sustainability

MDPI

Article

Research on Estimation Model of Carbon Stock Based on Airborne LiDAR and Feature Screening

Xuan Liu [1,2], Ruirui Wang [1,2,*], Wei Shi [3], Xiaoyan Wang [1,2] and Yaoyao Yang [1,2]

[1] College of Forestry, Beijing Forestry University, Beijing 100083, China; liuxuan2021@bjfu.edu.cn (X.L.); wangxy23@bjfu.edu.cn (X.W.); yangyaocici@bjfu.edu.cn (Y.Y.)
[2] Beijing Key Laboratory of Precision Forestry, Beijing Forestry University, Beijing 100083, China
[3] Beijing Ocean Forestry Technology Co., Ltd., Beijing 100083, China; bnuvictor@hotmail.com
* Correspondence: ruiwang@bjfu.edu.cn

Abstract: The rapid and accurate estimation of forest carbon stock is important for analyzing the carbon cycle. In order to obtain forest carbon stock efficiently, this paper utilizes airborne LiDAR data to research the applicability of different feature screening methods in combination with machine learning in the carbon stock estimation model. First, Spearman's Correlation Coefficient (SCC) and Extreme Gradient Boosting tree (XGBoost) were used to screen out the variables that were extracted via Airborne LiDAR with a higher correlation with carbon stock. Then, Bagging, K-nearest neighbor (KNN), and Random Forest (RF) were used to construct the carbon stock estimation model. The results show that the height statistical variable is more strongly correlated with carbon stocks than the density statistical variables are. RF is more suitable for the construction of the carbon stock estimation model compared to the instance-based KNN algorithm. Furthermore, the combination of the XGBoost algorithm and the RF algorithm performs best, with an R^2 of 0.85 and an MSE of 10.74 on the training set and an R^2 of 0.53 and an MSE of 21.81 on the testing set. This study demonstrates the effectiveness of statistical feature screening methods and Random Forest for carbon stock estimation model construction. The XGBoost algorithm has a wider applicability for feature screening.

Keywords: LiDAR; feature screening; carbon stock; bagging; random forest; forests; model

Citation: Liu, X.; Wang, R.; Shi, W.; Wang, X.; Yang, Y. Research on Estimation Model of Carbon Stock Based on Airborne LiDAR and Feature Screening. *Sustainability* **2024**, *16*, 4133. https://doi.org/10.3390/su16104133

Academic Editor: Ali Bahadori-Jahromi

Received: 19 March 2024
Revised: 27 April 2024
Accepted: 8 May 2024
Published: 15 May 2024

1. Introduction

Forests play an important role in regulating the balance of GHGs in the atmosphere not only as a sink for GHGs due to their use of CO_2 in photosynthesis but also as a source of CO_2 via wildfires [1]. Forest ecosystems account for the largest proportion of terrestrial ecosystem components [2,3]. Forest ecosystems contain about 80% of aboveground carbon and 40% of belowground terrestrial carbon [4]. At the same time, carbon storage is also an important indicator of regional ecosystem service function [5]. The accurate estimation of a regional ecosystem's carbon storage and the exploration of its spatial distribution and influencing factors are of great significance for ecosystem carbon sink function enhancement and management [6]. At present, there are many methods for estimating forest carbon stock, including the sample inventory method, carbon flux observation method, model simulation method based on remote sensing technology, and so on. The required ground data are mainly based on the national forest inventory data, and the biomass and carbon stock of the sample plots are closely related to the storage capacity.

The traditional forest resources survey is based on sampling theory, with a ground survey as the main method; however, the ground measurements of forest resource surveys not only constitute a large amount of workload and consume a long period of time but are also difficult to take consecutive samples over a large area [7]. In recent years, remote sensing technology has developed rapidly; it has been applied by more and more experts and scholars for its ability to accurately, conveniently, and quickly carry out large-scale forest resource surveys in real time. It changes the original traditional forest resource survey

methods, and in a short period of time, it can obtain large-scale data, providing a significant advantage in acquiring metrics such as forest biomass [8,9], stocking capacity [10], crown density [11], etc. Compared to the traditional carbon stock inventory methods that are time-consuming, labor-intensive, and inefficient, the combination of LiDAR technology and machine learning methods can efficiently and accurately estimate forest carbon stocks. Zhang et al. [12] extracted 23 variables based on airborne LiDAR data, such as depression, maximum height, height percentile, crown width percentile, etc. They then constructed a biomass estimation model combining the AGB data using Random Forest and Support Vector Machine algorithms. As a result, the prediction accuracies of the two models for the AGB were both high. Chen et al. [13] extracted variables such as the mean height of the forest, crown density, mean leaf area density, etc. Then, they established a multiple linear regression model and power function model to estimate carbon stock based on UAV-LiDAR point cloud data, and the results showed that the four-parameter nonlinear model had the best fitting effect. Based on airborne LiDAR data, Mu et al. [14] extracted characteristic variables such as the percentile height variable, maximum height, and the percentile density variable of sample plots, and used two models (multivariate linear stepwise regression and Random Forest regression) to estimate carbon stock. Their results showed that the training accuracy and prediction accuracy of the Random Forest regression model were higher than in the multivariate linear stepwise regression model, and the lowest estimation accuracy of the Random Forest model was lower than that of the multivariate linear stepwise regression model. The accuracy of the result is lower than the lowest estimation results of the multiple linear stepwise regression model.

In the construction of carbon stock estimation models, redundant feature variables can reduce the accuracy of data analysis and lead to the incorrect training process of the model due to the reuse of this part of the data. Similarly, the elimination of feature variables that are more related to the target variable will also reduce the accuracy and fitting effect of the model. Therefore, variable screening plays a crucial role in improving model efficiency and even enhancing model performance. In 2008, Fan et al. [15] proposed a sure independence screening (SIS) method based on a correlation study using ultra-high dimensional linear models, whereby feature variables with a weak correlation with the response variable are eliminated to achieve data accuracy and fit. With the rapid development of machine learning technology, variable screening plays an increasingly important role in data analysis and data-driven modeling [16–19]. Li et al. [20] used SAR and Landsat5 TM images to screen remote sensing feature variables using two methods, stepwise regression and Bootstrap, and found that the variable modeling effect of Bootstrap screening is better than that of stepwise regression.

In order to explore the applicability of the combination of different variable screening methods with various machine learning methods to the forest carbon stock estimation model, this study took the forest land in the Zengcheng forestry farm in Guangzhou City, Guangdong Province as the study area. We used airborne LiDAR data to extract the variables, then constructed the carbon stock estimation model using two variable screening methods in combination with three machine learning methods, and compared the effectiveness of the combination of different variable screening methods and machine learning modeling methods. In this study, we introduced the Spearman's Correlation Coefficient (SCC) method and Extreme Gradient Boosting (XGBoost) algorithm to screen the LiDAR variables and found that variable screening to remove redundant features can make the carbon stock estimation model fit at a high level. We compared and analyzed the accuracy of carbon stock estimation models constructed by combining different variable screening methods and machine learning modeling methods to provide a basis for local forest resource inventory and the formulation of forest resource management measures.

2. Materials and Methods

2.1. Site Description

The study area of this paper is located in the Zengcheng District Forestry Farm (23.292°–23.369° N, 113.681°–113.815° E), Guangzhou City, Guangdong Province, which is a state-owned forest farm with a total area of about 2777.55 hm^2. This area is located in the eastern part of Guangzhou City and has a subtropical oceanic monsoon climate, with an average annual temperature of about 22.1 °C, making it warm and rainy, with long summers and short winters. The average annual rainfall is 2039.5 mm, the average annual relative humidity is 78.8%, and the annual sunshine is 1715.4 h. This natural condition is very suitable for tree growth, so there are many kinds of trees and rich vegetation in the forest, with tree species mainly including Masson pine (*Pinus massoniana*), Eucalyptus (*Eucalyptus robusta Smith*), Camphor (*Cinnamomum camphora* (L.) *Presl*), Fir (*China fir*), etc. The area of plantation in the forest is large. Due to this large forest plantation area and the wide distribution of different tree species, this study selected part of the area for the carbon stock study that covers an area of 6.83 hm^2. The geographical location of the study area is shown in Figure 1.

