IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 23, 2019, accepted November 7, 2019, date of publication November 15, 2019, date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2953807

A Contrastive Study of Machine Learning on Energy Firm Value Prediction

CHUQING ZHANG[1], HAN ZHANG[2], AND DUNNAN LIU[1]
[1]School of Economics and Management, North China Electric Power University, Beijing 102206, China
[2]School of Cyber Science and Technology, Beihang University, Beijing 100191, China

Corresponding author: Han Zhang (zhhan@buaa.edu.cn)

This work was supported by the China Postdoctoral Science Foundation under Grant 2019M650610.

ABSTRACT For decades, a high prediction error rate of firm value assessment has been reported by using traditional financial evaluation methods, therefore develop a suitable assessment tool to improve firm value prediction accuracy is in urgent. This paper provides a comprehensive review and statistical comparison of six machine learning models: K-Nearest Neighbor, Decision Trees, Support Vector Regression, Artificial Neutral Network, AdaBoost, and Random Forest in oil firm and power firm value prediction. Based on nearly 5000 M&A items, this paper finds that for both oil and power industries, the prediction error of ANN is the lowest in all the three measurement terms. ANN performs better than the other five ML models by 18% at least for oil industry, and outperforms the others by 19% for power industry. It shows that ANN models can produce both accurate and reasonably understandable prediction results. ANN can be applied to a wide range of M&A decisions and value assessment for energy firms.

INDEX TERMS Firm value, machine learning, ANN, energy firm, M&A.

I. INTRODUCTION

Mergers and acquisitions (M&A) is the combination of the assets and liabilities of two companies to form a new business entity. Accurately evaluating the value of target companies not only facilitates the deal completion, but also brings great economic value to the acquirer. Yet, determining the value of target companies is not easy. Many studies have shown that predicted firm value results in failure with huge divergence to realistic value. Energy firm value prediction is one of such area that poses a lot pressure on evaluators' judgment. In many cases, even with rigorous logical reasoning and statistical regression, decision maker can hardly make reliable predictions. Part of the reason for misestimation lies in the unsuitable analytical tool. Generally speaking, there are two methodologies in predicting firm price: (1) traditional linear regression methods, and (2) machine learning approach. The traditional linear regression methods, including Balance-Sheet-Based methods, Income Statement and Market-Based methods, Discounted Cash Flow method, often reach its limit constraint by strict statistical assumptions, such as normality, linearity, and independence. It makes them unable to discover unseen patterns under complex scenarios and often falls short in nonlinear problems. Therefore, it is essential to discover an accurate value prediction tool for energy firm. For these reasons, the second category, ML approach is gaining more and more popularity recently.

As a computer-based interdisciplinary tool, machine learning(ML) has outstanding ability in processing large quantity and high dimensional data [14], [38]. Such ability has made machine learning models widely used in many cases, for example, in pattern recognition, classification, data mining, and forecasting [10], [43]. In financial area, ML has shown its strength in stock price prediction [18], [21], [22], house price prediction [31], [35], firm value prediction [24], [26], [28] and so on. ML methods are able to discover the unknown function, dependency or structure between inputs and outputs which are impossible to be represented by explicit algorithms [43]. Yet which ML method, like K-Nearest Neighbors, Decision Trees, Neural Networks, Support Vector Machines is more proper applied to firm value filed is still up in the air. Reference [24] proves Ensemble Approach performs better than CART and ANN in prediction corporate dividends. Reference [26] reveals that Decision Trees and the ensemble models outperform ANN in firm value prediction. Reference [28] demonstrates BPANN provides superior outcomes than linear regression method in firm market value assessment and influential factors selecting.

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

This paper aims to discover an accurate prediction model for energy firm value assessment by providing a comprehensive review and statistical comparison of ML methods. By using the actual M&A transaction price as bench mark, it figures out the most accuracy prediction model with lowest prediction error and identify the related variables which influence the deal price most. This paper compares the prediction performance of four supervised learning models, namely K-Nearest Neighbor (KNN), Decision Trees (DT), Support Vector Regression (SVR), Artificial Neural Networks (ANN), and two ensemble models-AdaBoost (AB), Random Forest (RF), and the results show that the ANN could predict both oil and power firm values more accurately than its competitors on all error terms stably. It is the most suitable tool for energy firm value assessment. Besides, it is also believed that this paper is the first to provide a comprehensive review of KNN, DT, SVR, ANN, AB, and RF in firm value evaluation field.

Thus, this paper makes the following main contributions:

- This paper tests the performance of four supervised machine learning models and two ensemble machine learning methods on a new scenario-firm value assessment. The results reveal that ensemble ML methods, i.e. AB, RF cannot provide reliable prediction results than those single learning models in energy firm value prediction scenario.
- Based on the real-life data, this paper finds that for oil industry, the prediction error of ANN is 5.15 on average, which is about 18% lower than other models. For power industry, the lowest prediction error is achieved by ANN at 8.64, 19% lower than other models.

This paper is organized as follows. Sec. II, is a brief introduction to the six commonly used prediction ML methods. Sec. III is sample description with data preparation. Sec. IV describes research design. Sec. V provides with experiment results and analysis model. Finally, concludes the paper in Sec. VI.

II. BACKGROUND AND RELATED WORK

In the following part, this paper provides a comprehensive literature review of the six ML models-KNN, DT, SVR, ANN, AB and RF, as well as their application in firm value assessment.

