Weekly Sales Prediction for One of the Largest Retailer “Walmart”

Samina Islam Eva, Souradip Dey, Md Rubayatur Bhuyian
Department of Electrical & Computer Engineering
University of Waterloo
sislamev@uwaterloo.ca, s25dey@uwaterloo.ca, mrrbhuyian@uwaterloo.ca

Abstract-

Time series modeling collects and studies the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series analysis and its applications has significant importance in different areas and domains. Predicting the sales of any business is one of them. Big fishes in the business industry always try to have prior knowledge about their sales and demand of the products. In this paper, an approach has been made to predict the store wise weekly sales of Walmart having 45 different stores located in different regions. Several algorithms were implemented such as liner regression (LM), Neural Network (NN); Random Forest Regression (RF) to predict the price of each store and a comparison has been shown between each model to find the optimum one.

***Keywords*— Machine learning; linear Regression (LM); Neural Network (NN); Random Forest Regression (RF)**

I. Introduction

Forecast of sales is very important for companies because they need to plan based on the forecast. This paper shows how different Regression models can be used to forecast company sales. A model is used to forecast one year ahead of the total sale from a samples available from.

While the Forecasting System can automate the analysis of data, it is also required to figure out the model which can find the most effective forecasts. The root mean square error (RMSE) of the price variation are selected as the performance index. Testing and validation scores for different algorithms implemented in this paper are compared to find the best algorithm to find the optimum model for the given problem.

II. Data set

Walmart’s dataset from kaggle (link: https://www.kaggle.com/c/walmartrecruiting-store-sales-forecasting/data) has been used. There are three datasets named train.csv, store.csv and features.csv. These datasets contains the following information.

stores.csv: This file contains anonymized information about the 45 stores, indicating the type and size of store.

train.csv: This is the historical training data, which covers sales from 2010-02-05 to 2012-11-01. There are several other fields as follows: • Store - Store number • Dept - Department number • Date - Date • Weekly_Sales - Sales for the given department in the given store • IsHoliday - whether the week is a special holiday week or not.

features.csv: This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields: • Store - Store number • Date - Date • Temperature - Average temperature in the region • Fuel_Price - Cost of fuel in the region • MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data are only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA. • CPI - the consumer price index •Unemployment - the unemployment rate • IsHoliday - whether the week is a special holiday week or not.

These three data sets have been processed and merged together named as input_data_team03.csv to perform further analysis.

III. Data Preparation

The attributes are scattered and needed to be merged both for training and test purpose to build a successive model. The weeks are being extracted from the date so that the sales can be easily identified from the week number also it helps to run the model smoothly. Some weekly sales were negative which indicates that there were more return than sales for that particular department in that week. We aggregated the weekly sales data and built our forecasting model on that dataset. Moreover the missing vales (NA) in markdown features has been imputed with zero. The correlation between independent and dependent variable has been done to find out the most significant attributes which contains valuable information to predict the sales. As an example the store size is positively highly correlated with sales. Each store has 143 weeks of observation among which first 110 observation used to train different models and rest 33 were used as test set.

IV. Data Visualization

Before proceeding further with the processing of data it is very important to visualize the data also understand the relationship between different features and target variable. The distribution of sales price for store 1 is shown below.

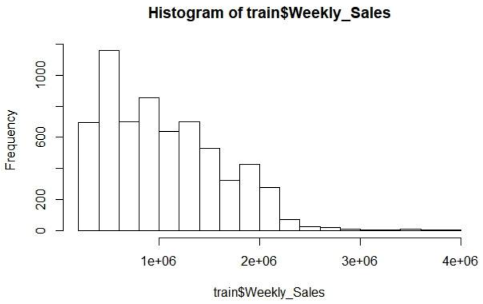

Figure 1. Distribution of SalesPrice of Store 1

Correlation with different features with target variable is shown in the below figure.

