Weekly Sales Prediction of Walmart Stores Based on Three Regression Models

Tianyang Lu *

Department of mathematics, Shandong University, Jinan, China

* Corresponding Author Email: 202023092010@mail.sdu.edu.cn

Abstract. In the information era, the sales of a supermarket are influenced by a lot of factors. It is very important for a supermarket to predict its expected sales according to several given data about the factors and then make corresponding plans. Regression method is a common method to do the prediction and it applies the known data to set up regression models to predict the future data according to some given information. This study will try to find out the relationship between the factors and the weekly sales of two Walmart stores and try to use the model to predict its future sales by using three regression models, the linear regression model, the elastic net model and the SVR model. Then, the paper will divide the raw data into a training set and a testing set and build the model with the data in the training set and find out the best one by calculating the MAPE of the testing set for every model. According to the analysis, three SVR models and one elastic net regression model perform best in four situations of the prediction and do a simple analysis to every factor about their negative or positive impact on the weekly sales and their weight. These results will offer a prediction about the future sales of the two Walmart stores for the market and can be used as a reference for both the management of the market and the consumers.

Keywords: Sales prediction; machine learning; elastic net regression; support vector regression.

1. Introduction

The prediction of the future sale of a supermarket plays a very important role for the management of the supermarket to make future plans. In the modern era, a lot of factors such as the weather, the CPI and the unemployment will influence the sales of a supermarket. So, when the information is given, it is very necessary to find out a model to predict the sales. In order to do the prediction, a lot of models are put forward. The linear regression model is a basic model for the prediction [1] and it can be used to determine the importance of every factor. The Lasso regression [2] and the ridge regression [3] apply the method of regularization to improve the fitting effect on the basis of the linear regression. The elastic net regression is a generalization and a combination of the Lasso regression, and the ridge regression as mentioned in [4]. Except the linear regression, some nonlinear methods can also be adapted for regression with better fitting effect such as the Support Vector Regression (SVR) put forward in [5]. These models can correct the error of the linear model because sometimes the MAPE or the R-Square value of the linear model is too large, so it is not so precise.

The prediction is helpful because they can analyze the importance of every influence factor and pay special attention to it, especially for the linear models it is very obvious to do this. It can also help the management to predict future work according to given information and make plans accordingly. Especially in the information era, the prediction regression method is widely used in lots of fields and find an efficient model is important to reply to the requirement. Sometimes a better model can increase great profit to a company and reduce the economic loss of a disaster. For example, in the medical field, previous scholars apply the multiple linear regression model to predict new active cases of COVID-19 pandemic [6] and apply the elastic net model to do some research about human cancers [7]. The application of nonlinear models can be found in Ref. [8] which uses M5 model tree and [9], which uses artificial neural networks (ANN). Although the simple linear regression model is imprecise, some models are based on the linear regression, and they are widely used. Two examples of the application of the principal component analysis with linear regression models [10, 11].

This paper will try to set up regression models for two Walmart stores' weekly sales according to four influence factors and do some analysis about the result. Two linear models, the linear regression and elastic net regression and a nonlinear model, the supporting vector regression will be used to set up models and find out the important factor which will influence the sales greatly. It can offer suggestions for the management of the two Walmart stores and can also give references to the analyst of the market sales. The Sec. 2 will provide a simple introduction to the three models and their parameters which can be an indicator to the result. Next, the Sec. 3 will show the results of the three models for both Walmart stores. It should be noted that this paper sets up separate models for two Walmart stores when it is on or off holiday because the influence of the four factors may change. The Sec. 4 will show an analysis to the final result and find out the best model for every case. Subsequently, the Sec. 5 will state its limitations and offer some suggestions for the future work which can improve the paper's work. Eventually, a brief summary will be given in Sec. 6.

