Sales Forecasting on Superstore Data Using ARIMA

Aman Gupta[1], Aman Sagar[2], Deepak yadav[3], Manvendra Singh[4], Shelly Gupta[5]

[1,2,3,4] Student CSE dept., IPEC Ghaziabad
[5] Assistant Professor CSE dept., IPEC Ghaziabad

---***---

Abstract - Sales forecasting is widely recognized and plays a major role in an organization's decision making. Continuous improvement in sales forecasting is a worthy goal for any organization. It helps the organization to plan for maximizing the profits by having the high demanding products in stock. Thus, by using the techniques of machine learning[1] we can predict in advance what products are sold in high quantity and at what period of the year. Those companies that have responded positively to the sales audit process have experienced significant improvement in their forecasting performance. Sales forecasting uses trends identified from historical data to predict future sales, enabling educated decisions including assigning or redirecting current inventory or effectively managing future production. So that they can change their strategy to improve sales. This helps in better management of resources like machine, money, and manpower also help in managing revenue and inventory accordingly. Using machine learning model helps in getting reliable and accurate results. The objective of this paper is to predict the future sales of products of a Super Store. For this ARIMA technique, SARIMA technique which takes seasonality of the data into the consideration. We then forecast the results with great degree of accuracy.

***Key Words*:** Sales Forecasting, Feature Extraction, ARIMA, SARIMA.

1. INTRODUCTION

Continuous improvement in sales forecasting is a worthy goal for any organization. It helps the organization to plan for maximizing the profits by having the high demanding products in stock. Thus, by using the techniques of machine learning we can already predict in advance what products are sold in high quantity and at what period of the year. Those companies that have responded positively to the sales audit process have experienced significant improvement in their forecasting performance. The historical data contains some trends which can be used to predict the sales. This will help the company officials to make informed decisions about their inventory and visualize a clear picture about their sales.

Some of the outcomes from this technique are improved decision-making about the future, reduction of sales pipeline and forecast risks. We can also expect benchmarks that can be used to assess trends in the future and give us the ability to focus a sales team on high-revenue, high-profit sales pipeline opportunities, resulting in improved win rates.

This would help the end-user to improve their overall sales and will help them modify the changes in their sales strategy. This analysis will also prove to be cost savior for the company as it would only sell more products that are high in demand.

The objective of this paper is to analyze the historical data and based on the trends in the sales predict the future sales based on some parameters and learning various techniques[2] for time series forecasting like ARIMA[3], SARIMA, etc.

Sales forecasting helps businesses to stock sufficient products by estimating customer demand in advance. This analysis would also help the companies to manage the resources effectively.

The model that we have used for the prediction also takes into the consideration, seasonality of the data therefore, we can say that our model is feasible and can be used for predicting the future sales more accurately.

This paper is discussed in the following manner - first we have the literature review followed by the proposed methodology and then reflecting the experimental results and after that finally concluding the whole summary of the paper.

2. LITERATURE REVIEW

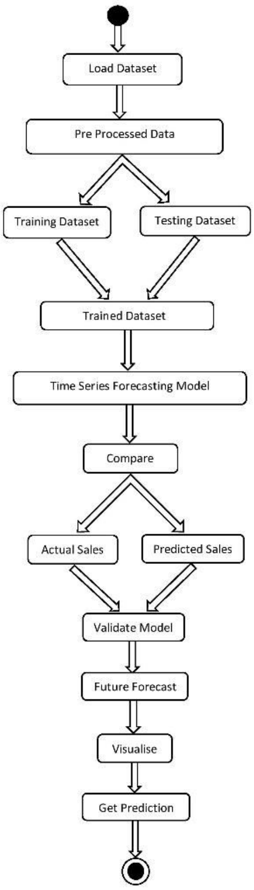
This section provides the brief description of various research papers studied for this study. The given below table 1 represents the summarization of various methods that can be applied for Frosting of Data.

