Walmart Sales Prediction Based on Machine Learning

Siming Yi*

Faculty of Science and Engineering, Nottingham University, Ningbo, China

*Corresponding author: ssysy10@nottingham.edu.cn

Abstract. Accurate sales forecasting can improve a company's profitability while minimizing expenditures. The use of machine learning algorithms to predict product sales has become a hot topic for researchers and companies over the past few years. This report features the machine learning sales prediction model that combines the ML algorithm and meticulous feature engineering processing to predict Walmart sales. The following regressions are analyzed in this paper: linear regression, random forest regression, and XGBoost regression. The regression analysis has been tested for the same time period every year for three years from 2010 to 2012 on a continuous time basis. The experiments show that XGBoost algorithm overperforms the other machine learning methods by examining the same evaluation metric (WAME). The findings can contribute to a better understanding of the development of new decision support for the retail industry e.g., Walmart retail stores. Moreover, this paper also represents a detailed procedure to rank the feature importance for the dataset. Within the next few years, the ML algorithm is destined to become an important approach for business forecasting. However, this strategy largely ignores the time series method for accuracy.

Keywords: Machine learning, Feature engineering, Walmart sales.

1. Introduction

1.1. Background

Currently, the volume of data is increasing exponentially. How to make reasonable and effective use of existing business data to support future business decision-making has become the focus of attention for many businesses. Sales projections can increase store earnings, which are crucial to the growth of contemporary and future commerce, by reducing the lack of hot-selling products and the accumulation of unpopular products. Hence, forecasting sales is especially necessary for companies to help reduce costs and increase profits.

1.2. Related Research

The effectiveness of operations in the retail industry and its supply chains can be enhanced by accurate estimates of future retail sales. Therefore, model-based forecasting of sales volume has received the attention of many domestic and international scholars. There are several related research results as follows. Chu and Zhang [1] applied time series approaches and regression approaches for aggregate retail sales forecasting. The results show that neural network model performance is superior to other regression methods and time series methods when dealing with strong seasonal variation datasets. Zhao et al. [2] implemented a visual clustering algorithm, specific regression algorithm, and time series method for electricity sales forecasting for power companies. Thomassey and Fiordaliso [3] predicted textile sales by combining clustering and classification tools. The experiment shows that this combined model is substantially more accurate than other single regression models in making medium-term forecasts. Johannesen and Kolhe [4] forecast the electric load for urban areas by using multiple regression tools. In the case of a large dataset itself, the influence of meteorological parameters is additionally considered. The result is that random forest regression provides a better prediction for the short-term load.

1.3. Objection

Based on the studies mentioned above, it is shown that there is a demand for making relevant forecasts in various fields. Moreover, machine learning methods are highly used in predicting models. The research topic of this paper primarily solves the problems of Walmart sales forecasting that is provided by the Kaggle competition. The rest of this paper is arranged as below: first, the methodology used in this paper will be introduced. In this section, feature engineering and machine learning method are detailly represented. Next, the procedures for the experiment are shown, including the comparison of different models with the same metric. Finally, the conclusion and future work are represented.

2. Methodology

2.1. Source of data

In this part, the dataset will be introduced.

The data is extracted from Kaggle website. There are four files called train.scv, features.scv, train.scv and test.scv. In this project, those documents are merged to create a dataset. It provides historical sales data for 45 Walmart stores which contain 81 departments in total in different regions. The project had access to 33 months of weekly sales between 2010 and 2012. The dataset consists of 20 variables which can be separated into two types. They are the hierarchy variable and exogenous variable. In this dataset, the sale of a week is regarded as the dependent variable.

The other 19 factors (not including weekly sales) may have an influence on the weekly sales to varying degrees. This project will analyze their impact on sales separately. For the hierarchy part, there are department ID, store ID, store type and store size, etc. Others for an external part, temperature, markdowns, fuel price, CPI (the consumer price index), holidays (whether there is a special holiday in that week), and unemployment index are taken into consideration.

2.2. Feature Engineering

In this part, the procedure for feature engineering will be discussed.