A. K-NEAREST NEIGHBOR REGRESSOR MODEL

KNN is a simple, effective non-parameters method. It calculates the similarity between a target object and the most similar k-nearest neighbors in the training sample set by Euclidean distance,

$$d(x, x_i) = \sqrt{\sum_{i=1}^{n}(x - x_i)^2} \tag{1}$$

where x is the target object and x_i is the i-th similar nearest neighbors. Then according to a majority vote of its neighbors, the target will be assigned to the most common class among its k-nearest neighbors. The classification $\delta(x_i, c_i)$ is for x_i with respect to class c_i can be expressed as:

$$\delta(x_i, c_i) = \begin{cases} 1, & \text{if } x_i \in c_i \\ 0, & \text{if } x_i \notin c_i \end{cases} \tag{2}$$

According to a majority vote of its neighbors, the target will be assigned to the most common class among its k-nearest neighbors. Reference [39] finds KNN is effective in rice price prediction. Reference [7] proves KNN performs well with stock data. Reference [4] shows depending on the actual stock prices data, the KNN prediction results are close and almost parallel to actual stock prices.

B. DECISION TREE REGRESSION MODEL

A regression tree is a decision tree that deals with a continuous target attribute. Regression tree technique constructs a single regression tree by repeatedly splitting the data into mutually exclusive groups or nodes by means of an efficient recursive partitioning algorithm. Theoretically, when the tree construction is finished, all the instances in a node are of the same class [29]. In reality, the optimized tree is to achieve balance between complexity (size) and prediction capacity [15], therefore pruning the tree is also necessary. Reference [40] attempts to use a decision tree approach to deal with the problem of stock selection. It is shown that the tree approach performs significantly better than those built by simple stock screening and ranking models. Reference [13] replaces linear regression approach in housing price prediction with decision tree, and proves it to be an important statistical pattern recognition tool by offering relationship between house prices and housing characteristics. Furthermore, [32] and [23] also confirm that decision tree is useful in price prediction.

C. SUPPORTED VECTOR REGRESSION MODEL

SVRs are classification techniques based on statistical learning theory [11]. It can handle linear indivisible problems by using nonlinear kernel support vector machines. Based on the kernel function which can achieve a hyperplane that lies "close" to as many of the data points as possible [21], SVM maps non-linear input vectors into a high-dimensional feature space.

$$minimize \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\zeta_i + \xi_i) \tag{3}$$

subject to :

$$y_i - f(x) \le \epsilon + \xi_i \tag{4}$$

$$y_i - f(x) \le \epsilon + \xi_i \tag{5}$$

$$f(x) - y_i \le \epsilon + \zeta_i \tag{6}$$

$$\zeta_i, \xi_i \ge 0 \tag{7}$$

where C controls the penalty imposed on residuals, and the two slack variables ζ_i and ξ_i represent the distance from actual values to the corresponding boundary values. The training points that are closest to the optimal separating hyperplane are called support vectors [24]. Reference [31] compares the result of SVR with that of least square regression (LSR) and vector autoregressive (VAR), it proves that SVR is a good predictor of CPI (Consumer Price Index). Reference [18] also confirms with previous studies that SVR has predictive power in stock price prediction. Reference [8] testifies that SVR model performs much better than linear regression model in prediction accuracy.

D. ARTIFICIAL NEURAL NETWORK MODEL

Artificial Neural Network is a network composed with determined architecture, weights, biases and activation function [12], [46]. During training, a weighted sum of the inputs is calculated as

$$H_t = f(\sum_{i=1}^{n} w_{ti}x_i + b_t) \tag{8}$$

where w_{ti} is the weight on connection from the i-th to the t-th node; x_i is an input data from input node; N is the total number of input nodes; and b_t denotes a bias on the t-th hidden node [33]. When the neurons of the input layer are activated by the activation function, the activation values will propagate from the input layer to the intermediate layers. Then such process will iterate from hidden layer to output layer. At the end of the feedforward process, the network back-propagates calculated error between target and actual outputs for further iterative weight adjustments [28]. The processes of information forward propagation and error back propagation are the learning and training process of the neural network. This process continues until the output error reaches a specified value or a predetermined number of learnings is reached [25]. Reference [33] demonstrates that based on the values of RMSE, the ANN models achieve a better forecast performance than the regression models for debt ratio. Reference [9] compares ANN, Decision Trees and hybrid model in stock price prediction by finding out that ANN performs better than the other models. Reference [28] demonstrates BPANN provides superior outcomes than linear regression method in firm market value assessment and influential factors selecting. [2] shows that the ANN model possesses a good predictive ability for property valuation. Reference [3] shows that the ANN technique outperforms the HPM approach.

E. ADABOOST MODEL

AdaBoost works by re-weighted the misclassified samples iteratively. Namely, the misclassified samples of previous round with other new data are combined to form a new training samples, which will enter into a new training process in next round. Iterations stop when the specified number of iterations or the expected error rate is reached. In each iteration, the instances that are misclassified in the previous iteration will be assigned more weight. Reference [16], [45] finds that AdaBoost method is effective in predicting firm price and value. Reference [17] suggests that in companies' financial distress prediction, the model based on AdaBoost algorithm has a higher overall accuracy than the model based on Neural Network. Reference [44] shows that the AdaBoost model possesses much higher predictive capacity than a regression-based model in consumer demand prediction.

