Online Sales Forecasting using Machine Learning Techniques

M.L.S.N.S Lakshmi[1], Prasad Jones Christydass[2]

[1]*Associate professor, ECE Department QIS college of Engineering and Technology,Ongole,India*
E-mail Id:mlakshmi.0290@gmail.com

[2]*Associate professor, ECE Department QIS college of Engineering and Technology,Ongole,India*

Abstract:

As online product review forums become in popularity, more people are encouraged to share their thoughts and feelings about the products. It provides information about the product. These online reviews are crucial for predicting how well a product will sell. The ability to predict future sales is crucial when making goods purchases. The dynamic, global, and unpredictable business environment in which organisations now operate is one of the major problems they face. Manufacturers today cannot solely rely on their cost advantage over competitors due to rising customer expectations for pricing and quality. Machine Learning (ML) is an effective way for sales forecasting. Technological innovation helping to make huge changes to the organization's sales rate for securing business profitability. In this work we are implementing sales forecasting in Jupiter tool using Python. By Using the model ARIMA (Autoregressive integrated moving average) in Machine learning we can achieve better security and profitability in sales forecasting for different business applications.

Keywords: Sales forecasting, Machine Learning, Auto Regression, Time Series data, Stationary.

I. INTRODUCTION

Sales forecasting is nothing but prediction of future sales. These are based on historical data and current sales status. Mostly sales forecasting is used to estimate monthly, weekly and annual sales reports. Due to this we can plan our products to increase sales in a more effective way, we can also estimate what will happen in the future and plan accordingly as the population increases and mostly people are also willing to purchase by online shopping. Now in the present situations like virus which is spreading by crowd so people are more likely to buy their needs which are available in online. so we can say that the implementation of online sales forecasting forums invites the development of models that allows these information for decision support of the prediction.

In 2020, C. Lu, F. Wang, and G. Trajcevski from Xidian University released a paper on a crop forecast system employing a machine learning algorithm. The same year, a journal based on data mining prediction demand was also published [2,3]. In 2020, Martinez c. Schmuck conducted research on a machine learning system for predicting client purchases in non-contractual settings[4]. S. Wang, W. Zhao, Ji, and D. Guo had anticipated a module on the three stage XGBoost based model's applicability for 2019[5]. By combining sentiment analysis and machine learning, Martnez-Plumed, Contreras-Ochando, and Ferri [7,8] predicted suicides on social media in 2019. Amalina, F., and Hashem, I.A.T.[9,10] have introduced merging new technology in 2019. In the year 2018[11], Tyralis, H.; Papacharalampous created a new method for large-scale prophet assessment for multi-step predicting. Papacharalampous, G., Tyralis, and Koutsoyiannis suggested automatic time series forecasting methods to predict monthly precipitation and temperature in 2018[12]. "Sales forecast of Four Wheelers Unit (4W) with seasonal algorithm Trend Decomposition with Loess" was a presentation made by A. Telaga, A. Librianti, and U. Umairoh to the IOP conference in the field of materials and science engineer in the year 2019[14]. Punam, K., Pamula, R., and Jain had presented on the topic. a two-level statistical model for predicting big-mart sales. 2019 International Conference on Computing, Power, and Communication Technologies[15,16].

Atime series is a collection of pictures that were taken at regularly spaced-out intervals of time. It is therefore a collection of discrete-time data. You can better comprehend how a security, asset, or economic attribute changes over time by using time series analysis. The methodology used in this study is a parametric one, which presupposes that the underlying stationary stochastic process has a structure that can be modelled with a limited number of parameters, such as an autoregressive or moving average model. In order to prevent mistakes from the past and boost a company's sales by making wise decisions based on statistical data, keeping track of sales and projecting sales is essential. Predicting stock prices is one way to assess a company's value.

II. PROPOSED METHODOLOGY

Time series can be modelled using traditional statistical models like moving average, exponential smoothing, and ARIMA. These models are linear because future values must be linear functions of historical data. Due to their simplicity in understanding and implementation, linear models have received a lot of attention from academics over the past few decades. Demand is forecasted using the majority of time series forecasting techniques.

