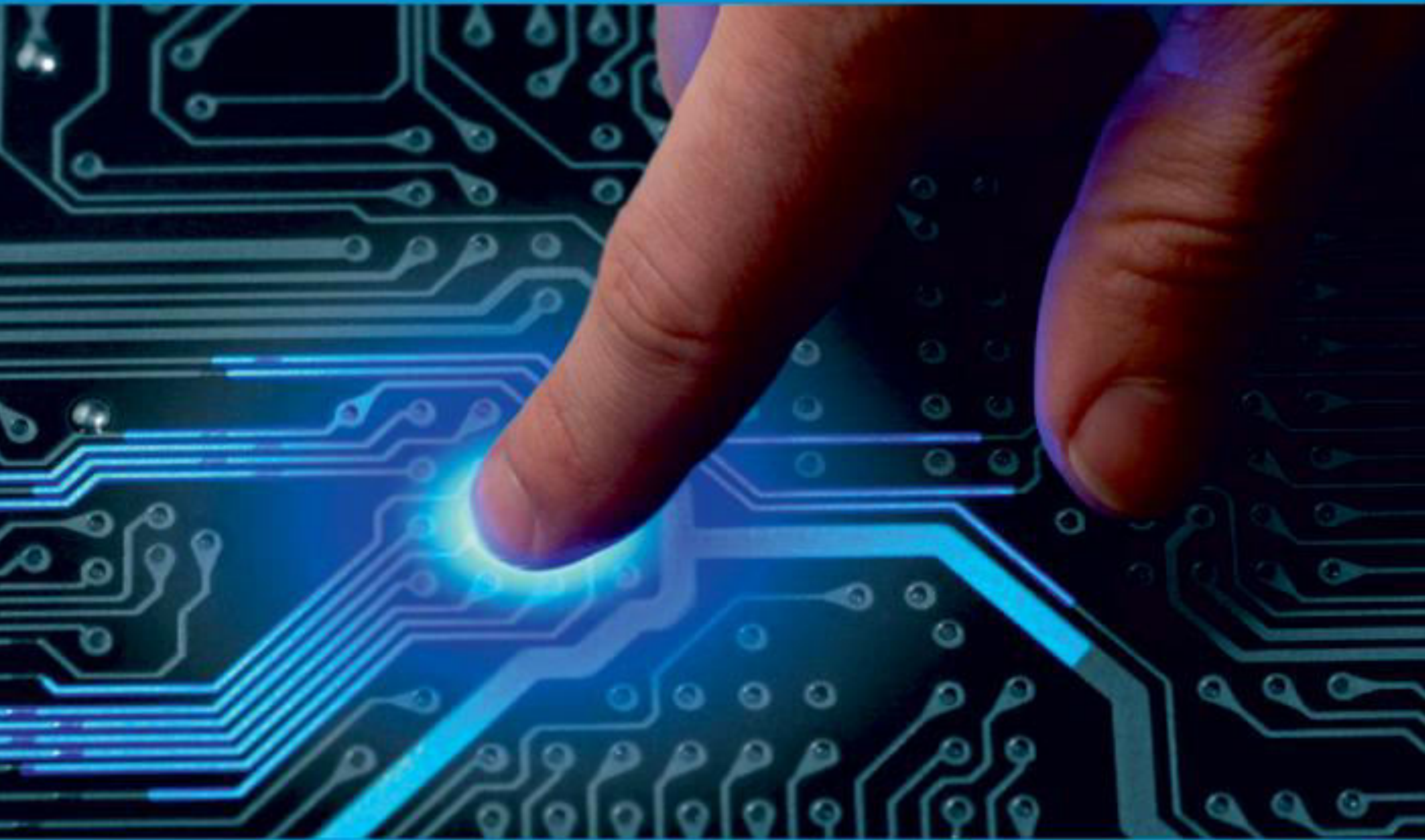
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

Time Series Analysis for Demand Forecasting using Autoregressive Integrated and Gradient Boosting Method

Ms K Gomathi M.E., Srinithi K, Sujitha T, Pradeepa J, Vaishnavi A R

Assistant Professor, Dept. of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India

