Sales Prediction of Walmart Sales Based on OLS, Random Forest, and XGBoost Models

Tian Yang *

School of Science and Engineering, The Chinese University of Hong Kong Shenzhen, Shenzhen, China

* Corresponding Author Email: 120090082@link.cuhk.edu.cn

Abstract. The technique of estimating future sales levels for a good or service is known as sales forecasting. The corresponding forecasting methods range from initially qualitative analysis to later time series methods, regression analysis and econometric models, as well as machine learning methods that have emerged in recent decades. This paper compares the different performances of OLS, Random Forest and XGBoost machine learning models in predicting the sales of Walmart stores. According to the analysis, XGBoost model has the best sales forecasting ability. In the case of logarithmic sales, R2 of the XGBoost model is as high as 0.984, while MSE and MAE are only 0.065 and 0.124, respectively. The XGBoost model is therefore an option when making sales forecasts. These results compare different types of models, find out the best prediction model, and provide suggestions for future prediction model selection.

Keywords: Sales prediction; Random Forest; XGBoost; OLS; Machine learning.

1. Introduction

The technique of estimating future sales levels for a good or service is known as sales forecasting. Companies can lessen the issue of overproduction by using accurate sales forecasts [1, 2]. In addition, it is important for retail companies to record and maintain sales-related data, e.g., the number of units sold and revenue generated relative to time. Then, one can= use the sales forecast method to reasonably plan future scenarios [3]. Sales forecasting methods are evolving in accordance with the changing times. Initially, qualitative methods were utilized, followed by time series methods, regression analysis, and econometric models. With the advent of big data and enhanced computing power, machine learning methods have emerged in recent decades. It is imperative to note that each method has a distinct impact.

There are numerous studies that utilize diverse techniques and models for sales predictions. To illustrate, one article accomplished a sales volume prediction of Amazon via employment of ARIMA model based on time series methodology [4]. Nevertheless, the issue with such models lies in their failure to encompass all the pertinent factors that impact sales, e.g., population and interest rates. In addition, OLS model was used to estimate the decisive factors affecting the life of houses and was compared with various machine learning models [5]. However, this method may require a large sample size and is also susceptible to outliers. Additionally, in the field of machine learning, the LightGBM model is employed to forecast air ticket sales with a high degree of consistency [6]. Moreover, another study employs a recurrent neural network model (RNN) to anticipate the sales of a medium-sized restaurant, with the aim of optimizing the staffing requirements of the restaurant [7].

Regarding the sales data utilized in this study from Walmart, some articles have employed machine learning techniques to predict sales. The XGBoost model, for instance, is used to forecast the weekly sales of Walmart retail locations. It has been demonstrated that the XGBoost model predicts more accurately than the Ridge regression and logistic regression [8]. Additionally, some scholars argue that employ the LightGBM model and neural network model for sales data of Walmart retail stores, with negligible forecasting errors in all of their results [9, 10]. However, previous research on retail sales data from Walmart stores has not fully considered the effects of numerical skewness and extreme individual values. Therefore, it is necessary to implement special processing techniques for retail data and observe the predictive abilities of each model after the data has been processed.

The aim of this paper is to contrast the forecasting outcomes of diverse machine learning models applied to an identical dataset and make some suggestions about the model selection based on the models' performance. The remaining part of the paper proceeds as follows. Part two will first introduce the data source and various variables and feature engineering, followed by an explanation of the model and parameters selected for this study. Finally, it will clarify how the model was evaluated and present some evaluation metrics. Part three will begin with a correlation analysis of the data, followed by a description of variable outcomes and an overview of the model's performance. Finally, some recommendations for selecting models will be provided. Part four will address the limitations of this study and prospects for future research. Part five will summarize the entire study.

