Exploring Forecasting Methods in Supply chain: A comparison of Time Series Analysis and Machine Learning Approaches for Demand Prediction

K.C.Bhanu[1], Dr.P.Uma Maheswari Devi[2]
[1]Research Scholar, Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry
[2]Associate Professor **,** Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry
bhanu1605@gmail.com,umadevi_4@yahoo.com

ABSTRACT

In this study, we use time series analysis and machine learning to forecast demand. The study on demand forecasting is focused on Walmart, a multinational American retailer. to guarantee that the set of inputs utilised to produce the final output are taken from the dataset in order to almost perfectly forecast Walmart's demand. To identify the best accuracy, we will use a number of methods, including the Lasso Regressor, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor, and Time Series Analysis. Iterations were performed to find the best parameters before building the model.

Keywords: -Machine Learning , Supply chain, Demand forecasting, Time Series Analysis, Feature selection.

I. INTRODUCTION

Making projections regarding a company's potential future sales is one of the most important aspects of strategic planning. Walmart is the ideal case study for novices to utilise since it has the biggest retail data collection. Walmart used this issue with the sales projection as a recruiting technique as well. We are tasked with predicting sales across the many business areas. There are many sections in each shop. It is important to remember that we also have access to external data, such the CPI, unemployment rate, and gasoline costs close to each shop, which should ideally support the development of a more comprehensive research.

Over time, it has been harder and harder to forecast what will happen in the future. But because of a number of developments, including the globalisation of supply chains, the proliferation of product variants, the shortening of product life cycles, and increased market competition, its importance has increased.

In recent years, academics have focused a lot of attention on using machine learning prognostic models to estimate product sales. A statistical tool for predictive modelling is the random forest regression method. It entails building several decision trees and aggregating the results of each to arrive at a final forecast. Applications for machine learning and data analysis often use this strategy.

Extreme gradient boosting is a machine learning method that use a boosting strategy to incrementally enhance a model's performance by fusing together many weak models into a stronger one. The XGBoost algorithm is an ensemble approach based on decision trees that uses a gradient boosting framework. By optimising the complexity penalty and loss function, XGBoost improves the optimisation of the objective function.

Support-vector regression (SVR) is the SVM, or support-vector machine. Support Vector Machine (SVM) is a mathematical approach that can do classification and regression tasks because it creates a hyperplane in a multidimensional space. Based on the given training data, the multiple linear regression model used x as the independent variables and y as the dependent variable, especially SBP.

A statistical technique for variable selection and regularisation in linear regression models is called the Least Absolute Shrinkage and Selection Operator (LASSO). When working with high-dimensional data sets, where the number of predictors is much more than the number of observations, this strategy is very helpful. In predictive modelling, the LASSO approach is used to reduce model complexity and avoid over-fitting.

II.LITERATURE SURVEY

The research article "Information sharing in a supply chain." was written by Lee, Hau L., and Seungjin Whang and published in 2000. The current research identifies the many types of information that are shared among the various parties involved, including sales, inventory, demand forecasts, order statuses, and production schedules. The paper also elaborates on three different information-sharing system types.

The research paper "Forecast of sales of Walmart store using big data applications" was written by "Harsoor, Anita S., and Anushree Patil" in 2015. This research analyses sales data from Walmart shops located across a variety of geographic locations in order to get insight into the major factors affecting the retail store business. The study will also look at how sales forecasting might help with efficient resource management inside retail networks.

The study "Walmart Gross Sales Forecasting Using Machine Learning" was released in 2021 by Mounika, S., et al. This study describes the use of machine learning methods that use data from the past to predict future outcomes. The use of algorithms allows for the forecasting of future sales. By adding additional characteristics, the algorithm that produces the best degree of accuracy is chosen and examined. Future sales may be predicted as a result of this method.

