International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Retail Sales Prediction

Dr.M.Sengaliappan[1],L.KarthigaiSelvan[2]

[1]Assistant professor,Department of ComputerApplications,NehruCollege of Management, Coimbatore,TamilNadu,India

[2]Student of IIMCA, Department of ComputerApplications, NehruCollege of Management, Coimbatore,TamilNadu, India

Abstract

Accurate retail sales forecasting is critical for optimizing inventory management, reducing operational costs, and improving strategic decision-making. Traditional statistical models often fail to capture complex consumer behaviors and market fluctuations, leading to inaccurate predictions. With advancements in machine learning (ML), predictive models now offer enhanced accuracy and adaptability in forecasting sales trends. This study evaluates the effectiveness of machine learning algorithms, specifically **k-Nearest Neighbor (kNN) regression, Multinomial regression, and Decision Tree with AdaBoost regression**, in predicting retail sales. The research assesses model performance using standard evaluation metrics, highlighting the advantages and limitations of each approach. The findings provide valuable insights for retailers, business strategists, and policymakers, offering a data-driven framework for improving sales forecasting and decision-making processes.

Keywords: Retail Sales, Forecasting, Machine Learning, Regression Models, Predictive Analytics, Business Intelligence

1. Introduction

1.1 Background and Significance

Retail sales forecasting plays a crucial role in business operations by enabling organizations to manage supply chains, optimize inventory levels, and plan promotional activities effectively. Poor sales predictions can result in stock shortages, overstocking, and revenue losses. Traditionally, retailers have relied on time-series analysis and regression-based statistical models for forecasting. However, these methods often struggle with handling large datasets and complex market dynamics. With the advent of machine learning, businesses can leverage advanced algorithms that identify patterns, detect trends, and provide more accurate sales predictions. ML- based models improve forecasting precision by dynamically adapting to external factors such as seasonality, economic trends, and customer preferences.

1.2 Research Objectives

The primary objectives of this study are:

- To explore the application of machine learning techniques in retail sales forecasting.
- To compare the performance of different ML regression models in predicting sales trends.
- To provide insights into the practical implementation of ML models for business decision-making.

2. Literature Review

2.1 Traditional Approaches to Sales Forecasting

Conventional methods, such as **autoregressive integrated moving average (ARIMA)** and **exponential smoothing**, have been widely used in sales prediction. While effective for short- term forecasts, these methods often fail in dynamic and high-volume retail environments.

2.2 Machine Learning in Retail Forecasting

Recent studies have shown that machine learning models outperform traditional statistical techniques in handling large-scale and unstructured retail data. Some of the widely used ML approaches include:

- **Supervised Learning Models**: Regression-based models such as Decision Trees and Multinomial Regression are commonly applied to predict sales based on historical data.
- **Ensemble Learning Methods**: Techniques like **AdaBoost** and **Random Forest** enhance prediction accuracy by combining multiple weak learners.

This research builds on previous findings by implementing and evaluating **kNN regression, Multinomial regression, and Decision Tree with AdaBoost regression** to determine the most efficient model for retail sales forecasting.

3. Methodology

3.1 Dataset and Preprocessing

The dataset used in this study includes historical sales data from multiple retail stores, with attributes such as:

- **Product category**
- **Store location and size**
- **Seasonal trends**
- **Discounts and promotions**

Data preprocessing involves handling missing values, normalizing numerical data, and encoding categorical variables.

3.2 Machine Learning Models

To evaluate the effectiveness of different ML techniques, the following models are implemented:

1. **k-Nearest Neighbor (kNN) Regression** – Predicts sales based on the similarity of past instances.
2. **Multinomial Regression** – A classification-based regression technique for modeling categorical sales data.
3. **Decision Tree with AdaBoost Regression** – An ensemble learning model that improves accuracy by combining multiple decision trees.

3.3 Model Evaluation Metrics

The models are assessed using key performance indicators:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **R^2 Score (Coefficient of Determination)**

4. Experimentation

The experiments were conducted by developing a simulation environment in python also using WEKA. Three machine learning algorithms k-Nearest Neighbour regression model, Multinomialregression model, Ada Boost regression model were implemented on the training dataset.

