e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

SALES FORECASTING FOR ONLINE & RETAIL SELLERS

Vinay Kamble [*1], Durvankur Kadam[*2], Laxmi Kattimani[*3], Praveen Patel[*4]

[*1,2,3,4]Mumbai University, IT, Bharati Vidyapeeth College of Engineering,
Navi Mumbai, Maharashtra, India.

ABSTRACT

In this era of digitization, the business owners or online sellers must understand and able to analyse accounting information and the trades information for the sustainability of their business. However, many of them are still unable to fully utilize accounting information. The trade data can be fully understood and utilize with the assistance of machine learning. This paper proposes a Software system which uses machine learning models & concepts which will help the retailers to utilize their trade data and to monitor their business growth or decline. The system consists of ARIMA & SARIMA models which uses time series data to either better understand the data set or to predict future sales. It is a statistical analysis model which will help the user to monitor the health of his/her business.

Keywords: Sales, Purchases, Time Series Analysis, Online Sales, Pandas, Sklearn, Trade, Regression, Forecasting, Classification

I. INTRODUCTION

Due to the increase in online marketing & businesses, a large amount of trades data has been generated and this data is being used for analysis of individual businesses and the conclusions derived from this data helps in the growth of the businesses. This paper proposes a Software system which uses machine learning models & concepts which will help the retailers to utilize their trade data and to monitor their business growth or decline. The system consists of ARIMA & SARIMA models which uses **time series data** to either better understand the data set or to predict future sales.

A time series is a sequence of data points that occur in successive order over some period. Forecasting sales is one of the essential uses of Machine Learning. It helps to identify benchmarks and determine incremental impacts of new initiatives, plan resources in response to expected demand, and project future budgets.

This forecasting technique comes under supervised learning method. In this technique, they generate predictions by finding trends and seasonality patterns using the past sales data.

II. METHODOLOGY

Method For this system, we are making of use of time series analysis of the data for sales forecasting, time series is a sequence of observations recorded over a certain period of time. A simple example of time series is how we come across different temperature changes day by day or in a month. There are two main types of sales forecasting:

1. Rule-based forecasting
2. Machine Learning forecasting

In this we are using Machine Learning Models for sales forecasting, we are making use of the AutoRegressive integrated moving average (ARIMA) or Seasonal AutoRegressive integrated moving average (SARIMA) models for predicting the future sales. There are various regression models available which will help in predicting the future sales but as the data of every other online seller or retailer will lack the necessary features required for training the model, so as per the most user's data we have selected the ARIMA model as it only required the past sales of the retailer. The development of this model is divided into four parts:

A. Data Pre-processing

In this we have used real time data of an online retailer, the data mainly consists of the trade details (i.e. invoice date, customer name, item name, supply state, quantity etc.). This data should be processed & analysed before feeding it to our model. For that Firstly we have removed the unnecessary columns from the dataset and only considered the date and total sale value column of the dataset. We have also cleaned the data by removing rows having missing values and converted the date column to Datetime type for further time series analysis.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

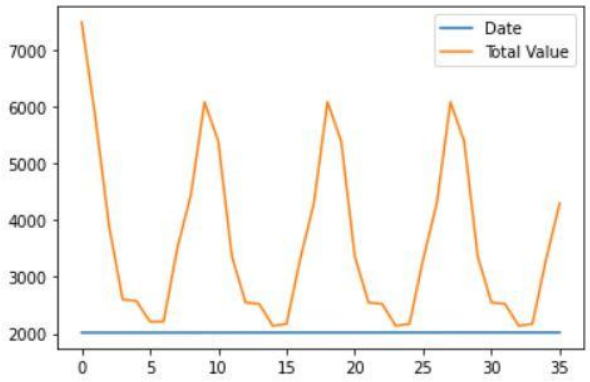
Secondly, we have grouped the sum of the sales of a whole month for every month of every year, as we are going to predict the monthly sales.

B. Data Visualization

In order to understand our data, to understand trends, outliers, and patterns in data, we have visualized our data in various charts and graphs. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. We are using matplotlib & seaborn to do the visualization.

