Walmart Sales Forecasting using Different Models

Chenghao Yu [*]

Department of mathematics, Imperial college London, London, United Kingdom

[*] Corresponding Author Email: cy1420@ic.ac.uk

Abstract. This research study compares three regression models, namely Random Forest, Linear Regression, and Lasso Regression, to determine which model has better performance on predicting Walmart sales. Through the analysis of historical sales data, including factors such as time, unemployment rate, CPI, and temperature, the dataset have training and testing sets. Random Forest is implemented and compared with Linear Regression, a traditional statistical method, as well as Lasso Regression, which includes a regularization term for feature selection and prediction accuracy improvement. Performance evaluation is conducted using mean squared error,and R-squared score. The results consistently show that Random Forest outperforms both Linear Regression and Lasso Regression in predicting Walmart sales, demonstrating its accuracy and robustness. This research offers insights into predictive modeling in retail sales forecasting and highlights the potential for using Random Forest as a reliable tool for inventory control, demand forecasting, and strategic planning at Walmart and similar retailers. Overall, this study contributes to the understanding of sales prediction in the retail industry, suggesting avenues for future research in exploring advanced machine learning algorithms and data preprocessing techniques to further improve accuracy.

Keywords: Walmart sales, forecast, linear regression, random forest, lasso regression.

1. Introduction

Walmart is a multinational retailer located in the United States. The Walton family founded it in 1962. Walmart operates 10,586 shops and clubs in 24 countries under 46 distinct identities until October 31, 2022. According to the Fortune Global 500 ranking for October 2022, Walmart is the world's largest company by revenue. With 2.2 million employees, Walmart is also the world's largest private employer [1].

In recent years, Walmart has actively pursued digital transformation, investing in e-commerce, and online sales platforms, and adopting advanced technologies to enhance operational efficiency.

The goal is to find which model can suit the data best and also seeks to determine whether holiday days can have the positive influence on the weekly sales, in order to that retailers can reset the selling combos to get a higher income.

In 2015, Harsoor and Patil collaborated on projecting Walmart Store Sales utilizing big data tools such as Hadoop, MapReduce and Hive to ensure that resources are managed efficiently [2]. The referenced study utilized the identical sales dataset employed in this analysis. However, their approach involved predicting sales for the subsequent 39 weeks through the application of Holt's Winter algorithm. The forecasted sales data was then visually depicted in Tableau, employing bubble charts for representation.

Michael Crown, examined a comparable dataset but focused on time series forecasting using non-seasonal ARIMA models to generate his forecasts [3]. He engaged in ARIMA modeling to generate one-year weekly forecasts by leveraging 2.75 years of sales data. The model incorporated various features, including store information, department details, date, weekly sales, and holiday data. The evaluation of model performance was conducted using the normalized root-mean-square error (NRMSE).

Rashmi had already done with the same dataset and used WMAE to determine which model suited best and make predictions, in 2021 [4].

2. Data introduction

The whole data was collected on Kaggle which offered historic weekly sales data for 45 Walmart locations across the country, as well as department-wide data for these locations.

There are 12 columns, as shown in table 1.

Table 1. different features

fearure name	Meaning
Store	Store number
Date	Week
Temperature	Average temperature in the region
Fuel price	Cost of fuel in the region
Markdown1,2,3,4,5	Anonymized data related to promotional markdowns that Walmart is running
CPI	The consumer price index
Unemployment	The unemployment rate
Is holiday	Whether the week is a special holiday week

It contains training and testing data; aside from the weekly sales data, the testing data is identical to the training data. From 2010-02-05 to 2012-11-01, the training dataset contains weekly sales data for stores and departments. It also has a column that tells whether a certain date is a holiday. The training dataset has 421570 rows, whereas the testing dataset has 115064 rows [4].

As shown in Figure 1, it is clear to see that there are 2 peaks in 3 years which are Dec 2010 and Dec 2011, from this, this study could realize that Thanksgiving and Christmas do imply a positive influence on weekly sales. People are more willing to shop during these periods.