2. Data and Method

The data is from the weekly sales of two Walmart stores from May 2,2010 to October 26, 2011. The weekly sales were influenced by the average local temperature for that week, the cost of fuel in the region in given week, the consumer price index (CPI) in given week, the local unemployment rate in given week and whether or not a holiday occurred in given week. Next, the paper will introduce three regression models to do the prediction. The paper will set up separate models for the weekly sales on and off holiday for every store and analyze the impact of the factors to the weekly sales in R. The paper set up the models with the data in the training set and evaluate the quality of the model based on the mean absolute percentage error (MAPE) of the testing set. The ratio of the data in training set and testing set is 70:30. The formula to calculate MAPE is shown in Eq. (1):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y_i}}{y_i}\right| \quad (1)$$

For the multivariate linear regression, the model assumes that there exists a linear relationship between the weekly sales and the factors. Thus, in this model the aim is to find the coefficients of the linear formula. Some factors can be omitted in the final model so different models can be built with different combination of the factors. The paper will compare their MAPE to find out the best one. The aim of the SVM regression method is to find out a gap from the regression line and minimize the sum of the square of the distance to the area of the points outside the gap. The parameter ε is the width of the gap and the parameter γ is the coefficient of the kernel function. In the model in this paper the function is the radial function. The elastic net regression is a combination of the Lasso regression and the ridge regression. The aim of the regression method is to find the proper value of α, λ and the vector β to minimize Eq. (2):

$$(Y - X\beta)^T(Y - X\beta) + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2) \quad (2)$$

3. Results

The paper will set up these three models to every condition and choose a best model with the least MAPE. For the first Walmart store, when there was no holiday in the week, the paper drew the Cp plot for all subsets regression in R and get Fig. 1. According to the graph, set up the three possible models listed in Table 1. So, the T-F-C-U model is the best one and the regression formula is shown in Eq. (3):

$$Weekly\_Sale = -2040448.3 - 1054.9Temperature - 7956.6Fuel_{Price} + 14769.4 * CPI + 68244.2Unemployment \quad (3)$$

For the SVR model, the paper found out the best value of γ as 0.25 and ε as 0.1 by R. The MAPE is 0.1280575 and the predict data is shown in Fig. 2. For the elastic net model, the paper found out the best value of α as 0.25 and λ as 33093.62 by R. The MAPE is 0.1327138 and the predict data is shown in Fig. 3. The regression formula of the elastic net model is shown in Eq. (4):

$$Weekly\_Sale = -365402.2548 - 752.2823Temperature + 16204.3166Fuel\_Price + 5576.6955CPI \quad (4)$$

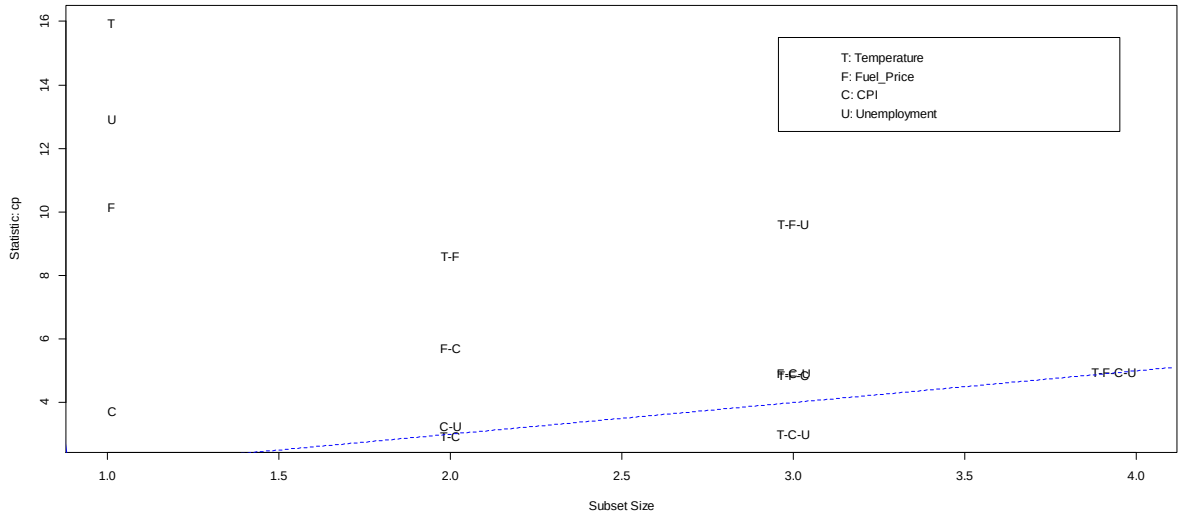

Fig. 1 the Cp plot for store 1 off holiday

Table 1. The coefficients and the MAPE of three linear regression models for store 1 off holiday.