TABLE I: METHODS THAT CAN BE APPLIED FOR SALES FORECASTING

S.NO	Title	Method	Description
1.	Linear Regression Analysis Study	Linear Regression	Kumari. K This paper explains the basic concepts and explains how we can do linear regression[4] calculations in SPSS and excel.
2.	Sales Forecasting using Regression and Artificial Neural Networks	Regression and ANN	The goal of this paper is to incorporate regression techniques and artificial neural network[5] (ANN) models to predict industry sales, which exhibit a seasonal pattern, by using both historical sales and non-seasonal economic indicators.
3.	Short Term Load Forecasting Using XGBoost	XGBoost	Raza Abid Abbas used XGBoost[6] for Electricity load prediction system, accurate prediction is needed for different perspectives, that are related to control, forwarding, planning and unit responsibility in a Power supply grid.
4.	Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea	ARIMA	Mohammed H. Alsharif This study aims to build a time series ARIMA model[7] to forecast daily and monthly solar radiation in Seoul, South Korea, based on the hourly solar radiation data obtained from the Korean Meteorological Administration (KMA) over 37 years. A time series ARIMA model is built to forecast the daily and monthly solar radiation of Seoul, South Korea, considering the accuracy, suitability, adequacy, and timeliness of the collected data
5	Forecasting Performance of Arima and Sarima Models on Monthly Average Temperature of Zaria , Nigeria	SARIMA	Time series analysis and forecasting has become a major tool in different applications for research[8]. In this study, we used Box-Jenkins Methodology to find an appropriate Model among Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) Model for temperature series of Zaria from 1993-2012.
6	A review on time series forecasting techniques for building energy consumption	Time series forecasting	Energy consumption forecasting for buildings has immense value in energy efficiency and sustainability research[9] . Accurate energy forecasting models have numerous implications in planning and energy optimization of buildings and campuses. For new buildings, where past recorded data is unavailable, computer simulation methods are used for energy analysis and forecasting future scenarios.
7	Machine learning vs statistical methods for time series forecasting: Size matters	ML vs statistical methods for time series forecasting	Show that these approaches systematically present a lower predictive performance relative to simple statistical methods[10]. In this work, we counter these results. We show that these are only valid under an extremely low sample size
8	Ensemble sales forecasting study in semiconductor industry	ML Model Comparison	Comparison of a sale forecasting tool to predict Intel's weekly CPU sale by lines of business[11]. The novel feature engineering helped to reduce the subjective bias and measurement errors, especially for economic indicators.
9	Machine-learning models for sales time series forecasting	Stacking Of ML models	A stacking approach for building regression ensemble of single models has been studied[12]. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting.
10	Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA and PROPHET	PROPHET and ARIMA Model	This study uses Facebook's Prophet Forecasting Model and ARIMA Forecasting Model[13] to compare their performance and accuracy on dataset containing the confirmed cases, deaths, and recovered numbers, obtained from the Kaggle website.
11	ARIMA Modelling using Machine Learning for selected Indian Automobile listed Companies	ARIMA modelling, Automobile sector	This study gives an approach as to how to execute an ARIMA model with the help of machine learning as new tool for researchers and also checks its suitability for forecasting share prices of three major companies of automobile sectors namely Mahindra& Mahindra, Maruti Suzuki and Tata Motors based on market capitalization.
12	Predictive analytics in Agriculture: Forecasting prices of Arecanuts in Kerala	Predictive analytics; SARIMA, RMSE	The fluctuations in prices of agricultural commodities have an adverse effect on the GDP of a country. The models SARIMA, HoltWinter's Seasonal method and their performance was evaluated based on the RMSE value on the arecanut dataset with prices.

3. METHODOLOGY

For this project, the following steps have been followed.



3.1 Data Collection and Data Description

To achieve the desired objective of study, the 'US superstore sales' data is selected, and the data set was collected from the Kaggle. The data includes the sales of many categories and sub-category under them, between the period of 2014-2017. The per day sales of these categories was given in the data.

The dataset contains various attributes like Order date, Country, City, State, Region, Category, Sub-categories, Sales of the product, Quantity etc. An important point to note here is that the various sub-categories that are present in the dataset can be grouped under 3 main categories which are Office supplies, Furniture and Technology.

We are going to forecast the sales of these 3 main categories for the next 24 months to see how that each category is doing in sales.

3.2 Data Preprocessing

Out of 20 Columns or Attributes, 2 attributes have been selected for the further work. The data contained many categories but for this project we chose only 3 categories named 'Furniture', 'Office Sales', and 'Technology'. Based on project requirements, per day sales of each individual category is converted into per month sales, which helped to analyze data more precisely. Here, Average of monthly sales have been calculated.

