Walmart Sales Prediction Based on Decision Tree, Random Forest, and K Neighbors Regressor

Bo Yao *

Department of Environment, Education and Development, the University of Manchester, Manchester, United Kingdom

* Corresponding author: bo.yao-3@postgrad.manchester.ac.uk

Abstract. Sales forecasting is a very important research direction in the business and academic fields, and sales forecasting methods are also in full bloom, such as time series model, machine learning model and deep neural network model. This paper will use three machine learning models: Decision Tree Regressor, Random Forest Regressor, and K Neighbors Regressor to predict Walmart Recruiting - Store Sales data. Using correlation, mean absolute error, and mean square error to evaluate the prediction results of these three models, it is found that the prediction effect of Random Forest Registrar performs the best of these three models. The R^2 value between the predicted sales volume of Random Forest Regressor and the sales volume of the test set is 0.937, the average absolute error is 1937.810, and the mean square error is 32993323.634. Therefore, Walmart can use Random Forest Regressor when forecasting the weekly sales of its own stores. At the same time, this paper provides a good model reference value (especially Random Forest Regressor) for other industries when researching the sales forecast, as well as methods for evaluating different model predictions. Overall, these results shed light on guiding further exploration of Sales forecasts for supermarkets.

Keywords: Sales prediction, Machine learning, Random Forest Regressor.

1. Introduction

Sales forecasting is a very important research direction in the business and academic fields, and sales forecasting methods are also in full bloom, such as the time series model, machine learning model and artificial neural network model. In the research direction of sales forecasting, the time series model starts early, but the time series model only focuses on the sales volume itself, and does not consider other relevant variables in the same period. For example, ARIMA model is used to forecast the sales of goods to meet the turnover demand of sellers for the inventory of goods sold and improve the turnover rate of inventory, so as to reduce the problem of capital turnover caused by overstock of inventory as much as possible [1].

In addition to ARIMA model, the Exponential Smoothing model and Holt's Linear model are used to forecast the sales of different new products [2], so as to maintain the problem of overstock of inventory caused by too many new products and the problem of sales reduction caused by too few products. With the continuous development of computer technology, machine learning models and artificial neural network models that require a lot of computation have entered people's vision, which plays a great role in promoting the research of sales forecasting. For example, random forest regression, regression tree and XGboost model are used to forecast the sales of VLISCO Group's distribution subsidiaries, and more accurately purchase appropriate products from suppliers [3]. Besides, the LEAST SQUARE SUPPORT VECTOR MACHINE model was used to predict the sales of network products, and good similarity calculation was achieved [4]. In the field of artificial neural network, BP artificial neural network is used to forecast the sales data of the company, and highly consistent prediction results are obtained [5].

For the Walmart Recruiting (i.e., Store Sales data studied in this article), many articles have carried out relevant research on this data, not only limited to the prediction of sales data. Through simple visual data analysis, consumers' consumption behavior can be understood to evaluate sales data [6]. When it comes to the sales forecast in terms of uncertain competition, the M5 forecast uncertainty competition solution is proposed. The combination of multi-level state space model and Monte Carlo

simulation is used to generate forecast scenarios and use the negative binomial distribution to model the observed sales [7]. There are also research papers that evaluate the actual situation of retailers around the country and help retailers understand their local market sales to predict [8-10].

The purpose of this article is to compare the prediction results of the models, find out which model is the most accurate and close to the sales data, and give some suggestions on sales prediction of Walmart retailers according to the prediction results. This paper will use four parts for analysis and elaboration. In the second section, the paper will first introduce the source of data and the basic information of each variable in the data, then the article will introduce the relevant data cleaning, feature engineering and results used in the data, followed by a description of the model principles used in this paper, and finally the article will introduce how the research process of this paper is carried out. In the third Section, the article will describe and explain the results of fitting and prediction in Python, and give a specific introduction to the research results. In the fourth Section, the limitations of this article will be introduced, and the future prospects will be given according to the existing limitations. In the fifth Section, the article will give some suggestions on the sales forecast of retailers.