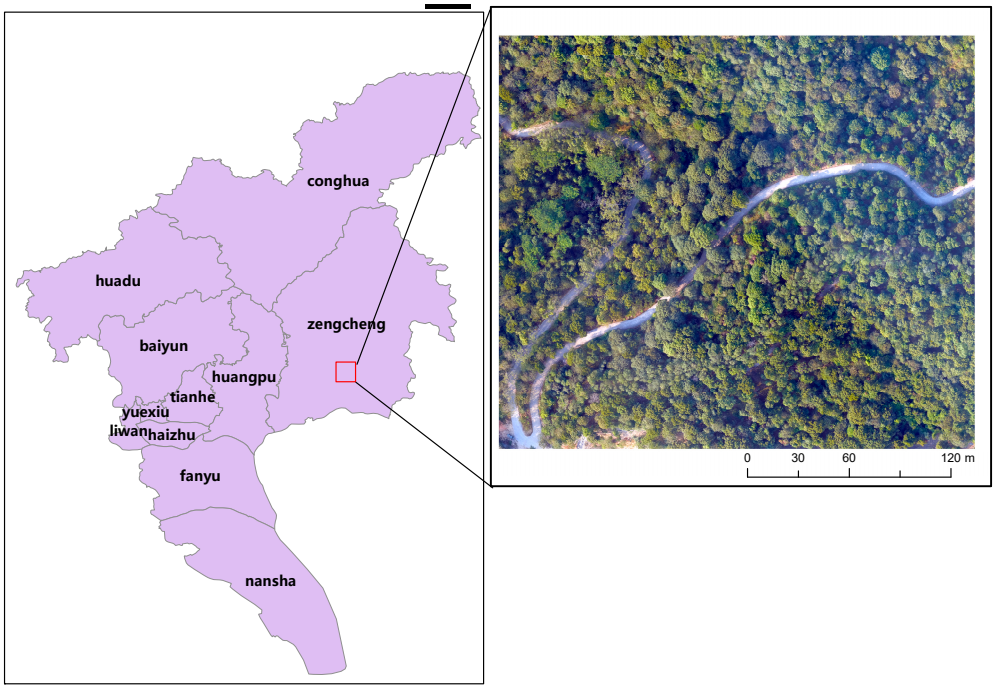

Figure 1. Map of the study area.

2.2. Data Sources and Preprocessing

2.2.1. Data Acquisition

The data used in this study were collected using a Bell helicopter between 18–19 November 2019. The parameter settings applied for the helicopter are as follows: a flight altitude of 500 m; a 45% side-to-side overlap of the flight paths; and a 65% heading overlap of the flight paths. The laser sensor carried by the helicopter is the Galaxy Prime Sensor. The data mainly consisted of airborne laser point cloud data. The settings of the sensor parameter are shown in Table 1. The weather conditions in the study area during the data collection operation were favorable, characterized by sunny and breezy conditions, no cloud cover, and suitable light conditions. Information in the forest field can be accurately obtained, which provides the basis for the study of single-tree segmentation, parameter acquisition, and the accurate estimation of carbon stock.

Table 1. Sensor parameters.

Parameter	Value
Altitude/m	500
Ground speed/kn	60
Mapping bandwidth/m	175
laser wavelength/nm	1064
Pulse Repetition Rate/kHZ	50–1000
Scanning Angle/(°)	10–60
Average density of point clouds/(pts/m^2)	180
Positioning and Orientation System (POS)	POS AV^{TM} AP60 (OEM) 220-channel dual-band GNSS receiver GNSS Receiver Antenna with Iridium Filter High Accuracy AIMU (Type 57)

2.2.2. Data Preprocessing

In this study, the data preprocessing mainly includes two parts: point cloud data denoising and point cloud data normalization. The LiDAR scanning process is easily affected by various factors, resulting in some non-negligible noise in the point cloud data, which affects the effectiveness of detection [21]. In this study, the neighborhood determination method is used to remove the noise points, the threshold for the number of field points is set to 10, and the distance threshold is set to 5 times the standard deviation of the mean elevation of the point cloud [22]. In this study, we use the elevation value of the point cloud data to subtract the corresponding DEM elevation value to achieve the normalization process, as shown in Figure 2.

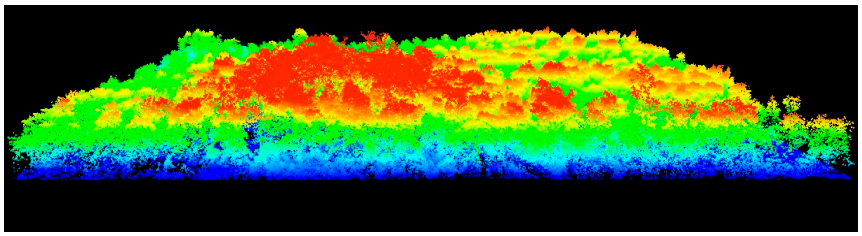

Figure 2. Preprocessed data.

2.3. Samples and LiDAR Variables

2.3.1. Sample Dataset Construction

Carbon stock can be obtained through the calculation of biomass and the carbon content coefficient, through photointerpretation and fieldwork. The tree species involved in this study area are Masson pine (*Pinus massoniana*), Eucalyptus (*Eucalyptus robusta Smith*), Camphor (*Cinnamomum camphora* (L.) *Presl*), and Fir (*China fir*). The calculation of individual tree biomass is from measured individual tree information. The model for calculating individual wood biomass using diameter at breast height is shown in Equations (1)–(4). The calculation of carbon stocks is based on the carbon content factor of different tree species and biomass calculated from the individual tree information. The tree carbon content coefficients are Masson pine 0.5254 [23], Eucalyptus 0.4731 [24], Camphor 0.5117 [25], and Fir 0.5003 [26], and the source of the parameters is the standard issued by the State Forestry Administration.

$$W_1 = 0.09949D^{2.40859} \quad (1)$$

$$W_2 = 0.12576D^{2.46209} \quad (2)$$

$$W_3 = 0.01159D^{3.04803} \quad (3)$$

$$W_4 = 0.07637D^{2.40393} \tag{4}$$

where W_1 is the Masson pine biomass, W_2 is the Eucalyptus biomass, W_3 is the Camphor biomass, W_4 is the Fir biomass, and D is the diameter at the breast height of individual trees.

2.3.2. LiDAR Variables Selection

Spearman's Correlation Coefficient (SCC) is a statistical quantity obtained by ordering the sample values of two random variables by the magnitude of the data and replacing the actual sample values with the ordered values The essence of the Spearman's Correlation Coefficient (SCC) method lies in the correlation analysis based on the ordering position of the original data [27], which has been widely used in carbon stock research, such as the accumulation of organic carbon in lakes [28] and the succession of secondary forests [29]. The formula of Spearman's Correlation Coefficient (SCC) is shown in (5):

$$\rho_s = \frac{\sum_{i=1}^{N}\left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\left[\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2 \sum_{i=1}^{N}\left(S_i - \overline{S}\right)^2\right]^{\frac{1}{2}}} \tag{5}$$

where R_i and S_i are the rank values taken for individual variables or data, respectively, $\overline{R}$ and $\overline{S}$ denote the average rank of the two variables, respectively, and N is the total number of observations.

XGBoost (Extreme Gradient Boosting) [30] is an algorithm based on the GBDT (Gradient Boosting Decision Tree) structure; the GBDT algorithm is a method of generating learners in the process of integrated learning. The idea behind the XGBoost algorithm is to construct the objective function to obtain its optimal value and thereby obtain the algorithm parameters. The objective function consists of a loss function and a canonical term, as shown in (6). The objective function is parameterized using Taylor's second-order expansion and then the tree structure is introduced into the objective function. The purpose is to construct the optimal tree using the optimal objective function and then obtain the algorithm parameters.

$$obj = \sum_{i=1}^{n} L(y_i, y) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where $\sum_{i=1}^{n} L(y_i, y)$ is the loss function of the model, as well as the distance between the predicted value and the measured value of the sample; y_i is the predicted value of the sample i; and y is the true value of the sample i. It is a regular term, which characterizes the complexity of the tree.