3.3 Data Normalization

In the dataset, due to the high range of the sales value, normalization was performed. Normalization helps to rescaling real valued numeric attributes into the range 0 and 1. For the normalization, the MinMaxScaler method has been used, which is defined under the sklearn library. This method rescales the value using below formula.

$$y = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Here,
y = new value after normalization,
x_i = value before normalization,
min(x) and max(x) is the minimum and maximum value of the column respectively.

3.4 Training and Testing

For the training purpose, 70% of the data has been used and 30% of the data has been used for the testing purpose. Model performance have been evaluated by root mean square errors (RMSEs).

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(x_t - x_0)^2}{n}}$$

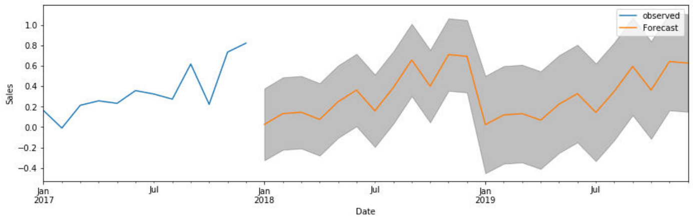
Here x_t is the forecasted observation and x_o is the actual observation.

3.5 Prediction

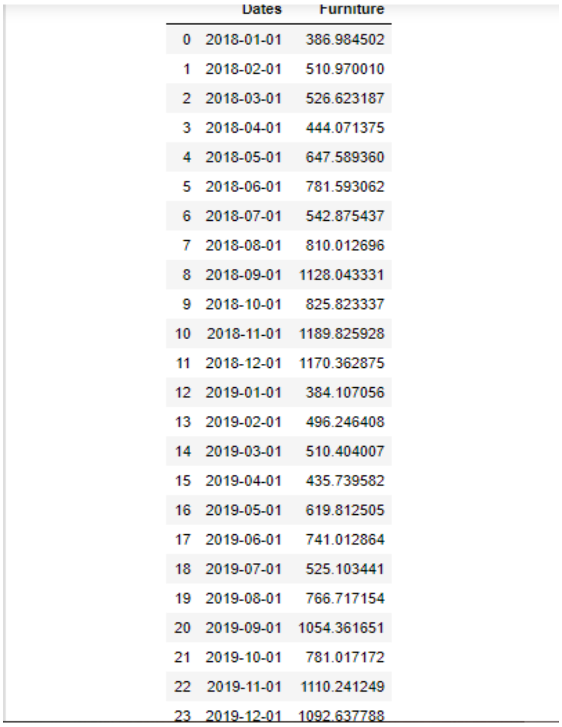
After the training and evaluation of ARIMA model, it is used for the prediction[14] of the future sales of the Super Store. The model is used to predict the sales for the next 24 months for each category. The predicted sales were in the range of 0 to 1 because of normalization performed on Sales attribute. To get the original value of sales, inverse_transform method has been used.

