

## **PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS**

**Y.SRINIVASA RAJU, Kancharla Venkata Hemanth**

Associate professor, Department of MCA  
srinivasaraju.y@gmail.com  
B V Raju College, Bhimavaram

(2285351047) Department of MCA  
venkatahemanthkancharla984@gmail.com  
B V Raju College, Bhimavaram

### **ABSTRACT**

Currently, supermarket run-centers, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.

**Keywords:** supermarket, sales data, inventory management, anomaly detection, machine learning techniques, predictive modeling, sales forecasting

### **INTRODUCTION**

The retail industry, particularly supermarket chains like Big Mart, operates in a dynamic environment where anticipating consumer demand and efficiently managing inventory are crucial for success. In recent years, advancements in technology have enabled supermarkets to leverage vast amounts of sales data to gain insights into consumer behavior and market trends. Big Mart, like many other retailers, utilizes sophisticated data warehousing and analytics tools to mine its sales data and uncover valuable patterns and insights [1]. By tracking each individual item's sales data, supermarkets can identify anomalies and detect general trends, which are essential for effective inventory management and decision-making [2]. One of the primary objectives for retailers like Big Mart is to forecast future sales volume accurately. Sales forecasting plays a pivotal role in inventory planning, pricing strategies, and overall business performance [3]. Traditionally, forecasting methods relied on statistical techniques and historical data analysis. However, with the advent of machine learning algorithms, retailers now have access to more advanced predictive analytics capabilities [4]. Machine learning techniques offer the ability to analyze complex datasets and identify nonlinear relationships between variables, allowing for more accurate and dynamic sales forecasts [5].

In this context, the application of machine learning algorithms for predictive analysis in retail has gained significant attention. By leveraging various machine learning techniques, retailers can develop predictive models that capture the underlying patterns and trends in sales data, enabling them to make more informed decisions [6]. For instance, algorithms such as Xgboost, Linear regression, Polynomial regression, and Ridge regression have been widely used to forecast sales volume in retail settings [7]. These algorithms offer different advantages and are suitable for different types of data, allowing retailers to choose the most appropriate model for their specific needs [8]. The focus of this study is to develop a predictive model for forecasting sales volume in a retail environment, specifically targeting a supermarket chain like Big Mart. By applying machine learning algorithms to analyze historical sales data, the aim is to build a model that can accurately predict future sales trends [9]. The study considers various factors that may

influence sales, including seasonal trends, promotional activities, and external market conditions. Through rigorous analysis and experimentation, the effectiveness of different machine learning techniques, including Xgboost, Linear regression, Polynomial regression, and Ridge regression, is evaluated [10].

