

Sales Analytics to drive profitability - A case study of a Fashion E-Commerce Retailer

Tharuka Gallekankanamge
REVA Academy for Corporate Excellence,
REVA University
Bengaluru, India
tharuka.ba07@race.reva.edu.in

JB Simha
REVA Academy for Corporate Excellence,
REVA University
Bengaluru, India
jb.simha@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence,
REVA University
Bengaluru, India
rashmi.agarwal@reva.edu.in

Abstract— This research is based on a global fashion retailer that entered India in 2021. This retailer specializes in value fashion format. The value fashion industry in India. Due to the vast diversity in the market, it has become vital that retailers have an Omni presence to cater to the needs of Indian consumers. Today, retailers are thriving both offline and online. Indian retail has shown that being a Direct to Consumer (D2C) brand is important. Also, social commerce is booming, and this is due to India having a younger population compared to the rest of the world.

This study lays the foundation for introducing an analytical culture to the organization. Various aspects of profitability have been covered in this study. The Indian value fashion market is heavily penetrated by competition. Hence it is vital to keep an eye on the Cost of Doing Business (CODB) and overall profitability. Value fashion brands work on seasonal collections and the Stock Keeping Units (SKUs) in a season are large to ensure all categories are served. Especially when the brand is new to the target market, it is important to find its pros and cons, and Unique Selling Propositions and create a niche in the industry.

A thorough analysis is conducted to find out what are the key drivers of e-commerce profitability. How are discounts run on the platform affecting the net profit? What are the top categories contributing to a healthy net profit? This study aims to answer all these questions and derive meaningful insights keeping in mind that there is a CODB in the online channel. Industry and business research are carried out to ensure the insights derived from this study and well thought through. It is vital to know the domain and gather enough knowledge of the industry to carry out this study. This is conducted by speaking to the department heads and business heads of the organization's Indian Branch.

The main aim is to develop a framework for predicting online sales and convincing the management to use analytical tools and techniques for the betterment of the organization. This study will be a stepping-stone in creating an analytical culture in the organization. The key insights derived from this study will be passed on to the e-commerce and operations teams that will deploy the suggestions derived from this study.

Forecasting techniques such as (Auto Regressive Integrated Moving Average) ARIMA and Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX) are utilized for sales prediction. Overall, a comprehensive approach is taken where most of the weightage is given to the Exploratory Data Analysis (EDA) to derive meaningful and insightful information. The reason for giving more weightage to EDA is to ease the analytical adaptation to

ensure concepts are simple and practical to the targeted audience.

Keywords— *E-commerce, Value Fashion, Trend Analysis, profitability, Sales Analytics, Retail, Sales Forecasting.*

I. INTRODUCTION

The retailer considered for this study is a private limited company that specializes in value fashion format in women's wear, men's wear, kids', footwear, and accessories. The value fashion industry in India is booming and India has become a very promising market for both local and international fashion retailers. Due to the vast diversity in the market, it has become vital that retailers have an Omni presence to cater to the needs of Indian consumers. Today, retailers are thriving both offline and online.

The sales data gathered are specific to one online portal which started its operations in January 2022. The brand is operating on "a marketplace model" with one of the major online giants in the country. This means the inventory is owned by the brand and the products are only listed on the online platform. This model is the most preferred mode of operating in e-commerce due to its easy scalability and high turnaround time of working capital.

The focus is given to creating a process for online sales forecasting and encouraging the management to adopt analytical ways of seeing data for the betterment of the organization. The insights and recommendations from this study will be passed on to the e-commerce and operations teams that will deploy the suggestions derived from this study.

ARIMA and SARIMAX forecasting techniques are thoroughly explored for predicting sales. A holistic approach is followed, and more focus and weightage are given to the EDA section. This helps in introducing analytical concepts within the company.

II. LITERATURE REVIEW

Both industry and subject-related research has been carried out to acquire complete knowledge on the main subject of this study.

A. Retail industry research

To succeed in formulating an effective social media strategy which is an exceedingly challenging task since the number of platforms the strategy needs to be adapted to has changed. So, by default, big organizations have the upper hand

when it comes to formulating a more impactful social media strategy [1].

New-gen companies leverage emerging tech such as Instagram and WhatsApp to sell and generate more revenue. Currently, the fashion industry is run on a year's old trends, and this is changing fast since social influencers are impacting near-time purchase choices via these mediums [2].

The online domain has emerged strongly in the past few years due to the digital revolution that is shaping up in India. India will be the world's most tech-savvy e-commerce market with exponential growth due to the rapid growth of internet users in the country [3].

