IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2020, accepted August 11, 2020, date of publication August 25, 2020, date of current version September 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019332

A Boosting Regression-Based Method to Evaluate the Vital Essence in Semiconductor Industry Performance

PING-YU HSU[1], I-WEN YEH[1], CHING-HSUN TSENG[2], AND SHIN-JYE LEE[3]
[1]Department of Business Administration, National Central University, Chungli 32001, Taiwan
[2]Department of Computer Science, The University of Manchester, Manchester M13 9PL, U.K.
[3]Institute of Management of Technology, National Chiao Tung University, Hinschu 30010, Taiwan

Corresponding author: Shin-Jye Lee (camhero@gmail.com)

This work was supported in part by the Ministry of Science and Technology Research Grant in Taiwan under Grant MOST 109-2221-E-009-098.

ABSTRACT In accordance with the statistical analysis, the industrial performance is usually related to research and development (R&D) intensity, and this factor indeed plausibly brings the biggest profit with patents and supporting products to the development of semiconductor industry. How to evaluate the completive performance of modern industries is an increasing issue, especially for the semiconductor industries in these decades. However, almost every traditional statistical model is deterred by the hypothesis of population and independent correlation among each feature, and this makes the result of typical regression model potentially lose reliability. To avoid this weakness, this article therefore applies a gradient boosting based method - XGBoost to evaluate the feature importance of semiconductor industries. In the simulation experiments, different findings revel certain information, apart from R&D intensity, actually sway the gross net value in the annual financial announcement of semiconductor industries. Moreover, this article proposes another concept to evaluate the essential factor contributing the development of semiconductor industries. Instead of only focusing on the effect of R&D intensity, this article also predicts the future growth rate (GR) of net value by applying the greedy search of XGBoost Regression.

INDEX TERMS XGBoost, boosting regression, semiconductor industry.

I. INOTRODUCTION

Research and development (R&D) management is one of important issues in the research field of human resource management (HRM), and how to make proper investment in the expense of R&D management has become an increasing issue. The most works reveal a common conclusion in the literature that R&D has a positive effect on the performance of industries, so the term ''R&D Intensity'' is primarily applied in the analysis of relevant works and defined as the key factor between ''R&D Expenditure'' and ''Sales Revenue''. In order to obtain more competitive advantages, the knowledge-intensive based firms usually spend substantial amounts of cost on R&D activities. Therefore, R&D Intensity is one among the key indices that can be applied to predict the future performance of industries, and the suitable deployment of R&D expenditures is also one of the essential issues to the industries. However, being R&D Intensive cannot completely guarantee the success of industries. For example, Nokia used to be the biggest brand in the mobile phone market from 1990 to 2008, as Nokia always invested a huge amount of cost in R&D projects. Its annual report stated that Nokia invested about €5.8 billion in R&D in 2010, which is more than 4 times of that to its competitor Apple. However, it lost its market share abruptly from 39% in 2008 to 25% in 2011, and the even worse trend is that Nokia was totally defeated by Apple and Samsung in the high-end smart phone market. Also, there are many similar big business examples of failure in different markets, such as Kodak and Xerox. Why is that?

In addition to the absence of awareness, the possible causes might ignore certain potential factors. As mentioned before, R&D Intensity has been recognized as a critical effect toward the firm performance but some arguments toward the cash equivalent, free cash flow, and so on. In the traditional methods, most research works applied the approach of linear

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

regression to fit the overall trend, and these works also tried to explain the pattern even though some potential flows are hidden in a linear function [1]–[3]. In addition, the approach of logistic regression possesses the merit of nonlinear ability to fit a variety of tasks, but those methods usually fail to fulfill real-world tasks. As a result of the sparsity distribution of the parameter in logistic regression, it hardly finds the high-impact feature based on this situation. Also, the overfitting problem is not rare in processing linear regression method. In order to prevent the overfitting problem, the typical way usually applies the fine-tuned L2 norms to find the best model fitting the objective, but it eventually cannot fit the similarity situation in the complex correlation. Moreover, this issue is discussed in the Part C of Section II in detail. Further, as for the controversial disadvantage of statistical technique, the collinearity and bias problems easily weaken the objectiveness of result.

Due to the highly promising benefit brought by artificial intelligence (AI), applying a variety of machine learning based methods to either support decision making or discover more specific pattern is currently a prevalent trend in these years. Instead of conventional statistical methods, this article applies a novel regression model based on the mechanism of machine learning, XGBoost, to effectively evaluate the high-impact factors of semiconductor industries. Different from the normal regression model, XGBoost is one of boosting models, which is an ensemble of multiple weak regression or classification models. Through minimizing the residual between real values and predicted values, each result will be affected by the former result. What also makes XGBoost shine is its robustness developed by the blessing of not only L2 norm but also the leaves restriction, which is represented as T in the algorithm (3) in the Part B of Section III. By introducing T in loss function, the overfitting problem can be relieved so it can potentially fit the objective and prevent the variance in the prediction. Besides, the performance of constructing a forest with a tree is also an advantage of XGBoost, so it can be recognized that a bewildering array of competitions are overcome by XGBoost. Due to the inherent mechanism of using gain to split the data into one node, this article applies XGBoost regression to fit the real trend of collecting financial announcement over two decades, and furthermore the mechanism of gain is clearly explained in the Part C of Section III. Extracting the gain in each node and summarizing it to compare the score of each feature. Therefore, the vital index affecting the firm performance can be recognized, as XGBoost regression fits the objective feature with a lot of models and the divergence of predicted value and real value is hence approximated as close as possible. The work ''Tree Boosting with XGBoost'' provides a detailed explanation about why XGBoost wins ''Every'' competition and also presents that XGBoost applies the boosting tree to automatically select feature and capture high-order interactions without breaking down [4], so XGBoost can be considered a robust method even facing the problem of curse of dimensionality.

In this article, this section describes the problem context as well as reviews the current barriers in traditionally examining the feature importance, and also the merit of XGBoost regression in data analysis is included. In Section II, the related works are reviewed, including R&D Intensity examination and machine learning works. The loss function and measurement of XGBoost are shown in Section III. In Section IV, it conducts XGBoost regression processing with the collected data and finds a different picture in assessing the firm performance. Finally, it reveals some insights and prediction in Section V.