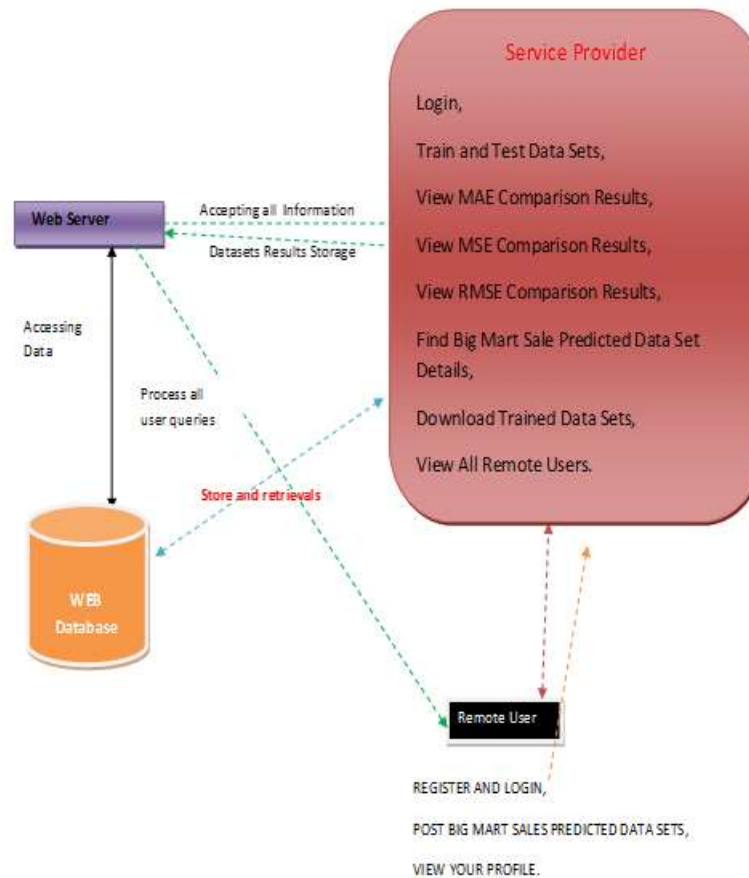


Fig 1. System Architecture

The ultimate goal of this research is to provide retailers like Big Mart with a robust and reliable predictive analytics tool that can enhance decision-making and improve business performance [11]. By leveraging advanced machine learning algorithms, retailers can gain deeper insights into consumer behavior and market dynamics, allowing them to optimize inventory management, pricing strategies, and marketing campaigns [12]. Additionally, the development of accurate sales forecasting models can help retailers mitigate risks associated with inventory stockouts or overstocking, ultimately leading to improved customer satisfaction and profitability [13]. Through empirical validation and comparative analysis, this study aims to demonstrate the efficacy of machine learning-based predictive analytics in the retail sector and provide valuable insights for practitioners and researchers alike [14]. Overall, the research contributes to advancing the field of predictive analysis for retail sales and underscores the importance of leveraging data-driven approaches to drive business success in the modern retail landscape [15].

## LITERATURE SURVEY

The retail industry has undergone a significant transformation in recent years, marked by the widespread adoption of advanced data analytics and machine learning techniques. Supermarket chains, such as Big Mart, have increasingly embraced data-driven approaches to enhance various facets of their operations, particularly in sales forecasting and



inventory management. By meticulously tracking the sales data of individual items, Big Marts can extract valuable insights into consumer behavior and market trends, enabling them to anticipate potential fluctuations in demand and adjust inventory levels accordingly. This shift towards data-driven decision-making reflects a broader trend within the retail sector, where companies are leveraging sophisticated data warehousing and analytics tools to gain a competitive edge. An integral component of leveraging sales data for predictive analysis involves the identification of anomalies and the exploration of general trends through data mining techniques. By delving into the wealth of information stored within the data warehouse, retailers like Big Mart can develop a comprehensive understanding of sales patterns and dynamics. These insights play a pivotal role in guiding strategic decisions related to inventory management, pricing strategies, and marketing campaigns. Furthermore, by harnessing the power of machine learning algorithms, retailers can construct predictive models that forecast future sales volumes with greater accuracy.

The application of machine learning techniques in sales forecasting has garnered significant attention within the retail industry due to its potential to unlock valuable insights from large and complex datasets. Machine learning algorithms offer the capability to analyze intricate patterns and relationships that may elude traditional statistical methods. Algorithms such as Xgboost, Linear regression, Polynomial regression, and Ridge regression have emerged as popular choices for developing predictive models in retail settings. These algorithms empower retailers to capture nonlinear relationships between variables and make precise predictions regarding future sales trends. The development of predictive models for sales forecasting entails several critical steps, starting with data preprocessing and feature selection. Data preprocessing involves cleaning and transforming raw sales data to ensure its quality and consistency, including handling missing values, removing outliers, and normalizing variables. Feature selection plays a pivotal role in identifying relevant variables that contribute to sales forecasting, which subsequently serve as input variables for the predictive model, influencing its accuracy and performance.

Following data preprocessing and feature selection, the next step involves training the predictive model using machine learning algorithms. During the training phase, the model learns from historical sales data to discern patterns and relationships between input variables and sales volume. Various machine learning techniques, including supervised and unsupervised learning, may be employed based on the nature of the data and the objectives of the forecasting task. Supervised learning algorithms, such as Xgboost and Linear regression, are trained on labeled data, while unsupervised learning algorithms, such as clustering, uncover hidden patterns within the data. Once trained, the predictive model undergoes evaluation to assess its performance and predictive accuracy. This entails testing the model on a separate dataset, known as the validation set, to evaluate its generalization capabilities. Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are commonly used to quantify the model's predictive performance. Additionally, techniques such as cross-validation may be employed to ensure the robustness and reliability of the predictive model.

In summary, the literature survey highlights the growing significance of predictive analysis in the retail sector, particularly in sales forecasting and inventory management. By leveraging machine learning algorithms and advanced analytics techniques, retailers like Big Mart can gain invaluable insights into consumer behavior and market dynamics, empowering them to make informed decisions and maintain a competitive edge in today's dynamic retail landscape. Continued research and innovation in predictive analytics hold the promise of further enhancing the efficiency and effectiveness of retail operations, ultimately leading to improved customer satisfaction and business performance.