B. Subject study-related research

This paper helps in understanding the practical application of the ARIMA model in a bona fide business problem. Even though it is on-demand forecasting of a company in the Food domain the theoretical aspect of modeling could be directly applied to this study's objectives. It highlights how historical data can be used to forecast future implications and how it affects the downstream verticals of an organization. The model has pitched well with ARIMA (1, 0, 1) and it is validated on historical demand data under similar situations. The results achieved agrees that this model can be used to predict future demand in the food industry under the same conditions. The results obtained prove that the model could be utilized to forecast the future demand in this food manufacturing[4]

The fashion industry in India is a heavily penetrated competitive market. Due to this reason, it is important to ensure accurate sales forecasting is done. It will also set the expectation for to supply chain department to get the products. The forecast will have to consider certain aspects that are unique to the fashion industry. There is a variety of forecasting methods out there to take care of these needs of ever challenging fashion industry. Computer-based predictive analytics is one of them. Various forecasting modelling techniques are evaluated and their application to the fashion industry is thoroughly examined. Even though there is a visible benefit of using predicting analytics models in sales forecasting in the fashion industry it is not widely accepted due to the inbuilt nature of the business. This study gives a good understanding of the fashion domain and provides vital insights and the future bottlenecks of predictive analytics in the fashion industry [5].

Sales forecasting is one of the most important predictive analytics tools utilized around the globe. The common approach to forecasting is to learn from historic data and predict the future. The assumption is if certain patterns are inbuilt into the data for a long time, they will be appropriate for the future as well. Since it is a generic approach, it can be easily applied to weather prediction, Sales forecasting, etc. Sales prediction will be influenced by quantity sold, inventory, cost of the goods, and the time considered for prediction. This study has predicted the sales quantity for ten years of time series data. The data belongs to Mahindra Tractors Company. The output of the ARIMA model predicts the sales quantity for the next five years [6].

III. METHODOLOGY

A. Business Understanding

This study is based on a data set that comes from the Indian fashion retail domain. Below are a few facts from the Indian fashion retail industry which is acquired by a fashion eCommerce report [7]. The projected revenue in the fashion segment is US\$19.69 billion in 2022. The expected CAGR is (2022-2025) 18.92%. The projected market volume is US\$33.11 billion by 2025. The expected number of users in the fashion segment is expected to be 446.2 million users by 2025. User penetration will be at 22.8% in 2022 and hit 30.9% by 2025. The average revenue per user (ARPU) will amount to US\$61.46.

The Online sales market share of fashion includes D2C sales of apparel (menswear, womenswear, and Kidswear), footwear, luggage, and bags, as well as accessories (hats and caps, watches, and jewellery) by a medium that is online. The mode of sales in this market share includes e-commerce retailers such as Myntra, AJIO, amazon, etc.

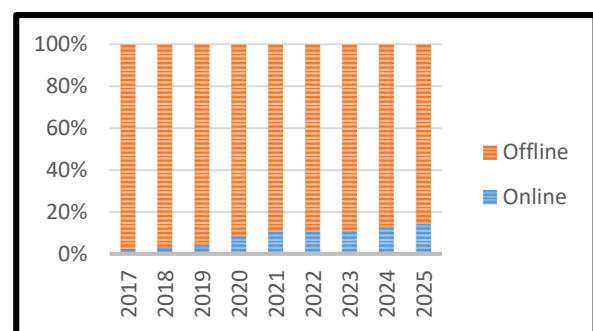


Fig. 1 Sales Market Share by Chanel [7].

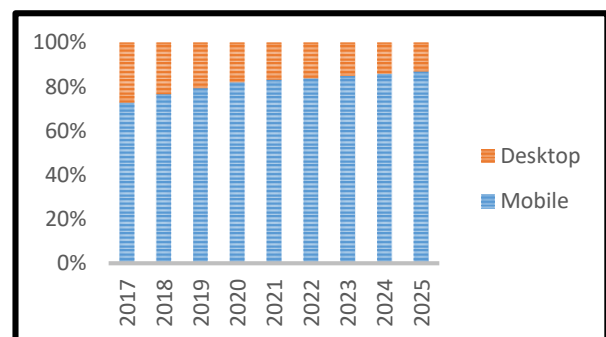


Fig. 2 Desktop vs. Mobile Market share [7].

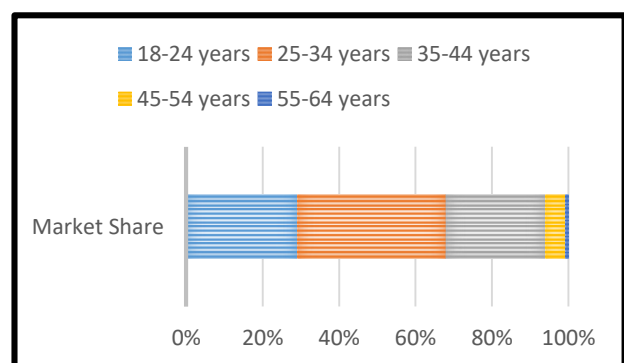


Fig. 3 Market share by age group [7].

