Journal of *Marine Science and Engineering*

MDPI

Article

Predicting the Sound Speed of Seafloor Sediments in the East China Sea Based on an XGBoost Algorithm

Mujun Chen [1,2], Xiangmei Meng [1,2], Guangming Kan [1,2,*], Jingqiang Wang [1,2], Guanbao Li [1,2], Baohua Liu [2], Chenguang Liu [1,2], Yanguang Liu [1,2], Yuanxu Liu [1,3] and Junjie Lu [1,2]

[1] Key Laboratory of Marine Geology and Metallogeny, First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China
[2] Laboratory for Marine Geology, Pilot National Laboratory for Marine Science and Technology (Qingdao), Qingdao 266237, China
[3] College of Underwater Acoustic Engineering, Harbin Engineering University (Harbin), Harbin 150001, China
* Correspondence: kgming135@fio.org.cn

Abstract: Based on the acoustic and physical data of typical seafloor sediment samples collected in the East China Sea, this study on the super parameter selection and contribution of the characteristic factors of the machine learning model for predicting the sound speed of seafloor sediments was conducted using the eXtreme gradient boosting (XGBoost) algorithm. An XGBoost model for predicting the sound speed of seafloor sediments was established based on five physical parameters: density (ρ), water content (w), void ratio (e), sand content (S), and average grain size (M_z). The results demonstrated that the model had the highest accuracy when n_estimator was 75 and max_depth was 5. The model training goodness of fit (R^2) was as high as 0.92, and the mean absolute error and mean absolute percent error of the model prediction were 7.99 m/s and 0.51%, respectively. The results demonstrated that, in the study area, the XGBoost prediction method for the sound speed of seafloor sediments was superior to the traditional single- and two-parameter regressional equation prediction methods, with higher prediction accuracy, thus providing a new approach to predict the sound speed of seafloor sediments.

Keywords: sound speed; seafloor sediments; XGBoost; the East China Sea

Citation: Chen, M.; Meng, X.; Kan, G.; Wang, J.; Li, G.; Liu, B.; Liu, C.; Liu, Y.; Liu, Y.; Lu, J. Predicting the Sound Speed of Seafloor Sediments in the East China Sea Based on an XGBoost Algorithm. *J. Mar. Sci. Eng.* **2022**, *10*, 1366. https://doi.org/10.3390/jmse10101366

Academic Editor: George Kontakiotis

Received: 7 August 2022
Accepted: 19 September 2022
Published: 24 September 2022

1. Introduction

The shallow sediments of the seafloor exhibit unique acoustic properties that provide the necessary basic data for seafloor acoustic field research, seafloor engineering geology, and marine petroleum geology and are important factors in determining the marine acoustic field environment [1]. They have important research value in the fields of seafloor sediment investigation, marine resource exploration and development, and marine environmental monitoring [2,3]. Seafloor sediments are generally considered a two-phase medium consisting of solid and liquid phases [4], and their acoustic properties are closely related to the physical properties of seafloor sediments.

As the basic element of seafloor acoustics research, the measuring methods of the sound speed of seafloor sediments primarily include in situ measurements, laboratory measurements, and the geoacoustic inversion method. In addition, prediction based on geoacoustic models is an important method for obtaining the sound speed of seafloor sediments. Therefore, it is extremely important to establish an accurate geoacoustic model to describe the relationship between the sediment sound speed and physical parameters. Many studies established theoretical models for predicting sound speed in seafloor sediments [5–11]. However, because of the complex and diverse marine sedimentary environment and sedimentary disturbance error in the determination process, the several input parameters of the theoretical model are difficult to measure. Concomitantly, many studies have established regressional equations of sound speed prediction in different sea areas by

examining the correlation between the sound speed and physical parameters of seafloor sediments [12–22]. However, the physical properties of sediments frequently require multiple parameters for accurate characterization. The empirical equation only establishes the correlation between the sediment sound speed and one or two physical parameters to an extent, and there are limitations and reduced accuracy in the sound speed prediction.

To supplement the existing literature, this study proposes a machine learning prediction method for the sound speed of seafloor sediments based on the eXtreme gradient boosting (XGBoost) algorithm, which can fuse multiple sediment physical parameters as well as effectively improve the problem of the previous empirical equations tending to overfit or underfit the measured sound speed of seafloor sediment Here, data preprocessing was performed on the data measured at laboratory of sediment samples from the East China Sea. The physical parameters with high correlation were then extracted using the Pearson correlation coefficient matrix, and the model was trained and verified. Then, using the test samples and traditional single- and two-parameter regressional equations to comprehensively evaluate the model, we analyzed the contribution of each physical parameter. The results demonstrated that the prediction model of the seafloor sediment sound speed based on the XGBoost algorithm was superior to the traditional prediction equation methods with higher prediction accuracy.

2. Study Area and Data Source

2.1. Location of the Study Area

The study area is located on the continental shelf of the East China Sea at a water depth of 26–107 m. The water depth of the western-inland region is shallow and gradually increases to the east. As one of the widest continental shelves worldwide, it has considerable terrigenous input and is an important area for studying land–ocean interactions and source–sink processes [23]. The deposition of terrigenous sediments in this area is primarily controlled by the coastal upwelling and downwelling of the southern continental shelf of the East China Sea [24]. Owing to the inflow of small coastal rivers such as the Yangtze River and the influence of the Yellow River, the continental shelf of the East China Sea has received a high input of terrigenous materials [25]. In the interaction between the Zhejiang–Fujian coastal current and Taiwan warm current, most sediments diffused from the north to the south were confined in the continental shelf, thus forming an intrashelf mud wedge [26]; however, most of the sediments in the study area were confined in the continental shelf and covered by sand [27]. The coastal sediments in the study area are parallel to the coast and distributed in a band, and the nearshore is mostly silty clay, whereas the outward is clay. The sediment then quickly changes to coarse-grained silt or fine sand.

2.2. Data Sources

Seafloor sediment samples were collected at 45 sites on the East China Sea shelf using box and gravity samplers during the open research cruise of the East China Sea supported by the National Science Foundation of China (NSFC) Shiptime Shearing Project in 2021. Among these samples, those that were ~400 cm in length were obtained from 16 stations in the East China Sea's silty clay and clayey silt areas, while most obtained from the sandy bottom in the study area were 20–200 cm in length. The sound speed and physical parameters were measured in the laboratory. First, the sample was cut and divided as per the actual requirements, generally cut into 30 cm long sections and placed on a cylindrical sample measurement platform. The laboratory measurement system of the acoustic property of seafloor sediment cores was used to measure the sound speed of the sediment samples using the transmission method. After the 30 cm long sediment measurement was completed, a length of 10 cm was cut off, and the sound speed of the remaining 20 cm long sediment was measured again to obtain as much data as possible. The sound speed was calculated as follows:

$$V_P = L / \left(10^3 \left(t - t_0\right)\right) \quad (1)$$

where V_P is the sound speed of the seafloor sediment sample (m/s), L is the column length of the sample (mm), t is the sound wave penetration time (μs), and t_0 is the correction value of zero sound time (μs).