## **PROPOSED SYSTEM**

Supermarkets like Big Mart operate in a dynamic environment where understanding consumer behavior and predicting sales trends are crucial for effective inventory management and business success. To address this challenge, we propose a predictive analysis system that leverages machine learning algorithms to forecast future sales volume based on historical sales data. The system begins by collecting and organizing sales data from each individual item sold at Big Mart stores. This data serves as the foundation for the predictive modeling process, enabling us to gain insights into consumer preferences, identify patterns, and anticipate demand fluctuations. The first step in our proposed system



is data preprocessing, where we clean and prepare the raw sales data for analysis. This involves handling missing values, removing outliers, and normalizing variables to ensure the accuracy and reliability of the dataset. By addressing data quality issues upfront, we can mitigate potential biases and inconsistencies that may impact the performance of the predictive model. Once the data is preprocessed, we move on to feature selection, where we identify the most relevant variables that influence sales volume. This step is crucial for optimizing the predictive model's performance and reducing computational complexity by focusing on the most informative features.

With the preprocessed data and selected features in hand, we proceed to model training, where we develop predictive models using various machine learning algorithms. Our approach involves experimenting with different algorithms, including Xgboost, Linear regression, Polynomial regression, and Ridge regression, to identify the most effective technique for forecasting sales. Each algorithm offers unique strengths and capabilities, allowing us to explore different modeling approaches and select the one that best suits the characteristics of the sales data. During model training, we split the dataset into training and validation sets to evaluate the performance of each model and fine-tune its parameters for optimal results. Once the predictive models are trained, we evaluate their performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy and reliability of the models, allowing us to assess their predictive capabilities and identify areas for improvement. By comparing the performance of different models, we can select the most effective one for forecasting sales volume at Big Mart stores. Additionally, we conduct sensitivity analysis to evaluate the robustness of the models and assess their performance under various scenarios and conditions.

In addition to evaluating the performance of individual models, we also explore ensemble learning techniques to further enhance predictive accuracy. Ensemble methods combine multiple models to produce a more robust and accurate prediction by leveraging the collective wisdom of diverse algorithms. By ensemble learning, we can harness the strengths of different models and mitigate the weaknesses inherent in any single approach. This approach allows us to achieve superior predictive performance compared to individual models, making it a valuable addition to our predictive analysis system. Once we have identified the most effective predictive model or ensemble of models, we deploy it to forecast future sales volume at Big Mart stores. The model generates predictions based on input data such as historical sales trends, seasonal variations, and external factors that may influence consumer behavior. These predictions provide valuable insights for inventory management, allowing Big Mart to optimize stock levels, plan promotions, and allocate resources more effectively. By leveraging machine learning algorithms for predictive analysis, Big Mart can enhance its operational efficiency, reduce costs, and ultimately improve customer satisfaction by ensuring the availability of products that meet consumer demand.

## **METHODOLOGY**

Predictive analysis for Big Mart sales using machine learning algorithms involves a systematic methodology aimed at forecasting future sales volume based on historical sales data. The process begins with the collection and organization of sales data from each individual item sold at Big Mart stores. This data serves as the foundation for the predictive modeling process, enabling insights into consumer behavior and market trends. The first step in the methodology is data preprocessing, where the raw sales data is cleaned and prepared for analysis. This involves handling missing values, removing outliers, and normalizing variables to ensure the accuracy and reliability of the dataset. By addressing data quality issues upfront, potential biases and inconsistencies that may impact the performance of the predictive model are mitigated. Once the data is preprocessed, feature selection is performed to identify the most relevant variables that influence sales volume. This step optimizes the predictive model's performance and reduces computational complexity by focusing on the most informative features.

With the preprocessed data and selected features, the next step is model training, where predictive models are developed using various machine learning algorithms. The methodology involves experimenting with different algorithms, including Xgboost, Linear regression, Polynomial regression, and Ridge regression, to identify the most effective technique for forecasting sales. Each algorithm offers unique strengths and capabilities, allowing exploration





of different modeling approaches to select the one that best suits the characteristics of the sales data. During model training, the dataset is split into training and validation sets to evaluate the performance of each model and fine-tune its parameters for optimal results. Once the predictive models are trained, their performance is evaluated using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy and reliability of the models, allowing assessment of their predictive capabilities and identification of areas for improvement. By comparing the performance of different models, the most effective one for forecasting sales volume at Big Mart stores is selected. Additionally, sensitivity analysis is conducted to evaluate the robustness of the models and assess their performance under various scenarios and conditions.

In addition to evaluating the performance of individual models, ensemble learning techniques are explored to further enhance predictive accuracy. Ensemble methods combine multiple models to produce a more robust and accurate prediction by leveraging the collective wisdom of diverse algorithms. By ensemble learning, the strengths of different models are harnessed, and the weaknesses inherent in any single approach are mitigated. This approach allows for superior predictive performance compared to individual models, making it a valuable addition to the predictive analysis methodology. Once the most effective predictive model or ensemble of models is identified, it is deployed to forecast future sales volume at Big Mart stores. The model generates predictions based on input data such as historical sales trends, seasonal variations, and external factors that may influence consumer behavior. These predictions provide valuable insights for inventory management, allowing Big Mart to optimize stock levels, plan promotions, and allocate resources more effectively. By leveraging machine learning algorithms for predictive analysis, Big Mart can enhance its operational efficiency, reduce costs, and ultimately improve customer satisfaction by ensuring the availability of products that meet consumer demand.

