Data Science based Supermarket Sales Forecasting for better Inventory Optimization

An innovative approach for Inventory Optimization

Prathamesh Kurunkar (*Author*)
International Institute of Information Technology
Pune, Maharashtra, India
kurunkarprathamesh7@gmail.com

Onkar Bharat Sawant (*Author*)
International Institute of Information Technology
Pune, Maharashtra, India
onkarsawant5613@gmail.com

Nikhil Varghese (*Author*)
International Institute of Information Technology
Pune, Maharashtra, India
nikhilajivarghese@gmail.com

Abstract**—**Sales Forecasting is a technique that indicates how much of an individual product is expected to be sold with respect to demand in the market in a stipulated future period at a certain price. Inventory Optimization is a process of managing a business' Inventory in the most optimal manner. Better Sales Forecasting and Analysis lead to businesses having better optimization of their respective inventories. Finer management of inventories is beneficial for businesses as it greatly impacts their profit margins and adds value to the organization. The model explained in this research piece uses SARIMAX(*Seasonal autoregressive integrated moving average with external variables*) method for performing Time Series Forecasting on Sales of a supermarket. Further we extend the model on a web application which showcases Sales analytics, insights through interactive visualizations and performs forecasting for improved inventory optimization.

Keywords — Time Series Forecasting, Sales Forecasting, Sales Analysis,Inventory Optimization, SARIMA, visualizations.

I. Introduction

Any Forecasting can be explained as a measure for what is probable in terms of prediction in specified future time. It suggests how much of a particular product will likely be sold in the specified future period in a specific market at a particular price. Accurate sales forecasting is essential for business nowadays to make sure it produces the required quantity of a particular product at the right time. Thus we can define sales forecasting as, estimation of type, quantity and quality of future sales from in depth analysis of existing data. Better Sales Forecasting and Analysis would lead to better inventory optimization. In elementary terms, inventory management is a combination of all those processes which businesses avail themselves of to oversee and structure their goods or materials in their premise. It's high time for businesses to streamline their Inventory Management processes if they are still using age-old methods of maintaining ledger systems.

In order to achieve a more efficient way of managing an Inventory, our project brings Sales Analysis, Visualization, and Forecasting for better Inventory Optimization. This implementation is based on a publicly available dataset of a Supermarket. This project will help in optimizing the inventory based on various factors considered while working on the available dataset, such as seasonal trends, discounts, in-demand products, external factors and so on. Optimizing the inventory of the Supermarket based on the analysis performed in this project will definitely help in making a profitable business. [9]

II. Related work

A. What is the need?

Forecasting is essential for several aspects of a contemporary business. According to several literary works the word forecasting can be defined in the following manner:

"Forecasting is predicting, projecting, or estimating some future event or condition which is outside an organization's control and provides a basis for managerial planning" [6]. Sales Forecasting has been one such application or branch of Time series Forecasting alone. Although a substantial number of research works are reported for Sales forecasting techniques and Inventory Optimization techniques,not many speak about applying both together. Being correlated, these topics tend to widely affect each other. For sales forecasting purposes, statistical techniques, such as ARIMA,exponential smoothing, Box & Jenkins model, regression models or Holt-Winters model, are usually applied.

Hybrid routines appear to be precise in income forecasting [1]. Xia and Wong [2] proposed the variations among classical methods which were primarily based on mathematical and statistical methods and suggested adapting to new contemporary heuristic methods. In the primary group, the call exponential smoothing, regression, Box–Jenkins[8], autoregressive included shifting moving average (ARIMA), generalized autoregressive conditionally heteroskedastic (GARCH) methods[3]. Most of these models are linear and are not able to cope with the asymmetric behaviour in most of the real-world sales data. Hence, we considered the seasonal tendency of Sales and applied a SARIMAX model,making the often used ARIMA model much more capable.

Bohdan M. Pavlyshenko [4]says that there exist a few obstacles during the time collection processes for income forecasting. Here are a number of them listed:

a)The need to have historic statistics for a long term length to seize seasonality is often forgotten. However, there are times when we cannot have the past data for a target parameter, for example in cases where we launch a fresh product. At the same time we have sales time analysis of a series for a similar product and we can expect that our novel product will have an equivalent sales effect. Sales data can have a lot of misplaced data that is difficult to gauge. The need to clean such adversative and exceptive data before using a time series approach is vital and unavoidable. We have to ease the outliers and interpolate statistics mentioned prior in order to make the data more comprehensive so as to pertain to a time collection approach.