As there exist a large amount of data and many variables are in our data, it is necessary to choose the moderate correlation features carefully which could reduce computational costs and obtain more precise prediction results. This paper mainly uses four steps to do with the feature, prepossession for data, data normalization, statistical features and heatmap, and feature selection. First, do some appropriate processing for the data. Then, extra the time characteristics from the timestamp, for example, the day of year and month. The second step is normalizing the data. By limiting the preprocessed data to a specific range, such as [0,1] or [-1,1], normalization aims to eliminate the undesired effects brought on by odd sample data. Next, statistical features in a certain period, like mean, minimum, maximum, standard deviation, etc. Meanwhile, visualizing the correlation between features by drawing a heatmap. The heatmap represents the correlation coefficient among those features, the numerical value of the correlation of the coefficient will be between -1 to + 1 [5]. The strength of the linear correlation between the two variables depends on how close the absolute value of the coefficient is to 1; conversely, the weaker the linear correlation between the two variables depends on how close the absolute value of the coefficient is to 0. Finally, applying ELI5 method to select moderate variables and remove unimportant features. ELI5 is a Python library that aids in explaining and debugging machine learning classifier predictions. ELI5 can be used to display feature importance, interpret predictions from decision trees and tree clusters, and interpret the weights and predictions of linear classifiers, scikit-learn regressors, and linear classifiers [6].

2.3. Models and Algorithms

Marching learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed [7]. The machine learning

algorithms are organized as supervised learning and unsupervised learning. Supervised learning that the various algorithms generate a function that maps inputs to desired outputs is used in this research.

In this study, several machine learning algorithms are applied to the prediction of sales. In this section, these methods are introduced.

2.3.1 Linear Regression

Linear regression is used to build a mathematical linear equation for the relationship between a dependent variable and independent variables [8]. The main idea of this model is to gain a line equation that fits the dataset best. The most appropriate line is when the error of the total prediction for all the data points is as small as possible. The model for a multiple linear regression model that relates a y-variable to (p -1) x-variables is written as Formula (1) below:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad (1)$$

β : coefficients, the slope of the line

ϵ : redundant, the intercept of the line

The subscript i of x refers to which independent variable x it is. β represents the different degrees of the contribution of each variable to the independent variable y. ϵ is the error.

2.3.2 Random Forest Regression

Random forest is an ensemble method based on the decision tree and bagging method. It contains a bunch of single decision trees, but all the trees are mixed randomly instead of separate trees growing individually. For bagging, it is short for bootstrapping aggregation. The principle of the bagging method is to construct multiple decision trees independently by randomly sampling and feature sampling from the data set, and to aggregate the predicted values by averaging the predicted values from each decision tree [9]. The final prediction can be obtained by majority voting:

$$\hat{Y} = \phi_{T,P}(X) = \frac{1}{B}\sum_{b=1}^{B} \phi_{T_b,m}(X) \quad (2)$$

T : the entire training set.

T_b: The subset that was randomly selected from the training set T (b = 1…, B)

2.3.3 XGBoost Regression

Boosting is another ensemble learning strategy[10]. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. GBDT builds a model consisting of multiple decision trees, the same as the random forest. However, in contrast to the bagging employed by random forests, boosting, is taken to start from a single weak model and iterate to generate a collection of strong models. Its final result is obtained by using a weighted average of all decision tree responses, and in each training session, the weights of each decision tree are reassigned considering the previous results, as the training process is continuous.

2.4. Evaluation Meric

To comprehensively evaluate the method selected from machine learning, that method should be compared by the same evaluation metric. In this paper, WMAE metric is employed which is named weighted mean absolute error. This metric is able to express the difference between actual and predicted values, meanwhile, assign different weights to different variables [11]. Formula (3) of WAME is as follows below:

$$WMAE = \frac{1}{\sum w_i}\sum_{i=1}^{n} w_i \lfloor y_i - \hat{y_i} \rfloor \quad (3)$$

where n is the number of rows, $\hat{y_i}$ is the predicted sales, y_i is the actual sales, w_i are weights. (w = 5 if the week is a holiday week, 1 otherwise) The lower the WMAE value, the better the prediction effect of the model.

