electronics

MDPI

Article

Decision Tree Application to Classification Problems with Boosting Algorithm

Long Zhao [1], Sanghyuk Lee [1] and Seon-Phil Jeong [2,*]

[1] Depart of Mechatronics and Robotics, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; long.zhao19@student.xjtlu.edu.cn (L.Z.); Sanghyuk.Lee@xjtlu.edu.cn (S.L.)
[2] Division of Science and Technology, CST Programme BNU-HKBU United International College, Zhuhai 519085, China
* Correspondence: spjeong@uic.edu.cn; Tel.: +86-756-362-0616

Abstract: A personal credit evaluation algorithm is proposed by the design of a decision tree with a boosting algorithm, and the classification is carried out. By comparison with the conventional decision tree algorithm, it is shown that the boosting algorithm acts to speed up the processing time. The Classification and Regression Tree (CART) algorithm with the boosting algorithm showed 90.95% accuracy, slightly higher than without boosting, 90.31%. To avoid overfitting of the model on the training set due to unreasonable data set division, we consider cross-validation and illustrate the results with simulation; hypermeters of the model have been applied and the model fitting effect is verified. The proposed decision tree model is fitted optimally with the help of a confusion matrix. In this paper, relevant evaluation indicators are also introduced to evaluate the performance of the proposed model. For the comparison with the conventional methods, accuracy rate, error rate, precision, recall, etc. are also illustrated; we comprehensively evaluate the model performance based on the model accuracy after the 10-fold cross-validation. The results show that the boosting algorithm improves the performance of the model in accuracy and precision when CART is applied, but the model fitting time takes much longer, around 2 min. With the obtained result, it is verified that the performance of the decision tree model is improved under the boosting algorithm. At the same time, we test the performance of the proposed verification model with model fitting, and it could be applied to the prediction model for customers' decisions on subscription to the fixed deposit business.

Keywords: decision tree; boosting algorithm; receiver operating curve; cross-validation

Citation: Zhao, L.; Lee, S.; Jeong, S.-P. Decision Tree Application to Classification Problems with Boosting Algorithm. *Electronics* **2021**, *10*, 1903. https://doi.org/10.3390/electronics10161903

Academic Editors: Jihoon Yang and Unsang Park

Received: 15 July 2021
Accepted: 5 August 2021
Published: 8 August 2021

1. Introduction

As a classification function approximation method, the decision tree is developed from the field of machine learning [1]. Recently, decision tree design methodology has been extended and proposed to raise accuracy via boosting algorithm addition. Numerous researchers have emphasized the related research [2–4]. Hunt et al. proposed that the concept learning system is the earliest decision tree algorithm [5]. Then, the decision tree algorithm gradually developed a series of algorithms, such as Iterative Dichotomizer3 (ID3) algorithm, C4.5 algorithm, C5.0 algorithm, Classification and Regression Tree (CART) algorithm, and so on [6]. The algorithms used in this paper are C5.0 algorithm and CART algorithm, both of which are evolved from the previous algorithm, and their comprehensive performance has been improved [6]. C5.0 algorithm is an intuitive and efficient classification method, but it has the problems of information gain rate calculation complexity, and is prone to overfitting and decision tree bias. To solve these problems, the calculation process of the information gain rate is simplified by formula transformation. In the pruning process, the combination of loss matrix and confidence interval is used to judge pruning, and the weights of multiple models are adjusted. A modified C5.0 algorithm with boosting method is proposed [7]. In the previous study, a classifier ensemble was proposed to enhance diversity, and it provided a near-optimal classifying system [8,9].

In previous studies, C5.0 algorithm and CART algorithm generally have overfitting problems or insufficient model performance optimization when they deal with imbalanced data. This causes problems with the decision-making mistakes, which are prone to unstable prediction when they are applied to real problems. In order to overcome these problems, this paper proposes, by adding the cost matrix and boosting algorithm, to improve these problems [6], and it verifies the decision results improvement through application to actual data.

At the same time, there is the problem with the classification error of different generation values when it is not treated differently with the decision tree C5.0 algorithm, which makes the cost of classification error higher. In this paper, we use the value of misconduct and cost matrix to reduce the high-cost error rate; we realize C5.0 under the condition that the overall error rate of the model changes is small. It is expected that the optimized model can reduce the high-cost error rate in the test data. The result is proven when the application effect of the cost matrix is obvious, and the general cost error rate could be reduced [10]. Finally, based on the C5.0 decision tree, a boosting algorithm is used in this paper, and a cost matrix is introduced for the comparison with the CART algorithm. According to the receiver operating curve model, the performance evaluation index and the decision tree algorithm cross-check are the result. Then, the model performance is comprehensively evaluated.