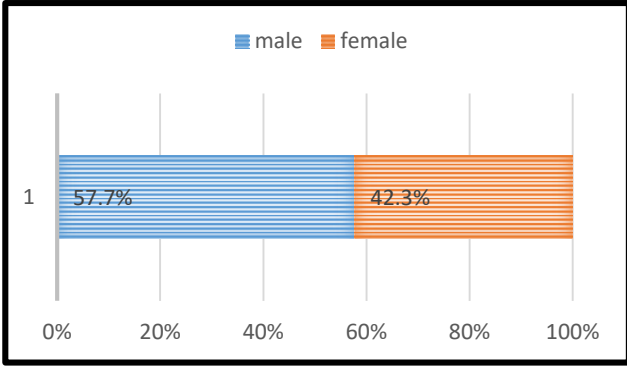


Fig. 4 Market Share by Gender [7].

The Indian fashion retail domain is expected to achieve great heights by 2025. As per Fig. 1, the Online market share is exponentially growing and continues to grow. Sales via mobile devices are increasing hence the desktop share will keep reducing and this is depicted in Fig. 2. This is a promising figure for the online sales channel.

Sixty-eight percent of the Indian shoppers are below the age of 34 years as per Fig. 3 and most shoppers are male which is 57.7% and this is shown in Fig. 4. This generation is tech-savvy and is most of the workforce hence the focus should be given to these age groups and the male segment.

Post understanding the industry and the domain it is also important to understand the organization. The data belongs to an international fashion retailer operating their business in India since 2021. The Brand operates via offline stores which are in south India and there is a rigorous expansion plan of opening 10+ stores in the South by end of 2022. The Brand's online presence is significant, and it contributes to 32% of the current business. The Brand is doing business with several well-known online aggregators such as Amazon, Ajio, and Myntra via the "marketplace" model.

B. Data Understanding

The objective of this step is to understand the attributes of the data and summarize and derive the essence of the data by identifying key characteristics, such as the volume of data and the total number of variables/attributes in the data. Understanding if there are any issues with the data, such as missing values, inaccuracies, and outliers is vital at this stage.

1). Data Source

The dataset is acquired from the organization's Order Management System (OMS). The OMS is capturing the B2C sales. There are 110 columns in the original data set. The data set is thoroughly examined to identify the important attributes and omit the rest. Also, the timeline considered for this study is six months (February 2022 – July 2022). New attributes are also derived from the existing data set to ensure more meaningful insights are captured throughout this study. The final data set has twenty-two columns with all missing values imputed.

Table I depicts a quick snapshot of the collected data. The total no. of records is 76,008. The total missing cells are 1,749. This dataset is a healthy dataset to consider since the missing value percentage is minuscule at 0.1%.

TABLE I DATASET STATISTICS

Dataset Statistics	
Number of variables	22
Sample size	76008
Missing cells	1749
Missing cells (%)	0.10%
Duplicate rows	80
Duplicate rows (%)	0.10%
Numeric Variables	9
Categorical Variables	13

The data source will be further pre-processed in the "Data Preparation" phase. This study identified two main missing elements in the data that could be easily captured. Geographical data and customer contact details. This request has been passed on to the Information Technology (IT) department of the organization and capturing customer and geographical data has been implemented since August 2022.

2). Data Exploration

The insights derived from this will shape the analytical culture of the organization. The aim is to derive actionable and practical insights. The dataset belongs to an organization where processes are new and being framed as you read this report. Hence it is critical to understand each element of the data and where it gets generated to provide meaning to the research carried forward by this study. Python's pandas package is extensively used in this section. Before getting into deep exploration pandas profiling report has been generated to identify important variables and their correlation with other variables. A few important aspects captured in the pandas profiling report are depicted in Table II.

TABLE II CORRELATION OVERVIEW OF IMPORTANT VARIABLES

No.	Variable	Highly Correlated Field
1.	Disc %	COGS
2.	Gross Sales	Discount %
3.	Net Profit	Gross Sales
4.	Category	Discount %

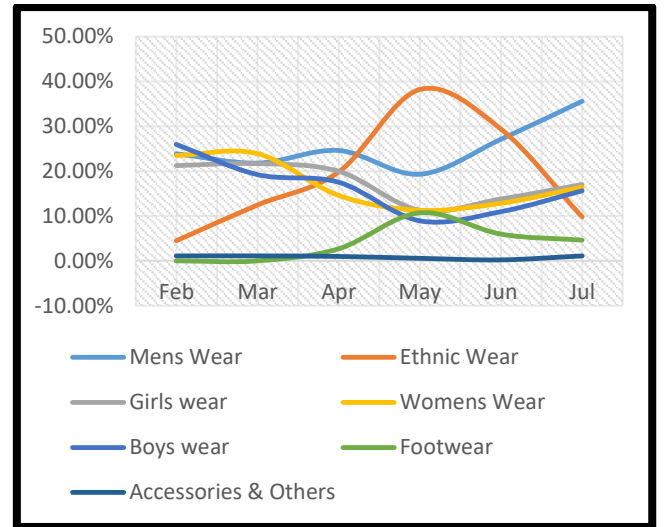


Fig. 5 Category-wise Gross Sales (Share)

Correlation depicts the relationship between two variables. As per the pandas profiling report, the four variables mentioned in Table II have a high correlation with the mentioned variable. This is an important insight to carry out further analysis and deriving insights.