After measuring the sound speed of the sample, the physical and mechanical properties of the sediment were measured [28]. The properties included density, water content, void ratio, sand content, silt content, clay content, and average grain size. During the laboratory measurements, two-frequency (25 and 100 kHz) acoustic transducers were used to measure the acoustic properties and physical parameters of each section of the sample. Because the acoustic data with a frequency of 100 kHz were of good quality and the covered physical parameters were more representative, although the sound speed values at different frequencies differed, the relationship between them and the physical properties of the sediments was the same. Thus, this study selected sound speed data of 100 kHz (292 groups). Table 1 lists the maximum, minimum, and average of the physical parameters and sound speeds.

Table 1. Statistics of sound speed and physical parameters of sediments in the study area.

	ρ/ (g/cm^3)	w/ %	e	S/ %	T/ %	Y/ %	M_z/ Φ	V_P/ (m/s)
Max	2.00	74.85	2.04	76.30	79.30	73.10	8.78	1695.38
Min	1.56	24.25	0.68	0.10	10.70	34.50	6.71	1492.86
Ave	1.72	52.07	1.43	10.34	55.15	7.60	5.9	1540.96

3. Methods

3.1. Data Preprocessing

3.1.1. Data Noise Removal

The experimental data were collected from a real scene and contained a lot of data noise; therefore, they could not be directly used for model training. The primary task of data noise removal is to remove incomplete or wrong data. Here, data cleaning was divided into outlier and missing value removal. Data points significantly far from the fitted curve are marked as outliers using a regression fit between the sound speed and each physical parameter. The missing sound speed values and physical parameters of each sample were simultaneously removed to improve the data quality and integrity.

3.1.2. Normalization Processing

The numerical units of different physical parameters differ; therefore, the data required normalization. The parameters were uniformly set to a value between 0 and 1. Normalizing the input data can prevent neuron saturation and increase the accuracy and generalization ability of the model prediction. The formula of normalization is as follows:

$$\overline{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

where $\overline{X}$ is the normalized data, X is the original data, and $X_{\min}$ and $X_{\max}$ are the maximum and minimum values of the original data, respectively.

3.1.3. Physical Parameter Extraction

In order to reduce the complexity of the model and improve the generalization ability of the model, some but not all parameters with relatively strong correlation were selected while ensuring accuracy during the data processing. Here, the data collected included the sound speed, basic information of the following sampling stations, and physical parameters of the sediments:

1. Geospatial information of the seafloor sediment sampling stations—longitude (Log), latitude (Lat), and depth (D);
2. Basic physical parameters—density (ρ), water content (w), and void ratio (e);

3. Grain composition—sand (S), silt (T), clay contents (Y);
4. Grain size coefficient—average grain size (M_z).

Among the acoustic parameters and physical and mechanical parameters of seafloor sediments, the sound speed exhibits a good correlation with the density, water content, and porosity [18]. The Pearson correlation coefficient can measure the degree of correlation and whether there is a linear correlation between two features [29]. Thus, the Pearson correlation coefficient was used to measure the correlation between the sound speed and other physical parameters and plot the correlation coefficient matrix (Figure 1). Using two variables—X and Y—the Pearson correlation coefficient between the variables is as follows:

$$f_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) + E^2(X)}\sqrt{E(Y^2) + E^2(Y)}} \quad (3)$$

where $f_{X,Y}$ is the correlation coefficient. A positive value indicates a positive correlation between the physical parameters and sound speed, and a negative value indicates a negative correlation.

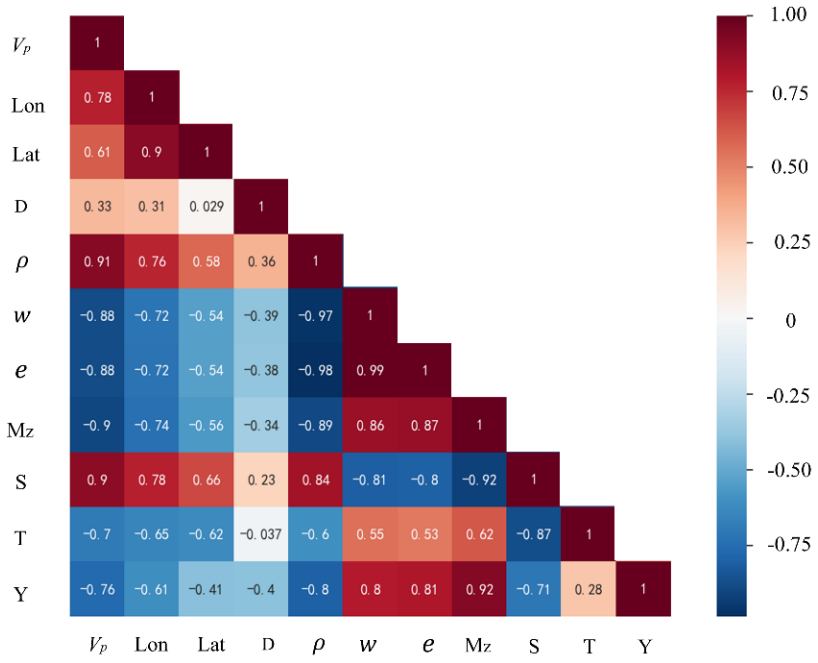

Figure 1. Half−matrix plot of Wilson's correlation coefficients between the sediment sound speed and physical parameters. The first column of the matrix is the distribution of the correlation coefficient between the sound speed and each physical parameter. The redder the matrix unit, the more evident the positive correlation, and the bluer the unit, the more evident the negative correlation.

As per the first column of the matrix diagram of the correlation coefficient between the sound speed and each physical parameter (Figure 1), it was concluded that the correlation coefficient between the sound speed and density, water content, void ratio, sand content, and average particle size of the seafloor sediments reached ~0.90, exhibiting strong or extremely strong correlation. The correlation coefficient with other parameters was below 0.80. During the training of the XGBoost algorithm model, the highest correlation between the input feature factor and target parameter was selected such that the predicted result was closer to the expected value. Similarly, the dimension of the model was reduced as much as possible while ensuring the accuracy of the model to reduce the complexity of the model. Therefore, the factors selected for training in this study were the density, water content, void ratio, sand content, and average grain size. Among these factors, the density and sand content positively correlated with the sound speed, indicating that the higher the density and sand content, the higher the sound speed. The water content, void ratio,

and average grain size negatively correlated with the sound speed, indicating that with an increase in each parameter, the sound speed decreased.

3.1.4. Data Division

After data preprocessing, 280 data sample groups were collected, and the samples were randomly divided into 200 training, 40 validation, and 40 test sample groups. After the training, validation, and test sets were determined, only then were the parameters of the learning algorithm adjusted to explore suitable parameters, screen suitable features, rapidly detect the algorithm performance, and guide the most important changes to the machine learning model. The training samples were used to train the sound speed prediction model, and the final model was trained by setting the parameters of the fitter. After training multiple models via the training set, the model with the best effect was selected from the validation set. The corresponding parameters can then be used to control the occurrence of model over-fitting. To measure the performance of the optimized model, the test samples were considered as a nonexistent data set—a data sample that did not participate in the entire model building process and was used to measure the performance of the optimized model. Figure 2 shows the training, validation, and test samples in the sediment triangulation. There were multiple types of sediments in the data samples, which improved the generalization ability of the model.