Model	Regression Coefficients					MAPE
	Temperature	Fuel_Price	CPI	Unemployment	Intercept	
C-U	0	0	14192	73460	-2052620	0.134989
T-C	-1212	0	9265	0	-347656	0.1302983
T-F-C-U	-1054.9	-7956.6	14769.4	68244.2	-2040448.3	0.1300908

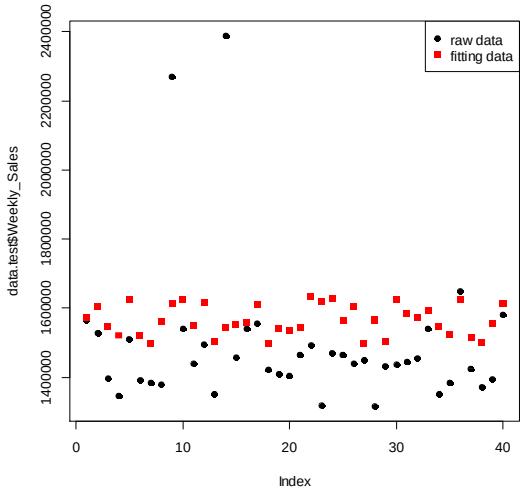

Fig. 2 The fitting data of the SVR model for store 1 off holiday

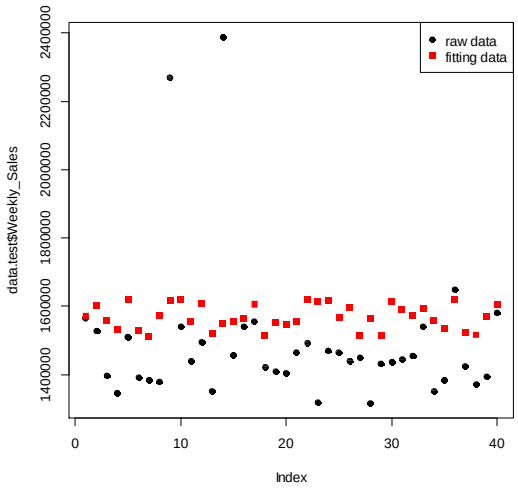

Fig. 3 The fitting data of the elastic net model for store 1 off holiday

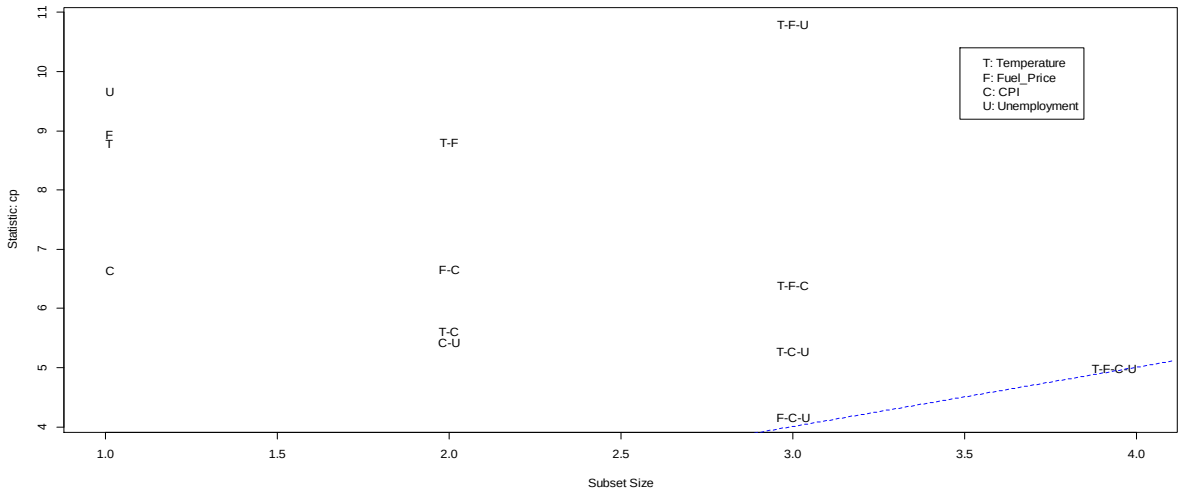

Fig. 4 the Cp plot for store 1 on holiday

Table 2. The coefficients and the MAPE of three linear regression models for store 1 on holiday

