Retail Giant Sales Forecasting using Machine Learning

Ashwini Rekha. Banjanagari, Vijaykumar. B

Abstract---Sales forecasting is widely recognized and plays a major role in an organization's decision making. It is an integral part in business execution of retail giants, so that they can change their strategy to improve sales in the near future. This helps in better management of their resources like machine, money and manpower. Forecasting the sales will help in managing the revenue and inventory accordingly. This paper proposes a model that can forecast most profitable segments at granular level. As most retail giants have many branches in different locations, consolidation of sales are hard using data mining. Instead using machine learning model helps in getting reliable and accurate results. This paper helps in understanding the sales trend to monitor or predict future applicable on different types of sales patterns and products to produce accurate prediction results.

Index Terms—Machine Learning, ARIMA, sale-forecasting, smoothing, COV and classical decomposition.

I. INTRODUCTION

Most of the organizations need forecasting the reason for such a forecast is the customers would need a sufficient lead time for the products to be delivered and they do not want to wait too long. So, it is very obvious the organizations need to predict what the future demand would be and they need to have sufficient stock on hand so as to reduce time to deliver the product to the customer. Time series analysis is used in this project which is based on time stamped data.

II. PROBLEM DEFINITION

Global Mart is an online store having widespread operations which takes orders and transport across the earth and deals with all the major product categories - consumer, corporate & home office. Sales manager responsibility is to finalize the plan for the next 6 months. Sales and the demand has to be forecasted for the next six months, that would help in managing the expenditure and inventory accordingly. The store caters to 7 different market segments APAC (Asian pacific), Africa, EU (European Union) Canada, Australia, EMEA (Europe, Middle East and Africa) and America and in the 3 major categories. Forecast should be at the granular level, so the data has to be subsetted into 21(7*3) buckets before analyzing it. But not all of these 21 market buckets are important from the store's point of view. So we need to find out 2 most profitable and consistent segment from these 21 and Estimate the purchase and demand for these segments.

Revised Version Manuscript Received on 10 September, 2019.

Ashwini Rekha. **Banjanagari**, Information Technology, G. Narayanamma Institute of Technology and Science, Hyderabad, India.
(Email: ashuashwini682@gmail.com)
Vijaykumar. B, Information Technology, G. Narayanamma Institute of Technology and Science, Hyderabad, India.
(Email: vijayballa504@gmail.com).

III. DATA UNDERSTANDING

Now the data is currently under input level data. There is only one data set for this project that is Global superstore this is a CSV file. This is obtained from the information provided by the global mart online store. It contains the information of attributes of that company like date and sales of the products like order date of the products in which market segment they ordered how much profit they received of those products of all 7 market segments. We just consider only relevant data attributes for the analysis then check the data. Pick important variables such that increase revenue and manage inventory we need to perform data cleaning.

TABLE I ATTRIBUTES AND THEIR DESCRIPTION

Attribute	Description
Order date	Order was placed on that particular day.
Segment	Customer belongs to that particular market section
Market	Topographical market segment where the customer belongs to.
Sales	Final dealing worth of the transaction.
Quantity	Amount of that product ordered.
Profit	Profit produced on that transaction.

Exploratory Data analysis has to be performed after data understanding in EDA we first segregate the market segment based on sales and quantity As mentioned in the problem statement we need to analysis 21 buckets to find the best 4 in terms of profitability. Either we can go for creating 21 subset of the main dataset or we can use some in build function for consumer level, corporate level and home office level.

Then combine the market and segment variable and check the levels of concatenate variable then first create a first subset of Africa Consumer variable. Perform the function of coefficient of variation calculation then order the matrix. Check the data frame if any data frame rows are empty lets remove those empty rows. Then order the matrix and find the most profitable segments using coefficient of variation.

IV. DATA PREPARATION

We have to convert the transaction-level data into time series. The Cov (coefficient of variation). Cov is equal to standard deviation value divided by the mean value. This is used to find the most profitable segments. Segments with least Cov values are the most profitable segments, hence APAC consumer and EU consumers are considered.