When the seasonal adjustment order is high or the diagnostics do not show that the time series is stationary after seasonal adjustment, it is frequently impossible to determine a model. In these situations, it is believed that the static parameters of the traditional ARIMA model are the main obstacle to accurately anticipating highly fluctuating seasonal demand. The need for several observations to find the best fit model for a data series is another drawback of the conventional ARIMA approach. An ARIMA model (p, d, q) is one that includes the following:

The number of autoregressive terms is given by p.

The number of variations is d.

The moving average and autoregressive process count is q. According to the autoregressivemodel,

The parameters mentioned above are linearly related to Yt. According to equation (1)

$$Yt = \alpha 1 Yt\text{ -}1 + \varepsilon t$$

A random component (random shock) plus a linear combination of earlier observations make up each observation. The self-regression coefficient, 1, is represented in this equation. the combined approach. The behaviour of the time series may be impacted by the cumulative impact of some processes. Consumption and supply, for instance, are continually adjusting stock status, but the average level of stocks is mostly influenced by the cumulative impact of the moment-to-moment changes throughout the time between inventories. Although the short-term stock prices may vary considerably from this average value, the long-term level of theseries will not change. A timeseries determined by an action's cumulative effect is one of the components of the class of integrated processes.

A timeseries that is defined by the whole impact of an activity is a member of the class of integrated processes. Even if a series exhibits chaotic behaviour, the variations between observations may still be negligibly tiny or even swing around a fixed value for a process observed over a variety of time scales. The stationarity of the series of differences for an integrated process is a crucial quality from the perspective of statistical analysis of the timeseries. Integrated processes are used to depict nonstationary series. In an order 1 differentiation, the difference between two future values of Yi is taken to be constant. Equation describes a seamless process (2)

$$Yt = Yt\text{-}1+\varepsilon t$$

Thereby we are developing a practical approach.Initially Run the ARIMA model Box-Jenkins method. Model identification is done in three iterative processes.

1. Data discovery
2. Data preparation and cleaning
3. Feature engineering

The most crucial aspect of machine learning is data. Data discovery is the process of collecting data from various sources and identifying the properties of the data[1].Data cleaning means removing null data, replacing it with mean, and making adjustments to the data.[2].Feature engineering is the process of extracting features from data using domain or field knowledge.[3]

The basic rule for identifying a model is that it is chronological Obtained from the ARIMA process.Arima can be used to find one or more potential models of given time data. Box and Jenkins recommended using them. The order of the ARIMA model is determined using the autocorrelation function (ACF) and partial autocorrelation function (PACF) of sample data as fundamental tools.

The mean and standard deviation, are statistical features of stationary time series and autocorrelationin this method we can use classic statistical methods for models in time series analysis the structure is constant over time. Usually, we apply Differentiation and power conversion to data Before removing the trend and stabilizing the variance now You can install the ARIMA model,After that, it will be easier to specify the model and determine the model's parameters. Model validation is then carried out, and if the model is insufficient, additional parameter estimations must be made. The development of new models can benefit from diagnostic information. When a high degree of satisfaction is attained, repeat the process as represented by the Box-Jenkins model to eliminate mistakeswhile using.now you are free to use this model to predict variables. Researchers can estimate parameters many observations are required. as a result, there are some restrictions on the use of ARIMA models. However, high quality of sales is achieved as soon as the ARIMA model is applied.