UG Students, Dept. of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India

ABSTRACT: A positive impact on the environment and the economy can result from improved sales forecasts for specific products at grocery stores. In the past, these predictions have been made using a combination of statistical data and knowledge. However, there has been interest in using machine learning to solve this issue due to the increasing processing power offered by modern computers. Inventory forecasting seeks to anticipate future demand for a certain product and reserve the appropriate quantity of that commodity based on the forecasting outcomes. An optimum goal is to produce and store just enough to meet current needs with certain par reserves. With adequate data, we can forecast consumer behaviour and demands for a specific product or service. With the help of time series analysis, we can recognize patterns and check for any correlation between storage and consumption. It incorporates seasonal fluctuations, consumer habits, and trends. Thus, if modelled adequately, it can predict future behaviour from the correlation observed and thus help find its causation, which will help management teams avoid excess stockings and reduce capital investment on standing assets and reserves.In proposed system, Gradient Boosting technique and statistical S-ARIMA are used, which excels for its prediction speed and accuracy, particularly with huge and complicated datasets, to create more accurate predictions and analysis. This algorithm has delivered the finest outcomes of any machine learning for business solutions. The MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error) deviations for these approaches were the lowest.

KEYWORDS: Gradient boosting; S-ARIMA; XGBoost; LightGBM

I. INTRODUCTION

Inventory Demand Forecasting involves utilising a suitable machine learning model to forecast customer demand for a grocery shop. For analysis, feature engineering, model construction, and performance evaluation, a grocery store's sales from prior months' dataset will be used.The dataset should be utilised in such a way that the project can forecast sales for the upcoming few months, allowing for the future production or purchase of the products. A particular method of examining a dataset gathered over a period of time is called a "time series analysis." Instead of just capturing the datasets intermittently or arbitrarily, time series analysts record the data points at regular intervals throughout a predetermined timeframe. But this kind of study involves more than just gathering data over time. To maintain consistency and dependability, time-series analysis often needs a lot of data. A large data collection guarantees that your analysis can sort through noisy data and that your sample size is appropriate. Additionally, it guarantees that any trends or patterns are not outliers and can take seasonal variation into consideration. Time-series data can also be utilised for forecasting, which is the process of making predictions about the future based on the past. Non-stationary data, or items that change over time or are impacted by time, are studied using time series analysis. Time series analysis is commonly used in sectors like finance, retail, and economics because currency and sales are always fluctuating.Time series analysis models include:

- Classification: The process of identifying and classifying the data.
- Curve fitting: Studies the correlations between variables within the data by plotting the data along a curve.
- Descriptive analysis: Finds trends, cycles, or seasonal variations in time series data.
- Explanatory analysis: Makes an effort to understand the data, its relationships, as well as cause and effect.
- Exploratory analysis: Draws attention to the key traits of the time series data, typically in a visual style.
- Forecasting: Estimates data for the future. This kind is built on earlier patterns. It forecasts possible outcomes along with possible plot points using previous data as a model for future data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

- Intervention analysis investigates how a condition can alter the information.
- Segmentation: Divides the data into sections to reveal the underlying characteristics of the information source

II. RELATED WORK

A wide range of issues must be addressed under the inventory management approach. The necessity for effective inventory management is essential given the e-commerce sector's massive growth. To improve the current inventory management strategies, more study is needed. Using Apache Spark-based probabilistic demand forecasting models, e-commerce giants define an end-to-end machine learning system. Large datasets are present in such e-commerce giants.

A. *Algorithm Used In Previous Work*:

Numerous methods have been developed to forecast sales, with neural networks and auto-regression being the most well-known. This is understandable given that the issue was essentially a time series issue. In the academic literature, Holt-winters forecasting method and Long Short-Term Memory (LSTM) have generated a lot of discussion and promising results. Regression models like Lasso, Support Vector Regression (SVR), and Random Forest have also demonstrated promising results, suggesting that this strategy could produce noteworthy outcomes. Regression analysis and neural networks were employed as the prediction models when building the model. The Adadelta method is used to set the learning rate in this model to be adaptable. The performance and validity of the result were examined using the 5-fold cross-validation. Traditional prediction error measurements like Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) between the expected sales amount and actual sales amount are used as performance metrics.

III. PROPOSED ALGORITHM

Machine learning not only increases the accuracy of demand forecasts, but it also automates large amounts of planner work and can process enormous data sets far more than any human planner would be capable of. To generate an accurate demand forecast, a system must be able to process an enormous amount of data on the wide range of variables that can potentially impact demand. With advancements in large-scale data processing and in-memory computing, modern demand planning systems can make millions of forecast calculations within a minute, taking into consideration more variables than ever before possible. Consider the three broad areas of variability that continuously impact demand: recurring variations in baseline demand patterns, your own internal business decisions, and external factors such as weather or local events. To predict the impact of business decisions, you must leverage machine learning algorithms that can process large amounts of retail data and integrate them into the baseline demand forecast to be accounted for.