Fig. 5 illustrates the sales share owned by each category of the organization. A few insights derived from Fig. 5 are,

- Men's wear category is the most stable and the biggest contributor to sales.
- Women's wear share has drastically dropped since April 2022.
- Kids' wear in both Girl's and Boy's categories have performed well in the first three months and there is a dip in these categories since April 2022.
- Ethnic wear had a slow start in the begging and contributed to the top line of the business in May and June 2022. This volatility is due to other factors and needs more exploration.
- non-Apparel categories such as footwear, accessories, and others are contributing not more than 6%.

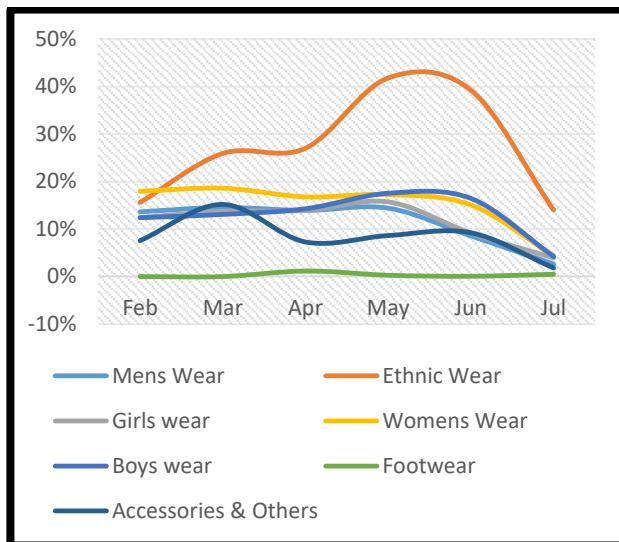


Fig. 6 Category-wise Average discount percentage

As per Fig. 6, it is evident that the more discounts you run on the online platforms more top line it will deliver. Also, Men's wear is achieving a steady share of business with minimum discounts run online, even though Ethnic wear highlighted promising top-line numbers in Fig. 5, it is evident in Fig. 6 that it is by running more discounts. All other categories have maintained a healthy discount percentage.

Another important aspect of profitability is the Cost of Goods Sold (COGS). The lesser the COGS better it is for the sales teams. This gives them more room to adapt offers, Promotions, etc. The simple logic of COGS is explained in equation (1).

$$Profit = Sales - Discounts - COGS \quad (1)$$

As per the main subject of this study, it is especially important to understand how Net profit is distributed across six months. Fig. 8 depicts the same.

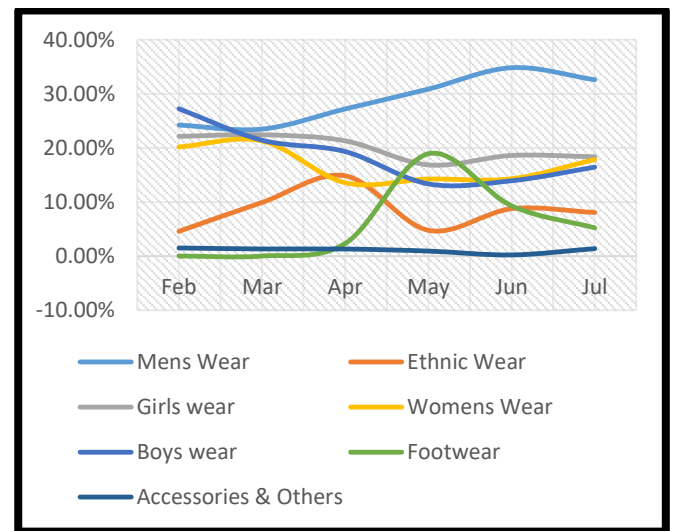


Fig. 8 Category-wise Net Profit

As per Fig. 8, it is evident that even though online is a discount-led channel it is important to keep an eye on the bottom line to ensure profitability is kept intact. As per Fig. 8, this study has found the below insights.

- Men's wear is achieving the highest share of business with low discounts and keeping the bottom line healthy.
- The ethnic wear category is the least profitable category now. The main reason for the net profit of ethnic wear to drop drastically is the high discounts run in May and June.
- All other categories have maintained a healthy Net profit margin.

Post analysing Gross Sales in Fig. 5, Average discount percentage in Fig. 6, and Net profit in Fig. 8, this study has derived some very insightful information that is affecting profitability. Apart from gross sales, Average discount percentage, and Net profit, some categorical variables directly influence Sales.

