

Article

Establishing a Multiple-Criteria Decision-Making Model for Stock Investment Decisions Using Data Mining Techniques

Kuo-Chih Cheng ¹, Mu-Jung Huang ^{1,*}, Cheng-Kai Fu ¹, Kuo-Hua Wang ², Huo-Ming Wang ² and Lan-Hui Lin ¹

¹ Department of Accounting, National Changhua University of Education, Changhua 500, Taiwan; h12343562@ms46.hinet.net (K.-C.C.); mjhuang8601@gmail.com (C.-K.F.); linlh@cc.ncue.edu.tw (L.-H.L.)

² Department of Finance, National Changhua University of Education, Changhua 500, Taiwan; wanglee1968@gmail.com (K.-H.W.); bikewang@gmail.com (H.-M.W.)

* Correspondence: mjhuang@cc.ncue.edu.tw

Abstract: This study attempts to integrate the decision tree algorithm with the Apriori algorithm to explore the relationship among financial ratio, corporate governance, and stock returns to establish a stock investment decision model. The sports and leisure related industries are employed as the research target. The data are collected and processed for generating decision tree and association rules. Based on the analysis outcome, an investment decision model is constructed for investors expecting to decrease their investment risks and further increase their profits. This stock investment decision model is one type of multiple-criteria decision-making model. This study makes three critical contributions to investors. (1) It proposes a systematical model of exploring related data through the decision tree algorithm and the Apriori algorithm to reveal the implicit investment knowledge. (2) An effective investment decision model is established and expected to provide a reference basis during stock-picking decisions. (3) The investment decision model is enhanced with implicit rules found among variables using association rules.

Keywords: data mining; apriori algorithm; multiple-criteria decision-making; association rules; decision tree



Citation: 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. <https://doi.org/10.3390/su13063100>

Academic Editor: Tsu-Ming Yeh

Received: 10 February 2021

Accepted: 2 March 2021

Published: 11 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, scholars have conducted a lot of research on stock prices or the rate of return. It can be divided into two categories, namely fundamental analysis and technical analysis. Fundamental analysis is to inspect a company's finances, internal operations, industrial market conditions, domestic and foreign political and economic situations, etc., including a company's financial statements and non-financial information, as long as you can find the embedded value of the stock, which can predict the stock price rise or fall. Technical analysis is to believe that history will repeat itself and stock prices will follow specific trends shifting to the equilibrium point [1,2]. As long as the past stock price and volume changes are analyzed, the stock price trend can be predicted. Although there are many factors that affect stock price, the most basic is the company's own value. The financial statements of each company reflect the most direct and quantitative information that is an important reference indicator for evaluating the value of a company. Therefore, this study tries to explore the relationship between corporate governance (presented by financial ratios) and stock price [3–5].

Concerning stock price, and in terms of return on investment, current major scholars still choose to use artificial neural networks for research. However, because the artificial neural network learns by establishing a numerical structure, its knowledge structure is implicit, lacks explanatory ability, and difficult to interpret. In contrast, decision tree, one of the data mining techniques, learns by inductive learning to establish symbols, and its knowledge structure is explicit and more explanatory [6,7]. Thus, this study decided to adopt the decision tree in data exploration, which is a method that is commonly used for

prediction and classification. In addition, this study will also use association rules—another data mining technique—to explore and to find highly-correlated rules and hidden rules in data, to integrate with the results of decision tree exploring.

Not only would such an approach lead to more accurate results throughout the huge amount of analyzed data, but it could also be useful to make predictions about the future. In view of this, this study intends to apply data mining techniques to explore the relationship between corporate governance and stock price. The main purpose of this study tries to use data mining techniques, decision tree, and association rules, to extract from relevant data on corporate governance and financial ratios to find implied rules and knowledge, which are used to construct investment decision models to help investors. It can be used as a basis for making stock selection decisions. This stock investment decision model is one kind of multiple-criteria decision-making model. The purpose of this study is summarized as follows:

1. Use decision tree analysis to find easy-to-understand classification rules among the data to construct an investment decision model.
2. The decision tree analysis may not be able to see the degree of mutual influence among the variables; therefore, this study joined another algorithm, the Apriori algorithm in association rules, to supplement the explanation of the mutual influence of various variables.

2. Literature Review

2.1. Data Mining

Data mining is an automatic or semi-automatic process, which integrates and analyzes raw data to obtain potential knowledge or a new relationship [8]. Data mining is also commonly referred to as knowledge mining in database. It is from a large database, to extract interesting knowledge that is meaningful in decision-making, and is also an application area that can provide significant competitive advantage to an organization [5,6,9]. Designing a framework for a knowledge discovery process is critical. Researchers have described a series of steps that constitute a knowledge discovery process, ranging from a few steps to more sophisticated models (e.g., the nine-step model suggested by Fayyad et al.) [10].

Data mining approaches, such as prediction, classification, clustering, etc., have been widely applied in solving many real-world problems, i.e., business [11] and finance [12]. Data mining is widely applied as it provides solutions to complex systems where conventional models are inappropriate [13].

Data mining tasks can generally be classified into five main categories: prediction, classification, association, estimation, and clustering. Prediction is commonly referred to as the act of telling the future. Classification, or supervised induction, is to analyze the historical data stored in a database and generate a model that can predict future behavior. Association, or association rule learning in data mining, is a well-researched technique for discovering interesting relationships among variables in big data. Two commonly used derivatives of association rule learning are link analysis and sequence mining. With link analysis, the linkage among many objects of interest is discovered. With sequence mining, relationships are examined in terms of their order of occurrence to identified association over time. Algorithms used in association rules include the popular Apriori and FP-Growth (Frequent Pattern) [6,14].

2.2. Decision Tree

Decision tree learning (a method of the data mining technique) is recommended because of its richness of classification arithmetic rules and the appeal of visibility. Decision tree learners are the workhorses of many practical applications of machine learning because they are fast and produce intelligible output that is often surprisingly accurate [15].