II. LITERATURE REVIEW

A. R&D INTENSITY AND FIRMS PERFORMANCE

Firms invest in R&D to develop new technologies and products to create competitive advantage, so R&D is critical for a firm to survive and sustain its competitive advantage in the dynamic environments. Prior research has shown that R&D expenditures are positively related to firm performance [4]–[6]. Moreover, firms investing more in R&D tend to perform better than firms investing less in R&D in competitive industries [7], although the positive effects occur often with lagged period of time [8] where some threshold effects exist [9]. R&D investment is then considered as a critical driving force of technological change and economic growth in modern countries [10]. According to the Schumpeterian growth model, long-run growth resulting from innovation and then innovation resulting from R&D investments, new innovations will finally replace old technologies [10]. That is, without R&D investment, firms will finally lose in the fierce technology competition. Additionally, R&D Intensity can be considered as a proxy of innovation capabilities, because it is examined to be a positive predictor of firm performance in the semiconductor industry. Hence, R&D-Intensive based firms are probably less sensitive to external shocks, since their products are not easily substituted with other cheaper alternatives [11]. However, this is not necessary the case always as in the Nokia example, as Nokia was unaware of failing to the market trend. It implies that R&D-Intensive based firms will perform better in a dynamic environment. The ''returns on invest (ROI)'' of R&D activity can be viewed as the potential performance in the future due to the delayed reaction by markets. Meanwhile, the theory of adaptive capability emphasizes that the adaptive capability allows organizations to identify and capitalize on the opportunities of emerging markets in a relatively quick and flexible sense [12]–[15], so organizations can then reconfigure resources and coordinate processes promptly to produce more innovative products [16], [17].

Further, the partial adjustment may offer new insights into the process of Schumpeterian competition in a dynamic environment. According to the partial adjustment theory, firms tend to increase their R&D investment to reinforce the strength of R&D Intensity gradually. The speed of adjustment varies widely across firms, and those with higher speed of

adjustment usually perform better in the technology competition. In dynamic environments, the speed of adjustment plays a critical role in enhancing the competitive power of firm performance, and the speed of adjustment also can also be considered as a measure of the adaptive capability of firms. In accordance with the information-based theory [18], firms usually tend to imitate each other, especially to imitate those are considered as possessing superior information. As a result, R&D Intensity of the firms competing in the same industry tends to be similar [10]. Meanwhile, the increasing competition will then encourage these firms to innovate aggressively so that they can take advantage of competition in the markets. Also, an effective innovation can alleviate the imitation of followers and brings profits to firms [18]. Firms are even more willing to increasingly invest on R&D activities in order to differentiate themselves from other competitors via innovation under the intensive competitions, especially for the firms that are in the leading position, i.e., frontier firms in the market [10], [19]. However, R&D expense usually costs much and may have negative impacts on the financial performance of firms in the short-term investment [20]. Also, the long-term investment that is difficult to return in less than one year may even deteriorate the financial status of firms. In addition, ''free cash flow (FCF)'' can then be an index to reflect the cash generating capability of previous long-term investments and it also can be considered as a monitoring criterion for firms to evaluate the corresponding strategy of long-term R&D investments. Potentially, FCF can be financially manipulated, and ambiguous results are usually found in the literature. For example, Brush *et al.* [15] argued that FCF is not profitable, while Kim and Bettis [16] claimed that ''cash is surprisingly valuable as a strategic asset''. However, a firm that cannot generate free cash in a long run may perform worse in the future. Moreover, Deb *et al.* [17] stated that cash is especially beneficial in ''highly competitive, research-intensive, or growth-focused industries'', such as semiconductor industry. Moreover, investment opportunities and revenue sources for firms in the semiconductor industry are strongly affected by the economic cycle. It is hard for them to keep their R&D Intensity at fixed targets or optimal levels, because R&D investment usually cost highly and it thus negatively impacts on the indicators of accounting performance in the short run, such as ''Return on Assets''.

However, an increase in R&D investment does not mean that there will be an increase in organizational risks [21]–[23]. Some research works revealed that an organization may be conservative in making changes as its performance is satisfactory [24]. Such a conservative strategy usually leads to the loss of markets in the coming future because of less innovation due to less investment on R&D. Actually, it also shows that firms unsatisfied with their performance will increase the proper investment on R&D [25]. Moreover, if a firm operates smoothly and FCF is accumulated to be at a higher level, e.g., 50% higher than the typical performance in the past, the enterpriser would have the potential to invest more for the future growth. Since FCF can be an index to reflect the operation performance of firms, it might be suitable to be incorporated with R&D Intensity as an index to evaluate whether being conservative or aggressive on R&D investments.

B. MACHINE LEARNING APPLICATION IN DECISION MAKING

According to T. Mitchell (1997) [26], ''Machine learning is the study of computer algorithms that improves automatically through experience.'', and it therefore implies that the hidden information and the potential pattern of the targeted entity or behavior can be discovered by iteratively training the mathematical models through past experience. As for the learning type of machine learning algorithms, it primarily comprises supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, and the purpose of each learning aims to address different problems based on various *a priori* knowledge. For example, the mechanism of famous AlphaGo and AlphaGo Zero is based on reinforcement learning algorithms, and it presents an excellent strategy of real-time decision making as the *a priori* knowledge is unknown. Meanwhile, effectively applying the cutting-edge technique of machine learning to enhance the industrial performance has become an increasing issue in the field of industrial engineering and management in these years, With the development of Industry 4.0, the semi-supervised learning algorithms have been usually applied to empower the performance of smart factory, such as smart manufacturing. Especially for the semiconductor industry, simplifying a manufacturing process or predicting an advanced recipe of more high-level chip by the refined semi-supervised learning algorithms are already being considered for production, and it is another type of decision making in the semiconductor industry. As for the machine learning models, it primarily comprises Fuzzy Systems, Neural Network, Decision Tree, Support Vector Machine, and Bayesian Network, and each inference model is processed based on its unique mechanism respectively. Basically, the so-called ''Deep Learning'' is an advanced development of neural network, and it is developed on the basis of more than two hidden layers of neural network. Moreover, the work ''Artificial Intelligence for Humans'' stated ''Problems that require more than two hidden layers were rare prior to deep learning. Two or fewer layers will often suffice with simple data sets. However, with complex datasets involving time-series or computer vision, additional layers can be helpful. The following table summarizes the capabilities of several common layer architectures.'' [25]. At present it has been widely applied to a variety of fields, including government policy [27], [28], medical diagnosis [29]–[32], business strategy [33], [34], financial trade [35], [36] [37], industrial engineering [38]–[40], and so on, and financial technology (FinTech) [41] is also one of popular applications. In addition, the series of decision tree models usually perform well in processing mid-size data, such as classification and regression trees (CART) [42], ID3 [43], and C4.5 [44]. As for the forest-type decision tree models, they are ensemble learning methods combining

weak learners to a powerful learner, which can also be customized in different tasks. In accordance with the sampling methods, the mainstream of ensemble learning can be classified into two branches: Bagging [45] and Boosting [46]. The mechanism of bagging is to randomly sample and then vote on the training samples, and the typical representative are random forest [47], rotation forest [48] and so on. Also, the mechanism of Boosting is to have weights in the sampling, which stands for gradient boost decision tree (GBDT) [49], XGBoost [50], etc. Thus, there are huge amounts of applications conducted by machine learning in a variety of decision-making purposes [51]–[53], and they are either promising or reliable nowadays.