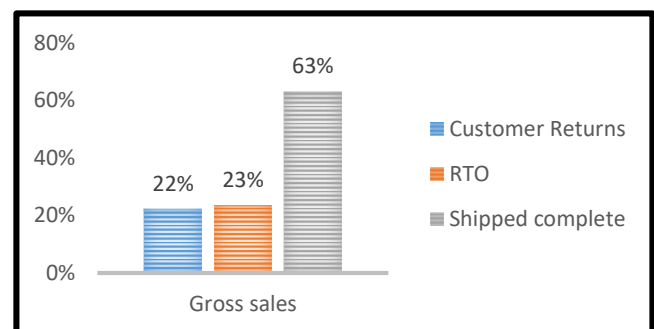


Fig. 9 Online Gross sales as per Order Status

Fig. 9 is capturing the actual sales that finally impacted the top line. Return To Origin (RTO) is Orders that got shipped out of the Organisation's warehouse but returned without reaching the customer and orders that are cancelled before it is shipped out from the warehouse. Customer returns are genuine returns where the customer has returned it post receiving the order. The numbers in Fig. 10 are alarming.

Forty-five percent of confirmed orders have become returns during the past three months. A thorough root cause analysis needs to be conducted to reduce this number as per the industry standard of <25%.

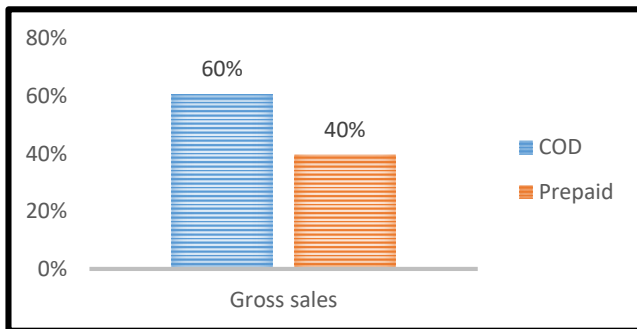


Fig. 10 Online Gross sales as per the mode of payment.

As per Fig. 10, it is evident that Cash of Delivery (COD) is the most preferred mode of payment by the customer. To understand this point further, domain experts are consulted, and it is told that it is the most preferred payment mode due to a lack of trust in order delivery and the ease of payment to the delivery partner via Unified Payments Interface (UPI) payments.

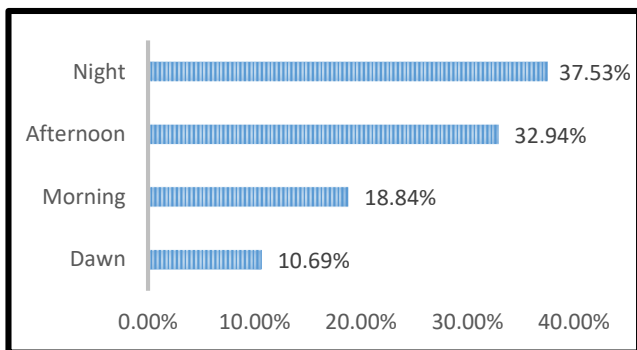


Fig. 11 Online Order flow as per the time of the day.

Fig. 11 depicts the order flow per the four timelines in a day: Dawn, Morning, Afternoon, and Night. Seventy percent of the orders flow in from noon to midnight. Also, 10% of the orders come in between twelve midnight to six in the morning which is a considerable number of orders.

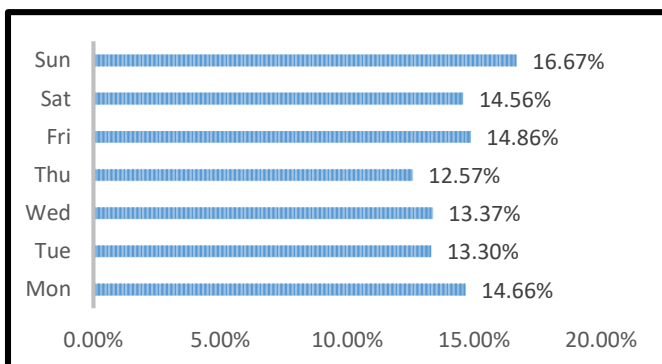


Fig. 12 Online Order flow as per the days of the week.

Fig. 12 illustrates the order flow that comes in each day of the week. The highest order flow is on Sundays followed by Fridays, Mondays, and Saturdays. This is evidence that most of the shopping happens on and around the weekend. Thursdays are the lowest order flow of the week followed by Tuesdays and Wednesdays.

As mentioned at the beginning of this study “Data Exploration” is given a lot of weightage in this project to ensure correct and directional insights are fetched via the dataset. All these insights are passed on to the e-commerce team to carry forward course correction immediately to ensure an optimum level of profitability is achieved.

C. Data Preparation

Data preparation is vital in sales forecasting, and it addresses the inconsistencies in data. Before starting pre-processing the data, the source must be identified and reviewed. Then the attributes are labelled as numeric, categorical and character, etc. to understand how to proceed with data preparation. The objective of this phase is to eliminate outliers, and discrepancies while amending the data to derive insights.