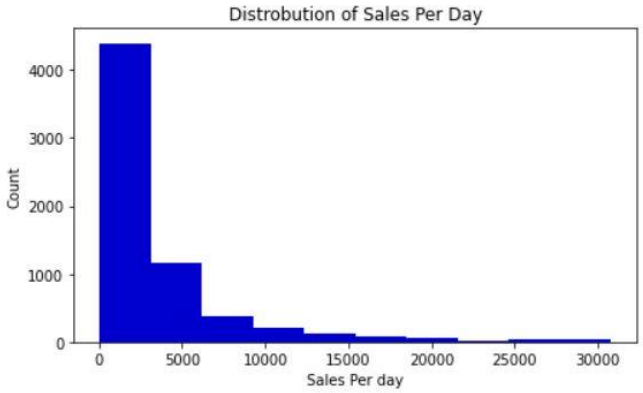
a). Normal Graph (Total Sales Vs Months)

In this graph we get any idea of the sales distribution



b). Distribution of Sales Per Day

this chart gives an idea about the daily sales of the business, by getting the count of the sales on a day for further analysis.



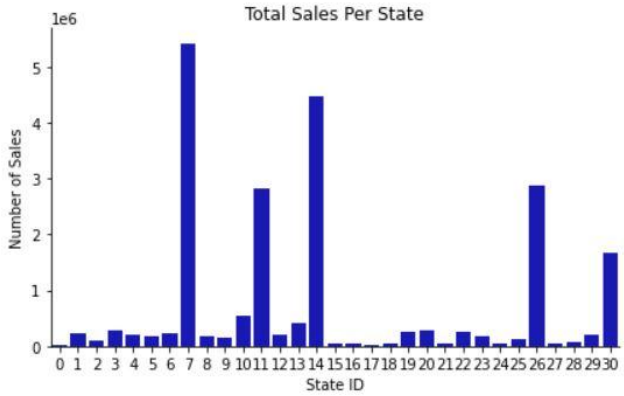
c). Distribution of Sales as per States

this bar chart gives us the insight regarding the sales done in various states, which gives us the idea of potential customer base in various states and manage the demand as per the knowledge obtained.

Categorical Feature to Numerical Features – For States					
0	Maharashtra	11	Madhya Pradesh	21	Odisha
1	Telangana	12	Delhi	22	Puducherry
2	Haryana	13	Goa	23	Punjab
3	West Bengal	14	Himachal Pradesh	24	Rajasthan
4	Others	15	Jammu & Kashmir	25	Sikkim
5	Karnataka	16	Gujarat	26	Uttar Pradesh
6	Andaman & Nicobar Islands	17	Kerala	27	Uttarakhand

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

7	Andhra Pradesh	18	Manipur	28	Arunachal Pradesh
8	Assam	19	Meghalaya	29	Tripura
10	Jharkhand	20	Nagaland	30	Tamil Nadu
31	Mizoram				



III. MODELING AND ANALYSIS

For the forecasting of the sales or predicting the future sales we are using the Auto Regressive Moving Averages Model (i.e. ARIMA). It is a type of time series model, which is based on the stationarity of the data. In the most intuitive sense, stationarity means that **the statistical properties of a process generating a time series do not change over time**.

Firstly, we are checking for the stationarity of our data, for that we are using the **Augmented Dickey fuller** test. ADF test belongs to a category of tests called 'Unit Root Test', which is the proper method for testing the stationarity of a time series. The presence of a unit root means the time series is non-stationary.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p} + e_t$$

where

- y(t-1) = lag 1 of time series
- delta Y(t-1) = first difference of the series at time (t-1)

A key point to remember here during the dickey fuller test : Since the null hypothesis assumes the presence of unit root, that is α=1, the p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis. Thereby, inferring that the series is stationary.

We are using the "**statsmodel**" package of python which contains "**adfuller"** function for implementing this test on our data.

```
adfuller_test(df['Sales'])
```

```
ADF Test Statistic : -1.8335930563276195
p-value : 0.3639157716602467
#Lags Used : 11
Number of Observations Used : 93
weak evidence against null hypothesis, time series has a unit root,
```

As our p-value is not less than 0.05 and the null hypothesis is not rejected so, our data is not stationary.