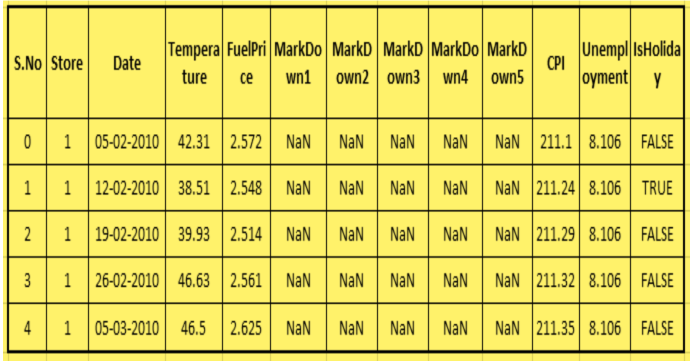
The research paper "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering" was written by Yiyang Niu and published in 2020. This essay suggests a fresh approach to forecast Walmart sales. In this research, the Walmart sales dataset from Kaggle was analysed using 'XGBoost' in combination with feature engineering methods. Empirical findings show that our method beats both the "Logistic Regression" and "Ridge" algorithms and is effective at extracting features of various dimensions. and examined the attributes' relative relevance rankings to acquire some intriguing and helpful advice.

III.RELATED WORK

The dataset was obtained through the Kaggle website, and it was maintained in an orderly manner. In this dataset, the variables listed below are

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 11, Nov 2022**



S.No	Store	Date	Temperature	FuelPrice	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	05-02-2010	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.1	8.106	FALSE
1	1	12-02-2010	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.24	8.106	TRUE
2	1	19-02-2010	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.29	8.106	FALSE
3	1	26-02-2010	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.32	8.106	FALSE
4	1	05-03-2010	46.5	2.625	NaN	NaN	NaN	NaN	NaN	211.35	8.106	FALSE

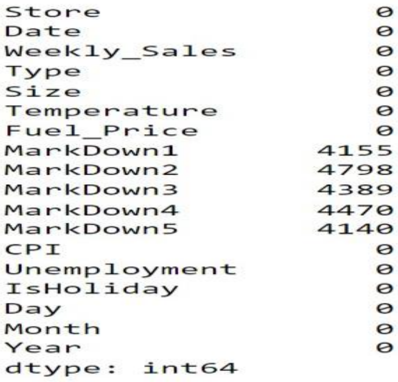
TABLE I. Info of Dataset

Stores:	**Features:**	**Sales:**
Store: The number of Walmart Stores in that area (1-45) . **Type:** The stores are marked as A,B and C based on type.. **Size:** This represent the number of products are there in the store from 34,000 to 210,000.	**Temperature:** The thermal conditions of area in that week. **Fuel Price:** Fuel Price in that region during that week. **MarkDown1-5 :** Represents the Type of markdown and what quantity was available during that week. **CPI:** Consumer Price Index during that week. **Unemployment:**The unemployment rate during that week in the region of the store	**Date:** The date of the week where this observation was taken. **Weekly_Sales:** The sales recorded during that Week. **Dept:**It shows the Department data of 1–99 stores department. **IsHoliday:** A Boolean value representing a holiday week or not.

IV. METHODOLOGY

A. Data Cleaning

Verify whether any of the data's values are null. We would eliminate null values if they were present in extremely small amounts since they have no effect on the data's content. If the fraction of null values is large, we handle them using the proper procedures. After handling the null values, we concentrate on the data needed for model deployment and dataset information verification.



```
Store              0
Date               0
Weekly_Sales       0
Type               0
Size               0
Temperature        0
Fuel_Price         0
MarkDown1       4155
MarkDown2       4798
MarkDown3       4389
MarkDown4       4470
MarkDown5       4140
CPI                0
Unemployment       0
IsHoliday          0
Day                0
Month              0
Year               0
dtype: int64
```

Fig:1

B. Data Pre-processing

Data processing is the collection and modification of data in digital form to provide meaning-laden information. It falls within the category of data processing, which is the change (processing) of data in any visible manner.

Since there were more missing values in this dataset for Markdown variables, we filled each one with a zero where it was needed. We also put into use imputation techniques as KNN imputation, Linear imputation, and Multiple imputation.

V. EXPLORATORY DATA ANALYSIS

"Exploratory Data Analysis" (EDA) is a typical technique used to examine and analyse data sets with the intention of distilling their key characteristics. Data representation techniques are often used in this procedure. The use of this methodology makes it easier to find the best ways to manipulate data sources in order to get the results you want. Finding tools for data professionals in pattern identification, anomaly detection, hypothesis testing, and assumption verification. The properties of the data may be used for complex data analysis or modelling when EDA is finished and insights have been extracted.Through visualisation, we discovered insights from the Walmart dataset in this. We loaded packages like Matplotlib and Seaborn into a Jupyter notebook to visualise. We discovered a better grasp of the variables, parameters, and relationships between them in the dataset using these techniques.