This study has directly gathered the data from the OMS backend. Even though it is a directly downloaded file there are a lot of calculated variables that are newly constructed for this operation. The date for the time series dataset is captured via the date and time stamp that gets recorded in the OMS system. The target variable for this study is “Gross Sales” which is “Maximum Retail Price (MRP) – Product discounts offered.” Please note that Gross sales are considered which is inclusive of Goods and Service Tax (GST).

The below inconsistencies in data have been dealt with before fitting the ARIMA model.

1. Dataset has 80 (0.1%) duplicate rows. The Duplicate rows have been removed from the data set.
2. No. of missing data attributes to 1,749 in the data set. The missing values are COGS. This information is available with the business in a separate file and the data has been mapped to all the 1,749 missing COGS values.

All Numerical and Categorical variables are utilized extensively in EDA in the “Data Understanding” section. For time series sales forecasting, only two variables are considered out of twenty-two variables.

1). Converting Month to Datetime

As the first step, the month is converted into “Datetime” format with the help of python’s Datetime package. This module helps in allowing variables to work with date and time. Datetime is treated as an object in Python.

2). Visualize the Data

After prepping the dataset as per the time series equivalent procedure, it is always good to visualize the data. Fig. 13 depicts the first instance of how the “Gross sales” are spread across six months before converting to Datetime format.

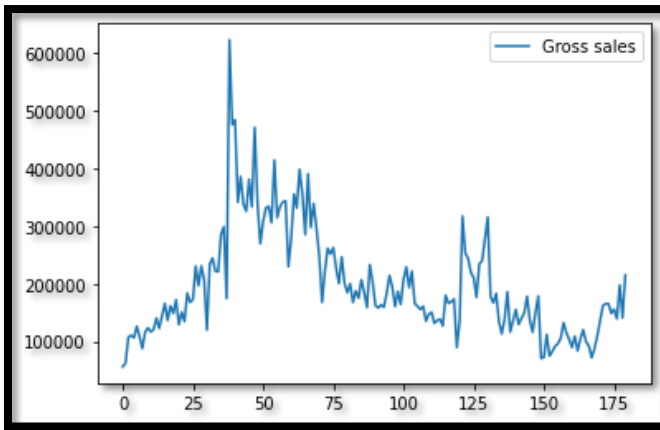


Fig. 13 Gross sales before converting to Datetime

Fig. 14 depicts Gross sales post-application of Datetime conversion. This transformation makes the data time series equivalent.

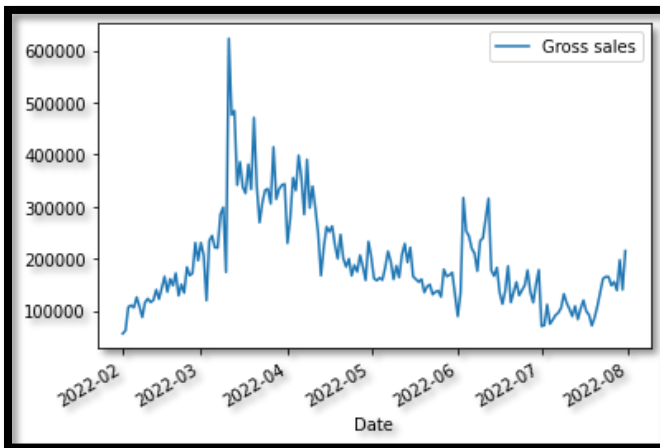


Fig. 14 Gross sales post converting to Datetime format

3). Testing For Stationarity and Differencing

If the data is stationary, it means that the statistical properties in the time series data do not change with time. The data need to be stationary since theoretical statistical models and tests depend on the data being stationary. Hence it is vital to evaluate if the data is stationary or non-stationary. “The Dickey-Fuller Test” have been conducted to check if the data is stationary or non-stationary.

The Dickey-Fuller test is the original statistical test formed to test the null hypothesis that a unit root is present in an autoregressive model of a given time series and that the process is thus not stationary. The first test treats the case of a simple lag-1 autoregression model. The “Dickey Fuller test” on python is conducted with the below hypothesis built to conduct Dicky-fuller test.

1. Null Hypothesis (Ho): It is non-stationary
2. Alternative Hypothesis (H1): It is stationary

```
ADF Test Statistic : -1.9091341775456034
p-value : 0.3277763816273437
#Lags Used : 12
Number of Observations Used : 167
```

Fig. 15 Hypothesis testing for Stationarity

As per Fig. 15 results, P-value is 32% which is greater than 5% hence the data is non-stationary. Statistical models operate with the assumption that the data is stationary hence Differencing is adapted to ensure data becomes stationary. “Sales First difference” is adapted and the “Dickey Fuller” test is re-conducted.

```
ADF Test Statistic : -3.4802686418808784
p-value : 0.00850853495400031
#Lags Used : 6
Number of Observations Used : 167
```

Fig. 16 Hypothesis testing for Stationarity post-Differencing

As per Fig. 16 results, P-value is 0.8% which is lesser than 5% hence the data is stationary. The sales data attracts a lot of noise and volatility. There can be a lot of reasons for the data to be non-stationary such as inventory, market conditions, seasonality, etc. In the Indian context, the buying patterns will differ even per the festivals of India. Due to the nature of the data and the domain, this study must make the data stationary before carrying out the ARIMA model.

