Sales Prediction of Walmart Based on Regression Models

Jiayuan Zhang[1,*]

[1]University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA

*jz116@illinois.edu

Abstract. In recent years, sales prediction remains a hot and interesting issues in fast sales industry. This study offers a deep dive into Walmart's sales prediction based on regression models, mainly focused on multiple linear regression models. The paper starts with a brief introduction to Walmart's history and operations. Subsequently, it shifts the focus to the importance of sales forecasting, prevailing studies, and current research about sales forecasting. Properly predicting future sales is important to a firm's success, and different methods have their own advantages and limitations. The study also analyzes the dataset, introducing the response and explanatory variables and the regression method used. Then, the paper gives a comprehensive analysis based on five tasks and a multiple linear regression model. After showing the result, the paper provides some insights into the data. Finally, the research offers limitations of the analysis and some future outlooks on sales forecasting. Overall, these results shed light on guiding further exploration of sales prediction.

Keywords: Sales Prediction, Walmart, Linear Model.

1 Introduction

Walmart is the world's largest chain retailer, headquartered in Arkansas, United States. As a growing global digital enterprise, it operates over 11,500 stores worldwide, spanning across the United States, Canada, the United Kingdom, and many other countries [1-5]. Walmart is a diversified company, offering a wide range of products and services, including groceries, home goods, electronics, and financial services, among others. Walmart started with one man: In 1962, Sam Walton began with just one store and one mission: to help people save money so they could live better. In Walton's autobiography, "Sam Walton: Made in America," he mentioned his childhood experience that taught him "learning to value a dollar." Thus, his vision was to provide more value to customers by offering high-quality products at lower prices compared to competitors. This approach propelled Walmart to remarkable success not only in the United States but also globally over the following decades. The company has maintained its leadership position in the global retail market by adopting cutting-edge retail technology and innovative business models. Through unique logistics and supply chain management techniques, Walmart has achieved new heights of operational efficiency, enabling it to sustain its low-price strategy while maintaining healthy profitability levels. According

A. Bhunia et al. (eds.), *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Advances in Economics, Business and Management Research 264,
https://doi.org/10.2991/978-94-6463-298-9_45

to the Wall Street Journal, the company's sales rose in the most recent quarter and this higher-than-expected first quarter performance lifted the company's outlook on profits for the full year.

One of the most crucial things a firm does is sales forecasting. It supports sales planning and is used across various enterprises for staffing and budgeting. Good sale forecasting can help a company to make the right strategy and keep sales on track for future years. Companies need a relatively accurate and detailed prediction to determine their overall plans, such as how many products to produce or how many people they hire or lay off. However, by using the wrong techniques, sales forecasting can also have negative impacts on a company's future profits.

Understanding what influences sales is essential for accurate sales forecasting. According to the econometrician, retail sales are part of the economic flows and can be traced in an endless circle. Based on this principle, sales prediction is the statistical application of this general econometric approach. Sales prediction is closely related to national income, and the purchase of goods is integrally related to disposable income. Thus, companies need to analyze customers' purchasing power before making predictions. Generally, the longer the life of the product and the higher its price bracket, the greater is the response in sales to changes in disposable income [6]. This principle underlies modern forecasting methods many companies used today.

There are many forecasting methods, and the correlation method is one of the most popularized ones. This method uses an equation to explain sales fluctuation in terms of presumably causal variables to solve for the sales value [7]. Though this method has obvious limitations, it is easily mastered and can be used as a forecasting aid. The most commonly used forecasting method involves statistical techniques that ignore "the dynamics or memory of the process involved and treat the data. Times series analysis provides a solution to serially correlated data, giving it an advantage over economic methods [8, 9]. Recently, with the development of modern technology, the combination of computer and time series analysis represents a breakthrough. Nowadays, there are more and more applications of modern technologies to aid sale forecasting.

