Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting

Stuti Raizada, Jatinderkumar R. Saini*
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University)
Pune, India

Abstract—**This study talks about how data mining can be used for sales forecasting in retail sales and demand prediction. Prediction of sales is a crucial task which determines the success of any organization in the long run. There are various techniques available for predicting the sales of a supermarket such as Time Series Algorithm, Regression Techniques, Association rule etc. In this paper, a comparative analysis of some of the Supervised Machine Learning Techniques have been done such as Multiple Linear Regression Algorithm, Random Forest Regression Algorithm, K-NN Algorithm, Support Vector Machine (SVM) Algorithm and Extra Tree Regression to build a prediction model and precisely estimate possible sales of 45 retail outlets of Walmart store which are at different geographical locations. Walmart is one of the foremost stores across the world and thus authors would like to predict the sales accurately. Certain events and holidays affect the sales periodically, which sometimes can also be on a daily basis. The forecast of probable sales is based on a combination of features such as previous sales data, promotional events, holiday week, temperature, fuel price, CPI i.e., Consumer Price Index and Unemployment rate in the state. The data is collected from 45 outlets of Walmart and the prediction about the sales of Walmart was done using various Supervised Machine Learning Techniques. The contribution of this paper is to help the business owners decide which approach to follow while trying to predict the sales of their Supermarket taken into account different scenarios including temperature, holidays, fuel price, etc. This will help them in deciding the promotional and marketing strategy for their products.**

Keywords*—*Sales forecasting; linear regression; random forest regression; KNN regression algorithm; SVM algorithm; supervised machine learning techniques

I. Introduction

Retail is considered as one of the most significant and fast-growing business domains in data science field because of its high-volume data and abundant optimization challenges for example, ideal prices, recommendations, discounts, stock levels which can be resolved by using different data analysis methods. When it comes to predicting the sales of commodities, it gets quite challenging in today's stimulating and ever-changing business environment. Only a few enhancements while sales prediction could help retailers in depressing operational costs and improving sales. And it could result in more customer satisfaction [1]. Prediction of correct sales at every outlet of retail is important for the accomplishment of each retailing company as it aids in management of inventory, results in right distribution of products across various stores, solves the problem of over and under stocking at each store in order to minimize losses, and maximize sales and satisfaction of customers [2]. Since there are many factors that come into play when one has to predict the sales, it has become a challenging issue to solve for all retail companies [3]. To add on, sales can also depend on a diversity of external factors such as weather, seasonal trends happening in a place where the store is located, competition from other retail stores and online shopping etc. It may include internal actions for example, promotions, discounts, pricing etc., which add to the intricacy of the problem.

There are various Machine Learning techniques which could be used for forecasting the sales of a Supermarket. Random Forest Regression is a supervised learning algorithm which incorporates ensemble learning method in it which helps in doing forecasting. A Random Forest works by constructing several decision trees first during the course of training and using the mean of the classes as the output for prediction of all the trees. A prediction from Random Forest Regressor is generated by taking an average of all the predictions produced by trees in the forest, which will increase the accuracy of the prediction. K-NN Regression is a technique that uses feature similarity to predict the value of a new data point. The new value is predicted based on how closely it lies to the points in the training dataset. SVR is a Supervised learning algorithm and works on the principle of Support Vector Machine. It is used in determining the best line of fit which has maximum no. of points lying on it. Extra Tree Regression is also an ensemble learning model with very minor difference to Random Forest Technique. Random Forest uses the replica of bootstrap i.e., doing the sub-sample of input data with replacement whereas Extra Tree Regressor takes into account the entire original sample of data. One for difference between the two techniques is selection of cut point. Extra Tree techniques choose the cut points randomly; however, Random Forest selects the optimum split.

After studying the literature related to Sales Forecasting, the methodology adopted has been defined. The model has been trained on various ML Algorithms such as Linear Regression, Random Forest Regression, KNN Regression, SVR and Extra Tree Regression. The Results obtained from each of the model have been discussed and finally the conclusion is obtained.

II. Literature Review

Microsoft Time Series Algorithm [4] provides regression algorithms that are optimized for forecasting of continuous

*Corresponding Author.

values such as product sales or demand over time. If one has to forecast for continuous variables like product sales, demand over time using regression algorithms Microsoft time series algorithm will help in that. It will not need any new additional columns to predict trends unlike decision tree algorithm, which can be considered as one of the significant advantages of Time Series Algorithm. It is capable of predicting any anomalies that we can face in the sales/demand based on the source data set that is fed into the model. As data keeps growing, one can simply update the data that is being used as input to the model, and the model will incorporate that and predicts accordingly. Cross Prediction can be performed using the Microsoft Time Series algorithm which is one of its unique features. The algorithm can be trained with two different, but related data sets or series, and the resulting model will be able to predict the outcome of one series based on the behavior of the other series by understanding the co-relation existing between them. For example, let's consider the problem statement as observed sales of one car can influence the forecasted sales of another car. Working of the Algorithm: When data related this problem, statement is given to the time series, it will be using Autoregressive Tree Models with Cross Prediction (ARTXP) and also Autoregressive Integrated Moving Average (ARIMA) and then combines the output of both algorithms which will help in improving the prediction accuracy. When it comes to predicting something for short-term, ARTXP algorithm will be used and for long-term predictions ARIMA.