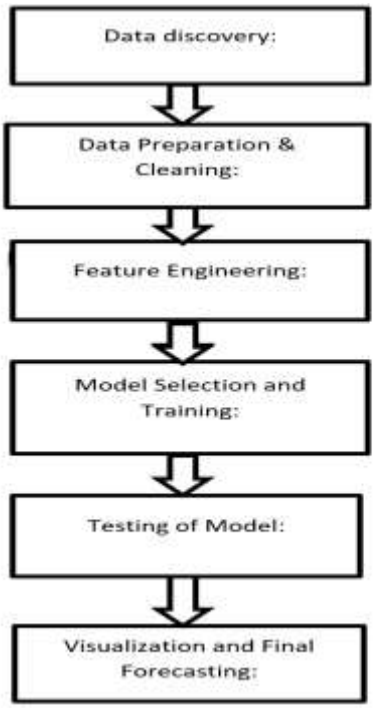

Fig: 1 Flowchart for sales forecasting

A machine learning model may be created by following a few simple steps. Feature engineering, data discovery, data preparation, training model, testing model, and visualisation. Seasonal ARIMA was used to create an inflation model that suggested seasonality in the time series.

III . RESULTS AND DISCUSSION

Step 1. Data accessing and discovery

The most crucial aspect of machine learning is data. Data discovery is the process of collecting data from many sources and identifying the features of the data.Data access is the ability to retrieve, edit, copy, or move data from IT systems on demand and with authorization.

```
import numpy as np
import pandas as pd
df=pd.read_csv(r'C:\Users\yeswanth jasti\Downloads\perrin-freres-monthly-champagne.csv')

df.head()
```

Fig: 2 Load the dataset

Fig: 2 shows load the data set into a python Jupiter notebook by using the module pandas than copy path of data set in "c-drive". Past the path of file in function "pd.read_csv()" the file is in the forms of csv(comma separated value).

Step 2. Data cleaning & visualization

Data cleaning entails eliminating null data, substituting mean for it, and making corrections to the data. Data visualisation is the process of representing information and data graphically. Using visual features like charts, graphs, and maps, data visualisation tools make it simple to spot and comprehend trends, outliers, and patterns in data.

```
# Convert Month into Datetime
df['Month']=pd.to_datetime(df['Month'])
```

```
df.tail()
```

	Month	Sales
100	1972-05-01	4618.0
101	1972-06-01	5312.0
102	1972-07-01	4298.0
103	1972-08-01	1413.0
104	1972-09-01	5877.0

Fig: 3 cleaning the data

```
df.plot()
```

```
<AxesSubplot:xlabel='Month'>
```

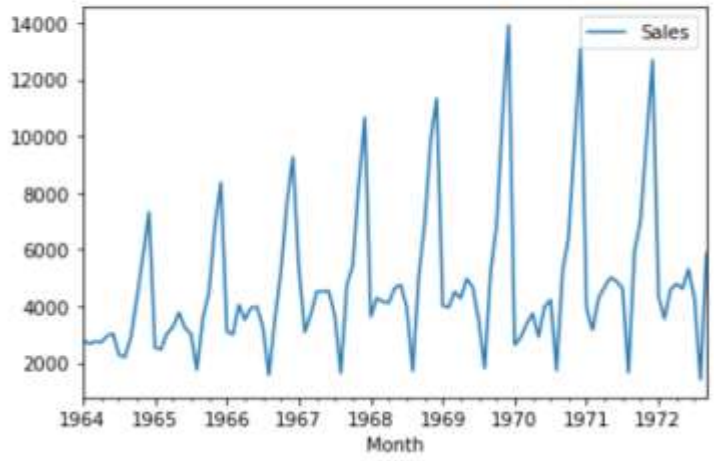

Fig: 4 data visualization

Fig: 3 shows clean the data set to make a we show time series data and data over timewhich show in Fig: 4 and the data is seasonal in character.

Step 3. Feature Engineering

The process of removing features from data is known as feature engineeringusing domain or field knowledge. Feature Engineering employs a variety of methods and tests to determine whether a dataset is suitable for model construction. For univariate time series we require stationary data, This implies that the "mean" and "standard deviation" shouldn't change over time.

We employed rolling statistics to check the stationarity of the data in the stocks dataset. For time series, rolling statistics are quite valuable. It's a type of window that performs operations on data of a specific size. We discovered that the data is nonstationary in this case.