2 Data & Method

2.1 Datasets

As one of the leading retail stores in the US, Walmart would like to predict sales and demand accurately. Many factors, e.g., certain events and holidays, will impact Walmart's daily sales. Thus, the company is facing the challenge of unforeseen demands, and occasionally there are stock shortages as a result of a bad machine learning system. An ideal multiple linear regression model will predict demand accurately and ingest factors like economic conditions, including CPI, Unemployment Index, etc. Walmart runs several promotional markdown events throughout the year. These markdowns precede major holidays, the four most significant: the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays, are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. This study uses the data from Kaggle, which contains Walmart’s historical sales data from 2010-02-05 to 2012-11-01. The dataset this article applied contains 45 Walmart stores. The response variable is the weekly sales for the given stores (Weekly_Sales). There are several explanatory variables, including the store number (Store), whether the week is a special holiday week (Holiday_Flag), the temperature on the day of the sale (Temperature), cost of fuel in the region (Fuel_Price), prevailing consumer price index (CPI), and prevailing unemployment rate (Unemployment). For variable “Holiday_Flag,” 1 represents the holiday, and 0 represents the non-holiday week [10].

2.2 Model and Programming Language Usage

The article utilizes R to analyze and predict Walmart's sales. The article first analysis on five questions, including which store has maximum sales, which store has a maximum standard deviation of sales, which store/s has a good quarterly growth rate in Q3'2012, which holiday has higher sales than the mean sales in non-holiday season for all stores together, and a monthly and semester view of sales in units. The five questions above give a glance at the dataset. Then, the article builds statistical prediction models to forecast demand using multiple linear regression. The article first creates a data frame using certain columns and removes any outliers. Then by creating a correlation plot, correlation matrix, and dummy variables, the article gains insights into which model is the best to predict sales at Walmart. When evaluating an MLR model, R-squared, adjusted R-squared, residual analysis, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) provide insights into its overall performance. R-squared and adjusted R-squared measure how much of the target variable's variance is explained by the model's predictors, and higher values indicate a better fit. Residual analysis helps identify patterns or violations of assumptions in the residuals, such as nonlinearity, heteroscedasticity, or outliers. Ideally, the residuals should be randomly scattered around zero. Finally, MSE or RMSE assesses the average.

3 Results & Discussion

There are mainly five analysis tasks for this dataset. After uploading, cleaning, and understanding the data, the analysis starts with which store has the maximum sales. The process begins with aggregating data by the “Store” category, with the subsequent calculation of the sum of “Weekly_Sales.” It then proceeds with a change in the column name of sales. In an effort to identify the store with the highest sales, as the graph shows, the stores are arranged based on their sales in descending order. The first store that appears in this order is then chosen.

To ensure the order remains constant for the graph, the “Store” column is converted into a factor. After these steps, a plot of “Store” versus “TotalSales” is generated. The insights derived from these analyses reveal that Store 20 tops the sales chart, with a total of 301.39 million. Store 4 follows closely in second place, with sales amounting to 299.54 million. At the bottom of the list is Store 33, which registers the least sales at 37.16 million.

The second analysis task is determining which store has the maximum standard deviation and the coefficient of mean to standard deviation. To accomplish this task, the data were first aggregated by "Store," and the standard deviation of "Weekly_Sales" was calculated. This was followed by a renaming of the columns. A "Store_Sales_Variation" dataframe was then created, representing the Coefficient of Variation (CV) for Sales by Store. The analysis was specifically designed to find the store with the highest standard deviation, which resulted in Store 14 being identified.

A density plot was also drawn for Store 14, providing further insight. Store 14 has the highest standard deviation, amounting to 317.5K, and a CV of 15.714. The sales for Store 14 are skewed to the right, indicating that the store had exceptionally high sales during a few weeks, which increased its standard deviation. Despite having the highest standard deviation, Store 14's CV is relatively low, suggesting that while there is variation in its weekly sales, this variation is small relative to the store's mean sales.

Moving on to which store/s has a good quarterly growth rate in Q3’2012, a new dataframe was created to facilitate data manipulation. This included the creation of a month-year column in data2. The data was then subsetted for Q3-2012 (i.e., July, August, September 2012) and Q2-2012 (i.e., April, May, June 2012). The next step was to aggregate the sales by store for Q3-2012 and change column names accordingly. This process was mirrored for Q2-2012, with sales aggregated by store and column names altered. Following these steps, data for the two quarters were merged by store. This allowed for creating a growth rate column for sales by the store in the newly formed dataframe, specifically targeting positive growth rates.