Linear Regression [5] is a technique to model the relationships between two variables by fitting a linear equation to observed data. One variable is termed as explanatory variable (predictor) and the other variable is the dependent variable (target). It is about finding the best line of fit for training as well as testing data. This technique has been used in predicting the demand of commodities by analyzing the sales of the stores. Sales Forecasting is an important aspect in Production and Supply Chain Management [6]. KNN Regression is a regression technique uses the similarity measure to predict the values after analyzing the past cases data. It extracts the features from the data and uses 'feature similarity' in order to predict the values of new data points. The value of new data is assigned by calculating the average of the nearest neighbors of the new data point. The other approach to KNN is by calculating an inverse distance weighted average of the K-nearest neighbors. It uses same distance functions as used for Classification – Euclidean, Minkowski and Manhattan. Association Rule Discovery [7]; because of its wide application, Association Rule Discovery has become a trending topic in Data Mining. It finds the frequent patterns among the datasets. The aim of Association rule mining is to extract interesting relations, common patterns, and correlations among sets of items in the data repositories. For Example, it can be seen that 80% of the customers in India who buy Mobile Phones also buy Headsets for better music quality. Shelke et al. [8] has discussed various Machine Learning algorithms which can be applied in multiple sectors of industry such as retail, marketing, logistics etc. based on the requirement. It concluded the study by indicating that Rule Induction (RI) is the most frequently used ML technique in data mining [9] [10]. The previous study on sales prediction have been performed using regression techniques as well as boosting techniques and boosting algorithms have resulted in better results as compared to regression techniques [11]. Zhan-Li Sun et.al [12] has used a neural network technique known as Extreme Learning Machine (ELM) to find out the relationship between sales amount and few crucial factors which affect demand using a real time dataset and found that their model outperform over the other sales forecasting methods using back propagation neural networks. Non-linear models are compared with linear model for sales forecasting and it was observed that neural networks perform well with de-seasonalized time series data whereas Regression models are effective with seasonal dummy variables [13]. Fashion Retail Industry has been considered as the most difficult in terms of predicting the sales due to shorter life span of products as the taste of the customer keeps changing. Statistical techniques have been applied to predict the sales such as Bayesian analysis, Exponential smoothening etc. Also, forecasting has been done using AI Methods of Artificial Neural Networks (ANN) and Evolutionary Neural Networks (ENN) [14]. Thomassey et al. [15] has proposed a forecasting model based on hybrid approach of combining fuzzy logic, neural networks, and evolutionary procedures. Manpreet et al. [16] have considered big data perspective while predicting the sales of Walmart. He has used the technologies such as Python API and Scala of the Spark framework.

Data is of no use if it cannot be examined, understood and applied in some context [17]. Harsoor and Patil et al. [18] have predicted the sales of Walmart stores using Big data applications such as Hadoop, MapReduce and Hive. For the purpose of analyses and visualizing of data, tools such as Hadoop Distributed File Systems (HDFS) [19], Hadoop Map Reduce Framework [20] and Apache Spark along with Scala, Python high level programming environments are used. Katal, Wazid and Goudar et al. [21] proposed Parallel programming like Distributed File System, Map Reduce and Spark as the prominent tools for dealing with Big Data. Sharma, Chauhan and Kishore did a comparative study between Hadoop, MapReduce and Spark and concluded that Spark is much better option for analyzing Big Data [22]. Spark has proven to be 100 times faster than other techniques of data analysis [23]. Omar et al. [24] has inspected the Back Propagation Neural Network for forecasting the sales of Walmart.

After studying the literature available for Sales Forecasting in various industries, it was identified that various algorithms have been used and the choice of algorithm is extremely crucial based on the dataset on which forecasting has to be made. For the data of short time period, statistical models could be used but they may not perform well with Big Data and thus, technologies such as Hadoop Distributed File System or Map Reduce is a better choice. This paper will help the retailers to decide which Machine Learning Algorithm will serve their purpose of sales forecasting without involving into the complexities of first choosing the algorithm and then implementing it.

III. Methodology

Here, the aim is to predict the sales of Walmart store using various Supervised ML Algorithms. The algorithms used are Linear Regression, Random Forest Regression, K-NN

Regression, Support Vector Machine and Extra Tree Regression as shown in Fig. 1. Linear Regression has been used for predicting the sales of several commodities in Walmart taking in consideration factors such as previous sales, Holiday week, Fuel price in that week, Unemployment rate etc. Considered the dataset of 45 retail outlets of Walmart store and did cleaning of data using Python. Further the dataset was divided into Training Data and Testing Data in the ratio of 80:20. Post that, Data Pre processing techniques have been applied in order make the data ready for feeding into the models. After feature selection, the model was trained and then the test data was given as an input and feature measurement has been performed. Lastly, the input was given to different model built for various algorithms and accuracy scores were obtained.