Reduce stationarity in time series data by differentiating or transforming the data. We used the differencing strategy, which involves deducting the value of one observation from another over time, to analyse our data.

```
df['Seasonal First Difference'].plot()
```

```
<AxesSubplot:xlabel='Month'>
```

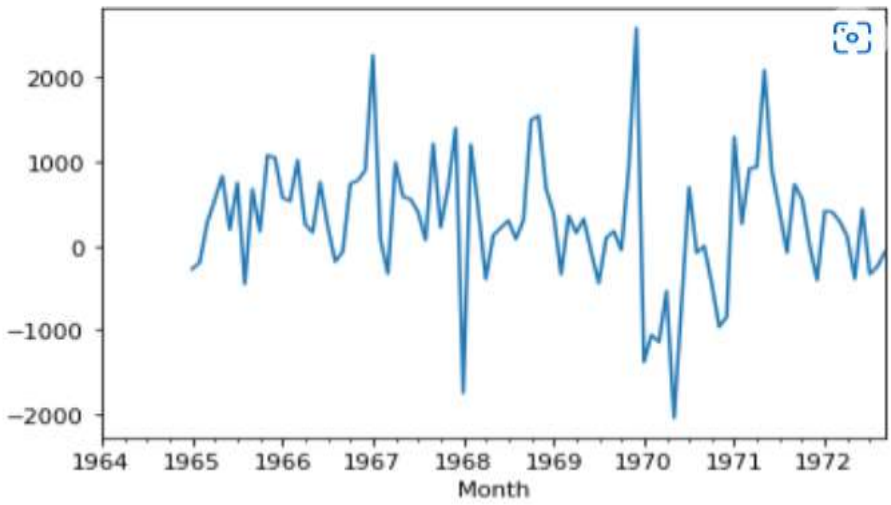

Fig: 5 visualize the stationary data

Step 4. Model Selection and Training

Since our data are both nonseasonal and seasonal in character, as shown in figure No. 4, we chose the ARIMA and SARIMA models. used the data from Figs. 6 and 7 to train the model.

ARIMA Model Results

Dep. Variable:	D.Sales	**No. Observations:**	104
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-951.128
Method:	css-mle	**S.D. of innovations**	2227.262
Date:	Tue, 14 Jun 2022	**AIC**	1910.251
Time:	16:27:28	**BIC**	1920.829
Sample:	02-01-1964	**HQIC**	1914.538
	- 09-01-1972		

	coef	std err	z	P>|z|	[0.025	0.975]
const	22.7844	12.405	1.837	0.066	-1.529	47.098
ar.L1.D.Sales	0.4343	0.089	4.866	0.000	0.259	0.609
ma.L1.D.Sales	-1.0000	0.026	-38.503	0.000	-1.051	-0.949

Fig: 6observed values of nonseasonal data

SARIMAX Results

Dep. Variable:	Sales	**No. Observations:**	105
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	**Log Likelihood**	-738.402
Date:	Tue, 14 Jun 2022	**AIC**	1486.804
Time:	16:27:29	**BIC**	1499.413
Sample:	01-01-1964	**HQIC**	1491.893
	- 09-01-1972		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2790	0.081	3.433	0.001	0.120	0.438
ma.L1	-0.9494	0.043	-22.334	0.000	-1.033	-0.866
ar.S.L12	-0.4544	0.303	-1.499	0.134	-1.049	0.140
ma.S.L12	0.2450	0.311	0.788	0.431	-0.365	0.855
sigma2	5.055e+05	6.12e+04	8.265	0.000	3.86e+05	6.25e+05

Fig: 7observed values of seasonal data

Step 5. Testing of Model

In the Fig 8 and 9, the model is tested on the final set of data in Figs. 8 and 9 to determine how accurate it is. The static sales prediction shown by the orange line indicates whether or not the model is reliable. We utilise MSE to determine the precision or line of bestfit of our model (mean squared error).

```
df['forecast']=model_fit.predict(start=90,end=103,dynamic=True)
df[['Sales','forecast']].plot(figsize=(12,8))
```

```
<AxesSubplot:xlabel='Month'>
```