TABLE II CALCULATING COV VALUES

Market	COV
APAC Consumer	0.6036
Africa corporate	1.6850
EU Corporate	0.6977
APAC Corporate	0.7407
LATAM Corporate	0.8909
LATAM Consumer	0.6889
US Corporate	1.0396
APAC Home office	1.0615
US Consumer	1.1085
EU Home Office	1.1281
US Home Office	1.02318
LATAM Home office	1.3599
Africa Consumer	1.4466
Africa Home Office	2.0139
EU Consumer	0.6553
EMEA Consumer	2.7499
EMEA Home Office	6.1402
EMEA Corporate	6.8618

Flow of paper

V. MODEL BUILDING

Once you arrive at the 2 most profitable segments, the next challenge is conjecture the commerce and quantity for the next 6 months. You are supposed for using classical decomposition and auto ARIMA for forecasting. Also, it is advised that you smoothen the data before you perform classical decomposition. Building a model on the smoothed time series using classical decomposition for that we need to convert the time series to a data frame then fit a multiplicative model with trend and seasonality to the data. The components of the time series are trend, seasonality, cyclicity and noise we need remove these from the data and make data stationary.

The seasonality will be modeled using a sinusoid function which means the function is like a sin function which is stretched or compressed using sin function. Now we need to look at the locally predictable series then we will model it as an ARMA series we check for Autocorrelation Factor (ACF) and partial autocorrelation factor (PACF) these are used to check the stationarity and seasonality of time series. Then we'll check if the residual series is white noise using augmented dickey fuller test (ADF) and Kwiatkowski Phillips Schmidt shin (KPSS) test.

VI. MODEL EVALUATION & RESULTS

Once you come up with a satisfactory model, the next step would be to prognosticate vending/demand for next 6 months using this model. For testing accuracy of your forecast, you must initially separate out the last 6 months values from your dataset after aggregating the transaction level data into the monthly data. Then check your 6 months forecast using the out-of-sample figures.

To determine the best performing model we evaluate the model using Mean Absolute Percentage Error (MAPE) and make a prediction for the next six months then compare our predictions with actual values, using MAPE. To get a visual feel of the fit we plot the predictions along with the original values till now we performed classical decomposition. Arima fit also performed and evaluated using MAPE. Same as the sales we need to perform for the demand and then we need to consider which model is better fit for forecasting this online store.

Fig 1. Forecasting Sales and demand- APAC consumer Sales

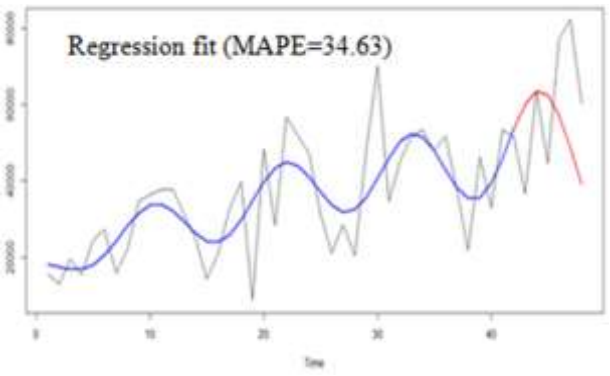

Fig 1.a Regression fit for APAC Consumer Sales

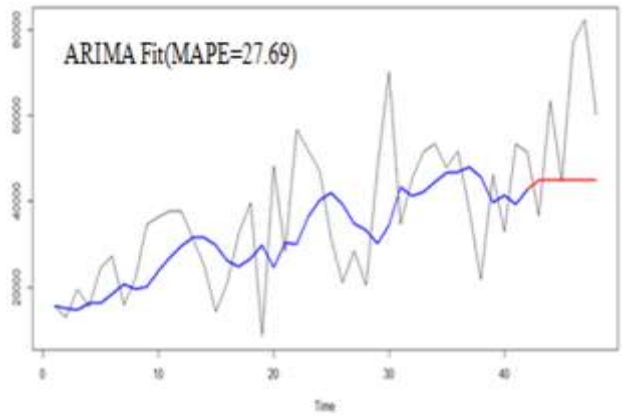

Fig 1.b ARIMA fit for APAC Consumer Sales Original Data

--- Original Data
--- Modeled Data
--- Forecasted Data.

In figure 1 between regression fit and ARIMA fit former looks a better fit visually. However, the ARIMA fit has lower MAPE value.