F. RANDOM FOREST MODEL

Random forest is the combination of several trees. By using the bootstrap method to generate multiple training sets and construct a decision tree for each training set, RF improves generalization ability of a single decision tree. Moreover, rather than choosing the best split among all attributes, it randomly selects a subset of the attributes and chooses the best split among them [37], which can successfully avoid the problem of overfitting problem. Reference [15] finds that the prediction errors obtained with RF are similar or even lower than those obtained with CART and Bagging methods in electricity price forecasting. Reference [6] indicates that RF is the top algorithm, which performs better than Support Vector Machines, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression in price prediction.

III. MATERIALS AND ANALYSIS

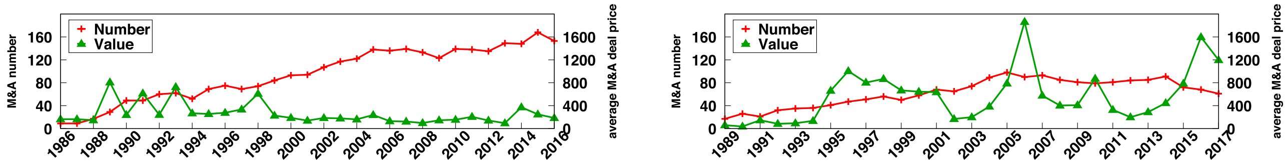
In this part, this paper offers a comprehensive introduction of the M&A dataset, as well as the variables included in the model.

A. DATASET DESCRIPTION

The Merger & Acquisition database covers more than 273,000 M&A transactions worldwide since the year 1983. The detailed information elements it provides regarding to target's firms information as well as deal information, including target firms' financial performance, firm type, industry, transaction dates, deal value, deal type, acquisition share. Since the energy industry has a very broad definition, including infrastructure, petroleum, electricity, renewable energy, and so on [27], firm value may vary greatly in different industry sectors. In this paper, two energy industries-oil industry and power industry are chosen as targets. For oil industry, this paper all together gets 3078 samples. For power industry, the sample size is 1834. Specific descriptions of the obtained data are shown in Fig.1(a) and Fig.1(b). Some sample data could be downloaded from [1].

B. VARIABLE DESCRIPTION

Firm value is the deal value, i.e. the actual evaluation price offered by the acquiring firm. *Firm size* is measured as the natural logarithm of the firm's total assets. *Asset Turnover Ratio* is measured by the ratio of total sales/total asset. *EBIT* is firm' s earnings before interest and taxes. *Net profit margin* equals to the value of net profit/revenue. *ROA* is the ratio of net income/total assets, and *ROE* is the ratio of net income/average shareholders' equity. From the perspective of financial profit and strength, ROA is commonly used to represent a firm' s financial outcome and effective utilization of input resources for market value creation, which represents



(a) Sample distribution of oil industry.

(b) Sample distribution of power industry.

FIGURE 1. Performance comparison of testing dataset.

TABLE 1. Indicators of firm value.

Variables	Variable explanation
Transaction value	Deal value
Firm size	Total assets
Asset Turnover Ratio	Total sales/ total asset
EBIT	Earnings before interest and taxes
Net profit margin	Net profit/revenue
ROA	Net income/total assets
ROE	Net income/ average shareholders' equity
Cash debt ratios	Cash/ pay debt
Capex	Firm's investment growth change
Total debt to assets	Total debts/ total assets
Firm type	Categorical variables
Nationality	The nationality of target firms
M&A type	Homogeneous M&A coded as 1, and others as 0
Acquisition year	The year the acquisition initiated
Share	The percentage that the acquire count in the deal

short-term operational performance. *Cash to pay debt ratio* is used as another financial tool to assess whether a firm has adequate cash to pay debts in order to cope with the high risk that SMEs are likely to face in the market [28]. Firm growth is measured by the logarithmic value of investment growth change in firm capital expenditures, i.e. the *Capex*. *Total Debt to Assets* is used as the firm leveraging indicator. Moreover, since there is substantial evidence that the firm type of the target plays an important role in firm value assessment, we included *organization form* as categorical variables. In this paper, we use 1-6 to represent government-based firm, joint venture; mutual firm; private firm; public firm and subsidiary. Since the intra-national M&A is a small potion, the influence of different country can be overlooked. Although we do not include it in the model, we still think the *nationality* of target firm as an important influencing factors. *M&A type*, i.e. whether the deal is made within the same industry or not is an indispensable factor in value assessment. Therefore, homogeneous M&A coded as 1, and heterogeneous M&A as 0. *Acquisition year* is also included in the model to mitigate the influence of inflation. The *share* that the transaction fee can buy also tend to affect firm value, hence it is necessary to include it into the model. The detailed information of these 15 attributes please refer to TABLE 1.

As a preliminary step toward empirical analysis, the descriptive statistics for oil and power industry is presented in TABLE 2 and TABLE 3 individually. Their kurtosis values (the oil industry is 735, and power industry is 215.61), indicates that the observations' distribution are concentrated as spike, which cannot satisfy the normal distribution for linear test. Therefore, linear regression model is not suitable in this occasion [42].

IV. EXPERIMENT DESIGN

The main task in this section includes preprocessing data, selecting prediction measurements and setting model specifications. In this section, this paper performs a systematic evaluation of the six ML methods on the SDC Merger & Acquisition Platinum.