Like the GBDT algorithm, the predicted value of the k models in the XGBoost algorithm is the sum of the predicted values of the first $k-1$ and the kth model currently trained. Therefore, the objective function can be rewritten as shown in (7):

$$obj = \sum_{i=1}^{n} L\left(y_i, {y_i}^{k-1} + f_k(x_i)\right) + \sum_{k=1}^{K-1} \Omega(f_k) + \Omega(f_k) \tag{7}$$

Expanding it to Taylor's second order and parameterizing it yields

$$obj = \sum_{j=1}^{J}\left[\sum_{i \epsilon I_j} g_i w_i + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) {w_j}^2\right] + \alpha T \tag{8}$$

where $\sum_{i \epsilon I_j} g_i w_i$ is the loss of samples on each leaf node; T stands for the number of leaf nodes; α, λ stand for the hyperparameters, which are used to control the degree of punishment; w stands for the value of each leaf node; and I_j denotes the samples on the j_{th} leaf node.

In this study, the variables screened using the above two methods and the original extracted variables were used to construct a carbon stock estimation model using three machine learning methods, respectively, and three sets of independent variable schemes were set up. The three schemes are: Scheme 1, Scheme 2 and Scheme 3. Scheme 1 is the combination of variables screened based on Spearman's Correlation Coefficient (SCC) method and machine learning methods, Scheme 2 is the combination of variables screened based on the XGBoost algorithm and machine learning methods, and Scheme 3 is the combination of all the point cloud feature variables extracted based on the LiDAR point cloud data and machine learning methods.

2.4. Model Construction and Accuracy Analysis

The Bagging algorithm's main idea is to randomly select a sample from the initial dataset, which contains S samples, and add it back to the dataset to create a sampling set. After conducting M random sampling operations, K sampling sets are obtained, each containing M training samples. A learner is a model obtained from the data by executing a learning algorithm and is an instantiation of the learning algorithm in the space of given data and parameters. In the process of integration learning, an individual learner is generated from the training data via an existing learning algorithm, and if the integration contains only individual learners of the same type, it is called a "base learner". A base learner is then trained based on each sampling set, and these base learners are combined to form a strong learner. When solving regression problems, the mean of the K model training results is used as the prediction result; when solving classification problems, the voting method is used to generate the results. The main steps of the KNN algorithm are calculating the distance between the test samples and the training samples, selecting the value of K, determining to which category the majority of the K training samples belong, and assigning the information of this category to the test samples. Random Forest (RF) is an algorithm operated by constructing an ensemble of decision trees using resampling techniques [31]. The basic principle is to construct multiple samples from the original training samples after randomly extracting the data using the put-back (Bootstrap) resampling technique, and then construct N decision trees via the random splitting of nodes for each of the resampled samples as the training set of the tree [32].

In summary, in this study, the variables extracted from LiDAR-based data after feature screening are the independent variables, and the carbon stock calculated from biomass after conversion of carbon content coefficients is the dependent variable. The carbon stock estimation model was constructed using three machine learning methods: Bagging, K-nearest neighbor (KNN), and Random Forest (RF), respectively. The estimation model-building process was performed on Python 3.9.10. The model estimation effect was analyzed by comparing the R^2 and RMSE metrics. The Bagging algorithm is regression realized by calling the BaggingRegressor function in Python 3.9.10. The KNN algorithm is realized by calling the KNeighborsRegressor function of the neighbors package in the Sklearn library. The Random Forest algorithm regression is realized by calling the RandomForestRegressor function of the ensemble package in the Sklearn library.

In this paper, we analyze the model's ability to explain the sample data based on the cross-validation method and use the "train_test_split" function of the Sklearn library in the Python language to divide the dataset containing information of 300 single-wood samples into a training set and a test set with the ratio of 7:3:210 samples are used as the training set and 90 samples are used as the test set. The single-wood sample information is the value of the independent variables, and the carbon stock on the single-wood scale is the value of the dependent variables. In this study, the validation accuracy of the model is evaluated using two metrics, the mean squared error (MSE) and the coefficient of determination (R^2). MSE is the average of the sum of squares of the difference between the predicted value and the true value, and the smaller the value of MSE is, the better the model fits. R^2 is a metric used to assess the effect of model fitting, which denotes the proportion of variance in the dependent variable that can be explained by the independent variable. R^2 is an indicator of

the explanatory power of the model. The value of R^2 ranges from 0 to 1, and the closer the value of R^2 is to 1, the better the model fitting effect is indicated to be.

In this study, we take advantage of Airborne LiDAR's ability to acquire information about the vertical structure of the forest canopy and use the variables extracted from point cloud data to construct a carbon stock estimation model via screening using statistically based and machine learning methods and comparing the effects.

3. Results

3.1. Variables Extraction

In this study, the variables were extracted based on the LiDAR point cloud data. According to previous studies, there is a correlation between height-related variables and density-related variables and carbon stocks. Therefore, this study extracted height statistical variables and density statistical variables based on airborne LiDAR data.

3.2. Variables Optimization

In this study, the point cloud feature variables that were selected using the SCC method and the XGBoost algorithm and all the variables extracted based on the airborne LiDAR data were used as the explanatory variables, and the carbon stocks that were calculated based on the biomass and carbon content coefficients were used as the response variables, in order to construct a model for carbon stock estimation. Figure 3 shows the correlation between forest carbon stocks and LiDAR height statistical variables using the SCC variable screening method, Figure 4 shows the correlation between carbon stocks and LiDAR density statistical variables using the SCC variable screening method.

The correlation matrix between the explanatory variables and the response variables calculated according to Spearman's correlation analysis is shown on the left side of all the following two figures, and the interval for correlation classification 0.2, is shown on the right side of all the figures.

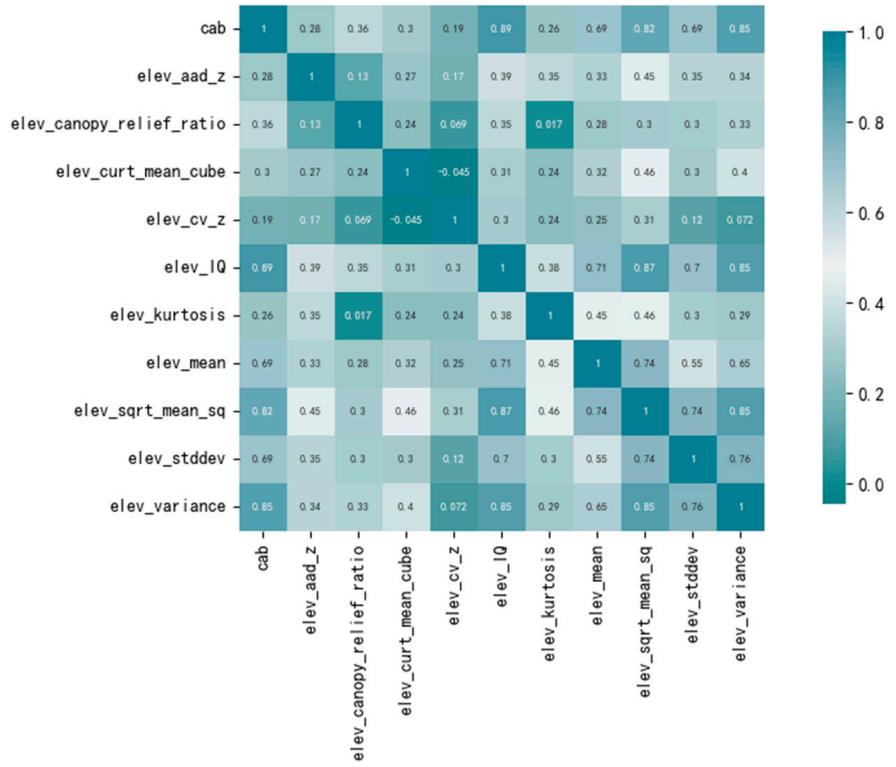

Figure 3. Correlation of forest carbon stocks with LiDAR height statistical variables.

