DEMAND FORECASTING
FOR INDUSTRIAL PRODUCTS

MINI PROJECT REPORT

Submitted by

ADITI KANNAN (203002004)

UEC1605

MACHINE LEARNING

Department of Electronics and Communication Engineering

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

Rajiv Gandhi Salai (OMR), Kalavakkam – 603 110

EVEN SEMESTER 2022-2023

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this mini project titled "**DEMAND FORECASTING FOR INDUSTRIAL PRODUCTS**" is the bonafide work of **Aditi Kannan (203002004) of VI Semester Electronics and Communication Engineering Branch during Even Semester 2022 – 2023 for UEC1605 Machine Learning**

Submitted for examination held on ________________________

INTERNAL EXAMINER

ABSTRACT

Demand forecasting is a crucial process that enables organizations to plan and manage their production, inventory, and supply chain activities effectively. The project aims to apply various statistical machine learning and deep learning techniques to build a robust demand forecasting model that can provide accurate predictions for future demand of industrial products, which helps organizations make informed decisions and optimize their resources to meet future demand. The dataset used for this project includes monthly time series data on the demand for abrasives, along with IIP data and special discount rate provided by the manufacturer. The models employed include ARIMAX, univariate LSTM, and multivariate LSTM. The results show that the ARIMAX model outperformed the LSTM models, with a MAPE of 17.746. Additionally, the study found that external variables, such as IIP and special discount rates, did not have any correlation with demand.

INTRODUCTION

Demand forecasting is an essential tool for organizations to make informed decisions and optimize their resources to meet future demand. Accurate forecasting of demand can help organizations plan and manage their production, inventory, and supply chain activities more effectively. In recent years, advances in statistical machine learning, and deep learning techniques have made it possible to build sophisticated models for demand forecasting. These models can take into account a wide range of factors, including historical demand, market trends, seasonal patterns, and external variables. This project aims to develop a demand forecasting model for industrial products using statistical, machine learning, and deep learning techniques. The study will focus on analyzing monthly time series data on the demand for abrasives, along with IIP data, to build a robust forecasting model that can provide accurate predictions for future demand. The project will compare the performance of various models, including ARIMAX, univariate LSTM, and multivariate LSTM, and evaluate the impact of external variables on demand. The findings of this project will provide insights into how organizations can use advanced techniques to improve their demand forecasting and optimize their resources to meet future demand.

OBJECTIVE

This project aims to investigate the effectiveness of different demand forecasting methods in predicting the future demand for industrial products. The project aims to compare traditional statistical methods with machine learning and deep learning techniques to determine which method provides the most accurate predictions. Additionally, the project will analyze the impact of external variables, such as economic indicators, on demand for industrial products. The findings of this project will help organizations improve their demand forecasting and optimize their resources to meet future demand, ultimately leading to better production planning, inventory management, and supply chain efficiency.

SYSTEM /MODEL DEVELOPED :

The demand forecasting system/model developed in this project utilizes a combination of machine learning, and deep learning techniques to predict the future demand for industrial products. The system is designed to analyze historical demand data, along with external variables, to predict future demand accurately.

The overall process flow of the demand forecasting for industrial products project involves several steps. The first step is to collect and preprocess the input data which includes the demand for abrasives, IIP data, and special discount rate. The collected data is then subjected to several preprocessing techniques such as stationarity, homogeneity, randomness, treatment for outliers, and imputation of missing data. The preprocessed data is then used to build the demand forecasting models. In this project, two models are employed – ARIMAX and LSTM. The ARIMAX model is used to capture the impact of external variables on demand while the LSTM model is used to capture the temporal dependencies in the demand time series. The trained models are then used to predict the future demand of industrial products. The prediction results are then analyzed to identify the accuracy of the models and to gain insights into the demand patterns.