B. Algorithm:

For analysis purposes we have to consider various dependent and independent factors which have an impact on sales. Waters (2003) [5]proposes a model for this type of usage, under some specific circumstances such as seasonality and trend in the demand. Demand can be divided into separate parts which more specifically are mentioned as: a) underlying value, which identifies the main demand that should be adjusted for trend and seasonality b) trend which is the change in demand, c) seasonality which is the variation occurring after a certain time interval around the trend and finally d) noise- which is considered to be a random effect. Taking several such points into consideration we decided to opt for a **SARIMAX model**.

III. Methodology

A. Brief Description of Algorithms Used:

Sarima(*Seasonal autoregressive integrated moving average*):

Autoregressive Integrated Moving Average abbreviated as ARIMA, this is one of the most popularly used forecasting methods when we deal with the data forecasting of univariate time series. Although the ARIMA can handle data with a specific trend, it does not support data with a seasonal component. For understanding the seasonal trend consider this example - On Sundays/Holidays(Christmas) sales are more than the regular days. So for the purpose of dealing with the seasonal component SARIMA or Seasonal ARIMA is used, it extends ARIMA by handling seasonal components on top of handling univariate time series data.

ARIMA expects data that is either not seasonal or has the seasonal factor removed, e.g. seasonally adjusted through strategies which include seasonal differencing.

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA is an extended version of ARIMA which manages to add three new hyperparameters; which are a). autoregression (AR) b).differencing (I) c). moving average (MA) for handling seasonal components of the time series, and also an additional parameter for the period of the seasonality.

The ARIMA model is considered a legit alternative for non-seasonal non-stationary bound data. Box and Jenkins proposed [6] to have a generalized approach for this version to cope with the seasonality factor. Their proposed version is called the Seasonal ARIMA (SARIMA) approach. In this method seasonal differencing of suitable order is used to get rid of non-stationary data from the existing training data. A first-order seasonal distinction is the distinction amongst a study and the corresponding finding for the preceding 12 months and is calculated as zt = yt -yt-s . For month-to-month time series s = 12 and for quarterly time series = 4. Thus,this model is typically termed as the SARIMA(p, d, q)× (P, D, Q)^s model.

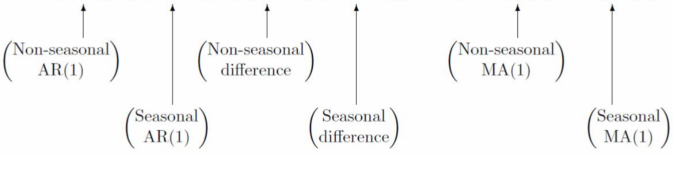
C. Equation:

To put SARIMA in mathematical equation we use (p,d,q)× $(P, D,Q)^s$ model in terms of lag polynomials[10]:

SARIMA (p,d,q)(P,D,Q)m
P (AutoRegressive)
D (Integrated)
Q (Moving Average)
M (seasonal factor)

p,d,q in lowercase is used for non seasonal
P,D,Q in uppercase is used for seasonal

Example:- SARIMA(1,1,1) (1,1,1) 4 :-



$$(1-\phi_1 B)\,(1-\Phi_1 B^4)\,(1-B)\,(1-B^4)y_t = (1+\theta_1 B)\,(1+\Theta_1 B^4)e_t.$$

Fig. 1: Equation of SARIMA Model [7]

D. Model diagram:

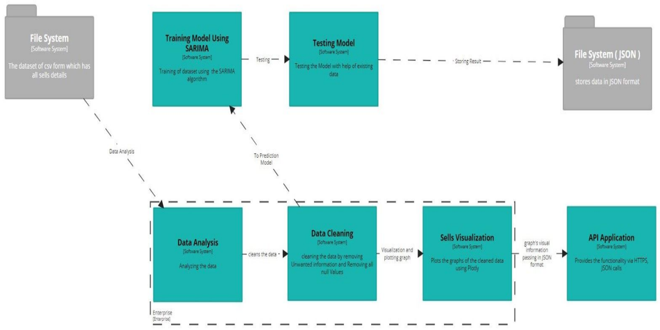

File System
[Software System]
The dataset of csv form which has all sells details
Training Model Using SARIMA
[Software System]
Training of dataset using the SARIMA algorithm
Testing
Testing Model
[Software System]
Testing the Model with help of existing data
Storing Result
File System (JSON)
[Software System]
stores data in JSON format
Data Analysis
To Prediction Model
Data Analysis
[Software System]
Analyzing the data
cleans the data
Data Cleaning
[Software System]
cleaning the data by removing Unwanted information and Removing all null Values
Visualization and plotting graph
Sells Visualization
[Software System]
Plots the graphs of the cleaned data using Plotly
graph's visual information passing in JSON format
API Application
[Software System]
Provides the functionality via HTTPS, JSON calls
Enterprise
[Enterprise]