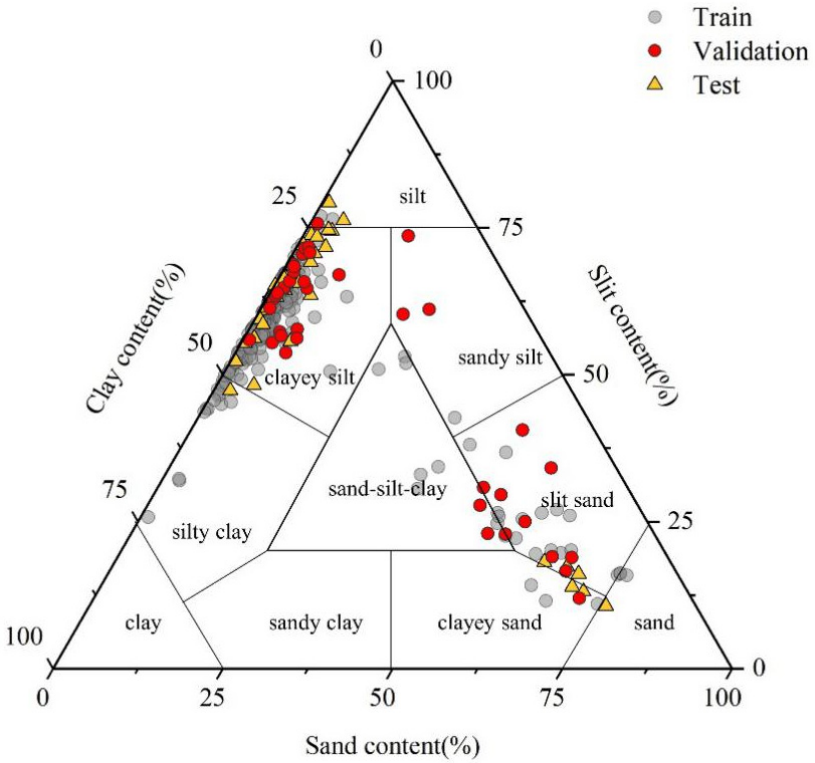

Figure 2. Triangular sediment classification diagram. Most of the sediment in the data samples included clay silt, followed by silt sand, a small amount of silt, silty clay, and clay sand. The grey dots, red dots, and yellow triangles indicate training, validation, and test samples, respectively.

3.2. XGBoost Algorithm

XGBoost is a boosted tree model that integrates multiple weak learners to build a strong learner [30,31]. The idea of the algorithm is to fit the negative gradient of the loss function in repeated iterations after optimizing the empirical loss function, select sample features to generate a basic learner, and continuously fit the previous residuals to minimize the objective function. We repeated this process to build hundreds of basic learners and integrated them in a comprehensive model (Figure 3).

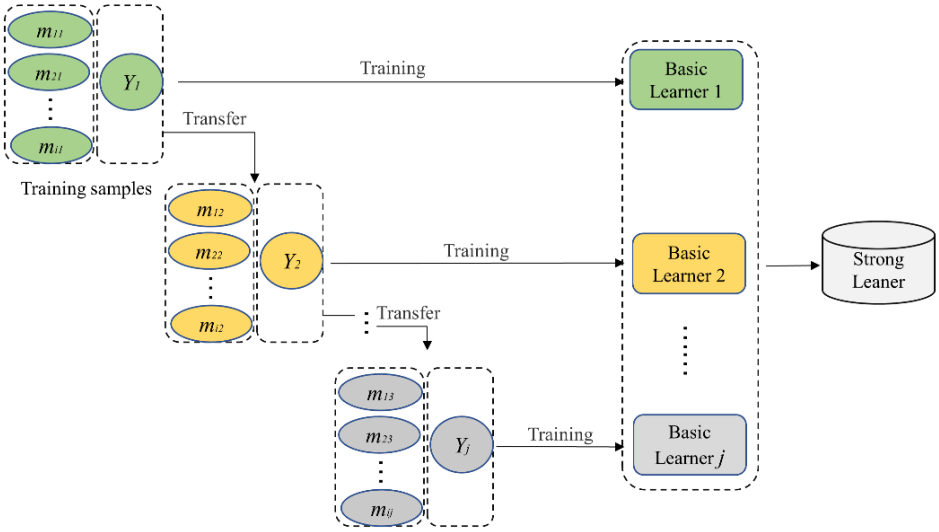

Figure 3. The boosting algorithm principle, where m is the characteristic factor of the training sample in the model, which is each physical parameter; m_{ij} is the i-th characteristic factor of the j-th base learner; and Y represents the target value of the training sample in the model, which is the sound speed of the seafloor sediment in this study.

The objective function of XGBoost comprises a loss function, regularization term, and constant term. The equation is as follows:

$$\mathrm{Obj}(\theta) = L(\theta) + \Omega(\theta) + C \tag{4}$$

The loss function was used to measure the quality of the model prediction, and the regularization term was used to control the complexity of the model and avoid overfitting.

3.2.1. Loss Function

The XGBoost algorithm can be considered as an additive model comprising K trees, as shown in Equation (5). The tree model used in this study is a regression tree:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{5}$$

where $\hat{y}_i$ is the sample prediction result, x_i, i is the i-th sample input, k is the sum of trees, f_k is each regression tree, and F is the set space with the regression tree.

The improvement of XGBoost over the Gradient Boosting Decision Tree (GBDT) is that it uses the first- and second-order derivatives and the Taylor expansion for approximation. If g_i is the first derivative and h_i is the second derivative:

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{6}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{7}$$

3.2.2. Regularization

The complexity of the regression tree can effectively control the overfitting of the model. It comprises two parts: the number and weight of leaf nodes. It was defined by the equation as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

where γ and λ are the normalization coefficients. Combining the two aforementioned parts, the objective function was rewritten as the following equation:

$$\mathrm{Obj}(\theta) \approx \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \tag{9}$$

where I_j is the sample on the j-th leaf node, and w_j is the weight of the j-th leaf node. The optimal objective function was then obtained:

$$\mathrm{Obj}(\theta) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j + \lambda} + \gamma T \tag{10}$$

The smaller the objective function, the smaller the prediction error, and the better the model effect.