The dataset was used to create the following graphics, which depict the data and show its patterns and trends.

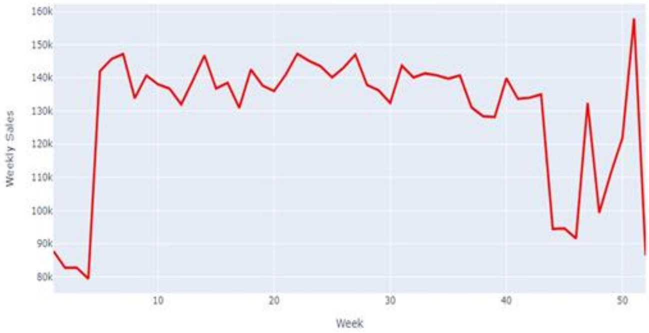


Fig:2 Weekly Sales

Using plot, we compared sales throughout the weeks in figure 2. As we can see, there was a little fall in sales during the first few weeks, which was followed by a rapid surge in sales that climbed substantially and likewise followed pattern up until 43 weeks. The sales finally reached their peak.

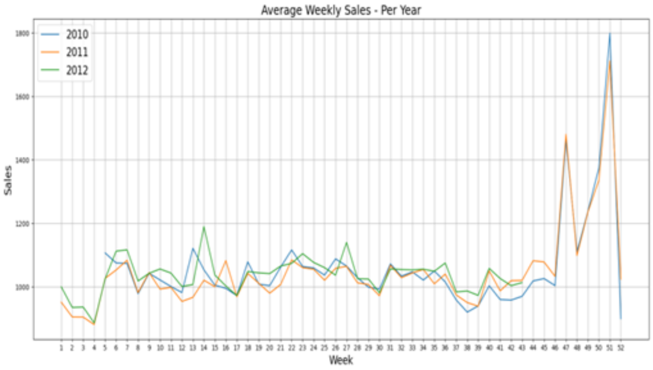


Fig-3: Average weekly sales per year

We represented the average weekly sales for the three years (2010, 2011 and 2012) in this graph. We saw that sales from three years ago followed the same trend over the course of a week, and eventually we discovered that sales across three years almost touched the same mean point.

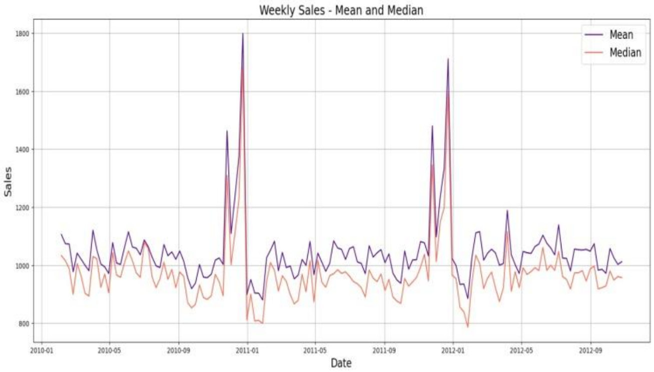


Fig-4:Weekly Sales-Mean and Median

We represented the mean and median weekly sales in Figure 4. Due to the festival season in the United States (USA), we have seen that the weekly sales are at their highest level in the final month of every year.

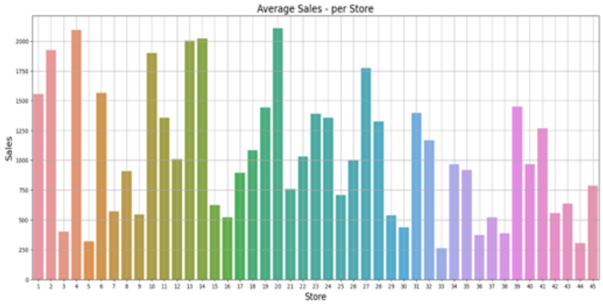


Fig-5: Average Sales – per store

We displayed the 45 shops' average sales in Figure 5. The top 22 shops out of 45 had higher average sales per store than the other half of the retailers. Store 33 had the lowest average sales.

