IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 12, December 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

Prediction of Backorder using Random Forest, Gradient Boosting Machine and Light Gradient Boosting Machine Techniques

R. Varadharajan [1], N. Marudachalam[2]

M.Phil., Student, Department of Computer Science, Presidency College, Chennai, Tamil Nadu, India[1]

Associate Professor, Department of Computer Science, Presidency College, Chennai, Tamil Nadu, India[2]

ABSTRACT: Supply Chain Management (SCM) refers to the management of the flow of products and services, including the transformation of raw materials and semi-finished products into finished goods. The Supply Chain Manager oversees all logistics aspects of the Supply Chain Management process. Backorder is a common problem in Supply Chain Management. Backorders can negatively impact Customer satisfaction, Customer trust, company revenue, and customer loyalty. Predicting the backorder with high accuracy allows us to take corrective action proactively to improve the Company's performance. The Supply Chain parameters can be analysed using Machine Learning algorithms to predict and identify likely backorder products. Light Gradient Boosting Machine (LightGBM), Random Forest (RF) and Gradient Boosting Machine (GBM) was used in this study to predict backordered products. Light GBM has the advantages of being able to train rapidly, with minimal memory usage, supports parallel learning, and can handle large amounts of data. Our analysis identifies potential backorder products before actual sales take place and compares the results of different algorithms. The LightGBM accelerates model construction by approximately 3 seconds. Random Forest takes about 12 seconds, and Gradient Boosting Machine takes about 70 seconds. In a short period of time, the LightGBM is also able to achieve almost similar accuracy. Using this algorithm and methods, supply chain management can predict backorders.

KEYWORDS: Supply chain, Prediction, Backorder, Machine Learning, LightGBM, Random Forest, Gradient Boosting Machine.

I. INTRODUCTION

In any given situation, if orders cannot be filled, they are called backorders. The backorders are caused by a shortage of supplies. The item may not be in the company's inventory at the moment, but may be in production, or the company may need to produce more. In the analysis of inventory management, backorders play a key role. Depending on where products were backordered and how long it took to fulfill these orders, it can be useful to evaluate how well the organization manages its inventory.

When the order volume and turnaround time are relatively manageable, it is generally an indication that an organization is doing well. However, longer wait times and large backorders can be a problem. The most common cause of backorders is unusual demand. As customer demands are unpredictable, the usual supply chain structure is less useful for forecasting demand, resulting in a backlog of goods. For a high probability of improved company performance, the industries should identify the parts with a high probability of deficiency before they occur to avoid backorders and shortages in supply chains [3].

The prediction of backorder products is provided with the help of RF, GBM and LightGBM algorithms. The study focused mainly on the accuracy of the backorder prediction with less training time and low memory usage for larger datasets.

The following is an overview of the paper. Introduction is provided in Section I, followed by descriptions of related research in Section II, datasets and exploratory analysis in Section III, and hypothesis testing in Section IV. The methodology and tools are provided in Section V, the results are given in Section VI, and the future directions are given in Section VII.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

II. RELATED WORK

Support Vector Regression (SVR) can be used to forecast demand in the supply chain industry [1]. Due to its emphasis on minimizing structural risks, it performs extremely well in the supply chain. As a result of support vector regression methods, SCM requirements are more accurate and less proportional to mean square error (MSE).

Author	Domain	ML Model	Backorder scenario?	Main Factors?	Relational Factors
Wang Guanghui	Forecast the demand	Support Vector Regression	No	-	-
C. Aguilar-Palacios, S. Muñoz-Romero and J. L. Rojo-Álvarez	Forecast the promotional sales	K-Nearest Neighbors	No	-	-
Samiul Islam & Saman Hassanzadeh Amin	Prediction of probable backorder	Distributed Random Forest, Distributed Gradient Boosting Machine	Yes	Yes	No
S. Luo and T. Chen	Prediction of Silicon Content in Blast Furnace system	eXtreme Gradient Boosting Light Gradient Boosting machine	No	-	-
F. Alzamzami, M. Hoda and A. E. Saddik	Sentiment classification on Short texts	Light Gradient Boosting Machine	No	-	-
Borg, A., Boldt, M., Rosander, O.	E-mail classification with machine learning and word embeddings for improved customer support	Long Short-Term Memory	No	-	-
Chen, RC., Dewi, C., Huang,SW.	Data classification using machine learning methods based on the selection of critical features	Random Forest Support Vector Machine K-Nearest Neighbours Linear Discriminant Analysis	No	-	-
Proposed work	Backorder Prediction using LightGBM	LGBM	Yes	Yes	Yes

Planned and forecasted promotions are an important way retailers engage consumers [2]. This algorithm is used to predict promotional sales within a neighborhood using K-nearest neighbors. Calculations are based on the similarity of promotional sales, which is the basis of the feature selection process. The data has been categorized into three categories, namely training datasets, validation datasets, and testing datasets.