A. DATA PREPROCESSING

Since the predicting items of SDC M&A database is composed of firm level information as well as deal level of information, normalization is needed. In this study, data are scaled into the interval of [0, 100] by using the following formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} * 100 \tag{9}$$

where x is the original value, x' is the scaled value, max(x) is the maximum value of feature x, and min (x) is the minimum value of feature x.

The model's prediction performance is measured by Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Deviation (MAE) [2]. MAPE, RMSE, MAE are the measures of the deviation between actual and predicted values. The smaller the values of MAPE, RMSE, MAE, the closer the predicted values to those of the actual value [22].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{p_i - \hat{p}_i}{\hat{p}_i}| \times 100\% \tag{10}$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \tag{12}$$

where n is the number of observations, p_i is the actual property price, $\hat{p}_i$ is the predicted property price from the model.

The machine learning software we use is scikit-learn [34] in Python. Scikit-learn is built on NumPy, SciPy, and matplotlib. It is a simple and efficient tool for data mining and data analysis. Scikit-learn contains six basic functions, namely classification, regression, clustering, data dimensioning, model selection and data preprocessing.

B. K-NEAREST NEIGHBOR

There is only one parameter-K in KNN model. K means how many neighbors are going to be selected into the model as

TABLE 2. Data description of oil industry.

Variables	Mean	S.D.	Min	Max	Skewness	Kurtosis	Shapiro-Wilks	P value
Value	357.75	2334.93	0.001	69445	25.240	735.004	0.101	0
ROA	-5.10	26.45	-114	380	-2.532	7.045	0.703	0
ROE	-9.78	50.57	-258	83	-2.839	9.891	0.693	0
M&A type	0.40	0.49	0	1	0.402	-1.842	0.622	0
Firm type	5.34	1.08	1	6	-2.132	4.209	0.628	0
M&A year	11.17	6.99	0	31	0.513	-0.695	0.951	0
Share	61.81	38.26	0	100	-0.247	-1.633	0.802	0
Nationality	123.99	78.98	1	216	-0.135	-1.599	0.857	0
Asset	7.13	3.21	0	16	0.088	-0.535	0.991	0
ART	1120.15	6397.69	0	59304	7.744	63.676	0.167	0
EBIT	8.88	0.98	-1	13	-3.974	55.824	0.392	0
NPM	1437.77	41114.79	-129143	1286056	29.412	914.04	0.032	0
CDR	1807.12	53636.60	0	1735477	32.315	1045.479	0.013	0
Capex	9.17	3.75	-3	17	-0.479	-0.122	0.98	0
TDA	301.11	1662.30	0	21736	8.84	93.505	0.177	0

TABLE 3. Data description of power industry.

Variables	Mean	S.D.	Min	Max	Skewness	Kurtosis	Shapiro-Wilks	P value
Value	644.61	2850.30	0.001	56266	12.813	215.614	0.196	0
ROA	1.98	14.91	-114	38	-4.409	26.579	0.535	0
ROE	1.77	37.34	-258	83	-4.116	22.863	0.561	0
M&A type	0.53	0.50	0	1	-0.102	-1.995	0.636	0
Firm type	5.26	1.12	1	6	-2.141	4.554	0.653	0
M&A year	11.37	6.52	0	28	-0.357	-0.786	0.964	0
Share	57.80	36.94	0	100	-0.049	-1.601	0.843	0
Nationality	124.42	74.62	1	219	-0.084	-1.546	0.88	0
Asset	8.70	2.63	0	16	-0.248	0.422	0.989	0
ART	140.11	2222.35	0	54697	21.55	504.454	0.037	0
EBIT	8.93	0.89	-1	13	-4.69	70.317	0.402	0
NPM	5812.83	166072.75	-59366	4528231	27.252	743.106	0.017	0
CDR	195.28	3425.88	0	87929	23.491	587.196	0.031	0
Capex	9.82	3.33	-7	17	-0.872	1.275	0.958	0
TDA	55.17	1218.88	0	33062	26.819	726.63	0.021	0

reference, which determines classification results of KNN. A smaller K may provide less approximation error, but it will lead to overfitting problem. Whereas a larger K value may generate more prediction error, the model will be simpler. In practice, trial and error are commonly used to determine the optimal K value. After rigorous testing, this paper sets K = 11 for oil industry, and K = 9 for power industry. In such condition, the model of KNN performs the best.

C. DECISION TREE REGRESSION

Decision Tree parameters are summarized in TABLE 4. This default rule often works well across a broad range of problems [5]. In details, most of the criterions for both oil firms and power firms are the same. The splitting criterion used is mse; the minimum sample leaf is 1; the min sample split is 2; the min weight fraction leaf is 0; the maximum leaf nodes is none; the minimum impurity split is none. The only difference lies in maximum depth. For oil industry, the maximum depth is 8, and for power industry it is 10.

TABLE 4. Training parameters of Decision Tree model.