4. Results

4.1. Training and Validation of Seafloor Sediment Prediction Model

This study was based on the XGBoost algorithm with the CART (Classification and Regression Tree) as the base learner. The 200 sets of training data obtained in the laboratory were used for model training, and the 40 sets of validation samples were used to assist the model training to obtain optimal parameters such that the model performance could reach the highest level. The selected training target parameter was the sediment sound speed; the characteristic factors were the density, water content, void ratio, average grain size, and clay content. During the model training process, the XGBoost algorithm had multiple hyperparameters, and it was impossible to adjust all the parameters. Therefore, this study optimized two parameters—n_estimator and max_depth, which were important to the model training accuracy—and the parameter adjustment and optimization followed the principle of "first importantly then weak, first coarse then fine." Here, the mean absolute error (MAE) and mean absolute percentage error (MAPE) were selected to compare the performance of different models. The equations of the two indicators are as follows:

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \tag{11}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{12}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of predicted samples.

The n_estimator in the XGBoost hyperparameters, i.e., the maximum number of trees, can be considered as the number of iterations that determines the learning ability of the model. The greater the number of trees, the stronger the learning ability of the model. When the number of trees in the XGBoost was low, the impact on the model was relatively high; when the number of trees was already high, the impact on the model was relatively low. The smaller the trees, the easier it was to cause the model to under-fit, and excess trees easily caused the model to over-fit. Therefore, selecting the appropriate n_estimator can impact the training accuracy of the model. Here, the optimal selection of the n_estimator was made within the range of [20, 200]. As shown in Figure 4, when the n_estimator exceeded 75, the training error of the model continually decreased, and the error of the validation sample gradually increased. Therefore, it was determined that when the n_estimator exceeded 75, the model was in the overfitting state, and so the n_estimator of the research model was 75, and the model effect was the best. At this time, the goodness of fit (R^2) of the model training was 0.917, the real value of the MAE of the training result was 8.15 m/s, and that of the validation result was 9.11 m/s.

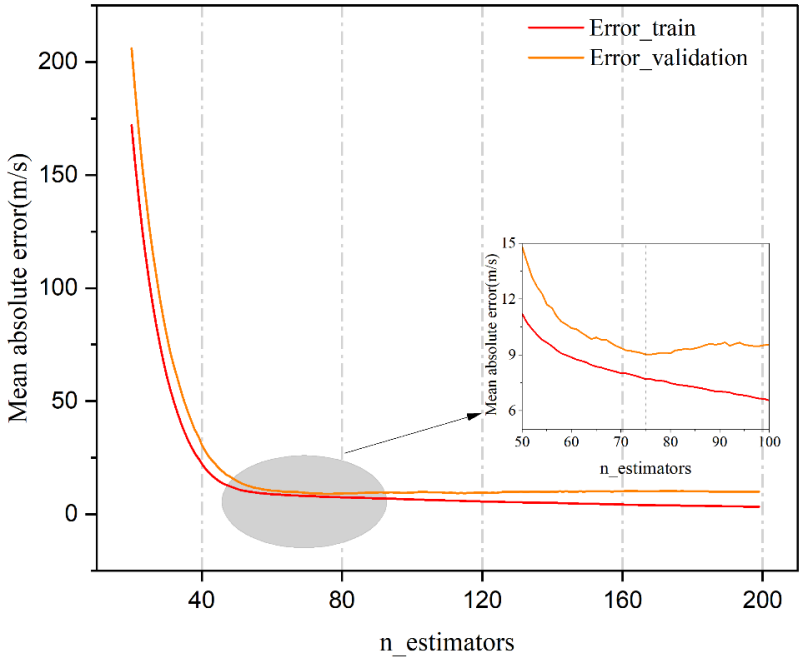

Figure 4. The training process of the XGBoost model is based on the hyperparameter—n_estimator. The red and orange solid lines indicate the iterative changes in the model training error and validation result, respectively.

The max_depth is the maximum depth of the tree in the model, which is used to control the model fitting state. The larger the max_depth, the more specific the model learns; however, if the max_depth is excessively large, the model overfits. Based on the premise that A in the model was 75, this study set the max_depth to perform optimal model training in [1,20]. As shown in Figure 5, when the max_depth was 5, the model had the highest validation accuracy. At this time, the R^2 of the model training was 0.923, the real value of the MAE of the training result was 7.79 m/s, and that of the validation result was 8.96 m/s.

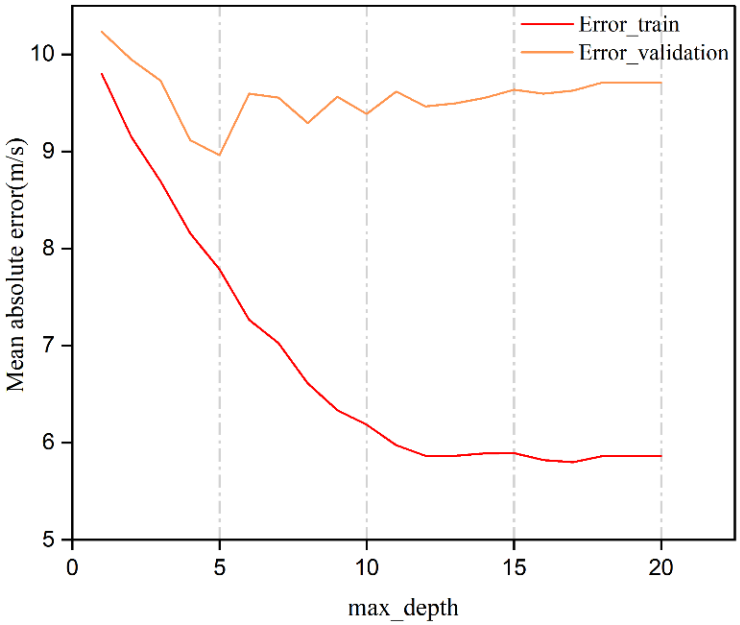

Figure 5. The training process of the XGBoost model was based on the hyperparameter—max_depth. The red and orange solid lines indicate the iterative changes in the model training and validation result errors, respectively.

After the continuous optimization of the model, along with the error performance of the training and validation results, the optimal model had an n_estimator and max_depth of 75 and 5, respectively. Figure 6 shows that the training results of the 200 training sample groups and validation results of 40 validation sample groups had a high degree of fit with the real values.

