the entropy value method and the TOPSIS method have a close correlation and obvious progressiveness in this study. Their combined application is not only necessary but also effectively improves the scientificity and credibility of the analysis results.

While the current study establishes a robust framework for estimating fetal Y chromosome concentration, several avenues warrant further exploration to enhance its clinical applicability and technical maturity. Future research will focus on the following directions:

Model Generalization and Multi-Center Validation: The predictive performance and stability of the proposed XGBoost model optimized by entropy-weighted TOPSIS need to be validated using large-scale, multi-center clinical cohorts. This will assess its generalizability across diverse populations and different NIPT technical platforms.

Incorporation of High-Dimensional Features: Future work could integrate more types of maternal physiological parameters (e.g., age, parity) and molecular features from sequencing data (e.g., fragment size distribution patterns) into the model. This may further capture subtle biological relationships and improve prediction accuracy for edge cases.

Dynamic Modeling and Real-Time Prediction: Developing dynamic models that can update predictions in real-time as new longitudinal data from pregnant women becomes available represents a significant future direction. This would align better with personalized prenatal care pathways.

Interpretability and Clinical Translation: Enhancing the model's interpretability using techniques like SHAP (SHapley Additive exPlanations) is crucial for fostering clinical trust and adoption. Furthermore, translating the model into user-friendly clinical decision support software could facilitate its practical application in assisting physicians to determine the optimal timing for NIPT.

Expansion to Other Prenatal Markers: The methodological framework proposed in this study has the potential to be adapted for the precise estimation of other fetal-specific biomarkers in maternal plasma, such as the concentration of fetal DNA or specific epigenetic markers, thereby broadening its impact in the field of non-invasive prenatal diagnosis.

Through continuous optimization and validation, this research approach is expected to provide more precise and reliable technical support for clinical prenatal screening and diagnosis.

References

[1] Owen A. "Do you want to know or not?" How prenatal providers manage clinical uncertainty related to chromosomal risk and noninvasive prenatal testing. [J]. Health (London, England: 1997),2025,13634593251377111.

[2] Zhou W, Liu F, Li S, et al. Novel Algorithm for Monogenic Noninvasive Prenatal Testing With Highly Similar Parental Pathogenic Haplotypes: A Representative Case of Congenital Adrenal Hyperplasia Pedigree[J]. Human Mutation,2025,2025(1):9990873-9990873.

[3] Yamamoto K, Suzumori N, Miura K, et al. Clinical Implications of Low Cell-Free DNA Fetal Fraction in Non-Invasive Prenatal Testing: A Retrospective Cohort Study of 40,716 Pregnancies. [J]. Prenatal diagnosis,2025.

[4] Cohen M S, Chen M, Sun L. Editorial: Advancements in prenatal diagnosis: from noninvasive prenatal tests to novel fetal imaging[J]. Frontiers in Medicine,2025,121682161-1682161.

[5] Yamashita Y, Shirato N, Ishii T, et al. Perinatal outcomes of fetal CNVs detected by genome-wide non-invasive prenatal testing in Japan[J]. Journal of Human Genetics, 2025, (prepublish):1-9.

[6] Sun F, Zhou Y, Zhao X, et al. Application of an optimized non-invasive prenatal testing for thalassemia based on change of haplotype doses. [J]. Journal of genetics and genomics = Yi chuan xue bao,2025.

[7] Non-Invasive Prenatal Testing Market Poised for Consistent Growth Through 2032, Fueled by Technological Advances and Rising Procedural Demand | DelveInsight[J].M2 Presswire,2025,

[8] O'Keeffe R, Doyle S. Non-invasive prenatal testing: Navigating the ethical, legal, and social frontiers of modern obstetric practice. [J]. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics,2025.

[9] Eekhout V A C J, Bax J C, Schuurman P V L, et al. Performance of non-invasive prenatal testing in vanishing-twin and multiple pregnancies: results of TRIDENT-2 study. [J]. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology,2025.

[10] Qi H, Chen H, Zhang Z, et al. Performance and clinical implications of non-invasive prenatal testing for rare chromosomal abnormalities: a retrospective study of 94,125 cases[J]. Frontiers in Molecular Biosciences,2025,121645223-1645223.