Sales Forecast of Walmart on Account of Multivariate Regression and Machine Learning Methods

Zhaoyu Chen *

* Corresponding author: 195139@stu.hebut.edu.cn

School of Artificial Intelligence and Data Science Hebei University of Technology Tianjin, China

Abstract—Sales prediction is an integral part of modern commercial sales problems. Machine learning, especially supervised machine learning algorithms, can be used to find complex and unpredictable trends in dynamic sales problems, including various potentially influential variables. The success of each sales company is closely tied to accurate sales predictions, which can help enterprises find potential risks and make more sensible decisions. This paper aims to predict the future weekly sales of Walmart based on three different regression models (Multiple Linear Regression, Elastic-Net Regression, and Polynomial Regression). In addition, statistical metrics (e.g., R2 and RMSE) are adopted to evaluate the quality of the model. According to the analysis, relevant variables that had an obvious effect on weekly sales are the Holiday, Date, Type, and Stores. In terms of the prediction model, the simplest Multiple Linear Regression Model makes the best sales prediction, which has a moderate R2-Score of about 0.933 and the RMSE with the smallest difference between the training and test sets. Overall, these results shed light on guiding the selection of appropriate models in sales prediction as well as providing suggestions to retail companies to make sales strategies.

Keywords-Sales prediction of Walmart; Multiple Linear Regression Model; Elastic-Net Regression Model; Polynomial Regression Model.

1 INTRODUCTION

Modern commercial sales can forecast sales in advance. A retailer in one of the areas of sales can improve its sales strategy based on different sales forecasts and customer needs to change their mindsets [1]. Walmart Stores, Inc. (WMT) is an American multinational retail company that operates department stores or discount stores of various sizes around the world. Based on income, Walmart is the largest retail company, which has headquarters located in Bentonville, Arkansas.

This paper aims to study the future weekly sales of Walmart by different models and find the impact of different independent variables on weekly sales. The data used in this paper come from the Kaggle dataset: Walmart Sales Forecasting [2]. Multiple Linear Regression Model, Elastic-Net Regression Model, and Polynomial Regression Model were used to predict future sales. According to the analysis, Type and Size have a great impact on Walmart's weekly sales, and when the size becomes larger, Walmart will get the higher weekly sales. Secondly, Store, Department and Holiday have a slight influence on weekly sales. Stores 10 and 35 have higher weekly sales values sometimes, but on average Store 20 and Store 4 take first and second place. Regarding department, department 92 has the highest weekly sales. As the study expected, Holiday average sales are higher than on normal dates, they are most significantly higher on dates around Thanksgiving. Thirdly, Date also has some influence, but this effect can be regarded as

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328455

related to holidays or seasonal changes. The remaining relevant variables such as CPI and temperature have little impact on weekly sales. Among the models selected in this paper, the Multiple Linear Regression Model has the best prediction result, with an R2-Score of about 0.933 and without a significant difference between RMSE in the Training set and Test set.

Sales prediction provides real businesses with information analysis and market research. On this basis, one can make business decisions to avoid product redundancy or shortages. It also provides theoretical and data support for improving sales profit and reducing inventory waste.

The rest part of the paper is organized as follows. The Sec. II will introduce the methodology, which includes the provided data set, the models and evaluation indexes that have been used. The Sec. III will analyze the results of this research. Eventually, a brief summary will be given in Sec. IV.

2 METHODOLOGY

2.1 Data

The data includes 421,570 weekly sales in 143 weeks for Walmart stores of three different sizes from '05-Feb-2010' to '26-Oct-2012'. There are 45 stores and 81 departments in total. The dataset provides a total of 14 features, which are regarded as independent variables X and are defined as follows [3]:

- Store: The store ID
- Size: The store size. Stores can be divided into A, B and C types according to different sizes.
- Dept: Store department.
- Date: Specified date. In the process of data preparation, the Date was split into ‘week’, ‘month’ and ‘year’ to find the relationship between the date and weekly sales more clearly.
- Temperature: The average temperature of the area (unit: °F).
- Fuel Price: Fuel prices in the region.
- Markdown1-5: Anonymous discounted goods. Markdown data is only available after November 2011 and is not always applicable to all stores. Any missing values are marked with Null.
- CPI: Consumer price index.
- Unemployment: The ratio of unemployment.
- Is_Holiday: Special holidays instead of a common day. Only four holidays are included: Super Bowl, Labor Day, Thanksgiving and Christmas.