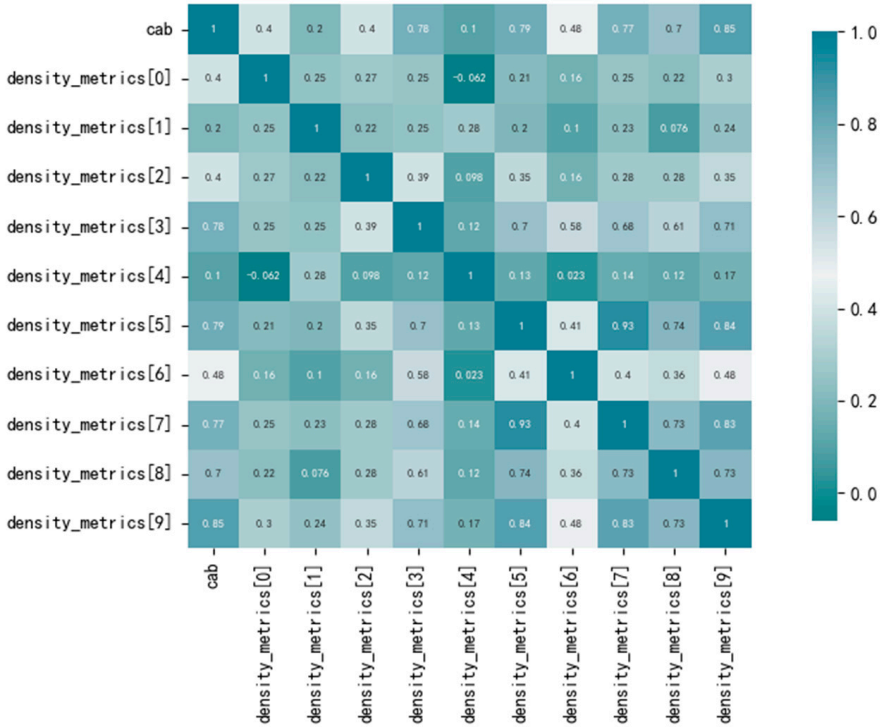

Figure 4. Correlation between carbon stocks and LiDAR density statistical variables.

The correspondence of the variables and the interpretation of the variables are shown in Tables 2 and 3.

Table 2. Correspondence between symbols and height statistical variables and interpretation of height statistical variables.

Variable	Interpretation
elev_aad_z	Average absolute deviation of height.
elev_canopy_relief_ratio	$\frac{mean-min}{max-min}$, where meanis the average height of all points in the cell, min is the minimum height value of all points in the cell, and max is the maximum height value of all points in the cell.
elev_curt_mean_cube	$\sqrt[3]{\frac{\sum_{i=1}^{n} z_i^3}{n}}$, where z_i is the height value of the i^{th} point in the statistical cell and n is the total number of points in the cell.
elev_cv_z	Coefficient of variation of height for all points within the cell.
elev_IQ	Quartile spacing of height percentiles.
elev_kurtosis	Flatness of the height distribution for all points within the cell.
elev_mean	The average of the heights of all points in the cell.
elev_sqrt_mean_sq	$\sqrt[2]{\frac{\sum_{i=1}^{n} z_i^2}{n}}$, where z_i is the height value of the i^{th} point in the statistical cell and n is the total number of points in the cell.
elev_stddev	Standard deviation of the heights of all points within the cell.
elev_variance	The variance of the height of all points within the cell.

According to the definition of the correlation coefficient, the closer the correlation coefficient is to -1, the greater negative correlation exists between the two variables; on the contrary, the closer the correlation coefficient is to 1, the greater positive correlation exists between the two variables. According to the statistical classification of correlation coefficients, a decrease in numerical value indicates a smaller correlation between variables. As can be seen from Figure 3, the correlation coefficients of the height statistical variables

with carbon stocks as a whole are high, and the correlation is at a medium level or above, with the correlation between elev_IQ, elev_sqrt_mean_sq, elev_variance and carbon stocks at the highest level.

Table 3. Correspondence between symbols and density statistical variables and interpretation of density statistical variables.

Variable	Interpretation
density_metrics[0] density_metrics[1] density_metrics[2] density_metrics[3] density_metrics[4] density_metrics[5] density_metrics[6] density_metrics[7] density_metrics[8] density_metrics[9]	The point cloud data are divided into ten slices of the same height from low to high, and the ratio of the number of echoes in each layer is the density variable of the corresponding layer (numbers in [ ] indicate the number of layers).

In this study, height statistical variables with correlation coefficient values of 0.8 or more were selected; the optimal carbon stock model construction variables selected based on the Spearman's Correlation Coefficient (SCC) method are shown in Table 4.

Table 4. Spearman's optimal variable selection results.

Categorization of Variables	Selection Results
Height variables	elev_IQ elev_sqrt_mean_sq elev_variance
Density variables	density_metrics[9] density_metrics[9] density_metrics[9] density_metrics[9] density_metrics[9]

In this study, the "model.feature_importances" function is utilized to score the feature importance (F score) of all the above LiDAR point cloud feature variables. The left side of the figure is labeled to correspond to the above extracted variables one by one in order, and the results are shown in Figure 5.

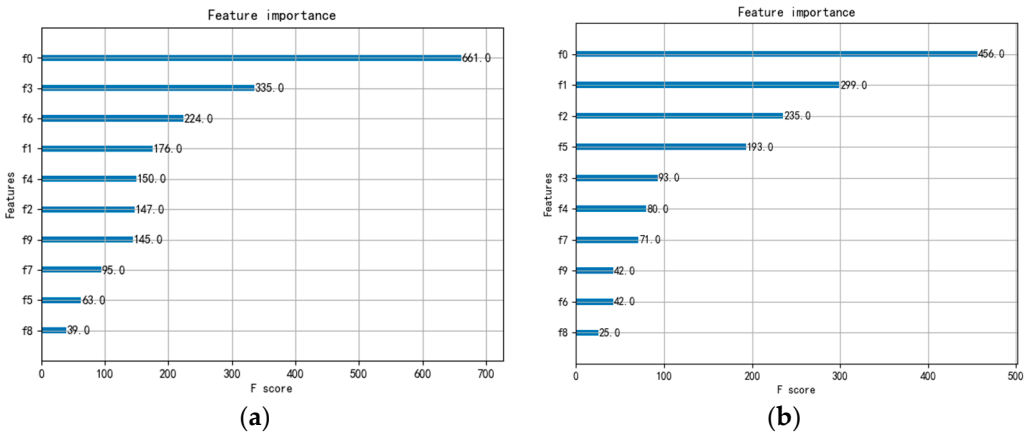

Figure 5. Correlation scores between carbon stocks and LiDAR variables. (**a**) Height statistical variables; (**b**) density statistical variables.

The correspondence of the symbol to height statistical variables and the interpretation of the variables are shown in Table 5. The correspondence of the symbol to density statistical variables and the interpretation of the variables are shown in Table 6.

Table 5. Correspondence between symbols and height statistical variables.

Symbol	Variable
f0	elev_add_z
f1	elev_canopy_relief_ratio
f2	elev_curt_mean_cube
f3	elev_cv_z
f4	elev_IQ
f5	elev_kurtosis
f6	elev_mean
f7	elev_sqrt_mean_sq
f8	elev_stddev
f9	elev_variance

Table 6. Correspondence between symbols and density statistical variables.

Symbol	Variable
f0	density_metrics[0]
f1	density_metrics[1]
f2	density_metrics[2]
f3	density_metrics[3]
f4	density_metrics[4]
f5	density_metrics[5]
f6	density_metrics[6]
f7	density_metrics[7]
f8	density_metrics[8]
f9	density_metrics[9]

According to the importance ranking of the scores of the above variables, in this study. In this study, the location of "the first inflection point" of the importance score and the variables whose scores are greater than their locations are selected as the final selection results. According to Figure 5, the inflection point of importance score in height statistical variables appeared between f3 and f6, so f6 and f0 were selected as the results of the selection of height statistical variables. Based on Table 5, we can see that they are elev_add_z and elev_kurtosis. The importance scores of density statistical variables appeared between f3 and f5, so f5, f2, f1, and f0 were selected as the final selection results. Based on Table 6, it can be seen that they are density_metrics[0], density_metrics[1], density_metrics[2], and density_metrics[5], respectively. The optimal carbon stock model-constructing variables selected based on the XGBoost algorithm are shown in Table 7.

Table 7. XGBoost optimal variable screening results.