Fig. 2: Level 3 Architecture Diagram

E. Dataset Used:

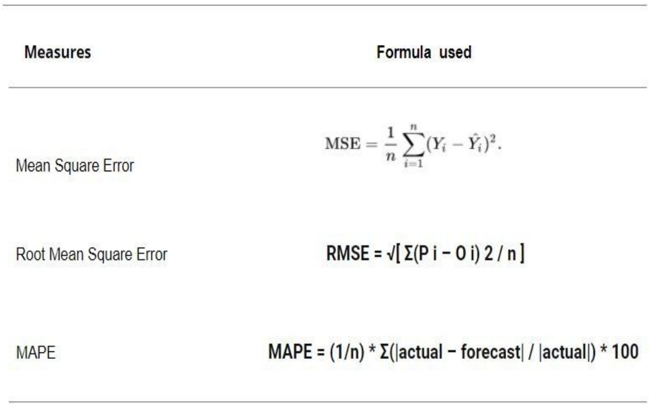
We used the dataset of American Supermarkets which was openly available. Dataset contained various parameters on which the sells are dependent such as Quantity, Discount given, profit, Fuel price, Temperature etc.

The dataset contains major 3 categories and 17 subcategories.

click here to visit dataset

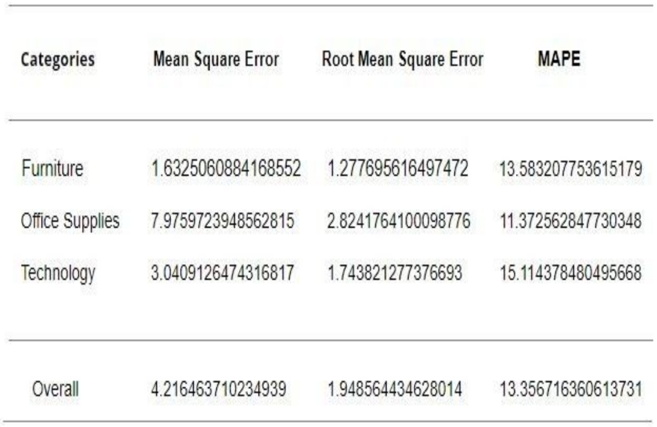
F. Accuracy measures:

1. Mean Square Error
2. Root Mean Square Error
3. MAPE



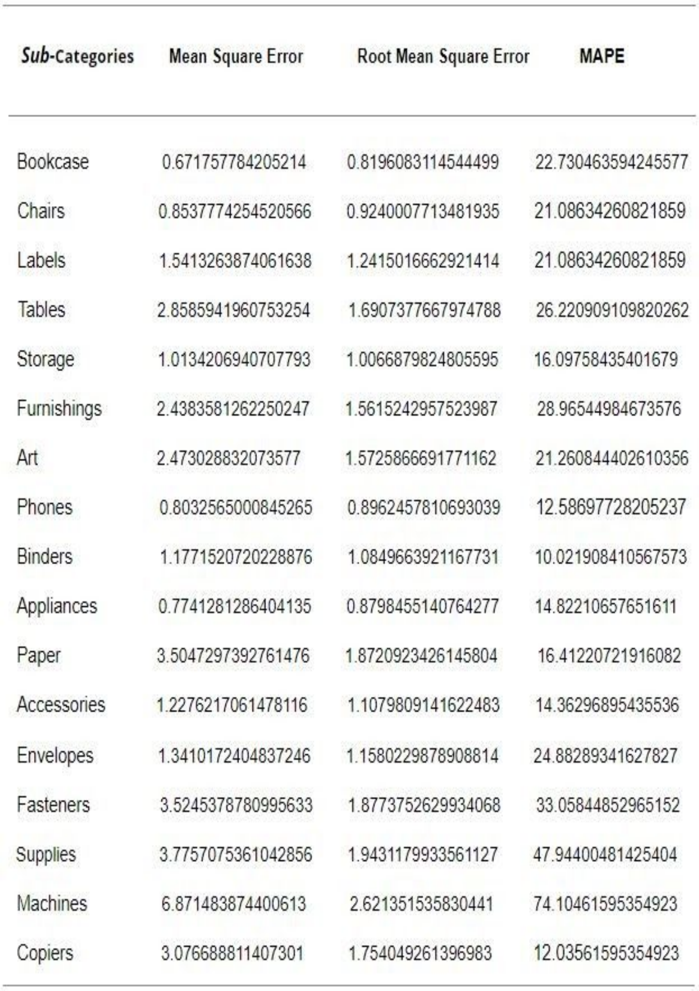
Measures	Formula used
Mean Square Error	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$
Root Mean Square Error	RMSE = √[Σ(P i – O i) 2 / n]
MAPE	MAPE = (1/n) * Σ(|actual – forecast| / |actual|) * 100