C. POTENTIAL DISADVANTAGE OF CONVENTIONAL REGRESSION MODELS

So far as a solid method is concerned, it both comprises the theoretic evidence as well as the practical evidence, and the practical evidence is convinced based on the theoretical evidence. The theoretical evidence means the method can be logically reasoned by mathematical inference or reasonably theoretical explanation. The practical evidence usually means the simulation experiment to evaluate the reliability of the proposed theory, and the practical evidence is strictly confirmed based on the theoretical evidence. The reason is, as the practical evidence is obtained without the theoretical evidence it is usually recognized as the coincidence. As for the theoretical evidence, the controversial disadvantage of the conventional statistical method is the collinearity and bias problems, and it can be proved by its mathematical mechanism. Also, the mechanism of machine learning algorithms usually applies random process, as it is black-box modelling and the *a priori* knowledge is completely unavailable. Compared with the mechanism of statistic and that of machine learning, the inference result processed by machine learning based methods has more objective than statistic as the casual relationship is not taken into consideration.

As the mathematical pros and cons of the conventional regression models (e.g., stepwise regression, lasso regression, multiple regression, etc.) can be reasoned by their mechanisms, the theoretical evidence has already been proved. Afterwards, the simulation experiment aims to prove the reliability of the theoretical evidence. Through processing the conventional regression models, the simulation experiment usually suffers with the multicollinearity problem. To address this problem, the simulation experiment applied L1, L2 norm, and even Lasso regression model to find a possible solution. Nonetheless, these approaches brought another problem to the simulation experiment, and that is famous ''curse of dimensionality''. Hence, these works naturally failed to examine the feature importance in such a high-dimension dataset, because a lot of coefficients in the regression are given in zero resulting in a very worse accuracy, which can be shown in the following Figure1, 2, and 3 as setting the L1 value to 0.1, 1, and 10 respectively in the Lasso edge regression.

```
Train_Acc: 0.014492753623188406
Test_Acc: 0.02857142857142857
Cofficient: [[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

FIGURE 1. L1: 0.1 in Lasso Edge Regression.

```
Train_Acc: 0.2608695652173913
Test_Acc: 0.0
Cofficient: [[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

FIGURE 2. L1: 1 in Lasso Edge Regression.

```
Train_Acc: 1.0
Test_Acc: 0.0
Cofficient: [[ 0.          0.          0.         ... 0.          0.
   0.        ]
 [ 0.          0.          0.         ... 0.          0.
   0.        ]
 [ 0.          0.          0.         ... 0.          0.
   0.        ]
 ...
 [ 0.          0.          0.         ... 0.          0.
   0.        ]
 [-0.82458557  0.          0.         ... 0.          0.
   0.        ]
 [ 0.          0.          0.         ... 0.          0.
   0.        ]]
```

FIGURE 3. L1: 10 in Lasso Edge Regression.

Through processing the grid searching to fine-tune the effect of L1 and L2, these approaches potentially ignore the original objective of the proposed work - providing an either efficient or robust method. In particular, Random Forest is one of machine learning algorithms, and it has also been applied to a variety of fields. It is a kind of ensemble learning, and its mechanism is operated on the basis of bagging strategy. Although a stable result can be potentially obtained, there exists an uncertainty that weak learners are always being chosen due to the negative property of bagging algorithm.

III. BOOSTING REGRESSION BASED METHOD

As for the semiconductor industry, what it faces is a fluctuating pattern, and it is a norm that the index and the object are not uncontained this kind of unstable symptoms. Meanwhile, the gradient regression methods are effective to process continuous data, and it converges through the residual iterations. Therefore, the outcome generally can be regarded as an either robust or reliable prediction. Among a series of gradient boost methods, XGBoost outshines in several AI applications [54]–[56], and the further discussion about the robustness of XGBoost can be realized in the work ''Tree Boosting with XGBoost'' [57]. Therefore, in this section, it will reveal the mechanism of XGBoost and explain why it is suitable to be applied in this work.

A. GRADIENT BOOSTING METHOD

As processing continuous data in the financial announcements, the mechanism of approximating the residual between observed values and predicted values is an appropriate method. The typical loss function of aforementioned residual can be shown as:

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

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and i is the index of the data.

However, in the gradient boosting, the loss function can be revealed as:

$$l\left(y_i, \hat{y}_i\right) = \frac{1}{2}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and i is the index of the data.

From the equation of (1) and (2), the effect of derivation, $\frac{1}{2}$, cannot sway different, but it does bring a good advantage to reduce the complexity of algorithm by averaging the residual summary in the early step and final step in XGBoost, which will be described in the following part.

B. XGBOOST

As a method of boosting algorithms, XGBoost processes in a similar manner. Nonetheless, instead of constructing a stump each time and add it up the residual to predict, XGBoost introduces a slightly larger tree with the leave restriction and normalization to avoid high variation and overfitting. The overall step in XGBoost can be described as follows:

Input:

Data $(x_i, y_i)\}_{i=1}^{n}$, and a differentiable Loss Function, as the algorithm (1): $l\left(y_i, \hat{y}_i = F(x)\right) = \frac{1}{2}(y_i - \hat{y}_i)^2$

Step 1:

Initialize model with a constant value: $F_0(x) = argmin \sum_{i=1}^{n} L(y_i, r)$

Step 2:

for m = 1 to M:

(1) Calculate $r_{im} = -[\frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}]_{F(x)=F_{m-1}(x)}$ for $i = 1 \ldots n$

(2) Fit a regression tree to the r_{im} values and build terminal regions R_{jm}, for $j = 1 \ldots J_m$

(3) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = argmin \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

(4) $F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} r_{jm} I(x \in R_{jm})$

Step 3:

Output $F_M(x)$

---

where i indicates the data index, n is the total data number, γ actually refers to the average of the observed data, m means the m_{th} tree, M is the total number of tree, j is the j_{th} residual in the m_{th} tree, and v is the learning rate or the distance of moving step toward the gradient of residual.

The entire process of XGBoost aims to fit the observed values by optimizing the loss function constantly. Due to the merit of restricting gradient with learning rate, the output can prevent overfitting but still make sure a low bias.

$$l^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2 \right] + \gamma T \tag{3}$$

where l is the number of leaves, $G_j = \sum_{i=1}^{n} g_i$, $w_j = f_t(x_i)$, $H_j = \sum_{i \in I_j} h_i$, $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$, λ is the Lagrange multiplier to penalize the L2 norm in order to prevent overfitting, w_i represents the score on the j-th leaf, γ means the number of the leaves, and T is the number of the nodes.