4. EXPERIMENTAL RESULTS

Furniture Future Sales

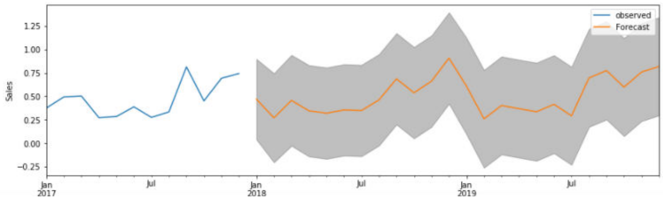


RMSE (Root Mean Square Error): 0.12521467915230222

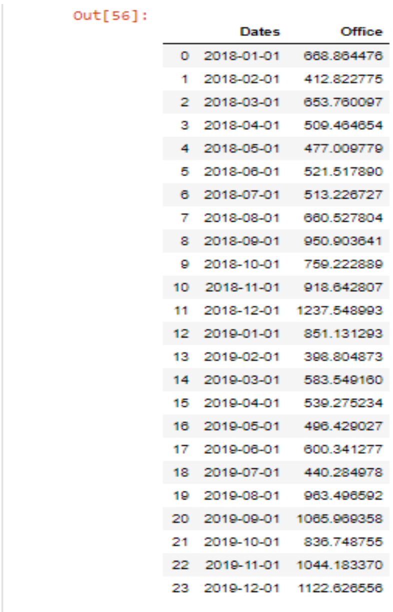


	Dates	Furniture
0	2018-01-01	386.984502
1	2018-02-01	510.970010
2	2018-03-01	526.623187
3	2018-04-01	444.071375
4	2018-05-01	647.589360
5	2018-06-01	781.593062
6	2018-07-01	542.875437
7	2018-08-01	810.012696
8	2018-09-01	1128.043331
9	2018-10-01	825.823337
10	2018-11-01	1189.825928
11	2018-12-01	1170.362875
12	2019-01-01	384.107056
13	2019-02-01	496.246408
14	2019-03-01	510.404007
15	2019-04-01	435.739582
16	2019-05-01	619.812505
17	2019-06-01	741.012864
18	2019-07-01	525.103441
19	2019-08-01	766.717154
20	2019-09-01	1054.361651
21	2019-10-01	781.017172
22	2019-11-01	1110.241249
23	2019-12-01	1092.637788

Office Supplies Future Sales



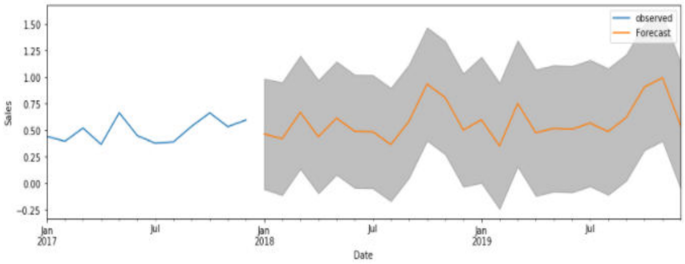
RMSE (Root Mean Square Error): 0.20697231345660533



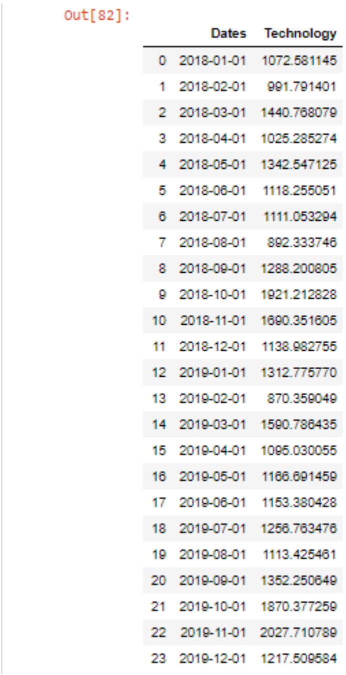
Out[56]:

	Dates	Office
0	2018-01-01	668.864476
1	2018-02-01	412.822775
2	2018-03-01	653.760097
3	2018-04-01	509.464654
4	2018-05-01	477.009779
5	2018-06-01	521.517890
6	2018-07-01	513.226727
7	2018-08-01	660.527804
8	2018-09-01	950.903641
9	2018-10-01	759.222889
10	2018-11-01	918.642807
11	2018-12-01	1237.548993
12	2019-01-01	851.131293
13	2019-02-01	398.804873
14	2019-03-01	583.549160
15	2019-04-01	539.275234
16	2019-05-01	496.429027
17	2019-06-01	600.341277
18	2019-07-01	440.284978
19	2019-08-01	963.496592
20	2019-09-01	1065.969358
21	2019-10-01	836.748755
22	2019-11-01	1044.183370
23	2019-12-01	1122.626556

Technology Future Sales



RMSE (Root Mean Square Error): 0.21206472246989572



Out[82]:

	Dates	Technology
0	2018-01-01	1072.581145
1	2018-02-01	991.791401
2	2018-03-01	1440.768079
3	2018-04-01	1025.285274
4	2018-05-01	1342.547125
5	2018-06-01	1118.255051
6	2018-07-01	1111.053294
7	2018-08-01	892.333746
8	2018-09-01	1288.200805
9	2018-10-01	1921.212828
10	2018-11-01	1690.351605
11	2018-12-01	1138.982755
12	2019-01-01	1312.775770
13	2019-02-01	870.359049
14	2019-03-01	1590.786435
15	2019-04-01	1095.030055
16	2019-05-01	1166.691459
17	2019-06-01	1153.380428
18	2019-07-01	1256.763476
19	2019-08-01	1113.425461
20	2019-09-01	1352.250649
21	2019-10-01	1870.377259
22	2019-11-01	2027.710789
23	2019-12-01	1217.509584

We have used RMSE (Root mean square Error) value as a parameter for determining the amount of error in the sales prediction for every category. If the RMSE value is between 0.2 and 0.5 then it can be said that our model can relatively predict the data accurately. As we can clearly see that the RMSE error for every category present here lies between the range of 0.2 and 0.5 or even less than that only. So, we can say that our model is really good for predicting the sales data.

