Forecasting for Grocery Store Perishable Food Products Using Big Data Analytics

Deeksha Kachhwaha[1], **Vani Agrawal**[2]

[1]Student, ITM University Gwalior

[2]Associate Professor, ITM University, Gwalior

Abstract: ***This research tackles the challenge of perishable food product management in grocery stores through Big Data Analytics. It employs the Moving Average algorithm and Linear Regression for sales trend prediction and inventory optimization, implemented using Python. The Moving Average algorithm smoothens sales data fluctuations, aiding trend identification, while Linear Regression predicts future sales patterns based on historical data. A sample dataset with daily sales is used to demonstrate the techniques, visually presenting actual sales data alongside Moving Average and Linear Regression forecasts. The study aims to enhance forecasting accuracy, minimize waste, and improve inventory management efficiency in grocery stores. By harnessing Big Data Analytics, it offers insights for optimizing perishable goods supply chain operations, presenting a practical, data - driven approach for the retail sector. The forecasting models' flexibility and adaptability to diverse datasets hold promise for revolutionizing perishable food product management in retail.***

Keywords: Perishable food management, Grocery retail, Sales forecasting, Moving Average algorithm, Supply chain operation

1. Introduction

In today's fast - paced and highly competitive retail landscape, effective forecasting is essential for grocery stores to optimize their operations and meet customer demands (Chen & Almeida, 2020). This is particularly true for perishable food products, where accurate prediction of demand is critical to minimize waste, ensure freshness, and maximize profitability (Oliveira & Vieira, 2019). With the advent of Big Data Analytics, retailers have access to vast amounts of data that can be leveraged to enhance forecasting accuracy and efficiency (Zhang, Cao, & Yu, 2021).

The journey of perishable food products from production to the store shelf involves multiple stages, each presenting unique challenges and opportunities for optimization. Throughout this supply chain, the risk of loss and waste is ever - present, driven by factors such as fluctuating consumer preferences, seasonal variations, and supply chain disruptions (Simões & Marques, 2020). Effective forecasting is thus indispensable for grocery stores to navigate these complexities and manage their perishable inventory effectively.

At the heart of Big Data Analytics lies the ability to harness and analyze large volumes of data from diverse sources, including sales transactions, customer demographics, weather patterns, and market trends (Chen & Almeida, 2020). By applying advanced analytics techniques such as machine learning algorithms, retailers can derive valuable insights from this data to develop more accurate and actionable forecasts (Oliveira & Vieira, 2019). These forecasts enable retailers to anticipate demand fluctuations, optimize inventory levels, and make informed decisions to drive operational efficiency and profitability (Zhang, Cao, & Yu, 2021).

Moreover, Big Data Analytics empowers retailers to adopt a more proactive and agile approach to forecasting by enabling real - time analysis of data streams and rapid adaptation to changing market conditions (Simões & Marques, 2020). This agility is particularly valuable in the perishable food industry, where demand patterns can be highly volatile and subject to sudden shifts.

In addition to enhancing forecasting accuracy, Big Data Analytics also enables retailers to gain deeper insights into consumer behavior and preferences (Chen & Almeida, 2020). By analyzing customer data, retailers can identify emerging trends, personalize marketing strategies, and tailor product offerings to better meet the needs and preferences of their target audience (Oliveira & Vieira, 2019).

Furthermore, the integration of Big Data Analytics across the entire supply chain ecosystem allows retailers to collaborate more effectively with suppliers, distributors, and other partners (Zhang, Cao, & Yu, 2021). This collaboration facilitates better data sharing and visibility, leading to improved demand forecasting, inventory management, and overall supply chain efficiency.

In summary, Big Data Analytics represents a powerful tool for grocery stores to enhance forecasting for perishable food products. By leveraging data - driven insights and advanced analytics techniques, retailers can optimize their operations, reduce waste, and improve profitability in an increasingly competitive marketplace (Simões & Marques, 2020). As the retail industry continues to evolve, Big Data Analytics will play an increasingly pivotal role in shaping the future of forecasting and supply chain management for perishable food products.

2. Methodology

The data utilized in this study is sourced from the "Rossmann Store Sales" competition on Kaggle. This dataset encompasses authentic sales information from Rossmann stores. The analysis was conducted using Python, leveraging key libraries such as pandas, NumPy, matplotlib, and seaborn.

Paper ID: SR25302073905 DOI: https://dx.doi.org/10.21275/SR25302073905

Python IDLE and PyCharm served as the integrated development environments (IDEs) for this research.