The main function of a decision tree is to create a tree structure by classifying known cases structures and to sum up certain laws in the cases from it; the resulting decision tree can use it to make out-of-sample predictions. In past studies, the evolution of decision

trees, in general classification, was often used; it is the most representative method [16] compared with other data mining methods. The data type limit that can be entered in the decision tree is comparison loose and the accuracy of prediction and explanatory power are similar with other methods. These advantages make the decision tree analysis widely used in various fields. The main algorithms of the decision tree include the decision tree analysis mode. The main algorithms include ID3, C4.5, and CART [17,18]. The use of the decision tree algorithm is more accurate than the traditional statistical segmentation method to have more correct analysis results. The use of the decision tree algorithm to present the problem relationship is clearer than using general traditional statistical methods [19,20].

2.3. Association Rules

Association rules were first developed by scholars—Agrawal, Imielinski, and Swami—in 1993 [21], and mainly used to find the relationship between items and attributes in the data, or some hidden relationships between data. Thus, association rules are often used to explore the sales relationships between different commodities or the consumption habits of customers. Apriori is the most widely used algorithm in association rules [14].

The form of association rules can be expressed as $X \rightarrow Y$. The analysis of association rules is mainly used to mine rules hidden in huge datasets. However, the number of rules generated is too large, or the rules do not have substantial meaning; thus, association rules usually need to be performed according to support and confidence to prune. This algorithm is used to operate based on the threshold of support and the threshold of confidence.

Predictive rules (found by the constrained association rule mining) are more abundant and have higher reliability than predictive rules induced by decision trees. Association rules, compared to decision trees, tend to have higher confidence; they involve larger subsets of the dataset, they work better with user-defined binning, and they are easier to interpret [22–24]. Association rule mining proves that the approach has higher classification accuracy than other famous techniques, such as SVM, Naive Bayes, neural networks etc. [25].

2.4. Stock Investment Using Data Mining Techniques

Stock forecasting is an important part of stock investment management. Wang and Hu (2019) [26] applied data mining techniques to select stock in investment management of commercial stock markets. In their study, stock was predicted using the data mining method, and back propagation neural network (BPNN) was taken as the basis. The genetic algorithm was used for optimization to obtain the improved BPNN algorithm, and then it was applied to stock prediction. Their study provides some theoretical support for the further application of data mining techniques in stock forecasting, which is helpful for investors to make correct stock choices, improve returns, and avoid risks.

Chen and Hsieh (2016) [27] proposed a domain-driven stock portfolio optimization approach based on the domain-driven data mining concept that can satisfy an investor's requests for mining an actionable stock portfolio using a genetic algorithms.

Ng and Khor (2017) [28] also proposed a stock profiling framework, StockProF, for building stock portfolios rapidly utilizing data mining approaches. The authors utilized the financial data of the plantation stocks listed on Bursa Malaysia and used 1-year stock price movements to evaluate the performance of the clusters and the outliers. The results showed that StockProF is effective as the profiling corresponded to the average capital gain or loss of the plantation stocks.

3. The Methodology

3.1. The Framework of a Stock Investment Decision Model

As shown in Figure 1, this study uses data mining techniques and algorithms in Weka 3.6 software to discover the implicit rules and knowledge within corporate governance, financial ratios, and stock returns, which are used to construct investment decisions model

to help investors use as a reference when making stock selection decisions. The architecture of this research mainly includes the following three parts:

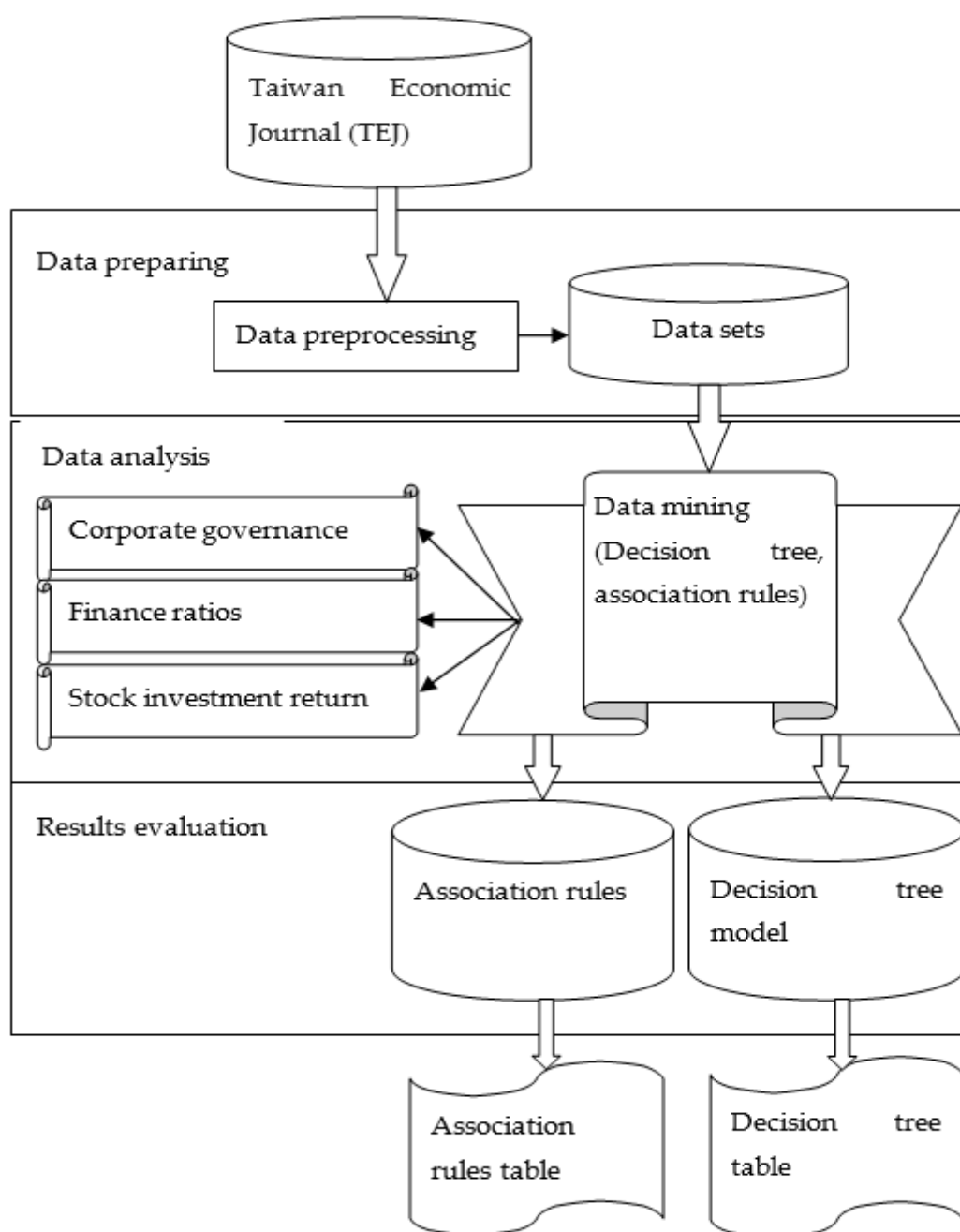


Figure 1. Research architecture.

3.1.1. Data Preparing

At first, this study extracts the related data from the Taiwan Economic Journal (TEJ) database. The collected data are filtered, and then the parts that do not meet the conditions, or incomplete data, are deleted. The data are converted into data types that can be used by subsequent analysis.

3.1.2. Data Analysis

The obtained datasets are then analyzed by decision tree and association rule algorithms. Both are described in the following:

1. Decision tree analysis: the pre-processed data are analyzed by using the J48 classifier established by C4.5 algorithm in Weka 3.6. Before performing decision tree analysis, the data are divided into two parts. One part is used as a training set for building the predict model, and the other is used as the test set for testing the accuracy of the model. In the process of data exploration, in order to ensure that the obtained model has better accuracy, a simple verification method was used to divide the original dataset into two parts: approximately 66.67% as training data to build the model, and approximately 33.33% as a test sample to test the accuracy of the model.
2. Analysis of association rules: in order to prevent the number of rules generated from being too large, or if the rules do not have substantial meaning, association rules usually need to be performed according to support and confidence to prune. In this study, the Weka 3.6 Apriori algorithm is used to set the threshold of support at 10% and the threshold of confidence at 85%. This algorithm is used to operate based on the threshold of support and the threshold of confidence.

3.1.3. Results Evaluation

Interpreting the decision tree model generated by the C4.5 algorithm and the association rules created by the Apriori algorithm, if any rule that is generated by the individual algorithm; it means that the value of the each rule is high. Finally, the each rule is evaluated whether it is useful.

3.2. Meanings of Variables

The variables are summarized in Table 1. The meanings of variables are described in the following [1,2,29].

Table 1. Definition of variables.

Variable Feature	Codename	Variable Name
Dependent variable	Return	Annual rate of return
Independent variable	Price	Stock price
	Year	Years on the market
	DS&F-holding	Directors, supervisors, and foreign shareholding ratio
	DS-Pledge	Pledge rate of shares held by directors and supervisors
	BIG4	Whether it is signed by one of the four large accounting firms
	R&D	Research and development expense rate
	EPS	Earnings per share
	ROE	Return on equity
	ROA	Return on assets
	FA-turnover	Turnover rate of fixed assets
	INV-turnover	Inventory turnover
	A/R-turnover	Accounts receivable turnover rate
	TA-turnover	Total asset turnover
	OR-growing	Revenue growth rate
	OE-growing	Net equity growth rate
	OI	Operating profit rate
	OP	Gross profit rate
	DR	Debt ratio
	CR	Current ratio

Source: Own elaboration.

First, in this study, the annual rate of return is the only one dependent variable: the annual rate of return used for the variable is calculated by taking the closing price at the end of the year. However, the calculation method is as follows: the i company's annual return on investment in year j is calculated according to Equation (1).

$$R_{ij} = \frac{(P_i(j+1) - P_{ij})}{P_{ij}} \quad (1)$$

Among them, P_{ij} is the average stock price of the i -th company in year j . The $j + 1$ means year $j + 1$. Second, there are 19 dependent variables in this study:

(1) Stock price: that is, the closing price refers to the price of a certain security before the end of a day's trading activities on the stock exchange. Because the closing price is the standard for the market quotation of the day and the basis for the opening price of the next trading day, the basis for investors to market can be used to predict the future securities market. Thus, investors are right in the market analysis—the closing price is generally used as the basis for calculation.

(2) Years on the market: companies that have experienced at least one to two business cycles may be more stable in operation and have better operating performance than companies that have just gone public. The three-year rate of return for holding shares of newly listed companies is significantly lower than that of non-newly listed company stocks.

(3) Directors, supervisors, and foreign shareholding ratio: the higher the shareholding ratio of directors, supervisors, and foreign the higher the confidence of the operators in their companies. This means that the interests of directors, supervisors, and foreign relationships with investors are also consistent. Therefore, the possibility of financial crisis for such companies will be lower.

(4) Pledge rate of shares held by directors and supervisors: the high pledge of shares held by directors and supervisors may mean that the directors and supervisors have taken their holdings to the bank for cash (to have reduced the risk). Companies with higher pledge ratios will have worse stock price performance and operational performance in the coming year than companies with low pledge ratios. Then, their companies face relatively higher operational risks.