By the application of boosting knowledge, Pang showed the C5.0 algorithm and the corresponding boosting technology in detail based on the decision tree C4.5 algorithm and embedded the boosting algorithm technology [11]. The personal credit rating model is established in a bank based on the C5.0 algorithm and the model is applied to perform a credit rating with the personal credit data of a German bank. By the comparison of the decision tree application with before results and after, the model parameters are adjusted. The experimental results show that the discrimination result with the decision tree after the parameter adjustment is better than before the parameter adjustment [11]. Furthermore, a modified *k*-mean clustering algorithm has been studied by Ahmad and Dey for the mixed numeric and categorical features, not only for numeric data [12]. Wang, Jiang, and Hui tried to increase the accuracy of the current stock prediction model, which is not high enough, but there are challenges such as overfitting or underfitting which are based on the analysis of the existing stock prediction methods. In the research, a CART-based decision tree was given for the stock forecasting method with boosting method, and it used boosting method which is cascaded to multiple decision trees to solve the fitness problem. By selecting seven indicators in the stock data, the mean square error (MSE) and the mean square standard deviation (RMSE) are used to evaluate the prediction accuracy. Experimental results show that the decision tree fitting effect and prediction accuracy rate after adding the boosting algorithm are higher than the original model [13]. Yao et al. researched and analyzed the new decision tree C5.0 algorithm. In predictive classification, the cost of misjudgment was considered in the decision tree modeling, and the value conditions for the substitution value of misjudgment were given, and a cost matrix was established to guide the modeling. The cost of the prediction classification error is minimized when the overall error rate of the model does not change much. In-depth study of the decision tree C5.0 algorithm based on the cost matrix and its application in the classification has been carried out for the patient classification problem in a Chinese hospital. From the final patient classification model, the model has a high classification error rate in the modeling data and test data, even though the model has the advantages of low risk and good stability [14]. In this paper, we add the boosting idea to the conventional decision algorithm and obtain high accuracy by the generation of a strong classifier to the corresponding data. The result also overcomes the overfitting problem, and optimal decision results are obtained for the given personal banking data by using a confusion matrix.

The paper is organized as follows: preliminary study on data processing and evaluation in Section 2. For the evaluation, accuracy and sensitivity are introduced with the confusion matrix. The considered boosting algorithm is introduced here. In Section 3,

C5.0 and CART algorithm are applied to empirical data. After data analysis, it is ensured that there is no need to perform principal component analysis. The decision results are carried out with conventional C5.0/CART model and by adding the boosting algorithm in Section 4. The results are discussed in Section 5. In the discussion, different considerations on positive prediction are investigated and illustrated. Finally, conclusions follow in Section 6.

2. Preliminaries and Methodology

In this section, the method of preliminary research on data analysis and the methodology of decision tree algorithm and boosting algorithm are explained.

2.1. Data Processing and Analysis

In statistics, data relation has been used with the help of correlation and covariance [15]. The variables with a correlation coefficient close to 0 are regarded as non-correlated, and close to 1 or -1 are regarded as having a strong correlation. Variance Inflation Factors (VIF) represent a measure of the severity of multi-collinearity characteristic in a multiple linear regression model. This shows the ratio between the variance of the regression coefficient obtained from estimator and the variance which is assumed that the independent variables are not linearly correlated. When the variance expansion factor is too large, it indicates that there is a strong correlation between the independent variables [1]. The specific steps of VIF inspection are as follows:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u \tag{1}$$

where k is the number of different VIF, and X_i's are variables. β_i,$i = 1, \ldots, k$ are the standard error of the estimates.

To calculate each VIF for the specific X_i, $i = 1, \ldots, k$, the following procedure is needed [1]: First, implement an ordinary least squares regression in which X_i is a function of all other explanatory variables in Equation (1). For $i = 1$, the equation satisfies in Equation (2):

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots + \alpha_k X_k + v \tag{2}$$

where α_1 is a constant and v is the error term.

Next, we calculate VIF by $\text{VIF} = \frac{1}{(1-R_i^2)}$, where ${R_i}^2$ is the coefficient of determination from the first step auxiliary regression. R_i is the correlation coefficient between X_i and other variables X_j, i, $j = 1, \ldots, k$. By analysis, the magnitude of multi-collinearity is obtained by calculating the size of the VIF. The value of VIF is greater than 1. The closer the VIF value is to 1, the smaller the multi-collinearity [1].

2.2. Model Performance Evaluation

In the field of machine learning, effective decision threshold value is considered by using a confusion matrix. It is used to provide an effective boundary to classify the data [16].

To simplify all the data, we use true positive (TP), false negative (FN), false positive (FP), and true negative (TN) as in Table 1, respectively. FN and FP are considered as Type I and Type II error. Four indicators are illustrated as the confusion matrix in Table 1 [17].

Table 1. Confusion matrix.