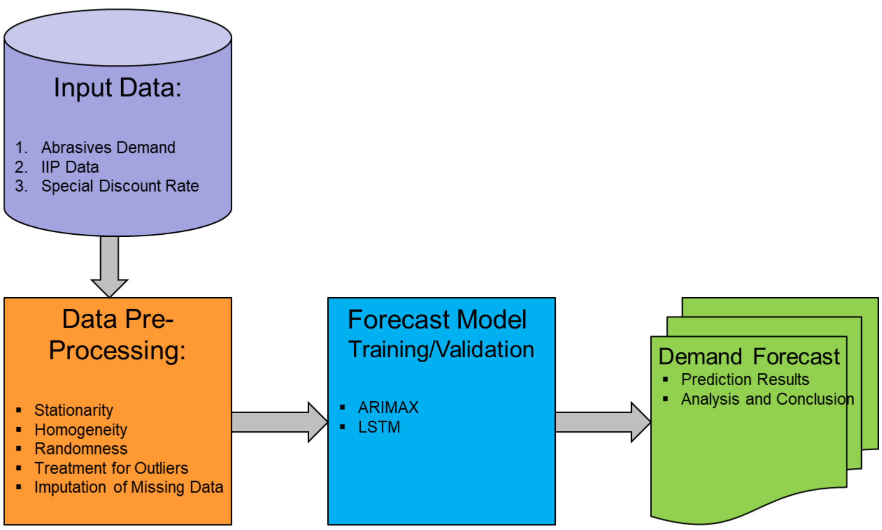

Fig 1: Block Diagram of forecasting system

ARIMAX Model:

ARIMAX (**A**uto**R**egressive **I**ntegrated **M**oving **A**verage with e**X**ogenous variables) is a statistical model that is used to capture the relationship between the demand for industrial products and external variables, such as IIP and special discount rates. The ARIMAX model is an extension of the ARIMA model and includes additional variables as input, which helps to improve the accuracy of the predictions. The ARIMAX model is expressed mathematically as:

$$Y_t = a_o + a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \\ \theta_o + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \\ \beta_o + \beta_1 X_1 + \dots + \beta_k X_k$$

- Y_t is the dependent variable at time t.
- a_1 to a_p are the autoregressive coefficients of order p.
- θ_1 to θ_q are the moving average coefficients of order q.
- ε_{t-1} to ε_{t-q} are the first q moving average terms
- β represents the differencing coefficients of own time series

The autoregressive part (AR) forecasts the variable of interest by using the past values of the variable The integrated part (I) represents the differencing of raw observations to allow the time series to become stationary. The moving average part (MA) uses past forecast errors in a regression-like model. The past forecast error is the difference between the actual data and fitted values. X represents exogenous variables that are independent of time-series but have effect on the time-series data. Exogenous variables are often used to capture the effect of external factors that may influence the behavior of the dependent variable

To choose an appropriate ARIMA model, the time series data must be analyzed to identify any trend, seasonality, or cyclical behavior. The stationarity of the time series data must also be tested, as ARIMA models work best with stationary time series data. Once the data has been preprocessed, the next step is to select the appropriate order of differencing (d), autoregressive (p), and moving average (q) terms. The order of differencing is determined by computing the difference between consecutive observations until the data becomes stationary. The values of p and q are then determined by analyzing the partial autocorrelation and autocorrelation functions of the differenced data. The selected ARIMA model must be validated by testing its goodness-of-fit to the data, using measures such as AIC, BIC, or RMSE. It is important to note that the selection of an appropriate ARIMA model requires expertise and experience, and it may involve several iterations until the best model is identified.