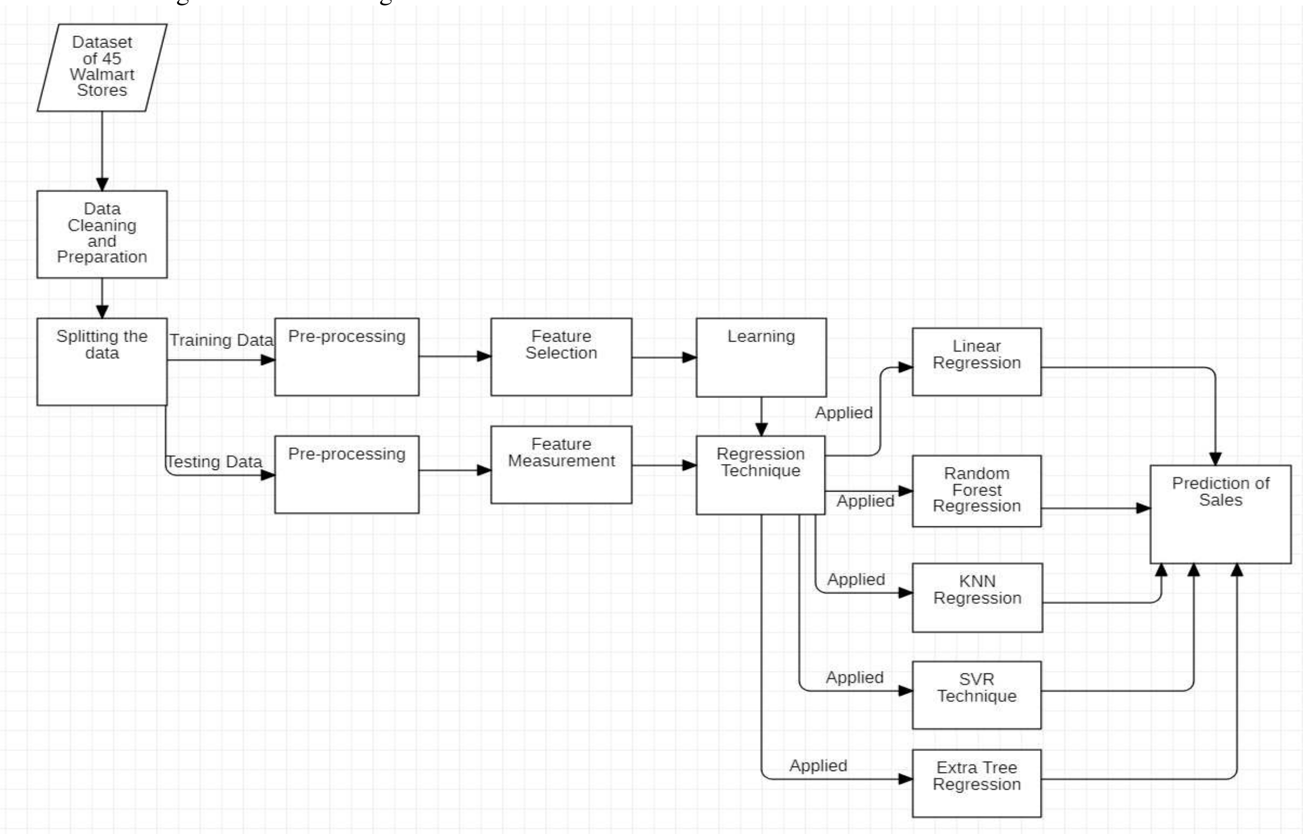
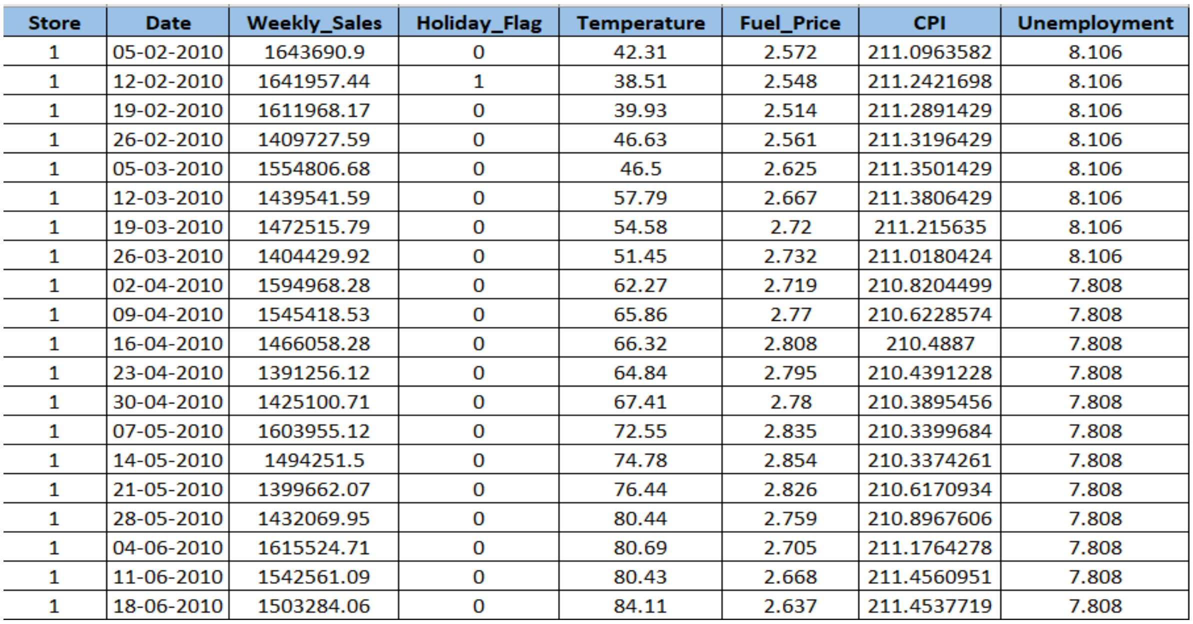


Fig. 1. Proposed Methodology.



Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	05-02-2010	1643690.9	0	42.31	2.572	211.0963582	8.106
1	12-02-2010	1641957.44	1	38.51	2.548	211.2421698	8.106
1	19-02-2010	1611968.17	0	39.93	2.514	211.2891429	8.106
1	26-02-2010	1409727.59	0	46.63	2.561	211.3196429	8.106
1	05-03-2010	1554806.68	0	46.5	2.625	211.3501429	8.106
1	12-03-2010	1439541.59	0	57.79	2.667	211.3806429	8.106
1	19-03-2010	1472515.79	0	54.58	2.72	211.215635	8.106
1	26-03-2010	1404429.92	0	51.45	2.732	211.0180424	8.106
1	02-04-2010	1594968.28	0	62.27	2.719	210.8204499	7.808
1	09-04-2010	1545418.53	0	65.86	2.77	210.6228574	7.808
1	16-04-2010	1466058.28	0	66.32	2.808	210.4887	7.808
1	23-04-2010	1391256.12	0	64.84	2.795	210.4391228	7.808
1	30-04-2010	1425100.71	0	67.41	2.78	210.3895456	7.808
1	07-05-2010	1603955.12	0	72.55	2.835	210.3399684	7.808
1	14-05-2010	1494251.5	0	74.78	2.854	210.3374261	7.808
1	21-05-2010	1399662.07	0	76.44	2.826	210.6170934	7.808
1	28-05-2010	1432069.95	0	80.44	2.759	210.8967606	7.808
1	04-06-2010	1615524.71	0	80.69	2.705	211.1764278	7.808
1	11-06-2010	1542561.09	0	80.43	2.668	211.4560951	7.808
1	18-06-2010	1503284.06	0	84.11	2.637	211.4537719	7.808

Fig. 2. Walmart Data Set for 45 Stores.