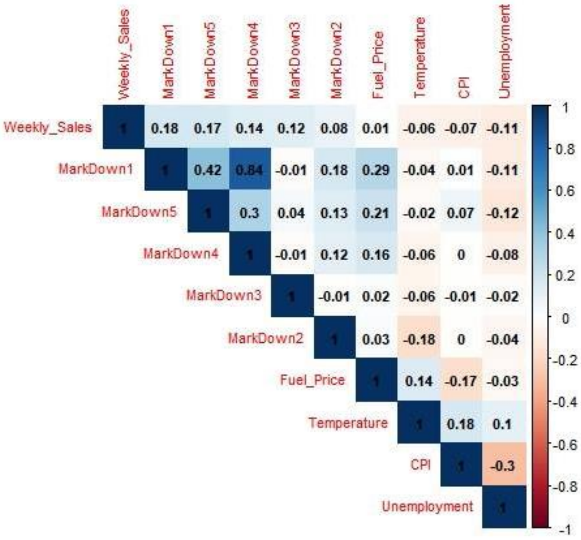

Figure 2. Correlation Matrix between different features of Store 1

V. Models

A. *Linear Regression (LR)*

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables x_1, x_2, ..., x_p is defined to be

$$y_i = ß_0+ß_1x_{i1}+ß_1x_{i1}+\ldots+ß_px_{ip}+\varepsilon_i \text{ for } i=1, 2,\ldots..n$$

Function Used: lm ()

For the 1st store we obtained following results after training the model on 1 to 110 observation and later tested the model on 111 to 143 observation.

Table 1. Train & Test RMSE for LM

Description	**Train**	**Test**
RMSE	140639.3	124939.5
R^2	.3007	-

B. *Random Forest Regression (RF)*

Random forest is a supervised learning algorithm. It creates a forest and makes it random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. A decision tree is built top down from a root node and involves portioning the data into subset that contain instances with similar values. On the other hand in a regression tree since the target variable is a real valued number it fits a regression model to the target variable using each of the independent variables. Then for each independent variable the data is split at several split points. Sum squared error is being calculated at each split point between the predicted value and the actual values. The variable resulting in minimum SSE is selected for the node.

Function: randomForest()

Parameters:
ntree: Number of trees to grow
mtry: Number of variables randomly sampled as candidates

Table 2. Train & Test RMSE for RF

Parameters		RMSE Train	RMSE Test
ntree	mtry		
1000	5	117850.1	74873.84
300	5	120060	74346.05
100	7	105957.9	71970.66
50	7	116726.8	74863.65

From above we can see that ntree=100 and mtry=7 gives lowest RMSE for test set.

C. *Auto Regressive Integrated Moving Average (ARIMA)*

Auto regressive moving average is one of the most commonly used model for time series data analysis. ARIMA models provide another approach to time series forecasting. It relates the present value of a series to past values and past prediction errors. The general equation of an ARIMA model is

After preparing our data it has been observed that there are total 143 weeks of sales are formed per store. So an ARIMA model has been built for store 1 first and it will be applied on other 44 store. During the selection of the ARIMA model several combination of parameters have been tried.

Function: ARIMA()

We varied different values of p & q based on the ACF and PACF plot. We have calculated log likelihood, AIC$_C$ and RMSE for different combination which are stated below:

Table 3. Train & Test RMSE for ARIMA

(p,d,q) (P,D,Q)$_{[52]}$	Log likelihood	AIC$_C$	RMSE (Train Set)
(0,0,0) (0,1,0)	90.64	-179.21	59323.59
(0,1,0) (0,1,0)	90.04	-178	58645.31
(1,0,1) (0,1,0)	105.72	-204.99	48303.89
(1,0,1) (0,1,1)	106.39	-204.03	35975.49
(1,1,1) (0,1,0)	103.93	-201.41	48322.89

Our main goal for ARIMA model is to maximize the log likelihood and minimize the AIC$_C$ value. Also we have to consider no of parameters such that it does not over fit our model. After playing with different parameters we have decided to with ARIMA(1,0,1)(0,1,1) with RMSE(Test Set) value of 71626.85.

D. Neural Network

Neural network maps one set of continuous inputs to another set of continuous outputs. Attributes will be fed into each node of the last hidden layer, and each will be multiplied by a corresponding weight. The sum of those products is added to a bias and fed into an activation function. Activation function can both be linear, sigmoid or RELU (rectified linear unit), RELU is commonly used and highly useful because it doesn't saturate on shallow gradients as sigmoid activation functions do. For each hidden node the activation function outputs an activation and the activations are summed going into the output node, which simply passes the activations sum through. So a neural network performing regression will have one output node, and that node will just multiply the sum of the previous layer's activations by 1. The result will be the network's estimate, the dependent variable that all attributes map to. To perform back propagation and make the network learn it simply compares the estimation to the observation and adjust the weights and biases of the network until error is minimized. Here the loss function is RMSE (root mean squared error). It can also be used for time series prediction.