C. GAIN

Gain is applied extensively, and it can find the optimal feature splitting data as a result of using greedy search in the mechanism. Generally, it can be considered as the benefit of applying the prediction to fit or separate data. This method can be processed in classification and regression tree, CART. Further, the algorithm of gain can be shown in below:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_L^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_L + \lambda} \right] - \gamma \tag{4}$$

where $G_L^2 = (\sum_{i \in I_L} g_i)^2$, $G_R^2 = (\sum_{i \in I_R} g_i)^2$, $H_L = \sum_{i \in I_L} h_i$, and $H_R = \sum_{i \in I_R} h_i$. Moreover, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are first and second order gradient statistics on the loss function, respectively.

Basically, the default value of γ is set with 0. If the gain is negative, the branch will be removed. Moreover, if the gain of the root with two leaves is negative, the root will be removed. It means the entire tree is dropped, and the output will take the original value as the prediction in this step. The aforementioned process is so-called ''pruned''.

IV. SIMULATION EXPERIMENT

A. EXPERIMENT DESIGN

In order to examine high-impact features in the semiconductor industry, the experimental data include over two decades financial announcements of famous semiconductor firms, 1993-2016, including UMC, ASE, SPIL, TSMC, MXIC, WINBOND, and SDI separately, and their capital range from 61.09 million US dollars (SDI) to 4.14 billion US dollars (TSMC), that the evaluated semiconductor firms comprise small-sized, medium-sized, and large-sized semiconductor firms. Also, the features stated in the financial announcements are FirmID, Name, Year, Cash_equiv, Total_asset, SH_equity, Revenue, Net_profit, NP_b4TI, NP_b4TID, RD_intensity, FCF_ratio, Depreciation, Amortization, Cash_gen, ROA_b4TID, ROA_b4I_aT, ROA_b4ID_aT, ROE, Gross_margin, Profit_rate, Net_margin_b4T, Net_margin, N_employees, Op_Expense_r, personnel_exp, CF_ratio, GR_revenue, GR_GrossPro, GR_profit, GR_profitb4T, GR_profitaT, GR_TotalAsset, and

GR_netvalue. Most importantly, the GR_netvalue is the objective feature to assess which feature is truly vital toward a good firm performance, especially for RD_intensity. In addition, the features with "GR (gross rate)" at front (prefix of GR), means that these features are associated with gross, and GR is the abbreviation of gross.

As for the data preprocessing, there are 168 data overall, and each firm keeps a record of 24 years data, from 1993 to 2016. Meanwhile, the column of NP_b4TI, NP_b4TID, and personnel_exp in ASE are Nan in 2018, so the complete data range is set from 1993 to 2017. Therefore, the range of training data is from 1993 to 2016, and the 2017 dataset is processed as the validation data. Finally, each feature will be normalized to follow the N (0, 1).

To obtain a more detailed finding, the simulation experiment is processed by two parts: (1) XGBoost gain is applied to examine the score of feature importance. (2) The data is divided by the period of four financial years in each firm. In the proposed model, a hundred of basic trees, same as the default number of cikit-learn library, are set in XGBoost, and the distribution of each feature is plotted in the Figure 4.

In accordance with the Figure 4, as the data is sorted by years, the distribution of each factor is clearly revealed. Also, it can be observed that the pattern of feature follows the objective trend, that GR_total_asset and GR_revenue have high impact on GR_netvalue. Aforementioned factors might fit the target well, but the explanatory ability will go to futile. This is so called multicollinearity. To obtain a more reliable result with the objectiveness, the problem of multicollinearity should be avoided as far as possible, so the proposed method, XGBoost can further examine the high-impact features without the highly related effect. Although multicollinearity means certain same-pattern factors will affect their similar indices, this simulation experiment still decides to remain this kind of factors, including Cash_equiv, Total_asset, SH_equity, Revenue, Net_profit, NP_b4TI, and NP_b4TID. The reason is that they might express certain important information to support the ability of fitting the target.

B. FEATURE IMPORTANCE

As described in the Section III, XGBoost can fit the object well with gradient, so the reliability of gain to measure feature importance is widely accepted. In addition, it also can be observed that the loss function with leave restriction and L2 norm can potentially prevent overfitting and high variance. Compared with other methods, such as Random Forest and Logistic Regression, the low bias indeed brings a clear explanation and good fitting capability to the prediction. As a result, each feature importance applies the greedy search to find the optimal splitting effect in the current situation, and the result can be illustrated in the following plot, as shown in Figure 5.

In Figure 5, a horizontal comparison is shown in the observed years, 1993-2017, and the percentage represents the average gain of feature in the overall gain at every tree node.

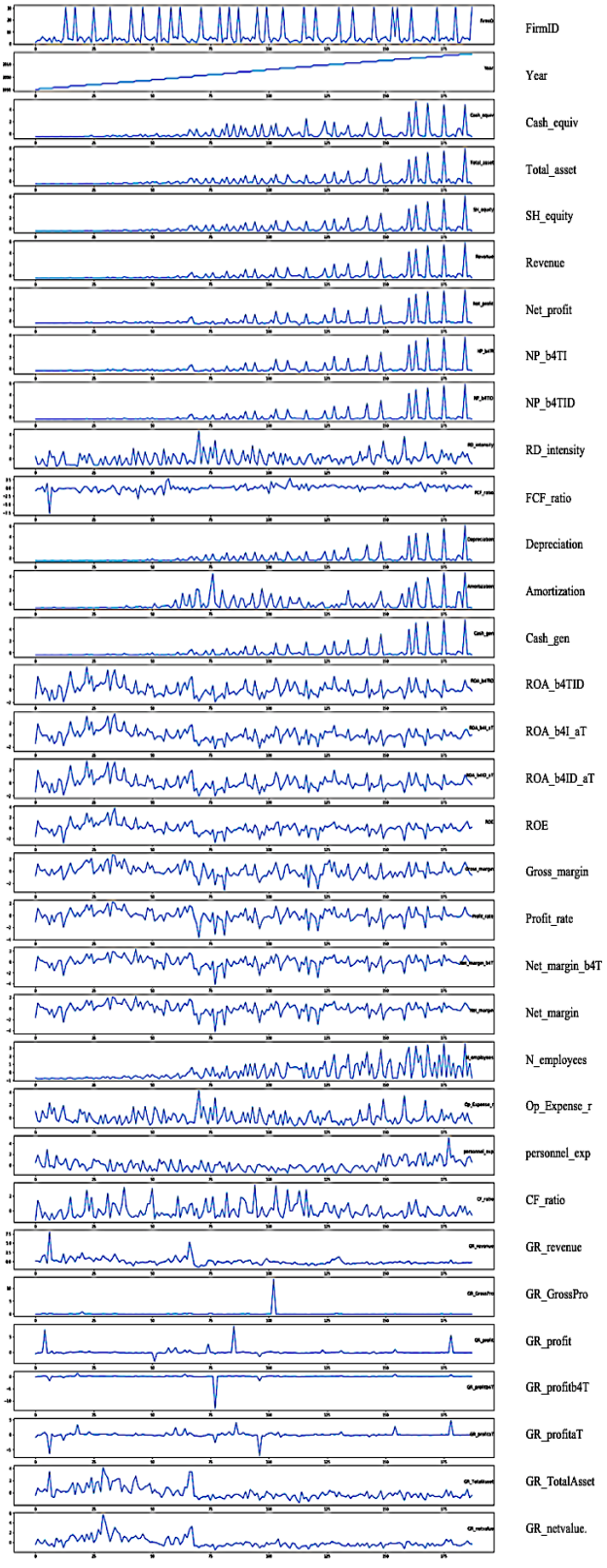

FIGURE 4. The distribution in each feature.