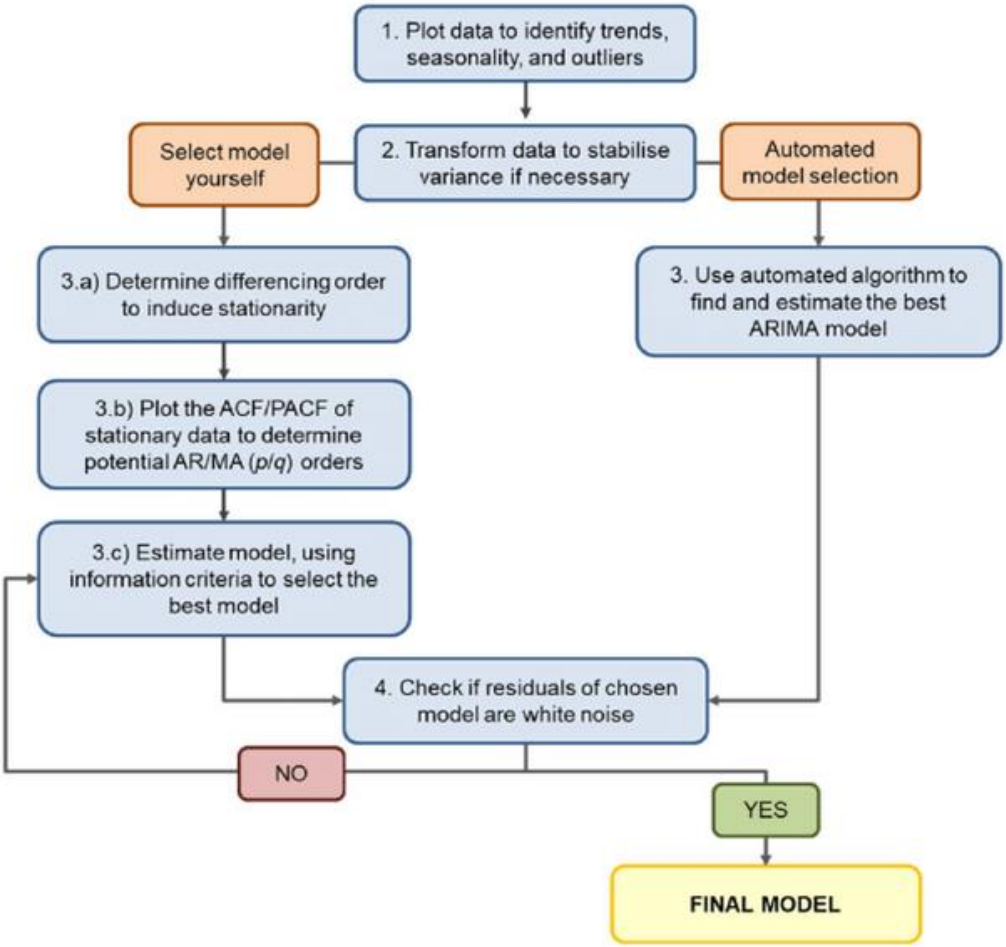

Fig 2: Steps to choose an appropriate ARIMA model

LSTM Model:

The LSTM (Long Short-Term Memory) model is a type of recurrent neural network that is designed to capture the complex nonlinear relationships between variables. The LSTM model is particularly useful when there are many variables or when the relationships are non-linear. The LSTM model can capture seasonal patterns and long-term dependencies in the demand time series data.

LSTMs work by maintaining a cell state that can store information over time, selectively updating and resetting the cell state based on input data. The cell state is controlled by gates, which are a set of sigmoid neural network layers that decide how much information to keep, how much to forget, and how much to output at each time step. LSTM architecture consists of four main components: the input gate, the forget gate, the output gate, and the cell state. The Input gate determines how much of the input should be added to the cell state. The Forget gate determines how much of the previous cell state should be forgotten. The Output gate determines how much of the cell state should be output. Cell state represents the memory of the LSTM that stores information over time. The Proposed Project uses a stacked LSTM with 50 hidden layers, 1 dense layer, Adam optimizer and MAE loss function.