2. Data & Method

2.1. Data

The data used in this paper is Walmart Recruiting - Store Sales data, which is sourced from the Kaggle website. Since there are three dataset files in the website, the data frames of the three files are aggregated according to the three variable fields of store, date and is holiday, and finally aggregated into a data frame. In this data frame, there are 16 variables in total.

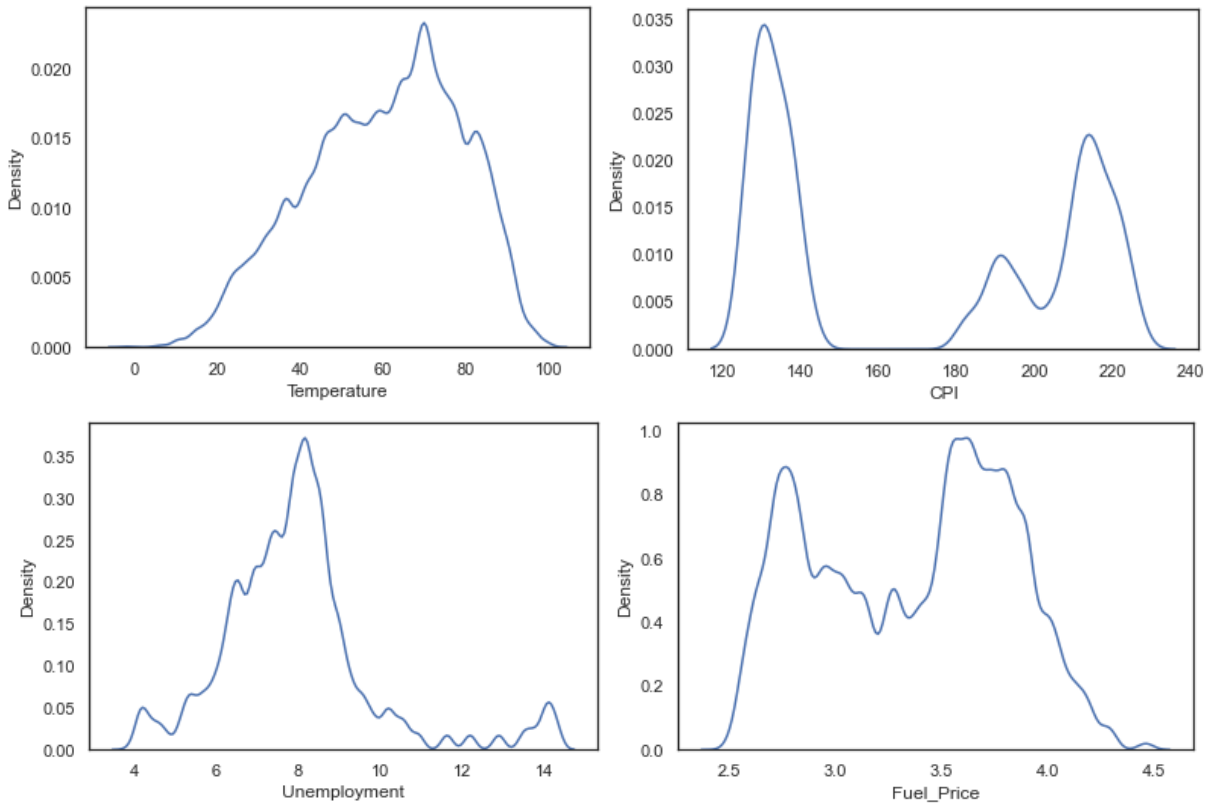

Figure 1. Density curves for four variables

The first one is store. This variable represents the number of the corresponding store and there are 45 stores in total; Date refers to the year, month, day and date of recorded data, and the interval between each date is one week; Temperature represents the average temperature of the area where the corresponding store is located in a week; Fuelprice indicates the oil price of the week in the

corresponding store area; Markdown1-markdown5 indicates that Walmart has five anonymous data hidden by Wal Mart; CPI refers to the consumer price index of the week; Unemployment refers to the unemployment rate of the week; Is today is a binary variable with the value fields of true and false. True indicates that there is a holiday in this week, and false indicates otherwise; Dept represents the department number of the corresponding store; Weekly sales represents the actual sales of the week; Type represents the store type of the corresponding store, including A, B and C stores; Size represents the floor area of the corresponding store.