VI. TIME SERIES

A time series is a collection of data or observations that were obtained or noted over a period of time. It is a group of time-stamped data points, to put it another way. Time series data are widely used in many areas, including economics and finance, banking, forecasting the weather, market research for investments, and signal processing.

Depending on the kind of data and the field in which it was collected, the time intervals between each data point in a time series might vary from minutes to years. Each "point" in a data collection is a discrete value or measurement of a significant temporal variable.

A. Stationary Test:

Stationary time series data have statistical features such as mean ,average, variance, and covariance that remain constant over time. A Stationary test is a statistical procedure that determines whether the time series data is stationary or not. For these data, the 'Augmented Dickey-Fuller (ADF)' test is performed. This is used to determine whether a time series contains a unit root.

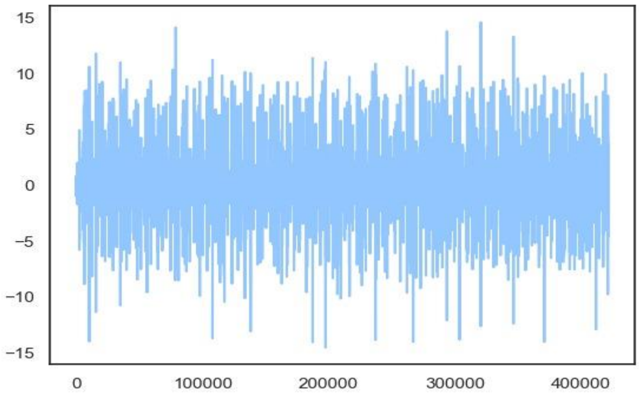


Fig-6:AD-Fuller Test

We used the AD-Fuller test to determine if the data in Figure 6 were stationary. The mean, variance, and covariance of the provided data remain constant across time. Therefore, the provided data is a stationary time series.

B. Decomposition:

Decomposition is a statistical method for separating the hidden components of a time series, which include seasonality, trend, and residual.

- **Trend:** The long-term trend that has been seen in the data collection over time. It may have an upward, downward, or flat tendency.
- **Seasonality:** A cyclical pattern that occurs across time, such as daily, weekly, monthly, or even annually. Seasonality includes other elements like holidays, weather, etc.
- **Residual:** This shows the irrelevant or erratic changes that cannot be accounted for by seasonality or trend.

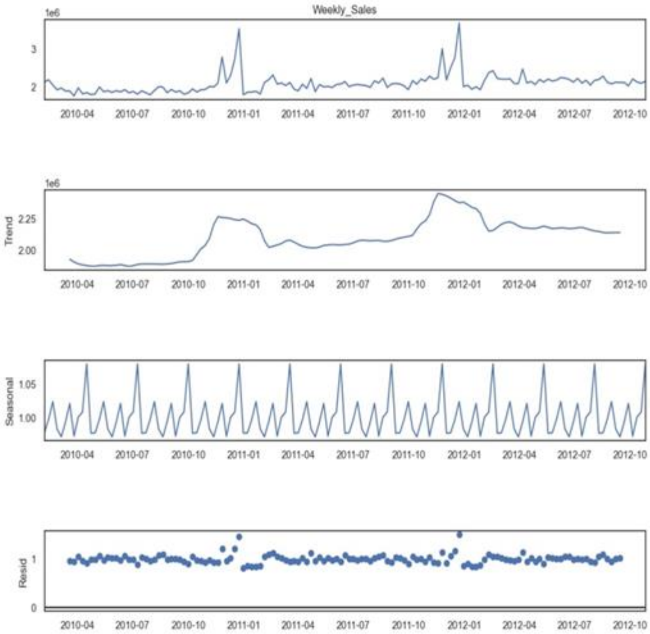


Fig-7:Seasonal Decomposition of Weekly Sales-4

We did a seasonal decomposition on the weekly sales-4 in Figure-7. We found that weekly sales data exhibits seasonality, residual, and trend.

C. SARIMA:

The most common models used for demand forecasting are seasonal autoregressive integrated moving average models (SARIMA). We can model trends using seasonal Arima models (SARIMA) if we are working with a time series where trends appear at regular intervals.

