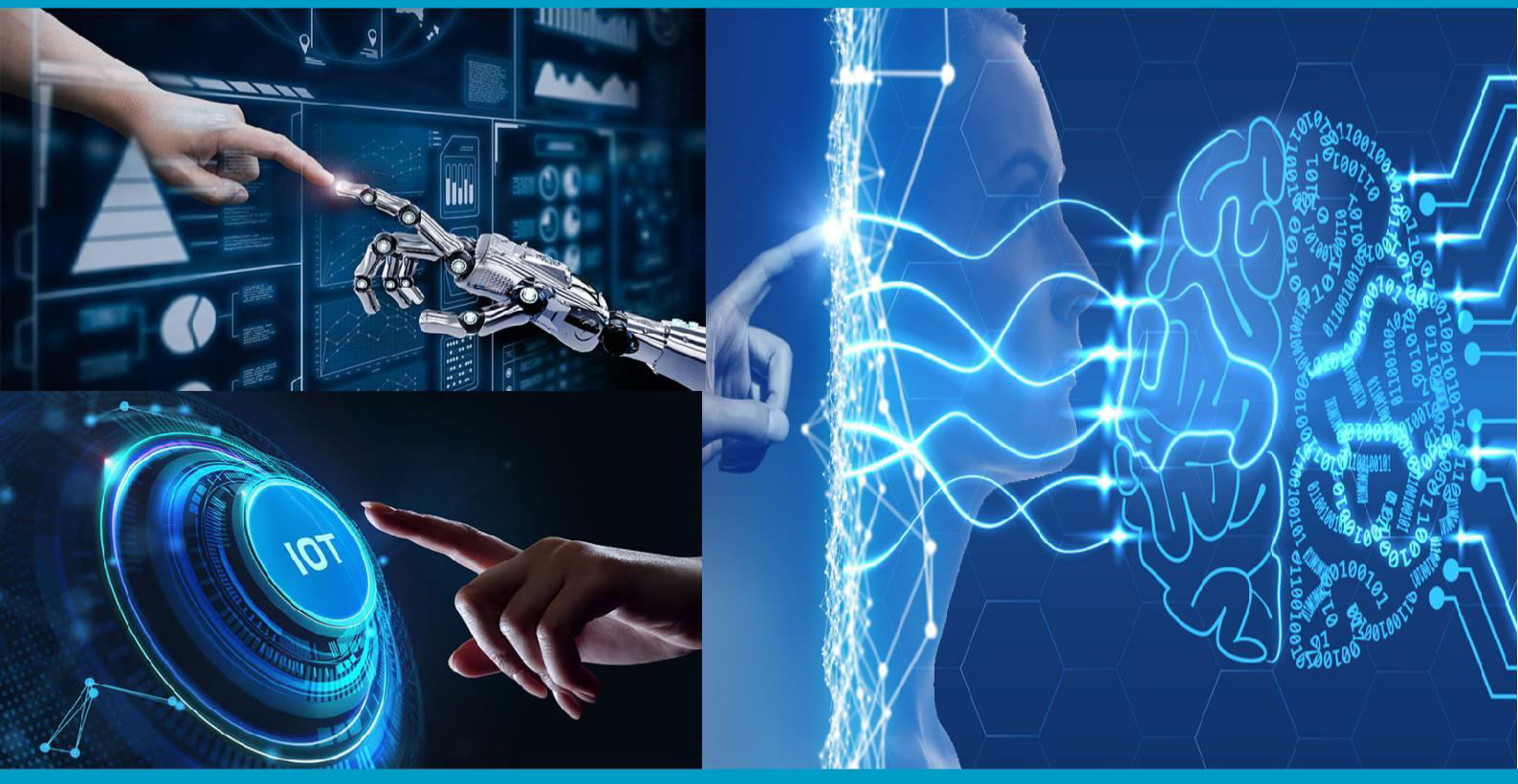
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 12, Issue 4, April 2025

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.214

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

Food Sales Prediction using Machine Learning

Subash N, Dr.K.Poornapriya, Dr.G.Aarthi, Mr.N.Kumaresan

Department of Master of Computer Applications, Vidyaa Vikas College of Engineering and Technology, Tiruchengode, Tamil Nadu, India

Principal, Vidyaa Vikas College of Engineering and Technology, Tiruchengode, Tamil Nadu, India

Head of the Department, Department of Master of Computer Applications, Vidyaa Vikas College of Engineering and Technology, Tiruchengode, Tamil Nadu, India