First, the missing values of the data are counted. It is found that the variables markdown1-markdown5 have a large number of missing values, while other variables do not have missing values. Therefore, the markdown1-markdown5 variables need to be deleted before model fitting, and only other variables need to be analyzed. Then, for several continuous variables: Temperature, Fuel_ Price, CPI and Unemployment draw density curves to observe the main distribution of their variables, as shown in Fig. 1. It can be found that the numerical distribution of temperature variable and Unemployment is somewhat close to the shape of normal distribution, but temperature variable shows a left-biased shape while Unemployment shows a right biased shape, while CPI and Fuel_ Price presents a groove-shaped distribution of multiple peaks.

2.2. Feature engineering

For the Walmart sales dataset after data aggregation and cleaning, this paper intends to use correlation analysis to conduct a feature analysis on this dataset. The purpose of correlation analysis is to delete one of the two highly correlated variables to reduce the complexity of model fitting. In general, two variables are considered to be low correlated if the absolute value of the correlation coefficient between them is less than 0.3, and highly correlated if the absolute value of the correlation coefficient between them is greater than 0.6, and one of the variables should be deleted before model fitting.

2.3. Models & Metrics

In the case that the linear relationship between the variables is not strong, this paper intends to use three models: regression tree, random forest regression and K nearest neighbor to fit and forecast the sales. Dividing the data collection into many small segments of data. The regression tree first divides the data set into many pieces of data that are conducive to establishing linear regression, next use our linear regression technology to model. If it is still hard to fit a linear model after the first segmentation, the regression tree will continue to segment the data until it can fit a linear regression model, and the regression uses the minimum mean square deviation to determine the optimal division of the regression tree, The partition criterion is to minimize the error variance of the subtree after the expected partition. The formula is as follows, where y represents the real value in the training set and represents the predicted value of the regression tree.

$$\mathrm{MSE} = \sum(\hat{y}_i - y_i)^2 \tag{1}$$

Random forest regression is an algorithm derived from the regression tree. It uses the calculation method of averaging multiple regression trees to fit the prediction. In Python, 'n_ Estimators' is used to control the number of selected regression trees and it can improve the number of regression trees and the fitting accuracy in the training set. However, with the increase of the number, the improvement of fitting accuracy will tend to be saturated. Increasing the number of regression trees again will not improve the fitting accuracy. At the same time, overfitting may occur in the test set.

The K nearest neighbor regression model does not need training parameters, but only needs to use the target values of the K nearest training samples around to make decisions based on the regression values of the samples to be tested. From this, different ways to measure the regression value of the sample to be tested are derived, namely, the ordinary arithmetic average algorithm and the weighted average considering the distance difference. After introducing the model, there are three indicators

will be used to evaluate the model in this paper: the correlation coefficient between the predicted value series and the actual value series, mean absolute error and mean squared error. The corresponding formula is as follows:

$$R^2 = \frac{\sum \hat{y}_i y_i - \sum \hat{y}_i \sum y_i}{\sqrt{\sum (\hat{y}_i - \bar{\hat{y}}_i)^2} \sqrt{\sum (y_i - \bar{y}_i)^2}} \tag{2}$$

$$\mathrm{MAE} = \sum \mid \hat{y}_i - y_i \mid \tag{3}$$

The three evaluation indicators have their own characteristics. The correlation coefficient can judge the similarity between the prediction series and the actual series, and the average absolute error can see how much the actual deviation is. The mean square error can reduce the error value of a prediction point that is close to the actual value, but it will also enlarge the error value of a prediction point that is far from the actual value.