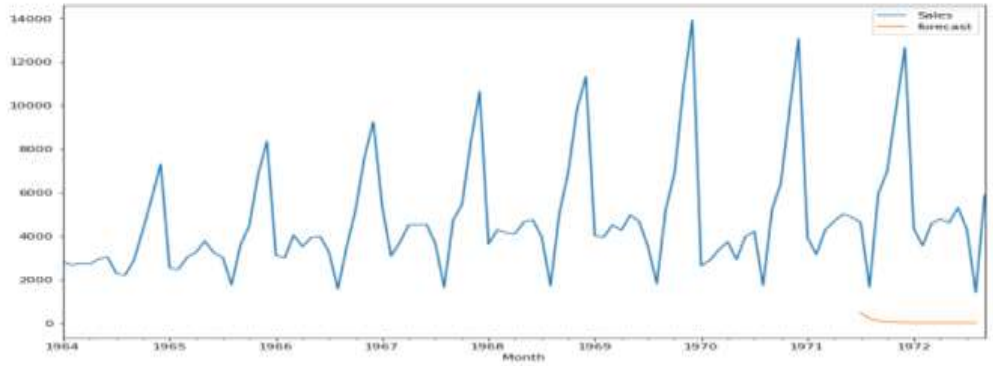

Fig: 8 It is fit for forecasting the future values of non-seasonal data

```
df['forecast']=results.predict(start=90,end=103,dynamic=True)
df[['Sales','forecast']].plot(figsize=(12,8))
```

```
<AxesSubplot:xlabel='Month'>
```

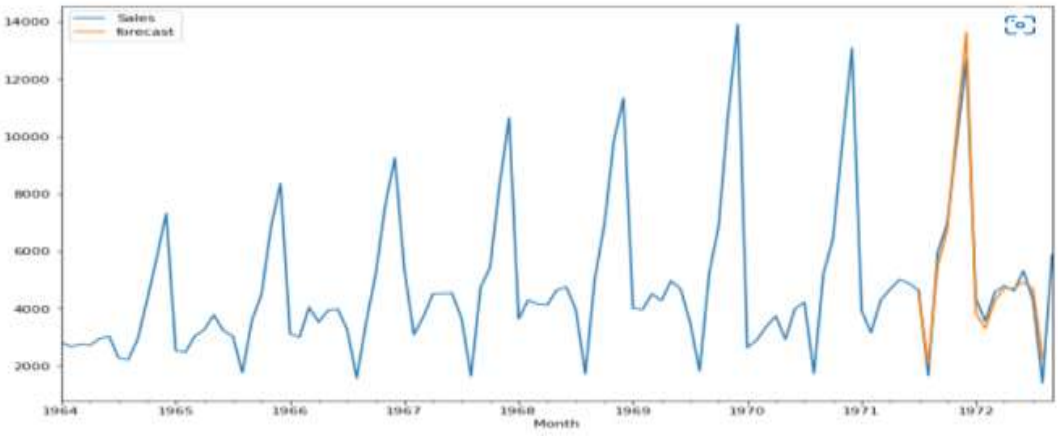

Fig: 9 It is fit for forecasting the future values of seasonal data

Step 6. Visualization and Final Forecasting

In this stage, we will anticipate the company's future sales using our tried-and-true approach.