(5) Whether it is signed by one of the four large accounting firms: based on the audit quality hypothesis, DeAngelo [30] believes that the larger the scale of the accounting firm, the higher the independence of the accountant, and the higher the quality of its audit, the higher the stocks of visa clients should have positive abnormal returns. If it is a company with a visa from the big four accounting firms, it will have a positive impact on the stock price.

(6) Research and development (R&D) expense rate: the higher the R&D expenses a company pays, the fewer the competitors who could also pay such high R&D expenses. Therefore, the R&D expense ratio is positively correlated with the company's current share price.

(7) Earnings per share (EPS): earnings per share is a company's profit indicator; with the company's stock price at a certain linkage, it is also one of the key elements for the company's existing shareholders (and potential investors) to measure the company's profit.

(8) Return on equity (ROE): the return on equity is based on how much net profit can be obtained for every dollar of the company's shareholders' equity. The higher the net profitability, the stronger the operating ability; thus, it is more beneficial to shareholders.

(9) Return on assets (ROA): the return on assets measures a company's profit for every one dollar of assets. Return on assets is often used as one of the indicators of a company's operating performance and represents the company's own profitability.

(10) Turnover rate of fixed assets (FA-turnover): the purpose of enterprises using fixed assets is to improve production efficiency and increase sales receipts. Therefore, the higher the turnover rate fixed assets a company has, the higher the utilization rate of the company's fixed assets, and the greater the benefits for the company.

(11) Inventory turnover (INV-turnover): inventory represents the backlog of company funds. Therefore, the more inventory turnover there is, the more efficient the company's inventory management.

(12) Accounts receivable turnover rate (A/R-turnover): accounts receivable provides customers with short-term credit (on credit) for the company. Accounts receivable turnover rate represents how quickly the company can receive loans. The faster the collection of accounts receivable, the more beneficial it is for the company's capital use.

(13) Total asset turnover (TA-turnover): the total asset turnover rate represents how much sales revenue the company can create for every dollar invested in assets, which means the efficiency of asset utilization.

(14) Revenue growth rate (OR-growing): the revenue growth rate is used to observe the degree of change in the company's operating income during a certain period of time. The higher the rate, the higher the company's sales volume growth.

(15) Net equity growth rate (OE-growing): the net equity growth rate is used to analyze whether the business is profitable and whether the net value of shareholders' equity increases; it reveals whether the goals of the company's management are consistent with shareholders' goals.

(16) Operating profit rate (OI): the operating profit rate is an indicator used to measure the efficiency of a company's operations, reflecting the ability of business managers to make profits.

(17) Gross profit rate (OP): the higher the gross profit rate, the stronger the company's profitability and its ability to control costs.

(18) Debt ratio (DR): an important indicator used to measure the capital structure of a company; it shows the ratio of external borrowing in all assets. The higher the ratio of external borrowing, the weaker the company's financial structure.

(19) Current ratio (CR): this is also called the working capital ratio, due to the principle of the rate of assets being converted into cash within 12 months. It is used to measure the company's short-term repayment ability to debt. Too low of a current ratio means that, if the company has problems, it is likely under the phenomenon of poor turnover.

3.3. The Decision Tree Algorithm

The decision tree induction algorithm is applied to find the decision tree rules. It is binary in that it creates a two-way branch at every split in the tree. The algorithm for mining decision tree rules consist of two principle types: classification tree and regression tree [17].

$$Entropy(A) = \sum_{i=1}^2 \frac{\sum_{j=1}^c f_{ij}}{R} \sum_{j=1}^c P(C_{ij}) \ln(P(C_{ij})) \quad (2)$$

where c is the number of outcome values, f_{ij} is the frequency of outcome j in branch i , R is the total number of records in both branches, and $P(C_{ij})$ is given by the following expression, Equation (3):

$$P(c_{ij}) = \frac{f_{ij}}{\sum_{k=1}^c f_{kj}} \quad (3)$$

For regression trees (i.e., numeric outcomes), the selection of the attribute for the split is done according to the ability of each attribute to produce branches with small deviations in the value of their outcomes. The normalized standard deviation (NSD) is used here for a split. The attribute with the lowest NSD is selected at every branching point. For discrete attributes, the value groups are split between the two branches to minimize the NSD . For numeric attributes, the two-way split is based on a numeric threshold that is derived to minimize the NSD . The NSD for any split is calculated according to the following equation [31]:

$$NSD(A) = \sum_{k=1}^2 \frac{R_k}{R} \sqrt{\frac{\sum_{i=1}^{R_k} V_{ik}^2 - \left(\sum_{i=1}^{R_k} V_{ik}\right)^2}{R_k}} \quad (4)$$

where V_{ik} is the numeric outcome value of record i in branch k , R_k is the number of records in branch k , and R are the total number of records in both branches.

3.4. Apriori Algorithm

The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters “support” and “confidence” are used. Support refers to the items’ frequency of occurrence; confidence is a conditional probability [17]. Items in a transaction form an item set. The algorithm begins by identifying frequent, individual items (items with a frequency greater than or equal to the given support) in the database and continues to extend them to larger, frequent itemsets.

The Apriori algorithm uses the downward closure property, i.e., all of the subsets of a frequent itemset are frequent, but the converse may not be true. The following are the main steps of the algorithm [14]:

1. Calculate the support of item sets (of size $k = 1$) in the transactional database (note that support is the frequency of occurrence of an itemset). This is called generating the candidate set.
2. Prune the candidate set by eliminating items with a support less than the given threshold.
3. Join the frequent itemsets to form sets of size $k + 1$, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support.

4. Case Study

This study uses only one industry to avoid the insignificant influence of independent variables on dependent variables, due to too many industries. In recent years, road running and cycling have gradually become popular. Therefore, this study uses Taiwan’s listed sports and leisure industry as the research object. The data of this study were taken from the 10 years’ data of listed companies in the Taiwan Economic Journal (TEJ). The data were collected from 2005 to the end of 2014 and excluded those with incomplete research variables that could not be calculated. Finally, 133 observations were used to construct the investment decision model.