		True Values	
	Total Population	**Positive**	**Negative**
Prediction	Positive	TP	FP
	Negative	FN	TN

From Table 1, TP + FN + FP + TN satisfies the total number of samples. Hence, we define accuracy as the closeness of the measurements to a specific value. So, it can be provided in Equation (3).

$$Accuracy\ (ACC) = \frac{\mathrm{TP}\ +\ \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{3}$$

The precision and sensitivity are the ratio of true positive value with respect to total positive predicted conditions and total amount of actual true values, respectively. These properties are illustrated in Equations (4) and (5). Both precision and sensitivity satisfy based on an understanding and measure of relevance.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

The true negative rate is also expressed as specificity and selectivity, and it is illustrated in Equation (6).

$$Specificity\ (SPC),\ Selectivity,\ True\ negative\ rate\ (TNR) = \frac{\mathrm{TN}}{\mathrm{FN} + \mathrm{TN}} \tag{6}$$

From Equations (4) and (5), another parameter F1 is defined in Equation (7).

$$\mathrm{F1}\ score = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{7}$$

F1 is represented as the harmonic mean of precision and sensitivity. So, F1 acts as a comprehensive indicator that is used to analyze whether the TP is large enough from two perspectives, subjective (predicted) and objective (actual).

For the classification model, the above evaluation indicators can be used to judge whether the classification model meets our requirements [16].

2.3. Decision Trees Model Fitting

In decision tree construction, C5.0 takes the information gain rate as the standard to determine the best grouping variable and segmentation point, and it considers the size of the information gain and the cost of obtaining information [18]. The higher the information gain rate of variables, the better it is to use them as grouping variables. Different from the C5.0 algorithm, the CART tree selects Gini coefficient as the split attribute and selects the feature with the largest Gini coefficient to divide [19].

In boosting technology, each step will produce a weak classification prediction model. In this paper, C5.0 and CART models are used as weak classifiers to perform weighted accumulation to obtain a new model. In this way, a model with weak classification prediction ability can be cascaded to obtain a model with strong classification prediction ability [20].

The basic idea of the algorithm is derived based on the given weak learning algorithm and training set such as Equation (8)

$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m). \tag{8}$$

First, initialize the distribution of the training set $D_1(i) = \frac{1}{m}$, then perform *T-round* training. In the *t-th* cycle, the weak learning algorithm is trained under the weight D_t to obtain the weak classifier h_t.

At the same time, calculate the error rate of the weak classifier with Equation (9) under the weight D_t:

$$\varepsilon_t = \sum_{i=1}^{N} D_t(x_i)[h_t(x_i) \neq y_i]. \quad (9)$$

Weight is updated with the error rate: $D_{t+1}(i) = D_t(i)\exp(-a_t y_i h_t(x_i))/Z_t$.

When the $\alpha_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ is satisfied; ε_t is the error rate of weak classifier h_t under weight D_t, and classifier is satisfied when $h_t(x_i) = y_i$, $y_i h_t(x_i) = 1$, otherwise $y_i h_t(x_i) = -1$, and Z_t is the normalization factor [16]. The final output strong classifier is expressed in Equation (10).

$$H(x) = sign\left(\sum_{i=1}^{T} a_t h_t(x)\right). \quad (10)$$

By application of the generated strong classifier to the corresponding data set, better prediction accuracy can be expected to be obtained [20].

3. Empirical Analysis

With the preprocessing, data are deleted or supplemented to be kept consistent and relevant for the data mining. Decision model fitting is considered with C5.0 and CART algorithm, and VIF calculation is carried out to find the possibility for the application of dimension reduction.

3.1. Data Introduction

Considered data comprise bank customer information, and we evaluate their credit by decision tree proposal. The data include 16 attributes with 7 continuous and 9 discrete variables, and the target variable is whether the customer is trustworthy or not. Specific data attribute information is shown in Table 2 [21].

Table 2. Variable description for customer's personal information.

Variable type	Serial Number	Attributes	Property Description
Continuous variables	1	age	Age
	2	balance	Annual average balance
	3	day	last contact day of the week
	4	duration	last contact duration
	5	campaign	Number of customer contacts during campaign
	6	pdays	number of days that passed by after the customer was last contacted from a previous campaign
	7	previous	number of customer contacts before this campaign
Discrete variables	8	job	job type
	9	marital	marital status
	10	education	education type
	11	default	has credit in default?
	12	housing	has housing loan?
	13	loan	has personal loan?
	14	contact	contact communication type

Table 2. *Cont.*

Variable type	Serial Number	Attributes	Property Description
	15	month	last contact month of the year
	16	poutcome	outcome of the previous marketing campaign
Target variable	17	y	has the client subscribed to a term deposit?(no or yes)

3.2. Additional Data Analysis

The correlation coefficient can be used to describe the correlation between quantitative variables, and the Pearson product difference correlation coefficient can be used to measure the degree of linear correlation between two quantitative variables. The Pearson correlation coefficient is used here to measure the correlation between continuous variables. The results are shown in Table 3.