Fig. 3: Error Measures and Formulae



Categories	Mean Square Error	Root Mean Square Error	MAPE
Furniture	1.6325060884168552	1.277695616497472	13.583207753615179
Office Supplies	7.9759723948562815	2.8241764100098776	11.372562847730348
Technology	3.0409126474316817	1.743821277376693	15.114378480495668
Overall	4.216463710234939	1.948564434628014	13.356716360613731

Fig. 4: Error measures and categorical results



Sub-Categories	Mean Square Error	Root Mean Square Error	MAPE
Bookcase	0.671757784205214	0.8196083114544499	22.730463594245577
Chairs	0.8537774254520566	0.9240007713481935	21.08634260821859
Labels	1.5413263874061638	1.2415016662921414	21.08634260821859
Tables	2.8585941960753254	1.6907377667974788	26.220909109820262
Storage	1.0134206940707793	1.0066879824805595	16.09758435401679
Furnishings	2.4383581262250247	1.5615242957523987	28.96544984673576
Art	2.473028832073577	1.5725866691771162	21.260844402610356
Phones	0.8032565000845265	0.8962457810693039	12.58697728205237
Binders	1.1771520720228876	1.0849663921167731	10.021908410567573
Appliances	0.7741281286404135	0.8798455140764277	14.82210657651611
Paper	3.5047297392761476	1.8720923426145804	16.41220721916082
Accessories	1.2276217061478116	1.1079809141622483	14.36296895435536
Envelopes	1.3410172404837246	1.1580229878908814	24.88289341627827
Fasteners	3.5245378780995633	1.8773752629934068	33.05844852965152
Supplies	3.7757075361042856	1.9431179933561127	47.94400481425404
Machines	6.871483874400613	2.621351535830441	74.10461595354923
Copiers	3.076688811407301	1.754049261396983	12.03561595354923