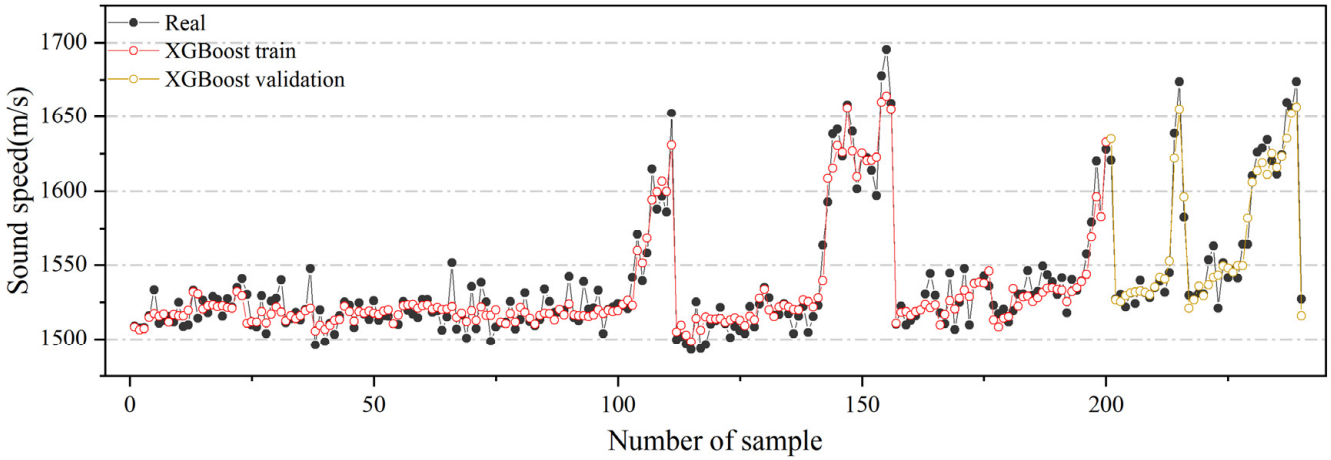

Figure 6. Training and validation results of the final model. The black line is the actual measured value, the red line is the training result of 200 sets of training samples, and the brown line is the validation result of 40 sets of validation samples.

4.2. Model Interpretation

SHapley Additive exPlanations (SHAP) can be used to explain the output of the machine learning model and build an additive explanation model, and all features are regarded as "contributors." For each predicted sample, the model produced a predicted value, and the SHAP value was the value assigned to each feature in that sample. If the i-th sample was x_i, the j-th feature of the i-th sample was x_{ij}, the predicted value of the model for the sample is y_i, the baseline of the entire model (usually the mean of the target variable of the samples) is y_{base}, and then the SHAP value obeyed the following equation:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ij}) \quad (13)$$

where $f(x_{ij})$ is the SHAP value of x_{ij}. Intuitively, it was the contribution value of the first feature in the i-th sample to the final predicted value. When $f(x_{i1}) > 0$, it implied that the feature improved the predicted value, showing a positive effect; otherwise, it showed that the feature reduced the predicted value, showing a negative effect.

The characteristic factors of the prediction model of the seafloor sediment sound speed in this study were the density, water content, void ratio, sand content, and average particle size. Figure 7 shows the characteristic factors arranged from top to bottom, and the successive decrease in their contribution to the model, i.e., the contribution of the physical parameters to the sound speed in this model. The degree of influence is arranged in terms of the water content, density, sand content, average particle size, and void ratio. Moreover, Figure 7 shows that the sound speed increased with an increase in the density and sand content and decreased with an increase in the water content, void ratio, and average grain size.

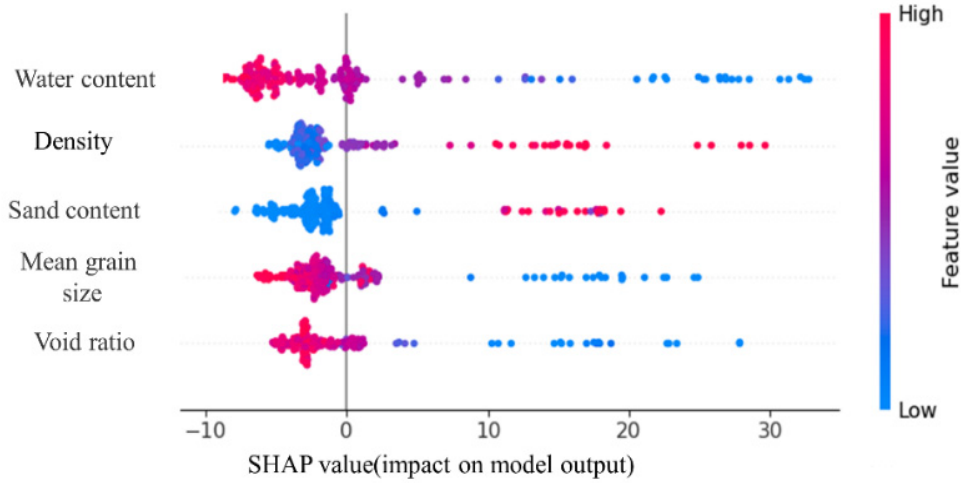

Figure 7. SHAP value distribution of characteristic factors in the model, which shows the contribution of the sample point of each characteristic factor to the model.

5. Discussion

5.1. Single-Parameter Prediction Equation

Based on the 200 training sample groups, a regression analysis was performed between the sound speed and each parameter—density, water content, void ratio, sand content, and average grain size. In previous relevant studies, the traditional single parameter method for sediment sound speed is basically dominated by the unary quadratic equation [16,18,19], and in our single parameter regression analysis and comparison, the fitting effect of the single parameter using the quadratic equation is the best. For example, it is found that the fitting degree of the cubic equation established by water content and average grain size is lower than that of the quadratic equation, and the fitting degree of the cubic equation of other parameters is not improved too. So, the unary quadratic regression method was employed to obtain the single parameter equation of the sound speed prediction (Table 2). The results demonstrated that the correlation coefficient between the physical parameters and sound speed was high ($R^2 \geq 0.76$); the density and void ratio had the highest correlation coefficient of 0.86, followed by the water content ($R^2 = 0.85$), sand content ($R^2 = 0.76$), and average grain size ($R^2 = 0.76$). Figure 8 shows the fitting curve of the sound speed and physical parameters.

Table 2. Expressions of the single-parameter equations and fitting correlation coefficients.

Related Parameters	Prediction Equation	R^2
ρ	$V_P = 1141.54\rho^2 - 3642.28\rho + 4419.83$	0.86
w	$V_P = 0.088w^2 - 11.74w + 1903.64$	0.85
e	$V_P = 123.96e^2 - 448.07e + 1916.74$	0.86
M_z	$V_P = 7.90{M_z}^2 - 136.04M_z + 2086.99$	0.76
S	$V_P = 2.97s - 0.015s^2 + 1515.95$	0.76

5.2. Two-Parameter Prediction Equation

Based on the 200 training sample groups, the bivariate quadratic regression prediction equation of the sound speed and two of the parameters among density, water content, void ratio, and average grain size was obtained using the principle of least square (Table 3). The results demonstrated that the correlation coefficient between the tow physical parameters and sound speed was high ($R^2 \geq 0.86$), the correlation coefficient of the equation obtained using the density and average grain size-based equation, density and void ratio-based equation, density and water content-based equation is 0.87, and water content and void ratio-based equation is 0.86. Figure 9 shows the 3D spatial distribution between the sound speed and parameters.