3. Results and Discussion

In this part, the process of the experiment will be introduced. This section starts with an overall introduction to the dataset used in this paper, which includes descriptive statistics, then followed by pre-processing of the dataset. The third step is the selection of features. Finally, the machine learning models are applied to the dataset and compared with the same metric to select the best model.

3.1. Statistic features

Fig. 1 shows the overview of weekly sales throughout the whole year. Basically, the curve fluctuation characteristics are roughly the same from year to year. This figure interprets that most weeks have sales around the median with a plunge around week 48 and a recovery for the holidays. It is worth noticing that by week 47 (the week of Thanksgiving) and week 51 (one week before Christmas) the sales rise up by a huge margin.

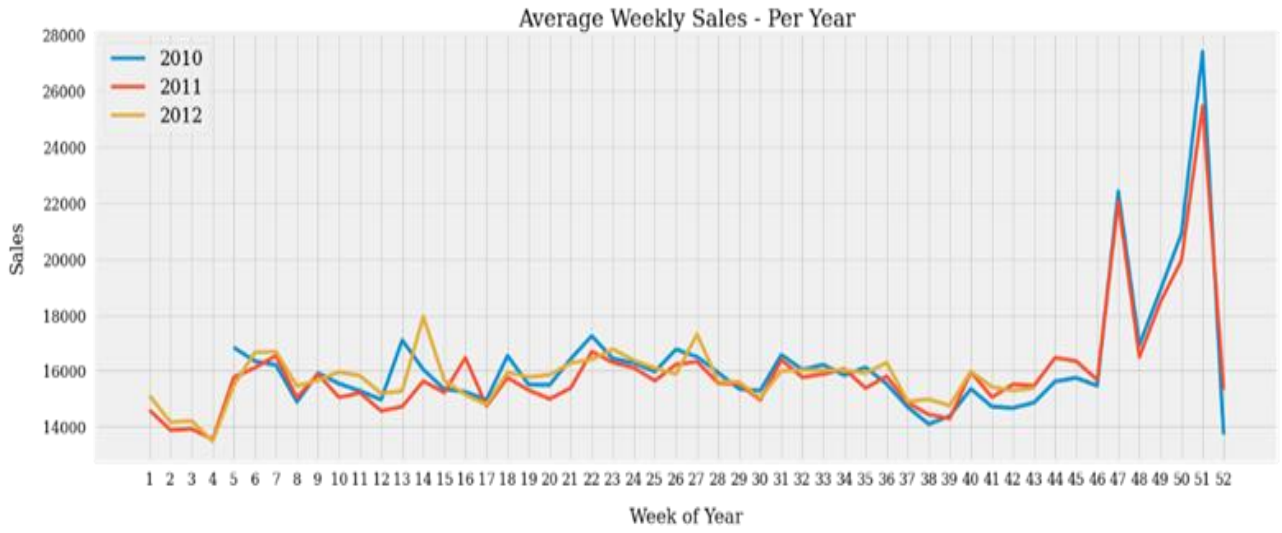

Fig. 1 average weekly sale- per year (Photo credit: Original)

Table 1 represents the stats of important numerical features in the dataset. For weekly sales, the value of the standard deviation is exceptionally large, which indicates that it has a large degree of dispersion and there is an amount of variation. The reason for the appearance of this phenomenon should be the significant increase in sales during the holidays, which is the source of the amount of variation. This can also be seen from the span between its value of 75 percent and the maximum value. Hence, further analysis for the weeks of holidays is required.