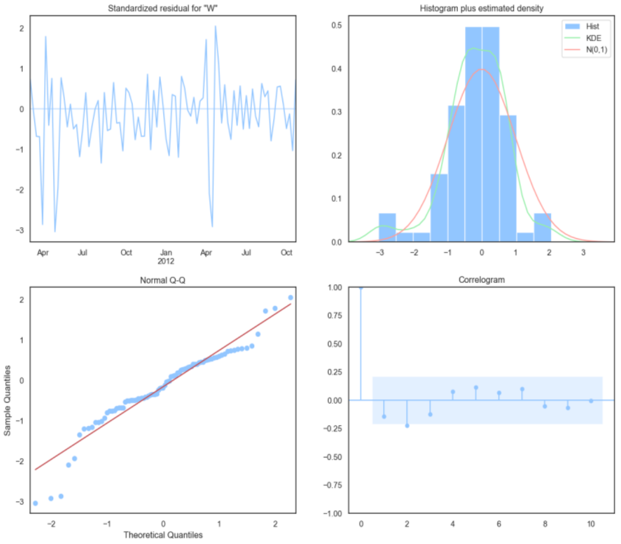


Fig-8:SARIMA

We used SARIMA to create Figure-8.We found that SARIMA—seasonal autoregressive integrated moving average models—are used to analyse the data.

D. Sales Forecasting:

Time series predictions are the process of using scientific hypotheses obtained from historical data with temporal markers. The procedure entails building models based on prior research, utilising them to draw conclusions, and using them to guide future tactical decisions.

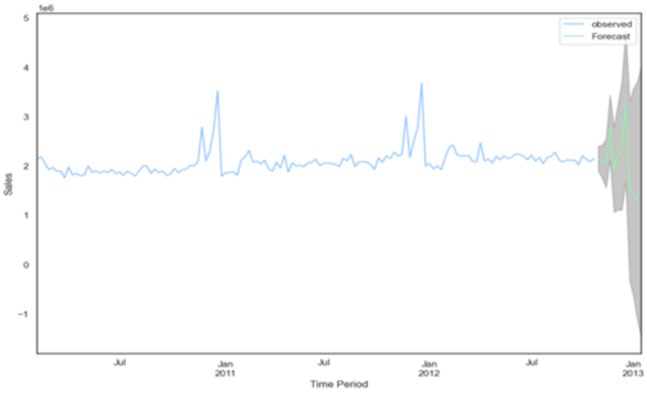


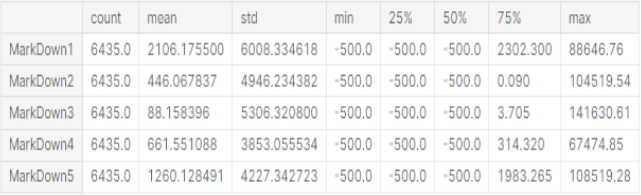
Fig-9:Sales Forecasting

Using the data from the preceding graph, we forecasted revenues for the next three months (i.e., 90 days) in Figure 9.

VII. IMPUTATION

A. KNN Imputation

The KN neighbours imputation uses the data of the dataset's nearest neighbours to estimate missing values.



	count	mean	std	min	25%	50%	75%	max
MarkDown1	6435.0	2106.175500	6008.334618	-500.0	-500.0	-500.0	2302.300	88646.76
MarkDown2	6435.0	446.067837	4946.234382	-500.0	-500.0	-500.0	0.090	104519.54
MarkDown3	6435.0	88.158396	5306.320800	-500.0	-500.0	-500.0	3.705	141630.61
MarkDown4	6435.0	661.551088	3853.055534	-500.0	-500.0	-500.0	314.320	67474.85
MarkDown5	6435.0	1260.128491	4227.342723	-500.0	-500.0	-500.0	1983.265	108519.28

More than 50% of the data here have missing values of -500.Therefore, using KNN for imputing on this dataset is not advised. because the R2 and accuracy are terrible.

MD1 vs Weekly Sales	14%
MD2 vs Weekly Sales-	41%
MD3 vs Weekly Sales-	20%
MD4 vs Weekly Sales-	34%
MD5 vs Weekly Sales-	08%

TABLE-II. R–Square Values of MD Vs Weekly_Sales

B. Linear Regression

By using the relationship between the parameter with missing values and other factors in the dataset, linear regression imputation predicts missing values.We imputed missing data in the Markdowns (1–5) using linear regression.This has a linear regression imputation with an R2 (test) of 93% and an R2 (train) of 94%.