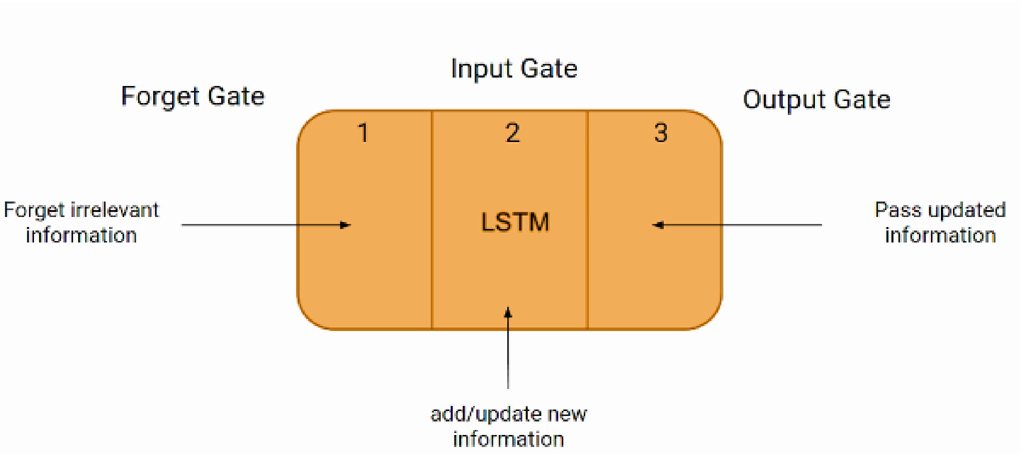

Fig 3: Structure of an LSTM model

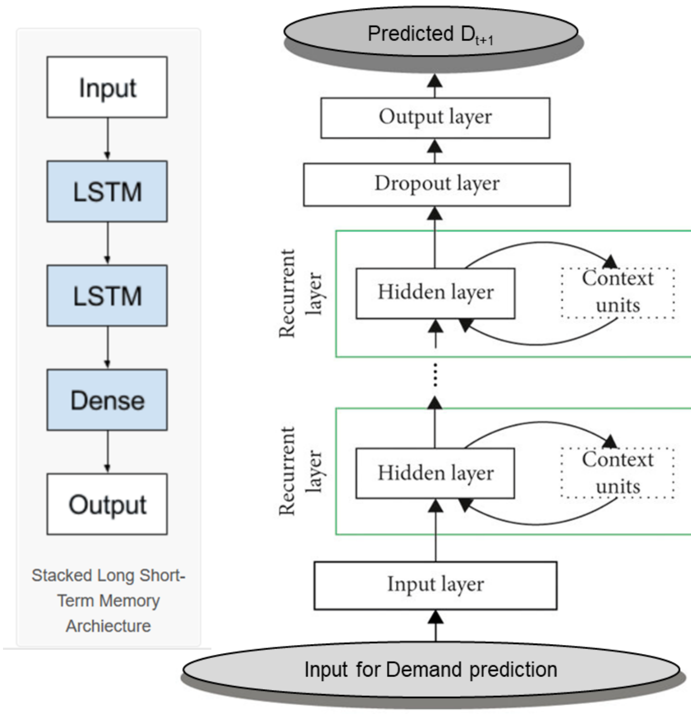

Fig 4: Block Diagram of stacked LSTM model used

Dataset Details:

The dataset used in this project includes monthly time series data on the demand for abrasives, along with IIP data. The demand data is from January 2009 to November 2021, the IIP data is from January 2014 to July 2021, and the Special Discount Rate data is from January 2009 to November 2021.

Data Availability Statement

		Availablity	
Sl. No	Description	From	To
I. Abrasives Demand			
	a) BC26	"2009-03-01"	"2021-11-01"
	b) BC27	"2009-03-01"	"2021-11-01"
II. Index of Industrial Production (IIP)			
	IIP_MFG	"2014-01-01"	"2021-07-01"
III. Special Discount Rate			
	a) BC26	"2009-03-01"	"2021-11-01"
	b) BC27	"2009-03-01"	"2021-11-01"