Table 3. Pearson correlation coefficient.

	Age	Balance	Day	Duration	Campaign	Pdays	Previous
age	1	0.0987	−0.0091	−0.0047	0.0048	−0.0238	0.0013
balance	0.0978	1	0.0045	0.0216	−0.0146	0.0034	0.0167
day	−0.0091	0.0045	1	−0.0302	0.1625	−0.093	−0.0517
duration	−0.0046	0.2016	−0.0302	1	−0.0846	−0.0016	0.0012
campaign	0.0048	−0.0146	0.1625	−0.0846	1	−0.0886	−0.0329
pdays	−0.2038	0.0034	−0.093	−0.0016	−0.0886	1	0.4548
previous	0.0013	0.1067	−0.0517	0.0012	−0.0329	0.4548	1

According to Table 3, the correlation coefficients between the variables are all less than 0.5, so it is concluded that there is no obvious correlation between the variables. Therefore, the original 17 variables are used, of which 16 are input variables and 1 is output. Obviously, the highest correlation is between pday and previous with 0.4548.

VIF can be used to judge the multi-collinear relationship between continuous variables. $\sqrt{\text{VIF}}$ indicates the degree to which each variable can be expanded to predictive variables, and the results are shown in Table 4.

Table 4. Variance inflation factor.

	Age	Balance	Day	Duration	Campaign	Pdays
$\sqrt{\text{VIF}}$	1.010472	1.010426	1.034284	1.00869	1.040001	1.015155

It can be seen from Table 4 that all variables $\sqrt{\text{VIF}} < 2$. Therefore, there is no problem of multi-collinearity between variables, and there is no need to perform PCA to reduce dimensionality to eliminate multi-collinearity.

3.3. Decision Model Fitting and Receiver Operating Characteristic

When the model is fitted with the C5.0 algorithm and CART algorithm, the number of all samples is 40,690 as shown in Table 5.

Table 5. C5.0 model fitting results.

Classification Tree	C5.0	CART
Number of samples	40,690	40,690
Number of predictors	16	2
Tree size	414	74

The CART algorithm's result is rather different from the decision tree generated by the C5.0 algorithm. The C5.0 algorithm uses most of the 16 attributes. Whereas the Gini coefficient is used by dividing attributes for the CART algorithm, only 2 of the 16 attributes are used, which are duration and poutcome. The size of the generated decision tree is 414 and 74 for C5.0 and CART; that is, the number of decisions satisfies 414 and 74, respectively.

4. Decision Model and its Evaluation

In this section, the decision tree model with boosting algorithm is implemented and applied to experiments. To obtain the optimal discrimination, the model has been evaluated through confusion matrix and cross-validation.

4.1. C5.0 with Boosting Algorithm

C5.0 decision tree with boosting algorithm is applied to actual data. The confusion matrix/cost matrix addition cases are also considered. The test data set results with confusion matrix are illustrated in the tables below.

From the 4521 test samples, Table 6 shows that the C5.0 model predicts 4094 (3853 + 241) samples accurately, and 427 (288 + 139) samples are incorrectly predicted with an error rate of 9.4%. Table 7 shows that the C5.0 model predicts 4051 (3679 + 372) samples after adding the cost matrix, and 470 (157 + 313) samples are incorrectly predicted with an error rate of 10.9%. Table 8 indicates that the C5.0 model predicts 4084 (3845 + 239) samples after adding the boosting algorithm, and 437 (290 + 147) samples are incorrectly predicted with an error rate of 9.6%. The error rate of the C5.0 model with the added cost matrix is slightly higher than the other error rates of the model, and after adding the boosting algorithm, the C5.0 model will not significantly improve the performance of the C5.0 algorithm. Next, we analyze the performance of C5.0 algorithm from the accuracy, precision, and sensitivity viewpoint.

Table 6. Confusion matrix with C5.0 model.

Actual Default	Predicted Default	
	No	**Yes**
No	3853	139
Yes	288	241

Table 7. Confusion matrix with C5.0 model and added cost matrix.

Actual Default	Predicted Default	
	No	**Yes**
No	3679	313
Yes	157	372

Table 8. Confusion matrix with C5.0 model with added boosting.

Actual Default	Predicted Default	
	No	**Yes**
No	3845	147
Yes	290	239

From the calculation results of Table 9, C5.0 model's accuracy, precision, and sensitivity are 0.9055, 0.4556, and 0.4558, respectively. The sensitivity is 0.4558, which means that 45.58% of potential customers are correctly classified. Whereas precision and sensitivity measures with cost matrix (CM) are increased around 25% compared with only C5.0 tree.

Table 9. C5.0, C5.0 + CM and C5.0 model with added boosting model performance metrics.