Through sorting the percentage, apart from certain highly related features with GR_netvalue, such as the feature with "GR" at front, R&D Intensity actually plays an essential role

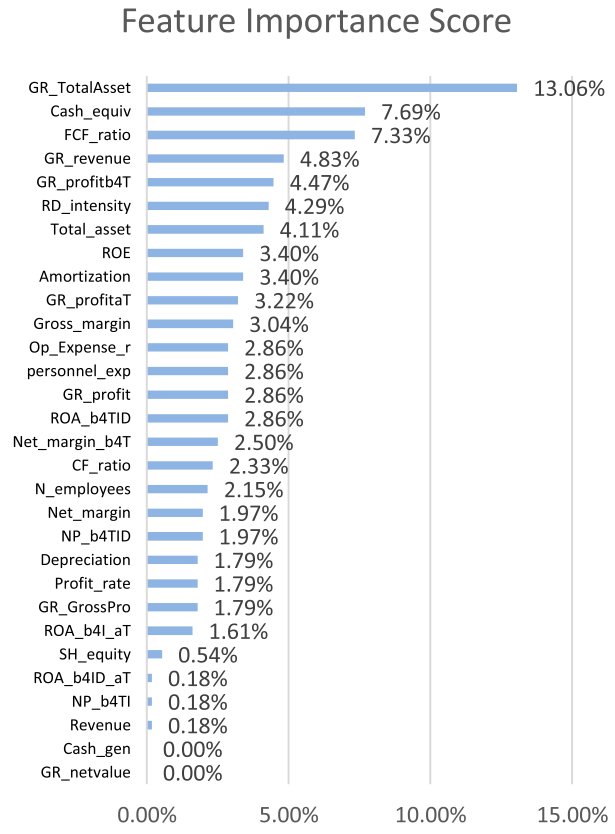

FIGURE 5. The overall feature importance in gain.

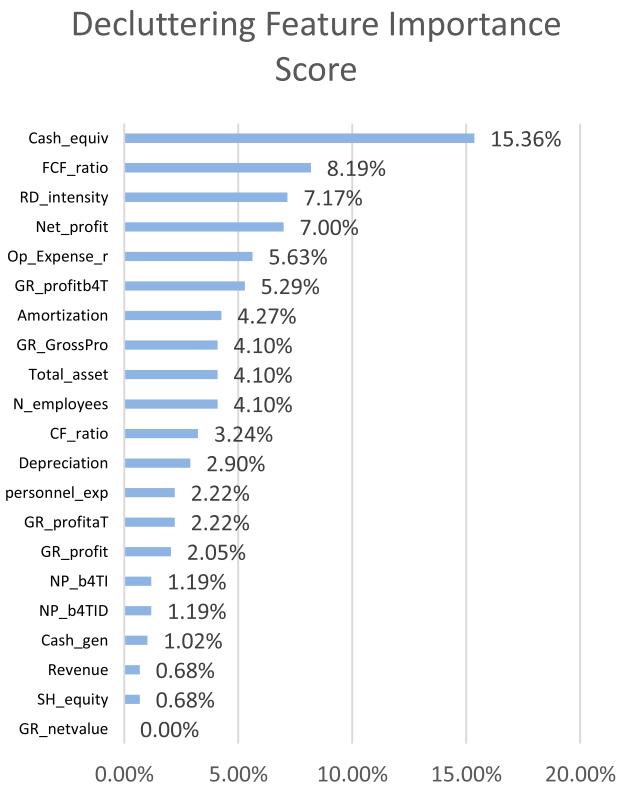

FIGURE 6. The feature importance score in gain after removing high related feature.

as using this index to fit the GR_netvalue. It therefore proves that the importance of R&D department investing.

Multicollinearity is also a general problem in regressions, and the feature unreliability usually happens as a result of the cross impact by high-related features. In this regard, removing high correlative features in regressions is an ideal scheme, and these features comprise GR_Total_Asset, GR_revenue, Net_margin, Net_margin_b4T, Profit_rate, Gross_margin, ROE, ROA_b4ID_aT, ROA_b4I_aT, and ROA_b4TID, as shown in Figure 7. As a result, the further score of feature importance after removing high related features are shown in the Figure 6, and it is obvious that the gain score of R&D Intensity steps to a higher rank, 3rd, which concretes the hypothesis even further. Besides, as shown in the Figure 6, the R&D Intensity is not the most important feature. Actually, the cash_equivalent, free_cash_flow, net_profit, and etc. are also essential factors as examining the firm performance. Moreover, the overall correlation score is shown in the Figure 7.

C. R&D INTENSITY TREND

During the two decades financial announcement, the firm performance has been fluctuated frequently. As described above, R&D Intensity indeed plays an essential role in the semiconductor industry, so the simulation experiment aims to evaluate further by observing the gain score of R&D Intensity and its rank over twenty years. As shown in the Figure 8, by the recommended divided period in four years based on the analysis of semiconductor industries of Morgan Stanley, the data is separated into five sections with four years in each part, and the high correlated features are already removed. In addition, as the stated data information in A of section IV, there are some Nan (Not a Number) in the data of 2018, so it is difficult to separate period in three years. Thus, only the data of 1993-2016 are taken into account in the simulation experiment. Through ranking the feature importance with gain in XGBoost regression, the trend is shown in the following Figure 9, Figure 10 and Table 1.

TABLE 1. The feature importance of RD_intensity in gain with rank.