A. *Problem Definition*:

To predict the sales of multiple grocery stores from historical data and forecast the future demands.

B .*Methodologies of Problem Solving*:

1. Analysis of the chosen data which consists of five-year sales data.
2. Applying a machine-learning algorithm to build a model capable of predicting demands and evaluating the accuracy of the model built.
3. Allowing the user to choose the suitable model according to their requirements

C. *Proposed Methodologies*

Seasonal Arima:

ARIMA is the abbreviation for Autoregressive Integrated Moving Average.Auto Regressive (AR) terms refer to the lags of the differenced series, Moving Average (MA) terms refer to the lags of errors and I is the number of difference used to make the time series stationary.

Assumptions of ARIMA model:

1. Data should be stationary – by stationary it means that the properties of theseries don't depend on the time when it is captured. A white noise series and series with cyclic behaviour can also be considered as stationary series.

2. Data should be univariate – ARIMA works on a single variable. Auto-regression is all about regression with the past values.

D. *Steps to be followed for ARIMA modeling*:

Exploratory analysis: Autocorrelation analysis to examine serial dependence: Used to estimate which value in the past has a correlation with the current value. Provides the p,d,q estimate for ARIMA models. Spectral

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

analysis: It is to examine cyclic behaviour: Carried out to describe how variation in a time series may be accounted for by cyclic components. Also referred to as a Frequency Domain analysis. Using this, periodic components in a noisy environment can be separated out. Trend estimation and decomposition: Used for seasonal adjustment. It seeks to construct, from an observed time series, a number of component series (that could be used to reconstruct the original series) where each of these has a certain characteristic. Before performing any EDA on the data, we need to understand the three components of a time series data: Trend: A long-term increase or decrease in the data is referred to as a trend. It is not necessarily linear. It is the underlying pattern in the data over time. Seasonal: When a series is influenced by seasonal factors i.e. quarter of the year, month or days of a week seasonality exists in the series. It is always of affixed and known period. E.g. – A sudden rise in sales during Christmas, etc. Cyclic: When data exhibit rises and falls that are not of the fixed period we call it a cyclic pattern.

E. *Boosting*:

In machine learning, boosting is an ensemble meta-algorithm for primarily reducing bias and variance in supervised learning. The principle behind the boosting algorithm is first we built a model on the training dataset, then the second model is built to rectify the errors present in the first model.

Gradient Boosting Algorithm:
This algorithm is mainly used to build models sequentially where these subsequent models try to reduce the errors of the previously built models. This can be done by building the new model on the errors of the previous model. When the target column is continuous, we use a Gradient boosting regressor.

Light Gradient Boosting Machine:
LightGBM is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. By including a kind of automatic feature selection and concentrating on boosting examples with greater gradients, LightGBM expands the gradient boosting technique. This can result in a dramatic speedup of training and improved predictive performance.LightGBM has become an algorithm for machine learning competitions when working with tabular data for regression and classification predictive modelling tasks.As a result, together with Extreme Gradient Boosting, it shares some of the responsibility for the rising popularity and widespread use of gradient boosting techniques in general. LightGBM, Microsoft first created the distributed gradient boosting framework for machine learning known as Light Gradient Boosting Machine(LightGBM), which is free and open-source. Its initial release was in the year 2016 since then it has been proved to be a successful model for forecasting in data science competitions and recent literature.LightGBM has many boost advantages, including sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. It has been actively used in online communities with promising results. A class of ensemble machine learning methods known as gradient boosting can be applied to classification or regression predictive modelling issues.Ensembles are constructed from decision tree models.To correct the prediction mistakes caused by earlier models, trees are added one at a time to the ensemble and fitted. Boosting is a term used to describe this kind of ensemble machine learning model. Any arbitrary differentiable loss function and the gradient descent optimization procedure are used to fit the models. This gives the technique its name, “gradient boosting” as the loss gradient is minimized as the model is fit, much like a neural network.Light Gradient Boosted Machine, or LightGBM for short is a free and open-source gradient boosting implementation that aims to be as effective as possible. LGBM is based on decision tree algorithms and is used for ranking, classifications, and other machine learning tasks. The development focus is on performance and scalability.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