C. Multiple Linear Regression

By building several complete datasets and separately analysing each dataset, taking into account the disagreement arising from missing values, multiple regression imputation generates many alternative values for data that is missing.

We imputed missing data in the Markdowns (1–5) using multiple linear regression. Multiple Linear Regression in this case has R2 (test) of 94%, R2 (train) of 94%, and (Adjusted R2 of 94%).We obtained high R square values for both the test and train sets of data in Multiple Regression Imputation after using the aforementioned three imputation approaches. As a result, we use multiple imputation to fill in the gaps in the data.

VIII. 'MACHINE LEARNING MODELS'

Creating algorithms and statistical models that can analyse data, learn from it, and make predictions is the focus of the discipline of artificial intelligence known as machine learning.

1) Linear regression

One or more independent variables and the dependent factor are linearly related using the machine learning process known as linear regression. When there is only one independent variable, univariate linear regression is appropriate, but multivariate logistic regression is appropriate when there are numerous independent features.

$$\boldsymbol{y = ax + b + e}$$

The above equation is a linear regression model, with the dependent variable being y, the independent variable being x, the slope being a, the intercept being b, and the error term being e

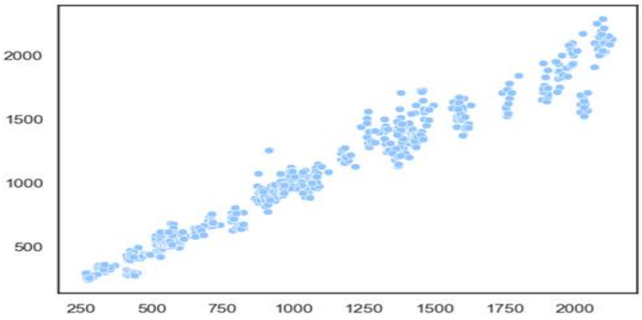


Fig-10.Scatterplot of Predicted Value for Linear Regressor Model

2) Decision Tree Regressor

A prominent ML strategy for dealing with regression and classification problems is the decision tree approach. Its success in these fields accounts for its appeal. The classifier is built up in the form of a tree, with each leaf node designating the classification result and inner nodes expressing the characteristics of the data set.

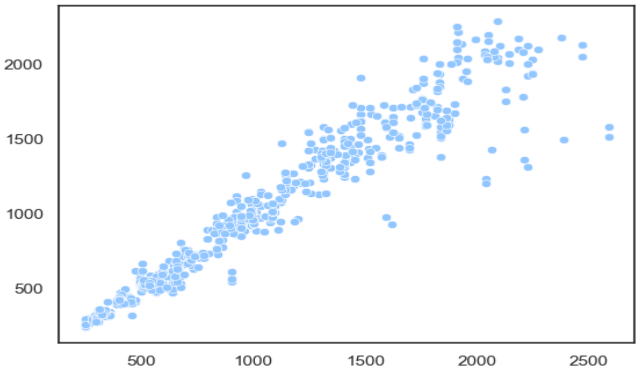


Fig-11.Scatterplot of Predicted Value for Decision Tree Model

3) Random Forest Regressor

ML methods that come under the heading of supervised learning include the Random Forest. This method may be used for both classification and prediction tasks. The idea behind this strategy is based on ensemble learning, a method that combines many classifiers to tackle challenging issues and enhance model effectiveness.

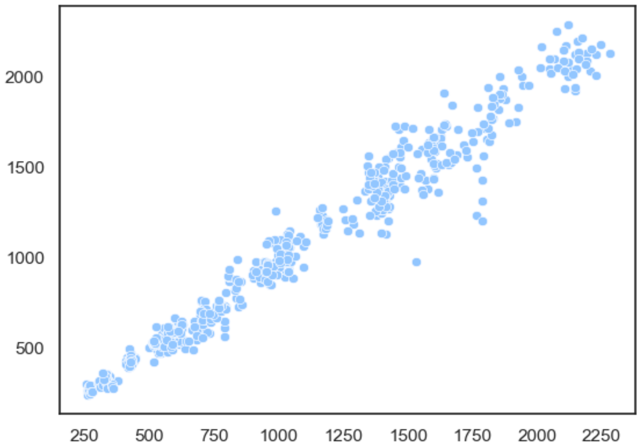


Fig-12.Scatterplot of Predicted Value for Random Forest Model

4) Support Vector Regressor

For both the classification and regression problems in supervised learning, the Support Vector Machine (SVM) method is often used. It mostly serves classification purposes in the context of machine learning. The goal of the technique is to create a decision boundary or ideal line that may efficiently divide an n-dimensional space into several classes, making it easier to accurately classify prospective data pieces. A hyperplane is a frequent name for the ideal decision boundary that successfully divides the data points into several groups.