Model Performance Metrics	C5.0	C5.0 + CM	C5.0 + Boosting
Accuracy	0.9056	0.8961	0.9033
Precision	0.4556	0.7034	0.4518
Sensitivity	0.4558	0.7032	0.6192
Specificity	0.9652	0.9215	0.9299
Kappa value	0.4793	0.5543	0.5224

However, the number of samples in each category is not often balanced in actual classification problems. If there is no adjustment on this kind of unbalanced data set, the model is easily biased towards the big category and the small category is ignored [22]. Hence, an index can be considered to punish the bias of the model to increase the accuracy in this time [23]. According to the calculation formula of *Kappa*, the more unbalanced the confusion matrix, the lower the illustrated *Kappa value*. It gives a low score to the model with strong biasedness. Therefore, the higher the *Kappa value* selected, the better the represented model performance [4].

The C5.0 model with cost matrix's sensitivity is 0.7032, which means that 70.32% of the customers who confirm the subscription deposit are correctly classified. The Kappa value of C5.0 model without the cost matrix is lower at 0.4793 compared with C5.0 + CM. Thus, the sensitivity and Kappa value of the results after fitting the C5.0 model with the cost matrix are significantly improved. In a brief summary, the C5.0 model with cost matrix can be improved more accurately and classify potential users, and is more suitable for dealing with imbalanced data sets.

For the accuracy point of view, accuracy of the model with boosting has not changed significantly. Together with the boosting algorithm, cross-validation is added to obtain the average performance of the model. The results are as follows:

It can be seen from the results in Table 10 that 8 candidate models were tested. The results show that trials = 1 provides the best performance according to the Kappa value; trials = 25 provides the best performance according to the accuracy rate, but the Kappa value is not ideal, so choosing a model with trials = 1 not only results in better computing performance but also reduces the possibility of overfitting.

Table 10. Adapted model with cross-validation.

Trials	Mean-Accuracy	Kappa
1	0.9044	0.4983
5	0.9030	0.4775
10	0.9045	0.4609
15	0.9053	0.4793
20	0.9064	0.4804
25	0.9066	0.4854
30	0.9065	0.4805
35	0.9065	0.4812

4.2. CART with Boosting Algorithm

The CART decision tree with boosting algorithm is considered and the test data set with confusion matrix are illustrated in the tables below.

Table 11 shows that the CART model predicts 4084 (3886 + 197) samples accurately, and the model prediction accuracy rate is 90.31%. The accuracy of the model fitting result shows not much difference from the C5.0 algorithm, and the correct classification ability of the model is satisfactory.

Table 11. Confusion matrix with the CART model.

Actual Default	Predicted Default	
	No	**Yes**
No	3886	106
Yes	332	197

Table 12 shows the result after the boosting algorithm is added to the CART model; the number of samples is predicted accurately at 4112 (3855 + 257) samples, and the model prediction accuracy rate is 90.95%. The accuracy rate has increased slightly from 90.31% to 90.95%.

Table 12. Confusion matrix with CART model and added boosting.

Actual Default	Predicted Default	
	No	**Yes**
No	3855	272
Yes	137	257

4.3. Cross-Validation

K-fold cross-validation is commonly used to evaluate model performance [24]. Cross-validation is a different approach from the repeated random sampling from the sample set. K-fold cross-validation divides all samples into K group separately; then each part is called a fold. When 10-fold cross-validation is adopted, we randomly divide the data set into 10 parts and use 9 of them for training and the other 1 for testing. This process is repeated 10 times. The process of training and testing the model is repeated 10 times, and the output results of 10 times are obtained with an average performance index [25,26].

After the model is fitted, 10-fold cross-validation is used for each of the five algorithms to obtain the accuracy of 10 model checks, and then the average accuracy is calculated. Compared with the accuracy of a model prediction obtained by the confusion matrix, the accuracy of the 10-fold cross-check is more suitable for evaluating the performance of the model.

As can be seen from the above results in Table 13, C5.0 model with CM sacrifices the accuracy of the model to improve the sensitivity of model fitting, thereby ensuring a more accurate classification of potential customers.

Table 13. Accuracy of 10-fold cross-validation.

Fold	Accuracy				
	C5.0	**C5.0 + CM**	**C5.0 + Boosting**	**CART**	**CART + Boosting**
1	0.910	0.898	0.904	0.910	0.92
2	0.905	0.899	0.903	0.904	0.906
3	0.904	0.907	0.905	0.903	0.907
4	0.906	0.901	0.906	0.906	0.908
5	0.901	0.906	0.903	0.902	0.9
6	0.901	0.904	0.905	0.901	0.906
7	0.902	0.905	0.906	0.902	0.91
8	0.900	0.901	0.902	0.89	0.905
9	0.897	0.900	0.906	0.895	0.908
10	0.902	0.895	0.905	0.902	0.901
Mean-Accuracy	0.903	0.901	0.905	0.902	0.907

As shown in Table 13, the performance of the ranking model according to the average accuracy is illustrated in Equation (11):

$$\text{CART} + \text{Boosting} > \text{C5.0} + \text{Boosting} > \text{C5.0} > \text{CART} > \text{C5.0} + \text{CM} \quad (11)$$

The accuracy rates of the five models are all high, all above 90%; this indicates that the models have better prediction performance for the sample data.