Parameter type	Ooil industry	Power industry
Splitting criterion	MSE	MSE
Maximum depth	8	10
Minimum samples leaf	1	1
Minimum samples split	2	2
Minimum weight fraction leaf	0	0
Maximum leaf nodes	None	None
Minimum impurity split	None	None

D. SUPPORTED VECTOR REGRESSION

The accuracy of SVR model depends on the chosen kernel function. Like most of other studies, RBF kernel is used in the SVR model. Two parameters, i.e. penalty parameter for the error C and kernel parameter γ are chosen by grid search method [19]. The range of C and γ are $C = \{1, 2^1, 2^2, 2^3, 2^4\}$, and $\gamma = \{2^{-3}, 2^{-2}, 2^{-1}, 1, 2^1, 2^2\}$. By multiple times of training, the best combination of the error C and kernel parameter γ are set respectively as 4 and 0.5 for oil industry and 16, 0.125 for power industry.

E. ARTIFICIAL NEURAL NETWORKS

In this study, three-layer fully connected back-propagation neural networks (BPNN) are used as benchmarks. In this paper, the input layer contains 14 nodes representing the 14 attributes, and one node for the output layer. After iteration, an ANN architecture of 14–50–1 (14 input variables,

TABLE 5. Training parameters of AdaBoost model.

Parameter type	Value for oil industry	Value for power industry
Base estimator	None	None
N_estimator	60	50
Learning rate	1	1
Loss	square	square
Random state	5	9

TABLE 6. Training parameters of Random Forest model.

Parameter type	Value for oil industry	Value for power industry
N_estimator	100	50
Maximum depth	10	7
Maximum feature	sqrt	sqrt
Oob_score	True	True
Random state	5	10

1 hidden layer with 50 neurons and 1 output) generated to be the best network for oil industry, and an architecture of 14–75–1 is the best for power industry.

F. ADABOOST

The parameters are summarized in TABLE 5. Base estimator indicating the algorithm is used. If none, then the base estimator is DecisionTreeRegression. N_estimator is the number of subtree. This paper sets 60 for oil industry and 50 for power industry. Learning rate shrinks the contribution of each classifier, and normally set as 1. The loss parameter is used to specify the type of model calculation error. In this paper, square is used for loss parameter. Random state represents the setting of the random number seed.

G. RANDOM FOREST

RF has several unique parameters. One is maximum feature, and the other is oob_score. Maximum feature is the number of features. Increasing maximum features generally improves the performance of the model, but reduces the diversity of the subtree. This paper sets the number of variables to the square root of the total number of predictors. Oob_score indicates whether the sample outside the bag is used to evaluate the quality of the model. This paper sets true both power and oil industry. Detailed parameter information is presented in TABLE 6.

V. EXPERIMENT RESULTS

In this section, the evaluation results of each single ML model and the two ensemble models are presented. Then, the efficiency and robustness of all these methods is evaluated.

A. PREDICTION RESULTS

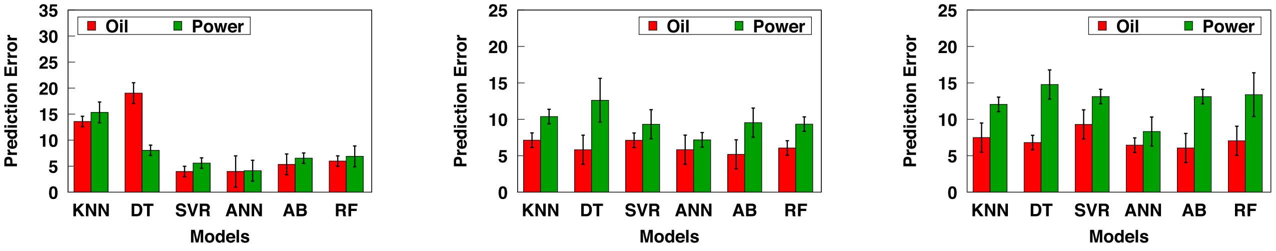
In this part, the prediction results under different metrics are provided. As the prediction results have close relationships with parameters selection, this paper splits the total data randomly into training and testing dataset with training data size counting 80%, and testing data size counting 20%, where the training dataset is used to fit parameters and the testing dataset is used to assess the models. In KNN, the errors mainly come from the nearest neighbors, since the overlarge value could lead to underfitting while small value opposites. To improve the degree of accuracy, this paper caps the value of it between 5 and 12. Decision tree learners can create over-complex trees that do not generalize the data well. To improve its performance, this paper caps the maximum depth of the tree at 10 to avoid this problem. In SVR, this paper adopts slack variable to reduce the deviations. The ANN model has many hidden layers which could lead to overfitting easily. This paper employs L2 regularization term to help in avoiding this by penalizing weights with large magnitudes. For AdaBoost, this paper caps the maximum number of estimators between 20 and 60 to avoid overfitting. Similar to decision-tree, this paper caps the maximum depth of the Random Forest at 10 to reduce deviations.

Fig. 2 shows the results of training dataset and Fig. 3 shows the corresponding testing dataset. In each experiment, this paper repeats the data splitting process 100 times, where the error bar paints the mean and mean ± standard deviation value. Comparing the two pictures, it can be seen that the results of testing dataset approximate to the training's, since the above technologies to avoid deviations are adopted when deriving the moderate parameters. Fig. 2(a) and Fig. 3(a) show the results of MAPE, which presents the average error ratio. It can be seen that ANN performs best, followed by SVR, KNN, RF, AdaBoost and DT. For power industry, the rankings are similar. Fig. 2(b) and Fig. 3(b) depict MAE, which presents the average absolute error. It shows that ANN performs close to AdaBoost and Random Forest and slightly better than others. Fig. 2(c) and Fig. 3(c) are the result of RSME, which is similar to MAE and MAPE.