Fig 5: Dataset Details

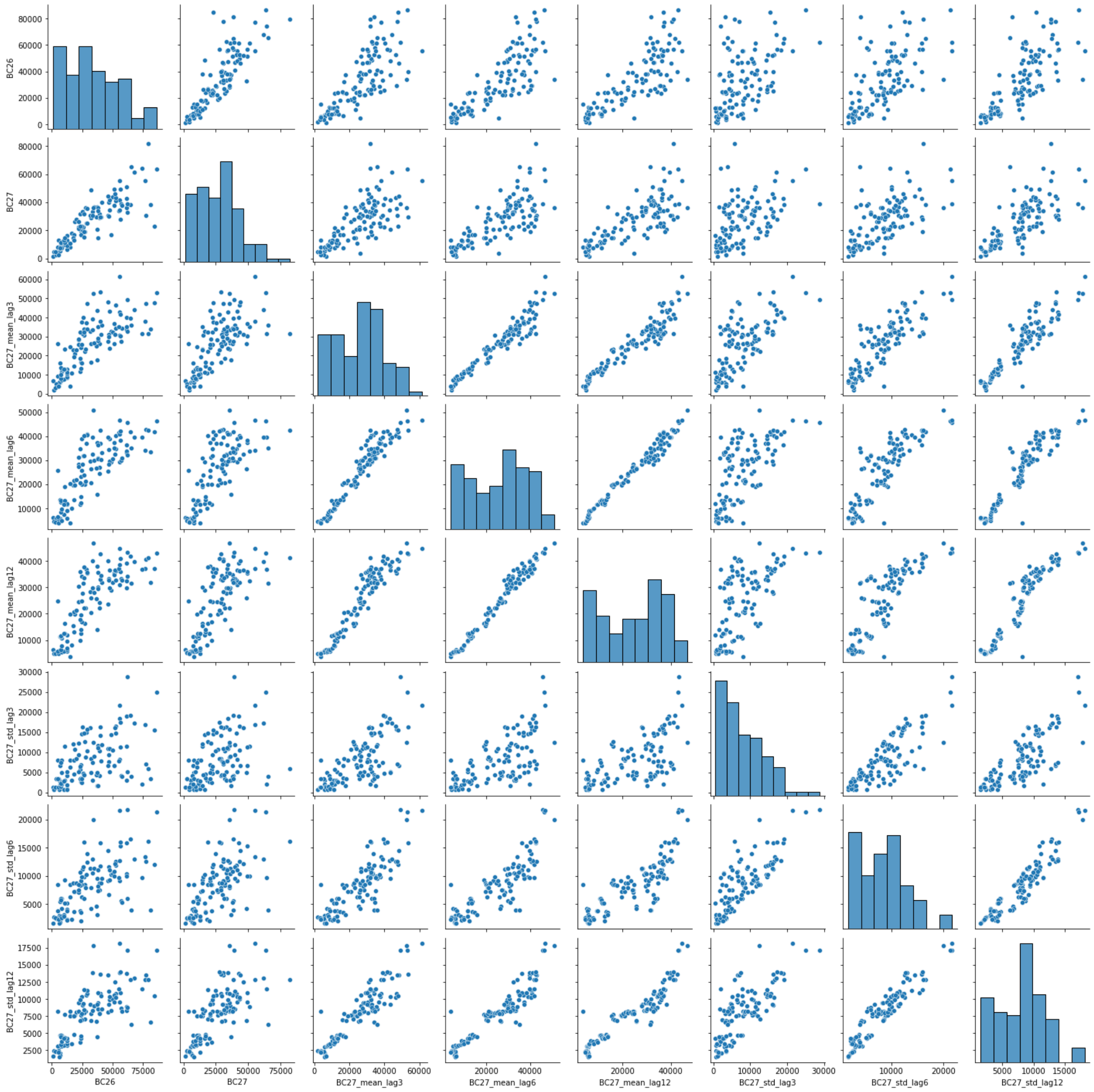

Fig 6: Pair plot depicting relationship between variables

Data Quality:

Data quality check are important measures to understand the quality of time series under consideration. The following statistical tests are performed to check whether the given time series is stationary, homogeneous, and random.

1. ADF and KPSS Test for Stationarity
2. Test for Homogeneity
3. Test for Randomness

The statistical test results are presented in the following tables:

Parent Item Code	ADF Statistic	p-value	Critical Value_1%	Critical Value_5%	Critical Value_10%
BC26	-7.179045579	2.68E-10	-3.505190196	-2.894232085	-2.584210123

Parent Item Code	KPSS Statistic	p-value	Critical Value_1%	Critical Value_5%	Critical Value_10%
BC26	0.705065883	0.01308	0.739	0.463	0.347

ADF	$p < 0.05$	the time series is stationary
KPSS	$p < 0.05$	the time series is non-stationary

Fig 7: ADF and KPSS test for Stationarity

Parent Item Code	h	Change Point	p-value	Test Statistic	Mean1 Before CP	Mean2 After CP
BC26	TRUE	38	0.00015	1046	45242.10526	66596.22642

Fig 8: Test for Homogeneity

Parent Item Code	Z-Value	Z-Critical	TS Random
BC26	1.369414	1.96	TRUE

Fig 9: Test for Randomness

Since KPSS test results shows non-stationarity and ADF test results indicates stationarity, the time series is difference stationary. Hence differencing is to be used to make the time series stationary. As regards the test for homogeneity, h=TRUE indicates that the time series is non-homogeneous and there is a change point detected at 38 position of the timeseries and the mean before and after CP are different. As regards the test for randomness, $|Z| <$ Z-critical indicates that the timeseries is random.