From the result, CART + boosting and C5.0 + boosting algorithms show satisfactory average accuracy; this means that the boosting algorithm can enhance the performance of the model.

5. Discussion

According to the positive prediction in Table 1, the calculated model evaluation index values could be different. Therefore, the different result could be derived. The positive class is considered insofar as it should be more concerned with practical applications. In this paper, *Positive = yes* means that bank customers who subscribe to fixed deposits are considered as positive, and bank customers who do not subscribe to fixed deposits are denoted as negative. In this category, more attention should be paid to the model's ability to correctly classify the potential users. Among them, precision, sensitivity, and specificity are assumed as evaluation indicators for calculating a certain classification characteristic.

Accuracy, F1 score, and model fitting time are the criteria for judging the overall classification model.

Evaluation indices of *Positive = yes* are illustrated in Table 14, and the overall prediction accuracy for each model shows a small amount of difference. The CART + boosting model represents the highest accuracy, reaching 90.95%, and the C5.0 + CM model has an accuracy of 89.60%, which is the lowest among the five models. At the same time, by comparing CART model and CART + boosting model, the precision increased from 37.24% to 65.23%, which means that the model's ability to predict potential customers has been improved. After the CART model was added to the boosting algorithm, the F1 score increased from 47.36% to 55.69%. Therefore, by adding the boosting algorithm to the CART model, the performance can be improved drastically, but the CART model after adding the boosting algorithm needs a long time to classify large data sets.

Table 14. *Positive = yes* confusion matrix evaluation index.

	C5.0	C5.0 + CM	C5.0 + Boosting	CART	CART + Boosting
Accuracy	0.9056	0.8960	0.9033	0.9031	0.9095
Precision	0.4560	0.7032	0.4518	0.3724	0.6523
Sensitivity	0.6324	0.5431	0.6192	0.6502	0.4858
Specificity	0.9305	0.9591	0.9299	0.9213	0.9657
F1 score	0.5305	0.6129	0.5224	0.4736	0.5569
Model fitting time	4 s	4 s	8 s	10 s	2 min

From the results of accuracy and F1 values, the model performance has not been improved after the C5.0 with boosting algorithm. Because the C5.0 algorithm is mainly strengthened by increasing the number of iterations according to the data in Table 13, the C5.0 model is the optimal model when *trials = 1*, so the improvement effect of the algorithm is not significant.

However, after adding the CM, although the model's total sample prediction accuracy decreased, the precision and F1 score have been improved. After considering the addition of the confusion matrix to the C5.0 model, the precision and F1 values surpassed the other four models, being 70.32% and 61.29%, respectively. Not only does the performance of correctly classifying potential users show the best, but the model fitting time also becomes shorter. Hence, it is the best model to classify potential users correctly. Therefore, in the case of *Positive = yes*, the C5.0 model illustrates the best performance by adding the CM.

In Table 15, the evaluation indices of *Positive = no* are illustrated. In this case, the bank customers who do not subscribe to the fixed deposit are considered as positive. By the comparison with Table 14, it can be found that the total sample prediction accuracy rate is unchanged, but precision and sensitivity have been greatly increased, while specificity has decreased. This is because in the overall sample, the number of customers who will not subscribe to fixed deposits is far greater than the number of customers who will subscribe to fixed deposits. By observing the four indices of accuracy, precision, sensitivity, and F1 score, it is found that the overall performance difference between the five models is very small. It is worth noting that the accuracy and F1 score of the CART model after the boosting algorithm are still improved; this shows that the boosting algorithm can indeed enhance the performance of the model, but the effect is not significant when *Positive = no*.

Table 15. *Positive = no* confusion matrix evaluation index.