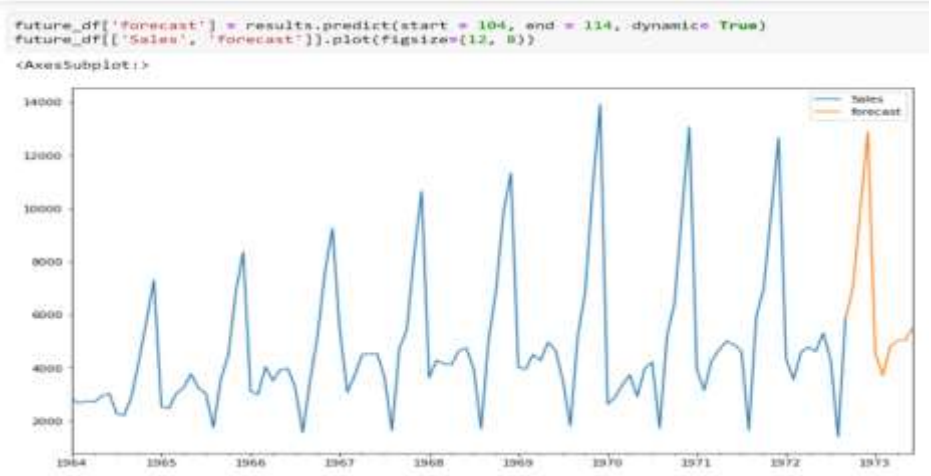

Fig: 10 prediction of 10 months sales.

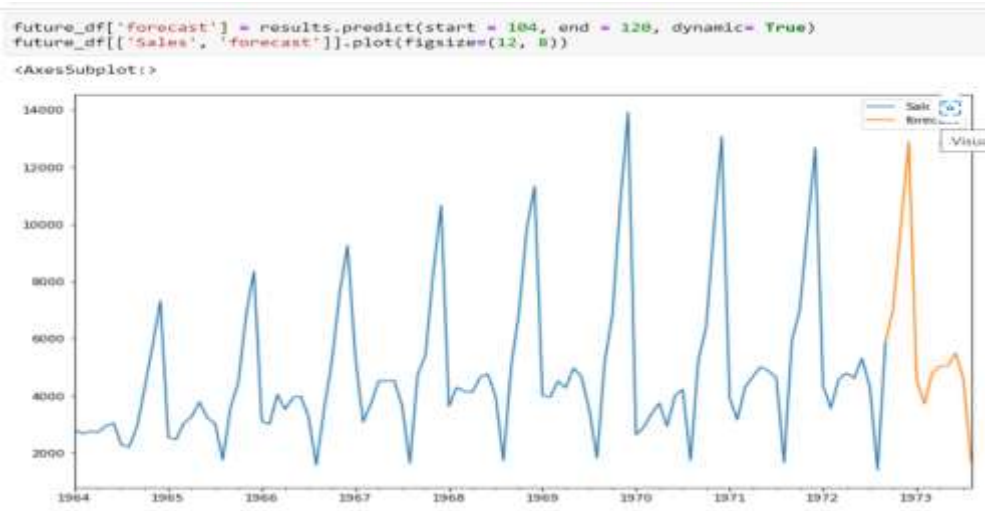

Fig: 11 prediction of 12 months sales.

```
future_df['forecast'] = results.predict(start = 104, end = 130, dynamic= True)
future_df[['Sales', 'forecast']].plot(figsize=(12, 8))
```

```
<AxesSubplot:>
```

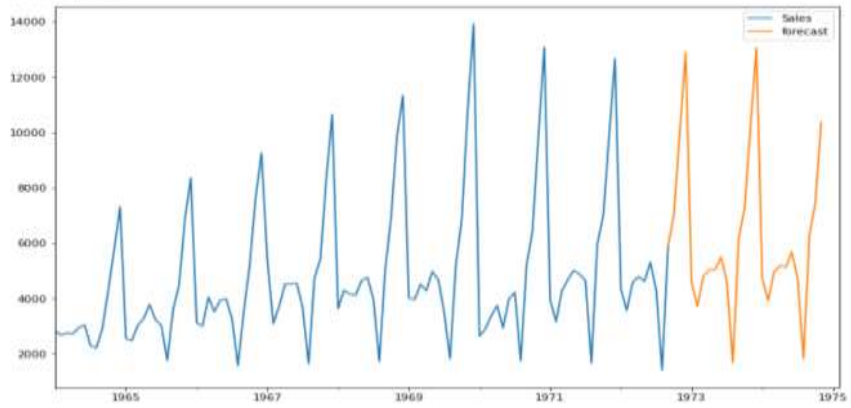

Fig: 12 prediction of 24 months sales.

```
future_df['forecast'] = results.predict(start = 104, end = 150, dynamic= True)
future_df[['Sales', 'forecast']].plot(figsize=(12, 8))
```

```
<AxesSubplot:>
```