Model	Regression Coefficients					MAPE
	Temperature	Fuel_Price	CPI	Unemployment	Intercept	
F-C-U	0	-423369	90372	428897	-19795936	0.2167581
T-F-C-U	-3064	-363991	86434	382133	-18588583	0.2354513

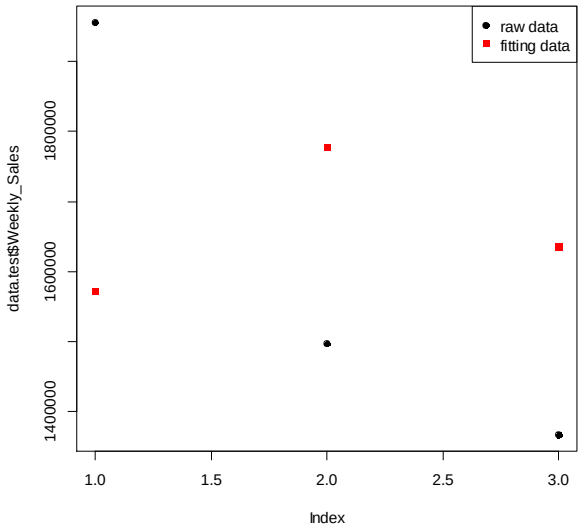

Fig. 5 The fitting data of the SVR model for store 1 on holiday

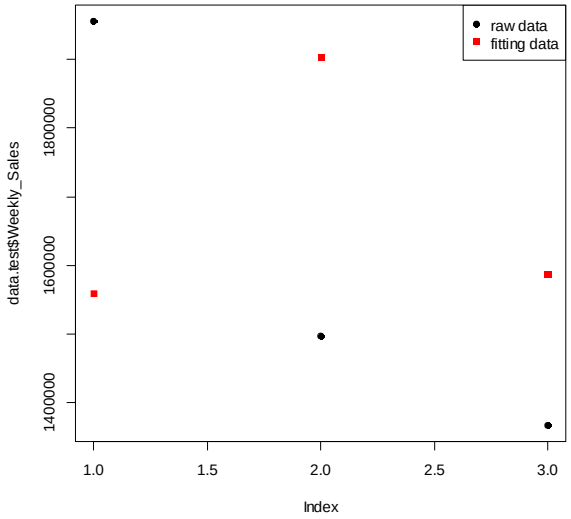

Fig. 6 The fitting data of the elastic net model for store 1 on holiday.

For the first Walmart store, when there was no holiday in the week, the paper drew the Cp plot for all subsets regression in R and get Fig. 4. According to the graph, set up the three possible models as follows in Table 2. So, the F-C-U model is the best one and the formula is shown in Eq. (5):

$$Weekly\_Sale = -19795936 - 423369Fuel\_Price + 90372CPI + 428897Unemployment \quad (5)$$

For the SVR model, the paper found out the best value of γ as 0.5 and ε as 0.1 by R. The MAPE is 0.1934705 and the predict data is shown in Fig. 5. For the elastic net model, the paper found out the best value of α as 1 and λ as 13206.89 by R. The MAPE is 0.2110606 and the predict data is shown in Fig. 6. The regression formula of the elastic net model is shown in Eq. (6):

$$Weekly\_Sale = -7409053.249 - 3120.758Temperature - 48451.796Fuel\_Price + 37967.042CPI + 161062.982Unemployment \quad (6)$$

Next, the paper will consider the second Walmart store. For the second Walmart store, when there was no holiday in the week, the paper drew the Cp plot for all subsets regression in R and get Fig. 7.