Assistant Professor, Department of Master of Computer Applications, Vidyaa Vikas College of Engineering and Technology, Tiruchengode, Tamil Nadu, India

ABSTRACT: Retail sales forecasting plays a crucial role in optimizing inventory management and maximizing revenue for businesses. In this study, we explored the application of two regression techniques, namely Ridge Regression and Support Vector Machine (SVM) Regression, to predict sales figures for a retail dataset. The dataset comprises various features such as item attributes, outlet information, and sales figures. We preprocessed the dataset by converting categorical variables into numerical representations and divided it into training and testing sets. Subsequently, we trained Ridge Regression and SVM Regression models on the training data and evaluated their performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) on the testing dataset. Our results indicate that Ridge Regression outperformed SVM Regression, demonstrating lower errors and higher R-squared values. These findings suggest that Ridge Regression is more effective in accurately predicting retail sales based on the provided features. This study contributes to the understanding of regression techniques' applicability in retail sales forecasting and provides insights for businesses seeking to enhance their sales prediction models.

KEYWORDS: Forecasting, Bigmart Sales, Machine Learning, Sales Prediction

I. INTRODUCTION

Retail sales forecasting is a vital aspect of inventory and revenue management for businesses, helping them make informed decisions and reduce operational inefficiencies. With the advent of machine learning, predictive modeling has become more accurate and scalable. In this study, we focus on building a predictive system for BigMart sales using regression techniques. The dataset includes features such as item attributes, outlet characteristics, and historical sales data. To prepare the data for modeling, preprocessing steps like encoding categorical variables are performed. Two regression models Ridge Regression and Support Vector Machine (SVM) Regression are implemented to forecast sales. The models are trained on a portion of the dataset and evaluated on unseen data using standard performance metrics. Our goal is to identify the more accurate model for retail sales prediction, aiding businesses in demand forecasting and strategic planning.

1.1 FORECASTING

Forecasting is the process of predicting future trends or events using historical data and current conditions. It plays a crucial role in helping individuals and organizations make strategic decisions by anticipating changes in demand, market trends, or external factors. Common forecasting methods include trend analysis, regression analysis, time series analysis, and simulation modeling. These techniques help identify patterns, relationships, and outcomes that guide future planning. The accuracy of forecasting depends on the quality of data and the method used. In fields like finance, weather, and business, forecasting aids in risk management and resource allocation. Using multiple forecasting approaches can improve prediction reliability. Ultimately, forecasting supports proactive and informed decision-making for future success.

1.2 BIGMART SALES

BigMart is a popular retail chain offering a diverse range of products such as food, apparel, electronics, and household items. The BigMart Sales dataset includes detailed information about item sales across various stores, with features like store size, location, item type, and item pricing. The primary objective is to predict item sales based on these variables, enabling better inventory planning, pricing strategies, and marketing decisions. Machine learning algorithms like linear regression, decision trees, and neural networks are often used to analyze such datasets. These algorithms possess strong

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

predictive power, uncovering hidden patterns in vast amounts of data. They also offer efficiency by processing data quickly and accurately, supporting timely decisions. Once trained, these models operate automatically, offering continuous insights into sales trends. Their flexibility allows them to adapt to different datasets and business requirements. Overall, BigMart Sales analysis helps improve operational performance through intelligent, data-driven decisions.

1.3 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence that focuses on creating algorithms that allow computers to learn from data and make decisions or predictions without explicit programming. It mimics human learning by improving its performance over time through experience. The main types of machine learning include supervised learning, where models are trained on labeled data; unsupervised learning, where patterns are discovered in unlabeled data; and reinforcement learning, where models learn through rewards and penalties. These techniques are widely used in applications such as image and speech recognition, natural language processing, and recommendation systems. Machine learning has significantly impacted industries like healthcare, finance, and transportation by enhancing diagnostics, risk prediction, and automation. Its success depends on the quality of data and the effectiveness of the models used. With the rapid increase in data availability and computing power, machine learning continues to evolve. As a result, it holds great potential to further transform the way we solve complex problems and make decisions in the future.

1.4 SALES PREDICTION

Sales prediction is a crucial component of business strategy that involves forecasting future sales of a product or service based on various data-driven techniques. It helps organizations plan production, manage inventory, allocate budgets, and design effective marketing strategies. By analyzing historical sales data, market trends, and consumer behavior, businesses can estimate demand and prepare accordingly. Accurate sales forecasting leads to better resource planning, financial stability, and improved customer satisfaction. It also enhances marketing by enabling targeted campaigns that resonate with consumer needs. Financially, it supports better cash flow management and investment planning. With reliable sales predictions, businesses can make smarter decisions, identify potential risks, and adapt quickly to market changes. Ultimately, it provides a competitive edge by helping companies meet demand proactively and strengthen customer relationships.