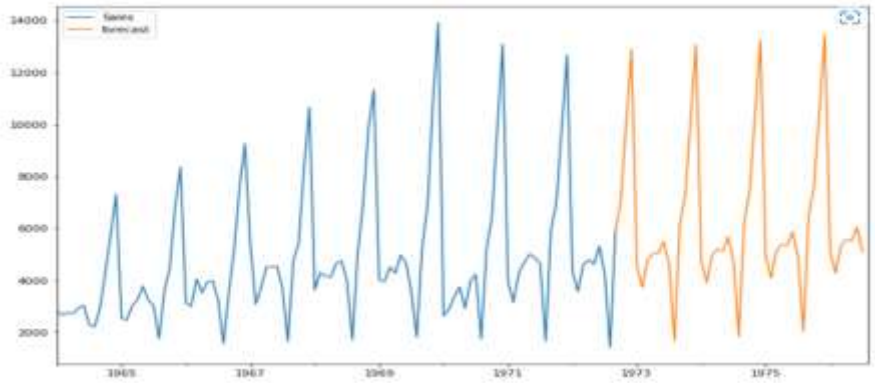

Fig: 13 prediction of 48 months sales.

	Sales	Sales First Difference	Seasonal First Difference	forecast
1972-06-01	5312.0	694.0	438.0	NaN
1972-07-01	4298.0	-1014.0	-335.0	NaN
1972-08-01	1413.0	-2885.0	-246.0	NaN
1972-09-01	5877.0	4464.0	-74.0	5878.711681
1972-10-01	NaN	NaN	NaN	7024.262856
1972-11-01	NaN	NaN	NaN	9996.420027
1972-12-01	NaN	NaN	NaN	12892.153458
1973-01-01	NaN	NaN	NaN	4561.508843
1973-02-01	NaN	NaN	NaN	3718.617239
1973-03-01	NaN	NaN	NaN	4792.360570
1973-04-01	NaN	NaN	NaN	5034.610209
1973-05-01	NaN	NaN	NaN	6047.986466
1973-06-01	NaN	NaN	NaN	5486.572784
1973-07-01	NaN	NaN	NaN	4593.628124
1973-08-01	NaN	NaN	NaN	1676.138729
1973-09-01	NaN	NaN	NaN	6148.223890
1973-10-01	NaN	NaN	NaN	7262.512114
1973-11-01	NaN	NaN	NaN	10194.735739
1973-12-01	NaN	NaN	NaN	13057.412651
1974-01-01	NaN	NaN	NaN	4731.200499
1974-02-01	NaN	NaN	NaN	3915.317327
1974-03-01	NaN	NaN	NaN	4961.390913
1974-04-01	NaN	NaN	NaN	5189.452637
1974-05-01	NaN	NaN	NaN	6119.496073
1974-06-01	NaN	NaN	NaN	6875.241966

Fig:14 Sales forecast in Future

IV. CONCLUSION

The purpose of this work is to convey the research's conclusions regarding potential future sales. An innovative technique is put forth in this work.An ARIMA model is used to approach the sales prediction. Finally, a new improved classification technique is implemented. Feature engineering is the process of extracting features from data using domain or field knowledge. to enhance the Sparse Representation Classifier approach. Several experiments are conducted with real-time reviewsto assess the effectiveness of the recently proposed strategy. We have consider an sales report data and applied ARIMA Model for that sales and predicted the forecasting for the sales for three months, five months and three years. Analysis have been done on that sales prediction to determine the manufacturing rate of that product, and get details success rate of that product.