A. Dataset and Experiment Discussion

The sales data which is considered for prediction model has been taken from 45 stores of Walmart. The historical data taken for prediction covers sales from February 5, 2010 to November 1, 2012 [25]. There are 3 separate data files corresponding to each year and the accuracy of models has been calculated accordingly.

The data set which is considered for the study contains the following fields:

1) Store - the number of stores as 45 stores are considered.

2) Date – We have considered date as the first date of the week of sales for time series forecasting.

3) Weekly Sales – Weekly sales for the given store.

4) Holiday Flag – to determine if the week is a special holiday week. It shows 1 for Holiday week and 0 for Non-holiday week. This will help in understanding the trends during the holidays.

5) Temperature – Temperature recorded on the day of sale.

6) Fuel Price – Cost of fuel in the region where the store is located.

7) CPI – Dominant consumer price index.

8) Unemployment – Dominant unemployment rate in the region where store is present.

Fig. 2 shows the snapshot of the dataset of Walmart store.

Following are the steps which are followed for experimentation**:**

1) The first step is importing the necessary libraries which would be used for building the model such as numpy, pandas, matplotlib, seaborn.

2) After that, loaded the dataset for every year 2010, 2011 and 2012 respectively to the IDE.

3) Once data has been loaded, prepared the data for experiment by converting date to datetime format.

4) Checked if there are any missing or null values.

5) Then, splitted the date column and created 3 columns namely day, month and year.

6) For building the prediction model, found outliers in the data by plotting Temperature, Fuel Price, CPI and Unemployment on X- axis.

7) The next step was to drop the outliers and considered the range in which outliers does not fall.

8) Then, again checked if the plot looks fine without outliers.

9) Imported sklearn library for building the model and selected features and target for X and Y axis to predict the sales.

*10)*Splitted the data into training and testing set in the ratio of 80:20.

*11)*Used Linear Regression, Random Forest Regressor, KNN Regressor, SVR, Extra Tree Regressor to predict the sales of Walmart store along Y-axis to do comparative analysis of Prediction Model.

*12)*Calculated the errors in the Prediction Model by finding Mean Absolute Error, Mean Squared Error and Root Mean Squared Error.

IV. Result

The results obtained from the prediction models fed with datasets of three years 2010, 2011 and 2012 have been summarized below in Tables I, II and III, respectively.

From the Tables I, II and III, it has been observed that the Mean Absolute Error is highest in case of Support Vector Regression for all the three years and minimum in case of Extra Tree Regression. Mean Absolute Error is the average magnitude of the error in prediction set. It is the average over the test sample of absolute difference between prediction and actual observation.

TABLE I. STATISTICAL MEASURES FOR DATASET OF YEAR 2010

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	424421.93	251400879887.53	501398.92
Random Forest Regression	80396.14	25126954749.18	158514.84
KNN Regression	281135.23	131980791838.85	363291.60
Support Vector Regressor	470857.85	320200129812.73	565862.28
Extra Tree Regression	48281.35	5534368506.39	74393.33

TABLE II. STATISTICAL MEASURES FOR DATASET OF YEAR 2011

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	409909.51	235782880216.47	485574.79
Random Forest Regression	43811.57	4031635222.35	63495.15
KNN Regression	272092.75	123596844025.18	351563.42
Support Vector Regressor	430737.43	267243666430.57	516956.15
Extra Tree Regression	42752.07	3840250218.75	61969.75

TABLE III. STATISTICAL MEASURES FOR DATASET OF YEAR 2012

Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	447155.52	269282898979.58	518924.75
Random Forest Regression	50487.45	9118432907.22	95490.48
KNN Regression	272433.64	122911201809.44	350586.93
Support Vector Regressor	474953.02	311212134485.71	557863.90
Extra Tree Regression	42218.75	3952869757.58	62871.85

Root Mean Squared error is the square root of the average of squared differences between predicted and actual observation. It is least in case of Extra Tree Regression as compared to other Regression Techniques.

Before building the prediction model, the outliers in the dataset have been identified and removed. Fig. 3 to Fig. 6 visualizes outlier detection for the datasets of year 2010, 2011 and 2012. Among these, Fig. 3, Fig. 4 and Fig. 6 depict the presence of outliers in Temperature data, Fuel Price data and Unemployment data respectively. Fig. 5 depicts that there is no outlier in Consumer Price Index.

After finding out the outliers, the next step was to remove the outliers for feeding the input to the Prediction model.

Fig. 7, 12 and 17 are obtained after performing Linear Regression on the data for the year 2010, 2011 and 2012 respectively. However, it has been observed that for all the three datasets the graph looks scattered and thus it is not advisable to predict the sales using Linear Regression Model.

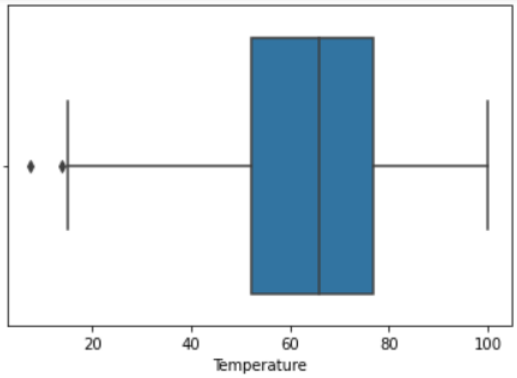


Fig. 3. Outlier in Temperature.