Categorization of Variables	Selection Results
Height variables	elev_add_z elev_kurtosis
Density variables	density_metrics[0] density_metrics[1] density_metrics[2]

3.3. Modeling of Carbon Stock Estimation

In this study, the collected sample data of single wood was randomly divided into a training set and a validation set in the ratio of 7:3. In this case, it is worth noting that

the test set is used to adjust the parameters and decide the time to stop training, and the prediction set is used to evaluate the generalization ability of the final model. In this study, the three models built based on machine learning methods were derived by comparing the R^2 and MSE metrics.

The results of model fitting based on Scheme 1 are shown below. Among them, Figure 6 shows the fitting effect of the training set, and Figure 7 shows the fitting effect of the test set. The red line is the fitted line, and the black dots are the sample data points.

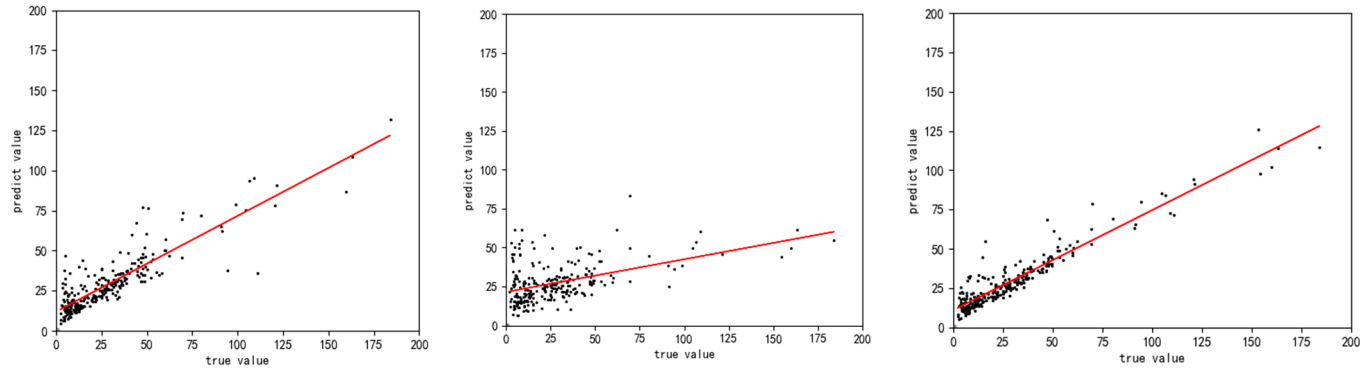

Figure 6. Schematic diagram of the fitting effect of Spearman's Correlation Coefficient method on the training set. (**a**) Training effect of the Bagging method; (**b**) training effect of the KNN method; (**c**) training effect of the Random Forest method.

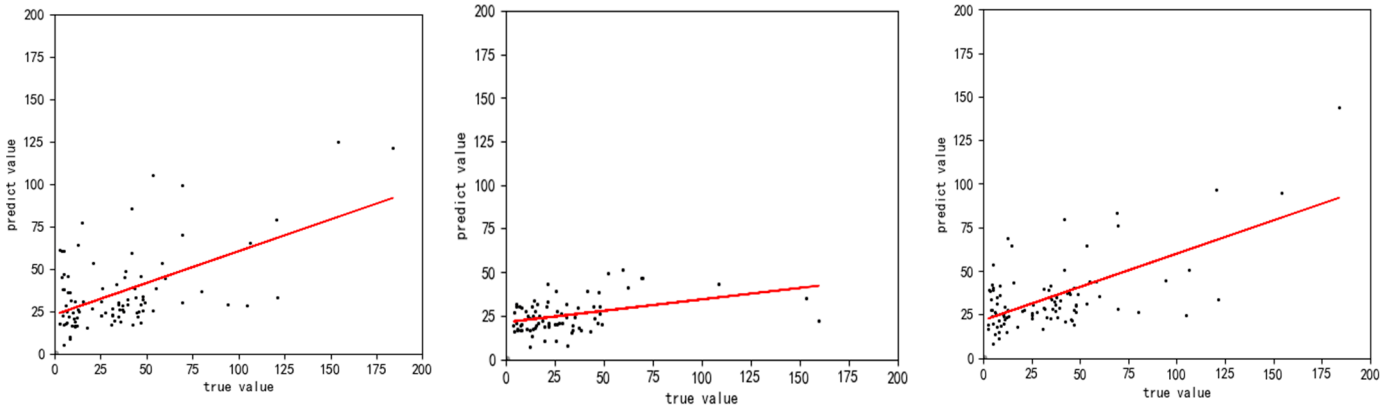

Figure 7. Schematic diagram of the fitting effect of Spearman's Correlation Coefficient method on the test set. (**a**) Testing effect of the Bagging method; (**b**) testing effect of the KNN method; (**c**) testing effect of the Random Forest method.

The effects of the variables screened using Spearman's Correlation Coefficient (SCC) method on the estimation of the model constructed with the three machine learning methods are shown in Table 8.

Table 8. Fitting effect of model based on SCC–Machine Learning.

	Bagging		**KNN**		**Random Forest**	
	train	test	train	test	train	test
R^2	0.73	0.35	0.21	0.12	0.82	0.42
MSE	13.94	24.22	26.46	27.98	12.22	19.69

The results of the model fitting based on Scheme 2 are shown below. Figure 8 shows the training set fitting results and Figure 9 shows the test set fitting results. The red line is the fitted line, and the black dots are the sample data points.

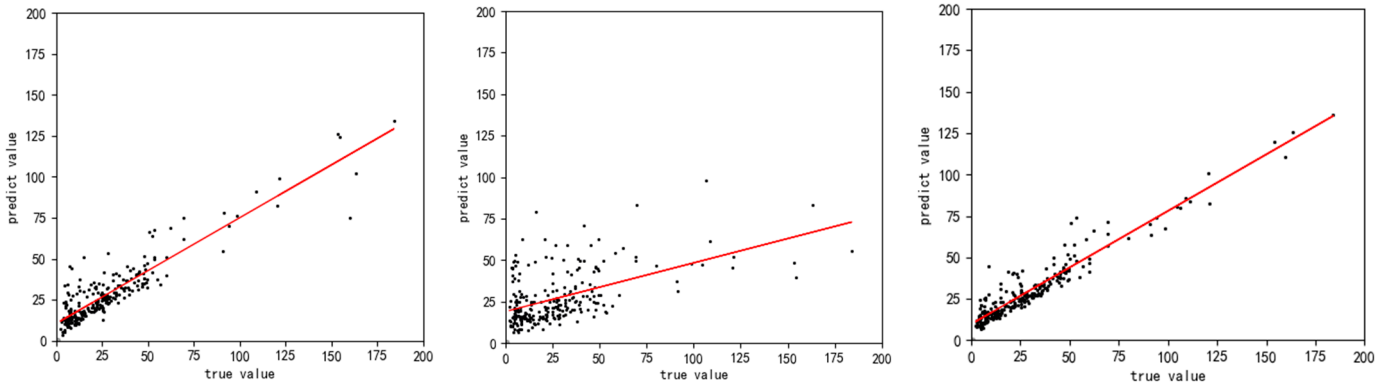

Figure 8. Schematic diagram of the fitting effect of the training set of the XGBoost algorithm. (**a**) Training effect of the Bagging method; (**b**) training effect of the KNN method; (**c**) training effect of the Random Forest method.