5. CONCLUSIONS

After applying ARIMA[15] and SARIMA[11] technique to this Super Store data, we have observed that the category technologies have the highest sales among all the categories followed by office sales with second highest sales and then furniture sales at bottom. From this we can understand that technology items have the most potential for sale and are more profitable than office supplies which are in turn more profitable than Furniture. From the forecasts we've seen for the three sales categories we can say that the data is spread wide around the regression line making it difficult to predict accurate sales values. This can be improved with help of model boosting and further cleaning the data. Seasonal

ARIMA forecasting uses time series data, and this helps in avoiding problems that are associated with multivariate models. Seasonal ARIMA model was very useful in getting proper analysis and in getting better results. The future predictions for every category are plotted in the code.

ACKNOWLEDGEMENT

We take this opportunity to thank our teachers and friends who helped us throughout the project. First and foremost, I would like to thank my guide for the for her valuable advice and time during development of project.

REFERENCES

[1] P. Mondal, L. Shit, and S. Goswami, "Study of effectiveness of time series modeling (arima) in forecasting stock prices," *IJCSEA*, vol. 4, no. 2, pp. 1066–1067, 2014, doi: 10.1016/S0140-6736(01)13488-4.

[2] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[3] C. I. Permatasari, W. Sutopo, and M. Hisjam, "Sales forecasting newspaper with ARIMA: A case study," *AIP Conf. Proc.*, vol. 1931, no. February, pp. 030017–10, 2018, doi: 10.1063/1.5024076.

[4] K. Kumari and S. Yadav, "Linear regression analysis study," *J. Mood Disord.*, vol. 4, no. 21, pp. 33–36, 2018, doi: 10.5455/jmood.20130624120840.

[5] G. H. Nguyen, J. Kedia, R. Snyder, R. D. Pasteur, and R. Wooster, "Sales Forecasting Using Regression and Artificial Neural Networks," *Midstates Conf. Undergrad. Res. Comput. Sci. Math.*, no. August 2015, 2013, [Online]. Available: http://www.researchgate.net/publication/280742365_Sales_Forecasting_Using_Regression_and_Artificial_Neural_Networks.

[6] R. A. Abbasi *et al.*, *Short term load forecasting using xgboost*, no. March. Springer International Publishing, 2019.

[7] M. H. Alsharif, M. K. Younes, and J. Kim, "Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of seoul, south korea," *Symmetry (Basel).*, vol. 11, no. 2, pp. 1–17, 2019, doi: 10.3390/sym11020240.

[8] J. Y. Kajuru and M. M. Muhammed, "Forecasting performance of arima and sarima models on monthly average temperature of zaria , nigeria," *J. sciecnce Technol. Educ.*, vol. 7, no. 3, pp. 205–212, 2019.

[9] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 74, no. November, pp. 902–924, 2017, doi: 10.1016/j.rser.2017.02.085.

[10] V. Cerqueira, L. Torgo, and C. Soares, "Machine learning vs statistical methods for time series forecasting: Size matters," *arXiv*, vol. 1, no. 1, pp. 1–9, 2019.

[11] Q. Xu and V. Sharma, "Ensemble sales forecasting study in semiconductor industry," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10357 LNAI, pp. 31–44, 2017, doi: 10.1007/978-3-319-62701-4_3.

[12] B. M. Pavlyshenko, "machine-learning models for sales time series forecasting," *Data*, vol. 4, no. 1, pp. 1–11, 2019, doi: 10.3390/data4010015.

[13] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia, and N. Hanafiah, "Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 524–532, 2021, doi: 10.1016/j.procs.2021.01.036.

[14] K. M. Sabu and T. K. M. Kumar, "Predictive analytics in Agriculture: Forecasting prices of Arecanuts in Kerala," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 699–708, 2020, doi: 10.1016/j.procs.2020.04.076.

[15] A. Ghangare and S. Singh, "ARIMA modelling using machine learning for selected indian automobile listed companies.," *Stud. Rosenethaliana*, vol. XII, no. IV, pp. 1–22, 2020.