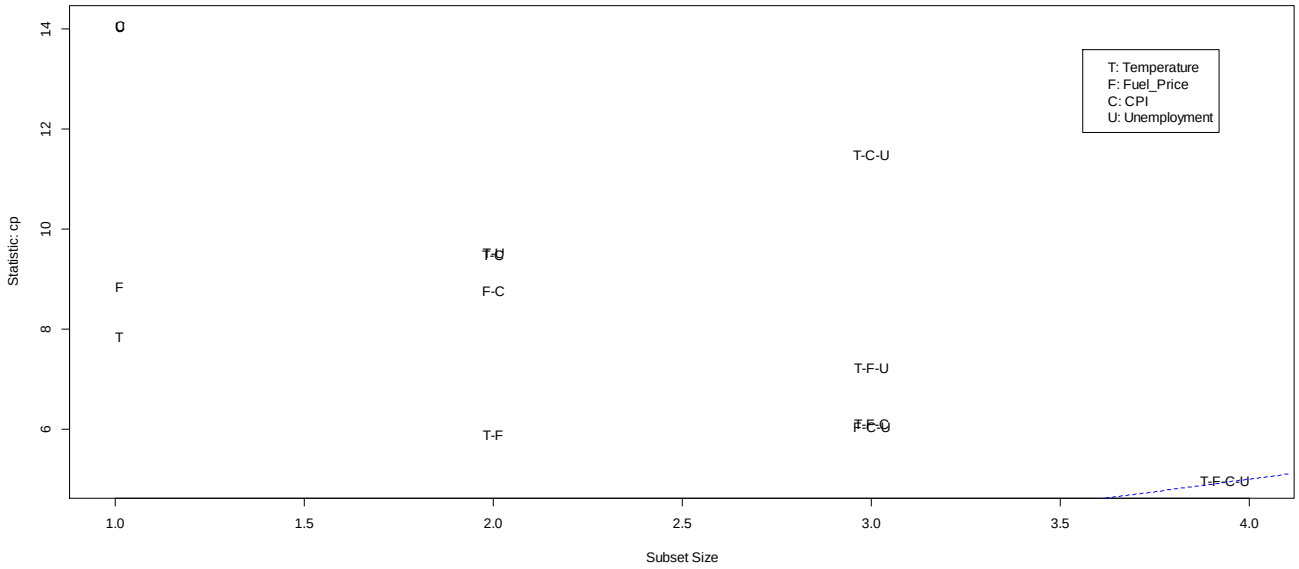

Fig. 7 The Cp plot for store 2 off holiday.

According to the graph, the T-F-C-U model is the best one and the regression formula is shown in Eq. (7), and its MAPE is 0.1170653:

$$Weekly\_Sale = -7180413.5 - 1715.4Temperature - 192011.6Fuel\_Price + 38766.0CPI + 195085.3Unemployment \quad (7)$$

For the SVR model, the paper found out the best value of γ as 0.25 and ε as 0.1 by R. The MAPE is 0.1093205 and the predict data is shown in Fig. 8:

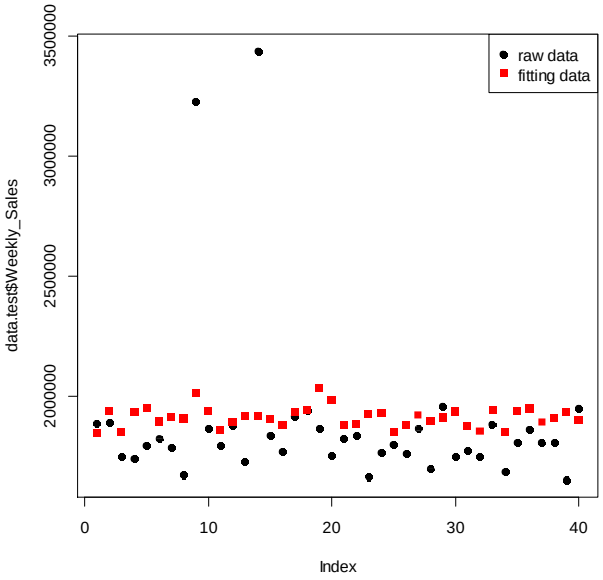

Fig. 8 The fitting data of the SVR model for store 2 off holiday

For the elastic net model, the paper found out the best value of α as 0.1 and λ as 2688.943 by R. The MAPE is 0.4294391 and the predict data is shown in Fig. 9. The regression formula of the elastic net model is shown in Eq. (8):

$$Weekly\_Sale = -2103537.429 - 1901.148Temperature - 137072.586Fuel\_Price + 18527.330CPI + 79896.955Unemployment \quad (8)$$

For the second Walmart store, when there was on holiday in the week, the paper drew the Cp plot for all subsets regression in R and get the Fig. 10. According to the graph, set up the three possible models as shown in Table 3.