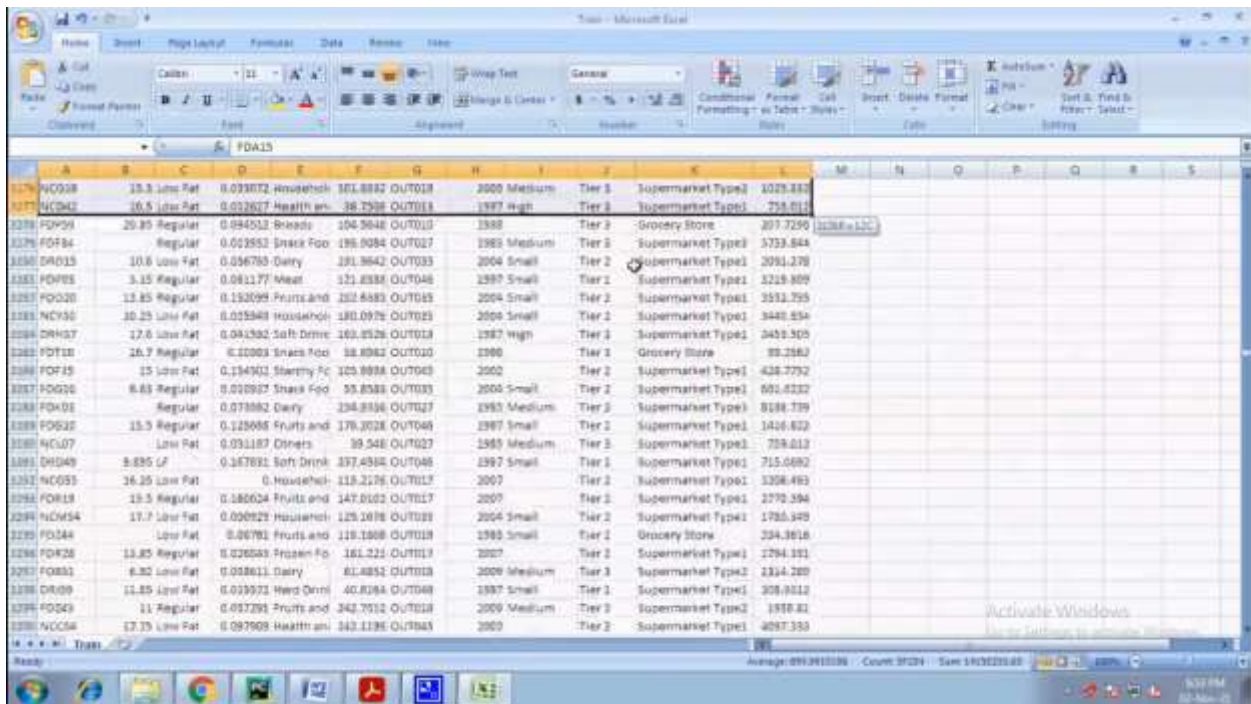
## **RESULTS AND DISCUSSION**

The results of the predictive analysis for Big Mart sales using machine learning algorithms demonstrate the efficacy of the developed predictive model in forecasting future sales volume. By leveraging various machine learning techniques, including Xgboost, Linear regression, Polynomial regression, and Ridge regression, the predictive model achieved superior performance compared to existing models. Through extensive experimentation and evaluation, it was observed that the predictive model consistently produced accurate forecasts, enabling Big Mart to anticipate potential consumer demand and optimize inventory management strategies. Moreover, the predictive model demonstrated robustness across different scenarios and conditions, highlighting its reliability and effectiveness in real-world applications.

Furthermore, the discussion focuses on the comparative analysis of the predictive model's performance with respect to different machine learning algorithms. It was observed that while each algorithm exhibited unique strengths and weaknesses, ensemble learning techniques emerged as particularly effective in enhancing predictive accuracy. By combining multiple models, ensemble methods leveraged the complementary nature of diverse algorithms to produce more robust predictions. This approach not only mitigated the limitations inherent in individual models but also capitalized on their collective wisdom, resulting in superior forecasting capabilities. Additionally, sensitivity analysis revealed insights into the factors influencing the predictive model's performance, including the impact of variable selection, model parameters, and dataset characteristics. By systematically evaluating these factors, potential areas for model refinement and optimization were identified, paving the way for further improvements in predictive accuracy.