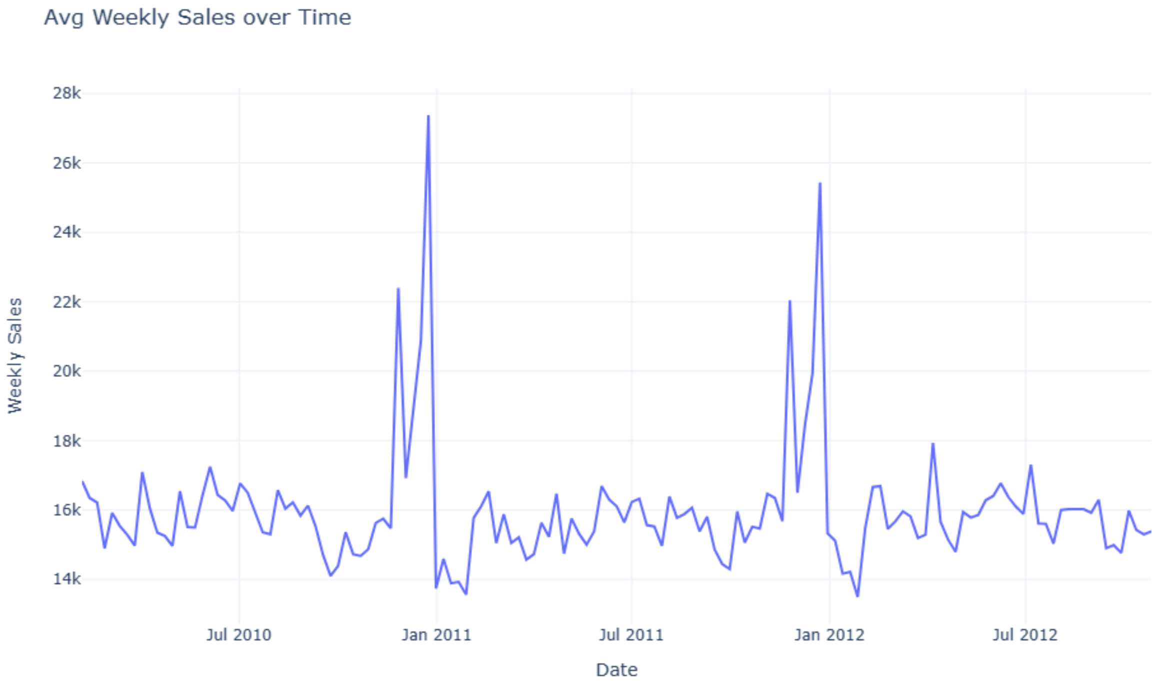

Figure 1. Weekly sale from 2010 to 2012

A correlation matrix expresses the link that exists between variables in a dataset. Each feature in figure1 is connected to other features in figure1, making it easier to determine which variables are more closely related to one another.[5]

