

Walmart Sales Prediction Using Multiple Linear Regression

Anting Lin*

Department of economic, Hangzhou Dianzi University, Hangzhou, 310000, China

*Corresponding author: 22150136@hdu.edu.cn

Abstract. The sales of Walmart are often considered to be related to many factors, and indeed this is the case. This study aims to examine the impact of objectively observable independent variables on Walmart's weekly sales. The paper employs the Z-score standardization technique to normalize the dimensions and utilizes linear regression analysis to process data from 6,435 weekly sales records of Walmart. The model is derived after preprocessing and standardizing the data in this paper. The findings reveal that most variables have either a minor or significant effect on sales. Notably, temperature and Fuel Price exert a certain influence on sales, while Holiday, Consumer Price Index (CPI), and Unemployment rates have a substantial impact. This investigation into the objective factors affecting Walmart's sales will benefit the management by providing insights to refine strategic planning through environmental surveys, thereby aiding in the enhancement of the shopping atmosphere and customer experience, ultimately aiming to boost sales revenues.

Keywords: Walmart weekly sales; linear regression; objective influencing factors.

1. Introduction

Shopping is an indispensable part of modern people's daily life and sales are regarded as one of the important indicators of supermarket work efficiency. As one of the largest supermarkets in the world, Walmart's sales often receive widespread attention. As of January 2023, Walmart is the largest retailer in the world, with 10,623 stores and 380 distribution facilities in 27 countries and its revenues since 2022 have exceeded \$600 billion a year [1]. In recent two years, Walmart U.S. comp sales increased 19.6% and eCommerce sales grew 35% [2]. Obviously, there are a lot of objective factors that affect Walmart's sales. Exploring the impact of various factors and predicting sales will also become the most important decision-making reasons for Walmart managers. Therefore, this paper aims to find out objective factors that have a greater impact on sales and build a mathematic model for forecasting Walmart sales.

The factors that affect the sales of Walmart are many and complex, involving a variety of subjective and objective factors. Niu put forward the XGBoost sale prediction model which combines XGBoost algorithm and meticulous feature engineering processing for predicting Walmart's sales problem [3]. The experimental results indicate that the sale forecast model based on XGBoost outperforms the other machine learning models. Xiao used multiple linear regression methods and built deep neural network models to analyze data and make predictions [4]. In this paper, in order to avoid reducing the multicollinearity problem, the correlation between attributes is taken into account, explaining the stronger correlation features between variables. At the same time, the concept of dimension is also introduced for data preprocessing. Singh et al. adopted big data analysis, including Spark and MapReduce [5]. In particular, those researchers leverage the Spark framework's Scala and Python API to understand Walmart's sales action strategy and explain consumer behavior through data visualization. Finally, under the analysis of factors such as temperature, fuel prices and holidays, and it is pointed out that more sales can be generated when fuel prices are within a reasonable range, for example.

In similar directions, Yao found the prediction effect of Random Forest Registrar works well and the study results shed light on guiding further exploration of Sales forecasts for supermarkets [6]. Wu et al. illustrated that Auto-Regressive Moving Average Model (ARMAV) with linear trend method is the best benchmark model to forecast sales data for new product with trend and with sales person's inputs [7]. The EMD-XGBOOST-RLSE model is used to solve the complex production process prediction problem by Xu et al. [8]. Harsoor et al. look at Walmart's sales at stores in different

geographic locations [9]. Some people think that the improved least squares support vector machine algorithm can effectively predict business information [10]. Jeswani's report shows that the gradient lift model provided the most accurate sales predictions and slight relationships were observed between factors such as store size, holidays, unemployment, and weekly sales [11].

In summary, after consideration and optimization, this paper will analyze five variables: Temperature, Holiday Flag, Fuel Price, customer price index (CPI), Unemployment to build a linear regression model for predicting Walmart sales.

2. Methods

2.1. Data Source

The dataset used in this paper is fetched from the Kaggle website (Walmart Dataset). It was from 2010 to 2012, in the file Walmart Store sales. This dataset contains 6435 groups of data, and this research selected all of them as samples. The original dataset remained in .csv format.