Interestingly, Store 7 had the highest growth rate of 13.33%, with Store 16 following at 8.49% and Store 35 at 4.47%. Seven additional stores exhibited positive growth rates: 26, 39, 41, 44, 24, 40, and 23. A visual representation of growth rates was also generated. However, Store 14 had the highest negative growth rate, underscoring the need for further examination to understand its sales performance.

The next analysis is about which holidays harm sales and which holidays have higher sales than the mean sales in the non-holiday season. This task involves creating a dataframe with holiday data, merging it with another dataframe, and replacing any null values in the event column with "No_Holiday." Then, creating a new dataframe that calculates the mean sales for both "No_Holiday" periods and different events is needed. It's also better to change the column names to ensure clarity and ease of interpretation. Based on the given analysis, it is determined that Christmas and Labor Day have a negative impact on sales, while Thanksgiving and the Super Bowl have a positive impact. To delve deeper into the investigation, examining the negative impact concerning specific holiday dates and non-holiday dates is recommended. The analysis filtered the data for holiday dates, calculated the mean of weekly sales during those periods, and calculated the mean of weekly sales for non-holiday periods.

Based on the insights gained from the analysis, several key observations can be made. Firstly, holidays such as the Super Bowl, Thanksgiving, and Labor Day demonstrate higher sales than the mean sales during non-holiday seasons. These holidays positively impact overall sales, indicating that customers tend to spend more during these periods. Secondly, specific dates, such as September 9th, 2011, September 10th, 2010, December 30th, 2011, and December 31st, 2010, have been identified as hurting sales. The sales figures fell below the mean sales on these dates, suggesting that certain holidays or events might not drive significant consumer spending.

Furthermore, the analysis reveals a consistent pattern: all dates related to Christmas exhibit lower sales than the mean sales. In comparison, all dates associated with the Super Bowl and Thanksgiving show higher sales than the mean. This indicates that Christmas has a negative impact on sales, whereas the Super Bowl and Thanksgiving have a positive impact. Lastly, it is interesting to note that Labor Day, despite having two days with sales below the mean value, still demonstrates an overall positive impact on sales. This suggests that the positive effect of Labor Day on the remaining days compensates for any temporary downturn. These insights provide valuable information for businesses, enabling them to identify the holidays that generate higher sales than the mean sales during non-holiday seasons for all stores combined. Armed with this knowledge, businesses can tailor their strategies to capitalize on peak sales periods, optimize inventory management, and effectively target marketing efforts to maximize profitability during holiday seasons.

Based on the analysis, several insights can be derived regarding the impact of different holidays on sales. Firstly, while the Super Bowl and Labor Day show higher sales than the mean, the difference is relatively small, indicating a moderate impact. In contrast, Thanksgiving stands out as a holiday that generates a significantly higher positive impact on sales than other holidays. Secondly, the analysis confirms that the Christmas holiday period generally exhibits lower sales compared to the mean sales during non-holiday periods. Specifically, the dates of December 30th, 2011, and December 31st, 2010, related to Christmas, have a negative impact on sales, further emphasizing the lower sales during this holiday. However, a noteworthy observation from the graph is that the week just before Christmas records the highest sales. This can be attributed to customers engaging in pre-Christmas shopping to prepare for the holiday season or celebrate the popular Advent tradition.

Overall, these insights shed light on the varying impacts of different holidays on sales. While some holidays show only slight deviations from the mean sales, Thanksgiving stands out as a holiday with a significantly positive impact on sales. On the other hand, Christmas is characterized by lower sales, but the week preceding the holiday witnesses the highest sales, suggesting a unique shopping behavior pattern during that period. Businesses can leverage these insights to optimize their strategies and effectively target their marketing efforts during different holiday seasons.