II. LITERATURE REVIEW

2.1 THE POWER OF SIMPLICITY: PROCESSING FLUENCY AND THE EFFECTS OF OLFACTORY CUES ON RETAIL SALES

Andreas Herrmann et.al.,has proposed in this paper Although ambient scents within retail stores have been shown to influence shoppers, real-world demonstrations of scent effects are infrequent and existing theoretical explanation for observed effects is limited. The current research addresses these open questions through the theoretical lens of processing fluency. In support of a processing fluency explanation, results across four studies show the complexity of a scent to impact consumer responses to olfactory cues. A simple (i.e., more easily processed) scent led to increased ease of cognitive processing and increased actual spending, whereas a more complex scent had no such effect. Implications for theory and retail practice are provided. © 2012 New York University. Published by Elsevier Inc. All rights reserved. Prior research has clearly demonstrated that olfactory cues can influence the perceptions and (sometimes) behaviors of consumers within retail settings. Despite the obvious commercial interest in these findings, research investigating the impact of scent on actual behavior, and identifying theoretical underpinnings for observed effects, has been limited, and indeed in some instances apparently equivocal.

2.2 AN EMPIRICAL STUDY ON HYPERPARAMETER TUNING OF DECISION TREES

Rafael Gomes Mantovani et.al.,has proposed in this paper Machine learning algorithms often contain many hyperparameters whose values affect the predictive performance of the induced models in intricate ways. Due to the high number of possibilities for these hyperparameter configurations, and their complex interactions, it is common to use optimization techniques to find settings that lead to high predictive accuracy. However, we lack insight into how to efficiently explore this vast space of configurations: which are the best optimization techniques, how should we use them, and how significant is their effect on predictive or runtime performance? This paper provides a comprehensive approach for investigating the effects of hyperparameter tuning on three Decision Tree induction algorithms, CART, C4.5 and CTree. These algorithms were selected because they are based on similar principles, have presented a high predictive performance in several previous works and induce interpretable classification models. Additionally, they contain many interacting hyperparameters to be adjusted. Experiments were carried out with different tuning strategies to induce models and evaluate the relevance of hyperparameters using 94 classification datasets from OpenML.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

Experimental results indicate that hyperparameter tuning provides statistically significant improvements for C4.5 and CTree in only one-third of the datasets, and in most of the datasets for CART.

2.3 A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S et.al.,has proposed in this paper Theoretical neurophysiology rests on certain cardinal assumptions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neuron. The velocity along the axon varies directly with its diameter, from < 1 ms-' in thin axons, which are usually short, to > 150 ms- ' in thick axons, which are usually long. The time for axonal conduction is consequently of little importance in determining the time of arrival of impulses at points unequally remote from the same source. Excitation across synapses occurs predominantly from axonal terminations to somata. It is still a moot point whether this depends upon irreciprocity of individual synapses or merely upon prevalent anatomical configurations. To suppose the latter requires no hypothesis ad hoc and explains known exceptions, but any assumption as to cause is compatible with the calculus to come.

2.4 THE MATHEMATICS OF DECISION TREES, RANDOM FOREST AND FEATURE IMPORTANCE IN SCIKIT-LEARN AND SPARK

X yang et.al.,has proposed in this paper This post attempts to consolidate information on tree algorithms and their implementations in Scikit-learn and Spark. In particular, it was written to provide clarification on how feature importance is calculated. There are many great resources online discussing how decision trees and random forests are created and this post is not intended to be that. Although it includes short definitions for context, it assumes the reader has a grasp on these concepts and wishes to know how the algorithms are implemented in Scikit-learn and Spark. Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The condition, or test, is represented as the "leaf" (node) and the possible outcomes as "branches" (edges). This splitting process continues until no further gain can be made or a preset rule is met, e.g. the maximum depth of the tree is reached. Random forests are a supervised Machine learning algorithm that is widely used in regression and classification problems and produces, even without hyperparameter tuning a great result most of the time. It is perhaps the most used algorithm because of its simplicity. It builds a number of decision trees on different samples and then takes the majority vote if it's a classification problem.