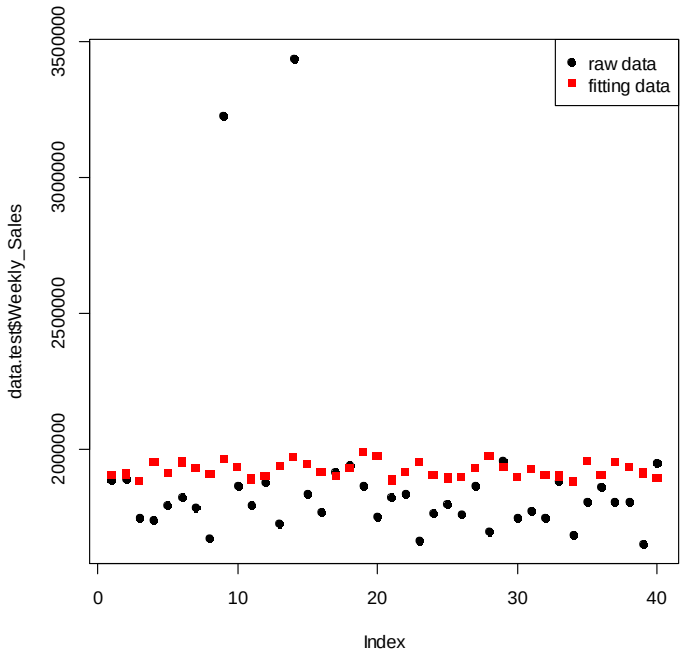

Fig. 9 The fitting data of the elastic net model for store 2 off holiday

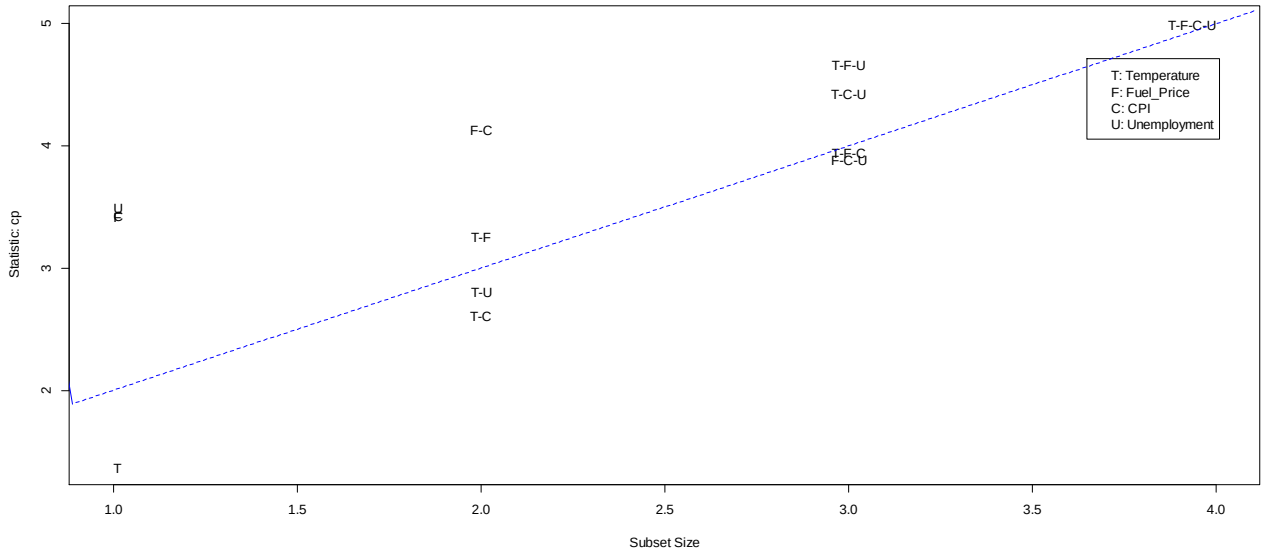

Fig. 10 the Cp plot for store 2 on holiday.

Table 3. the coefficients and the MAPE of three linear regression models for store 1 on holiday

Model	Regression Coefficients					MAPE
	Temperature	Fuel_Price	CPI	Unemployment	Intercept	
F-C-U	0	-954588	279350	1318751	-65243090	0.2677571
T-F-C	-8439	-431192	62036	0	-9446821	0.2293035
T-F-C-U	-5921	-739633	217758	944114	-49420741	0.2559335

So, the T-F-C model is the best one and the regression formula is shown in Eq. (9):

$$Weekly\_Sale = -9446821 - 8439Temperature - 431192Fuel\_Price + 62036CPI \quad (9)$$