Fig 2. Forecasting Sales and demand -EU consumer sales

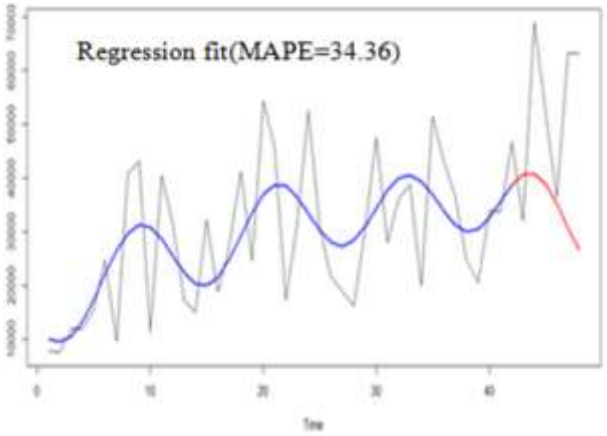

Fig 2.a Regression fit for EU consumer sales

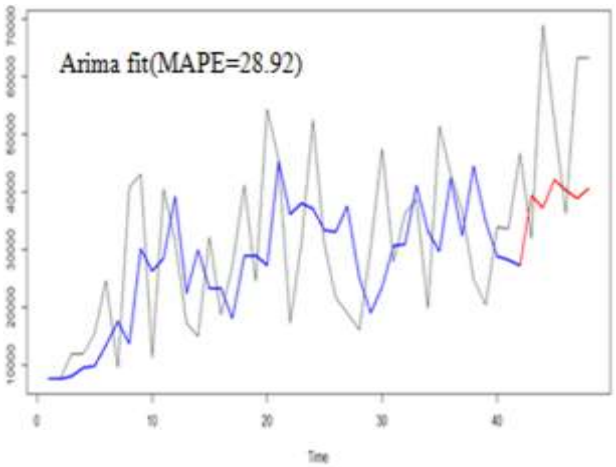

Fig 2.b ARIMA fit for EU consumer sales

In figure 2 visually both regression fit and arima fit look similar however arima fit has a lower mape value.

Fig 3. Forecasting Sales and Demand – APAC Consumer Demand

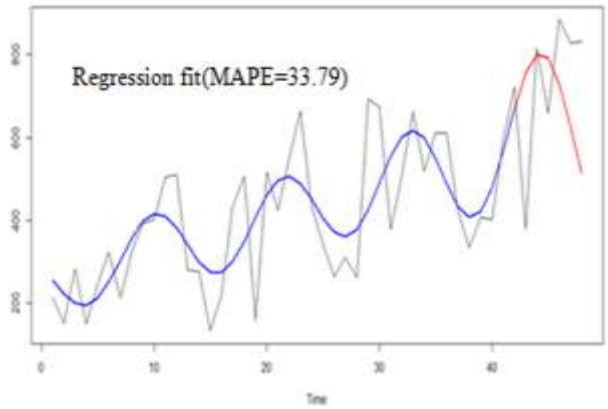

Fig 3.a Regression fit for APAC consumer demand

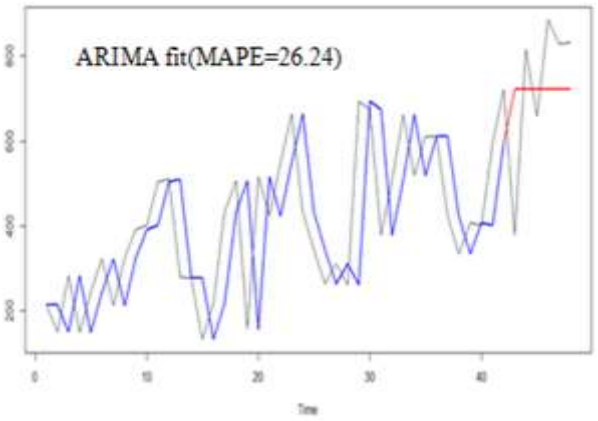

Fig3b. ARIMA fit for APAC consumer demand

In figure 3 between the regression fit and the ARIMA fit, the latter looks like a better fit visually. As expected, the ARIMA fit has a lower MAPE value.