D. Modeling

Post getting the data set prepped for conducting the modelling phase this study has fitted ARIMA and SARIMA models. ARIMA is a naïve model used across various business domains. The main reason to select the ARIMA sales forecasting technique is the easy understandability by the business experts. As explained at the beginning of the study, The organization is new to the analytical journey and operations in India. Hence to keep the objective simple yet powerful ARIMA forecasting techniques are utilized. This study has utilized the “stat models” package of python extensively in this chapter.

1). ARIMA

ARIMA is a group of models that predicts the target variable by utilizing its historical records. Auto-Regressive means that it utilizes “lag values” to forecast. And the Moving Average (MA) component utilizes “lagged forecast errors” for its prediction. Integrated (I) combines both “AR” & “MA” components together [8]. Due to the nature of the time series data where sales data captured contains seasonality SARIMA model is utilized for predictions.

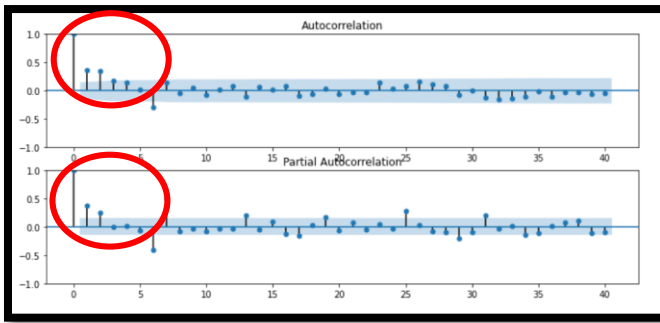


Fig. 17 ACF and PACF plots

Fig. 17 depicts the Auto Correlation Function (ACF) and the Partial Auto Correlation Function (PACF) plots post-fitting the ARIMA model. The ACF and PACF plots are utilized to determine the order of AR, MA, and ARMA models. The order of autoregression in the PACF plot is two. This means the lags do not cross the threshold post the second lag. This means the model's "AR" component has fitted well. And the order of autoregression in ACF is two as well. This translates that the model's "MA" component is fitting well too. In the "Model Evaluation" phase an upgraded version of ARIMA is fitted and evaluated to make the results more adaptable.

2). SARIMAX

SARIMAX is an uplifted version of the ARIMA model. Since the data has a seasonal aspect to it SARIMAX would be the best approach to predicting sales. Even external noise can be dialed down by this approach. Hence, the SARIMAX forecasting technique is used to forecast sales in this study [9].

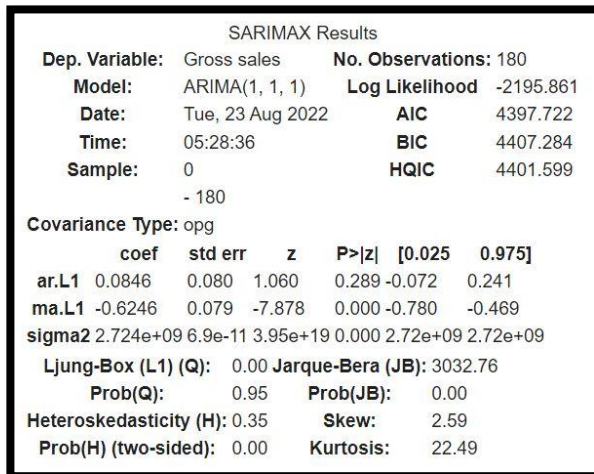


Fig. 18 SARIMAX results

As per Fig. 18, AR lag one is non-significant since the p-value is greater than 5%. However, MA lag one is significant since the p-value is lesser than 5%. As per the "Ljung-Box" Statistical approach, this study can achieve a probability score of 0.95. Hence, it translates that the model has fitted well.

3). Forecasting the sales

As per Fig. 19, this study concludes that SARIMAX is a good forecasting technique to be adapted in the fashion industry. Also, it is a gateway forecasting technique that non-technical business owners and leaders can also adapt easily.