2.5 AN OVERVIEW OF STATISTICAL LEARNING THEORY

Yang et.al.,has proposed in this paper Statistical learning theory was introduced in the late 1960's. Until the 1990's it was a purely theoretical analysis of the problem of function estimation from a given collection of data. In the middle of the 1990's new types of learning algorithms (called support vector machines) based on the developed theory were proposed. This made statistical learning theory not only a tool for the theoretical analysis but also a tool for creating practical algorithms for estimating multidimensional functions. This article presents a very general overview of statistical learning theory including both theoretical and algorithmic aspects of the theory. The goal of this overview is to demonstrate how the abstract learning theory established conditions for generalization which are more general than those discussed in classical statistical paradigms and how the understanding of these conditions inspired new algorithmic approaches to function estimation problems. This article presents a very general overview of statistical learning theory. It demonstrates how an abstract analysis allows us to discover a general model of generalization. According to this model, the generalization ability of learning machines depends on capacity concepts which are more sophisticated than merely the dimensionality of the space or the number of free parameters of the loss function (these concepts are the basis for the classical paradigm of generalization).

III. EXISTING SYSTEM

Existing sales prediction systems often rely on traditional algorithms such as XGBoost, Linear Regression, Polynomial Regression, or simple moving averages. These basic models are commonly used but have limitations, especially when dealing with complex, seasonal, or rapidly changing sales patterns. Moving averages may perform adequately in flat demand scenarios but react too slowly to trends or seasonal shifts. In many cases, managers override baseline predictions with their intuition, which can lead to errors due to information overload, lack of experience, or human oversight. Traditional statistical methods like exponential smoothing also struggle with adapting to dynamic market conditions. Additionally, spreadsheet-based tools like Microsoft Excel are frequently used for forecasting but lack the scalability and efficiency required for handling large or complex datasets. These methods often require manual input and don't incorporate external factors like holidays, weather, or economic shifts. As a result, such systems may fall short in providing accurate and timely sales forecasts for modern retail environments.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

IV. PROPOSED SYSTEM

The proposed system focuses on building a robust and accurate sales prediction model for BigMart stores using advanced machine learning techniques. Utilizing the provided BigMart Sales dataset, the system will begin with thorough data preprocessing, which includes converting categorical variables into numerical format to ensure compatibility with regression models. The dataset will then be split into training and testing sets, containing key features such as item identifiers, item weight, item visibility, item type, item MRP, outlet ID, outlet establishment year, outlet size, location type, and outlet type. Two regression algorithms—Ridge Regression and Support Vector Machine (SVM) Regression—will be employed to train on the dataset and capture the underlying patterns between input features and the target variable, which is item sales. After training, the models will predict sales values for the test set. These predictions will be evaluated using performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. Through these evaluations, the system will identify which model offers greater accuracy and reliability. It is anticipated that Ridge Regression will outperform SVM Regression by producing lower prediction errors and higher R^2 values. The results will offer insights into the strengths of each model and guide future improvements. Ultimately, this predictive system aims to support BigMart in improving its inventory planning, marketing strategies, and overall sales performance through data-driven forecasting.

4.1 BIGMART SALES DATASET

The BigMart Sales Dataset is the core dataset used for this project. It contains various attributes such as item identifier, item weight, item visibility, item type, outlet identifier, outlet establishment year, outlet size, outlet location type, and outlet type. The target variable in this dataset is the Item_Outlet_Sales, which represents the sales of each item in the BigMart stores.

4.2 DATASET AFTER CATEGORICAL TO NUMERICAL CONVERSION

After preprocessing the BigMart Sales Dataset, categorical variables are converted into numerical representations. This transformation enables the utilization of machine learning algorithms, as most algorithms require numerical input data. By converting categorical variables to numerical form, the dataset becomes suitable for training regression models to predict sales based on various factors.

4.3 TRAINING DATASET WITHOUT CLASS LABEL

The training dataset consists of a subset of the preprocessed data, excluding the class label, which in this case represents the sales figures. This dataset is utilized for training machine learning models, allowing them to learn patterns and relationships between input features (e.g., item attributes, outlet information) and the target variable (sales).

4.4 TRAINING DATASET CLASS LABEL VALUES ONLY

In contrast to the previous dataset, this module contains only the class label values corresponding to the training dataset. These values represent the actual sales figures associated with the training data instances. They serve as the ground truth against which the predictions of the regression models are evaluated during training and testing phases.

4.5 TESTING DATASET WITHOUT CLASS LABEL

Similar to the training dataset, the testing dataset comprises a subset of the preprocessed data without the class label (sales figures). This dataset is used to evaluate the performance of trained regression models on unseen data. By providing input features without corresponding sales figures, it simulates where predictions are made based on new observations.