ML is a cutting-edge method for predicting sales rate to increase business profitability. The performance of the organization saw significant improvements as a result of technological innovation and the employment of Python and Jupiter software. The development of new technologies makes it possible to obtain a complete data set and eliminate data complexity in order to estimate sales. The usage of various technologies and algorithms significantly altered how efficiently resource plans were carried out within the business. The effect of ML on sales procrastination has been determined using a secondary data collection technique. The section on recommendations made it possible to completely rethink the computational algorithms used to drastically alter the sales rate.

REFERENCES

[1] C. Lu, F. Wang, G. Trajcevski, Y. Huang, S. Newsam and L. Xiong, "The 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2020)", SIGSPATIAL Special,vol. 12, no. 3, pp. 3-6, 2021. Available:10.1145/3447994.3447997.

[2] "Crop Prediction System Using Machine LearningAlgorithm", Journal of Xidian University, vol. 14, no. 6,2020. Available: 10.37896/jxu14.6/009.

[3] D. V., "Data Mining based Prediction of Demand inIndian Market for Refurbished Electronics", Journal of Soft Computing Paradigm, vol. 2, no. 2, pp. 101-110,2020. Available:10.36548/jscp.2020.2.007

[4] Martínez, C. Schmuck, S. Pereverzyev, C. Pirker and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting", European Journal of Operational Research, vol.281, no. 3, pp. 588-596, 2020. Available:10.1016/j.ejor.2018.04.034.

[5] S. Ji, X. Wang, W. Zhao and D. Guo, "An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise", Mathematical Problems in Engineering, vol. 2019, pp. 1-15, 2019. Available:10.1155/2019/8503252.

[6] "Suicide Prediction on Social Media by Implementing Sentiment Analysis along with Machine Learning", International Journal of Recent Technology and Engineering, vol. 8, no. 2, pp. 4833-4837, 2019. Available: 10.35940/ijrte.b3424.078219.

[7] "A Machine Learning Based Method for Customer Behavior Prediction", Tehnickivjesnik – Technical Gazette, vol. 26, no. 6, 2019. Available: 10.17559/tv-20190603165825.

[8] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C.,Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R. and Flach, P.A., 2019. CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering.

[9] Amalina, F., Hashem, I.A.T., Azizul, Z.H., Fong, A.T.,Firdaus, A., Imran, M. and Anuar, N.B., 2019. Blending big data analytics: Review on challenges and a recent study. Ieee Access, 8, pp.3629-3645.

[10] Li, X., Huang, X., Li, C., Yu, R. and Shu, L., 2019. EdgeCare: leveraging edge computing for collaborative data management in mobile healthcare systems. IEEE Access, 7, pp.22011-22025.

[11] Tyralis, H.; Papacharalampous, G.Variable selection in time series forecasting using random forests.Algorithms 2017,10, 114. [CrossRef].

[12] .Tyralis, H.; Papacharalampous, G.A. Large-scale assessment of Prophet for multi-step ahead forecasting ofmonthly streamflow. Adv. Geosci. 2018,45, 147–153. [CrossRef].

[13] Papacharalampous, G.; Tyralis, H.; Koutsoyiannis, D. Predictability of monthly temperature and precipitationusing automatic time series forecasting methods. Acta Geophys. 2018,66, 807–831. [CrossRef].

[14] A. Telaga, A. Librianti and U. Umairoh, "Sales prediction of Four Wheelers Unit (4W) with seasonal algorithm Trend Decomposition with Loess (STL) in PT. Astra International, Tbk", IOP Conference Series: Materials Science and Engineering, vol. 620, p. 012112, 2019. Available: 10.1088/1757-899x/620/1/012112

[15] B. Pavlyuchenko, "Machine-Learning Models for Sales Time Series Forecasting", Data, vol. 4, no. 1, p. 15, 2019. Available: 10.3390/data4010015..

[16] Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart salesprediction. In: 2018 International Conference on Computing, Power and Commu-nication Technologies (GUCON). pp. 617–620. IEEE (2018).