Machine learning models help businesses make better business decisions. When calibrated properly, Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM) algorithms are more effective at predicting backorder items. Researchers predicted the probable backorder by analyzing inventory, lead time, sales, and forecasted sales data. These algorithms can be used to analyze any supply chain situation.

Light Gradient Boosting Machine (LightGBM) is a derivative algorithm of Gradient Boosting Decision Tree (GBDT), which was developed by Microsoft in 2017. The prediction status of a blast furnace is determined more

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

effectively with boosting algorithms than with traditional methods, including Lasso, Random Forest, and Support Vector Machine (SVM). LightGBM has the advantage of high training speed and low memory usage when compared to conventional algorithms [4].

To determine the best split candidates, Light Gradient Boosting Machine (LightGBM) uses a histogram-based method. To indicate which data instances are most important, LightGBM uses a sampling algorithm called Gradient-based One Side Sampling (GOSS). Furthermore, LightGBM includes an Exclusive Feature Bundling (EFB)-based sparsity handling algorithm [5].

The e-mails received from customers are categorized using machine learning. The categories are easy to select, which helps the user make a quick decision. In the telecom company's case, the Long Short-Term Memory (LSTM) network helped with classification of e-mails with a F1 score of 0.91 [6].

Based on their importance scores, the feature selection identifies the most important and least important variables in the dataset. Once the unimportant variables are eliminated, the training time will automatically decrease. With Random Forest, there is a better chance of selecting features [7]. Our study is summarized in Table I along with a few other research papers.

III. DATASET AND EXPLORATORY ANALYSIS

This research used data from Kaggle, one of the largest data science communities. There are two different parts to the data set, namely training and testing. Among the attributes are integers, decimal numbers and strings.

There are 1687861 records in the training dataset, and 242075 records in the testing dataset. The dataset has 23 attributes in which the stock keeping unit (SKU) contains the unique value, so the column was dropped. In addition to inventory, lead time, sales, forecast, minimum stock, and local buyout quantity, our predicting variables and our target variables are lead time, sales, forecast, and went on backorder.

Therefore, our study belongs to the machine learning binary classification problem since the target variable is labelled with two classes, yes or no. Our study indicates yes and no as 1 and 0. Since 0.7% of the overall dataset has the "Yes" class and 99.3% has the "No" class, the target variable is imbalanced. Figs. 1, 2, and 3 illustrate the data distribution.