2.2. Variable Selection

The original dataset has a large amount of data. But there is no null for variables, so this literature directly preprocessed the dataset. By pre-processing the above variables through Z-score normalization, a dataset with a unified dimension can be obtained. The formula for Z-score normalization is provided here:

$$z = \frac{x_i - \bar{x}}{Std} \quad (1)$$

The dataset contains 5 variables (Temperature, Holiday Flag, Fuel Price, customer price index (CPI), Unemployment) and one dependent variable (Weekly sales). The specific description of this dataset is shown in Table 1:

Table 1. List of Variables

Variable	Logogram	Meaning
Temperature	x_1	Temperature on the day of sale
Holiday Flag	x_2	Some special period may deeply influence sales
Fuel Price	x_3	Whether the week is a special holiday week 1 – Holiday week 0 – None holiday week
CPI	x_4	Prevailing consumer price index
Unemployment	x_5	Prevailing unemployment rate
Weekly sales	Y	Sales for the given store

2.3. Method Introduction

This paper employs the method of multiple linear regression for analysis. The initial effort focused on identifying relationships among various variables, which were then used to construct a linear regression model for predictive purposes.

Multiple linear regression is a statistical method used to analyze the linear relationship between two or more independent variables (explanatory variables) and one dependent variable (response variable). The multiple linear regression model describes the relationship between the independent and dependent variables through an equation that typically includes an intercept term and multiple slope parameters, each corresponding to an independent variable. The model equation is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (2)$$

And the parameters in the model, including the intercept and slopes, are usually estimated using the least squares method. This approach aims to minimize the difference between the observed values

and the predicted values of the model. After estimating the model parameters, significance tests are commonly performed to determine if the independent variables have a statistically significant effect on the dependent variable. To ensure the validity of the model, regression diagnostics are conducted to check whether the data adheres to the basic assumptions of the model, such as the independence of error terms, normality, and homogeneity of variance. Overall, multiple linear regression is a powerful tool that helps researchers understand complex relationships among multiple variables and plays a role in prediction and decision-making.

3. Results and Discussion

3.1. Descriptive Analysis

Initially, a box plot of Walmart's weekly sales can be generated, which provides a rough idea of the range of its data. Below is the resulting box plot presentation:

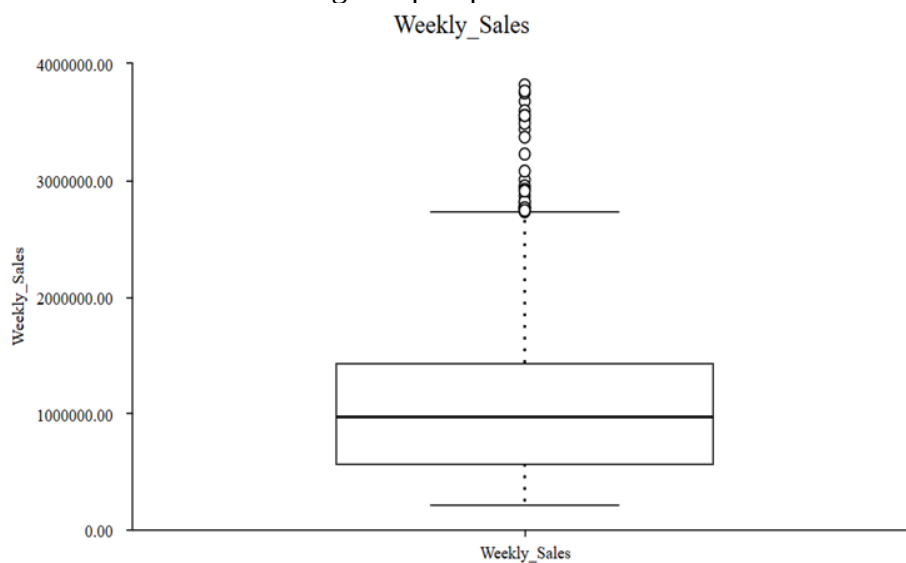


Fig. 1 Walmart Weekly Sales box plot

As can be seen from the Fig. 1, Walmart's weekly sales are mostly distributed around 1,000,000, roughly exhibiting a skewed normal distribution. It can also be observed that there are many outliers in the data, indicating a need for preprocessing of the data.