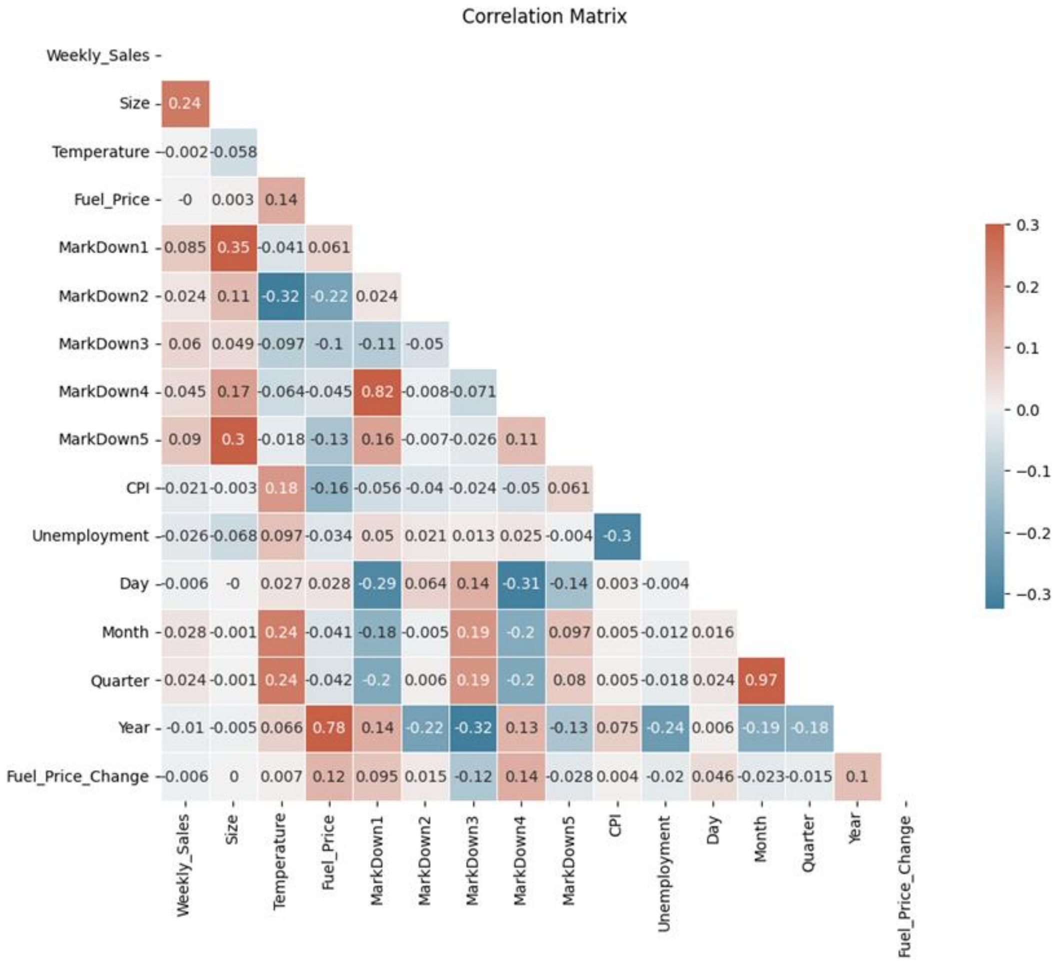

Figure 2. Correlation matrix

As shown in figure 2 MarkDown4 and MarkDown1 have high correlations. MarkDown1 and MarkDown5 have a positive correlation with Size. They could be exclusive offers for bigger stores MarkDown2 have a negative correlation with Temperature MarkDown2 could be offered during winters/holiday offers

3. Model selection and application

Regression analysis is a statistical approach used to model the relationship between one or more independent variables (or predictors) and a dependent variable (or target variable). The goal is to comprehend the nature of the relationship so that it may be used to create predictions or infer causal links.

In the context of predicting sales, regression analysis helps us understand how various factors influence sales figures. Regression analysis provides a powerful tool for understanding and predicting sales in the retail industry, enabling businesses to make informed decisions and stay competitive in a dynamic market.

To compare the models, this paper use means square error and r squared to compare

The mean squared error (MSE) of an estimator in statistics measures the average of the squares of the errors—that is, the average squared difference between the estimated and actual values. [6].

$$\boldsymbol{MSE = \frac{1}{N}\sum_{I=1}^{n} Y_i - Y} \tag{1}$$

3.1. Linear regression

Linear regression is a statistical method that fits a linear equation to observed data to represent the connection between a dependent variable (target) and one or more independent variables (predictors). The basic equation for simple linear regression is:

$$\boldsymbol{Y = \beta + \alpha X + \varepsilon} \tag{2}$$

In this situation, Y is the weekly sales, X is the independent variables (eg: markdown1, 2, 3, 4, 5…), β is intercept, α is slope coefficient and ε represents the error term.

Advantages:

1. Linear regression is easy to understand and interpret, making it accessible to a wide audience.

2. The coefficients in the regression equation provide clear insights into the magnitude and direction of the relationship between variables.

3. Linear regression models are relatively quick to implement, making them useful for initial analyses and exploratory data analysis.

Drawbacks:

1. For non-linear data, linear regression may provide inaccurate predictions.

2. Linear regression can be sensitive to outliers, which can disproportionately influence the model.

3. Linear regression may not capture complex; non-linear relationships present in the data.

Putting the dataset into linear regression, the MSE and R-squared were calculated as table 2;

Table 2. The MSE and R-squared

	MSE	R^2
Linear regression	179048962.5995	0.6566463

3.2. Random forest

Random forests are an ensemble learning method for classification, regression, and other problems that works by generating a large number of decision trees during training. For classification problems, the random forest output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks [7].

Advantages:

Random Forest reduces overfitting and gives more reliable predictions on unknown data by pooling the predictions of numerous trees.

Random Forest is capable of capturing complex relationships within data, making it suitable for datasets with non-linear patterns.

3.The algorithm includes a measure of feature relevance. It can determine which characteristics are more influential in making predictions

Drawbacks:

It is difficult to interpret when facing the large number of trees.

Using a large number of trees and deep trees can be computationally and time-consuming. The algorithm's scalability may become an issue for massive datasets or real-time applications.

While Random Forests are generally robust to overfitting, they can still overfit to noisy data, particularly if the dataset is noisy or contains outliers. The diversity introduced by the ensemble might not be sufficient in some cases.