4.6 ACTUAL CLASS LABEL VALUES FOR TESTING DATASET

This module contains the actual sales figures corresponding to the instances in the testing dataset. These values serve as a reference for evaluating the accuracy of predictions generated by regression models. By comparing the predicted sales figures with the actual values, the performance of the models can be assessed using metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared.

4.7 RR Predicted class label values for Testing Dataset

After training a Ridge Regression model on the training dataset, predictions are made for the testing dataset. The predicted sales figures are compared against the actual values to assess the model's performance. Metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared are calculated to quantify the accuracy and efficacy of the Ridge Regression model in predicting sales.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

4.8 SVM Predicted class label values for Testing Dataset

Similarly, after training a Support Vector Machine (SVM) Regression model on the training dataset, predictions are generated for the testing dataset. The predicted sales figures are then compared with the actual values to evaluate the performance of the SVM model. Metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared are computed to gauge the accuracy and effectiveness of the SVM Regression model in predicting sales.

V. RESULT ANALYSIS

The result analysis of the implemented Ridge Regression and Support Vector Machine (SVM) Regression models provides clear insights into their predictive capabilities for retail sales. Ridge Regression consistently outperformed SVM Regression across all evaluation metrics, including MAE, MSE, RMSE, and R-squared. The lower error values and higher R-squared score of Ridge Regression indicate a better fit and greater accuracy in predicting sales data. This suggests that Ridge Regression effectively captures complex relationships between item-level and outlet-level features and the corresponding sales. In contrast, SVM Regression displayed higher errors and even a negative R-squared value, indicating weaker predictive performance. These results highlight the importance of choosing suitable regression models in sales forecasting tasks. Ridge Regression proved more reliable and robust for the BigMart dataset. This analysis supports the continued use and further optimization of Ridge Regression through hyperparameter tuning and advanced feature engineering to enhance future prediction models.

VI. CONCLUSION

In conclusion, the application of machine learning techniques, particularly Ridge Regression and Support Vector Machine (SVM) Regression, demonstrates promising results in enhancing sales forecasting accuracy within the retail sector. Through the analysis of the BigMart Sales Dataset and the evaluation of the trained regression models on testing data, it is evident that Ridge Regression outperforms SVM Regression in terms of predictive performance metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared. These findings underscore the potential of Ridge Regression as an effective tool for predicting retail sales based on item attributes, outlet details, and historical sales data. Moving forward, continued research and refinement of machine learning models can further improve sales forecasting accuracy, enabling retailers to make data-driven decisions and optimize their operations for increased profitability and customer satisfaction.

VII. FUTURE WORK

For future work, there are several avenues to explore in enhancing sales forecasting within the retail sector using machine learning methodologies. Firstly, incorporating more advanced regression techniques and ensemble learning approaches could improve predictive accuracy further. Additionally, exploring the integration of external data sources such as weather patterns, economic indicators, and consumer trends could provide valuable insights into sales fluctuations and seasonal patterns. Moreover, leveraging deep learning architectures, such as neural networks, may uncover complex nonlinear relationships within the data, leading to more accurate sales predictions.