4.1. Descriptive Statistics

This study conducts a narrative statistical analysis of the collected data. Table 2 shows descriptive statistics of all samples. First, from the perspective of the average annual return rate of concept stocks in leisure-related industries, is 17%. Often we can see whether company managers have confidence in their companies and whether these directors and supervisors are earnestly operating the companies. Their average number of directors, supervisors, and foreign shareholding ratios, and pledge rate of shares held by directors and supervisors, are 27.82% and 4.88% respectively. Overall, sports and leisure related industries need to be strengthened in this respect. As for the research and development expense ratio, the average figure is 1.71%. From this, we know that the investments in research and development expenses are low on average. The average number of years of listing of these companies is 12.8 years, of which 75% are checked by the big four accounting firms.

Let us look at the variables related to financial ratios. The average debt ratio and current ratio are 46.85% and 247.15%, respectively. In other words, the solvency of sports and leisure related industries is considered good. Then we look at the company’s operating capabilities. The average turnover rate, fixed asset turnover, inventory turnover, accounts receivable turnover, and total asset turnover rate are 6.8%, 5.32%, 7.06%, and 1.17%, respectively. In terms of company profitability, the average operating profit ratio, gross profit rate, earnings per share, return on equity, and return on assets are 3.11%, 17.14%, 2.36%, 9.55%, and 5.18%. Finally, see the growth rate. The average revenue growth rate and net equity growth rate are 12% and 7.27%, respectively.

Table 2. Descriptive statistics.

Variable	Sample Number	Max	Min	Average	Standard Deviation
Annual rate of return	133.00	−0.63	2.62	0.17	0.54
Stock price	133.00	2.69	281.00	37.95	45.96
Years on the market	133.00	2.00	25.00	12.80	5.44
Directors and supervisors and foreign and shareholding ratio	133.00	0.00	77.26	27.82	19.88
Pledge rate of shares held by directors and supervisors	133.00	0.00	61.84	4.88	10.60
Whether it is signed by one of the 4 large accounting firms	133.00	0.00	1.00	0.75	0.43
Research and development expense rate	133.00	0.00	6.72	1.71	1.38
Earnings per share	133.00	−3.06	11.20	2.36	2.97
Return on equity	133.00	−36.78	39.92	9.55	13.61
Return on assets	133.00	−32.84	21.87	5.18	8.02
Turnover rate of fixed assets	133.00	0.67	77.12	6.80	10.34
Inventory turnover	133.00	1.54	30.86	5.32	3.48
Accounts receivable turnover rate	133.00	2.86	16.49	7.06	3.14
Total asset turnover	133.00	0.23	2.03	1.17	0.39
Revenue growth rate	133.00	−46.71	680.14	12.00	61.18
Net equity growth rate	133.00	−36.27	98.44	7.27	15.92
Operating profit rate	133.00	−59.66	27.57	3.11	10.96
Gross profit rate	133.00	−14.84	35.36	17.14	6.31
Debt ratio	133.00	2.46	74.62	46.85	13.97
Current ratio	133.00	72.91	3663.80	247.15	369.81

Source: Own elaboration.

4.2. Decision Tree Analysis

This study adopts the J48 algorithm in Weka 3.6 data exploration software. The dataset is divided into two parts. Moreover, 66.67% is used as training data to build the predict model, and the remaining, approximately 33.33%, is used as the test data to test the accuracy of the model. According to the decision prediction model produced by the training data, the result is transformed into a decision table, as shown in Table 3. We obtained eight rules and explained the rules in the following, for investors to use as a reference when making stock selection decisions.

Table 3. Decision table.

No	DS&F-Holding	DS-Pledge	Year	OI	FA-Turnover	OP	OE-Growing	Return
1	<45	-	≤14	-	-	-	-	Bad
2	<45	-	>14	≤6.48	≤6.48	-	-	Bad
3	<45	-	>14	≤6.48	>6.48	-	-	Good ≥ ave *
4	<45	-	>14	>6.48	-	≤16.52	≤7.52	Bad
5	<45	-	>14	>6.48	-	≤16.52	>7.52	Good ≥ ave
6	<45	-	>14	>6.48	-	>16.52	-	Good ≥ ave
7	≥45	≤33	-	-	-	-	-	Good ≥ ave
8	≥45	>33	-	-	-	-	-	bad

Source: Own elaboration. * ave is the abbreviation of average.

There are eight rules generated through decision tree algorithm. Each rule is described as follows:

Rule 1: DS&F-holding (Directors, supervisors, and foreign shareholding) <45 and Year ≤ 14 → Return = Bad

Rule 1 indicates that the return rate of individual stocks is lower than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is less than 45%, and years on the market is less or equal to 14 years.

Rule 2: $DS\&F\text{-holding} < 45$ and $Year > 14$ and $OI \leq 6.48$ and $FA\text{-turnover} \leq 6.48 \rightarrow$ Return = Bad.

Rule 2 indicates that the return rate of individual stocks is lower than the average return rate of the industry, when directors, supervisors, and the foreign shareholding ratio is less than 45%, years on the market is more than 14 years, operating profit rate is less than or equal to 6.48%, and fixed assets turnover is less than or equal to 6.48%.

Rule 3: $DS\&F\text{-holding} < 45$ and $Year > 14$ and $OI \leq 6.48$ and $FA\text{-turnover} > 6.48 \rightarrow$ Return = Good.

Rule 3 indicates that the return rate of individual stocks is larger than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is less than 45%, years on the market is more than 14 years, operating profit rate is less than or equal to 6.48%, and fixed assets turnover is more than 6.48%.

Rule 4: $DS\&F\text{-holding} < 45$ and $Year > 14$ and $OI > 6.48$ and $OP \leq 16.52$ and $OE\text{-Growing} \leq 7.52 \rightarrow$ Return = Bad.