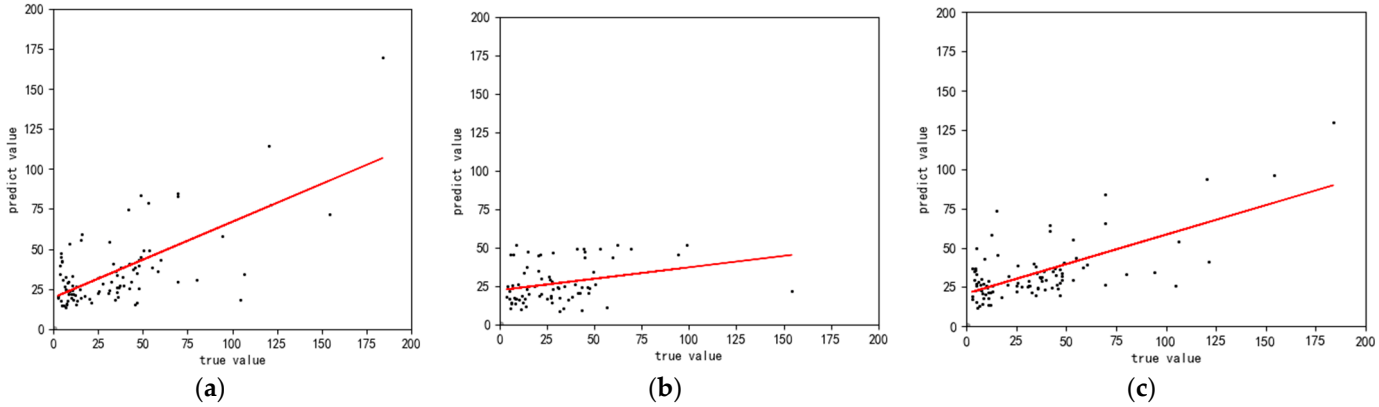

Figure 9. Schematic diagram of the fitting effect of the test set of the XGBoost algorithm. (**a**) Testing effect of the Bagging method; (**b**) testing effect of the KNN method; (**c**) testing effect of the Random Forest method.

The effects of the variables screened using Spearman's Correlation Coefficient method on the estimation of the model constructed with the three machine learning methods are shown in Table 9.

Table 9. Fitting effect of model based on XGBoost–Machine Learning.

	Bagging		**KNN**		**Random Forest**	
	train	test	train	test	train	test
R^2	0.80	0.42	0.27	0.15	0.85	0.53
MSE	12.61	22.46	23.63	24.36	10.74	21.81

The results of model fitting based on Scheme 3 are shown below. Figure 10 shows the training set fitting results and Figure 11 shows the test set fitting results. The red line is the fitted line, and the black dots are the sample data points.

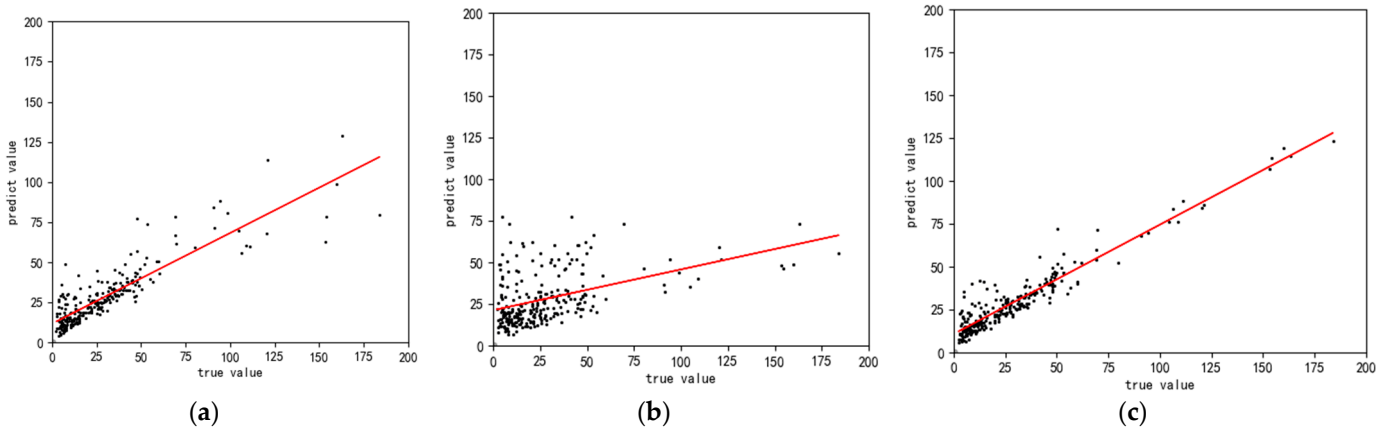

Figure 10. Schematic representation of the effect of fitting the training set of Scheme 3. (**a**) Training effect of the Bagging method; (**b**) training effect of the KNN method; (**c**) training effect of the Random Forest method.

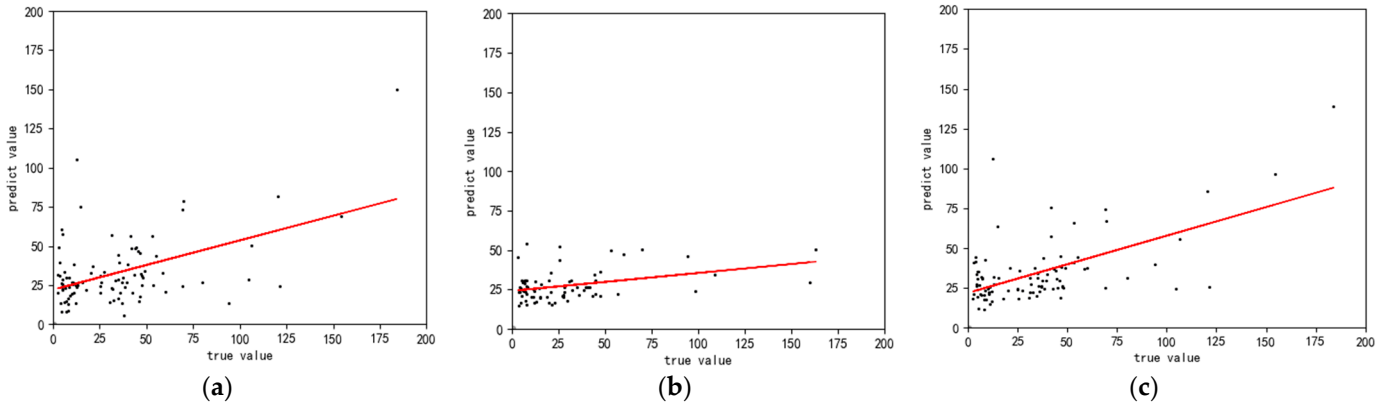

Figure 11. Schematic representation of the effect of fitting the test set of Scheme 3. (**a**) Testing effect of the Bagging method; (**b**) testing effect of the KNN method; (**c**) testing effect of the Random Forest method.

The effects of the variables screened using Scheme 3 on the estimation of the model constructed with the three machine learning methods are shown in Table 10.

Table 10. Fitting effect of model based on Scheme 3.

	Bagging		**KNN**		**Random Forest**	
	train	test	train	test	train	test
R^2	0.72	0.31	0.20	0.13	0.82	0.36
MSE	14.38	24.28	25.48	28.97	12.07	23.61

In comparing the fitting results of the three methods, it can be seen that some differences exist in the accuracy of the carbon stock estimation model constructed based on the above two variable screening methods. In the training set, the carbon stock estimation model constructed via the combination of the XGBoost–Random Forest methods has a better fitting effect, which is the same as the model fitting results in the test set, indicating that the combination of the XGBoost–Random Forest methods has a better feasibility for the construction of carbon stock estimation model.

4. Discussion

Forest carbon sinks exhibit significant ecological and economic worth, making them an indispensable approach to combatting global climate change [33]. Learning how to

accurately and rapidly assess forest carbon stocks is the major research focus of many scholars. With the rapid development of remote sensing technology, the use of remote sensing platforms for carbon stock estimation is becoming extremely consistent and reliable. As a result, remote sensing methods have become a major tool for quantifying forest carbon stocks on a wider scale [34].

LiDAR is an active remote sensing technology that uses short-wavelength laser pulses to penetrate forest canopies and obtain vertical structure information [35]. In contrast to Zhou et al. [36]'s study on categorizing feature variables based on the Random Forest algorithm and the SCC method, this study compares the effectiveness of an integrated algorithm of gradient boosting by iteratively training a series of weak learners and combining them with a statistical approach based on monotonic equations evaluating correlations of multiple statistical variables for the preferential selection of feature variables.

In this study, three machine learning methods, Bagging, KNN, and Random Forest, are used to construct a carbon stock estimation model. Compared with Mateus's study on estimating carbon stock based on machine learning methods with the Generalized Linear Modeling approach (GLM) [37], we compare the estimation effectiveness of the constructed models among the three machine learning methods. Our approach demonstrates that, for the prediction of carbon stock, the method constructed and combined with the multiple machine learners to accomplish the learning task significantly outperforms the instance-based machine learning approach, and the value of R^2 reaches 0.84. Comparing the estimation results of the Bagging algorithm and the Random Forest algorithm, the Random Forest algorithm slightly outperforms the Bagging algorithm, a phenomenon that is particularly prominent in the test set. This also proves that Random Forest is an improved version of Bagging, as the Random Forest algorithm further introduces random attribute selection during the training process of the decision tree, based on building a Bagging integration with the decision tree as the base learner.