	C5.0	C5.0 + CM	C5.0 + Boosting	CART	CART + Boosting
Accuracy	0.9056	0.8960	0.9033	0.9031	0.9095
Precision	0.9652	0.9216	0.9632	0.9734	0.9341
Sensitivity	0.9305	0.9591	0.9299	0.9213	0.9657
Specificity	0.6342	0.5431	0.6192	0.6502	0.4858
F1 score	0.9475	0.9399	0.9463	0.9466	0.9496
Model fitting time	4 s	4 s	8 s	10 s	2 min

6. Conclusions

This paper introduces the basic principles of the C5.0 algorithm model and the CART algorithm model and uses the personal information data of 45,211 customers of the Bank of Portugal, seven continuous variables, and nine discrete variables to conduct an empirical study on whether they subscribe to fixed deposits. The matrix confusion method and cross-validation method are used to compare the performance of the model. This paper fits two basic models, namely, C5.0 algorithm model and CART algorithm model. Based on each algorithm, a boosting algorithm is added, and a cost matrix (CM) is added to the C5.0 algorithm for model fitting. In the final comparison of models, the accuracy, F1 score, and the average accuracy of the 10-fold cross-check are used to evaluate the overall performance of the model. According to the recall, precision, and specific indicators, a certain classification feature is calculated to evaluate the specific classification of the model performance for the given banking data. The test results show:

(1) The performance improvement of C5.0 algorithm after combining with the boosting algorithm is not significant. This is because the experimental data set is an unbalanced data set (the number of customers who do not subscribe to the time deposit is much higher than the number of customers who subscribe to the time deposit). Experiments on this kind of unbalanced data set, if the model is not adjusted, are easy to bias towards the big category and give up on the small category. Table 10 experiments show that when the number of iterations is 1 (*trials = 1*), it is the C5.0 algorithm itself. The highest Kappa value indicates that the C5.0 algorithm has the lowest bias. Compared with the model after adding the boosting algorithm, the ability to deal with imbalanced data sets is improved. Therefore, in dealing with the problem of unbalanced data classification, the performance improvement of C5.0 algorithm combined with the boosting algorithm is not significant.

(2) Among all the fitted models, the sensitivity of the model fitted by the C5.0 algorithm by adding the CM is shown to be 13% and 54% higher than CM + boosting and CM only, respectively. The results are illustrated in Table 9. Therefore, we must consider the problem comprehensively, and we need to choose the model for the consideration of accuracy, sensitivity, or others. After the requirements are clarified, the model is further fitted and compared; the enhancement algorithm is a combination of multiple weak classifier models, which has some fitting effects to the better model.

The boosting algorithm may not significantly improve the performance. Therefore it is required to choose a model with lower computational complexity and better fitting. For example, if a boosting algorithm is added to the ID3 algorithm, the effect will be more significant.

(3) The bank customer classification problem is carried out as an example. In an actual decision problem, the speed of model fitting is also a factor that needs to be considered. On the one hand, this article conducts a classified evaluation on whether bank customers will subscribe to fixed deposits. For customers who subscribe to time deposits, it is recommended to use the C5.0 model with CM because the higher sensitivity can improve the performance of the model for classifying potential users. It predicts more customers who will subscribe to time deposits, and will facilitate the bank's business development. Furthermore it is also necessary to make predictions for users who will not book fixed deposits. The banking business covers a wide range, and other financial services can be promoted. Because the data set is large, it is recommended to use the C5.0 model to make predictions. The time is shorter, the model performance difference is small, and the accuracy rate is rather high.

Finally, the analysis of the proposed methodology can provide a more reliable basis for decision makers. How to set other better indicators to measure model performance, and how to determine whether the model to be compared is comprehensive are all issues that need to be discussed later in this article, and more in-depth research would be expected.

Author Contributions: Conceptualization, S.-P.J. and S.L.; methodology, L.Z.; software, L.Z.; validation, S.-P.J. and S.L.; resources, L.Z.; writing—review and editing, L.Z. and S.L.; supervision, S.-P.J.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Centre for Smart Grid and Information Convergence (CeSGIC) at Xian Jiaotong—Liverpool University.

Data Availability Statement: The data that support the findings of this study are available from the open source with blind information and they are processed without personal information.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Quinlan, J.R. Quinlan, Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
2. Rabcan, J.; Levashenko, V.; Zaitseva, E.; Kvassay, M.; Subbotin, S. Application of Fuzzy Decision Tree for Signal Classification. *IEEE Trans. Ind. Inform.* **2019**, *15*, 5425–5434. [CrossRef]
3. Sun, R.; Wang, G.; Zhang, W.; Hsu, L.-T.; Ochieng, W.Y. A gradient boosting decision tree based GPS signal reception classification algorithm. *Appl. Soft Comput.* **2020**, *86*, 105942. [CrossRef]
4. Drucker, H.; Cortes, C. Boosting Decision Trees. In *Advances in Neural Information Processing Systems 8*; NIPS: Stockholm, Sweden, 1995; pp. 27–30.
5. Hunt, E.B.; Marin, J.; Stone, P.J. *Experiment in Induction*; Academic Press: New York, NY, USA, 1966.
6. Michie, D.; Spiegelhalter, D.J.; Taylor, C.C. *Machine Learning, Neural and Statistical Classification*; Oversea Press: New York, NY, USA, 2009.
7. Watanabe, T.; Suzuki, E. Outlier Detection Based on Decision Tree and Boosting. In Proceedings of the 16th Annual Conference of Japanese Society for Artificial Intelligence, Tokyo, Japan, 29–31 May 2002.
8. Parvin, H.; MirnabiBaboli, M.; Alinejad-Rokny, H. Proposing a classifier ensemble framework based on classifier selection and decision tree. *Eng. Appl. Artif. Intell.* **2015**, *37*, 34–42. [CrossRef]
9. Niu, H.; Khozouie, N.; Parvin, H.; Alinejad-Rokny, H.; Beheshti, A.; Mahmoudi, M.R. An Ensemble of Locally Reliable Cluster Solutions. *Appl. Sci.* **2020**, *10*, 1891. [CrossRef]
10. Tanaka, T.; Kasahara, R.; Kobayashi, D. Efficient logic architecture in training gradient boosting decision tree for high-performance and edge computing. *arXiv* **2018**, arXiv:1812.08295. Available online: https://arxiv.org/abs/1812.0829 (accessed on 20 December 2018).
11. PANG, S.L.; GONG, J.Z. C5.0 classification algorithm and its application on individual credit score for banks. *Syst. Eng. Theory Pract.* **2009**, *29*, 94–104. [CrossRef]
12. Ahmed, A.; Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **2007**, *63*, 503–527. [CrossRef]