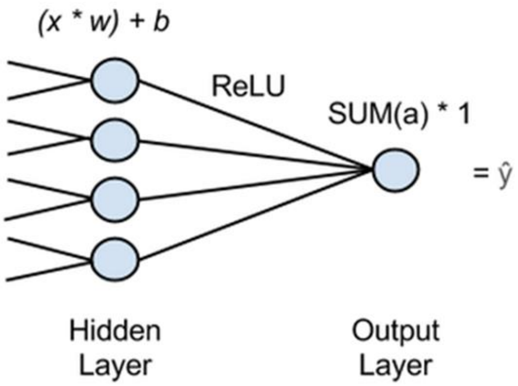

Figure 3. Regression with Neural Network

Function: nnetar() having several parameters like number of seasonal lags used as input, number of nodes in the hidden layer. We varied the size of the network to achieve the least RMSE but the layer remains constant as we only consider feed forward networks with one hidden layer.

Table 4. Train & Test RMSE for NN

Size	**RMSE Train**	**RMSE Test**
3	35767.15	68746.41
5	25328.13	85742.50
10	9559.229	100140.873
20	3193.408	118265.849

VI. Result Comparison and Model Selection:

The results comparison has been done on the basis of RMSE score evaluation.

Table 5. RMSE Comparison between different models

Model	**RMSE (Train)**
Linear Regression	140639.3
Random Forest	105957.9
ARIMA	35975.49
Neural Network	35767.15

Based on the comparison it has been found that Neural Network model works best in this case when neuron size is 3. Increase of the size of neuron increases the chance for over fitting. Also from different literature review it has been seen that for large dataset having multiple attributes neural network always works better than any other machine learning algorithm.

Conclusion

The main focus is to implement several statistical and machine learning algorithms to predict sales price. Several stores have different number of departments and corresponding attributes in this dataset. The stores are from different region having different scale of temperature, unemployment rate, customer price index etc. So it is logical to predict the weekly sales of the stores separately. We tried to demonstrate several statistical methods as well as machine learning algorithms to predict the weekly sale of a particular store labeled as store 1. Later based on these models weekly sales of other 44 stores can be predicted as well. As several methods have different algorithm and parameters to work with so the comparison between RMSE values should be the justified evaluation technique. Though RMSE value indicates that neural network performs well it also has some drawbacks. The training time is comparatively high for neural network and also it works like a black box. With higher number of nodes and layers it’s always difficult to keep the track of what’s going on in between.

References

[1] W. T. Lim, L. Wang, Y. Wang and Q. Chang, "Housing price prediction using neural networks," 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, 2016, pp. 518-522.

[2] L. Li and K. H. Chu, "Prediction of real estate price variation based on economic parameters," 2017 International Conference on Applied System Innovation (ICASI), Sapporo, 2017, pp. 87-90. doi: 10.1109/ICASI.2017.7988353

[3] Nissan Pow, Emil Janulewicz, and Liu (Dave) Liu, “Prediction of real estate property prices in Montreal”

[4] Gautam Shine, Sanjib Basak “Sales Prediction with Time Series Modeling.”

[5] Beatrice Ugiliweneza, “Use of arima time series and regressors to forecast the sale of Electricity.”

Appendix

A. *Linear Regression*

For the linear regression model we have obtained following results.

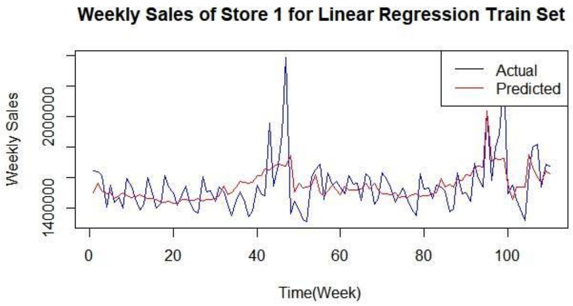

Figure 4. Weekly Sales of Store 1 (Train Set)