Variable screening plays a crucial role in improving model efficiency, model accuracy, and stability. Long et al. [38] constructed a vegetation carbon stock estimation model and also found that constructing a vegetation carbon stock estimation model based on bandwidth preference was the most effective.

However, comparing the variables selected using Scheme 3 with the first two methods, respectively, it can be seen that the predictive effect of the model constructed based on the variables of Scheme 3 has receded on the test set. And, the R^2 value is 0.1–0.2 lower than that of the model constructed by the first two methods. The R^2 value of the model constructed by combining the KNN algorithm on the test set is down to 0.13. This suggests that the predictive effect of the model constructed using the KNN algorithm is not significant in the case of the variables that are optimal. This is true even when they are selected based on different methods or when they have a high correlation with the dependent variable; accumulating them does not give the optimal estimation, and there is a tendency for them to decrease with the increase of the feature variables. However, it should be noted that this study was tested on a limited dataset, and there are various factors affecting the effectiveness of the carbon stock estimation model construction; in order to draw more generalized and stable conclusions, it is necessary to carry out the study on a larger number and variety of sample datasets, such as the type of tree species and the scale of the sample plots. The dataset of this experiment was obtained at a small to medium scale in the forest field, and sample datasets need to be obtained at a larger scale such as the national scale or even a larger scale to verify its applicability at different regional scales.

In this paper, the carbon stocks calculated based on biomass and carbon content coefficients in the sample plots are used directly as the dependent variable, and the height and density variables extracted from the point cloud data are used as the independent variables to construct the carbon stock estimation model. This is in contrast to the studies conducted by other scholars on the correlation between the construction of the aboveground biomass estimation model and the remotely sensed characterization factors or the combination of remotely sensed characterization factors and the point cloud characterization variables. For

example, Wang et al. [39] extracted vegetation indices based on Sentinel2 image data to construct an aboveground biomass estimation model, and Du et al. [40] extracted point cloud feature variables such as canopy cover and leaf area index, as well as texture features such as variance and mean, to construct an aboveground biomass model based on Landsat image data and airborne LiDAR point cloud data; the methodology of this paper has great potential to study the direct correlation between point cloud feature variables and carbon stock.

In this paper, a carbon stock estimation model is constructed based on LiDAR technology through different machine learning methods, and the variables extracted based on the LiDAR data are utilized for the construction of the estimation model. Compared with the previous estimation of forest stock only based on remote sensing imagery, this paper utilizes the characteristics of the LiDAR data that can reflect the vertical information of forests to highlight the role of vertical variables in the construction of the carbon stock estimation model. This shows that the machine learning method is less affected by human interference and has a stronger learning ability, which can effectively construct the carbon stock estimation model.

Among the carbon stock models constructed based on Spearman's Correlation Coefficient (SCC) method, the XGBoost algorithm, and the three machine learning methods of Bagging, KNN, and RF, the XGBoost algorithm, Bagging method, and Random Forest method have the best performance. This study shows that the combination of variable screening and the integration of machine learning algorithms to estimate carbon stock is more effective and suitable for the estimation of carbon stock. This study shows that the estimation of carbon stock via variable screening combined with the integrated learning algorithm in machine learning is more effective and has the best applicability. However, the sample dataset of this study is insufficient, and there are many factors affecting carbon stock such as climate conditions and soil type, so this study only constructs a carbon stock estimation model applicable to a specific period of time.

5. Conclusions

This study compares the effectiveness of Spearman's Correlation Coefficient (SCC) in statistical methods and the XGBoost algorithm in machine learning algorithms in screening carbon stock estimation models and combines the three machine learning methods with the above two feature screening methods to explore the optimal model for carbon stock estimation. The results are as follows:

1. Compared with density statistical variables, height statistical variables have a higher correlation with carbon stocks;
2. Comparing and analyzing the performance of the three algorithms in fitting the model, the Bagging algorithm and RF algorithm fit the model with a larger R^2 and a smaller MSE, which is a better fitting effect; on the contrary, the KNN algorithm fits the model with the smallest R^2 and a higher MSE, which is a poor fitting effect. Therefore, compared to instance-based machine learning algorithms, the integrated learning method has better applicability for the construction of a carbon stock estimation model, especially the Random Forest method;
3. The accuracies of the carbon stock estimation models constructed using the two variable screening methods are at a high level but comparing the effects of the carbon stock estimation models constructed on the basis of the two variable screening methods in the training and test sets, the XGBoost algorithm performs optimally.

In summary, comparing and analyzing the performance of the three algorithms in fitting the model, the combination of Spearman's algorithm and the Random Forest algorithm has the best fitting effect; on the contrary, the KNN algorithm does not have a high accuracy in fitting the model, and the fitting effect is not good. Therefore, the XGBoost–Random Forest combination method is more suitable for building a carbon stock estimation model in this study area.

Author Contributions: Conceptualization, X.L. and R.W.; methodology, X.L.; software, X.L.; validation, X.L., R.W., W.S., X.W. and Y.Y.; writing—original draft preparation, X.L.; writing—review and editing, X.L.; visualization, X.L.; supervision, X.L.; project administration, R.W.; funding acquisition, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41971376, and the APC was funded by "biomass precision estimation model research for large-scale region based on multi-view heterogeneous stereographic image pair of forest".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Wei Shi was employed by the company Beijing Ocean Forestry Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Beijing Ocean Forestry Technology Co., Ltd. had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Lee, S.J.; Kim, J.R.; Choi, Y.S. The extraction of forest CO_2 storage capacity using high-resolution airborne lidar data. *GISci. Remote Sens.* **2013**, *50*, 154–171. [CrossRef]
2. Harris, N.L.; Gibbs, D.A.; Baccini, A.; Birdsey, R.A.; de Bruin, S.; Farina, M.; Fatoyinbo, L.; Hansen, M.C.; Herold, M.; Houghton, R.A.; et al. Global maps of twenty-first century forest carbon fluxes. *Nat. Clim. Chang.* **2021**, *11*, 234–240. [CrossRef]
3. Watson, R.T.; Noble, I.R.; Bolin, B.; Ravindranath, N.H.; Verardo, D.J.; Dokken, D.J. *Land Use, Land-Use Change and Forestry: A Special Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2000.
4. Dixon, R.K.; Winjum, J.K.; Andrasko, K.J.; Lee, J.J.; Schroeder, P.E. Integrated Land-Use Systems: Assessment of Promising Agroforest and Alternative Land-Use Practices to Enhance Carbon Conservation and Sequestration. *Clim. Chang.* **1994**, *27*, 71–92. [CrossRef]
5. Gong, J.; Liu, D.Q.; Zhang, J.X.; Xie, Y.C.; Cao, E.J.; Li, H.Y. Tradeoffs/synergies of multiple ecosystem services based on land use simulation in a mountain-basin area, western China. *Ecol. Indic.* **2019**, *99*, 283–293. [CrossRef]
6. Sun, B.Q.; Du, J.Q.; Chong, F.F.; Li, L.J.; Zhu, X.Q.; Zhai, G.Q.; Song, Z.; Mao, J.L. Spatio-Temporal Variation and Prediction of Carbon Storage in Terrestrial Ecosystems in the Yellow River Basin. *Remote Sens.* **2023**, *15*, 3866. [CrossRef]
7. Pang, Y.; Li, Z.; Yu, T.; Liu, Q.; Zhao, L.; Chen, E. Status and Development of Forest Carbon Storage Remote Sensing Satellites. *Spacecr. Recovery Remote Sens.* **2022**, *43*, 1–15. (In Chinese)
8. Xu, H.; Pan, P.; Ning, J.; Zang, H.; Ouyang, X.; Xiang, Y.; Wu, Z.; Guo, R.; Gui, Y. Remote Sensing Estimation of Forest Aboveground Biomass Based on Multiple Linear Regression and Neural Network Model. *J. Northeast. For. Univ.* **2018**, *46*, 63–67. (In Chinese) [CrossRef]
9. Zheng, D.; Xia, C.; Wang, H.; Chen, J.; Hou, R. Study on biomass estimation model of masson pine based on SPOT-7 image. *J. Cent. South Univ. For. Technol.* **2018**, *38*, 82–88. (In Chinese) [CrossRef]
10. Liu, B.; Li, C.; Guo, R.; Liu, S.; Ma, T. Building Forest Volume Estimation Model Using GF-1 Image Spectral and Texture Information. *J. Northeast. For. Univ.* **2020**, *48*, 9–12+28. (In Chinese) [CrossRef]
11. Li, Q.; Wang, Z.; Wang, Y.; Liu, M.; Yang, Y. Study on canopy density inversion of Picea schrenkiana forest based on GF-2 remote sensing image. *J. Cent. South Univ. For. Technol.* **2019**, *39*, 48–54. (In Chinese) [CrossRef]
12. Zhang, P.; Ma, Q.X.; Lv, J.; Ji, J.; Li, Z. Application of machine learning algorithms in estimation of above-ground biomass of forest. *Bull. Surv. Mapp.* **2021**, *12*, 28–32. [CrossRef]
13. Chen, Z.; Liu, Q.; Li, C.; Li, M.; Zhou, X.; Yu, Z.; Su, K. Comparison in linear and nonlinear estimation models of carbon storage of plantations based on UAV LiDAR. *J. Beijing For. Univ.* **2021**, *43*, 9–16. (In Chinese) [CrossRef]
14. Mu, X.Y.; Liu, Q.W.; Pang, Y.; Hu, K.; Zhang, Q. Forest Aboveground Carbon Storage Using RF Algorithmic Model and Airborne LiDAR Data. *J. Northeast. For. Univ.* **2016**, *44*, 52–56. [CrossRef]
15. Fan, J.Q.; Lv, J.C. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 849–911. [CrossRef] [PubMed]
16. Wang, J.Z.; Wu, L.S.; Kong, J.; Li, Y.X.; Zhang, B.X. Maximum weight and minimum redundancy: A novel framework for feature subset selection. *Pattern Recognit.* **2013**, *46*, 1616–1627. [CrossRef]
17. Xin, L.; Zhu, M. Stochastic Stepwise Ensembles for Variable Selection. *J. Comput. Graph. Stat.* **2012**, *21*, 275–294. [CrossRef]
18. Yan, Y.; Wang, L.J.; Wang, T.; Wang, X.; Hu, Y.H.; Duan, Q.S. Application of soft computing techniques to multiphase flow measurement: A review. *Flow Meas. Instrum.* **2018**, *60*, 30–43. [CrossRef]