We are using the differencing method, in this we take a difference between the data points. So, let us say, your original time series was:

```
X1, X2, X3,...........Xn
```

series with difference of degree 1 becomes:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

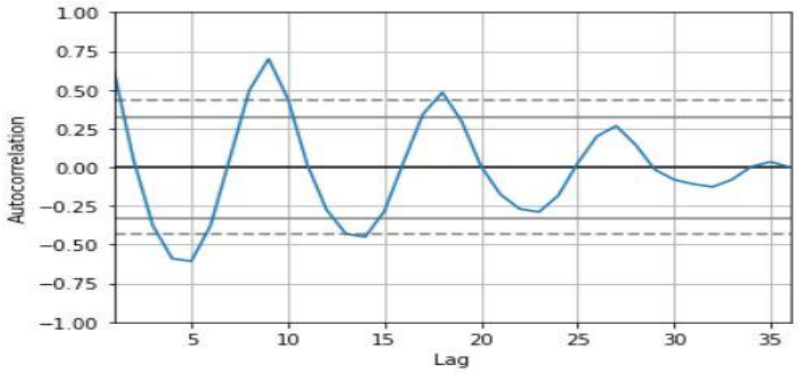
```
(X2 - X1, X3 - X2, X4 - X3,.......Xn - X(n-1)
```

We can set the degree of difference as per our data need. Looking towards the data plot we can say that our data is seasonal, so we are using degree of difference as 12.

Monthly Sales Before Differencing:

Monthly Sales After Differencing:

Consider if we want to predict today's sale, we have to know how many previous day data we have to consider. For that we plot an autocorrelation plot, as autocorrelation measures the relationship between a variable's current value and its past values. It tells us the correlation between points separated by various time lags.



Training the ARIMA Model

An ARIMA model requires p, d & q, we have to give those values to the model where,

p - AR model lags

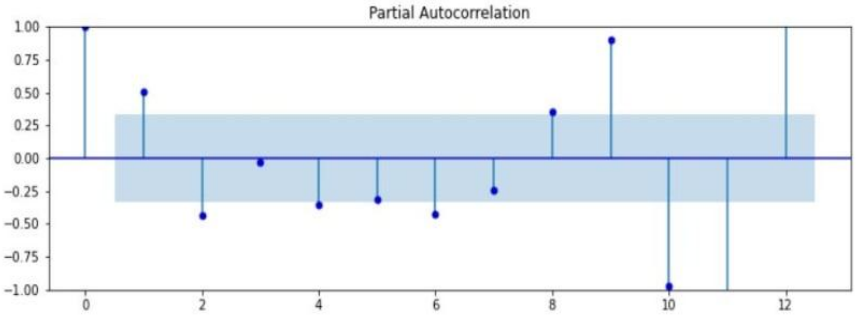
d – differencing

q – moving average lags

Identification of an Autoregressive model is often best done with the PACF.

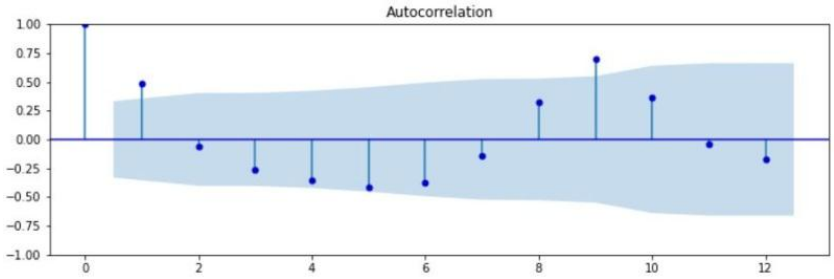
For an AR model, the theoretical PACF "shuts off" past the order of the model. The phrase "shuts off" means that in theory the partial autocorrelations are equal to 0 beyond that point. Put another way, the number of non-zero partial autocorrelations give the order of the AR . model. By the "order of the model" we mean the most extreme lag of x that is used as a predictor.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com



Identification of an MA model is often best done with the ACF rather than the PACF.

For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner. A clearer pattern for an MA model is in the ACF. The ACF will have non-zero autocorrelations only at lags involved in the model.