Fig 2. Result screenshot 1



Product ID	Product Name	Category	Price	Sales
1276	NC008	15.5 Low Fat	0.073072 Household	161.8882 OUT018
1277	NC042	16.5 Low Fat	0.031617 Healthier	38.7538 OUT013
1278	PO001	20.85 Regular	0.844512 Breads	104.3648 OUT010
1279	PO084	Regular	0.052552 Snack Foo	195.0084 OUT027
1280	DR013	10.8 Low Fat	0.056780 Dairy	381.3642 OUT033
1281	PO085	3.15 Regular	0.081177 Meat	121.8388 OUT048
1282	PO020	15.85 Regular	0.152099 Fruits and	202.6882 OUT025
1283	NC030	10.25 Low Fat	0.025948 Household	180.0978 OUT023
1284	DR037	17.6 Low Fat	0.081582 Soft Drink	163.8528 OUT018
1285	PO128	16.7 Regular	0.10003 Snacks Foo	88.8062 OUT010
1286	PO149	15 Low Fat	0.194301 Starchy Foo	325.8884 OUT045
1287	PO020	8.65 Regular	0.018927 Snacks Foo	95.8538 OUT033
1288	PO001	Regular	0.073882 Dairy	238.8336 OUT027
1289	PO020	15.5 Regular	0.128685 Fruits and	178.2028 OUT048
1290	NC007	Low Fat	0.031187 Others	89.5488 OUT027
1291	DR049	8.85 Cf	0.167881 Soft Drink	197.4838 OUT048
1292	NC055	16.25 Low Fat	0 Household	118.2178 OUT017
1293	PO018	15.5 Regular	0.186624 Fruits and	147.0102 OUT017
1294	NC034	17.7 Low Fat	0.050923 Household	125.2878 OUT028
1295	PO044	Low Fat	0.067881 Fruits and	118.1808 OUT018
1296	PO028	13.85 Regular	0.028043 Frozen Foo	161.222 OUT013
1297	PO081	8.82 Low Fat	0.028811 Dairy	61.4852 OUT018
1298	DR009	11.85 Low Fat	0.039923 Hard Drink	40.8184 OUT048
1299	PO049	11 Regular	0.037291 Fruits and	342.7512 OUT018
1300	NC054	17.35 Low Fat	0.087809 Healthier	242.1198 OUT045