2. Data & Method

2.1. Data

The data used in this study was obtained from the publicly available dataset "Walmart Recruiting - Store Sales" on the Kaggle website. After processing the dataset, a data frame was obtained, which includes a total of 16 variables. Each variable has 421570 data points. The dependent variable in this study is Weekly_Sales, which represents the weekly sales of each store. The first independent variable is Store, which represents the store number. There are a total of 45 stores, so the maximum value for Store is 45. The second independent variable is Dept, which represents the department number of the corresponding store. The third independent variable is Date, which represents the date of the store, with a weekly interval for each date. The next variable is IsHoliday, which indicates whether a certain week is a holiday for the store. The next variable, Type, represents the type of store. Then, Size represents the size of the store. Temperature represents the average temperature of the store within a week. Fuel_Price represents the average fuel price for the week. Markdown1-5 represents the total amount of markdowns in the store during holidays. CPI represents the consumer price index. Unemployment represents the unemployment rate in the area where the store is located.

2.2. Feature Engineering

The first issue to address is the missing values in the data. There are over 1 million missing values in Markdown1 - Markdown5 variables. One simple approach is to delete all these missing values. However, this approach is not always recommended. In this dataset, it is appropriate to replace all missing values with 0. This indicates that there were no holiday promotions or markdowns occurring in the store on the corresponding dates, and therefore the total amount of markdowns is 0. The next issue to address is the outlier values in the data. Regarding the "Weekly_Sales" variable, it is not feasible for it to have negative values. As observed from the graph, only a minor fraction of data falls below 0. Therefore, data points below 0 must be eliminated, and the lowest possible value should be set to 1. It should also be noted that the data includes a few extremely high values and has a right-skewed distribution. To prepare the data for a regression model, it is essential to normalize the data distribution as much as possible. Therefore, a simple way to do this is to apply a logarithmic transformation to the "Weekly_Sales" variable. In addition, the preprocessed dataset is partitioned into a training set and a test set at a ratio of 3:1.

2.3. Models

This paper aims to employ the OLS regression model, random forest model, and XGBoost model to fit and predict sales data. It is worth noting that the primary objective of this study is to compare the performance differences between the models, rather than exploring their specific details. The fundamental concept of the OLS model is to determine parameter estimates that minimize the sum of squares of the difference between the actual value and the model estimate. In other words, the objective of the OLS model is to discover the optimal function that fits the data by minimizing the sum of squares of the errors. In this paper, the dependent variables in the dataset were subjected to logarithmic transformation to satisfy the assumptions of the classical linear model as much as possible,

rendering the OLS model appropriate for analysis. The random forest algorithm generates a collection of decision trees. When making predictions for a given sample, each tree in the forest generates its own prediction. The final prediction for the sample is then determined by aggregating the predictions from all the trees in the forest, often through a majority voting method [11]. The relevant parameters are shown in Table 1. XGBoost is a popular gradient boosting algorithm for regression and classification tasks. It operates by iteratively training weak models, typically decision trees, and improving the predictions by minimizing the loss function. To prevent overfitting and enhance the generalization ability of the model, the algorithm incorporates regularization techniques [12]. The relevant parameters are shown in Table 2.

Table 1. Parameter Values of Random Forest Model

Parameter	Value
n_estimators	20
max_depth	None
min_samples_split	2
min_samples_leaf	1
max_features	‘auto’
bootstrap	True
oob_score	False

Table 2. Parameter Values of XGBoost Model

Parameter	Value
learning_rate	0.1
max_depth	10
n_estimators	200
objective	'reg:squarederror'
booster	gbtree
reg_alpha	0
reg_lambda	1

2.4. Evaluations

This paper employs three indices to evaluate the model performance, namely the mean square error (MSE), the mean absolute error (MAE), and the coefficient of determination (R2). Their formulae are as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{1}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|(y_i - \hat{y}_i)| \tag{2}$$

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(\bar{y}_i - y_i)^2} \tag{3}$$

$$R^{2'} = 1 - \frac{(1-R^2)(n-1)}{n-P-1} \tag{4}$$

Each of these three metrics serves a distinct purpose. It should be noted that in this paper, the adjusted R-squared is used for the OLS model. P represents the number of independent variables and n represents the number of data points in the dataset. The coefficient of determination evaluates the degree of model fit. MAE provides a straightforward measure of the discrepancy between predicted and actual values. MSE places a greater emphasis on significant deviations by penalizing larger errors more severely. Thus, when the predicted values deviate significantly from the actual values, MSE is a more effective metric to evaluate model performance.