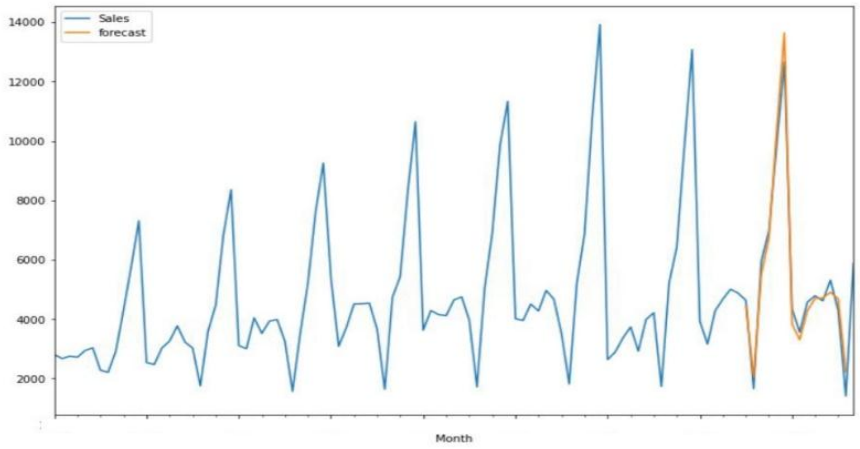


Now looking at the Autocorrelation & Partial Autocorrelation plots, we set the value of p=1, q=1 and d=1 as we have done the differencing once. As our data is seasonal, we are going to use the seasonal ARIMA model. We are using the **"statsmodel"** python package which consists of the SARIMAX model.

We are setting the values of p, q & d and an additional 4th column that represents the seasonal value and we are training our model with our sales data.

IV. RESULTS

Prediction For known data



For unknown data we create a dataset of 10 months

```
In [37]: from pandas.tseries.offsets import DateOffset
         future_dates=[df.index[-1]+ DateOffset(months=x)for x in range(0,24)]
```

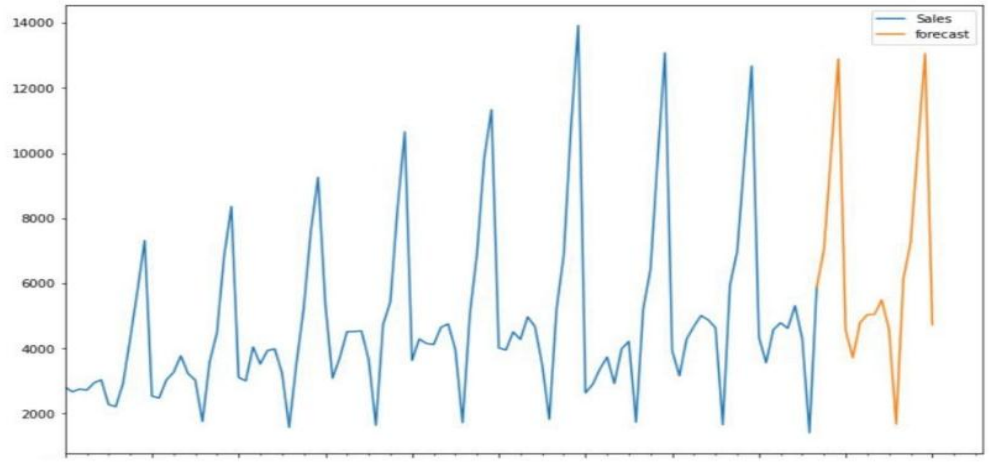
```
In [38]: future_datest_df=pd.DataFrame(index=future_dates[1:],columns=df.columns)
```

We then concatinate this data with our existing data and forecast the sales for new values

```
In [40]: future_df=pd.concat([df,future_datest_df])
```

Forecast for Unknown Values

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com



V. CONCLUSION

Due to growth of online shopping it has become essential for organizations to utilize the data to find the desired information resources, and to track and analyse the sales transaction. Patterns of the type of products that are sold frequently can be found out using data mining of sales data.

The Sales forecasting can be helpful in identifying the supply and demand of various products and helps in better management of the business.

VI. REFERENCES

[1] Z. Zhao, J. Wang, H. Sun, Y. Liu, Z. Fan and F. Xuan, "What Factors Influence Online Product Sales? Online Reviews, Review System Curation, Online Promotional Marketing and Seller Guarantees Analysis," in IEEE Access, vol. 8, pp. 3920-3931, 2020.

[2] Z. Pirani, A. Marewar, Z. Bhavnagarwala and M. Kamble, "Analysis and optimization of online sales of products," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-5.

[3] Z. Mu, "Fruit Online Chain Sales System Based on B/S," 2019 International Conference on Robots & Intelligent System (ICRIS), 2019, pp. 238-241.

[4] J. Wang, "A hybrid machine learning model for sales prediction," 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 2020, pp. 363-366.