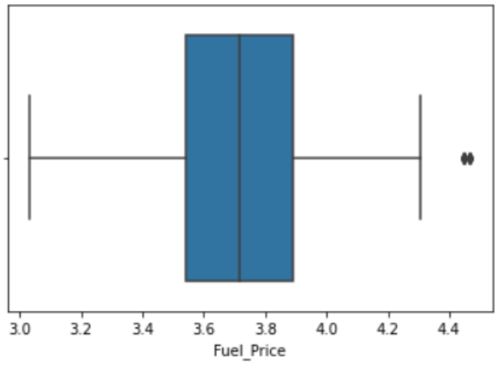


Fig. 4. Outliers Present in Fuel Price.

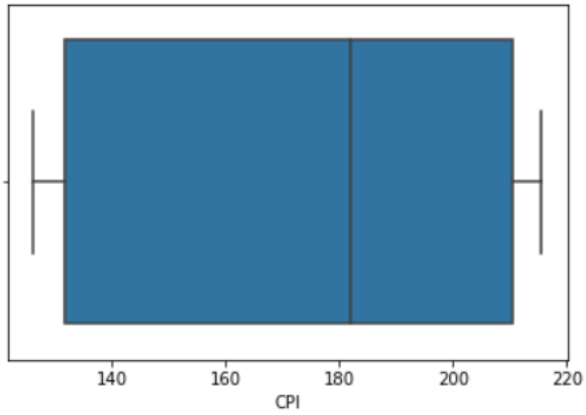


Fig. 5. No Outliers in CPI.

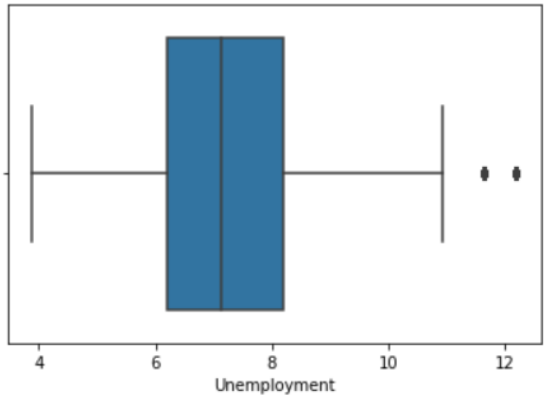


Fig. 6. Outliers Present in Unemployment.

Fig. 8, 13 and 18 forecasts the sales of the products in Walmart Store using Random Forest Technique against the weekly sales of the stores. Various factors taken into consideration on X-axis are Store number, Fuel Price, Unemployment, CPI, day, month and year. The graph in this case is almost concentrated on the best line of it and Random Forest provides good accuracy scores for all 3 datasets.

Fig. 9, 14 and 19 forecasts the sales using KNN Regression Technique. The graph is not much concentrated but it performs better than Linear Regression and provides the accuracy of around 50 to 60%.

Fig. 10, 15 and 20 are obtained after applying Support Vector Regression technique and it clearly demonstrate the worst performance amongst all other techniques used for predicting the sales of Walmart store.

Fig. 11, 16 and 21 are obtained after predicting the sales from Extra Tree Regressor Model and it performed best amongst all the models discussed so far. The graph looks somewhat similar to Random Forest Technique however, it is more accurate and all the data points are almost falling on the best of fir providing the accuracy of 98% in all three cases. This is because both these techniques use ensemble model of learning and averages the output obtained from several decision trees to provide better performance.

Results obtained for 2010 year dataset are shown below.

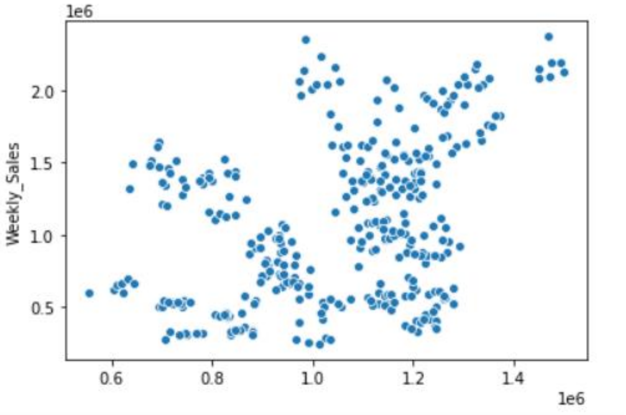


Fig. 7. Sales Prediction using Linear Regression Model (2010).

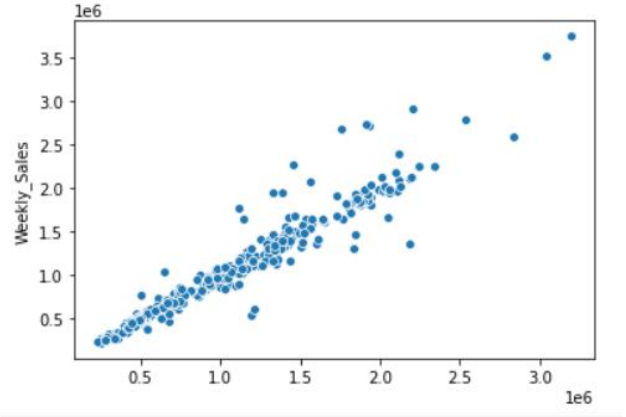


Fig. 8. Sales Prediction using Random Forest Regression Model (2010).

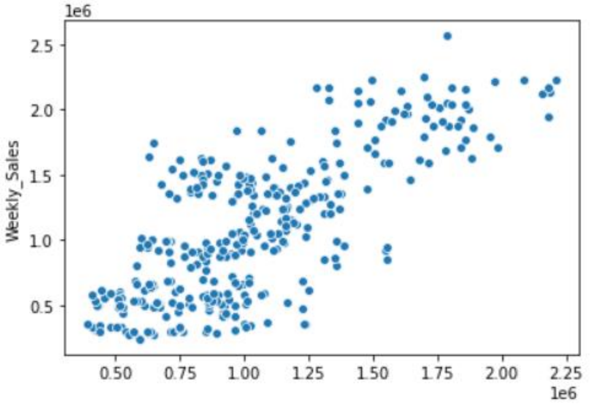


Fig. 9. Sales Prediction using KNN Regression Model (2010).