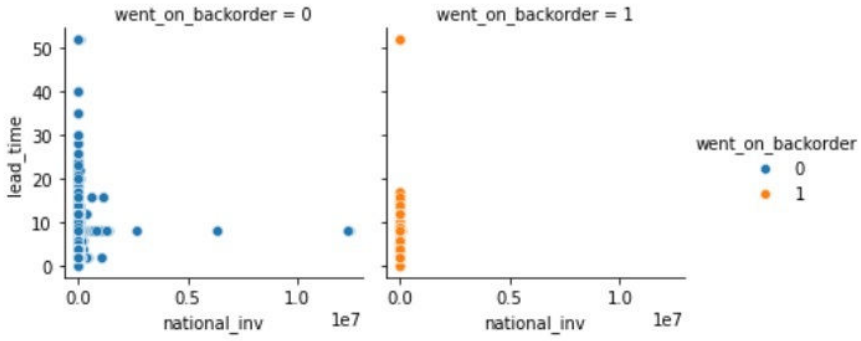

Fig. 1 Inventory and lead time vs went to backorder

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

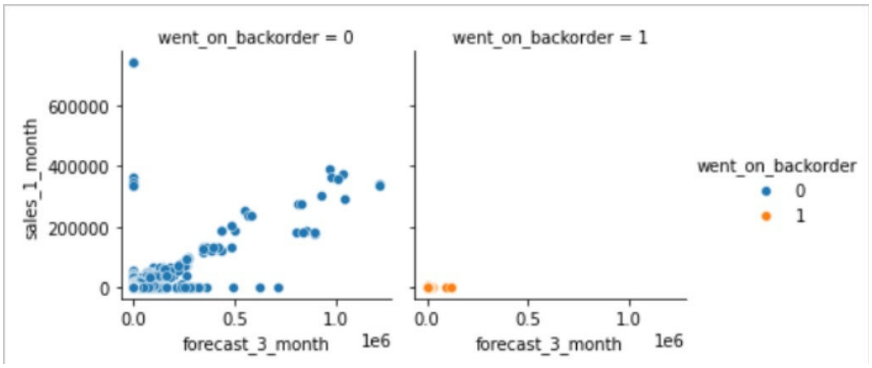

Fig. 2 Forecast and sales vs went to backorder

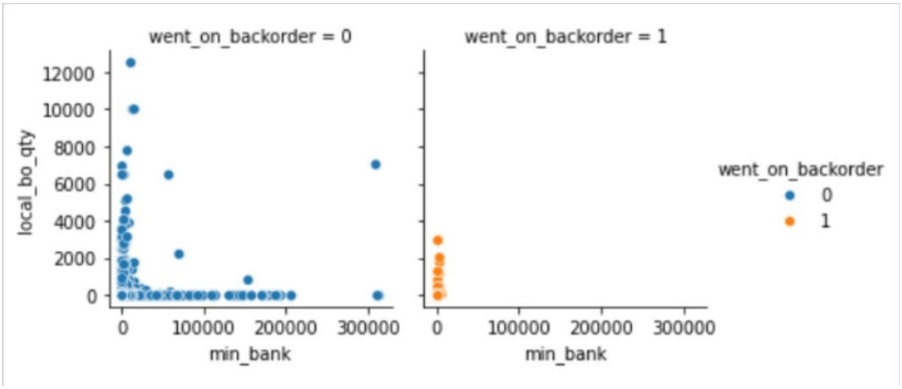

Fig. 3 Minimum bank and local buyout quantity vs went to backorder

Inventory indicates the amount of product that is available, and it has both positive and negative numbers. Leading time attributes are associated with the time elapsed between placing a product order and it being delivered to the customer. Our study indicates that lead times can range between zero and 52 weeks. The sales feature displays the sales quantity. A forecast shows the anticipated sales. There is a minimum bank which indicates the recommended minimum quantity of materials. Overdue stock orders are indicated by the local buyout quantity feature.

IV. HYPOTHESIS TESTING

Hypothesis testing is generally used to validate a relationship between two or more variables. It helps to determine whether samples provide strong evidence. A one-sample t-test was used as the hypothesis test in our study. As part of our investigation, we would like to determine whether some predicting variables of our dataset are related to the variables on backorder.

The null hypothesis refers to the fact that two variables do not correlate or one variable is not affected by another. Alternative hypotheses are those which are viable when the null hypothesis is not proven correct. A relationship exists between two variables when one variable affects the other. As a probability value, a p-value represents how likely it was that the significance occurred by accident. When the Alpha level is less than 5% (0.05), it strongly implies that the null hypothesis cannot be accepted because it has a less than 5% probability of being true.

At the first stage, we are interested in seeing whether the predicting variable current inventory has an effect on the target variable or not. When the inventory of the product reaches zero, the product goes into backorder. It is noted that the p-value is less than the significance level of 0.05, which tells us that this null hypothesis is not significant and the null hypothesis is false in this case.

The objective of the second scenario is to determine if the lead time has any impact on the target variable. In the absence of lead time data, assuming the null hypothesis as products went backordered. Null hypothesis does not have statistical significance because the p-value falls below the significance level. This means the null hypothesis is false.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

For the third scenario, we would like to see if the forecast has any effect on the target variable. Assumed the null hypothesis is that when forecasts are too high, products go on backorder. Null hypothesis has no statistical significance since p-value is below significance level. Thus, the null hypothesis is false.

We would like to determine whether actual sales affect the target variable in the fourth scenario. The null hypothesis was assumed, as products would be backordered if the number of products sold per month was higher than average. This null hypothesis does not have statistical significance due to a p-value below the significance level. So, the null hypothesis cannot be true.