The dependent variable Y that the study wants to predict is ‘weekly sales’. Subsequently, the paper mainly aims to find the relationship between 14 variables and weekly sales. In order to obtain the correlation between each feature more clearly, this research first made the correlation matrix, as shown in Figure 1 [4]. Based on the results, some features (e.g., temperature, unemployment, and CPI) almost have no significant effect on weekly sales, so they have been removed and draw the correlation matrix with fewer related features. In addition, Markdown 4 and 5 are highly correlated with Markdown 1. To avoid multicollinearity problems, Markdown 4 and 5 will also be removed. Therefore, the new correlation matrix after removing irrelevant terms is obtained, as shown in Fig. 2. It is obvious that size and type are highly correlated with weekly sales. Also, department and store are correlated with weekly sales.

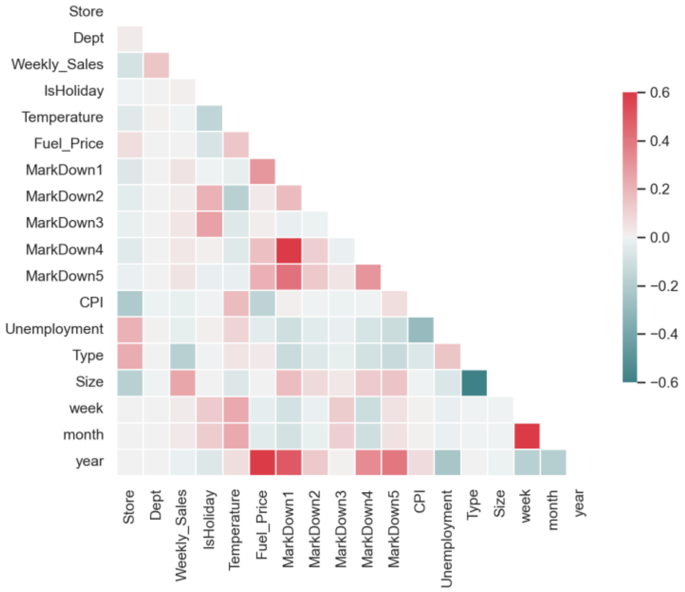

Figure 1. Correlation Matrix

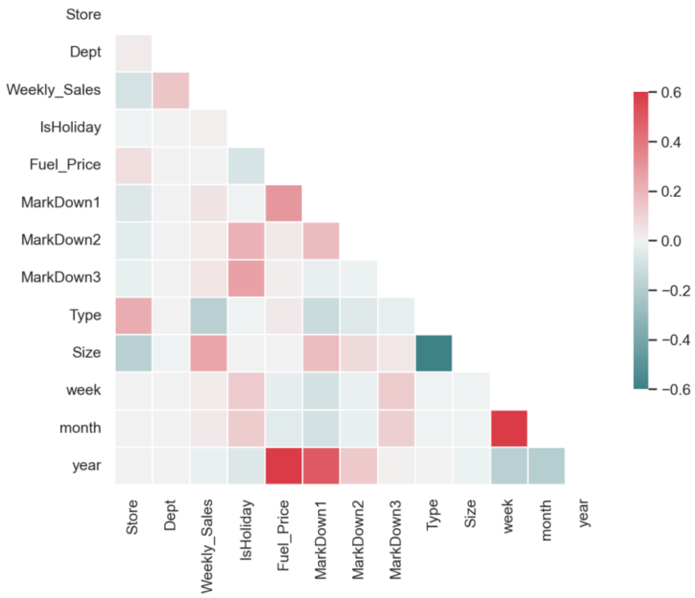

Figure 2. Correlation Matrix Without Irrelevant Terms

2.2 Model

Generally, the Multiple Linear Regression Model can be mathematically described as follows:

$$Y = \beta_0 + \sum_i \beta_i X_i \tag{1}$$

where Y is the dependent variable and Intercept, X_i is the Independent variable, and β_i are the coefficients. Elastic-Net combines Lasso and Ridge into a single model with two penalty factors: one proportional to the L1 norm and the other proportional to the L2 norm [5]. The resulting model using this approach is as sparse as Lasso Linear regression, but has the same regularization capability as ridge Linear regression provides [6]:

$$Y = \sum\left(\widehat{Yi} - Yi\right)^2 + \lambda\sum\beta^2 + \lambda\sum|\beta| \tag{2}$$

As for the polynomial regression model:

$$Y = \sum\nolimits_{i=0-N} \beta_i X^i \tag{3}$$

2.3 Metrics

To clearly determine which model works best, R2-Score and RMSE are used to measure the quality of the model. R2-Score, the coefficient of determination, reflects the proportion of the total variation of the dependent variable that can be explained by the independent variable through regression. The closer R2-Score is to 1, the better the model. Its formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\widehat{yi})^2}{\sum_{i=1}^{n}(yi-y^2)^2} \tag{4}$$