13. Wang, H.; Jiang, Y.; Wang, H. Stock return prediction based on Bagging-decision tree. In Proceedings of the 2009 IEEE International Conference on Grey Systems and Intelligent Services(GSIS 2009), Nanjing, China, 10–12 November 2009. [CrossRef]
14. Yao, X.; Li, X.; Su, Q. Study on the customer relationship management and its application in Chinese hospital. In Proceedings of the 2005 International Conference on Services Systems & Services Management, Chongqing, China, 13–15 June 2005; Volume 1, pp. 188–192. [CrossRef]
15. Coffman, D.L.; Maydeu-Olivares, A.; Arnau, J. Asymptotic distribution free interval estimation: For an intraclass correlation coefficient with application to longitudinal data. *Methodology* **2008**, *4*, 4–9. [CrossRef]
16. De Saint-Exupery, A.; Capra, F. Meta-analytic design patterns. In *Meta-Analytics*; Wiley: Hoboken, NJ, USA, 2019. [CrossRef]
17. Zheng, H.; Wang, R.; Yu, Z.; Wang, N.; Gu, Z.; Zheng, B. Automatic plankton image classification combining multiple view features via multiple learning. In Proceedings of the 16th international Conference on Bioinformatics (InCoB 2017): Bioinformatics, Shenzhen, China, 20–22 September 2017; Volume 18, p. 570.
18. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; CRC Press: New York, NY, USA, 1999.
19. Gehrke, J.; Ramakrishnan, R.; Ganti, V. RainForest—A Framework for Fast Decision Tree Construction of Large Datasets. *Data Min. Knowl. Discov.* **2000**, *4*, 127–162. [CrossRef]
20. Cheng, K.-C.; Huang, M.-J.; Fu, C.-K.; Wang, K.-H.; Wang, H.-M.; Lin, L.-H. Establishing a Multiple-Criteria Decision-Making Model for Stock Investment Decisions Using Data Mining Techniques. *Sustainability* **2021**, *13*, 3100. [CrossRef]
21. Mukherjee, S.; Deyasi, A.; Bhattacharjee, A.K.; Mondal, A.; Mukherjee, A. Role of Metaheuristic Oprimization in Portfolio Management for the Banking Sector: A Case Study. In *Metaheuristic Approaches to Portfolio Optimization*; Ray, J., Mukherjee, A., Dey, S.K., Klepac, G., Eds.; IGI Global: Hershey, PA, USA, 2019; Chapter 9; pp. 198–220.
22. Polaka, I.; Tom, I.; Borisovs, A. Decision Tree Classifiers in Bioinformatics. *Sci. J. Riga Tech. Univ.* **2010**, *44*, 119–124. [CrossRef]
23. Rokach, L.; Maimon, O. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2015.
24. Buffington, J.; Elliott, R.J. Regime Switching and European Options. In *Stochastic Theory and Control*; Springer: Berlin, Germany, 2002; Volume 280, pp. 73–82.
25. Wada, A.; Tsuruta, K.; Irie, R.; Kamagata, K.; Maekawa, T.; Fujita, S.; Koshino, S.; Kumamaru, K.; Suzuki, M.; Nakanishi, A. Differentiating Alzheimer's Disease from Mementia with Lewy Bodies using a Deep Learning Technique based on Structral Brain Connectivity. *Magn. Reson. Med. Sci.* **2019**, *18*, 219–224. [CrossRef] [PubMed]
26. Everitt, B.S.; Skrindal, A. *The Cambridge Dictionary of Statics*, 4th ed.; Cambridge University Press: Cambridge, UK, 2010.