Rule 4 indicates that the return rate of individual stock is lower than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is less than 45%, years on the market is more than 14 years, operating profit rate is more than 6.48%, gross profit rate is less than or equal to 16.52%, and net equity growth rate is less than or equal to 7.52%.

Rule 5: $DS\&F\text{-holding} < 45$ and $Year > 14$ and $OI > 6.48$ and $OP \leq 16.52$ and $OE\text{-Growing} > 7.52 \rightarrow$ Return = Good.

Rule 5 indicates that the return rate of individual stock is larger than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is less than 45%, years on the market is more than 14 years, operating profit rate is more than 6.48%, gross profit rate is less than or equal to 16.52%, and net equity growth rate is more than 7.52%.

Rule 6: $DS\&F\text{-holding} < 45$ and $Year > 14$ and $OI > 6.48$ and $OP > 16.52 \rightarrow$ Return = Good.

Rule 6 indicates that the return rate of individual stock is larger than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is less than 45%, years on the market is more than 14 years, operating profit rate is more than 6.48%, and the gross profit rate is more than 16.52%.

Rule 7: $DS\&F\text{-holding} \geq 45$ and $DS\text{-Pledge} \leq 33 \rightarrow$ Return = Good.

Rule 7 indicates that the return rate of individual stock is larger than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is larger than or equal to 45%, and the pledge rate of shares held by directors and supervisors gross profit rate is less than 33%.

Rule 8: $DS\&F\text{-holding} \geq 45$ and $DS\text{-Pledge} > 33 \rightarrow$ Return = Bad.

Rule 8 indicates that the return rate of individual stock is less than the average return rate of the industry when directors, supervisors, and the foreign shareholding ratio is larger than or equal to 45%, and the pledge rate of shares held by directors and supervisors is larger than 33%.

The decision table shows that when the pledge rate of shares held by company's directors and supervisors is low, the return on individual stocks will exceed the industry average rate of return. Next is the number of years the company has been listed. We can see that, when a company has been listed for a long time, for more than 14 years, the probability of the return rate of individual stocks larger than the average return rate of the industry will increase significantly.

The last four important variables are the operating profit rate, gross profit rate, fixed assets turnover, and net equity growth rate. Operating profit rate and gross profit rate represent the company's profitability and fixed assets turnover, and net equity growth

rate represent the company's operating capacity. A company's profitability and operating capability are all important factors that affect stock returns.

4.3. Association Rules

The Apriori algorithm in Weka 3.6 data exploration software is used to analyze the association rules. First, in order to ensure the degree of association of the obtained rules, we set the minimum support to 10%, and the minimum confidence level to 90%. This standard is used for the exploration of association rules. After the analysis of the Apriori algorithm, we screened out the variables related to the decision tree model, and got a total of six rules. The association rules results were converted into a variable correlation table (Table 4), so that investors can better understand the relationship between variables.

Table 4. Association rules.

Rule No	Condition	Result	Confidence
1	$ROE \geq \text{ave}^*$ and $OI \geq \text{ave}$	$ROA \geq \text{ave}$	100%
2	$ROE \geq \text{ave}$ and $ROA \geq \text{ave}$	$OI \geq \text{ave}$	100%
3	$ROE \geq \text{ave}$	$ROA \geq \text{ave}$ and $OI \geq \text{ave}$	100%
4	$ROA \geq \text{ave}$ and $BIG4 = Y$	$OI \geq \text{ave}$	98%
5	$ROA \geq \text{ave}$ and $OI \geq \text{ave}$	$ROE \geq \text{ave}$	96%
6	$ROA \geq \text{ave}$	$ROE \geq \text{ave}$ and $OI \geq \text{ave}$	94%

Source: Own elaboration. * ave is the abbreviation of average.

The meanings of the six association rules are described in the following:

- Association rule 1 indicates that if the return of equity is larger than or equal to the average of the industry, and the operating profit rate is larger than or equal to the average of the industry, then the return of assets is larger than or equal to the average of the industry.
- Association rule 2 indicates that if the return of equity is larger than or equal to the average of the industry, and the return of assets is larger than or equal to the average of the industry, then the operating profit rate is larger than or equal to the average of the industry.
- Association rule 3 indicates that if the return of equity is larger than or equal to the average of the industry, then the return of assets is larger than or equal to the average of the industry, and the operating profit rate is larger than or equal to the average of the industry.
- Association rule 4 indicates that if the return of assets is larger than or equal to the average of the industry, and the company is signed by one of the four large accounting firms, then the operating profit rate is larger than or equal to the average of the industry.
- Association rule 5 indicates that if the return of assets is larger than or equal to the average of the industry, and the operating profit rate is larger than or equal to the average of the industry, then the return of equity is larger than or equal to the average of the industry.
- Association rule 6 indicates that if the return of assets is larger than or equal to the average of the industry, then the return of equity is larger than or equal to the average of the industry, and the operating profit rate is larger than or equal to the average of the industry.

4.4. Discussion

The contents of the Table 3 decision table and Table 4 association rules are summarized in Table 5 as stock investment criteria. The implications of management and governance are described in the following. First, according to rule 3 in Table 5, a company must try hard to make the turnover rate of fixed assets higher or equal to 6.48%, to be a good company for the investors. Second, according to rules 5 and 6 in Table 5, a net equity

growth rate larger than 8.52% and operating profit rate larger than 8.52% are important for a company to achieve. Third, rules 7 and 8 in Table 5—that directors, supervisors, and foreign shareholding ratio be larger than 45% and pledge rate of shares held by directors and supervisors be under 33%—should be kept in mind for a company's directors and supervisors.

Table 5. Summarizing decision tree rules and association rules as stock investment criteria.