F. *System Architecture:*

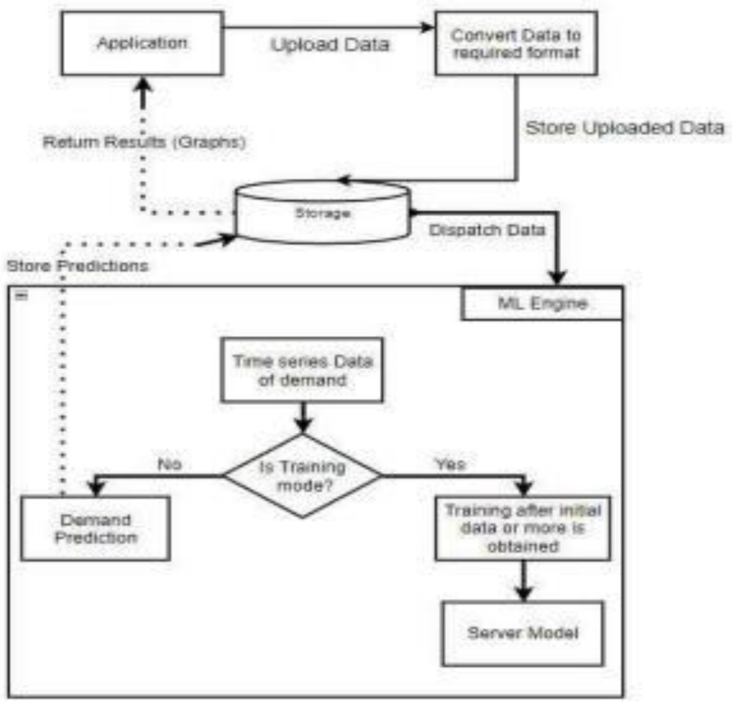


Fig_1_Architecture diagram

G. *Advantages of Proposed Methodologies*:

Light GBM is a histogram-based approach for quicker training and more efficiency. It discretely classifies continuous feature data into buckets to speed up the training process.
Lower memory usage: Replaces continuous values with discrete bins which results in lower memory usage. Better accuracy than any other boosting algorithm: It produces much more complex trees by following leaf The key to getting improved accuracy is to use a wise split technique rather of a level-wise strategy. However, it can sometimes lead to fitting which can be avoided by setting the max_depth_parameter. Compatibility with Large Datasets: It is capable of performing equally well with large datasets with a significant reduction in training time as compared to other algorithms. Support of Parallel, Distributed, and GPU learning.

IV. RESULTS AND DISCUSSION

To evaluate and compare the different models fairly, the choice of evaluation metrics was important as each metric has different characteristics. It was also important to include several metrics since different metrics could display different flaws or benefits in the models. Root mean squared error (RMSE), mean absolute error (MAE), and Mean Absolute Percentage Error (MAPE) have been used extensively in the academic literature and could, therefore, be deemed to be the most useful. In comparison with these, when analysing the online communities, data science competitions, and sources outside the academic literature, it was clear that Symmetric Mean Absolute Percentage Error (SMAPE) can be beneficial when comparing the models. By utilizing multiple performance metrics, with different characteristics, as specified above, the chances of locating the best algorithm for a specific outcome are increased.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

Metrics	Values
SMAPE	14.5199
MAE	7.4117
RMSE	10.0063

Table -1-Metrics Comparison_lgbm

Model Name	Execution Time
LightGBM	0 days 00:06:20.552760
XGBOOST	0 days 00:22:26.235582

Table -2-Execution time_lgbm

S-ARIMA:

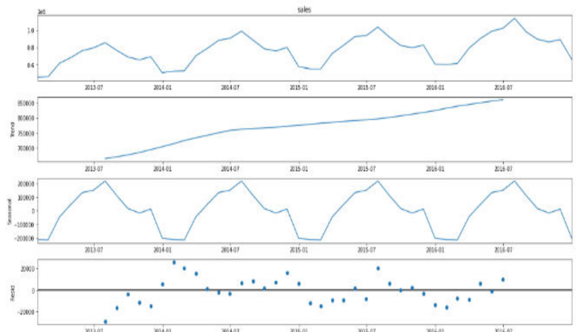


Fig_2_Results analysis_S-Arima

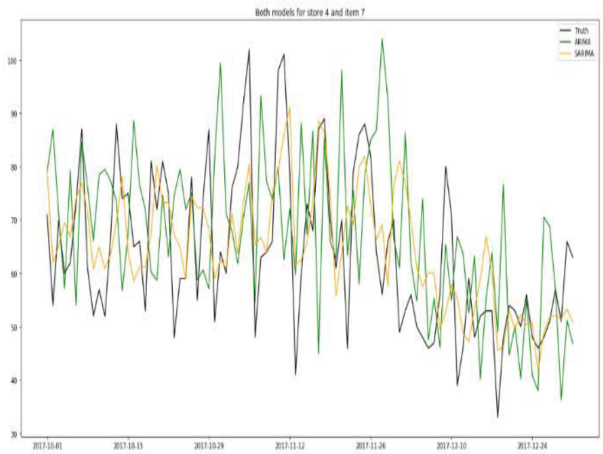


Fig_3_Output_S-Arima

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

Light GBM:

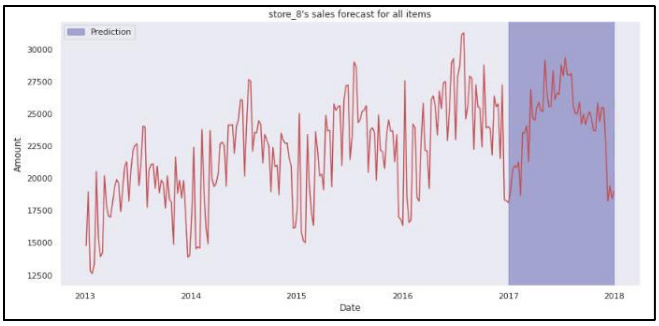


Fig_4_Output 1-lgbm

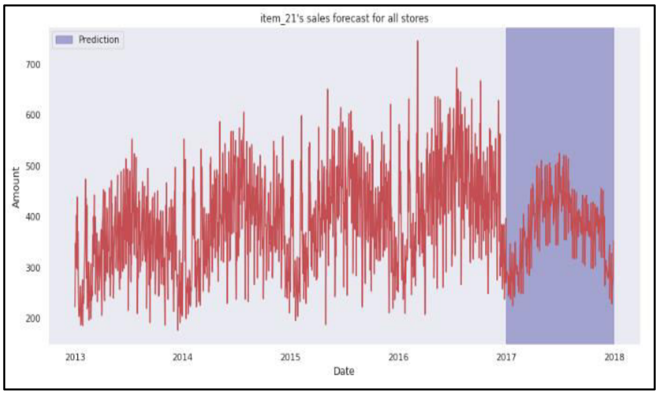


Fig_5_Output 2-lgbm

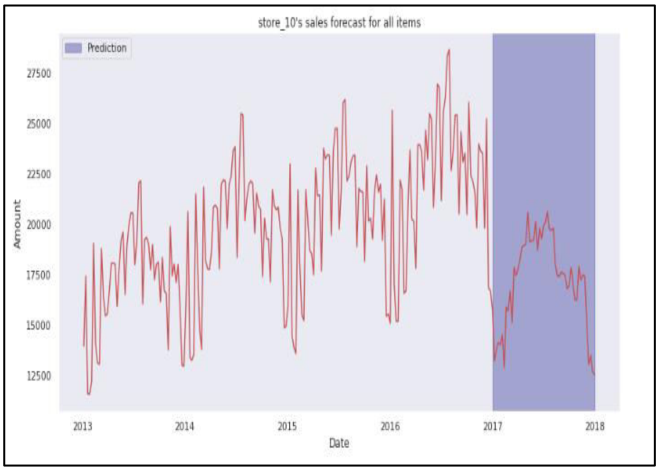


Fig_6_Output 3-lgbm

V. CONCLUSION

The LGBM regressor method can produce accurate output for the given dataset, based on the above-mentioned considerations. For desired items and regions of operation, the system can be successfully used in a variety of corporate sectors for demand forecasting, inventory optimization, and budget optimization. For producing predictions, no machine learning system can be considered perfect. As a result, predictions are made based on a time-series analysis of a dataset of product sales and inventory details using machine-learning algorithms such as S-Arima, XGBoost and LightGBM. For the model employed during the work, the system had an average accuracy of 92.5%.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006107 |

VI. FUTURE WORK

The system can be scaled to meet the needs of multinational organizations and customized to the preferences of the company's key decision makers. The technology can also assist in the examination of user sentiments toward various items at various times in time, based on sales numbers. A web application which allows users to upload their own dataset and to see the demand forecasting results of that dataset can be created.The system can also be expanded to generate an appropriate geographical location, such as a warehouse, for keeping additional products in real-time to reduce the company's transportation expenses.