RESULTS & DISCUSSION

ARIMAX (3,0,2) has been identified as the best model for timeseries forecast. The following are the parameters of the trained ARIMAX model used for 1-step ahead forecast.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   BC26   No. Observations:                   91
Model:                 ARIMA(3, 0, 2)   Log Likelihood               -1044.340
Date:                Mon, 08 May 2023   AIC                           2102.681
Time:                        00:35:55   BIC                           2120.257
Sample:                    01-31-2014   HQIC                          2109.772
                         - 07-31-2021
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       5.768e+04   1.82e+04      3.177      0.001    2.21e+04    9.33e+04
ar.L1          0.0775      0.179      0.432      0.665      -0.274       0.429
ar.L2          0.8858      0.081     10.941      0.000       0.727       1.044
ar.L3          0.0002      0.121      0.001      0.999      -0.237       0.237
ma.L1          0.1053      0.243      0.433      0.665      -0.372       0.583
ma.L2         -0.8837      0.169     -5.225      0.000      -1.215      -0.552
sigma2      5.771e+08      0.222    2.6e+09      0.000    5.77e+08    5.77e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                14.60
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               4.79   Skew:                             0.49
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.70
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 5.95e+25.
Standard errors may be unstable.
"""
```

The following figures show PACF and ACF plots for the selected timeseries (BC26). These plots are in tune with the model parameters selected.

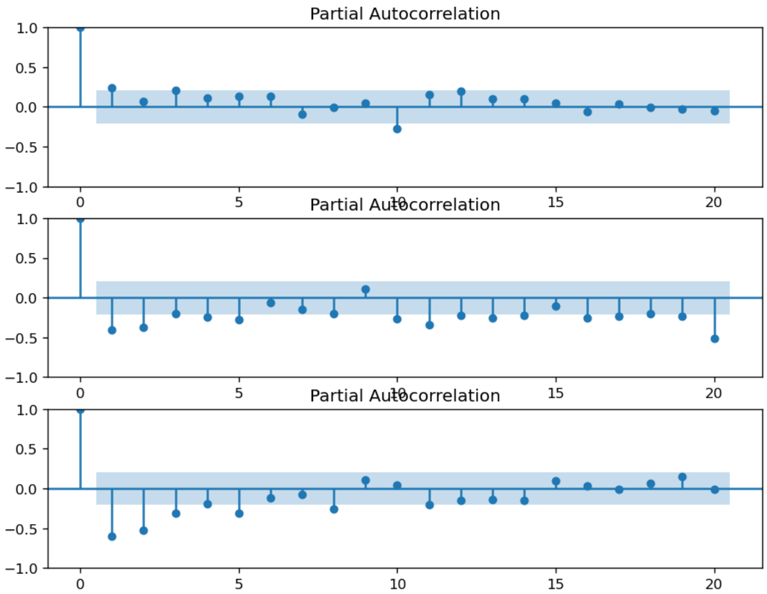

Fig 10: PACF plot for original, difference 1, and difference 2 timeseries

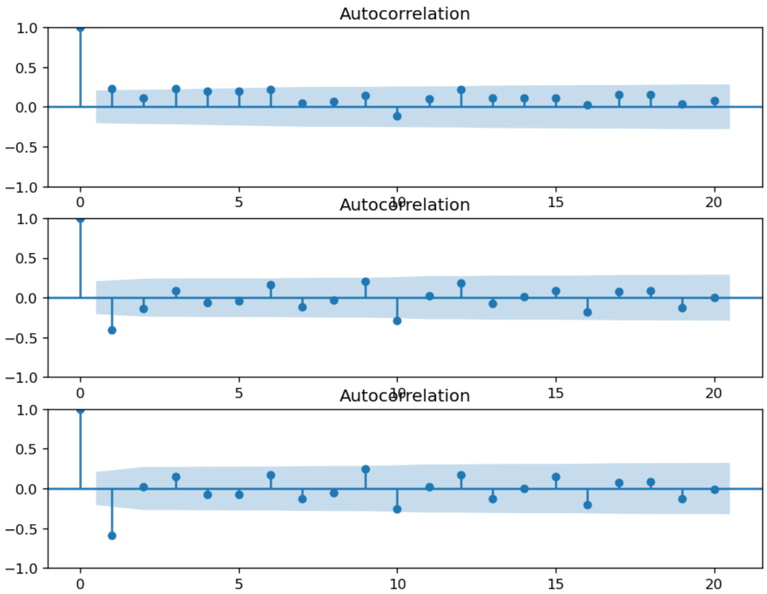

Fig 11: ACF plot for original, difference 1, and difference 2 timeseries