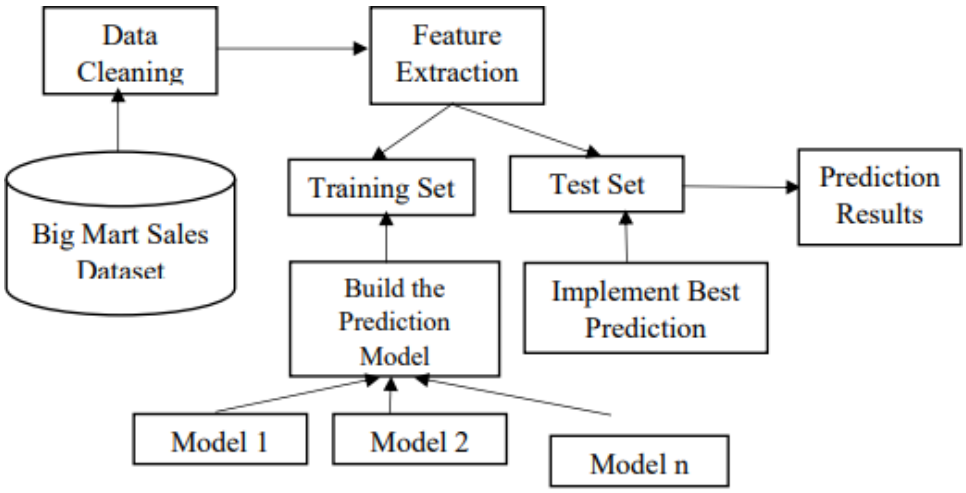

Fig 1. Proposed Flow of Work

k-Nearest Neighbour (k-NN) Regression Model: The k-NN algorithm is a non-parametric strategy used for regression analysis. In this model the input consists of k nearest training examples in thefeature space. The output value can be obtained by the average of k nearest neighbour values. In kNN algorithm it uses three distance measures Euclidean Distance, Manhattan Distance and Minkowski Distance.

The Euclidean Distance can be represented as

$$\text{Distance} = \sqrt{\sum_{j=1}^{k} (x_j - y_j)^2}$$

The Manhattan Distance can be represented as

$$\text{Distance} = \sum_{j=1}^{k} |x_j - y_j|$$

The Minkowski Distance can be represented as

$$\text{Distance} = \left(\sum_{j=1}^{k} (|x_j - y_j|)^q \right)^{1/q}$$

Table 1. Dataset Description

S. No.	Variable	Description
1	Item_Identifier	Unique product ID
2	Item_Weight	Weight of product
3	Item_Fat_Content	Whether the product is low fat or not
4	Item_Visibility	The % of total display area of all products in a store allocated to the particular product
5	Item_Type	The category to which the product belongs
6	Item_MRP	Maximum Retail Price (list price) of the product
7	Outlet_Identifier	Unique store ID
8	Outlet_Establishment_Year	The year in which store was established
9	Outlet_Size	The size of the store in terms of ground area covered
10	Outlet_Location_Type	The type of city in which the store is located
11	Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
12	Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

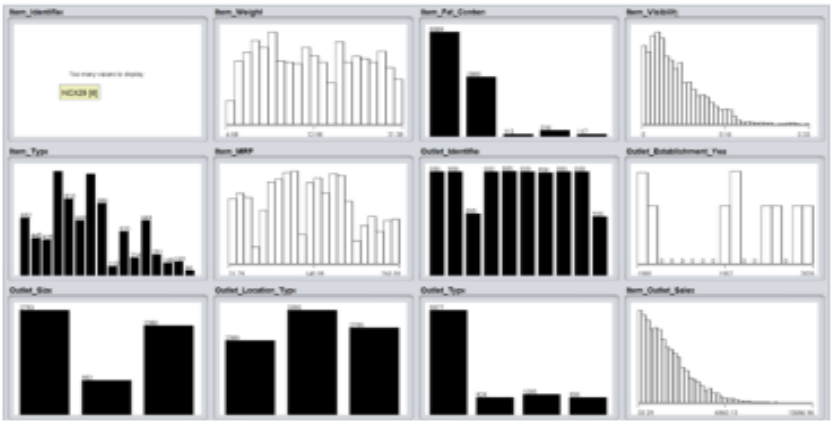

Fig 2. Attribute Visualization of Training set

Source: Experimental Setup (WEKA3.8.3)

5. Results and Discussions

The three machine learning models—K-Nearest Neighbors (KNN) Regression, Multinomial Regression, and Decision Tree Regression with AdaBoost—were implemented on the training dataset. The accuracy levels observed for these models were 85%, 100%, and 100%, respectively. These results indicate that both the Multinomial Regression and Decision Tree Regression with AdaBoost performed optimally on the training data, achieving perfect accuracy, whereas the KNN Regression model had slightly lower accuracy. The high accuracy of the latter two models suggests that they fit the training data very well. However, further evaluation on a test dataset is necessary to assess their generalization performance and to ensure that the models do not suffer from overfitting. The performance of these models on unseen data will help determine their practical applicability in real-world sales forecasting scenarios.