Here, yi represents the real observed value, y^2 represents the average value of the real observed value, and $\widehat{yi}$ represents the predicted value. Root Mean Square Error (RMSE) measures the deviation between the observed value and the real value, that is, to compare the training set and the test set, the formula is [7]:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(yi - \widehat{yi})^2} \tag{5}$$

It can be seen from Table 1 that the Polynomial Regression Model (PNR) has the highest R2-Score of about 0.979, while MLR and ENR only have 0.933 and 0.814 respectively, which means that PNR has the best understanding of the data set.

Table 1 Parameter estimation results

Models	**R2**	**Train-RMSE**	**Test-RMSE**
MLR	0.933	147439.805	1.544e+05
ENR	0.814	245518.948	2.509e+05
PNR	0.979	82175.399	2.701e+14

At the same time, in order to judge whether these models will cause over-fitting, another evaluation index Root Mean Squared Error (RMSE) is used in this paper. The model with the lowest RMSE value has the best performance. Although PNR performed slightly better than MLR in R2-Score comparisons, it does not look so good in RMSE comparisons. Apparently, Polynomial Regression is over-fitted, because the training set and test set have the greatest difference, as shown in Figure 3 below. Therefore, one can conclude that the simple Multiple Linear Regression Model (MLR) gives the best results by comparison.

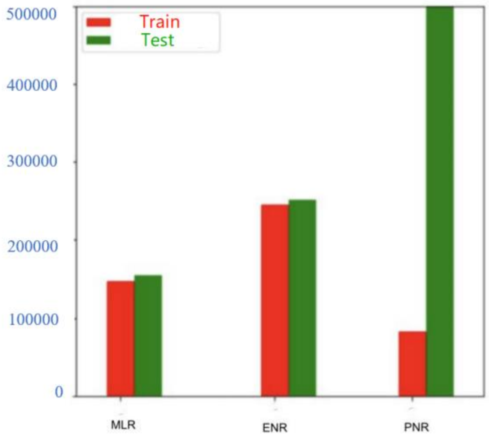

Figure 3. RMSE Comparison for Different Regression Models

3 RESULTS & DISCUSSION

Primarily, it was found that holiday average sales were higher than normal dates. However, not all holidays increase weekly sales. For example, Labor Day and Christmas do not. There is a positive effect on sales in the Super bowl, but Thanksgiving has the most significant influence on the improvement of weekly sales. That may be one of the reasons for the highest sales in December. These characteristics are presented in Figure 4.

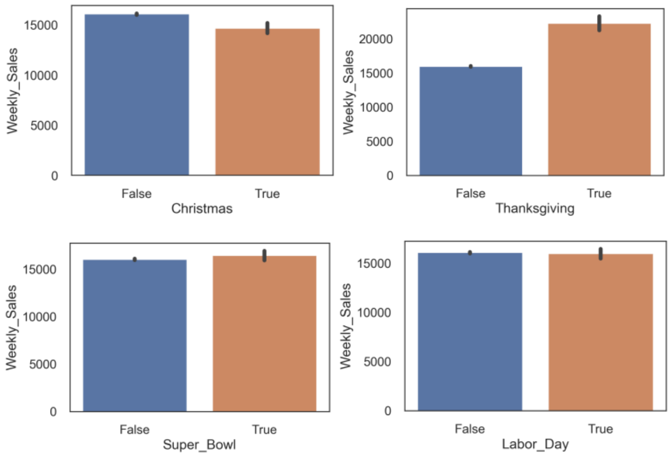

Figure 4. Holiday Effect

Seen from Fig. 5, the highest weekly sales are in the Thanksgiving week between holidays. Meanwhile, for all holidays Type A stores has the highest sales. Stores are divided into 3 types A, B and C according to their sizes. The size of the stores is consistent with sales, as expected. Almost half of the stores are bigger than 150000 and categorized as type ‘A’. According to type, sales of the stores are changing. With the decreased size of A, B and C types, the weekly sales become lower, too. Weekly sales in the past three years are generally very close. There is no data for the last two months in 2012. If added the two months' weekly sales, 2012 may take the first place among the past three years. The weekly sales in December are the highest, while January sales are significantly less than in other months. This is the result of November and December’s high sales. After two high sales months, people prefer to pay less in January. It is obvious that the 47th week and 51st weeks have significantly higher averages as Christmas, Thanksgiving and Black Friday effects, and the 49th and 50th weeks follow them. Interestingly, the fifth top sales belong to the 22nd week of the year. This result shows that Christmas, Thanksgiving and Black Friday are very important than other weeks for sales and the fifth most important time is 22nd week of the year and it is the end of May when schools are closed. Most probably, people are preparing for the holiday at the end of May.