3. Results & Discussion

3.1. Correlation Analysis

Based on Pearson correlation calcualtions, the correlation between the dependent variable and other independent variables is no more than 0.24. This means that the dependent variable has a weak correlation with the independent variable. Furthermore, some of the independent variables are highly correlated with each other. For example, the correlation between Markdown1 and Markdown4 is as high as 0.84. This high correlation can give rise to problems such as multicollinearity. Therefore, this data set is completely inapplicable to the OLS model before the logarithmic transformation. In other words, in order to use the OLS model to fit combined forecast data in this paper, logarithmic transformation must be performed on the dependent variable. The fitting effect of the model will be presented in Section 3.2.

3.2. Statistic Descriptions & Model Performances

Combining Fig. 1 and Table 3, R2 of the scatter plot of the OLS model is 0.725, which indicates y_real(Real sales data in the data set) has a general correlation with y_predictied(Sales data predicted by the model). R2 of the scatter plot of the Random Forest model is 0.978, which indicates y_real has a strong correlation with y_predictied. R2 of the scatter plot of the XGBoost model is 0.984, which also indicates y_real has a strong correlation with y_predictied. Table 3 shows that the OLS model has significantly higher MSE and MAE values compared to the other two models（Considering the logarithmic transformation of the dependent variable, the difference is quite large）. Additionally, the R-squared value of the OLS model is markedly lower than the other two models. In comparison with the Random Forest model, the XGBoost model has slightly lower MSE and MAE values, while its R-squared value is slightly higher.

Table 3. Regression fitting.

	MSE	MAE	R2
OLS	1.1135	0.6728	0.7257
Random Forest	0.0887	0.1432	0.9781
XGBoost	0.0655	0.1246	0.9838

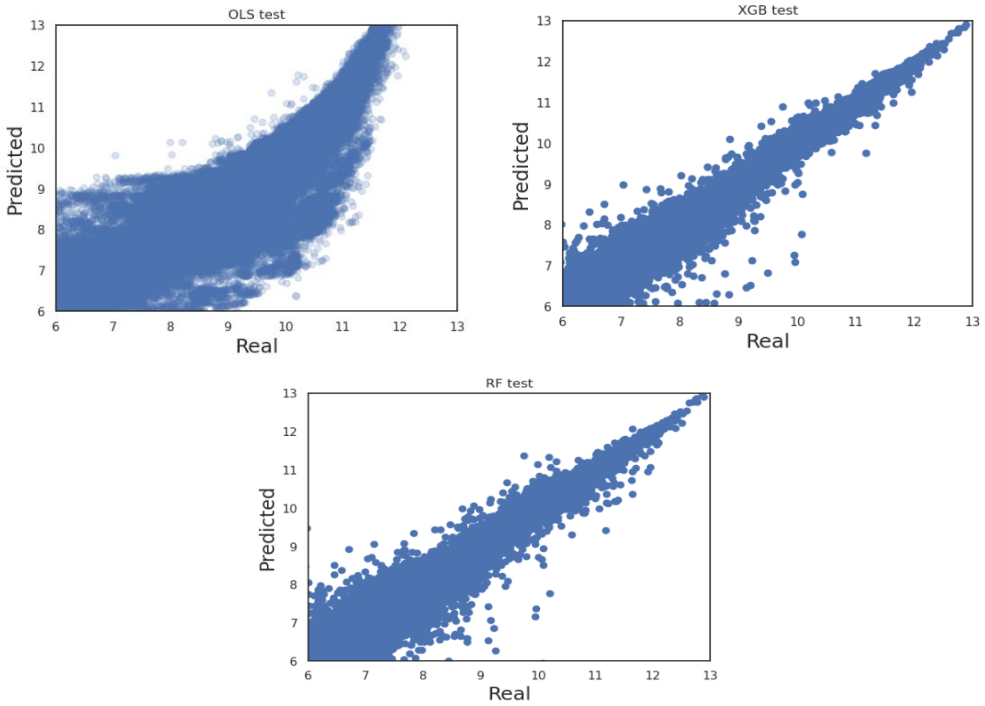

Fig 1. Scatter plots of predicted and actual values of different models

Based on the aforementioned data, the XGBoost model exhibits the best predictive performance, while the Random Forest model ranks second. However, the difference in performance between the XGBoost and Random Forest models is quite small. The OLS model performed the worst. The reason why the MSE of OLS model is larger than MAE may be that the predicted and actual errors on some values are too large. This result is also consistent with the actual situation on the scatter plot. The OLS model scatter plot basically doesn't match y_real=y_predictied, which means the model is quite different from the actual one. Conversely, the other two models exhibit a strong correlation, and the number of outliers on the scatter plot is minimal.