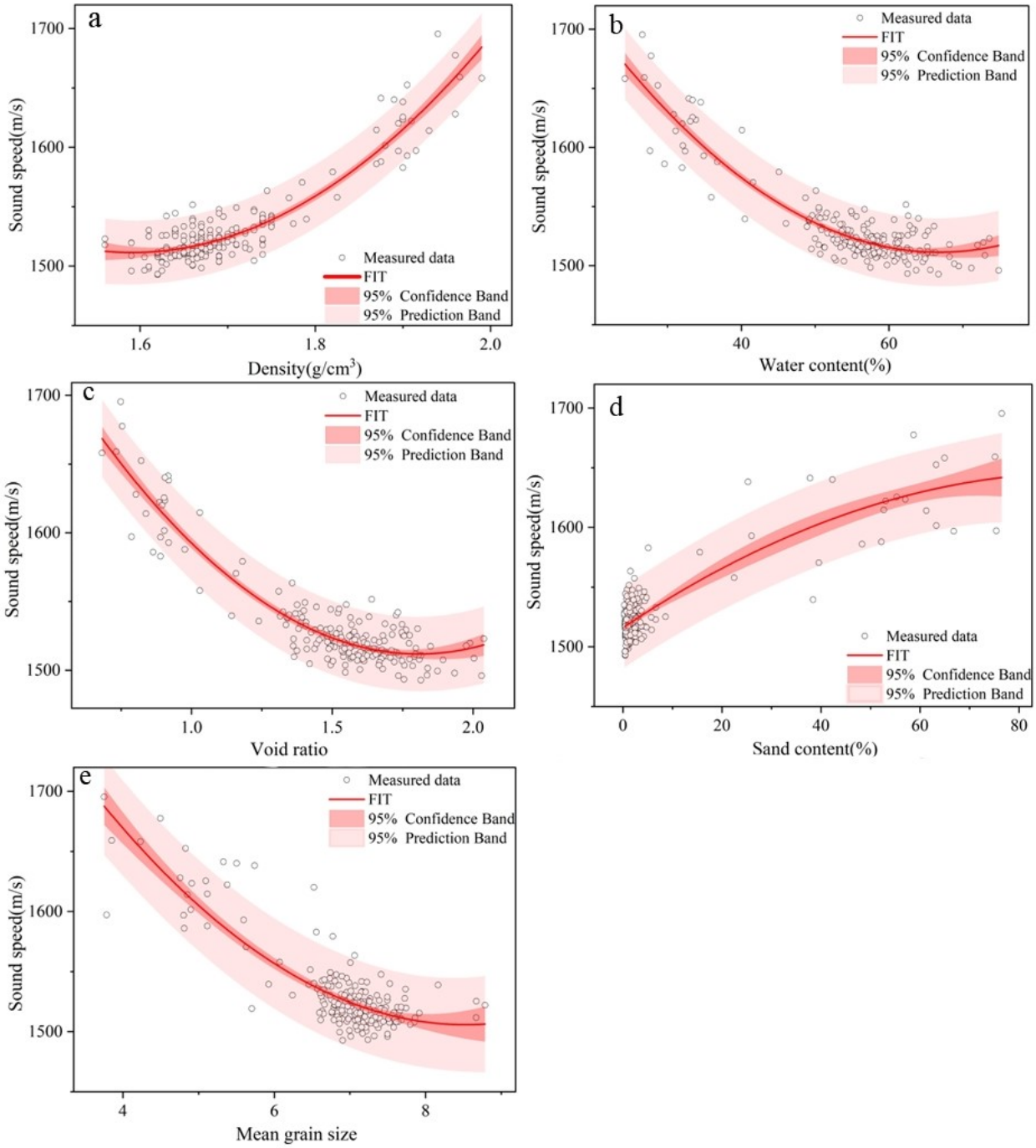

Figure 8. The single-parameter equation regression fitting of the sound speed and physical parameters: (**a**) density, (**b**) water content, (**c**) void ratio, (**d**) sand content, and (**e**) average grain size curves. The black hollow point, red solid line, red area, and pink area indicate the 200 training sample groups, fitted curve, confidence interval at 95%, and prediction interval at 95%, respectively.

Table 3. Expressions of double-parameter equations and fitting correlation coefficients.

Related Parameters	Prediction Equation	R^2
ρ, w	$V_P = 966.58\rho^2 + 0.012w^2 - 3123.61\rho - 1.95w + 4111.86$	0.87
ρ, e	$V_P = 852.79\rho^2 + 34.85e^2 - 2804.28\rho - 146.72e + 3968.22$	0.87
w, e	$V_P = 258.35e^2 - 0.099w^2 - 898.03e + 12.13w + 1921.24$	0.86
ρ, M_z	$V_P = 917.54\rho^2 + 1.3{M_z}^2 - 2904.8\rho - 18.79M_z + 3889.6$	0.87

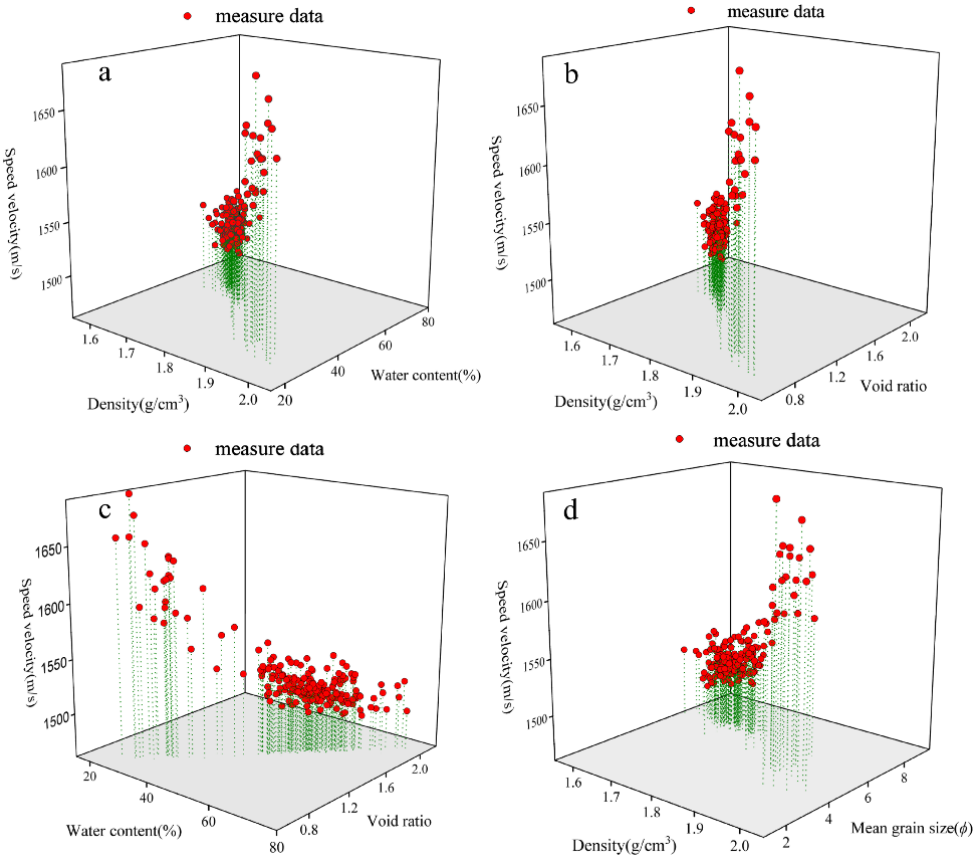

Figure 9. Three-dimensional scatter distribution of the sound speed and two parameters: (**a**) density and water content, (**b**) density and void ratio, (**c**) water content and void ratio, and (**d**) density and average grain size. The red dots indicate the 200 sets of training samples, and the green dotted lines indicate the projections of the dots on the coordinate plane formed by the parameters.