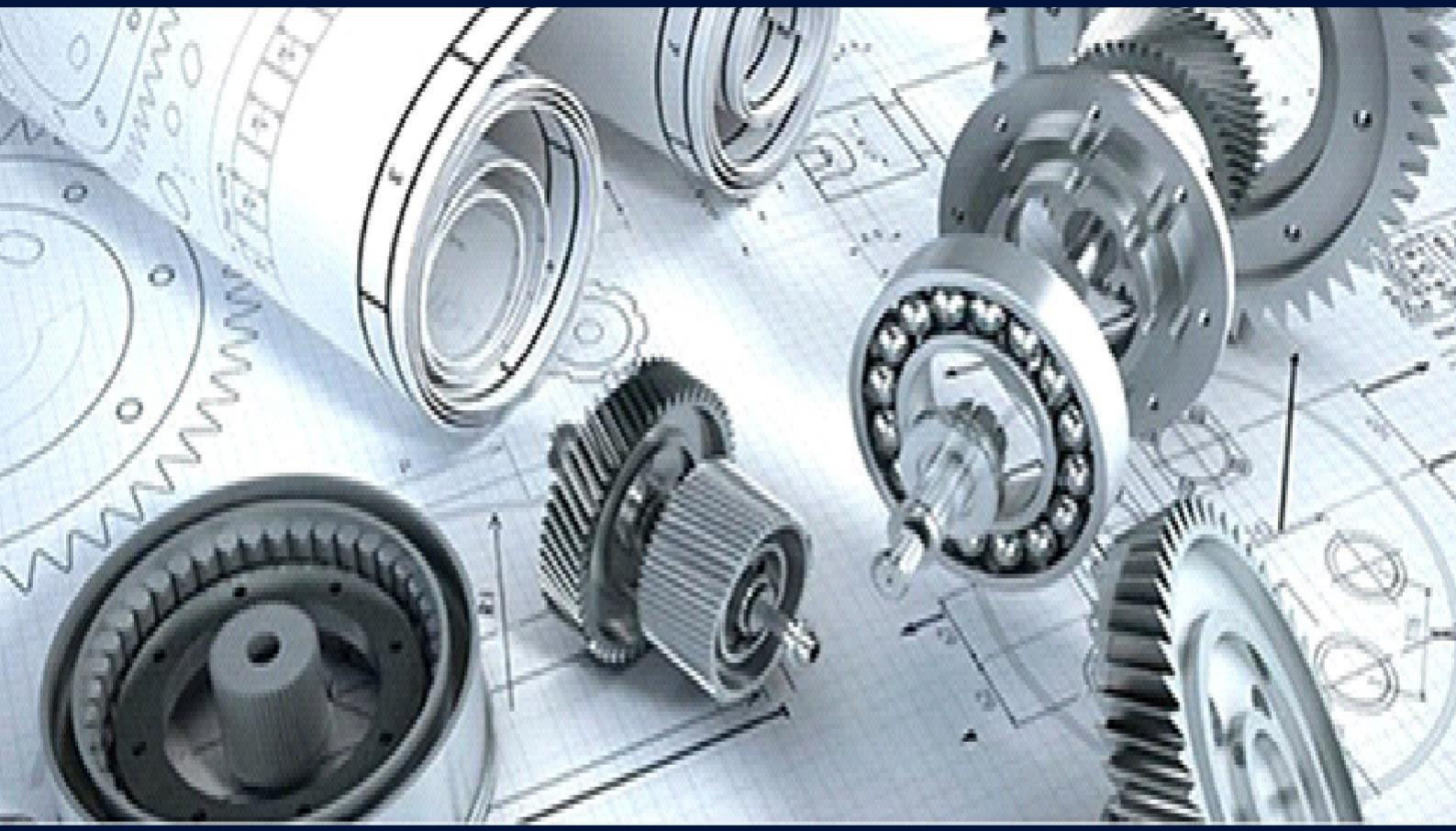
REFERENCES

1. "Sales-forecasting of Retail Stores using Machine Learning Techniques," IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, IEEE Xplore, 160-166 (2020), Akshay Krishna, Akhilesh V, Animikh Aich, and Chetana Hegde.
2. "Intelligent Sales Prediction Using Machine Learning Techniques," International Conference on Computing, Electronics & Communications Engineering (ICCECE), IEEE Xplore, 53-58, Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa (2021).
3. "A Multi-Task Prediction Framework for Sales Prediction," Chenhui Lu, Shuo Feng, Jiahao Huang, and Xiaojun Ye (2021), International Conference on Computing, Electronics & Communications Engineering (ICCECE), IEEE Xplore, 194-19
4. "Sales forecasting by combining clustering and machine-learning techniques for computer retailing," by I-Fei Chen and Chi-Jie Lu (2019). 105–112 in Neural Comput & Applications, Vol. 18
5. Xie dairu1,Zhang Shilong1 (2021), " Machine Learning Model for Sales Forecasting by Using XGBoost" IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE 2021),IEEE, 480-483
6. A two-level statistical model for big-mart sales prediction was presented by Punam, K., Pamula, R., and Jain, P. K. (2019) at the 2018 International Conference on Computing, Power, and Communication Technologies (GUCON), IEEE, 617–620. At the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE,

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

pages 57–61, Machandra H V, Balaraju G, Rajashekar A, and Harish Patil (2021) present "Machine Learning Application for Black Friday Sales Prediction Framework."
7. "Demand forecasting in restaurants using machine learning and statistical analysis," by Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, and Takeshi Takenakam (2019). CIRP Procedia, 79:679–683.
8. Chu Luo, Yuehui Zhang, Yanshan Tian, and Xu Ma (2018), "Using big data to predict future restaurant patrons," International Conference on Machine Learning and Cybernetics (ICMLC), volume 1, IEEE, 269–274.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com