Table 2 Comparison of Accuracy of Models

S.no	Regression Model	Accuracy
01	K nearest neighbour Regression Model	85%
02	Multinomial Regression Model	100%
03	Decision tree Regression –AdaBoost Model	100%

The following diagram shows the graphical representation of comparison of accuracy of three models.

Fig 3. Graphical Representation of Comparison of Accuracy of Models

Source: Experimental Results

Multinomial and Decision Tree with Ada Boost regression models results in with an accuracy of 100%. So, Decision Tree with Ada Boost regression model was implemented on test set to predict the outlet sales. The following table shows the sample output. Here we randomly considered some outlets to visualize the output and we considered two digits after decimal point in the predicted values.

Table 3. Sample Output from Decision Tree with Ada Boost regression model

S. No.	Item Identifier	Outlet Identifier	Item_Outlet_Sales
1	FDW58	OUT049	1827.07
2	FDY38	OUT027	6373.43
3	FDC48	OUT027	2133.95
4	FDQ56	OUT045	1445.86
5	DRL59	OUT013	837.94
6	DRC12	OUT018	2973.97
7	FDG52	OUT046	788.68
8	FDX22	OUT010	499.05
9	FDE21	OUT035	1959.97
10	NCR06	OUT018	523.79

11	**FDC26**	**OUT013**	**1837.78**
12	**NCS41**	**OUT035**	**3135.96**
13	**FDU34**	**OUT046**	**2088.55**
14	**FDM03**	**OUT013**	**1769.50**
15	**FDV44**	**OUT027**	**5146.45**
16	**FDT04**	**OUT013**	**611.97**
17	**FDW12**	**OUT035**	**2500.40**
18	**NCY42**	**OUT027**	**3495.73**
19	**FDA14**	**OUT013**	**2267.96**
20	**FDU58**	**OUT013**	**3087.11**
21	**FDB23**	**OUT046**	**3767.15**
22	**FDF47**	**OUT027**	**6236.36**
23	**NCA30**	**OUT045**	**3029.83**
24	**DRM48**	**OUT035**	**669.02**
25	**FDS08**	**OUT049**	**2729.96**
26	**FDC53**	**OUT019**	**263.36**

Source: Experimental Results

From the above table it is clear that the regression model predicted the sales of a particular item indicated with item identifier in a particular outlet indicated with outlet identifier. Hence this model best predicts the outlet sales prediction. Using these results, Big Mart will try to understand the properties of products and outlets which play a key role in increasing sales.

4. Conclusion

Machine learning applications are increasing day-by-day in business data processing and analytics area. In this work the results derived from the machine learning based algorithms are more precise, accurate and confidently use for decision making than other models. Compared to three models the last two models are giving hundred per cent accurate results. The results of this

study can boost confidence of retailers to implement machine learning in their business data processing and analysis. It also useful for the managers associated with retail sector for developing suitable competitive marketing strategies. Last but not least it helps the policy makers to make different estimations and make suitable policies relating to retail sector.

References

1. Arif, A. I., Sany, S. I., Nahin, F. I., Shahariar, A. K. M., & Rabby, A. (2019). Comparison Study : Product Demand Forecasting with Machine Learning for Shop.

2. B. Srinivasa Rao (2018), Butterfly Customers: Strategies and Technology for Marketers, International Journal of Engineering & Technology, 7 (3.24) (2018) 512-516

3. Baba, N., Science, I., City, K., Prefecture, O., & Suto, H. (2000). for Constructing an Intelligent Sales Prediction. 565–570.

4. Behera, G. (2019). A Comparative Study of Big Mart Sales Prediction A Comparative Study of Big Mart Sales Prediction. (October).

5. Behera, G., & Nain, N. (2019). Grid Search Optimization (GSO) Based Future Sales Prediction For Big Mart. 172–178. https://doi.org/10.1109/SITIS.2019.00038

6. Cheriyan, S., Ibrahim, S., & Treesa, S. (2018). Intelligent Sales Prediction Using Machine Learning Techniques. 53–58.

7. Dr. Sujatha Kamepalli and Dr. Srinivasa Rao Bandaru (2018) Implementation Framework of

8. Artificial Intelligence in Financial Services, International Journal of Research and Analytical Reviews, November 2018, Volume 5, Issue 4.