Year	score	rank
1993-1996	4.3%	8
1997-2000	6.1%	6
2001-2004	4.6%	9
2005-2008	5.5%	7
2009-2012	7.6%	6
2013-2016	7.1%	6

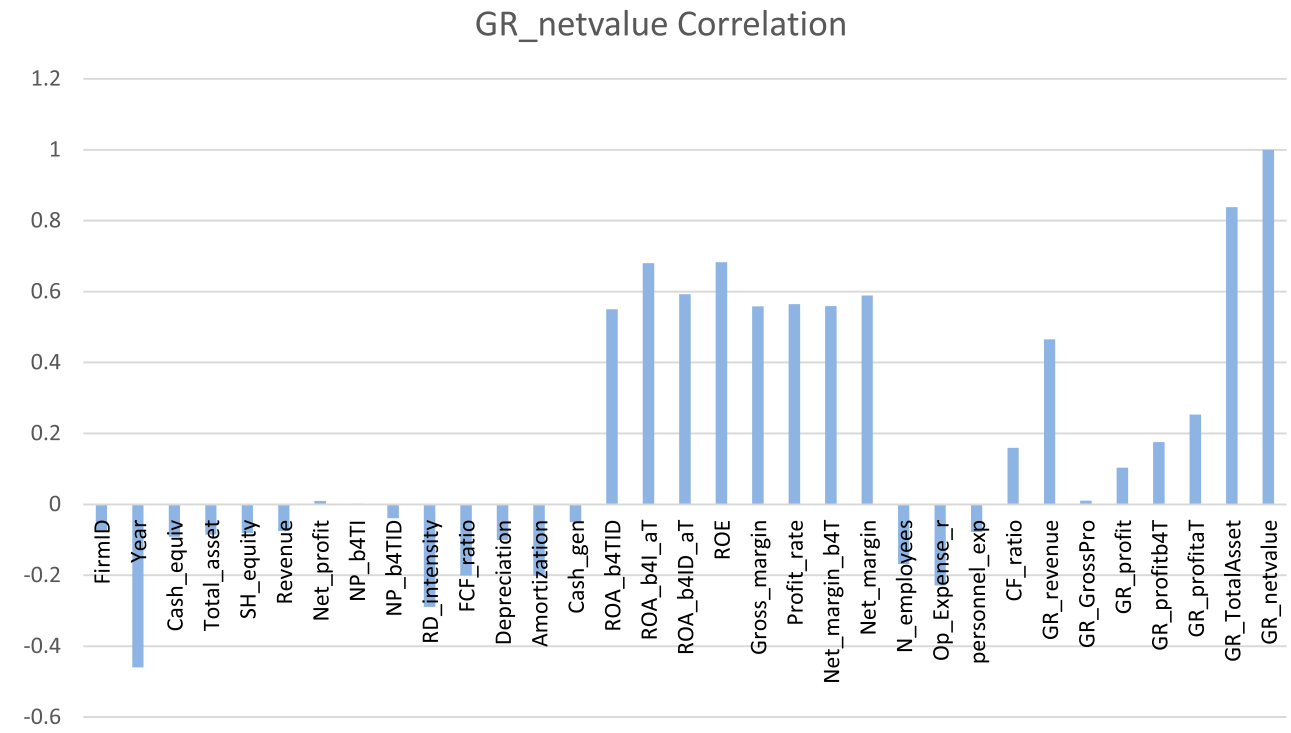

FIGURE 7. The overall each feature's correlation toward GR_netval.

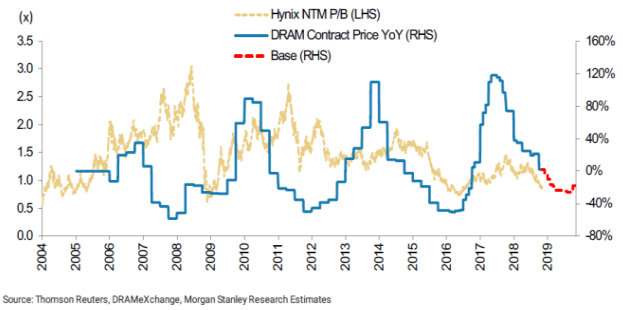

FIGURE 8. The life cycle of Semiconductor Companies (Left Axis: Return, Right Axis: Return Growth Rate).

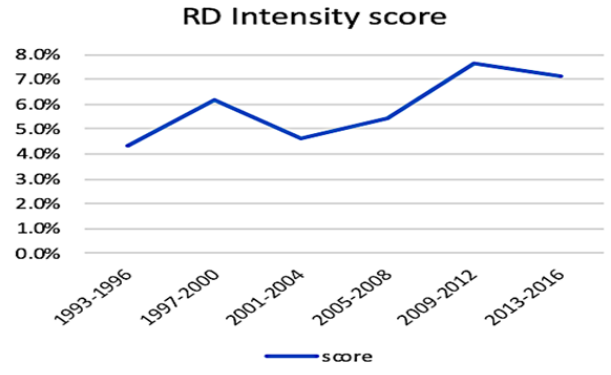

FIGURE 9. The feature importance of RD_intensity trend in gain.

In accordance with the above information, among all 19 features in the financial announcement, R&D Intensity occupies the top one-third positions in the list. Although

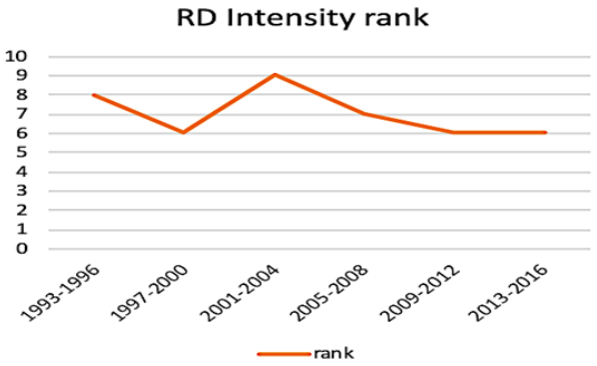

FIGURE 10. The feature importance of RD_intensity rank.

its rank declined in 2001-2004, afterwards the trend was mounting back. Eventually, the overall rank of R&D Intensity is staying around the sixth position, and it can be deduced that some features might not be always more important than R&D Intensity but R&D Intensity constantly stayed in a key position in a long period of years. Through the proposed method, the important factors in a fluctuating period of time can be observed with the objectiveness. Thus, the proposed method can effectively analyze the potential high-impact factors in the semiconductor industries, and it can positively provide a reliable decision making for the strategy design to the semiconductor industries.