2.4. Procedure

In this paper, first, the value range of the Type variable will be changed from (A, B, C) to an integer variable (0, 1, 2) to meet the requirements of the integer or floating-point type required by the model. Second, calling the sklearn third-party library in Python. First, divide the cleaned data set into the training set and test set in a 4:1 ratio. After the division, the variables in the training set data are used to model the dependent variable weekly sales. After fitting the three models, use the independent variables of the test set to predict the weekly sales and compare them with the actual weekly sales. Among them, the three indicators mentioned above are used to evaluate the quality of the model fitting: the correlation coefficient between the predicted value series and the actual value series, mean absolute error and mean squared error.

3. Results & Discussion

First, the correlation analysis is conducted for each variable of the dataset, and the results of the correlation analysis are shown in Table. 1. Seen from the results, one can notice the maximum absolute value of the maximum correlation coefficient between the two variables does not exceed 0.3, the maximum value is 0.2999 and most of the figure is less than 0.01. It can be considered that the linear relationship between the variables is not significant, so it is not necessary to conduct another feature screening because of the high correlation between variables when using variables. At the same time, the linear relationship between each variable and the weekly sales is also weak. Because of this, the machine learning model will not be directly selected to establish a linear relationship model, such as the generalized linear regression model and the SVM model.

Table 1. Correlation analysis

	Store	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday	Dept	Weekly_Sales	Size
Store	100%	-5%	7%	-21%	21%	0%	2%	-9%	-18%
Temperature	-5%	100%	14%	18%	10%	-16%	0%	0%	-6%
Fuel_Price	7%	14%	100%	-16%	-3%	-8%	0%	0%	0%
CPI	-21%	18%	-16%	100%	-30%	0%	-1%	-2%	0%
Unemployment	21%	10%	-3%	-30%	100%	1%	1%	-3%	-7%
IsHoliday	0%	-16%	-8%	0%	1%	100%	0%	1%	0%
Dept	2%	0%	0%	-1%	1%	0%	100%	15%	0%
Weekly_Sales	-9%	0%	0%	-2%	-3%	1%	15%	100%	24%
Size	-18%	-6%	0%	0%	-7%	0%	0%	24%	100%

Table 2. Regression fitting

	R^2	MAE	MSE
Decision Tree Regressor	0.905	2377.970	49832386.628
Random Forest Regressor	0.937	1937.810	32993323.634
K Neighbors Regressor	0.594	8199.393	213246328.556

Then, regression tree, random forest regression and K-nearest neighbor regression models are used for the data set of the training set, and regression prediction is made through the independent variables of the test set. The results of regression fitting are shown in Table 2. In the evaluation of R2 value, Random Forest Registrant is the highest, Decision Tree Registrant is slightly lower than Random Forest Registrant, while K Neighbors Registrant is the lowest and has a large gap with the other two models; Among the evaluation values of MAE, Random Forest Regressor is the lowest, Decision Tree Regressor is slightly higher than Random Forest Regressor, while K Neighbors Regressor is the highest and the difference between the two models is nearly 4 times; In the MSE evaluation, the value of Random Forest Regressor is the lowest, Decision Tree Regressor is slightly higher than Random Forest Regressor, while K Neighbors Regressor is the highest and the difference between the values of the other two models is nearly 7 times.

The three evaluation indicators all point out that Random Forest Regressor performs best in the prediction regression of the three models, Decision Tree Regressor performs second, and K Neighbors Regressor performs worst. The three indicators may be very consistent because the weekly sales value itself is large and the error value of many prediction points is large, which leads to the same indicator response of MSE and MAE. The reason for the poor effect of K Neighbors Regressor may be that the distance quantitative indicators used are not suitable for forecasting Wal Mart's data, and the effect of fitting may be improved when other evaluation indicators instead of ordinary arithmetic mean algorithm and weighted average considering distance difference are used.