19. Zhang, C.X.; Zhang, J.S.; Yin, Q.Y. Early stopping aggregation in selective variable selection ensembles for high-dimensional linear regression models. *Knowl.-Based Syst.* **2018**, *153*, 1–11. [CrossRef]
20. Li, M.; Yu, X.; Gao, Y.; Fan, W. Remote sensing quantification on forest biomass based on SAR polarization decomposition and Landsat data. *J. Beijing For. Univ.* **2018**, *40*, 1–10. [CrossRef]
21. Cai, Z.; Jin, C. Object Contour Recognition Based on 2D Lidar Point Cloud. *Appl. Laser* **2020**, *40*, 513–518. (In Chinese)
22. Su, T. *Three-Dimensional Segmentation of Single Wood in Power Line Corridor Based on Airborne LiDAR Data*; Beijing Forestry University: Beijing, China, 2020. (In Chinese) [CrossRef]
23. *LY/T 2263-2014*; Stumpage Biomass Model and Carbon Measurement Parameters—Masson pine. China Standard Press: Beijing, China, 2014.
24. *DB45/T 2751-2023*; Stumpage Biomass Model and Carbon Measurement Parameters—Eucalyptus. Guangxi Zhuang Autonomous Region Market Supervision and Administration Bureau: Nanning, China, 2023.
25. *DB44/T 2177-2019*; Stumpage Carbon Measurement Models and Parameters for Three Native Broadleaf Species, including Camphor tree. Guangdong Provincial Administration for Market Regulation: Guangzhou, China, 2019.
26. *LY/T 2264-2014*; Stumpage Biomass Model and Carbon Measurement Parameters—Fir. China Standard Press: Beijing, China, 2014.
27. Liu, H.; Song, F.; Lei, L.; Zheng, H.; Li, H.; Chen, K. Correlation of Clinical Features with Immunohistochemical Indices of 1,267 Cases of Breast Cancer. *Chin. J. Clin. Oncol.* **2011**, *38*, 656–659. [CrossRef]
28. Zhang, F.J.; Yao, S.C.; Xue, B.; Lu, X.X.; Gui, Z.F. Organic carbon burial in Chinese lakes over the past 150 years. *Quat. Int.* **2017**, *438*, 94–103. [CrossRef]
29. Cortés-Calderón, S.; Mora, F.; Arreola-Villa, F.; Balvanera, P. Ecosystem services supply and interactions along secondary tropical dry forests succession. *For. Ecol. Manag.* **2021**, *482*, 118858. [CrossRef]
30. Liang, W.Z.; Luo, S.Z.; Zhao, G.Y.; Wu, H. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics* **2020**, *8*, 765. [CrossRef]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Li, G.; Li, J.; Zhang, L.; Xin, Y.; Deng, H. Feature Selection Method Based on Ant Colony Optimization and Random Forest. *Comput. Sci.* **2019**, *46*, 212–215.
33. Li, Q.; Xia, X.L.; Kou, X.M.; Niu, L.; Wan, F.; Zhu, J.H.; Xiao, W.F. Forest Carbon Storage and Carbon Sequestration Potential in Shaanxi Province, China. *Forests* **2023**, *14*, 2021. [CrossRef]
34. Xu, L.; Lai, H.Y.; Yu, J.G.; Luo, S.L.; Guo, C.S.; Gao, Y.Q.; Zhou, W.; Wang, S.; Shu, Q.T. Carbon Storage Estimation of *Quercus aquifolioides* Based on GEDI Spaceborne LiDAR Data and Landsat 9 Images in Shangri-La. *Sustainability* **2023**, *15*, 11525. [CrossRef]
35. Xi, Z.L.; Xu, H.D.; Xing, Y.Q.; Gong, W.S.; Chen, G.Z.; Yang, S.H. Forest Canopy Height Mapping by Synergizing ICESat-2, Sentinel-1, Sentinel-2 and Topographic Information Based on Machine Learning Methods. *Remote Sens.* **2022**, *14*, 364. [CrossRef]
36. Zhou, G.L.; Ni, Z.Y.; Zhao, Y.B.; Luan, J.W. Identification of Bamboo Species Based on Extreme Gradient Boosting (XGBoost) Using Zhuhai-1 Orbita Hyperspectral Remote Sensing Imagery. *Sensors* **2022**, *22*, 5434. [CrossRef]
37. Schuh, M.; Favarin, J.A.S.; Marchesan, J.; Alba, E.; Berra, E.F.; Pereira, R.S. Machine learning and generalized linear model techniques to predict aboveground biomass in Amazon rainforest using LiDAR data. *J. Appl. Remote Sens.* **2020**, *14*, 034518. [CrossRef]
38. Long, Y.; Jiang, F.; Sun, H.; Wang, T.; Zou, Q. Estimating vegetation carbon storage based on optimal bandwidth selected from geographically weighted regression model in Shenzhen City. *Acta Ecol. Sin.* **2022**, *42*, 4933–4945. (In Chinese)
39. Wang, P.; Tan, S.; Zhang, G.; Wang, S.; Wu, X. Remote Sensing Estimation of Forest Aboveground Biomass Based on Lasso-SVR. *Forests* **2022**, *13*, 1597. [CrossRef]
40. Du, C.; Fan, W.; Ma, Y.; Jin, H.-I.; Zhen, Z. The Effect of Synergistic Approaches of Features and Ensemble Learning Algorithms on Aboveground Biomass Estimation of Natural Secondary Forests Based on ALS and Landsat 8. *Sensors* **2021**, *21*, 5974. [CrossRef] [PubMed]