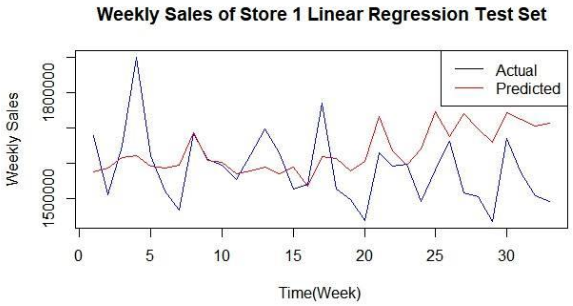

Figure 5. Weekly Sales of Store 1 (Test Set)

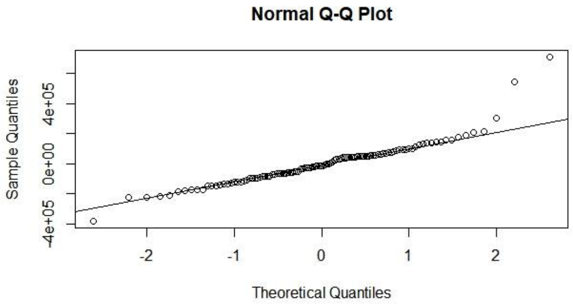

Figure 6. Normal Q-Q plot (Train Set)

From the above plot we can see that the residuals are normally distributed.

B. *Random Forest Regression (RF)*

For random forest several parameters were tuned and found ntree=100 & mtry=7 gives lowest RMSE for test set. Result are shown below:

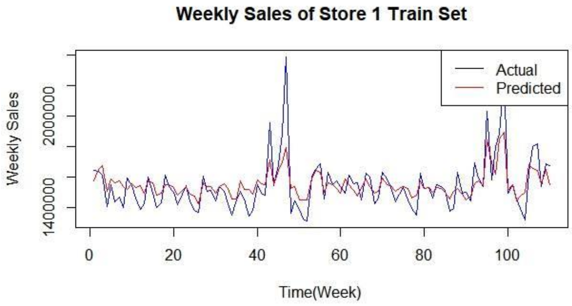

Figure 7. Weekly Sales of Store 1 (Train Set)

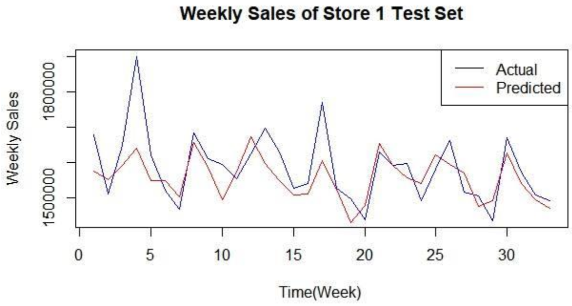

Figure 8. Weekly Sales of Store 1 (Train Set)

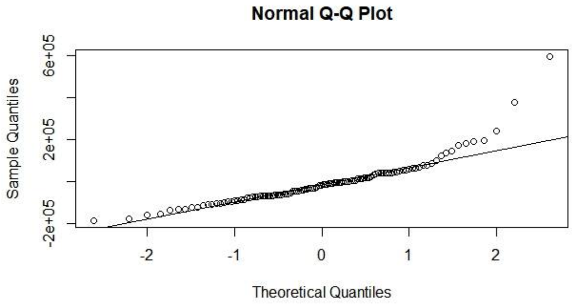

Figure 9. Normal Q-Q plot (Train Set)

C. *Auto Regressive Integrated Moving Average (ARIMA)*

For ARIMA we have trained various combinations to come up with the optimum one (1,0,1)(0,1,1)$_{[52]}$. Below ACF & PACF plot of the training data has been shown.

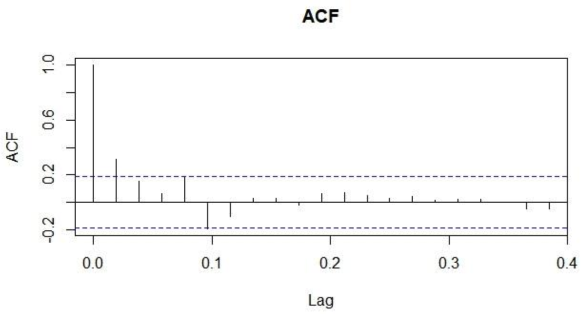

Figure 10. ACF of training data.