It can be concluded that if Wal Mart stores want to forecast the weekly sales, they can use the temperature of that week, CPI, Unemployment, oil price, whether it is a holiday, which department, and store type are used to forecast the sales of the current week, so as to enhance the operation capacity of the store itself and improve the turnover capacity of goods.

4. Limitations & Prospects

There are still some limitations in this paper. The first is the evaluation of model indicators. Each model only uses one evaluation indicator when fitting: MSE is used for regression tree fitting in the training set, so is the regression subtree in the random regression forest, and only one evaluation indicator is used for K nearest neighbor.

The second is the parameter setting of the model. The "n_ Estimators" set by the random regression forest in the process of training set fitting has only 20 values, that is, 20 regression trees are selected for fitting. Since there is no more parameter comparison in the paper, the setting of different parameters can be studied in future research, which can get better sales forecast results.

Finally, in terms of model selection, the three models selected in this paper are classic machine learning models, and the generalized linear regression model and SVM model are directly excluded from the correlation analysis.

In the future, research can compare the training fitting evaluation indicators of these models. Maybe after using different indicators, the sales forecast regression results of K nearest neighbor may be better than the other two models. At the same time, in terms of parameter selection, people can choose to use the GridSearch aspect to traverse and select a better parameter. finally, the artificial neural network model can be compared with these models at the same time to find the best model for predicting sales.

5. Conclusion

In summary, this study compared three classic machine learning models: Decision Tree R, Random Forest, and K Neighbors Regressor for the prediction of Walmart sales. According to the anlaysis, it is found that the random regression forest model had the best prediction effect through three evaluation indicators: the correlation number between the predicted value series and the actual value series, mean absolute error, and mean squared error. At the same time, this paper also has several limitations: multiple iteration indicators for model fitting are not selected, and the parameter settings of Random Forest Registrar are not compared. However, the shortcomings do not outweigh the weaknesses. The research results of this paper can play a very important role in predicting the sales volume of each Walmart store this week, to enhance the operation ability of the store itself and improve the turnover ability of goods. Overall, these results offer a guideline for sales forecasts for supermarkets.

References

[1] Hu W, Zhang X. Commodity sales forecast based on ARIMA model residual optimization. 2020 5th International Conference on Communication, Image and Signal Processing (CCISP). IEEE, 2020: 229 - 233.

[2] Wu L, Yan, J. Y., Fan Y. J. Data Mining Algorithms and Statistical Analysis for Sales Data Forecast. Proceedings of the 2012 Fifth International Joint Conference on Computational Sciences and Optimization June 2012, pp. 577 – 581.

[3] Comlan M, Koulo E. Sales Forecast and Design Generation for Textile Products Using Machine Learning. Future of Information and Communication Conference. Springer, Cham, 2022: 183 - 197.

[4] Zhou M, Wang Q. The on-line electronic commerce forecast based on least square support vector machine. 2009 Second International Conference on Information and Computing Science. IEEE, 2009, 2: 75 - 78.

[5] Li Z, Li R, Shang Z, et al. Application of bp neural network to sale forecast for H company. Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2012: 304 - 307.

[6] Singh M, Ghutla B, Jnr R L, et al. Walmart's Sales Data Analysis-A Big Data Analytics Perspective. 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE). IEEE, 2017: 114 - 119.

[7] de Rezende R, Egert K, Marin I, et al. A white-boxed ISSM approach to estimate uncertainty distributions of Walmart sales. International Journal of Forecasting, 2021.

[8] Harsoor A S, Patil A. Forecast of sales of Walmart store using big data applications. International Journal of Research in Engineering and Technology, 2015, 4 (6): 51 - 59.

[9] Thornley D J, Zverev M, Petridis S. Machine learned regression for abductive DNA sequencing. Sixth International Conference on Machine Learning and Applications (ICMLA 2007). IEEE, 2007: 254 - 259.

[10] Aruna M, Anjana M, Chauhan H, et al. Optimized Hyperparameter Tuned Random Forest Regressor Algorithm in Predicting Resale Car Value based on Grid Search Method. ResearchGate 2021.