The fifth scenario looks at whether the minimum stock impacts the target variable. When the minimum stock maintained for the products is less than the average, the null hypothesis holds that the products will go to backorder. In this case, the p-value is below the significance level for the null hypothesis. In that case, the null hypothesis cannot be true. We would like to know whether the local buyout quantity affects the target variable in the last scenario. If the product has the local buyout quantity, we assumed the null hypothesis because the products went to backorder. As the p-value is below the significance level, this null hypothesis is not statistically significant. The null hypothesis is therefore false.The summary of the hypothesis test results is available in below table:

Question	Hypothesis	Result Analysis	Decision
i) Are the products' inventory resulting backorders when it reaches 0?	H_0: $Backorder_{yes}$ = $Inventory_{very\ low}$	p-value = 1.62e-176, for α=0.05	The alternative hypothesis is true
ii) Are the products' lead time resulting backorders when it is less than a week?	H_0: $Backorder_{yes}$ = $Lead\ time_{min}$.	p-value = 0.01, for α=0.05	The alternative hypothesis is true
iii) Are the products' forecast resulting backorders when it is very high?	H_0: $Backorder_{yes}$ = $Forecast\ quantiry_{very\ high}$	p-value = 5.05e-12, for α=0.05	The alternative hypothesis is true
iv) Are the actual sales resulting backorders when it is greater than the average?	H_0: n > average sales; where n = actual sales	p-value = 0.0, for α=0.05	The alternative hypothesis is true
v) Are the minimum stock resulting backorders when it is maintained less than the overall average?	H_0: n < average minimum stock; where n = minimum stock	p-value = 7.69e-15, for α=0.05	The alternative hypothesis is true
vi) Are the local buyout quantity producing backorders?	H_0: n > 0; where n = local buyout quantity	p-value = 7.36e-28, for α=0.05	The alternative hypothesis is true

V. METHODOLOGY

Tools

Models based on machine learning are constructed using the Google Colaboratory. It is also known as Colab, which allows you to write Python code in a browser without requiring any configuration. In order to conduct research analysis and visualize data, we can harness the full power of Python libraries.

Machine Learning Model Construction

We constructed the Machine Learning model using the LightGBM algorithm in this study, which can predict the likely scenarios of backorder. As compared with Random Forest or Gradient Boosting Machine, LightGBM is the newest, and it provides several advantages such as faster training, better accuracy, support for parallel learning, and can process large datasets. Meanwhile, LightGBM's trees grow vertically while other tree-based algorithms grow horizontally.

Due to the increasing size of data, it is becoming more and more difficult for the existing algorithms to give faster results. Due to its ability to handle large amounts of data and its dedication to accuracy, LightGBM is so popular. According to the LightGBM algorithm, inventory is the most important category, followed by sales, minimum banks, forecasts, lead times, and local buyout quantities, as shown in Figure 11.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

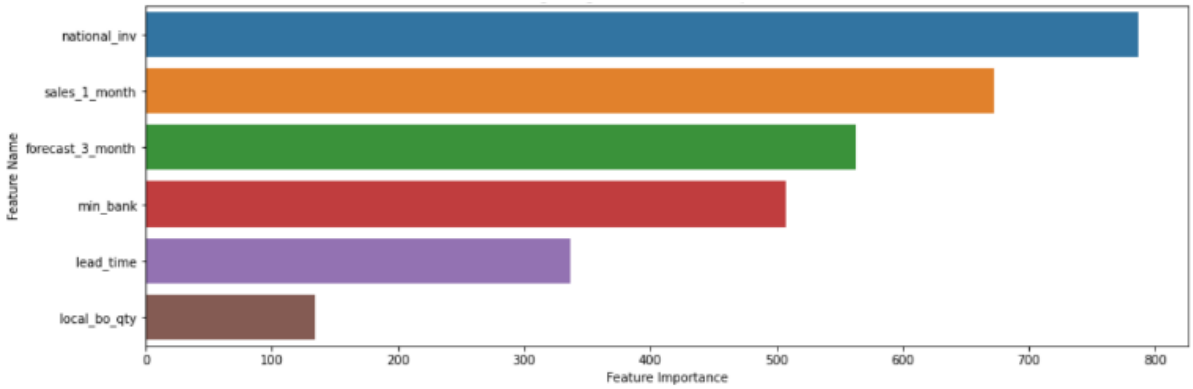

Fig. 12 LightGBM - Feature Importance

The LightGBM algorithm takes very less time to construct the machine learning model compared to Random Forest and Gradient Boosting Machine while achieving similar accuracy.

Model Evaluation

As the evaluation metrics, in this study considered the Model time, Accuracy score and the F1 weighted score.

Accuracy:

A thumb rule is that accuracy gives us insight into how the model is trained and how it performs on the validation dataset. Comparison with the training phase shows a slight decrease in the accuracy and F1 scores. In the context of a machine learning algorithm, such a slight variation is tolerable. The live data may contain different features, null values, and missing values in a production environment. The data scientist should be consulted for data cleaning and feature engineering in this case.