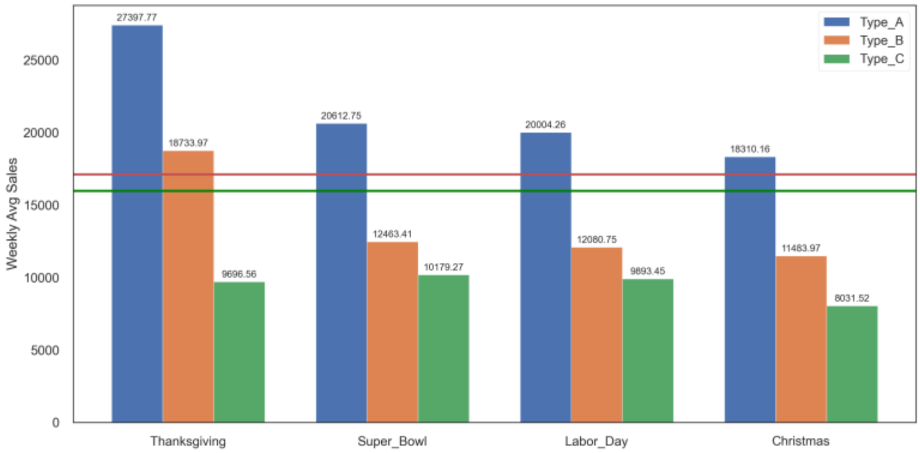

Figure 5. Holiday-Type Effect

In the total sales graph department 72 has higher weekly sales values, but in the average graph, department 92 is the best. It shows us, that some departments have higher values as seasonal like Thanksgiving. It is consistent when we look at the top 5 sales in data, all of them belong to department 72 at Thanksgiving holiday time. Although stores 10 and 35 have higher weekly sales values sometimes, in general, average store 20 and store 4 are in the first and second rank. It means that some areas have higher seasonal sales. CPI, temperature, unemployment rate, and fuel price have no pattern on weekly sales.

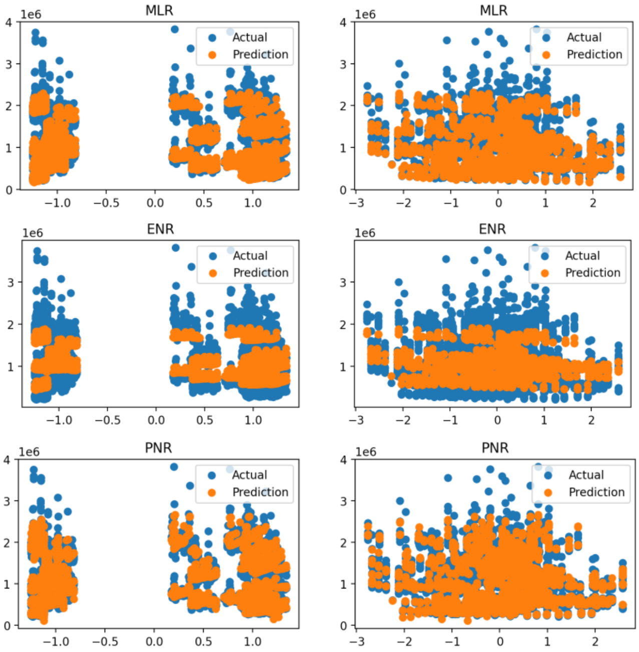

Figure 6. Comparison of Prediction Accuracy

As for the prediction results of the three models, in the second part about metric, this research makes a conclusion that the MLR and PNR are more accurate, with the R2-score approximately 0.933 and 0.979 respectively. Besides, MLR is selected for prediction according to the better performance in RMSE, as shown in Fig. 6 [8].

There are some limitations to this analysis. First, the influence of 14 variables on weekly sales will be inaccurate because of multicollinearity, which will lead to the increase of variance of least-squares estimation or sensitivity to sample changes, thus reducing test reliability. This research removed some significant multicollinearity variables. In the future, the Principal Component Analysis (PCA) method and Ridge Regression model can be added to improve the study. Furthermore, it is not enough to use linear regression to predict the trend of sales volume by reading other relevant sales forecasting literature. Stacking is a method of combining models that often get more accurate predictions than a single model. In the prediction model of Machine Learning, there are other commonly used algorithms, including Neural Network, Random Forest, XGBoost, and the prediction method based on time series, such as ARIMA, SARIMA etc. [9].