No	Condition	Result	Remark
1	DS&F-Holding < 45 and year ≤ 14	Bad	Decision tree rules
2	DS&F-Holding < 45 and year > 14 and OI ≤ 6.48 and FA-turnover ≤ 6.48	Bad	
3	DS&F-Holding < 45 and year > 14 and OI ≤ 6.48 and FA-turnover > 6.48	Good ≥ ave *	
4	DS&F-Holding < 45 and year > 14 and OI > 6.48 and OP ≤ 16.52 and OE-growing ≤ 7.52	Bad	
5	DS&F-Holding < 45 and year > 14 and OI > 6.48 and OP ≤ 16.52 and OE-growing > 7.52	Good ≥ ave	
6	DS&F-Holding < 45 and year > 14 and OI > 6.48 and OP > 16.52	Good ≥ ave	
7	DS&F-Holding ≥ 45 and DS-Pledge ≤ 33	Good ≥ ave	
8	DS&F-Holding ≥ 45 and DS-Pledge > 33	bad	
9	ROE ≥ ave and OI ≥ ave	ROA ≥ ave	
10	ROE ≥ ave and ROA ≥ ave	OI ≥ ave	
11	ROE ≥ ave	ROA ≥ ave and OI ≥ ave	
12	ROA ≥ ave and BIG4 = Y **	OI ≥ ave	Association rules
13	ROA ≥ ave and OI ≥ ave	ROE ≥ ave	
14	ROA ≥ ave	ROE ≥ ave and OI ≥ ave	

Source: Own elaboration. * ave is the abbreviation of average. ** Y means yes.

Rules 9 and 14 in Table 5 are from association rules. These financial indicators, such as return on equity (ROE), return on assets (ROA), and operating profit rate (OI), influence each other. A company needs to make great efforts to enhance these financial indicators over the average of industry. Since decision tree analysis is less able to see the degree of mutual influence between variables, this research utilized association rules analysis to make up for this shortcoming, so that investors can better understand the changes. From the association rules table, we can find that return on equity, return on asset, and operating profit rate will affect each other; moreover, when a company is signed by one of the four large accounting firms, the operating profit rate will be greater than or equal to the industry average.

If one wants to understand the profitability of a company, then the net profit ratio is the important indicator, because the net profit ratio is used to measure a company's profitability and its cost control ability. If one wants to see whether a company has operating efficiency, refer to the company's return on assets (ROA), but if one wants to compare across industries, one needs to observe the company's return on equity (ROE). Generally, if the annual ROE exceeds 20%, or the quarterly ROE exceeds 5%, this company is considered a good company by investors. In order to make the investment risk of investors lower and the probability of obtaining returns higher, this study suggests that, when investors use the decision table to assist in stock selection decisions, they must also consider measuring the association rules, so that the investment made stock investment decisions are more secure and profitable.

In Table 5, this study summarized decision tree rules and association rules as stock investment criteria for stock investors' investment decisions. According to rules 1, 2, 4,

and 8, stock investors are able to avoid bad conditions when making investment decisions. On the other hand, rules 3, 5, 6, and 7 help stock investors select good investment targets. Data mining techniques are deployed to use past and present data to build the predict model. The investment performance must be above average when the model built by this study is used to select stocks.

It is well-known that stocks have a survival bias problem. How much does the survival bias problem impact the results of this study? Survivorship bias or survivor bias is the tendency to view the performance of existing stocks in the market as a representative comprehensive sample without regarding those that have gone bust. Survivorship bias can result in the overestimation of historical performance and general attributes of a fund or market index. Generally, investors consider a company's basic aspects, technical aspects, and news aspects for stock investment. Basic aspects are the primary aspects. This study is based on the basic aspects for stock investment. When investors have learned the correct concept of survivor bias, there is little impact on the results of this study.

5. Conclusions

This study attempts to construct an investment decision model with the decision tree analysis. Since the decision tree analysis cannot see the degree of mutual influence among variables, to compensate for this shortcoming, this study uses association rules to reveal the important rules not in decision trees, to assist the decision tree results. In order to avoid the influence of different types of industries, this research only uses the sports and leisure industry as the research object. Through empirical analysis, the research conclusions are summarized, and recommendations are provided for subsequent research references. This study makes three critical contributions to investors: (1) it proposes a systematical model of exploring related data through the decision tree algorithm and the Apriori algorithm to reveal the implicit investment knowledge. (2) An effective investment decision model is established and expected to provide a reference basis during stock-picking decisions. (3) The investment decision model is enhanced with implicit rules found among variables using association rules.

From the findings, this study has established a multiple-criteria decision-making model based on corporate governance variables and financial-related variables for stock investment decisions. This study integrated the decision tree algorithm with the Apriori algorithm in data mining techniques to find the predict model. In future research, other data mining techniques, such as the Bayesian network or the deep learning algorithm, may also be considered.

Author Contributions: For this research article, we had a team with several authors to conduct the research. Individual contributions are described as follows: conceptualization, K.-C.C., M.-J.H. and C.-K.F.; methodology, K.-C.C., M.-J.H., C.-K.F. and K.-H.W.; software, M.-J.H. and C.-K.F.; validation, K.-H.W., H.-M.W. and L.-H.L.; formal analysis, K.-H.W. and L.-H.L.; investigation, K.-H.W. and H.-M.W.; resources, K.-H.W. and L.-H.L.; data curation, K.-C.C., K.-H.W. and H.-M.W.; writing—original draft preparation, M.-J.H., C.-K.F. and K.-H.W.; writing—review and editing, K.-C.C. and H.-M.W.; project administration, K.-C.C., M.-J.H. and C.-K.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study uses Taiwan's listed sports and leisure industry as the research objects. The data of this study were taken from the 10-year data of listed companies in the Taiwan Economic Journal (TEJ). The data are collected from 2005 to the end of 2014 and excluded those with incomplete research variables that cannot be calculated.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: This paper does not contain any studies with human participants or animals performed by any of the authors.