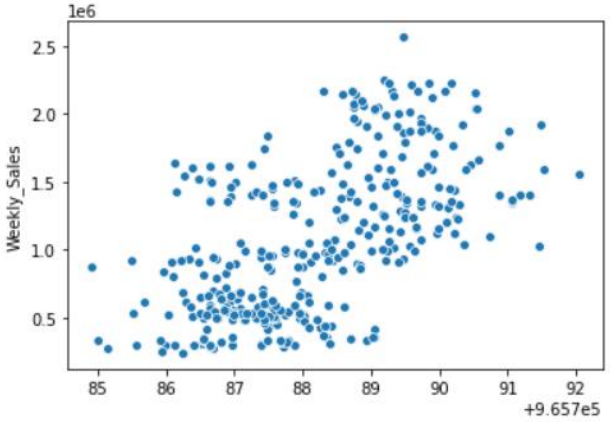


Fig. 10. Sales Prediction using SVR Model (2010).

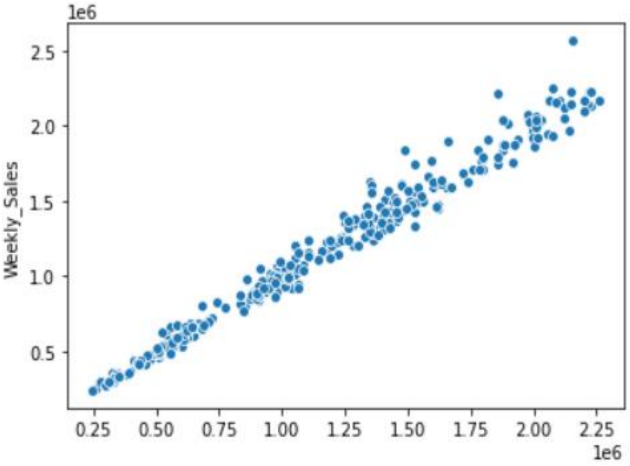


Fig. 11. Sales Prediction using Extra Tree (2010).

The results obtained for the dataset of year 2011 are as follows:

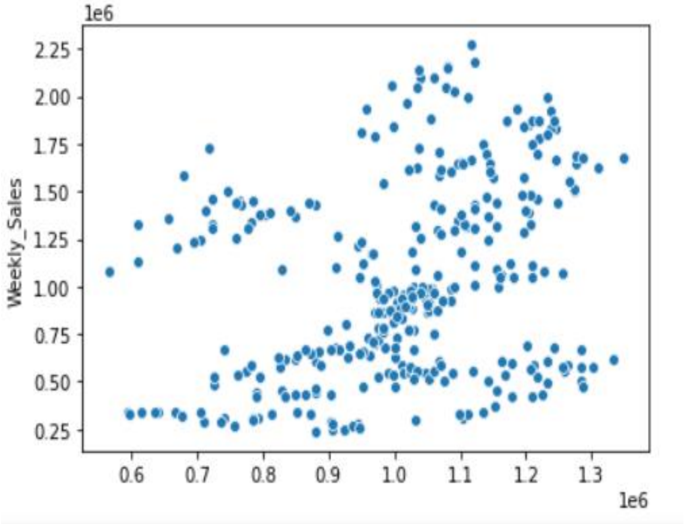


Fig. 12. Sales Prediction using Linear Regression Model (2011).

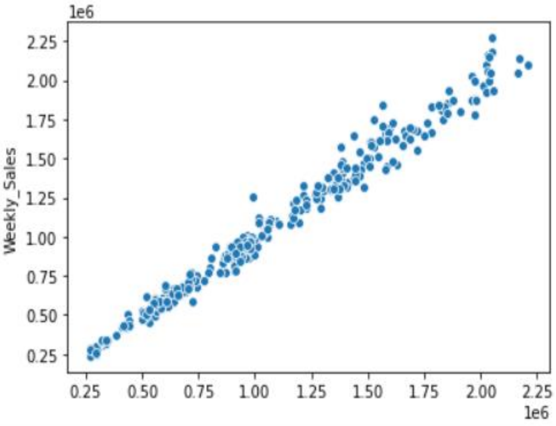


Fig. 13. Sales Prediction using Random Forest Regression Model (2011).

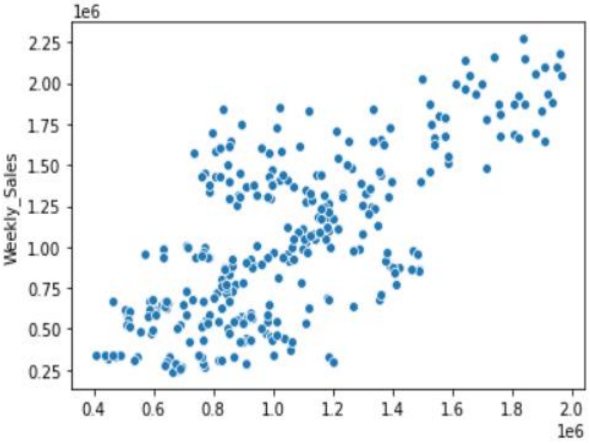


Fig. 14. Sales Prediction using KNN Regression Model (2011).

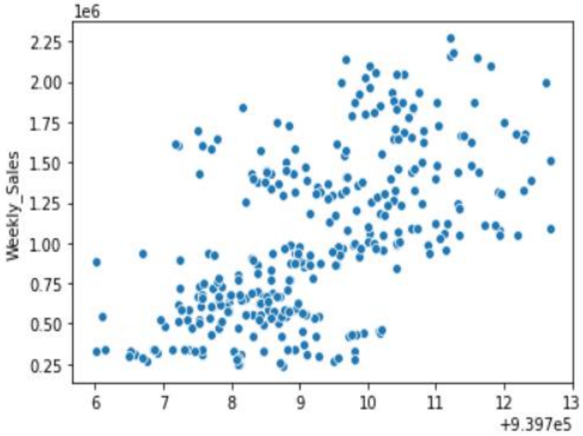


Fig. 15. Sales Prediction using SVR Model (2011).