Table 1. statistics of important features

	count	mean	std	min	25%	50%	75%	max
Store	421570	22.200546	12.785297	1	11	22	33	45
Dept	421570	44.260317	30.492054	1	18	37	74	99
Weekly_Sales	421570	15981.25812	22711.18352	-4988.94	2079.65	7612.03	20205.8525	693099.36
Temperature	421570	60.090059	18.447931	-2.06	46.68	62.09	74.28	100.14
Fuel_Price	421570	3.361027	0.458515	2.472	2.933	3.452	3.738	4.468
MarkDown1	150681	7246.420196	8291.221345	0.27	2240.27	5347.45	9210.9	88646.76
MarkDown2	111248	3334.628621	9475.357325	-265.76	41.6	192	1926.94	104519.54
MarkDown3	137091	1439.421384	9623.07829	-29.1	5.08	24.6	103.99	141630.61
MarkDown4	134967	3383.168256	6292.384031	0.22	504.22	1481.31	3595.04	67474.85
MarkDown5	151432	4628.975079	5962.887455	135.16	1878.44	3359.45	5563.8	108519.28
CPI	421570	171.201947	39.159276	126.064	132.022667	182.31878	212.416993	227.232807
Unemployment	421570	7.960289	1.863296	3.879	6.891	7.866	8.572	14.313
Size	421570	136727.9157	60980.58333	34875	93638	140167	202505	219622

3.2. Data Processing

3.2.1 Missing Value

The markdown columns were given by Walmart to see the effect of markdowns on sales. In fact, there are many NaN values for markdowns, more than 250000 in each markdown column (i.e., there are many missing values in markdowns). The solution for them is to fill the missing value with 0. It shows that there is no markdown for that week.

3.2.2 Encoding Categorical Data

The type of Is_holiday is bool and contains true and false which define into numerical values as 1 and 0. Meanwhile, replace the type of stores from A, B, and C with 1, 2, and 3.

3.2.3 Data Normalization

The normalization method used in this paper is the min-max normalization. This method restricts the unnormalized data to a specified upper and lower bound by linear operations. [12] The linear function converts the original data linearization method to the range of [0, 1], and the calculated result is considered normalized data. The linear formula (4) is shown below:

$$x'_{i,n} = \frac{x_{i,n} - min(x_i)}{\max(x_i) - min(x_i)}(nMax - nMin) + nMin \quad (4)$$

where max and min represent the maximum and minimum values of i-th feature separately.

3.3. Feature Selection

3.3.1 Multifactor Analysis

Fig. 2 which is the correlation heatmap for these features, shows the relationships between numerical variables as a visual. One noticed that among those features, Department, Store size, and Store type have a moderate correlation with weekly sales. Only Store type is negatively correlated with sales, the rest are positively correlated. Markdown1-5 has a very weak correlation with the weekly sales, so these columns may be left out. Meanwhile, Temperature, Fuel price, CPI, and Unemployment are very weakly correlated with weekly sales.

As holiday weeks have higher sales than non-holiday weeks, Is Holiday will be taken into account in the research to come. For the holidays, this dataset only recorded four important holidays during the year. There are Thanksgiving, Christmas, Superbowl, and Labor Day. What this paper has done is that extract the days of the week containing each of these holidays and name them as new variables for the next step.

In addition to the correlation with sales, the relationship between individual variables is also worth exploring and can filter out more variables. Two variables Month and Day will be left out as this information is already contained in the WeekOfYear. Fuel price and year have a strong correlation. In this situation, they would each carry data that is comparable to the model, thus one of them must be eliminated. Also, Markdowns 4 and 5 highly correlated with Markdown 1. They can be dropped when analyzed. For further research, these other variables with weak correlation to weekly sales can be investigated to determine whether they are relevant.