For the SVR model, the paper found out the best value of γ as 0.5 and ε as 0.1 by R. The MAPE is 0.1977356 and the predict data is shown in Fig. 11. For the elastic net model, the paper found out the best value of α as 0.7 and λ as 132137.1 by R. The MAPE is 0.184584 and the predict data is shown in Fig. 12. The regression formula of the elastic net model is shown in Eq. (10):

$$Weekly\_Sale = -2229250.852 - 2609.948Temperature \quad (10)$$

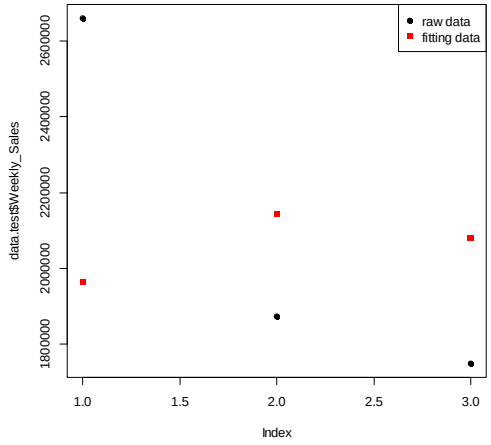

Fig. 11 The fitting data of the SVR model for store 2 on holiday.

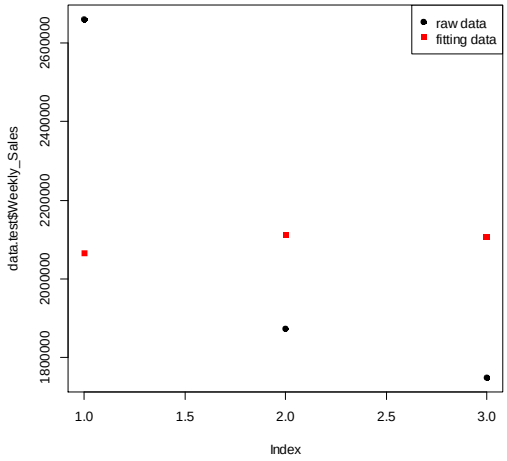

Fig. 12 The fitting data of the elastic net model for store 2 on holiday.

4. Discussion

From the results above, the best models can be found as shown in Table 4. it can be seen that the SVR is the most widely used model while the linear regression is relatively the coarsest model. The value 0.1 is the best value of the gap and for both Walmart stores. For the first Walmart store, the coefficient of the kernel function of the model on holiday is larger than that off holiday. The parameters of the two SVR models for the two Walmart stores off holiday is the same, which means the performance of the four factors off holiday is similar to some extent. For the elastic net model of the second Walmart store on holiday, it shows that the temperature is an important factor for the weekly sales and when the temperature is higher the weekly sales is fewer. For the two linear models, including the linear regression and the elastic net regression, the coefficients of these four factors can be seen as weight so the importance of the four factors is different. The coefficients are shown in Table 5. From this table, it can be concluded that for the first Walmart store, when it is off holiday, there is a negative correlation between the temperature and the weekly sales and a positive relationship between the CPI and the weekly sales. For the first Walmart store on holiday, there is a negative relationship between the fuel price and the weekly sales while there is a positive relationship between the CPI and the weekly sales and a positive relationship between the unemployment rate and the weekly sales. The unemployment rate is more important than the CPI. For the second Walmart store, when it is off holiday, if the temperature and the fuel price is higher the weekly sales are fewer, and the weight of the fuel price is higher. However, if the CPI and the unemployment rate is higher than the weekly sales are higher, and the weight of the unemployment rate is higher. For the second Walmart on holiday, it can be found that there exists a negative relationship between the temperature and the weekly sales. From the prediction of the test data, it can be seen that the weekly sales of the 9th week and the 14th week of both the two Walmart off holiday are two outliners, much higher than the prediction. It can be checked whether the record was wrong or there were some other reasons for these abnormal weekly sales such as a shopping festival.