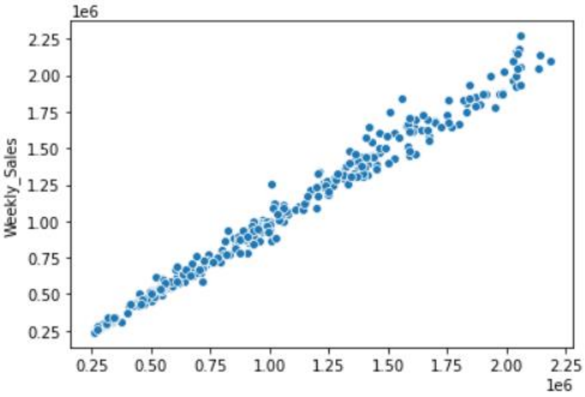


Fig. 16. Sales Prediction using Extra Tree (2011).

The below graphs are obtained for the data of year 2012:

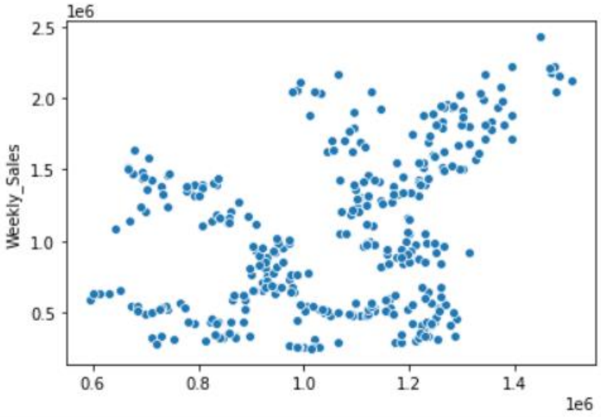


Fig. 17. Sales Prediction using Linear Regression Model (2012).

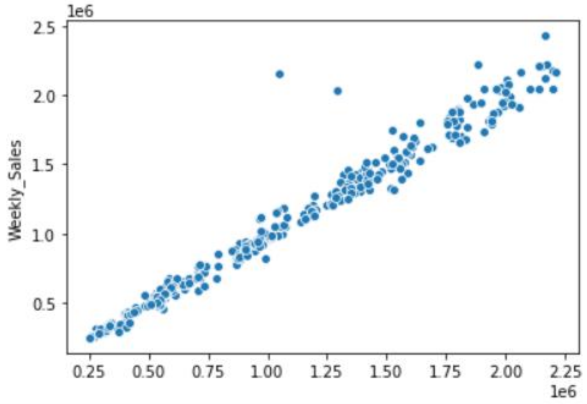


Fig. 18. Sales Prediction using Random Forest Regression Model (2012).

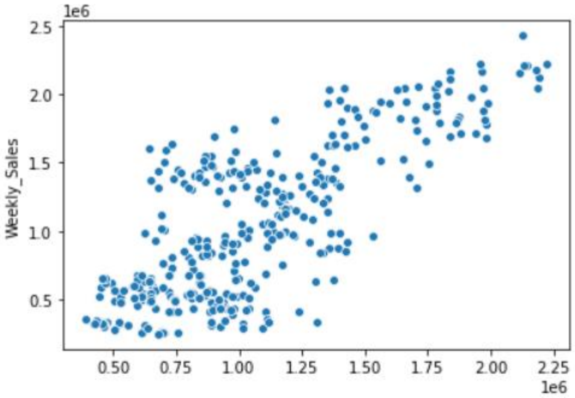


Fig. 19. Sales Prediction using KNN Regression Model (2012).

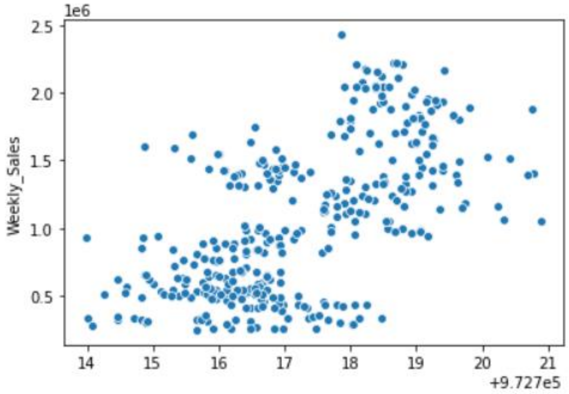


Fig. 20. Sales Prediction using SVR Model (2012).

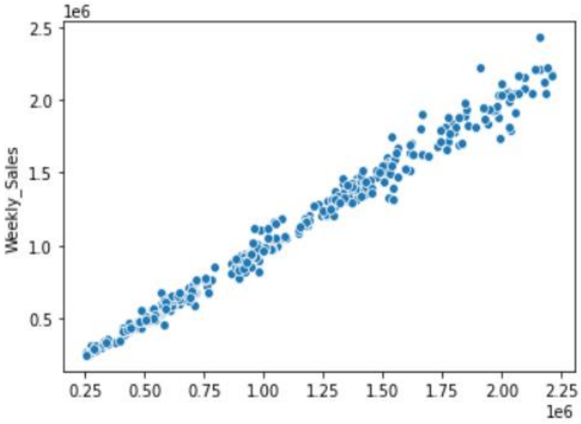


Fig. 21. Sales Prediction using Extra Tree (2012).