Fig. 5: Error Measures and Sub-Categorical Results

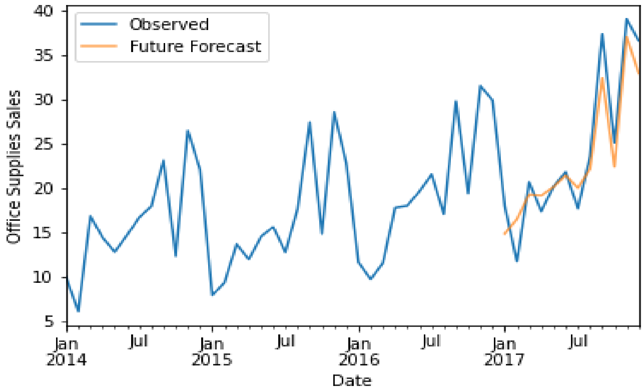


Fig. 6: Comparing Present And Actual Results

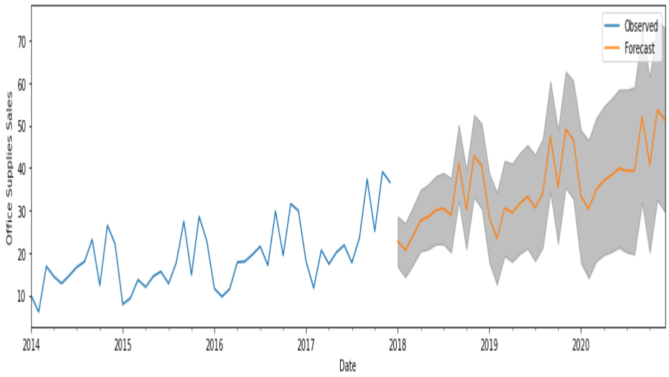
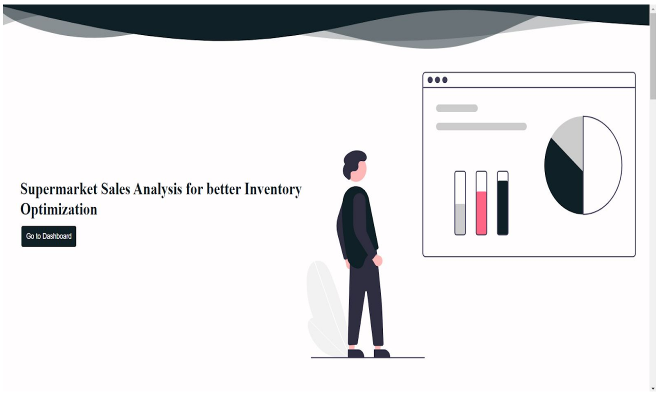
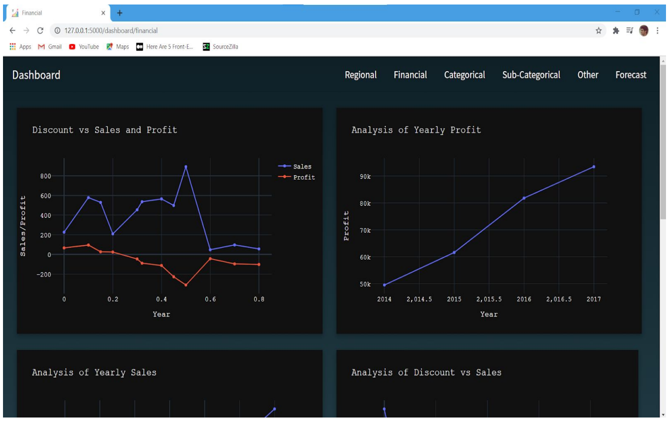


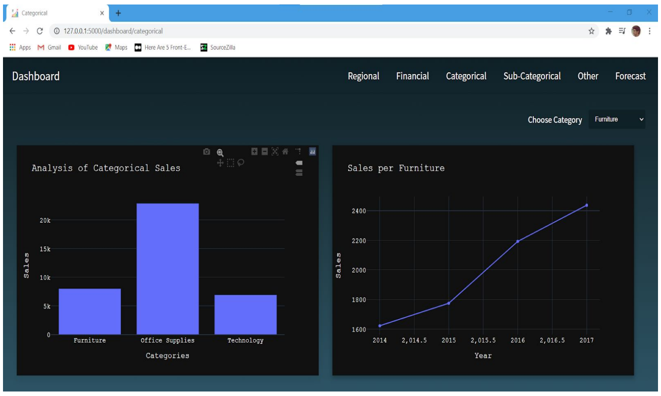
Fig. 7: Predicting Results of Next 3 Years

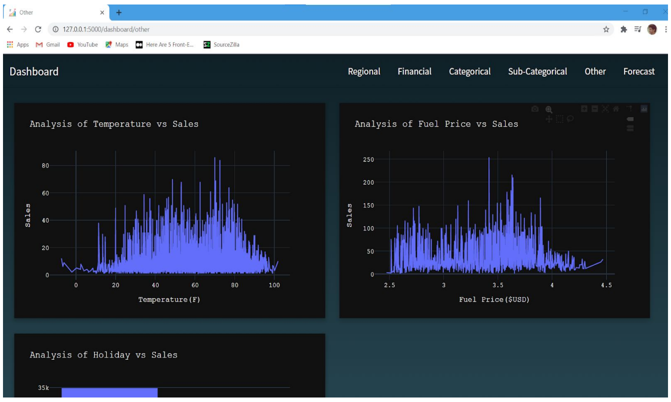
G. RESULTS

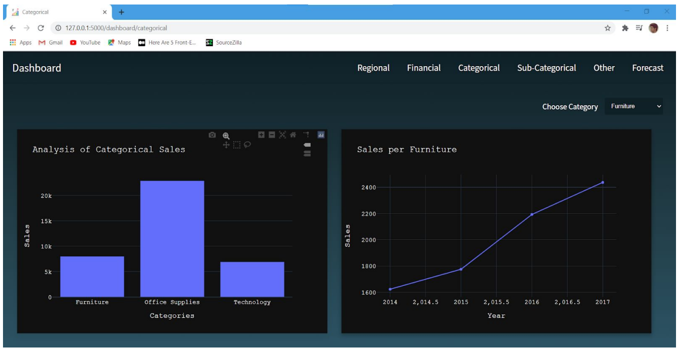
Here are some images from our web application.