The last analysis task is about monthly and semester views of sales. The data manipulation process began with converting the 'Date' column into a factor. To achieve this, the original format of the date was defined, followed by the definition of the desired format. The data was then aggregated by 'Month - Year,' and the sum of 'Weekly_Sales' was calculated and converted into a dataframe. The 'Year-Month' column was converted into a factor to ensure the order remained unchanged for plotting purposes.

The analysis started with comparing sales versus semester using the "lubridate" package and required converting the relevant data into a date format. To aid this process, a 'semester' column inclusive of the year was created. A new dataframe 's' was then formulated, containing total sales for every semester. An additional column was added that rewrote the semester and year in a different format. Subsequently, a graph was plotted to visualize the relationship between semester and sales. The insights derived from this analysis revealed that sales peaked in December and were at their lowest in January. Moreover, the sales for the second semester of each year were generally higher. However, the plot indicated a drop in sales for the second semester of 2012 (S2-2012) and the first semester of 2010 (S1-2010). This discrepancy is attributable to the lack of January data for S1-2010 and missing data for November and December 2012 in the S2-2012 semester.

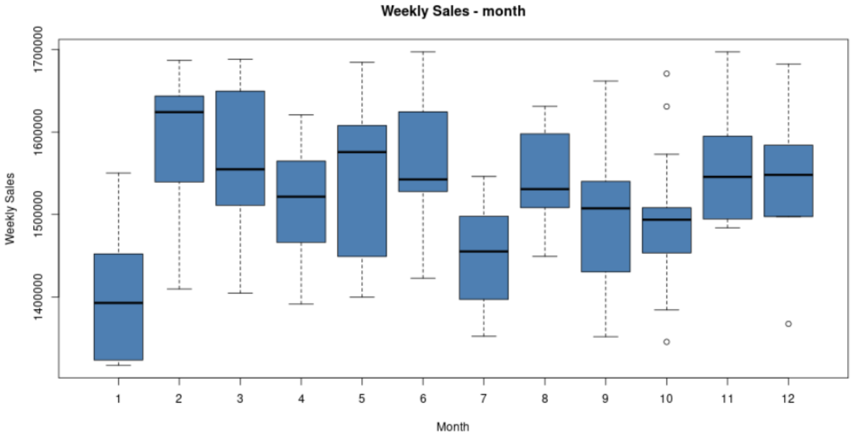

Fig. 1. Weekly sales monthly.

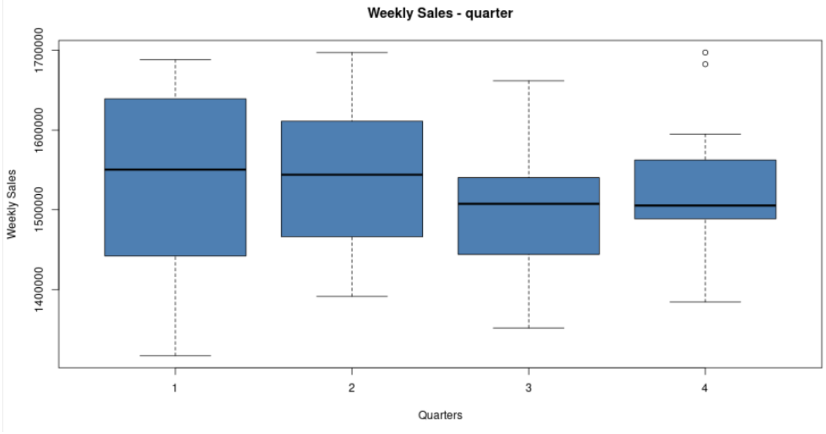

Fig. 2. Weekly sales quarter.

After analyzing the five tasks, the statistical model uses multiple linear regression. For the purposes of manipulation, a copy of the original data was created. The focus was then narrowed to only the first store, as predictions were solely required for this entity. The 'Date' column in the dataframe was then converted to a date format and arranged in ascending order. Additionally, a 'Week Number', 'Month', and 'Quarter' column were created within the dataframe, effectively adding quarter and month details. Next, a 'Holiday_date' vector was created and assigned a date format. An 'Events' vector was also constructed, and a new dataframe was created with 'Events' and 'Date'. These two dataframes were subsequently merged to form a unified dataset. Any null values in the 'Event' column were replaced with 'No_Holiday' to ensure data completeness. A scatter plot was then generated for data visualization, offering a clear and comprehensive overview of the trends and patterns within the dataset. This visualization allows for a better understanding of the relationships within the data and informs further predictive modeling.