B. ROBUST CHECK OF CLASSIFIERS

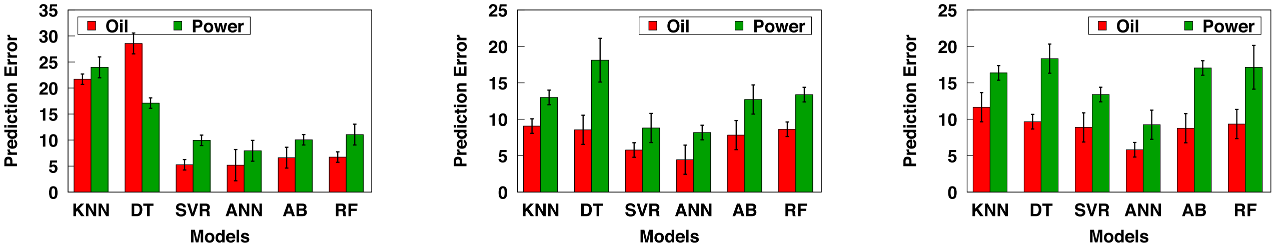
In this part, models' robustness under different level of training data size is checked. Specifically, from TABLE 7 it can be observed that all the machine learning methods achieve high training accuracy in predicting oil firm value. The predictive errors of all the models decrease progressively from the smallest to the biggest size group. The value prediction results for power industry follows the same pattern (as shown in TABLE 8). All the three indicators decrease with the enlarging of data size.

From the above analysis it can be seen that no matter for oil firm value prediction, or for power firm evaluation, ANN provides the most satisfying results. These results are similar to the findings of [9], [25] and so on. All these studies report that the ANN technique possesses a reliable predictive ability that can address the non-linearity of property values and property attributes [2]. It excels in nonlinear modeling even when the theoretical basis with regard to associated variables is weak [28]. It is suitable for solving complex problems of internal mechanisms and extracting hidden relationships between variables based on its self-learning and self-adaptive capabilities. Therefore, ANN algorithms are the



(a) MAPE (b) MAE (c) RMSE

FIGURE 2. Performance comparison of training dataset.



(a) MAPE (b) MAE (c) RMSE

FIGURE 3. Performance comparison of testing dataset.

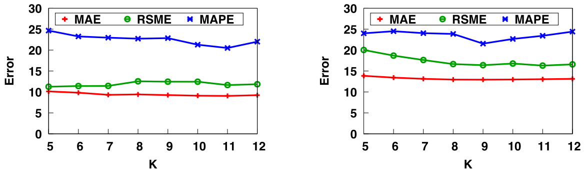
TABLE 7. Model robustness for oil firm value prediction.

	Measure of accuracy	KNN	DT	SVR	ANN	AdaBoost	RT
50% data as training data	MAPE	16.392	20.486	10.291	3.649	8.119	9.346
	MAE	9.203	6.488	5.395	7.086	5.360	6.358
	RMSE	8.984	7.425	7.547	7.059	6.211	7.850
60% data as training data	MAPE	16.454	21.049	9.885	3.957	8.536	9.422
	MAE	8.524	6.393	4.684	6.646	5.520	6.777
	RMSE	8.863	7.392	6.688	6.903	6.374	7.992
70% data as training data	MAPE	14.545	19.536	7.280	4.010	6.926	7.342
	MAE	7.585	5.929	7.157	6.061	5.177	6.648
	RMSE	7.781	6.590	10.299	6.129	5.675	7.635
80% data as training data	MAPE	13.586	19.034	3.983	3.973	5.346	5.982
	MAE	7.140	5.828	7.129	5.837	5.191	6.167
	RMSE	7.496	6.812	9.301	6.455	6.067	7.053

TABLE 8. Model robustness for power firm value prediction.

	Measure of accuracy	KNN	DT	SVR	ANN	AdaBoost	RT
50% data as training data	MAPE	20.773	4.556	9.288	4.545	9.161	7.923
	MAE	11.650	12.064	10.555	10.077	10.601	9.910
	RMSE	16.489	12.733	15.382	12.561	14.153	14.223
60% data as training data	MAPE	19.460	6.180	6.765	5.374	8.933	7.422
	MAE	12.410	12.746	9.411	9.368	10.231	9.777
	RMSE	18.004	13.443	14.367	13.233	13.831	13.900
70% data as training data	MAPE	15.023	6.953	6.480	4.578	8.413	7.342
	MAE	10.966	13.460	9.514	8.699	10.016	9.800
	RMSE	15.384	14.550	13.328	12.353	13.251	13.537
80% data as training data	MAPE	15.333	8.044	5.596	4.133	6.541	6.883
	MAE	10.379	12.615	9.320	7.183	9.541	9.341
	RMSE	12.049	14.778	13.122	8.325	13.115	13.388

most appropriate for this study [26], [36], [41]. SVR model performs good as well [20]. The average prediction error it provides for oil industry is around 6.30, and for power at 7.87. Both ANN, SVR are not sensitive to data distribution, and have a strong application ability [30]. Out of our expectation, the two ensemble models, AdaBoost and RF doesn't provide accuracy prediction results. The unbalanced data, and the outliers may lead to a decrease in its prediction accuracy. KNN model is average in model performance. And under different K values, its prediction accuracy also changes a



(a) Oil (b) Power

FIGURE 4. Performance comparison of KNN.

lot (Fig.4). DT is the most unsuitable model in our scenario. The low accuracy may partly caused by using binary segmentation to process continuous variables.