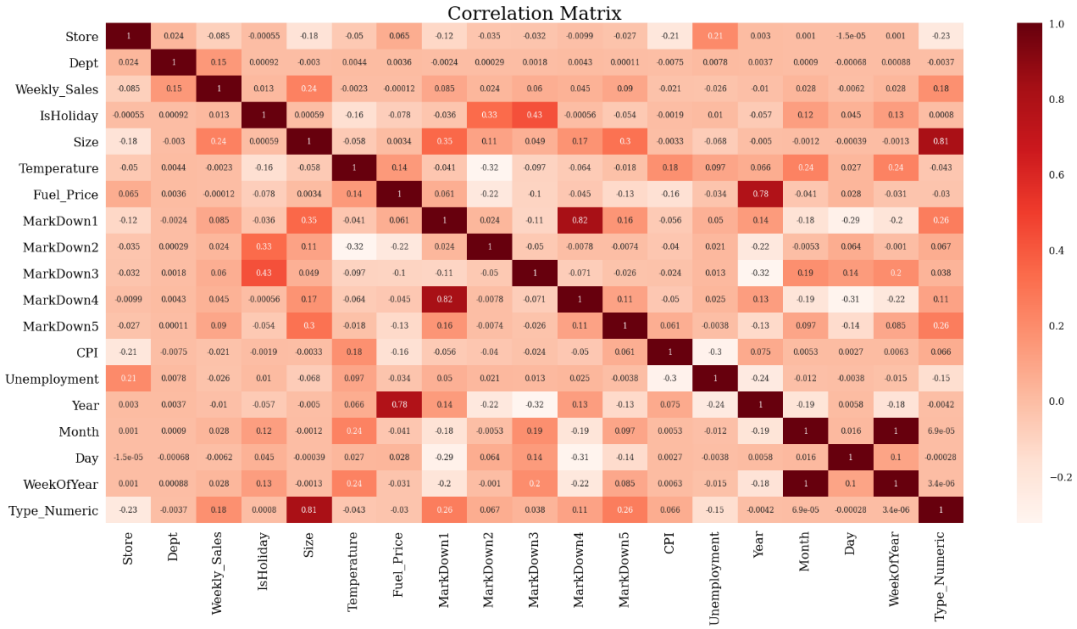

Fig. 2 Heatmap (Photo credit: Original)

3.3.2 Single factor analysis

To identify the degree of influence of each independent variable on the dependent variable, univariate analysis was conducted in this experiment. By applying the ELI5 method, the weights of features are represented in Table 2. For holidays, in order to confirm the effect of each of the four holidays on sales, the experiment singles out the weeks belonging to different holidays to form new variables. What stands out in this table is that it seems to be Dept, Store, Size, CPI, and Week, are the top 5 features.

Table 2. The weight of features

	Weight	Feature
0	1.6476 ± 0.0457	Dept
1	0.4512 ± 0.0220	Size
2	0.1091 ± 0.0068	Store
3	0.0494 ± 0.0044	CPI
4	0.0310 ± 0.0068	Week
5	0.0107 ± 0.0203	Tranksgiving
6	0.0088 ± 0.0004	Type
7	0.0080 ± 0.0017	Unemployment
8	0.0075 ± 0.0055	MarkDown3
9	0.0071 ± 0.0011	Temperature
10	0.0064 ± 0.0027	Day
11	0.0018 ± 0.0019	Fuel_Price
12	0.0018 ± 0.0010	IsHoliday
13	0.0008 ± 0.0003	MarkDown5
14	0.0004 ± 0.0007	MarkDown4
15	0.0003 ± 0.0001	MarkDown1
16	0.0003 ± 0.0002	Month
17	0.0002 ± 0.0001	Year
18	0.0002 ± 0.0001	MarkDown2
19	0.0002 ± 0.0001	LaborDay
20	0.0002 ± 0.0000	MarkdownsSum
21	0.0000 ± 0.0000	SuperBowlWeek
22	0 ± 0.0000	Christmas

3.4. Result

For the evaluation of machine learning model performance, around 80 percent of the dataset is treated as the training set, and the rest of them are the testing set. For the result of prediction, the best approach is Xgboost Regression. The detailed prediction for Xgboost is shown in Fig. 3. The WAME for each regression model is represented in Table 3. Normal denotes the result without feature engineering. It can be seen that there is a slight reduction in the error after feature selection.

Table 3. The values of WAME for each model

	normal	feature engineering
linear	0.031499	0.031425
xgboost	0.016339	0.016638
random forest	0.020919	0.019612