Table 4. The best model

Store	On/off holiday	Best model	Parameters
1	Off	SVR	γ=0.25, ε=0.10
	On	SVR	γ=0.50, ε=0.10
2	Off	SVR	γ=0.25, ε=0.10
	On	Elastic Net	α=0.7, λ=132137.1

Table 5. The coefficients of the linear model and the elastic net model

Store	On/off holiday	model	Regression Coefficients			
			Temperature	Fuel_Price	CPI	Unemployment
1	Off	Linear	-1054.9	-7956.6	14769.4	68244.2
		Elastic Net	-752.2823	16204.3166	5576.6955	0
	On	Linear	0	-423369	90372	428897
		Elastic Net	-3120.758	-48451.796	37967.042	161062.982
2	Off	Linear	-1715.4	-192011.6	38766.0	195085.3
		Elastic Net	-1901.148	-137072.58	18527.330	79896.955
	On	Linear	-8439	-431192	62036	0
		Elastic Net	-2609.948	0	0	0

5. Limitations and Prospects

This paper only applies three models for the regression, so it is possible to find a better model other than these three models. On some occasions, the results given by different models is very different so if another model can be applied, maybe a more comprehensive analysis can be done about the factors. As seen in Table 5, the MAPE of some models is not very low which shows there are better models. Moreover, the size of the sample is not so big because the paper only chooses two Walmart markets to analyze their data. If the information about the weekly sales of more Walmart stores in the same region can be given, a more generalized conclusion about the four factors' influence to the weekly sales can be drawn. Finally, the paper only chooses the MAPE as the criterion to assess the fitting effect of a regression model. However, there are some other indicators such as MSE or R-squared which can also show the efficiency of a certain model and they are not applied in this paper.

In the future, in order to make up for these limitations in this paper, first more regression models especially nonlinear regression models can be applied for prediction. There are some other regression methods other than the SVR in machine learning such as random forests and decision tree. These regression models can also be used for prediction according to the given data. Besides, more indicators can also be used to evaluate a model. By combining the different indicators of a model, a more overall evaluation can reflect the efficiency of the model better. Different indicators' focus on a model is different and they can show different aspects of a model. Finally, for the elastic net model and the SVR model in this paper, the selection of the parameters is not perfect because the paper tries to get the best parameters by trying to set up models with some groups of given parameters and choose a best one. Future researchers can try to find out an algorithm to determine the best parameters according to the raw data instead of exhaustion.

6. Conclusion

In summary, by setting up regression models for the two Walmart stores, the paper finds out an algorithm for the Walmart stores to predict the future sales if the information of the four factors is given. Three regression models including the linear regression model, the elastic net model and the SVR model are used to set up models and the MAPE is used as the indicator of the fitting effect of a certain model. Finally, by comparing the MAPE the paper finds out four best models for the two Walmart stores when it is on or off holiday separately. Among the four models, three are the SVR model and one is the elastic net model. By analyzing the linear models, the paper also finds some factors which have a positive correlation with the weekly sales and factors which have a negative correlation with the weekly sales and factors and compare their impact to the weekly sales. However, the paper only considers three models for the regression so it is possible that some other models with better fitting effect are omitted, and some other indicators of the fitting effect are not used in this paper, which could be included in future works. These results show an example of the application of regression to the market sales, and it can be used as a reference for the management of the Walmart

stores to make future arrangements accordingly and the market analysts to predict the influence of the four factors to the market.

References

[1] Olive D J, Olive D J. Multiple linear regression. Springer International Publishing, 2017.

[2] Chan-Lau M J A. Lasso regressions and forecasting models in applied stress testing. International Monetary Fund, 2017.

[3] Saleh A K M E, Arashi M, Kibria B M G. Theory of ridge regression estimation with applications. John Wiley & Sons, 2019.

[4] Mouret G, Brault J J, Nia V P. Generalized elastic net regression. GERAD HEC Montréal, 2013.

[5] Support vector machines: theory and applications. Springer Science & Business Media, 2005.

[6] Rath S, Tripathy A, Tripathy A R. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & metabolic syndrome: clinical research & reviews, 2020, 14(5): 1467-1474.

[7] Das J, Gayvert K M, Bunea F, et al. ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers. BMC genomics, 2015, 16(1): 1-13.

[8] Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W. Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal, 2014, 2014.

[9] Behzadi M R B, Aslaminejad A A. A comparison of neural network and nonlinear regression predictions of sheep growth. Journal of Animal and Veterinary Advances, 2010, 9(16): 2128-2131.

[10] Bakar N M A, Tahir I M. Applying multiple linear regression and neural network to predict bank performance. International Business Research, 2009, 2(4): 176-183.

[11] Çamdevýren H, Demýr N, Kanik A, et al. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecological Modelling, 2005, 181(4): 581-589.