VI. DISCUSSION AND CONCLUSION

This paper is motivated to provide a suitable data mining techniques to discover hidden patterns for firm value assessment. Based on the SDC M&A data, this paper selects two industries as targets: oil industry and power industry. It first compares four single machine learning methods, i.e. KNN, DT, SVR, ANN with two ensemble machine learning models-AdaBoost, and Random Forest on firm value prediction. Results show that for oil industry, the average prediction error of ANN is 5.15, the lowest from matching with real M&A transaction prices. SVR, AB, RF, KNN and DT achieve prediction error at 6.31, 7.74, 8.24, 14.16, and 15.59 respectively. The accuracy rate for ANN is 18% higher than others at least. For power industry, the lowest prediction error is achieved by ANN at 8.64, followed by SVR, AB, RF, KNN and DT at 10.72, 13.27, 13.85, 17.78, and 17.84 respectively. The accuracy rate for ANN is 19% higher than others on average. It is shown that ANN models can produce both accurate and reasonably understandable prediction results. It can be applied to a wide range of M&A decisions and value assessment for energy firms.

REFERENCES

[1] *M&A Data Sample*. [Online]. Available: https://github.com/zhanghan1990/M-A-datasample

[2] R. B. Abidoye and A. P. C. Chan, "Modelling property values in Nigeria using artificial neural network," *J. Property Res.*, vol. 34, no. 1, pp. 36–53, Jan. 2017.

[3] R. B. Abidoye and A. P. C. Chan, "Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network," *Pacific Rim Property Res. J.*, vol. 24, no. 1, pp. 71–83, Jan. 2018.

[4] K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock price prediction using k-nearest neighbor (KNN) algorithm," *Int. J. Bus., Hum. Technol.*, vol. 3, no. 3, pp. 32–44, 2013.

[5] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2387–2403, Dec. 2013.

[6] M. Ballings, D. Van Den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, Nov. 2015.

[7] T. Ban, R. Zhang, S. Pang, A. Sarrafzadeh, and D. Inoue, "Referential KNN regression for financial time series forecasting," in *Proc. Int. Conf. Neural Inf. Process.* Springer, 2013, pp. 601–608.

[8] Y. Cao, B. Ashuri, and M. Baek, "Prediction of unit price bids of resurfacing highway projects through ensemble machine learning," *J. Comput. Civil Eng.*, vol. 32, no. 5, Sep. 2018, Art. no. 04018043.

[9] T.-S. Chang, "A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14846–14851, Nov. 2011.

[10] A. Choubineh, H. Ghorbani, D. A. Wood, S. Robab Moosavi, E. Khalafi, and E. Sadatshojaei, "Improved predictions of wellhead choke liquid critical-flow rates: Modelling based on hybrid neural network training learning based optimization," *Fuel*, vol. 207, pp. 547–560, Nov. 2017.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[12] W. P. S. Dias and R. L. D. Weerasinghe, "Artificial neural networks for construction bid decisions," *Civil Eng. Syst.*, vol. 13, no. 3, pp. 239–253, Jun. 1996.

[13] G.-Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: A decision tree approach," *Urban Stud.*, vol. 43, no. 12, pp. 2301–2315, Nov. 2006.

[14] H. Ghorbani, D. A. Wood, A. Choubineh, A. Tatar, P. G. Abarghoyi, M. Madani, and N. Mohamadian, "Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared," *Petroleum*, to be published.

[15] C. González, I. Juárez, and J. Mira-McWilliams, "Important variable assessment and electricity price forecasting based on regression tree models: Classification and regression trees, Bagging and Random Forests," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 11, pp. 1120–1128, Aug. 2015.

[16] Z. Guoying and C. Ping, "Forecast of yearly stock returns based on Adaboost integration algorithm," in *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, Nov. 2017, pp. 263–267.

[17] M. Hemmatfar and S. A. Hosseinipak, "Prediction of firms'financial distress using Adaboost algorithm and comparing its accuracy to artificial neural networks," *Revista QUID*, vol. 1, no. 1, pp. 2151–2158, 2017.

[18] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Stock price prediction using support vector regression on daily and up to the minute prices," *J. Finance Data Sci.*, vol. 4, no. 3, pp. 183–201, Sep. 2018.

[19] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study," *Decis. Support Syst.*, vol. 37, no. 4, pp. 543–558, 2004.

[20] H. Ince and T. B. Trafalis, "Kernel principal component analysis and support vector machines for stock price prediction," *IIE Trans.*, vol. 39, no. 6, pp. 629–637, Mar. 2007.

[21] H. Ince and T. B. Trafalis, "Short term forecasting with support vector machines and application to stock price prediction," *Int. J. Gen. Syst.*, vol. 37, no. 6, pp. 677–687, Dec. 2008.

[22] L.-J. Kao, C.-C. Chiu, C.-J. Lu, and J.-L. Yang, "Integration of nonlinear independent component analysis and support vector regression for stock price forecasting," *Neurocomputing*, vol. 99, pp. 534–542, Jan. 2013.

[23] P. Kaur, M. Goyal, and J. Lu, "Pricing analysis in online auctions using clustering and regression tree approach," in *Proc. Int. Workshop Agents Data Mining Interact.* Springer, 2011, pp. 248–257.