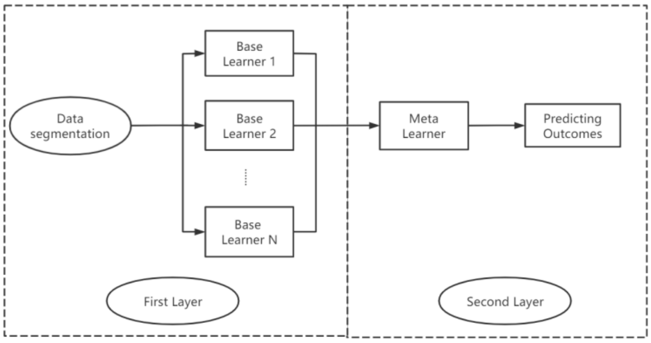

Figure 7. Two-layer Stacking Model

Therefore, it can be assumed that ARIMA, Random Forest, and BP Neural Network are used as the base learner to construct the Stacking prediction model. The algorithm fully considers different models and combines the advantages of different algorithms. The periodicity of data is fitted by using the time series prediction algorithm, the nonlinearity of data is fitted by Stacking Random Forest and BP Neural Network, and the strengths of three heterogeneous algorithms are combined to complement each other. The original linear regression algorithm is used in the second layer of Meta learner to ensure stability and generalization performance [10]. The structure is shown in Figure 7. It is speculated that with stacking the model, the prediction indexes will be lower than the individual indexes, enabling the model to fully incorporate the advantages of the three base learners to make better predictions.

4 CONCLUSION

In summary, this paper investigates the sales prediction of Walmart based on different Linear regression models. Firstly, the initial data set provided by the Kaggle platform has been cleaned. Subsequently, through the comparison of separate features, it is found that Size, Date, Store and other features have different effects on weekly sales. Afterwards, three different regression models are used to make sales predictions and found that with the more complex Elastic-Net Regression Model and Polynomial Regression Model, the Multiple Linear Regression Model performed better, which has the moderate R2 values and the RMSE with the smallest difference between the training and test sets.

In the future, more predictive models will be tried and apply the new models for other sales problems. This research has profound implications for the inventory preparation and sales planning of retailers. For companies and managers, it is recommended to deeply understand the conclusions based on experts in identifying potential sales problems and make their employees fully understand the market needs, and integrate their knowledge into the decision-making process. Overall, these results offer a guideline for sales prediction questions to choose appropriate models as well as provide suggestions to Walmart and similar companies to develop sales strategies according to influencing factors.

REFERENCES

[1] S. Kohli, G. T. Godwin, and S. Urolagin. "Sales prediction using linear and KNN regression." Advances in Machine Learning and Computational Intelligence. Springer, Singapore, vol. 1, 2021, pp. 321-329.

[2] A. Ahmedov, Walmart Sales Forecasting, 2022. [online] Kaggle.com. Available at: <https://www.kaggle.com/code/aslanahmedov/walmart-sales-forecasting/notebook#Markdown-Columns> [Accessed 8 May 2022].

[3] D. Thakur, Walmart Sales Prediction, 2022. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/divyajeetthakur/walmart-sales-prediction> [Accessed 10 May 2022].

[4] F. Lasso, Cracking the Walmart Sales Forecasting challenge, 2022. [online] Kaggle.com. Available at: <https://www.kaggle.com/code/fernandol/cracking-the-walmart-sales-forecasting-challenge> [Accessed 10 May 2022].

[5] M. Schmidt, "Least squares optimization with L1-norm regularization." CS542B Project Report vol. 504, 2005, pp. 195-221.

[6] G. C. McDonald, "Ridge regression." Wiley Interdisciplinary Reviews: Computational Statistics vol. 1.1, 2009, pp. 93-100.

[7] D. Chicco, M. J. Warrens, and G. Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science vol. 7, 2021, e623.

[8] Kaggle.com. 2022. Walmart Sales Prediction - (Best ML Algorithms). [online] Available at: <https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms> [Accessed 10 May 2022].

[9] S. Cheriyan, et al. "Intelligent sales prediction using machine learning techniques." 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 2018.

[10] H. Wang, and C. Li, "Stacking integrated learning methods in Sales Forecasting." Computer Applications & Software vol. 8, 2020.