Afterward, Detection and removal of outliers using Bi Variate Box Plots are needed. The data analysis began with generating a boxplot to identify potential outliers, followed by their removal to create a more accurate dataset. To this end, a new dataframe was created specifically for outlier treatment. Given that the focus of the analysis was on predicting sales, outliers were removed based on a variety of parameters. These parameters included temperature, where five outliers were identified and removed; the Consumer Price Index (CPI), where one outlier was found and removed; unemployment rate, which had three outliers that were subsequently eliminated; and fuel price, from which two outliers were removed. Additionally, outlier treatment was conducted for the 'Holiday Flag' category, but no outliers were found. The 'Month' category had four outliers that were removed, while the 'Quarter' category had 2. In total, 17 observations were removed, representing approximately 11.8% of the data frame. The order of outlier removal was as follows: Temperature (5 outliers), CPI (1 outlier), Unemployment (3 outliers), Fuel price (2 outliers), Month (4 outliers, as presented in Fig. 1), and Quarter (2 outliers, seen from Fig. 2). No outliers were found for the Holiday Flag category. Following this step, unnecessary columns were removed, and the structure of the 'Events' category was altered to better fit the refined data frame (as given in Fig. 3).

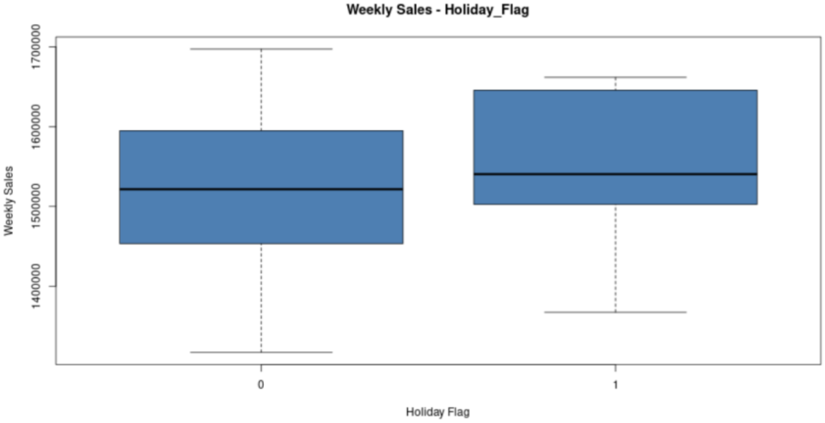

Fig. 3. Holiday flag.

Moving on to the correlation plot and correlation matrix analysis (seen from Table. 1), it was observed that there was a very low correlation between Temperature and Weekly Sales. This suggests that temperature has little to no influence on Weekly Sales, and hence, this parameter could potentially be omitted from future analyses. Similarly, a low correlation was found between the variables month, quarter, and Holiday Flag with Weekly Sales. This may be attributed to the treatment of these categorical variables as continuous variables, which could distort their relationship with Weekly Sales. Consequently, these factors may not be as significant in predicting Weekly Sales as initially thought.

Table 1. Correlation table.

	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Week_Number	month	quarter
Weekly_Sales	1.00	0.16	-0.03	0.28	0.31	-0.25	0.26	-0.07	-0.15
Holiday_Flag	0.16	1.00	-0.16	-0.07	-0.01	0.07	-0.01	0.09	0.03
Temperature	-0.03	-0.16	1.00	0.23	0.14	-0.15	0.19	0.45	0.42
Fuel_Price	0.28	-0.07	0.23	1.00	0.76	-0.50	0.79	-0.03	-0.04
CPI	0.31	-0.01	0.14	0.76	1.00	-0.84	0.98	0.11	0.09
Unemployment	-0.25	0.07	-0.15	-0.50	-0.84	1.00	-0.81	-0.05	-0.06
Week_Number	0.26	-0.01	0.19	0.79	0.98	-0.81	1.00	0.19	0.17
month	-0.07	0.09	0.45	-0.03	0.11	-0.05	0.19	1.00	0.96
quarter	-0.15	0.03	0.42	-0.04	0.09	-0.06	0.17	0.96	1.00