[24] H. S. Kim and S. Y. Sohn, "Support vector machines for default prediction of SMEs based on technology credit," *Eur. J. Oper. Res.*, vol. 201, no. 3, pp. 838–846, Mar. 2010.

[25] K. Kumar and S. Bhattacharya, "Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances," *Rev. Accounting Finance*, vol. 5, no. 3, pp. 216–227, Jul. 2006.

[26] C. Kuzey, A. Uyar, and D. Delen, "The impact of multinationality on firm value: A comparative analysis of machine learning techniques," *Decis. Support Syst.*, vol. 59, pp. 127–142, Mar. 2014.

[27] O. Kwon, S. Lim, and D. H. Lee, "Acquiring startups in the energy sector: A study of firm value and environmental policy," *Bus. Strategy Environ.*, vol. 27, no. 8, pp. 1376–1384, Dec. 2018.

[28] J. Lee and H.-B. Kwon, "Progressive performance modeling for the strategic determinants of market value in the high-tech oriented SMEs," *Int. J. Prod. Econ.*, vol. 183, pp. 91–102, Jan. 2017.

[29] R.-Z. Li, S.-L. Pang, and J.-M. Xu, "Neural network credit-risk evaluation model based on back-propagation algorithm," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 4, Jun. 2002, pp. 1702–1706.

[30] S.-J. Lin, "Integrated artificial intelligence-based resizing strategy and multiple criteria decision making technique to form a management decision in an imbalanced environment," *Int. J. Mach. Learn. Cyber.*, vol. 8, no. 6, pp. 1981–1992, Dec. 2017.

[31] P. Liu, J. Sun, L. Han, and B. Wang, "Research on the construction of macro assets price index based on support vector machine," *Procedia Comput. Sci.*, vol. 29, pp. 1801–1815, Jan. 2014.

[32] O. Özsoy and H. Şahin, "Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model," *Int. J. Housing Markets Anal.*, vol. 2, no. 2, pp. 167–178, 2009.

[33] H.-T. Pao, "A comparison of neural network and multiple regression analysis in modeling capital structure," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 720–727, 2008.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[35] V. Plakandaras, R. Gupta, P. Gogas, and T. Papadimitriou, "Forecasting the U.S. real house price index," *Econ. Model.*, vol. 45, pp. 259–267, Feb. 2015.

[36] J. L. M. Ramos, A. G. Exposito, J. Santos, A. T. Lora, and A. R. M. Guerra, "Influence of ANN-based market price forecasting uncertainty on optimal bidding," in *Proc. PSCC Power Syst. Comput. Conf.*, 2002.

[37] L. Rokach, "Decision forest: Twenty years of research," *Inf. Fusion*, vol. 27, pp. 111–125, Jan. 2016.

[38] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Comput. Oper. Res.*, vol. 106, pp. 210–224, Jun. 2019.

[39] D. Sinta, H. Wijayanto, and B. Sartono, "Ensemble K-nearest neighbors method to predict rice price in Indonesia," *Appl. Math. Sci.*, vol. 8, pp. 7993–8005, Nov. 2014.

[40] E. H. Sorensen, K. L. Miller, and C. K. Ooi, "The decision tree approach to stock selection," *J. Portfolio Manage.*, vol. 27, no. 1, pp. 42–52, Oct. 2000.

[41] B. Szkuta, L. Sanabria, and T. Dillon, "Electricity price short-term forecasting using artificial neural networks," *IEEE Trans. Power Syst.*, vol. 14, no. 3, pp. 851–857, Aug. 1999.

[42] S. Tonidandel and J. M. Lebreton, "Relative importance analysis: A useful supplement to regression analysis," *J. Bus. Psychol.*, vol. 26, no. 1, pp. 1–9, Mar. 2011.

[43] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, May 2017.

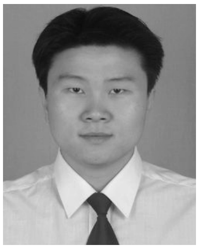
[44] K. W. Walker and Z. Jiang, "Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach," *J. Acad. Librarianship*, vol. 45, no. 3, pp. 203–212, May 2019.

[45] S. Yutong and H. Zhao, "Stock selection model based on advanced AdaBoost algorithm," in *Proc. 7th Int. Conf. Modeling, Identificat. Control (ICMIC)*, Dec. 2015, pp. 1–7.

[46] C. Zeng, C. Wu, L. Zuo, B. Zhang, and X. Hu, "Predicting energy consumption of multiproduct pipeline using artificial neural networks," *Energy*, vol. 66, pp. 791–798, Mar. 2014.

CHUQING ZHANG was born in Shijiazhuang, Hebei, China. She received the Ph.D. degree from Tsinghua University, China. She is currently with the School of Economic and Management, North China Electric Power University. Her research interests include project evaluation, technology innovation, and machine learning.

HAN ZHANG received the B.S. degree in computer science and technology from Jilin University and the Ph.D. degree from Tsinghua University. He is currently with the School of Cyber Science and Technology, Beihang University. His research interests include computer networks, network security, and AI. He has published more than 30 articles in his area.



DUNNAN LIU received the bachelor's, master's, and Ph.D. degrees in electric engineering from Tsinghua University. He is currently with the School of Economics and Management, North China Electric Power University. His research interests include project evaluation, project prediction, and AI. He has published more than 100 articles in his area.