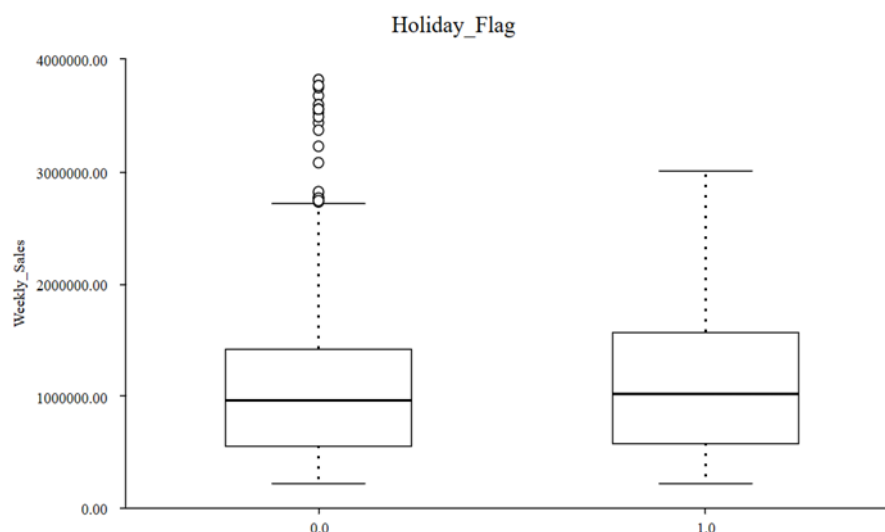


Fig. 2 Weekly Sales & Holiday Flag box plot

In Fig. 2, it's evident that Walmart's sales during holidays are significantly higher than on non-holidays, which aligns with common sense. In summary, the factors affecting Walmart's weekly sales are diverse and complex, encompassing both subjective and objective elements. This paper focuses on studying the impact of objective factors on Walmart's weekly sales. However, to avoid the issues of interaction between objective factors and multicollinearity, a correlation analysis will be conducted among the independent variables below.

3.2. Model Results

Fig. 3 is a heatmap of Pearson correlation coefficients. Clearly, there are no very strong correlations among the independent variables studied in this paper. The Pearson correlation coefficients between each pair of random variables do not exceed 0.3, indicating that there is no strong linear relationship between the variables, thus avoiding the problem of multicollinearity.

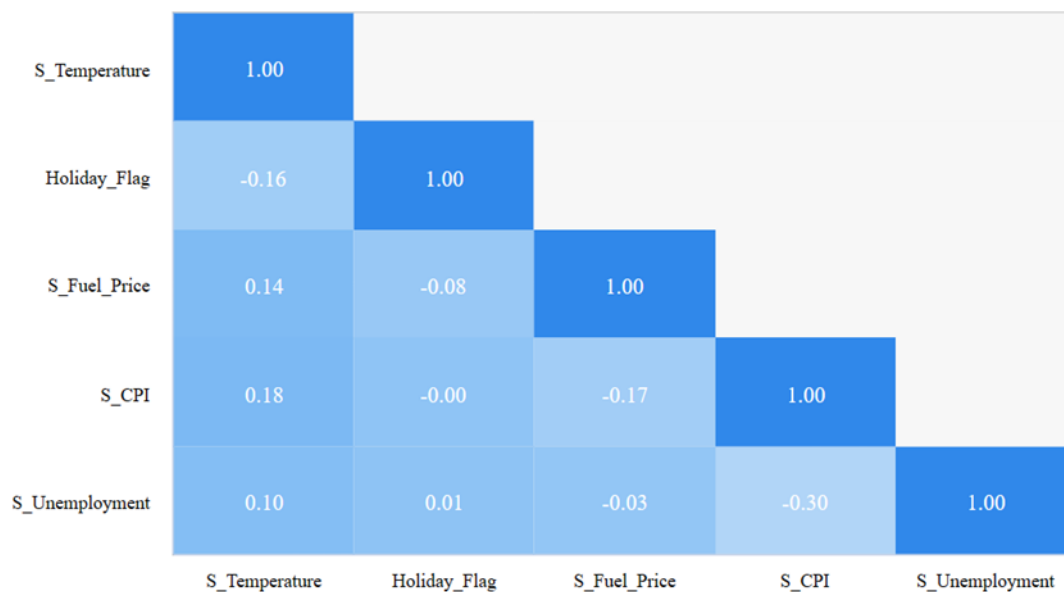


Fig. 3 Pearson correlation coefficient

Table 2. Regression coefficient table

	B	S.E.	Beta	t	p	VIF	Tolerance
Constant	-0.009	0.013	-	-0.726	0.468	-	-
Temperature	-0.024	0.013	-0.024	-1.808	0.071	1.130	0.885
Holiday Flag	0.133	0.049	0.034	2.710	0.007**	1.029	0.972
Fuel Price	-0.008	0.013	-0.008	-0.645	0.519	1.084	0.922
CPI	-0.111	0.014	-0.111	-8.194	0.000**	1.221	0.819
Unemployment	-0.138	0.013	-0.138	-10.460	0.000**	1.150	0.869
R2	0.025						
F	F (5,6429) =33.570, p=0.000						
D-W	0.114						