9. Gaku, R., & Takakuwa, S. (2015). Big data-driven service level analysis for a retail store. (2008), 791–799.

10. Gao, Y. F., Liang, Y. S., Liu, Y., Zhan, S. Bin, & Ou, Z. W. (2009). A neural-network-based forecasting algorithm for retail industry. Proceedings of the 2009 International Conference on MachineLearningandCybernetics,2(July),919–924. https://doi.org/10.1109/ICMLC.2009.5212392

11. Gao, Y., Liang, Y., Tang, F., Ou, Z., & Zhan, S. (2010). A demand forecasting system for retail industry based on neural network and VBA. 2010 Chinese Control and Decision Conference, CCDC 2010, 3786–3789.https://doi.org/10.1109/CCDC.2010.5498506

12. Gao, Y., Liang, Y., Zhan, S., Ren, X., & Ou, Z. (2011). Realization of a demand forecasting algorithm for retail industry. Proceedings of the 2011 Chinese Control and Decision Conference, CCDC 2011, 4227–4230.https://doi.org/10.1109/CCDC.2011.5968968

13. Gopalakrishnan, T., Choudhary, R., & Prasad, S. (2018). Prediction of Sales Value in online shopping using Linear Regression. 5–10.

14. İşlek, İ. (2015). A Retail Demand Forecasting Model Based on Data Mining Techniques. 55– 60.

15. Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2018). Sales-forecasting of Retail Stores using Machine Learning Techniques. 160–166.

16. Liang, Y., Li, J., & Chen, M. (2019). Online Shop Daily Sale Prediction Using Adaptive Network-Based Fuzzy Inference System.

17. Lv, H. R., Bai, X. X., Yin, W. J., & Dong, J. (2008). Simulation based sales forecasting on retail small stores. Proceedings - Winter Simulation Conference, (1968), 1711–1716. https://doi.org/10.1109/WSC.2008.4736257

18. Majhi, B. (2009). Efficient sales forecasting using PSO based adaptive ARMA model. 1333– 1337.

19. Majhi, R., Panda, G., Majhi, B., Panigrahi, S. K., & Mishra, M. K. (2009). Forecasting of retail sales data using differential evolution. 2009 World Congress on Nature and Biologically InspiredComputing,NABIC2009-proceedings,1343–1348. https://doi.org/10.1109/NABIC.2009.5393740

20. Meulstee, P., & Pechenizkiy, M. (2008). Food sales prediction: "If only it knew what we know." Proceedings – IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008, 134–143.https://doi.org/10.1109/ICDMW.2008.128

21. Ohrimuk, E. S., & Razmochaeva, N. V. (2020). Study of Supervised Algorithms for Solve the Forecasting Retail Dynamics Problem. 441–445.

22. Ping, X. (2018). Particle Filter Based Time Series Prediction of Daily Sales of an Online Retailer.

23. Punam, K., Pamula, R., & Jain, P. K. (2018). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018– 2021.

24. Schwenke, C., Ziegenbalg, J., & Dresden, D.-. (2012). Simulation based Forecast of Supermarket Sales Chair forTechnical Information Systems.

25. Singh, M., Ghutla, B., Jnr, R. L., Mohammed, A. F. S., & Rashid, M. A. (2017). Walmart ' s Sales Data Analysis- ABig Data Analytics Perspective. 114–119. https://doi.org/10.1109/APWConCSE.2017.00028

26. Sujatha Kamepalli, Bandaru Srinivasa Rao (2019), International Journal of Innovative Technology and ExploringEngineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-6C2, April 2019

27. Thiesing, F. M., Middelberg, U., & Vornberger, O. (1995). Short term prediction of sales in supermarkets. IEEE International Conference on Neural Networks - Conference Proceedings, 2, 1028–1031. https://doi.org/10.1109/icnn.1995.487562

28. Tsoumakas, G. (2018). A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review. https://doi.org/10.1007/s10462-018-9637-z

29. Wang, J., & Liu, L. (2019). A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry. 317–320.

30. Wei, D., Geng, P., Ying, L., & Shuaipeng, L. (2014). A Prediction Study on E-commerce Sales Based on Structure Time Series Model and Web Search Data. i, 5346–5351.

31. Wu, C. M., Patil, P., & Gunaseelan, S. (2018). Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data.

32. Yang, Y., & Huiyov, C. (2007). S V R mathematical model and methods for sale prediction. 18(4), 769–773