REFERENCES

1. Mehdi Khashei and Mehdi Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, Applied Soft Computing, Volume 11, Issue 2, 2011
2. L. Wu, J. Y. Yan, and Y. J. Fan, ”Data Mining Algorithms and Statistical Analysis for Sales Data Forecast,” 2012 Fifth International JointConference on Computational Sciences and Optimization, Harbin, 2012
3. T. Choi, C. Hui, and Y. Yu, ”Intelligent time series fast forecasting for fashion sales: A research agenda,” 2011 International Conference on Machine Learning and Cybernetics, Guilin, 2011, pp. 1010-1014.
4. A.D. Lacasandile, J. D. Niguidula, and J. M. Caballero, ”Mining the past to determine the future market: Sales forecasting usingTSDM framework,” TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, 2017, pp. 461- 466.
5. P. Sobreiro, D. Martinho and A. Pratas, ”Sales forecast in an IT company using time series,” 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 2018, pp. 1-5. DOI: 10.23919/CISTI.2018.8399191
6. G. Nunnari and V. Nunnari, ”Forecasting Monthly Sales Retail Time Series: A Case Study,” 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, 2017, pp. 1-6.
7. B. Siregar, E. B. Nababan, A. Yap, U. Andayani and Fahmi, ”Forecasting of raw material needed for plastic products based inincome data using ARIMA method,” 2017 5th International Conference on Electrical, Electronics and Information Engineering(IEEE), Malang, 2017, pp. 135- 139.
8. F. M. Thiesing, U. Middelberg, and O. Vornberger, ”Short term prediction of sales in supermarkets,” Proceedings of ICNN’95- International Conference on Neural Networks, Perth, WA, Australia, 1995, pp. 10281031 vol.2.
9. C. Chatfield, ”Time series forecasting with neural networks,” Neural Networks for Signal Processing VIII. Proceedings of the1998 IEEE Signal Processing Society Workshop (Cat. No.98TH8378), Cambridge, 1998, pp. 419-427.
10. D. H. F. Yip, E. L. Hines, and W. W. H. Yu, ”Application of artificial neural networks in sales forecasting,” Proceedings of International Demand Forecasting For Inventory Optimization
11. M. Gurnani, Y. Kirke, P. Shah, S. Udmale, V. Sambhe, and S. Bhirud, ”Forecasting of sales by using a fusion of machine learning techniques,” 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), Pune, 2017, pp. 93- 101.
12. AiyunZheng, Weimin Liu, and Fanggeng Zhao, ”Double trends time series forecasting using a combined ARIMA and GMDHmodel,” 2010 Chinese Control and Decision Conference, Xuzhou, 2010, pp. 18201824
13. Jaiswal R. C. and Danish khan, “Arduino based Weather Monitoring and Forecasting System using SARIMA Time-Series Forecasting”, Journal of Emerging Technologies and Innovative Research (JETIR), Open Access, Peer Reviewed and refereedJournal, ISSN-2349- 5162, Impact Factor:5.87, Volume 7, Issue 11, pp. 1149-1154, November 2020
14. Jaiswal R.C. and Lokhande S.D, “A Novel Approach for Real Time Internet Traffic Classification”, ICTACT Journal on Communication Technology, September 2015, volume: 06, issue: 03, pp. 1160-1166.(Print: ISSN: 0976- 0091, Online ISSN:2229-6948 (Impact Factor: 0.789 in 2015).
15. Jaiswal R.C. and Lokhande S.D, “Statistical Features Processing Based Real Time Internet Traffic Recognition and Comparative Study of Six Machine Learning Techniques”, IMCIP- International Multi Conference on Information Processing-(ICCN- International Conference on Communication Networks-2014, PP-120-129, ISBN: 9789351072515, University Visvesvaraya College of Engg. Department of Computer Science and Engineering Bangalore University, Bangalore.
16. AiyunZheng, Weimin Liu and Fanggeng Zhao, ”Double trends time series forecasting using a combined ARIMA and GMDHmodel,” 2010 Chinese Control and Decision Conference, Xuzhou, 2010, pp. 18201824.
17. https://machinelearningmastery.com/time-seriesforecasting-long-short-term-memory-network-python/

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details