Table 2 shows the regression coefficients for the multiple linear regression equation model. It is evident that the p-values for Holiday Flag, CPI, and Unemployment are less than 0.05, indicating that

these factors have a significant impact on Walmart's weekly sales. Moreover, based on the above data, the relevant multiple linear regression equation can be derived and compared:

$$y = -0.009 - 0.024x_1 + 0.133x_2 - 0.008x_3 - 0.111x_4 - 0.138x_5 + \varepsilon \quad (3)$$

This model was derived after preprocessing and standardizing the data in this paper, with a focus on studying the correlation between various independent variables and dependent variables. At the same time, it has also minimized the issue of multicollinearity to a considerable extent, making it more aligned with real scenarios.

4. Conclusion

The research presented in this paper focuses on examining the correlations between various variables, aiming to describe these relationships through the methodologies of linear regression. The findings indicate that Walmart's weekly sales are influenced by factors such as temperature, holidays, fuel prices, consumer price index (CPI), and unemployment rate. Among these, holidays, CPI, and the unemployment rate have a more substantial impact on sales.

It is undeniable that the factors affecting Walmart's weekly sales are diverse and complex. This is not only due to objective factors but also many subjective elements that alter sales to some extent, such as the service attitude of store employees and local shopping preferences. Due to the limitations in data volume and a lack of subjective awareness regarding the impact factors, it is challenging for linear models to provide a perfect explanation for the prediction of actual results. This also significantly affects the accuracy of the model. Nevertheless, this paper still possesses certain strengths and feasibility. On one hand, the adopted method of linear regression allows for a more direct presentation of how each independent variable impacts the dependent variable, with correlation analysis effectively demonstrating the trends in the dependent variable and predicting its trajectory. On the other hand, the use of visual graphics to illustrate the data makes the results more intuitive.

The purpose of this paper is to provide Walmart operators with suggestions regarding the factors that influence Walmart's sales. The objective factors studied in this paper will assist them in understanding the current operating environment and timely improve their business strategies, thereby enhancing selling efficiency and offering customers a better shopping experience.

References

- [1] Hu Xiao. Wal-Mart Retail Sales Data Forecast Analysis. *Statistics and Applications*, 2023, 12(6): 1683-1695.
- [2] 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), 2017.
- [3] Yao B. Walmart Sales Prediction Based on Decision Tree, Random Forest, and K Neighbors Regressor. *Highlights in Business, Economics and Management*, 2023, 5: 330-335.
- [4] Wu L, Yan J Y, Fan Y J. Data Mining Algorithms and Statistical Analysis for Sales Data Forecast. *Fifth International Joint Conference on Computational Sciences & Optimization*, IEEE, 2012, 577-581.
- [5] Xu Z G. Complex production process prediction model based on EMD-XGBOOST-RLSE. *Proceedings of 2017 9th International Conference on Modelling Identification and Control (ICMIC)*, 2017.
- [6] Harsoor A S, Patil A. Forecast of Sales of Walmart Store Using Big Data Applications. Working paper, 2024.
- [7] Zhou M, Wang Q. The On-line Electronic Commerce Forecast Based on Least Square Support Vector Machine. *International Conference on Information & Computing Science*. IEEE, 2009.
- [8] Jeswani R. Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard. Rochester Institute of Technology, 2021.
- [9] Zhu Xiaohua. Impact and Analysis of the Implementation of the New Income Standards on the Recognition of Revenue in Cooperative Supermarkets. *Shanghai Business*, 2023, 2: 34-36.

- [10] Left Liangjun, Zhang Lijuan. Analysis of the Impact of Agricultural Supermarket Operation on the Agricultural Industry Chain. Rural Economy, 2003, 2.
- [11] Bin Murong, Zhou Faming. Economic analysis and promotion ideas for the operation of agricultural product chain supermarkets. National Business Situation and Economic Theory Research, 2007.