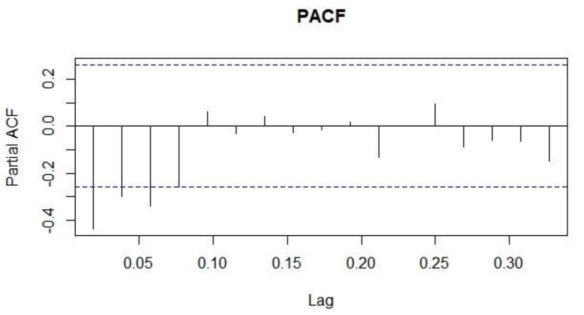

Figure 11. PACF of training data.

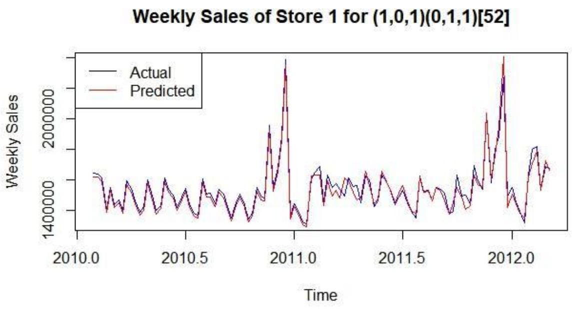

Figure 12. Weekly Sales of Store 1 [Training Data]

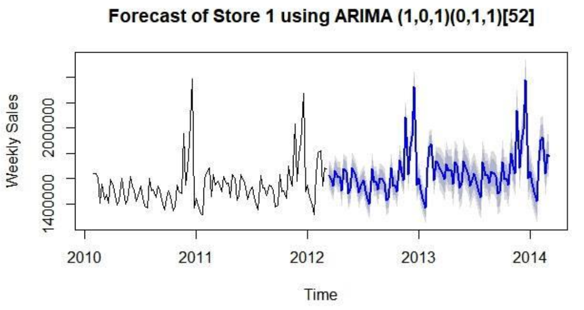

Figure 13. 2 year forecast (ARIMA model)

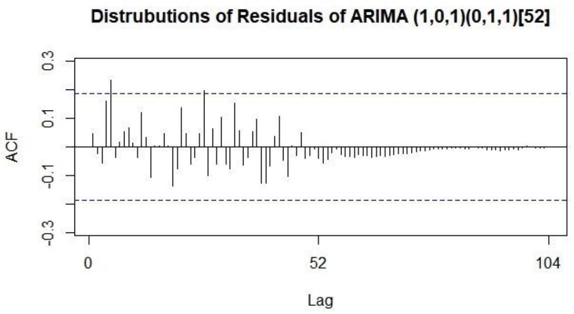

Figure 14. ACF of Residual.

D. Neural Network

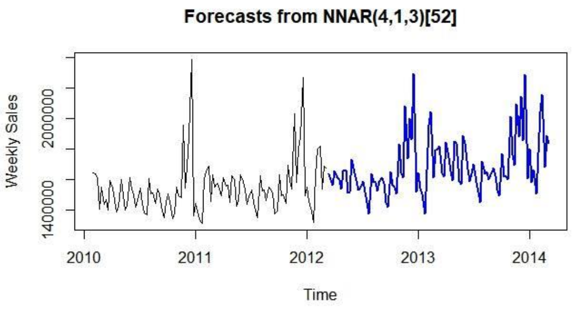

Figure 15. 2 year forecast NNAR(4,1,3)[52]

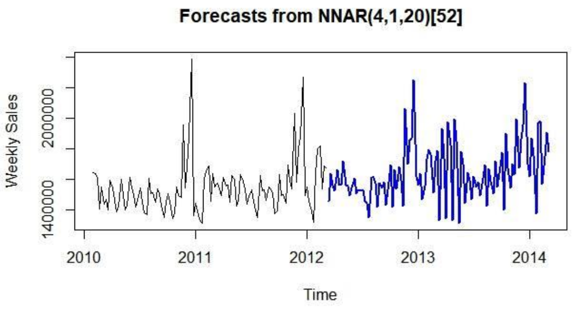

Figure 16. 2 year forecast NNAR(4,1,10)[52]