V. CONCLUSION

To avoid the potential disadvantage brought by the conventional regression models, this article provides a machine

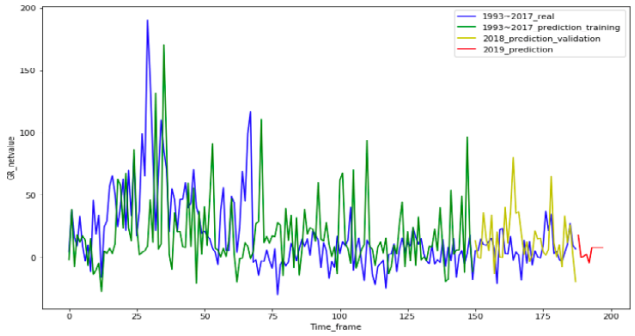

FIGURE 11. The 2019 GR_netvalue prediction.

learning based method to objectively evaluate the feature importance in the financial announcement of semiconductor industries. Generally, the financial report is usually analyzed by statistical methods to explain the current situation, but the statistical analysis is not always convinced as a result of the collinearity and bias problems. Instead of conventional statistical methods, this article applies a novel regression method based on the machine learning mechanism, XGBoost, to effectively evaluate the high-impact factors of semiconductor industries. In accordance with the simulation results obtained by the proposed method, it can be positively observed that the feature importance varies in the continuous type. Especially in fitting the GR_netvalue, R&D Intensity indeed plays a key role in the semiconductor industry, and it therefore implies that investing decent fund in the R&D activities can potentially bring a positive effect to guarantee a good firm performance. However, the R&D Intensity is not the only essential factor contributing to the development of semiconductor industries, as a range of non-multicollinearity factors, such as cash_equivalent, free_cash_flow_ratio, net_profit, and operation_expense, also play essential roles in affecting the firm performance. In accordance with the practical evidence based on the experimental results, these aforementioned factors bring different effects to different-sized semiconductor firms, and the R&D Intensity has a great effect upon the medium-sized semiconductor firms in particular. As a whole, the factors "R&D Intensity, cash_equivalent, free_cash_flow_ratio, net_profit, and operation_expense" are evaluated as the general high-impact factors of semiconductor industries by the proposed method, but their effects differ in different-sized semiconductor firms. Moreover, the XGBoost regression is also capable of predicting values of long period of time to realize the potential trend in the future, as shown in the Figure 11.

In the above figure, the GR_netvalue of 2019 in each company is predicted based on the financial announcement, 1993-2018. In addition, the limitation of this work is if the information provided by the financial announcement is not complete and precise sufficiently the prediction accuracy may decline. Thus, although it is an intelligent scheme to provide the industries an opportunity to think about how to make an optimal deployment of business plan for the coming future, it is further believed that the proposed scheme with the features of high-reliable gain score can potentially predict toward the trend of next year by taking more sophisticated mechanism into consideration, such as the related technology policy designed by the government and the trend of semiconductor industries in the world. Further, the proposed method can be applied to evaluate companies from different sectors, especially if they have the same high technology and high development characteristics.

REFERENCES

[1] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1525–1531.
[2] R. S. Menjoge and R. E. Welsch, "A diagnostic method for simultaneous feature selection and outlier identification in linear regression," *Comput. Statist. Data Anal.*, vol. 54, no. 12, pp. 3181–3193, Dec. 2010.
[3] S. Luo and Z. Chen, "Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces," *J. Stat. Planning Inference*, vol. 143, no. 3, pp. 494–504, Mar. 2013.
[4] G. K. Morbey and R. M. Reithner, "How R&D affects sales growth, productivity and profitability," *Res.-Technol. Manage.*, vol. 33, no. 3, pp. 11–14, 1990.
[5] G. Erickson and R. Jacobson, "Gaining comparative advantage through discretionary expenditures: The returns to R&D and advertising," *Manage. Sci.*, vol. 38, no. 9, pp. 1264–1279, 1992.
[6] K. Ito and V. Pucik, "R&D spending, domestic competition, and export performance of Japanese manufacturing firms," *Strategic Manage. J.*, vol. 14, no. 1, pp. 61–75, 1993.
[7] L. Gu, "Product market competition, R&D investment, and stock returns," *J. Financial Econ.*, vol. 119, no. 2, pp. 441–455, 2016.
[8] C. L. Lee and H. C. Wu, "How do slack resources affect the relationship between R&D expenditures and firm performance?" *R&D Manage.*, vol. 46, no. S3, pp. 958–978, 2016.
[9] Y. Chen and O. W. Ibhagui, "R&D-firm performance nexus: New evidence from NASDAQ listed firms," *The North Amer. J. Econ. Finance*, vol. 50, Oct. 2019, Art. no. 101009.
[10] P. Aghion, U. Akcigit, A. Bergeaud, R. Blundell, and D. Hemous, "Innovation, income inequality, and social mobility," in *Proc. Vox CEPR's Policy Portal*, 2015, p. 28.
[11] J. Xu and J.-W. Sim, "Characteristics of corporate R&D investment in emerging markets: Evidence from manufacturing industry in China and South Korea," *Sustainability*, vol. 10, no. 9, p. 3002, 2018.
[12] C. B. Gibson and J. Birkinshaw, "The antecedents, consequences, and mediating role of organizational ambidexterity," *Acad. Manage. J.*, vol. 47, no. 2, pp. 209–226, Apr. 2004.
[13] J. Sydow and U. Staber, "The institutional embeddedness of project networks: The case of content production in german television," *Regional Stud.*, vol. 36, no. 3, pp. 215–227, May 2002.
[14] C. L. Wang and P. K. Ahmed, "Dynamic capabilities: A review and research agenda," *Int. J. Manage. Rev.*, vol. 9, no. 1, pp. 31–51, Mar. 2007.
[15] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: An application of large-scale online learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 681–688.
[16] R. F. Hurley and G. T. M. Hult, "Innovation, market orientation, and organizational learning: An integration and empirical examination," *J. Marketing*, vol. 62, no. 3, pp. 42–54, Jul. 1998.
[17] U. Akgun, E. A. Albayrak, G. Aydin, W. Clarida, F. Duru, V. Khristenko, A. Moeller, J. Neuhaus, Y. Onel, J. Wetzel, D. Winn, and T. Yetkin, "Quartz plate calorimeter prototype with wavelength shifting fibers," *J. Instrum.*, vol. 7, no. 7, Jul. 2012, Art. no. P07004.
[18] M. B. Lieberman and S. Asaba, "Why do firms imitate each other?" *Acad. Manage. Rev.*, vol. 31, no. 2, pp. 366–385, Apr. 2006.
[19] P. Aghion, R. Veugelers, and C. Serre, *Cold Start for the Green Innovation Machine*. Brussels, Belgium: Bruegel Policy Contribution, Dec. 2009.
[20] J. Barney and T. Felin, "What are microfoundations?" *Acad. Manage. Perspect.*, vol. 27, no. 2, pp. 138–155, May 2013.

[21] T. J. Christensen and J. Snyder, "Progressive research on degenerate alliances," *Amer. Political Sci. Rev.*, vol. 91, no. 4, pp. 919–922, Dec. 1997.

[22] S. R. Grenadier and A. M. Weiss, "Investment in technological innovations: An option pricing approach," *J. Financial Econ.*, vol. 44, no. 3, pp. 397–416, Jun. 1997.