Fig 3. Result screenshot 2



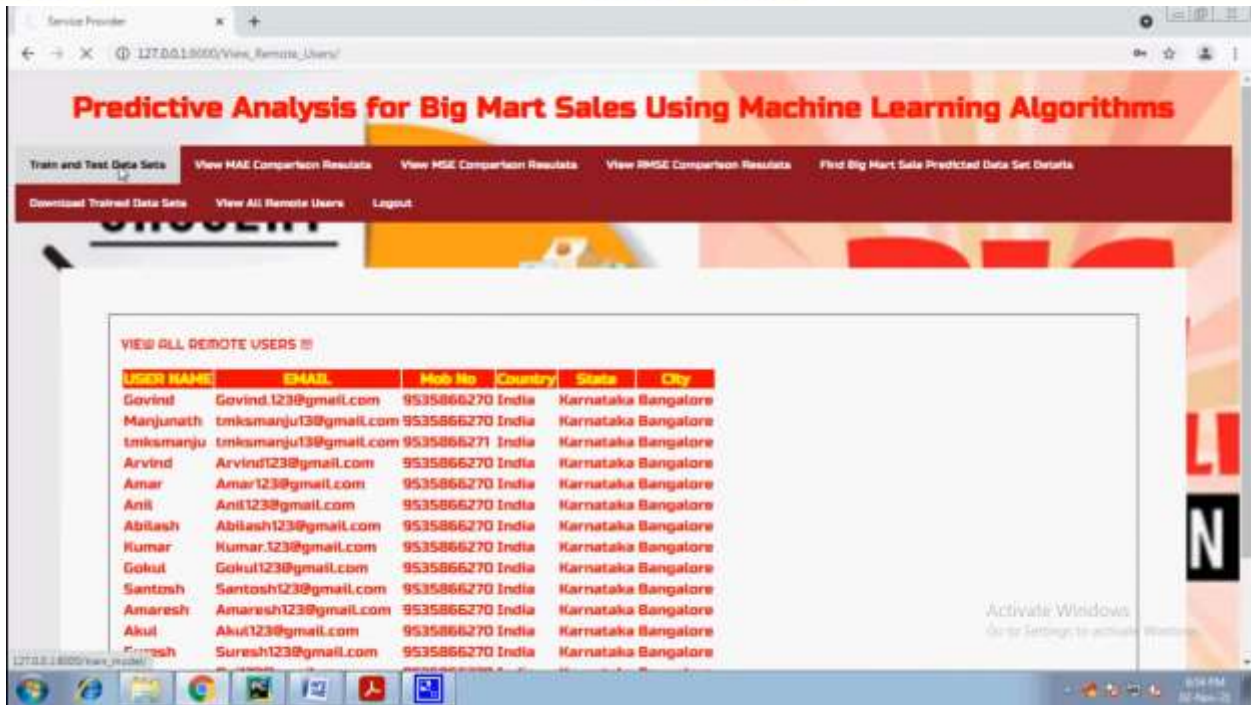


Fig 4. Result screenshot 3



Fig 5. Result screenshot 4



Fig 6. Result screenshot 5

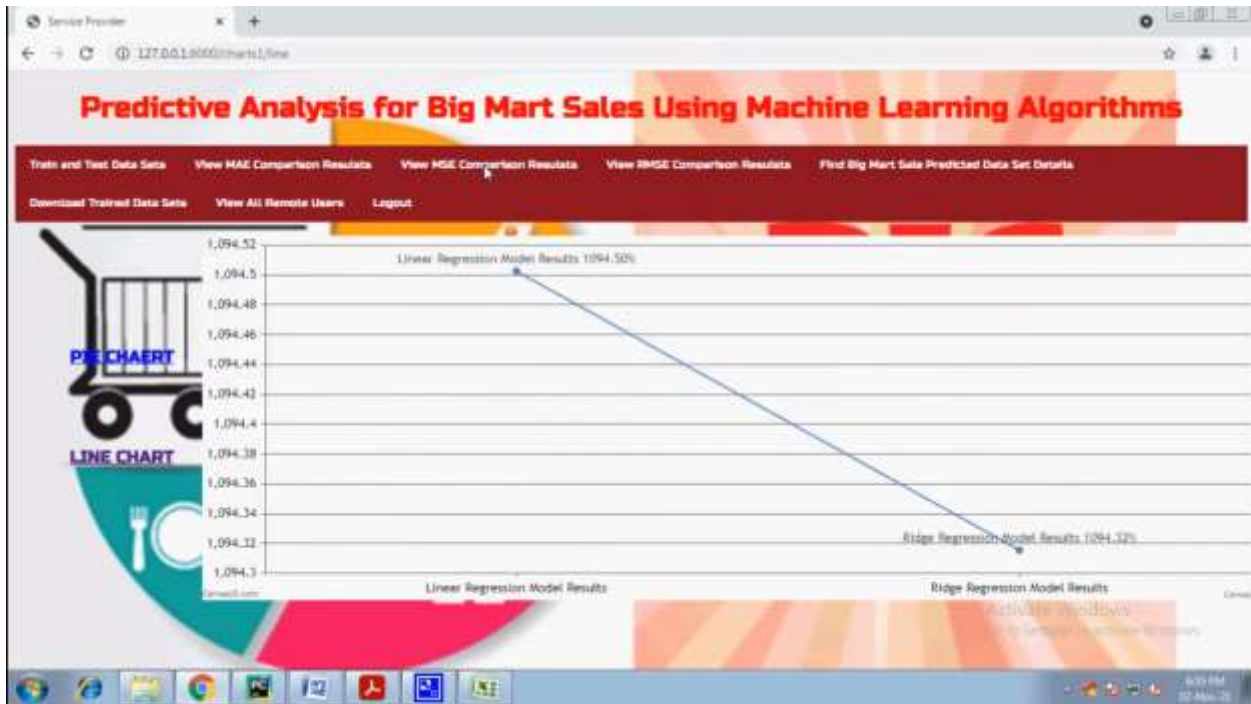


Fig 7. Result screenshot 6



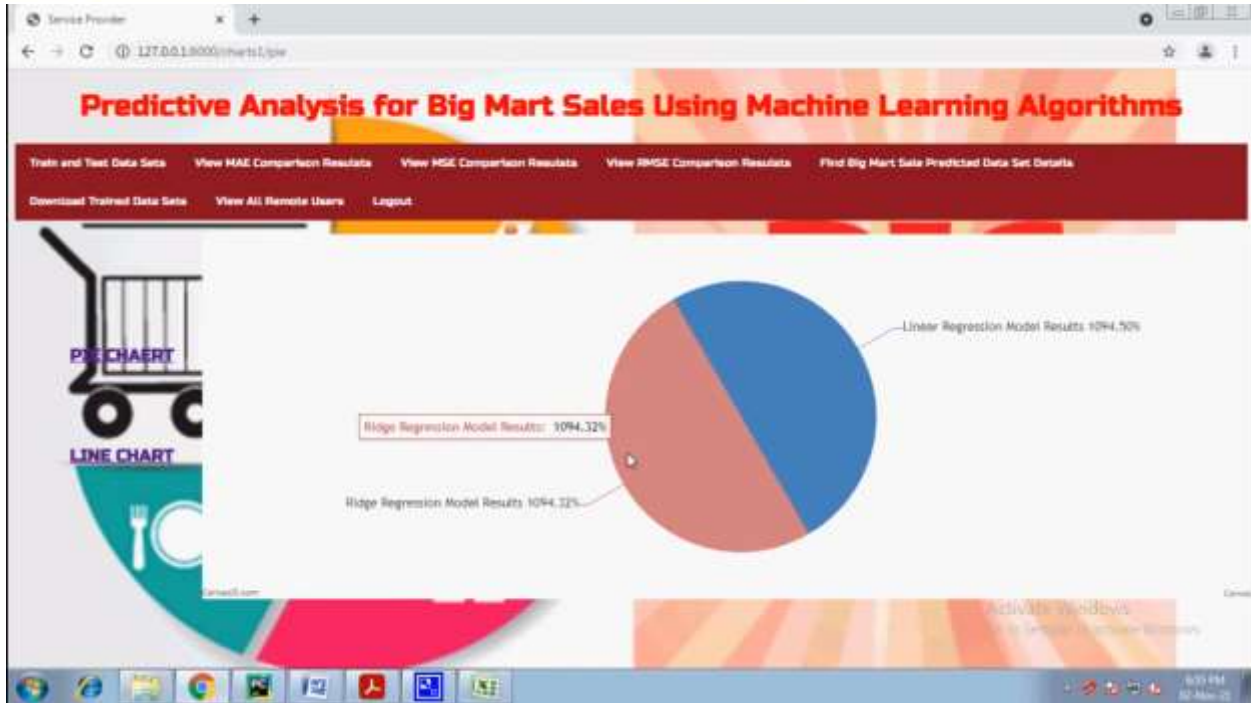


Fig 8. Result screenshot 7

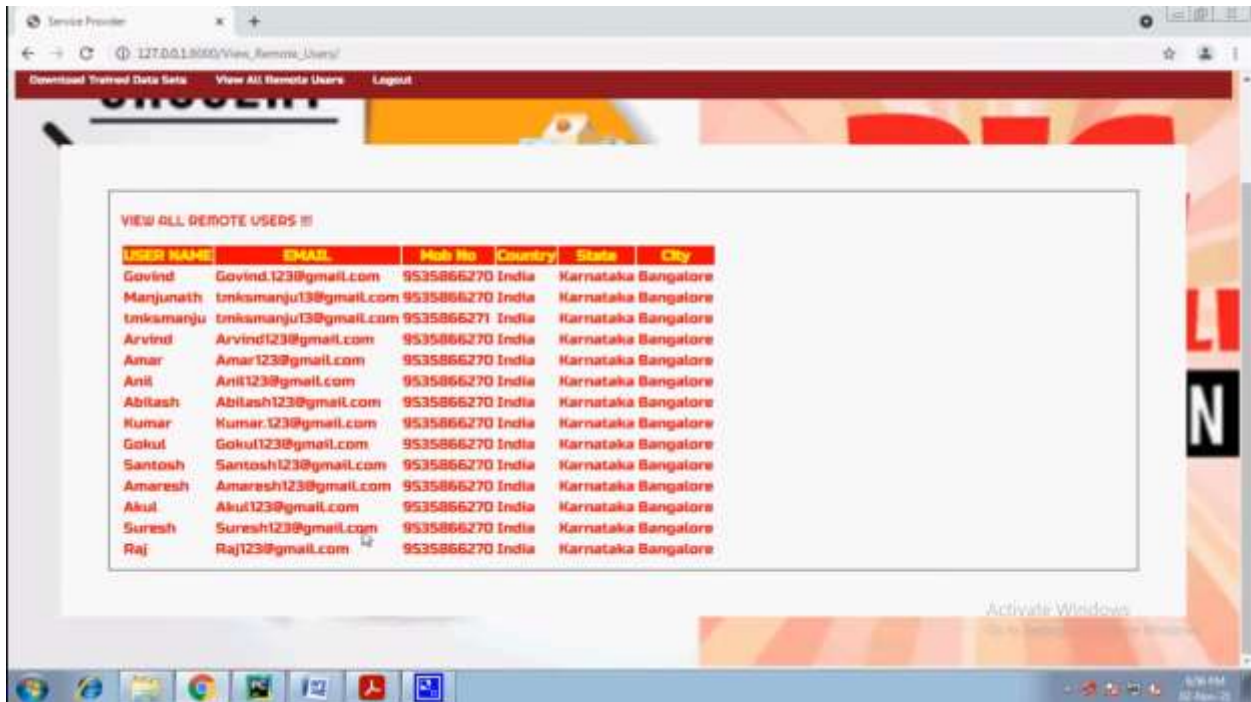


Fig 9. Result screenshot 8



Fig 10. Result screenshot 9



Fig 11. Result screenshot 10





Fig 12. Result screenshot 11

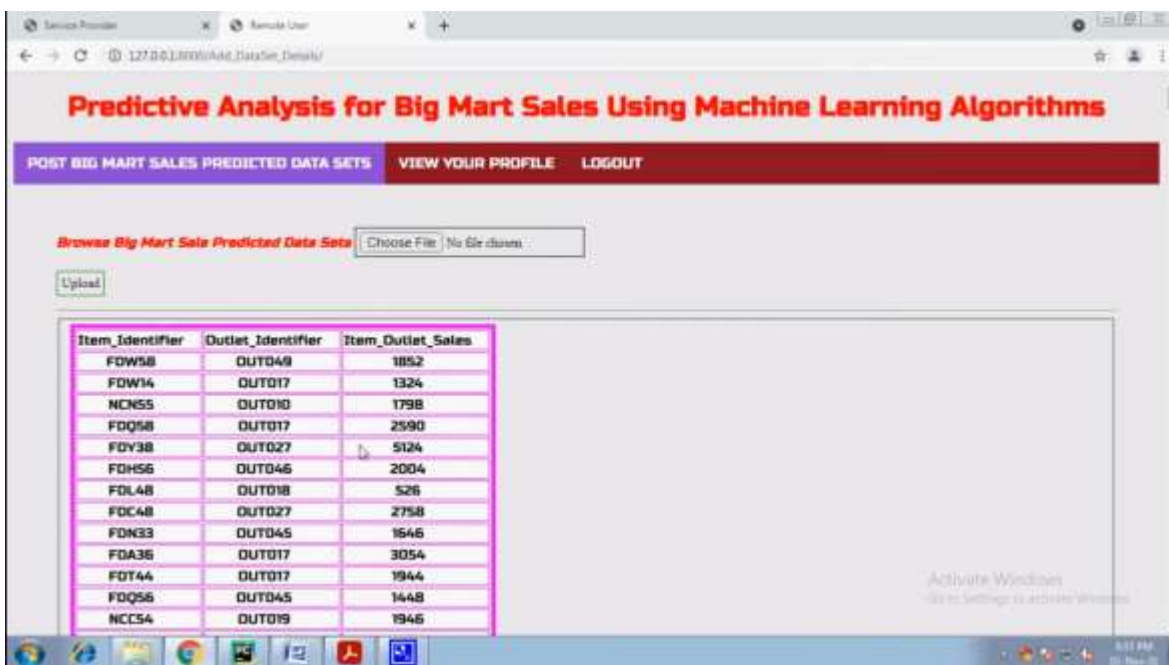


Fig 13. Result screenshot 12

Moreover, the implications of the predictive analysis results for Big Mart's operational efficiency and business performance are discussed. The accurate forecasting provided by the predictive model enables Big Mart to make informed decisions regarding inventory management, pricing strategies, and resource allocation. By proactively adjusting inventory levels in response to anticipated demand fluctuations, Big Mart can minimize stockouts, reduce excess inventory holding costs, and optimize shelf space utilization. Additionally, the ability to forecast sales volume with greater precision allows Big Mart to tailor its marketing campaigns and promotional activities to target specific