After using the random forest, the MSE and R-squared were calculated for the training and test sets, as shown in table 3:

Table 3. The MSE and R-squared

	MSE	R^2
Random Forest	22432271.3524	0.9569827

3.3. Lasso regression

Lasso Regression is a linear regression technique with a regularization term in its objective function. To overcome some of the drawbacks of traditional linear regression, it was introduced as a method for both variable selection and regularization.

Consider a sample of N cases, each of which has p variables and a single result. Let Yi be the outcome and $\mathrm{x_i} \coloneqq (\mathrm{x_1}, \mathrm{x_2} \ldots \mathrm{xp})_\mathrm{i}^\mathrm{T}$be the covariate vector for the ith case. Then the objective of the lasso is to solve

$$\min\left\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\} \text{ subject to } \sum_{j=1}^{p}(\beta_j) \le t, \text{[8]} \tag{3}$$

Here β_0 is the constant coefficient, $\beta \coloneqq (\beta_1, \beta_2, \ldots, \beta_\mathrm{P})$ is the coefficient vector, and t is a prespecified free parameter that determines the degree of regularization.

Advantages:

Lasso can automatically select the most relevant features by setting the coefficients of less important features to zero.

The regularization term in Lasso helps prevent overfitting and improves the model's generalization performance.

The sparsity induced by Lasso makes the model more interpretable by focusing on a subset of important features.

Drawbacks:

Lasso's ability to perform variable selection, setting some coefficients to exactly zero, can lead to instability when the data changes slightly. Small variations in the dataset or noise might result in different variables being selected.

Lasso is affected by the magnitude of the features. Larger scale features may have a stronger impact on the regularization term, potentially influencing variable selection. Standardizing or normalizing features is often recommended to mitigate this issue.

In situations where multiple variables are highly correlated, Lasso tends to arbitrarily select one variable over another. The choice might depend on the specific optimization path taken during the fitting process, leading to a lack of consistency.

After using Lasso model in python, the MSE and R-squared were calculated and shown in table 4:

Table 4. The MSE and R-squared

	MSE	R^2
Lasso regression	179047725.1763	0.6566487

3.4. Model Comparison and Analysis

Random Forest has smallest MSE and largest R squared, so it is the best model among these three. Holidays, especially thanksgiving and Christmas will increase the sales.

However, the dataset was too old. The data is collected 10 years ago and the shopping method changed rapidly through these 10 years. The factors will be more complicated since online shopping has hugely developed in this period of time.

4. Conclusion

In conclusion, Walmart's exploration of diverse sales forecasting models, including linear regression, lasso regression, and random forest, has been pivotal in elevating their predictive capabilities. The comparative analysis reveals that while linear regression and lasso regression provided valuable insights, the implementation of the random forest model emerged as the most effective solution for Walmart's sales forecasting needs.

The linear regression model, with its simplicity and interpretability, offered a foundational understanding of the relationships between variables. Lasso regression, with its feature selection capabilities, enhanced model efficiency by identifying and prioritizing influential factors. However, the random forest model demonstrated superior performance, leveraging the strength of ensemble learning to capture complex patterns and interactions within the data.

The resilience of the random forest model in handling non-linearity and accommodating a multitude of variables positions it as the optimal choice for Walmart's sales forecasting. Its ability to mitigate overfitting, handle diverse data types, and adapt to evolving market dynamics underscores its superiority in delivering accurate and robust predictions.

As Walmart continues to refine its forecasting strategies, the reliance on the random forest model signifies a commitment to leveraging advanced analytics for strategic decision-making. This journey reinforces the significance of selecting the most suitable models tailored to the specific nuances of the retail industry, ensuring Walmart's sustained competitiveness and operational excellence.

References

[1] Wikipedia contributors. Walmart. Wikipedia. 2023.

[2] Crown, M. Weekly sales forecasts using non-seasonal arima models. 2016. Retrieved on December 12 2023.Retrieved from http://mxcrown.com/walmart-sales-forecasting/.

[3] 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.

[4] Jeswani R. Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard. Rochester Institute of Technology, 2021.

[5] Glen, S. Elementary statistics for the rest of us. 2016.

[6] Mean Squared Error (MSE)". 2023.

[7] Ho, T K. “The Random Subspace Method for Constructing Decision Forests.” IEEE Trans. Pattern Anal. Mach. Intell. 1998, 20: 832 - 844.

[8] Tibshirani, Rt. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society. Series B (Methodological), 1996, 58 (1): 267 – 88.