References

1. Kane, A.; Marus, A.J.; McDonald, R.L. Debt Policy and the Rate of Return Premium to Leverage. *J. Financ. Quant. Anal.* **1985**, *20*, 479–499. [\[CrossRef\]](#)
2. Demsetz, H.; Villalonga, B. Ownership structure and corporate Performance. *J. Corp. Financ.* **2001**, *7*, 209–233. [\[CrossRef\]](#)
3. Myers, S.C.; Majluf, N.S. Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have. *J. Financ. Econ.* **1984**, *13*, 187–221. [\[CrossRef\]](#)
4. Lang, L.; Ofek, E.; Stulz, R.M. Leverage, Investment, and Firm Growth. *J. Financ. Econ.* **1996**, *40*, 3–29. [\[CrossRef\]](#)
5. Martikainen, T. Stock Returns and Classification Pattern of Firm-Specific Financial Variable: Empirical Evidence with Finnish Data. *J. Bus. Financ. Account.* **1993**, *20*, 537–558. [\[CrossRef\]](#)
6. Huang, M.J.; Sung, H.S.; Hsieh, T.J.; Wu, M.C.; Chung, S.H. Applying data-mining techniques for discovering association rules. *Soft Comput.* **2019**, *24*, 8069–8075. [\[CrossRef\]](#)
7. Bose, I.; Mahapatra, R. Business data mining-A machine learning perspective. *Inform. Manag.* **2001**, *39*, 211–225. [\[CrossRef\]](#)
8. Huang, M.J.; Chen, M.Y.; Lee, S.C. Integrating Data Mining with Case-based Reasoning for Chronic Diseases Prognosis and Diagnosis. *Expert Syst. Appl.* **2007**, *32*, 856–867. [\[CrossRef\]](#)
9. Kopun, D. A review of the research on data mining techniques in the detection of fraud in financial statements. *J. Account Manag.* **2018**, *8*, 1–18.
10. Fayyad, U.M.; Piatstsky-Shapiro, G. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54.
11. Olson, D.; Shi, Y. *Introduction to Business Data Mining*; McGraw-Hill/Irwin Englewood Cliffs: New York, NY, USA, 2007.
12. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Exp. Syst. Appl.* **2007**, *32*, 995–1003. [\[CrossRef\]](#)
13. Ladas, A.; Ferguson, E.; Aickelin, U.; Garibaldi, J. A data mining framework to model consumer indebtedness with psychological factors. In Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014.
14. Yanga, X.; Lina, X.; Lin, X. Application of Apriori and FP-growth algorithms in soft examination data analysis. *J. Intell. Fuzzy Syst.* **2019**, *37*, 425–432. [\[CrossRef\]](#)
15. Witten, F.; Hall, P. *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann, Inc.: San Francisco, CA, USA, 2017.
16. Questier, F.; Put, R.; Coomans, D.; Walczak, B.; Vander, H.Y. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemomater. Intell. Lab.* **2005**, *76*, 45–54. [\[CrossRef\]](#)
17. Cherfi, A.; Nouira, K.; Ferchich, A. Very Fast C4.5 Decision Tree Algorithm. *Appl. Artif. Intell.* **2018**, *32*, 119–137. [\[CrossRef\]](#)
18. Singh, N.; Singh, P. A novel Bagged Naïve Bayes-Decision Tree approach for multi-class classification problems. *J. Intell. Fuzzy Syst.* **2019**, *36*, 2261–2271. [\[CrossRef\]](#)
19. Chang, N.; Olivia, R.; Sheng, O. Decision-Tree-Based Knowledge Discovery: Single- vs. Multi-Decision-Tree Induction. *Inform. J. Comput.* **2008**, *20*, 46–54. [\[CrossRef\]](#)
20. Alos, A.; Dahrouj, Z. Decision tree matrix algorithm for detecting contextual faults in unmanned aerial vehicles. *J. Intell. Fuzzy Syst.* **2020**, *38*, 4929–4939. [\[CrossRef\]](#)
21. Agrawal, R.; Imielinski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 26–28 May 1993; Volume 22, pp. 207–216.
22. Ordóñez, C.; Zhao, K. Evaluating association rules and decision trees to predict multiple target attributes. *Intell. Data Anal.* **2011**, *15*, 173–192. [\[CrossRef\]](#)
23. Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*, 1st ed.; Morgan Kaufmann: San Francisco, CA, USA, 2001.
24. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning*, 1st ed.; Springer: New York, NY, USA, 2001.
25. Li, Z.; Li, L.; Yan, K.; Zhang, C. Automatic image annotation using fuzzy association rules and decision tree. *Multi Syst.* **2017**, *23*, 679–690. [\[CrossRef\]](#)
26. Wang, Q.; Hu, X. Stock Selection in Investment Management of Commercial Stock Market: Prediction by Data Mining. *J. Comput.* **2019**, *30*, 260–268.
27. Chen, C.; Hsieh, C. Actionable Stock Portfolio Mining by Using Genetic Algorithms. *J. Inf. Sci. Eng.* **2016**, *32*, 1657–1678.
28. Ng, K.; Khor, K. StockProF: A stock profiling framework using data mining approaches. *Inf. Syst. E-Bus. Manag.* **2017**, *15*, 139–158. [\[CrossRef\]](#)
29. Drobetz, W.; Wanzenried, G. What Determines the Speed of Adjustment to the Target Capital Structure? *Appl. Financ. Econ.* **2006**, *16*, 941–958. [\[CrossRef\]](#)
30. DeAngelo, L.E. Auditor size and audit quality. *J. Account Econ.* **1981**, *3*, 183–199. [\[CrossRef\]](#)
31. Attar Software Limited *XpertRuler Miner: Knowledge from Data*; Attar Software Limited Inc.: Manchester, UK, 2002.