From the Tables IV, V and VI, it is evident that Support Vector Regression model is the poorest and could not predict the sales of Walmart stores correctly. However, Extra Tree Regression Model performs best on the data for all three years when compared to other supervised Machine Learning techniques and predicts the sales with 98% accuracy and thus could be relied upon for Sales forecasting when the parameters considered are Fuel Price, Unemployment, Holiday and CPI.

TABLE IV. PERFORMANCE METRICS FOR YEAR 2010

Performance Metric Name	**Linear Regression Score**	**Random Forest Score**	**KNN Regression Score**	**SVR Score**	**Extra Tree Regression Score**
Accuracy	13.95%	92.73%	57.00%	- 4.32%	98.20%

TABLE V. PERFORMANCE METRICS FOR YEAR 2011

Performance Metric Name	**Linear Regression Score**	**Random Forest Score**	**KNN Regression Score**	**SVR Score**	**Extra Tree Regression Score**
Accuracy	10.23%	98.45%	52.71%	- 2.24%	98.53%

TABLE VI. PERFORMANCE METRICS FOR YEAR 2012

Performance Metric Name	**Linear Regression Score**	**Random Forest Score**	**KNN Regression Score**	**SVR Score**	**Extra Tree Regression Score**
Accuracy	14.15%	97.01%	59.77%	- 1.85%	98.70%

V. Discussion

Based on the above experimentation, it has been observed that Simple Regression techniques for building the prediction models may not be the best choice for sales prediction if the management is trying to predict the sales for lesser duration and have historical data only for few years. This is because the accuracy is good only for ensemble learning techniques which involves averaging of results obtained from multiple decision trees. Therefore, the business owner should choose Ensemble Learning Models.

One limitation of this study is that based on the variance in training data, the predictions obtained from a specific algorithm may vary. So, the owner has to decide the algorithm effectively given his requirements.

VI. Conclusion

Based on the dataset used, it can be said that Extra Tree Regression Technique is the best to predict the sales of Walmart Store in future followed by Random Forest Regression Technique. This result could be useful for other retail store owners as well in order to determine their sales and they could directly opt for Sales Prediction using Extra Tree Regression Technique or Random Forest Approach rather than spending time in doing analysis using other Supervised Machine Learning Algorithms. The other retailers could also be benefitted by doing the demand analysis on the similar grounds. This study contributed in understanding the fact that external factors, such as Unemployment rate, Holiday Week, CPI, etc. also plays a vital role while predicting the sales of any retail store.

References

[1] Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. San Diego, California: UC San Diego Jacobs School of Engineering.

[2] Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

[3] Wayne, L. (2014). Winston. Analytics for an Online Retailer: Demand Forecasting and Price Optimization.

[4] Mekala, P., & Srinivasan, B. (2014). Time series data prediction on shopping mall. Int. J. Res. Comput. Appl. Robot, 2(8), 92-97.

[5] Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2021). Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change, 163, 120480.

[6] Vahid Sohrabpour, Pejvak Oghazi, Reza Toorajipour, Ali Nazarpour. (2021). Export sales forecasting using artificial intelligence, Technological Forecasting and Social Change, Volume 163.

[7] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender systems: an introduction. Cambridge University Press.

[8] Shelke, R. R., Dharaskar, R. V., & Thakare, V. M. (2017). Data mining for supermarket sale analysis using association rule. Int. J. Trend Sci. Res. Dev, 1(4).

[9] Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. Information & management, 39(3), 211-225.

[10] Punam, K., Pamula, R., & Jain, P. K. (2018, September). A two-level statistical model for big mart sales prediction. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 617-620). IEEE.

[11] Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2018, December). Sales-forecasting of retail stores using machine learning techniques. In 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) (pp. 160-166). IEEE.

[12] Zhan-Li Sun, Tsan-Ming Choi, Kin-Fan Au, Yong Yu. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing, Decision Support Systems, Volume 46, Issue 1.

[13] Ching-Wu Chu, Guoqiang Peter Zhang. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting, International Journal of Production Economics,Volume 86, Issue 3.

[14] Govindan, Kannan, Liu, Na, Ren, Shuyun, Choi, Tsan-Ming, Hui, Chi-Leung, Ng, Sau-Fun. (2013). Sales Forecasting for Fashion Retailing Service Industry: A Review, Mathematical Problems in Engineering, Hindawi Publishing Corporation.

[15] S. Thomassey, M. Happiette, and J.-M. Castelain. (2005). "A global forecasting support system adapted to textile distribution," International Journal of Production Economics, vol. 96, no. 1, pp. 81–95, 2005.

[16] M. Singh, B. Ghutla, R. Lilo Jnr, A. F. S. Mohammed and M. A. Rashid, "Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2017.

[17] D. Silverman, "Interpreting Qualitative Data: Methods for Analyzing Talk Text and Interaction", Text and Interaction, Sage Publications Ltd: Methods for Analyzing Talk, 2006.

[18] A. S. Harsoor and A. Patil, "Forecast of sales of walmart store using Big Data application", International Journal of Research in Engineering and Technology, vol. 4, pp. 6, June 2015.

[19] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, et al., "Fast and interactive analytics over Hadoop data with Spark", U senix - The Advanced Computing Systems Association, 2012.

[20] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters", Association for Computing Machinery, 2008.

[21] A. Katal, M. Wazid and R. H. Goudar, Big Data: Issues Challenges Tools and Good Practices, 2013.

[22] M. Sharma, V. Chauhan and K. Kishore, "A review: MapReduce and Spark for Big Data analysis", 5th International Conference on Recent Innovations in Science. 5: Engineering and Management, June 2016.

[23] H. Pandey, Is Spark really 100 times faster on stream or its hype?, vol. 2, Sept 2016.

[24] Omar, H. A., & Liu, D. R. (2012, January). "Enhancing sales forecasting by using neuro networks and the popularity of magazine article titles." Sixth International Conference on Genetic and Evolutionary Computing (ICGEC) (pp. 577-580).

[25] https://www.kaggle.com/input/retail-analysis-with-walmart-data - Dataset used for modelling .