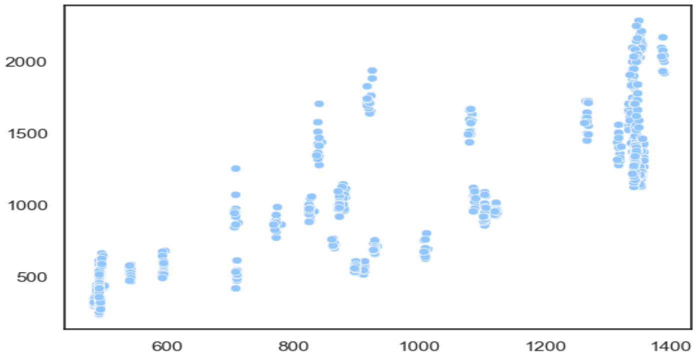


Fig-13.Scatterplot of Predicted Value for Support Vector Regressor Model

5) XGBoost Regressor

The XGBoost is a learning method that combines the forecasts from many weak models to create a more accurate forecast. Due to its ability to handle enormous data sets and achieve efficacy for a variety of machine learning tasks, including classification and regression, XGBoost, an acronym for "Extreme Gradient Boosting," has become a widely used machine learning algorithm. A key feature of the programme is its ability to handle absence values well, allowing it to manage real-world data containing absent values with little processing.

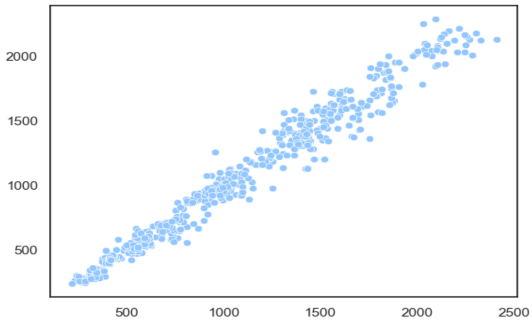


Fig-14.Scatterplot of Predicted Value for XGBoost Model

6)Lasso Regression

A normalisation technique is called lasso regression. It is used in place of regression methods to increase the accuracy of prognostications. The term "shrinking" refers to the phenomena of bringing data values closer to the mean. It demonstrates a bias for models with few parameters, or parsimonious models. This kind of regression works well for models with high degrees of multicollinearity or where certain model selection processes, such as removing parameters and selecting variables, should be automated.

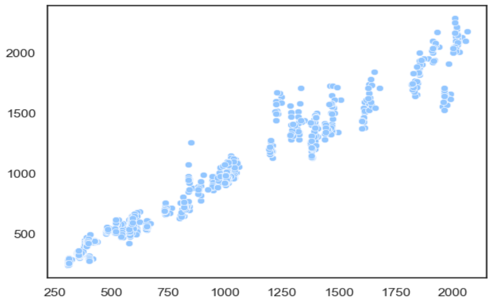


Fig-15.Scatterplot of Predicted Value for Lasso Regression Model

IX. CONCLUSION

Autoregressive and moving average features are used in the SARMIA Time Series model to find patterns and seasonality in a dataset. We predicted sales over the next three months using this methodology. This helps us understand the demand for the product. As a result, by satisfying client requests in the next months, companies may raise their profits.

S.No	Machine Learning Model	MAE	MSE	RMSE
1	Linear regression	71.67	10837.5	104
2	Decision Tree Regressor	93	25871.6	160.8
3	Random Forest Regressor	64	8913	94.4
4	Support Vector Regressor	233	100109	316.4
5	XGBoost Regressor	64.4	8454.6	92
6	Lasso Regression	80.6	12774	113

TABLE-III. RMSE Scores

We used the aforementioned machine learning models to forecast, and out of the six models, the XGBoost Regressor had the lowest Root Mean Square Error (RMSE).The XGBoost Regressor is the best model for the dataset, we conclude.