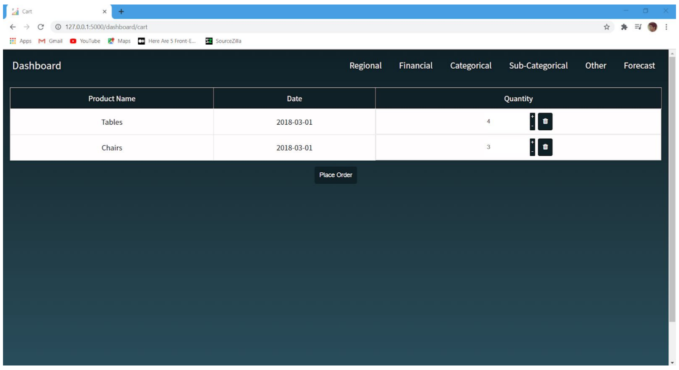

Supermarket Sales Analysis for better Inventory Optimization



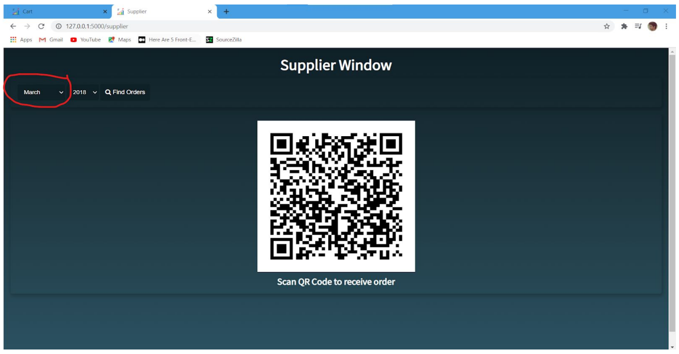



Other
127.0.0.1:5000/dashboard/other
Dashboard
Regional
Financial
Categorical
Sub-Categorical
Other
Forecast
Analysis of Temperature vs Sales
Sales
Temperature (F)
Analysis of Fuel Price vs Sales
Sales
Fuel Price ($USD)
Analysis of Holiday vs Sales





Product Name	Date	Quantity
Tables	2018-03-01	4
Chairs	2018-03-01	3



Acknowledgement

Thanks to Prof. Prashant Gadakh Assistant Professor Computer Engineering, Hope Foundation's International Institute of Information Technology, for immense support and guidance.

References

Here are some references we referred during our Research

[1] Aburto & Weber; Improved supply chain management based on hybrid demand forecasts;2007

[2] Xia and Wong; Time Series Forecasting with Arima Models. Beijing

[3] Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time Series Analysis: Forecasting and Control; John Wiley & Sons: Hoboken, NJ, USA, 2015 Chatfield, C. Time-Series Forecasting; Chapman and Hall/CRC: Boca Raton, FL, USA, 2000. 5. Brockwell, P.J.; Davis, R.A.; Calder, M.V. Introduction to Time Series and Forecasting; Springer: Cham, Switzerland, 2002; Volume 2

[4] Bohdan M. Pavlyshenko; Machine-Learning Models for Sales Time Series Forecasting;2018

[5] Waters (2003), Mentzer, J.T.; Moon, M.A. Sales Forecasting Management: A Demand Management Approach; Sage: Thousand Oaks, CA, USA, 2004.

[6] Golden J. et.al, 1994, p.33

[7] time series - SARIMA model equation - Cross Validated (stackexchange.com)

[8] J. Lee, "Univariate time series modelling and forecasting (Box-Jenkins Method)", Econ 413, lecture 4 He, S.Y., 2004. Applied Time Series Analysis. 1st Edn., Peking University Press, Beijing.

[9] Cryer, J. D. and K.S. Chan, 2008. Time Series Analysis with Application in R. 2nd Edn., Springer, New York, ISBN-10: 0387759581, pp: 491

[10] Chatfield, C 2004, The Analysis of Time Series: An Introduction, 6th ed., Chapman & Hall/CRC, Boca Raton, Fla