5.3. Comparison of XGBoost Prediction Models with Predictions of Single-and Two-Parameter Equations

Based on the trained XGBoost model, single-parameter prediction equations, and two-parameter prediction equations, the sound speed of the seafloor sediments was predicted for 40 groups of test samples to confirm the accuracy and superiority of the model (Figure 10). Table 4 shows the correlation error statistics between the prediction results and true values of each model. The results demonstrated that the XGBoost model exhibited the highest prediction accuracy, and the MAE, MAPE, and max absolute error are 7.99 m/s, 0.51%, and 29.27 m/s, respectively. This was followed by the density–void ratio, density, average grain size, void ratio, water content, density–water content, water content–void ratio, density–average grain size, and sand content equation. Compared with the prediction results of the traditional single- and two-parameter equations, the XGBoost model reduced the MAE and MAPE by 2.47–7.73 m/s and 0.16–0.49%, respectively. At the same time, it is found that the XGBoost model has a good performance in decreasing the max absolute error and max absolute percentage error. Compared with the traditional single- and two-parameter equation, the max absolute error and max absolute percentage error decreased by 6.54–19.56 m/s and 0.47–1.06%, respectively. Especially the max absolute error of specific density, water content, average grain size, density–water content, density–average grain size, and water content–void ratio equation decreased by more than 10 m/s. It is proven that the model has better performance in controlling errors and improving the prediction accuracy of the sediment sound speed.

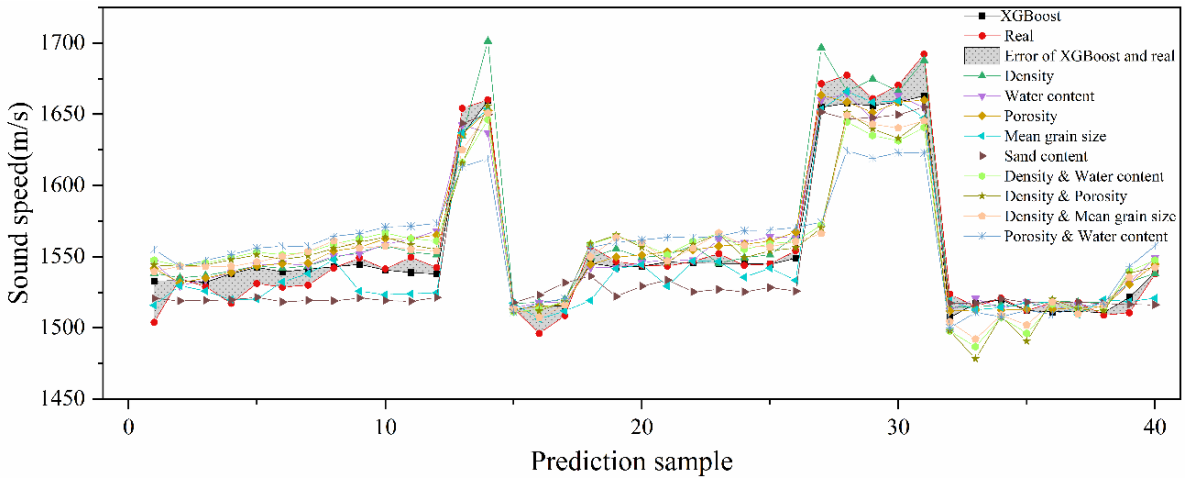

Figure 10. Comparison of prediction results of 40 groups of test samples of the sediment prediction model of the XGBoost algorithm and single- and two-parameter equations. The red dotted line indicates the real sound speed value, the black dotted line indicates the prediction result of the XGBoost model, the gray area indicates the error range between the XGBoost model and the real value, and the prediction results of the single-and two-parameter equations.

Table 4. The error analysis of the prediction results of the XGBoost model and single- and two-parameter equations.

Prediction Model	Max Absolute Error (m/s)	Max Absolute Percentage Error (%)	MAE (m/s)	MAPE (%)
ρ	41.29	2.49	10.53	0.67
w	42.03	2.79	12.30	0.79
e	37.88	2.52	11.97	0.77
M_z	48.83	2.65	10.57	0.67
S	37.24	2.20	15.72	1.00
ρ, w	39.36	2.38	13.89	0.89
ρ, e	35.81	2.37	10.46	0.67
w, e	39.39	2.62	14.36	0.93
ρ, M_z	41.42	2.75	14.77	0.95
XGBoost	29.27	1.73	7.99	0.51

6. Conclusions

Here, the sound speed prediction of seafloor sediments in the East China Sea was conducted based on the XGBoost algorithm. The optimal model super parameters were determined using 240 groups of samples from the East China Sea. Finally, the prediction accuracy of 40 groups of test samples was compared with the traditional single- and two-parameter regressional equations, and the contribution degree of the characteristic factors of the model was studied. The main conclusions of this study are as follows:

1. The XGBoost machine learning method exhibited high prediction accuracy and generalization ability when applied to the prediction of the sound speed of sediments in the East China Sea. When the n_estimator of the model was 75 and the max_depth was 5, the performance of the model was excellent, the goodness of fit (R^2) was 0.923, the MAE of the training results and true values was 7.79 m/s, and the MAE of the validation results and true values was 8.96 m/s.
2. Compared with the traditional single- and two-parameter models, the seafloor sediment model exhibited a higher goodness-of-fit and prediction accuracy. The MAE, MAPE, max absolute error, and max absolute percentage error of the prediction results were 7.99 m/s, 0.51% and 29.27 m/s, 1.73%, respectively, which were 2.47–7.73 m/s, 0.16–0.49%, 6.54 m/s–19.56 m/s, and 0.47–1.06% lower than those of the traditional

single- and two-parameter equations. It is proven that the model has better performance in controlling error and the prediction accuracy of the sound speed of the seafloor sediment improved.