Similarly, LSTM univariate and LSTM multivariate models are build with 100 hidden layers, 1 output layer, 'MAE' loss function, and Adam optimizer for training the model.

This project makes use of MAPE, or Mean Absolute Percentage Error, which is a commonly used performance metric in forecasting models. It measures the average absolute percentage difference between the actual and predicted values. MAPE provides a clear understanding of the accuracy of the model, with lower values indicating higher accuracy. It is a relative measure and is expressed as a percentage.

Formula

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error

n = number of times the summation iteration happens

A_t = actual value

F_t = forecast value

Fig 12: Performance metrics used

Table 1: Model Accuracy

Model	MAPE
ARIMAX	17.75
LSTM - Univariate	95.91
LSTM - Multivariate	29.77

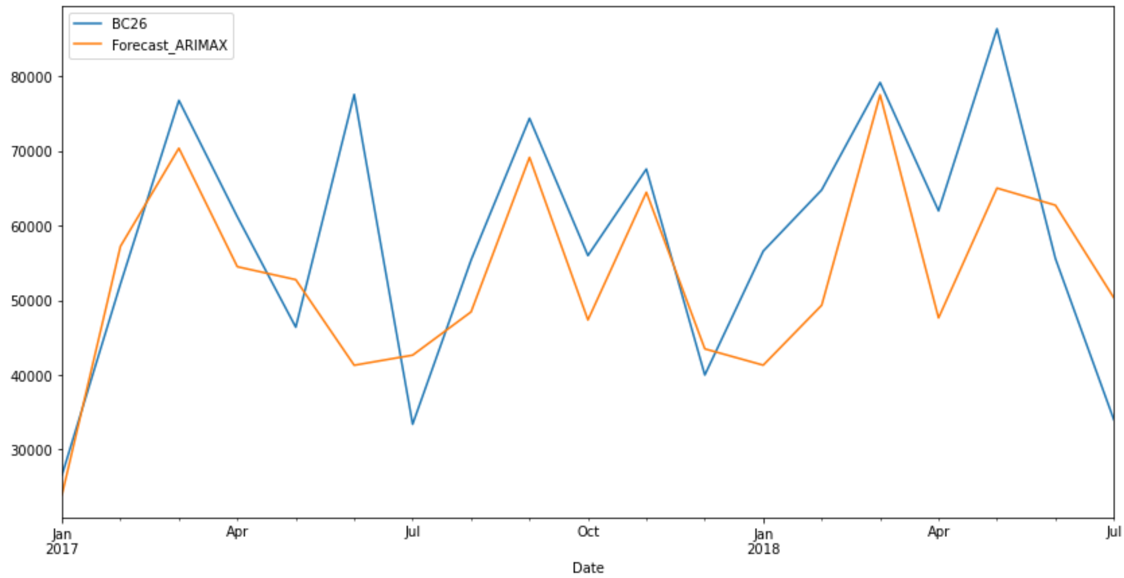

Fig 13: Comparison of observed and predicted demand for ARIMAX(3,0,2)(0,0,0)[0]

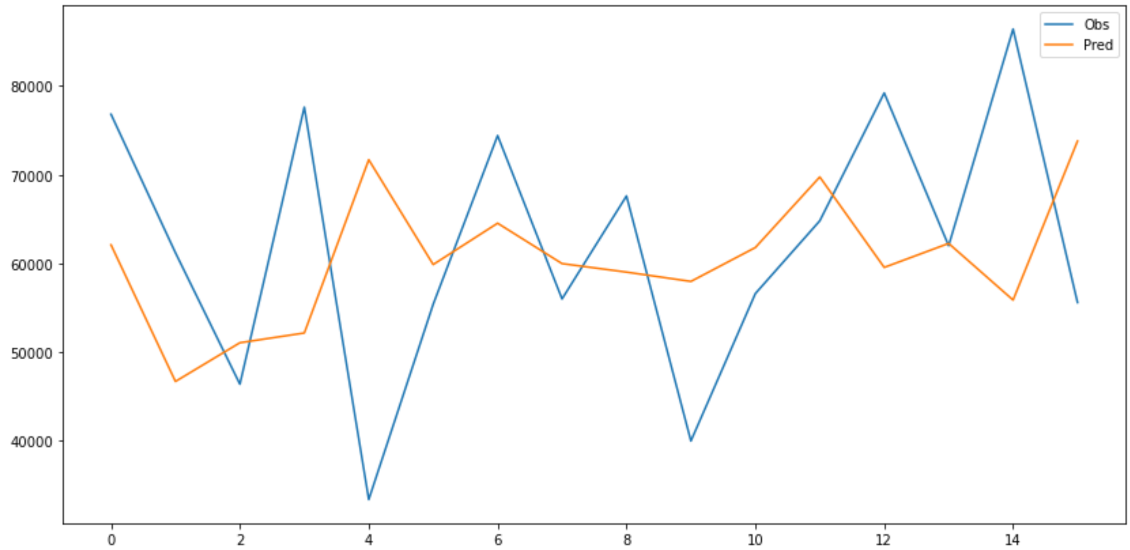

Fig 14: Comparison of observed and predicted demand for Univariate LSTM

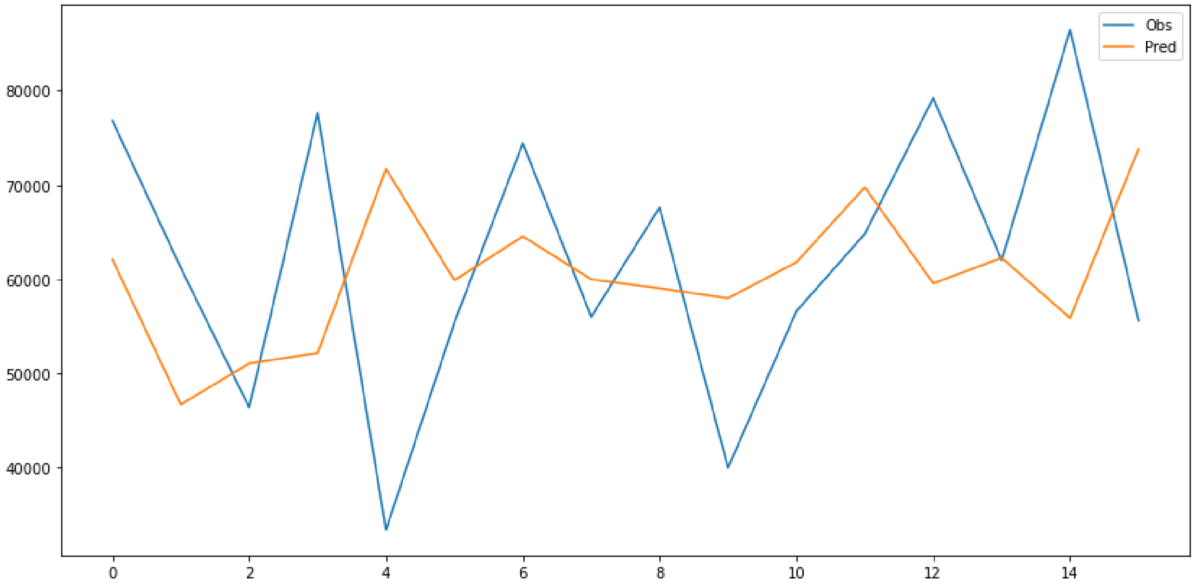

Fig 15: Comparison of observed and predicted demand for Multivariate LSTM

The ARIMAX model performs better than both the univariate and multivariate LSTM models, mainly due to the optimized parameters (p, d, q) used for forecasting. The ARIMAX model is able to capture the correlation structure between the predictor and predictand, resulting in more accurate predictions. However, despite trying various input combinations, the LSTM models did not significantly reduce the MAPE, indicating that the long and short-term memory did not impact the output significantly.

CONCLUSION