4 Limitations & Prospects

Using a regression model to predict Walmart's sales may have some limitations. First, the model's performance is directly related to data quality, and the prediction may be inaccurate if the data contains errors or is missing key information. Moreover, the feature selected is crucial for the model's credibility, and it is necessary to ensure that the features chosen fully reflect the key factors affecting sales. Regression models usually assume a linear relationship between the input and output, but many relationships may be nonlinear. Also, standard regression models may fail to capture time-series effects in sales data. At the same time, the model might be affected by many other factors, such as economic conditions, customer behavior, competitor strategies, etc., and loss of stabilities. Finally, the regression model might establish based on some false statistical assumptions in the real world, which will also affect the model's predictability. Thus, though the regression model could help Walmart predict sales, considering those limitations, Walmart needs to cooperate with other models or methods, such as time series models, deep learning models, etc., to achieve more accurate predictions. With the continuous development of data science and AI technology, Walmart's regression models could be improved by improving techniques for feature selection, introducing complex models to handle nonlinear relationships, and adopting new methods for handling time series data. Meanwhile, considering the complexity of sales forecasting, we also plan to explore regression models in combination with time series models such as ARIMA, state space models, or deep learning models such as long short-term memory (LSTM) networks and convolutional neural networks (CNN). Besides, developing a more robust and flexible forecasting system is necessary to cope with the challenges brought about by changes in the market environment, consumer behavior, and competitor strategies. In general, the regression model has the potential to do Walmart's sales forecasting, and it's promising to better utilize the advantages of this model through continuous exploration and research.

5 Conclusion

In summary, sales forecasting is essential to companies but also complex to be accurate. The result of using a regression model to predict sales for Walmart will depend on various factors such as data quality, feature selection, and model parameters. Ideally, suppose the model is built properly and the training data is comprehensive and accurate. In that case, the model might generate relatively accurate sales forecasts, helping Walmart develop more effective sales strategies and improve its profitability. However, considering the regression model's limitation in dealing with nonlinear relationships, time-series effects, and the high requirements for data quality and feature selection, non-effectively processed models may lead to inaccurate predictions and even wrong decisions. Finally, it is worth noting that any forecasting model is subject to some error. Therefore, one should be cautious when using a regression model to predict sales, not rely too much on the model, and at the same time, check and update the model regularly to adapt to changes in the market environment, consumer behavior, and other factors.

References

1. Walmart Homepage, https://careers.walmart.com/history, last accessed 2023/7/1.
2. Bonanno, A., Goetz, S. J.: WalMart and local economic development: A survey. Economic Development Quarterly, 26(4), 285-297 (2012).
3. Walton, S., Huey, J.: Sam Walton: Made in America. Bantam (1993).
4. LeCavalier, J.: The rule of logistics: Walmart and the architecture of fulfillment. U of Minnesota Press (2016).
5. Chan, A. (Ed.). Walmart in China. Cornell University Press (2019).
6. Kapoor, S. G., Madhok, P., Wu, S. M.: Modeling and Forecasting Sales Data by Time Series Analysis. Journal of Marketing Research, 18(1), 94–100 (1981).
7. McLaughlin, R. L.: The Breakthrough in Sales Forecasting. Journal of Marketing, 27(2), 46–54 (1963).
8. Gould, J. M.: Sales Forecasting. Journal of Marketing, 15(3), 357–361 (1951).
9. Ferber, R.: Sales Forecasting by Correlation Techniques. Journal of Marketing, 18(3), 219–232 (1954).
10. Kaggle Homepage, https://www.kaggle.com/code/sarvaninandipati/analysis-prediction-of-walmart-sales-using-r#Some-holidays-have-a-negative-impact-on-sales.-Find-out-

holidays-which-have-higher-sales-than-the-mean-sales-in-non-holiday-season-for-all-stores-together, last accessed 2023/7/1.