[23] H. R. Greve, "A behavioral theory of R&D expenditures and innovations: Evidence from shipbuilding," *Acad. Manage. J.*, vol. 46, no. 6, pp. 685–702, 2003.

[24] T. K. Lant and D. B. Montgomery, "Learning from strategic success and failure," *J. Bus. Res.*, vol. 15, no. 6, pp. 503–517, Dec. 1987.

[25] G. J. Lucas, J. Knoben, and M. T. Meeus, "Contradictory yet coherent? Inconsistency in performance feedback and R&D investment change," *J. Manage.*, vol. 44, no. 2, pp. 658–681, 2018.

[26] T. M. Mitchell, "Artificial neural networks," *Mach. Learn.*, vol. 45, pp. 81–127, Oct. 1997.

[27] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, "Identifying IT purchases anomalies in the brazilian government procurement system using deep learning," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 722–727.

[28] R. Alkhaddar, T. Wooder, B. Sertyesilisik, and A. Tunstall, "Deep learning approach's effectiveness on sustainability improvement in the UK construction industry," *Manage. Environ. Qual., Int. J.*, vol. 23, no. 2, pp. 126–139, Feb. 2012.

[29] N. Bien, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002699.

[30] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," in *Proc. SPIE*, vol. 9785, Dec. 2016, Art. no. 97850Z.

[31] J. De Fauw, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018.

[32] P. R. Jeyaraj and E. R. Samuel Nadar, "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm," *J. Cancer Res. Clin. Oncol.*, vol. 145, no. 4, pp. 829–837, Apr. 2019.

[33] V. Singh and N. K. Verma, "Deep learning architecture for high-level feature generation using stacked auto encoder for business intelligence," in *Complex Systems: Solutions Challenges Economics, Management and Engineering*. Springer, 2018, pp. 269–283.

[34] Y.-W. Chang and C.-Y. Tsai, "Apply deep learning neural network to forecast number of tourists," in *Proc. 31st Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2017, pp. 259–264.

[35] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: Deep portfolios," *Appl. Stochastic Models Bus. Ind.*, vol. 33, no. 1, pp. 3–12, Jan. 2017.

[36] A. Navon and Y. Keller, "Financial time series prediction using deep learning," 2017, *arXiv:1711.04174*. [Online]. Available: http://arxiv.org/abs/1711.04174

[37] A. L. Calvez and D. Cliff, "Deep learning can replicate adaptive traders in a Limit-Order-Book financial market," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1876–1883.

[38] H. Yan, J. Wan, C. Zhang, S. Tang, Q. Hua, and Z. Wang, "Industrial big data analytics for prediction of remaining useful life based on deep learning," *IEEE Access*, vol. 6, pp. 17190–17197, 2018.

[39] W. Ma, F. Cheng, and Y. Liu, "Deep-Learning-Enabled on-demand design of chiral metamaterials," *ACS Nano*, vol. 12, no. 6, pp. 6326–6334, Jun. 2018.

[40] J. Deutsch and D. He, "Using deep learning-based approach to predict remaining useful life of rotating components," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 1, pp. 11–20, Jan. 2018.

[41] W. Serrano, "The random neural network with a genetic algorithm and deep learning clusters in fintech: Smart investment," in *Proc. Int. Conf. Artif. Intell. Appl. Innov.*, 2018, pp. 297–310.

[42] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. vol. 37, no. 15. Belmont, CA, USA: Wadsworth Int. Group, 1984, pp. 237–251.

[43] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[44] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

[45] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[46] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, vol. 96, 1996, pp. 148–156.

[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[48] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2006.

[49] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 12, pp. 1189–1232, Oct. 2001.

[50] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[51] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decis. Support Syst.*, vol. 104, pp. 38–48, Dec. 2017.

[52] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with EMRs," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2014, pp. 556–559.

[53] B. Kratzwald, S. Iliä, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018.

[54] L. Zhang and C. Zhan, "Machine learning in rock facies classification: An application of XGBoost," in *Proc. Int. Geophys. Conf.*, Qingdao, China, Apr. 2017, pp. 1371–1374.

[55] K. Baraniak, "ISMIS 2017 data mining competition: Trading based on recommendations-XG boost approach with feature engineering," in *Intelligent Methods and Big Data in Industrial Applications*. Springer, 2019, pp. 145–154.

[56] L. Podlodowski and M. Kozlowski, "Application of XGBoost to the cybersecurity problem of detecting suspicious network traffic events," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2019, pp. 5902–5907.

[57] D. Nielsen, *Tree Boosting With Xgboost-Why Does Xgboost Win 'Every' Machine Learning Competition*. Trondheim, Norway: NTNU, 2016.

PING-YU HSU graduated from the CSIE Department, National Taiwan University, in 1987. He received the master's degree from the Computer Science Department, New York University, in 1991, and the Ph.D. degree from the Computer Science Department, UCLA, in 1995. He is currently a Professor with the Business Administration Department, National Central University, Taiwan, and the Secretary in Chief of the Chinese ERP Association. He is currently the Dean of the School of Management, National Central University. His research interests include business data related applications, business analytics, data mining, business intelligence, and adoption issues of enterprise systems. He has published more than 100 journal articles and conference papers. His articles have been published in *Decision Support Systems*, *European Journal of Information Systems*, the IEEE TRANSACTIONS, *Information Systems*, *Information Sciences*, and various other journals.

I-WEN YEH is currently pursuing the Ph.D. degree with the Department of Business Administration, National Central University, Taiwan. Her research interests include enterprise resource planning (ERP) and big data analysis.

CHING-HSUN TSENG received the master's degree from the Institute of Management of Technology, National Chiao Tung University, Taiwan, in 2019. He is currently pursuing the M.Phil. degree with the Department of Computer Science, The University of Manchester, U.K. Before that, he was working as a Data Engineer with AI Company, Taiwan. His research interests include developing innovated machine learning algorithms, especially in semi-supervised learning, and deep learning.

SHIN-JYE LEE received the M.Sc. (Eng) degree from the Department of Computer Science, The University of Sheffield, U.K., in 2001, the M.Phil. degree from the Judge Business School, University of Cambridge, U.K., in 2011, and the Ph.D. degree from the School of Computer Science, The University of Manchester, U.K., in 2012. He is currently an Associate Professor with the Institute of Technology Management, National Chiao Tung University, Taiwan. Before that, he was the Professor of National Pilot School of Software, Yunnan University, China, and he also made his academic career in Poland, in 2012 winter. In addition, he also had practical experiences in Fujitsu and Microsoft, from 2002 to 2005. His research interests include machine learning, computational intelligence and decision support systems, operational research, and technology policy, especially for the climate change issues and energy prediction.