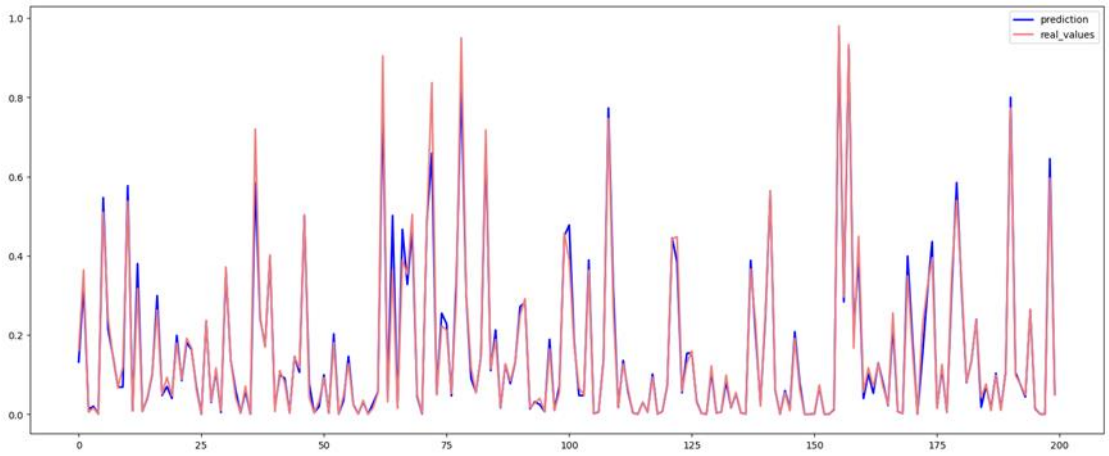

Fig. 3 The prediction result for xgboost model (Photo credit: Original)

3.5. Limitations

This paper only takes into consideration of simple machine-learning models. Some studies found that complex machine learning (ML) methods could outperform simple models. Therefore, in the future, it is better to try to use some complex method to do the prediction.

Also, it may be better to ensemble or do some cross-learning with several ML methods to get a more precise prediction. Alternatively, some time series approaches such as ARIMA can be applied in combination with machine learning.

The author also considers that the outcomes could be enhanced by incorporating other predictive variables. As an example, the impact of more external factors on sales can be taken into consideration, such as information for the location of the mall, whether it is in a place with a high or low population flow, or whether there are other competitors present in the market. The benefits generated by real-time push advertising on the Internet, for example, can also help to observe sales trends over a recent period of time and improve sales strategies based on that. Text mining techniques can be used for this purpose, boosting the management team's existing work.

Furthermore, the methods applied in this paper underestimate uncertainty in prediction. The accurate estimation of uncertainty should be taken into consideration.

Finally, if the model is applied to the current sales forecast is not accurate enough, because now under the influence of the covid-19, large amounts of stores are temporarily closed or even directly closed, which seriously affects the situation of offline sales. Therefore, if the model is to be applied, at this time, proper consideration of online sales should be added to this model.

4. Conclusions

In this paper, the sales of Walmart retail outlets are examined using regression tools, including linear regression, random forest regressor, and XGBoost regressor. Using the dataset provided by Walmart in Kaggle. The experiment shows that the xgboost regression provides higher performance and accuracy compared to other regression models.

The variables involved in this experiment are very limited and should include more external factors such as the location of the supermarket, and more internal factors such as the sales of each product for future study.

The author also believes that it would be beneficial to evaluate the performance of the same strategies in the same situation but with historical data. In circumstances when time series data is available, such as in the case of continuity products, it seems relevant to examine the potential of these models.

References

[1] Chu, C. W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of production economics, 86(3), 217-231.

[2] Zhao, J., Tang, W., Fang, X., Wang, J., Liu, J., Ouyang, H., ... & Qiang, J. (2015, October). A novel electricity sales forecasting method based on clustering regression and time-series analysis. In Proceedings of the 2015 International Conference on Artificial Intelligence and Software Engineering.

[3] Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems, 42(1), 408-421.

[4] Johannesen, N. J., Kolhe, M., & Goodwin, M. (2019). Relative evaluation of regression tools for urban area electrical energy demand forecasting. Journal of cleaner production, 218, 555-564.

[5] Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. Critical reviews in analytical chemistry, 36(1), 41-59.

[6] Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access, 8, 187814-187823.

[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.

[9] Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., & Lei, G. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression. Journal of Petroleum Science and Engineering, 196, 107801.

[10] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2), 139-157.

[11] Cleger-Tamayo, S., Fernández-Luna, J. M., & Huete, J. F. (2012, September). On the Use of Weighted Mean Absolute Error in Recommender Systems. In RUE@ RecSys (pp. 24-26).

[12] Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524.