X.REFERENCES

1. Lee, Hau L., and Seungjin Whang. "Information sharing in a supply chain." International journal of manufacturing technology and management 1.1 (2000): 79-93.
2. Harsoor, Anita S., and Anushree Patil. "Forecast of sales of Walmart store using big data applications." International Journal of Research in Engineering and Technology 4.6 (2015): 51-59.
3. Mounika, S., et al. "Walmart Gross Sales Forecasting Using Machine Learning." Journal of Advanced Research in Technology and Management Sciences (JARTMS)

3.4 (2021): 22-27.

4. Niu, Yiyang. "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering." 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2020
5. Grean, Michael, and Michael J. Shaw. "Supply-chain partnership between P&G and Wal-Mart." E-Business management: Integration of web technologies with business models (2002): 155-171.
6. Boone, Tonya, et al. "Forecasting sales in the supply chain: Consumer analytics in the big data era." International Journal of Forecasting 35.1 (2019): 170-180.
7. Kesavan, Saravanan, Vishal Gaur, and Ananth Raman. "Do inventory and gross margin data improve sales forecasts for US public retailers?." Management Science 56.9 (2010): 1519-1533.
8. Stone, Kenneth E. "Analyzing the impact of Wal-Mart supercenters on local food store sales." American journal of agricultural economics 88.5 (2006): 1296-1303.
9. Lin, Ruonan. "The importance of successful inventory management to enterprises: A case study of Wal-Mart." 2019 International Conference on Management, Finance and Social Sciences Research (MFSSR 2019). London: Francis Academic Press. 2019.
10. Govindan K, Cheng TCE, Mishra N, Shukla N. Big data analytics and application for logistics and supply chain management. Transport Res Part E Logist Transport Rev. 2018;114:343–9. https://doi.org/10.1016/J.TRE.2018.03.011.
11. Bohanec M, Kljajić Borštnar M, Robnik-Šikonja M. Explaining machine learning models in sales predictions. Expert Syst Appl. 2017;71:416–28. https://doi.org/10.1016/J.ESWA.2016.11.010.
12. Ali MM, Babai MZ, Boylan JE, Syntetos AA. Supply chain forecasting when information is not shared. Eur J Oper Res. 2017;260(3):984–94. https://doi.org/10.1016/J.EJOR.2016.11.046.
13. Bian W, Shang J, Zhang J. Two-way information sharing under supply chain competition. Int J Prod Econ. 2016;178:82–94. https://doi.org/10.1016/J.IJPE.2016.04.025.

14. Nguyen T, Zhou L, Spiegler V, Ieromonachou P, Lin Y. Big data analytics in supply chain management: a state-of-the-art literature review. Comput Oper Res. 2018;98:254–64. https://doi.org/10.1016/J.COR.2017.07.004.
15. Huang YY, Handfield RB. Measuring the benefits of erp on supply management maturity model: a “big data” method. Int J Oper Prod Manage. 2015;35(1):2–25. https://doi.org/10.1108/IJOPM-07-2013-0341.
16. Chuang Y-F, Chia S-H, Wong J-Y. Enhancing order-picking efficiency through data mining and assignment approaches. WSEAS Transactions on Business and Economics. 2014;11(1):52–64.
17. Jun S-P, Park D-H, Yeom J. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technol Forecast Soc Chang. 2014;86:237–53. https://doi.org/10.1016/J.TECHFORE.2013.10.021.
18. Varela IR, Tjahjono B. Big data analytics in supply chain management: trends and related research. In: 6th international conference on operations and supply chain management, vol. 1, no. 1, p. 2013–4; 2014. https://doi.org/10.13140/RG.2.1.4935.2563.
19. Agrawal S, Singh RK, Murtaza Q. A literature review and perspectives in reverse logistics. Resour Conserv Recycl. 2015;97:76–92. https://doi.org/10.1016/J.RESCONREC.2015.02.009.
20. Hofmann E, Rutschmann E. Big data analytics and demand forecasting in supply chains: a conceptual analysis. Int J Logist Manage. 2018;29(2):739–66. https://doi.org/10.1108/IJLM-04-2017-0088.