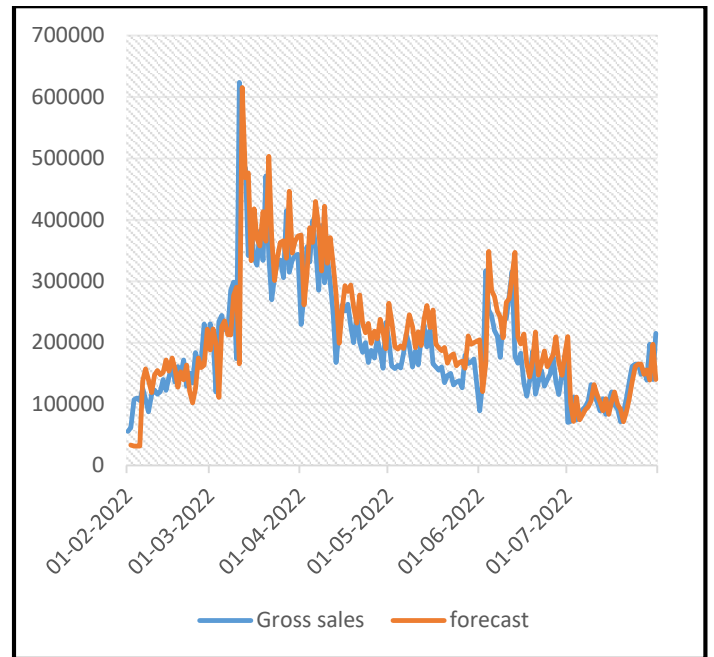


Fig. 19 SARIMAX (Actual vs. Predicted Sales).

IV. CONCLUSION

As mentioned at the beginning of this paper the organization the data has derived from is new in adapting to an analytical culture. Their forte lies in operations such as "Product Sourcing," "Sales and marketing," etc. Hence, this study's main objective is to derive meaningful and directional insights the business can seamlessly apply. At the beginning of the "Data Understanding" phase, it is identified that the customer information and where the order is generated are not captured.

This study concludes that SARIMAX is the best sales forecasting technique to adopt since it smoothens out most of the volatility in the data and most of all takes care of the seasonality aspect. This Study's main approach is to convince the organization's management of the advantages of adopting an analytical culture in the organization. The best way is to provide meaningful insights which this study can achieve to a larger extent.

V. FUTURE SCOPE

Since the business is new to its operations, The dataset captured for this study is limited to six months of data. Hence, In the future scope of this study, it is recommended to conduct this activity every month by adding new data so the model and learn better.

Once the data set is more mature and the business is also admirably adapting to the recommendations and start seeing the positive impact the model is bringing in forecasting sales more forecasting techniques such as Long Short-Term Memory (LSTM), Linear Regression, and Random Forest can be explored.

ACKNOWLEDGMENT

We would like to convey heartfelt gratitude to all the mentors at Reva Academy of Corporate Excellence, Mithun Dolthody Jayaprakash, and Ratnakar Pandey for their continuous support throughout the learning journey. We

would like to express a special thanks to Dr. Shinu Abhi, Director of REVA Academy of Corporate Excellence for her amiable support, helpful guidance, and information at various phases of this study, which helped in completing it.

REFERENCES

- [1] McKinsey & Company, "The State of Fashion 2022," 2022. Accessed: Aug. 15, 2022. [Online]. Available: <https://www.mckinsey.com/~media/mckinsey/industries/retail/our%20insights/state%20of%20fashion/2022/the-state-of-fashion-2022.pdf>
- [2] Deloitte Digital, "Apparel Trends 2025," 2022. Accessed: Aug. 15, 2022. [Online]. Available: <https://www.deloittedigital.com/content/dam/deloittedigital/us/documents/blog/blog-20200610-apparel-trends.pdf>
- [3] Wazir Advisors, "Wazir Report - The Road to 2025," pp. 1–32, 2022, Accessed: Aug. 15, 2022. [Online]. Available: <https://wazir.in/pdf/Wazir%20Report%20-%20The%20Road%20to%202025.pdf>
- [4] J. Fattah, L. Ezzine, Z. Aman, H. el Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, vol. 10, Oct. 2018, doi: 10.1177/1847979018808673.
- [5] J. E. P. Bug, "Application of predictive analytics to sales forecasting in fashion business," 2016. [Online]. Available: <https://www.researchgate.net/publication/325100494>
- [6] S. P. Shakti, M. K. Hassan, Y. Zhenning, R. D. Caytiles, and I. N.Ch.S.N, "Annual Automobile Sales Prediction Using ARIMA Model," *International Journal of Hybrid Information Technology*, vol. 10, no. 6, pp. 13–22, Jun. 2017, doi: 10.14257/ijhit.2017.10.6.02.
- [7] Statista Digital Market Outlook, "Fashion eCommerce report 2021," Jul. 2021. Accessed: Aug. 18, 2022. [Online]. Available: <https://www.statista.com/study/38340/ecommerce-report-fashion/>
- [8] Tony Yiu, "Understanding ARIMA (Time Series Modeling)," Apr. 26, 2020. <https://towardsdatascience.com/understanding-arima-time-series-modeling-d99cd11be3f8> (accessed Aug. 23, 2022).
- [9] YUGESH VERMA, "Complete Guide To SARIMAX in Python for Time Series Modeling," Jul. 30, 2021. <https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/> (accessed Aug. 23, 2022).