Author Contributions: Conceptualization, G.K.; methodology, M.C.; software, X.M.; validation, M.C.; formal analysis, M.C.; investigation, X.M. and G.K.; resources and data curation, J.W.; writing—original draft preparation, M.C., Y.L. (Yuanxu Liu) and J.L.; writing—review and editing, G.K., M.C. and B.L.; visualization, G.L.; supervision, C.L. and Y.L. (Yanguang Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China under Grant [numbers 42049902, 41676055, 41706062]; the Central Public-Interest Scientific Institution Basal Research Fund under Grant [number GY0220Q09].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Data acquisition and sample collections were supported by the National Natural Science Foundation of China Open Research Cruise (Cruise No. NORC2021-02+NORC2021-301), funded by the Shiptime Sharing Project of the National Natural Science Foundation of China. This cruise was conducted onboard R/V "XiangYangHong 18" by The First Institute of Oceanography, Ministry of Natural Resources, China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pan, G. *Research on the Acoustic Characteristics of Seafloor Sediments in the Northern South. China Sea*; Tongji University: Shanghai, China, 2003.
2. Jin, X.L. The development of research in marine geophysics and acoustic technology for submarine exploration. *Prog. Geophys.* **2007**, *22*, 1243–1249.
3. Wen, D.D.; Wang, N.; Zhou, X.H. Interdisciplinary study of acoustics and marine sedimentology. *Adv. Mar. Sci.* **2006**, *24*, 392–396.
4. Zhu, Z.Y.; Wang, D.; Zhou, J.P.; Wang, X.M. Acoustic wave dispersion and attenuation in marine sediment based on partially gas-saturated Biot-Stoll model. *Chin. J. Geophys.* **2012**, *55*, 180–188.
5. Biot, M.A. Theory of propagation of elastic waves in a fluid-saturated porous solid: II. Higher frequency range. *J. Acoust. Soc. Am.* **1956**, *28*, 179–191. [CrossRef]
6. Buckingham, M.J. Wave propagation, stress relaxation, and grain-to-grain shearing in saturated, unconsolidated marine sediments. *J. Acoust. Soc. Am.* **2000**, *108*, 2796–2815. [CrossRef]
7. Buckingham, M.J. Theory of acoustic attenuation, dispersion, and pulse propagation in unconsolidated granular materials including marine sediments. *J. Acoust. Soc. Am.* **1997**, *102*, 2579–2596. [CrossRef]
8. Buckingham, M.J. Theory of compressional and shear waves in fluidlike marine sediments. *J. Acoust. Soc. Am.* **1998**, *103*, 288–299. [CrossRef]
9. Stoll, R.D. *Sediments Acoustics*; Springer: New York, NY, USA, 1989.
10. Stoll, R.D. Acoustic waves in ocean sediments. *Geophysics* **1977**, *42*, 715–725. [CrossRef]
11. Wood, A.B.; Lindsay, R.B. A Textbook of Sound. *Phys. Today* **1956**, *9*, 37. [CrossRef]
12. Fu, S.S.; Tao, C.H.; Prasad, M.; Wilkens, R.H.; Frazer, L.N. Acoustic properties of coral sands, Waikiki, Hawaii. *J. Acoust. Soc. Am.* **2004**, *115*, 2013–2020. [CrossRef]
13. Fu, S.S.; Wilkens, R.H.; Frazer, L.N. In situ velocity profiles in gassy sediments: Kiel Bay. *Geo-Mar. Lett.* **1996**, *16*, 249–253. [CrossRef]
14. Hamilton, E.L.; Bachman, R.T. Sound velocity and related properties of marine sediments. *J. Acoust. Soc. Am.* **1982**, *72*, 1891–1904. [CrossRef]
15. Hamilton, E.L. Prediction of in-situ acoustic and elastic properties of marine sediments. *Geophysics* **1971**, *36*, 266–284. [CrossRef]
16. Hamilton, E.L. Geoacoustic modeling of the sea floor. *J. Acoust. Soc. Am.* **1980**, *68*, 1313–1340. [CrossRef]
17. Guangming, K.; Yuexia, Z.; Guanbao, L.; Guozhong, H.; Xiangmei, M. Comparison on the sound speeds of seafloor sediments measured by in-situ and laboratorial technique in Southern Yellow Sea. *Ocean. Technol.* **2011**, *30*, 52–56.
18. Kan, G.; Su, Y.; Li, G.; Liu, B.; Meng, X. The correlations between in-situ sound speeds and physical parameters of seafloor sediments in the middle area of the southern Huanghai Sea. *Acta Oceanol. Sin.* **2013**, *35*, 166–171.
19. Dapeng, Z.; Baihai, W.; Bo, L. Analysis and study on the sound velocity empirical equations of seafloor sediments. *Acta Oceanol. Sin.* **2007**, *29*, 43–50.

20. Orsi, T.H.; Dunn, D.A. Sound velocity and related physical properties of fine grained abyssal sediments from the Brazil Basin (South Atlantic Ocean). *J. Acoust. Soc. Am.* **1990**, *88*, 1536–1542. [CrossRef]
21. Orsi, T.H.; Dunn, D.A. Correlations between sound velocity and related properties of glacio-marine sediments: Barents Sea. *Geo-Mar. Lett.* **1991**, *11*, 79–83. [CrossRef]
22. Richardson, M.D.; Briggs, K.B. In situ and laboratory geoacoustic measurements in soft mud and hard-packed sand sediments: Implications for high-frequency acoustic propagation and scattering. *Geo-Mar. Lett.* **1996**, *16*, 196–203. [CrossRef]
23. Liu, X.; Li, A.; Dong, J.; Lu, J.; Huang, J.; Wan, S. Provenance discrimination of sediments in the Zhejiang-Fujian mud belt, East China Sea: Implications for the development of the mud depocenter. *J. Asian Earth Sci.* **2018**, *151*, 1–15. [CrossRef]
24. Xu, F.J.; Li, A.C.; Huang, J.L. Research progress in the mud deposits along the Zhemin coast of the East China Sea continental shelf. *Mar. Sci. Bull.* **2021**, *31*, 97–104.
25. Liu, S.; Shi, X.; Fang, X.; Dou, Y.; Liu, Y.; Wang, X. Spatial and temporal distributions of clay minerals in mud deposits on the inner shelf of the East China Sea: Implications for paleoenvironmental changes in the Holocene. *Quat. Intern.* **2014**, *349*, 270–279. [CrossRef]
26. Zhang, K.; Li, A.; Huang, P.; Lu, J.; Liu, X.; Zhang, J. Sedimentary responses to the cross-shelf transport of terrigenous material on the East China Sea continental shelf. *Sediment. Geolog.* **2019**, *384*, 50–59. [CrossRef]
27. Lim, D.I.; Choi, J.Y.; Jung, H.S.; Rho, K.C.; Ahn, K.S. Recent sediment accumulation and origin of shelf mud deposits in the Yellow and East China Seas. *Prog. Oceanogr.* **2007**, *73*, 145–159. [CrossRef]
28. Meng, X.M.; Liu, B.H.; Kan, G.M.; Li, G.B. An experimental study on acoustic properties and their influencing factors of marine sedi-ment in the southern Huanghai Sea. *Acta Oceanol. Sin.* **2012**, *34*, 74–83.
29. Dong, Y.H.; Liu, L. An improved ID3 algorithm based on correlation coefficients. *Comput. Eng. Sci.* **2016**, *38*, 2342–2347.
30. Qian, N.; Wang, X.; Fu, Y.; Zhao, Z.; Xu, J.; Chen, J. Predicting heat transfer of oscillating heat pipes for machining processes based on extreme gradient boosting algorithm. *Appl. Therm. Eng.* **2020**, *164*, 114521. [CrossRef]
31. Shi, J.; Zhang, J. Load forecasting based on multi-model by stacking ensemble learning. *Proc. CSEE* **2019**, *39*, 4032–4042.