3.3. Suggestions

According to the analysis, if the correlation between the initial dependent variable and independent variable is weak, it is not advisable to utilize the OLS linear regression model. This is due to the possibility of obtaining poor results even after performing extensive data processing. Moreover, while the XGBoost model typically outperforms the Random Forest model, the latter has the benefit of easier tuning. Therefore, the selection of a suitable model must be based on a comprehensive consideration of the data characteristics and individual circumstances.

4. Limitations & Prospects

There are still some limitations. First of all, logarithmic transformation was performed on the dependent variable in the final result, which made it difficult to directly obtain the original data and required exponential processing. In addition, this makes it difficult to measure coefficients in raw units. Secondly, this study employed only one set of parameters for each of the Random Forest and XGBoost models, which means that the results may be subject to contingency. For further studies, researchers may seek to strike a balance between modifying the original data and facilitating the measurement of coefficients in their original units. Furthermore, when comparing models that require tuning, such as XGBoost or Random Forest, it may be worthwhile to experiment with several additional sets of parameters to obtain more accurate results. Finally, it may be beneficial to incorporate the neural network method in the scope of model comparison to obtain the most reasonable prediction model.

5. Conclusion

In summary, this paper compares the predictive capabilities of the OLS, Random Forest, and XGBoost models in forecasting Walmart sales. After conducting various operations (e.g., logarithmic transformation and parametric modeling), the final result shows that the XGBoost model performs the best among the three evaluation indexes, followed by the Random Forest model and the OLS model. However, there are limitations to the study, such as the difficulty in measuring the original unit coefficient due to the modification of the original data and the use of only one set of parameters for the XGBoost and Random Forest models, which may result in random outcomes. Future research could explore ways to balance modifying the original data while maintaining the ability to measure the original unit coefficient, use multiple parameter sets to fine-tune the models, and incorporate neural network models to achieve the best prediction accuracy. The significance of this study lies in its comparison of different model types to identify the best prediction model and provide model selection advice for future sales forecasting.

References

[1] Islam S, Amin S H. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. Journal of Big Data, 2020, 7: 1-22.

[2] Nguyen H D, Tran K P, Thomassey S, Hamad M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 2021, 57: 102282.

[3] Khan R A, Quadri S M. Business intelligence: an integrated approach. Business Intelligence Journal, 2012, 5(1): 64-70.

[4] Singh B, Kumar P, Sharma N, Sharma K P. Sales forecast for amazon sales with time series modeling. In 2020 first international conference on power, control and computing technologies (ICPC2T) 2020: 38-43.

[5] Kim E M, Kim S B, Cho E S. Using mechanical learning analysis of determinants of housing sales and establishment of forecasting model. Journal of Cadastre & Land InformatiX, 2020, 50(1): 181-200.

[6] Tang X, Gao S, Jiang Z. A Blending Model Combined DNN and LightGBM for Forecasting the Sales of Airline Tickets. In Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence 2019: 150-154.

[7] Schmidt A, Kabir M W U, Hoque M T. Machine learning based restaurant sales forecasting. Machine Learning and Knowledge Extraction, 2022, 4(1): 105-130.

[8] Niu Y. Walmart Sales Forecasting using XGBoost algorithm and Feature engineering. In 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2020: 458-461.

[9] Qiao Z. Walmart Sale Forecasting Model Based on LightGBM. In 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) 2020: 76-79.

[10] Chen J, Koju W, Xu S, Liu Z. Sales forecasting using deep neural network and SHAP techniques. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) 2021: 135-138.

[11] Biau G, Scornet E. A random forest guided tour. Test, 2016, 25: 197-227.

[12] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016: 785-794.