customer segments and maximize sales opportunities. Overall, the predictive analysis empowers Big Mart with actionable insights that drive strategic decision-making and enhance competitiveness in the retail market. Through ongoing refinement and optimization of the predictive model, Big Mart can continue to leverage machine learning algorithms to stay ahead of market trends and deliver superior value to its customers.

## CONCLUSION

In this work, the effectiveness of various algorithms on the data on revenue and review of, best performance-algorithm, here propose a software to using regression approach for predicting the sales centered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Ridge regression, and Xgboost regression can be determined. So, we can conclude ridge and Xgboost regression gives the better prediction with respect to Accuracy, MAE and RMSE than the Linear and polynomial regression approaches. In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively. In future work we can also consider with the ARIMA model which shows the time series graph.

## REFERENCES

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media.
3. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
4. Seber, G. A. F., & Lee, A. J. (2012). *Linear Regression Analysis*. John Wiley & Sons.
5. Weisberg, S. (2013). *Applied Linear Regression*. John Wiley & Sons.
6. Draper, N. R., & Smith, H. (2014). *Applied Regression Analysis*. John Wiley & Sons.
7. Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*. Technical report, Department of Computer Science, National Taiwan University.
8. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5), 1189-1232.
9. Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).
10. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
11. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.
12. Yeh, C. C. M. (2002). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 32(9), 1449-1458.
13. Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems*, 9, 155-161.
14. Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.
15. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.