Fig 4. Sales and Demand – EU Consumer Demand Forecasting

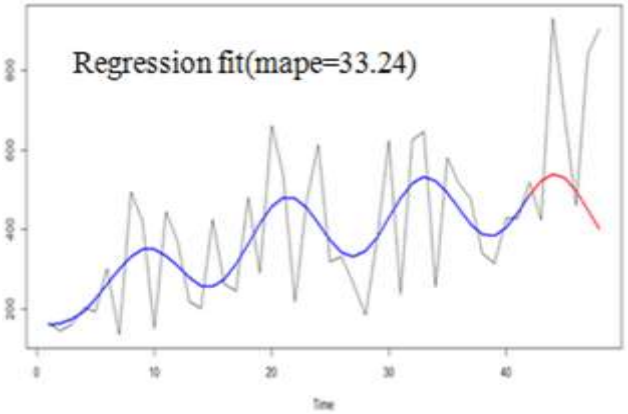

Fig 4.a. Regression fit for EU consumer Demand

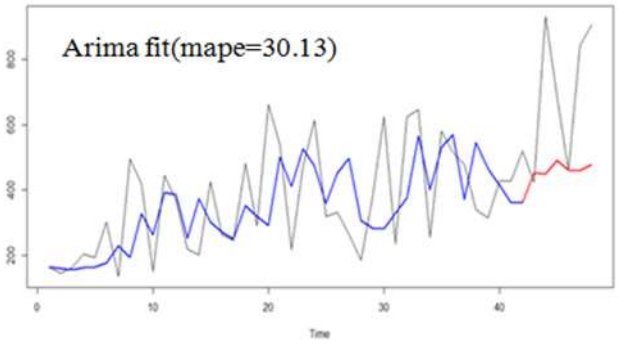

Fig 4.b ARIMA fit for EU consumer Demand

In figure 4 both the regression fit and the ARIMA fit looks like very bad fits, when looked at visually however, the ARIMA fit has a lower MAPE value.

VII. MOST PROFITABLE SEGMENTS

APAC and EU consumer segments seem to be the most profitable ones. So these 2 we need to consider first and manufacture more compare to other.

TABLE III APAC AND EU MAPE VALUES

Market Segment	Model	Sales	Demand
APAC	ARIMA fit (MAPE)	27.69	26.24
	CD fit (MAPE)	34.63	33.79
EU	ARIMA fit (MAPE)	28.92	30.13
	CD fit (MAPE)	34.36	33.24

VIII. CONCLUSION

APAC and EU consumer segments seem to be the most profitable ones. So in stock keep these items a bit more than other. Inventory levels should be kept as predicted by the ARIMA model (around 400 units) for the case of EU consumer segment, since the ARIMA model's predictions had a low MAPE value. However, the regression model should be used for predicting inventory requirement for the APAC consumer segment, as it is the only one of the two that is able to capture the seasonal behavior of sales and demand for this segment. In general, a buffer of at least 25% should be kept on inventory levels, as none of the models used was extremely accurate (lowest MAPE value was 42.24).

IX. ACKNOWLEDGMENTS

I would like to convey my special thanks of gratefulness to my guide B. Vijay Kumar sir, is an Assistant Professor in G. Narayanamma Institute of Technology and Science, as well as our head of the Department Information technology. Dr. I. Ravi Prakash Reddy in G. Narayanamma Institute of Technology and Science, who gave me a valuable opportunity to do this wonderful project, which also helped me in doing a lot of Research and I came to know new things. I am really thankful to them. Secondly. I would also like to thank my parents and friends who helped me a lot in finalizing this paper within the limited time frame.

REFERENCES

1. Pine, B.J. & Gilmore, J.H. The Experience Economy[M]. Harvard Business School Press, Boston, Massachusetts, 2000.
2. Liu Guohao, "Enterprise innovation system analysis of the competitiveness of the market [J]", The vitality of enterprises, pp. 30-31, 2003.
3. C.W.J. Granger, R. Ramanathan, "Improved Methods of Forecasting". Journal of Forecasting 1984,(3) pp. 197-204
4. Heng Liu "The analysis in Direct Marketing of E-Commerce mall in cosmetic", No. 9, 2011, Serial No. 207.
5. Mohit Gurnani , Yogesh Korkey, Prachi Shah, Sandeep Udmale, Vijay Sambhe, and Sunil Bhirud" Department of Computer Engineering and Information Technology",
6. J.Contreras, R. Espinola, F. J. Nogales and A. J. Conejo, "ARIMA models to predict next-day electricity prices", in IEEE 1014-1020 Transactions on Power Systems, vol. 18, no. 3, pp. Aug. 2003.
7. Mohit Gurnani , Yogesh Korkey, Prachi Shah, Sandeep Udmale Vijay Sambhe, and Sunil Bhirud Department of Computer Engineering and Information Technology, VJT, Mumbai,Zeal Education Society, Pune, India, Feb 24-26,
8. H.K. Temraz, M. M. A. Salama and V. H. Quintana "Application of the decomposition technique for forecasting the load of a large electric power network," in IEEE Proceedings - Generation, Transmission and Distribution, vol . 143, no. 1, pp. 13-18, Jan 1996.