In conclusion, this project aimed to develop a robust demand forecasting model for industrial products using various statistical machine learning, and deep learning techniques. The ARIMAX model outperformed both the univariate and multivariate LSTM models, providing more accurate predictions. However, the external variables such as IIP and special discount rate did not have any correlation with demand as it did not have any specific annual pattern. Overall, the developed model can help organizations in planning and managing their production, inventory, and supply chain activities effectively.

REFERENCES

[1] P. Pawar, S. Hatcher, L. Jololian and T. Anthony, "Demand Forecasting using Machine Learning," 2019 SoutheastCon, Huntsville, AL, USA, 2019, pp. 1-8, doi: 10.1109/SoutheastCon42311.2019.9020393.

[2] M. E. Hoque, A. Thavaneswaran, S. S. Appadoo, R. K. Thulasiram and B. Banitalebi, "A Novel Dynamic Demand Forecasting Model for Resilient Supply Chains using Machine Learning," 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2021, pp. 218-227, doi: 10.1109/COMPSAC51774.2021.00040.

[3] K. Wanchoo, "Retail Demand Forecasting: a Comparison between Deep Neural Network and Gradient Boosting Method for Univariate Time Series," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033651.

[4] Bontempi, G., Taieb, S. and Borgne, Y. (2013). Machine Learning Strategies for Time Series Forecasting. Brussels, Belgium: SpringerVerlag. Available at: https://www.academia.edu/20245534/Machine_Learning_Strategies_for_Time_Series_Forecasting.

[5] Bandar, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q. and Seaman, B. (n.d.). San Bruno, USA. Available at: https://www.academia.edu/38165760/Sales_Demand_Forecast_in_Ecommerce_using_a_Long_ShortTerm_Memory_Neural_Network_Methodology.

[6] S. Mehrmolaei and M. R. Keyvanpour, "Time series forecasting using improved ARIMA," 2016 Artificial Intelligence and Robotics (IRANOPEN), Qazvin, Iran, 2016, pp. 92-97, doi: 10.1109/RIOS.2016.7529496.

[7] S. Siami-Namini, N. Tavakoli and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 1394-1401, doi: 10.1109/ICMLA.2018.00227.