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

Precision:

Precision is how much we predicted correctly out of all the classes and the precision should be as high as possible. Precision is the ability to a classifier to not label a true negative observation as positive.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Recall:

Out of all the positive classes, recall is how much we predicted correctly. It is also called sensitivity or true positive rate (TPR). Recall is the real positives. Recall also should be as high as possible.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

F-1 Score:

F1 is the measure which combines the precision and recall scores. The F1 score is good when we have low false positive and low false negative. An F1 score is called very good for the model when it is 1 and complete failure if it is 0.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

VI. EXPERIMENTAL RESULTS

To analyse the data in the testing dataset, we used the model we had constructed during model construction or it can be called as training phase. We also calculated the accuracy, F1 score, precision and recall scores for the training and testing phases, and the results are listed below.

Performance metrics: Training phase

ML Algorithm	Model Time (sec)	Accuracy Score	F1 Score	Precision Score	Recall Score
LightGBM	3	99.16	0.99	0.99	0.99
Random Forest	12	99.04	0.99	0.99	0.99
GBM	70	99.02	0.99	0.99	0.99

Performance metrics: Testing phase

ML Algorithm	Accuracy Score	F1 Score	Precision Score	Recall Score
LightGBM	98.68	0.98	0.98	0.99
Random Forest	98.38	0.98	0.98	0.99
GBM	98.18	0.98	0.98	0.99

VII. CONCLUSION

Identifying backorder products improves the performance and profitability of the organization. A LightGBM algorithm was used to identify backorder products in this study. The use of live data can pose some challenges such as bias and missing values. We found that LightGBM's construction time was much shorter than traditional methods like Random Forests and Gradient Boosting Machines, as well as its accuracy was similar.

Backorder probabilities have been determined by analyzing inventory, sales, forecasts, lead times, as well as relational factors such as minimum stock levels and local buyout quantities. Demand is one factor that impacts backorders. Feedback from customers and user experiences also play a role. In the future, research could focus on these abilities that can help supply chain managers avoid backorders, thus improving the efficiency of the supply chain.

REFERENCES

[1] Wang Guanghui “Demand Forecasting of Supply Chain Based on Support Vector Regression Method”, Procedia Engineering, Volume 29, Pages 280-284, 2012
[2] C. Aguilar-Palacios, S. Muñoz-Romero, and J. L. Rojo-Álvarez “Forecasting Promotional Sales Within the Neighbourhood”, IEEE, Volume 7, Pages 74759-74775, 2019
[3] Samiul Islam & Saman Hassanzadeh Amin “Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques”, Springer, Volume 7, Article 65, 2020
[4] S. Luo and T. Chen, "Two Derivative Algorithms of Gradient Boosting Decision Tree for Silicon Content in Blast Furnace System Prediction” IEEE, Volume 8, Pages 196112-196122, 2020
[5] F. Alzamzami, M. Hoda and A. E. Saddik, "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation", IEEE, Volume 8, Pages 101840-101858, 2020
[6] Borg, A., Boldt, M., Rosander, O “E-mail classification with machine learning and word embeddings for improved customer support” Springer - Neural Computing and Applications, Volume 33, Pages 1881–1902, 2021.
[7] Chen, RC., Dewi, C., Huang, SW. “Selecting critical features for data classification based on machine learning methods” Springer – Journal of Big Data, Volume 7, 2020
[8] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches", IEEE, Volume 8, Pages: 17722-17733, 2020
[9] Ethem Alpaydin “Introduction to Machine Learning”, 3rd edition, PHI Learning Private Limited, Delhi – 2016

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 12, December 2021 ||

| DOI:10.15680/IJIRSET.2021.1012074|

[10] J. Lee, W. Wang, F. Harrou and Y. Sun, "Wind Power Prediction Using Ensemble Learning-Based Models", IEEE, Volume 8, Pages: 61517-61527, 2020
[11] M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions", IEEE, Volume. 7, Pages: 137184-137206, 2019
[12] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques", IEEE, Volume 8, Pages: 67899-67911, 2020
[13] S. Kim, S. Ku, W. Chang and J. W. Song, "Predicting the Direction of US Stock Prices Using Effective Transfer Entropy and Machine Learning Techniques" , IEEE, Volume 8, Pages: 111660-111682, 2020
[14] Sarker, I.H., Kayes, A.S.M. & Watters, P. "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage", Springer Volume 6, Article 57, 2019

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details