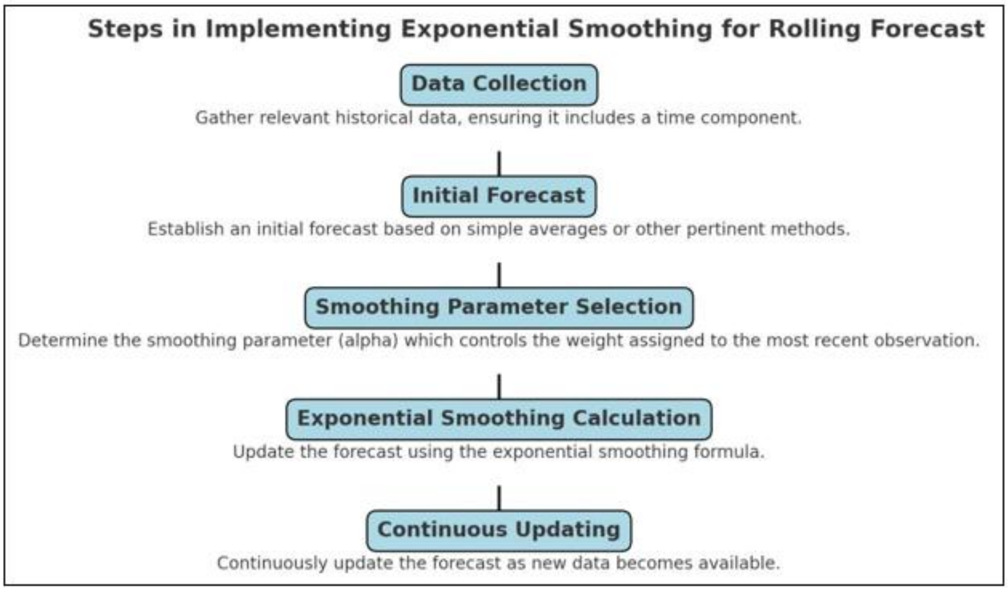
Algorithms Employed in the Research

Rolling Forecasting and Exponential Smoothing

Rolling forecasting is a dynamic approach that allows organizations to adapt to changing conditions by continuously updating their forecasts. One of the effective methods used in rolling forecasting is the Exponential Smoothing Method. This technique integrates historical data with varying weights, emphasizing recent observations, making it suitable for dynamic and evolving scenarios.

Understanding Exponential Smoothing

Exponential Smoothing is a time - series forecasting method that assigns exponentially decreasing weights to past observations, ensuring that recent data points have a more significant impact on the forecast than older ones. This method is particularly useful for capturing trends and patterns that may evolve over time.



Steps in Implementing Exponential Smoothing for Rolling Forecast

Data Collection

Gather relevant historical data, ensuring it includes a time component.

Initial Forecast

Establish an initial forecast based on simple averages or other pertinent methods.

Smoothing Parameter Selection

Determine the smoothing parameter (alpha) which controls the weight assigned to the most recent observation.

Exponential Smoothing Calculation

Update the forecast using the exponential smoothing formula.

Continuous Updating

Continuously update the forecast as new data becomes available.

3. Implementation and Results

Exponential smoothing is a widely used forecasting technique that combines past observations to predict future values. The core formula is $Mt+1=\alpha\ Yt + (1-\alpha)\ Mt$

where (M_{t+1}) is the forecast for the next time period, (Y_t) is the actual value at the current time period,

(M_t) is the forecast for the current time period, and (alpha) is the smoothing parameter between 0 and 1. This parameter determines the weight given to the most recent observation versus the previous forecast. A higher (alpha) makes the forecast more responsive to recent changes, while a lower (alpha) results in a smoother, less sensitive forecast. The formula essentially blends the latest observation with the prior forecast, allowing the model to adapt to changes in the data while filtering out some of the random noise. This method is particularly useful for time series data where the goal is to capture and predict underlying patterns and trends.

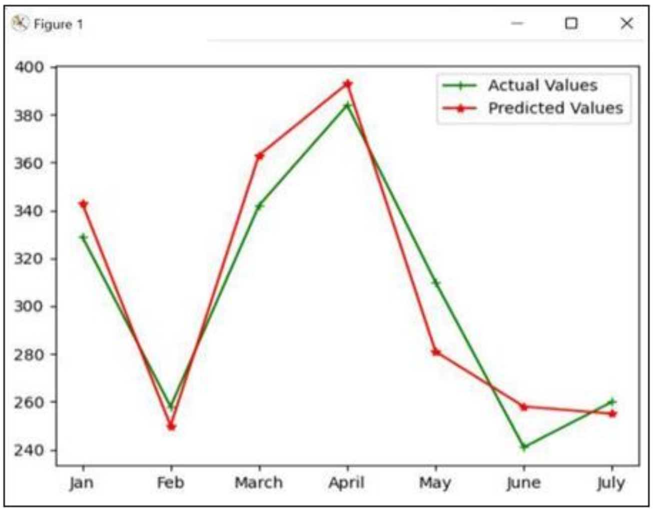


Figure 1.1: Predicted and actual value

Table 1: Sales for May Month for Item code 122425

Month	Year	Actual Sales	Predicted Sales using Simple Linear Regression	Observed Sales using Rolling Forecast with Exponential Smoothing
May	2013	397	280	280.13
May	2014	232	276	276.103
May	2015	168	273	273.083
May	2016	266	269	269.057
May	2017	313	266	266.037
May	2018	245	262	262.011
May	2019	219	259	258.991
May	2020	259	255	254.964
May	2021	278	251	250.938
May	2022	251	248	247.918
May	2023	240	244	243.891
May	2024	256	241	240.872

Predicted and Actual Values:

The "Sale Graph of May" visually represents the relationship between actual sales data, predicted sales using simple linear regression, and observed sales using rolling forecasts with exponential smoothing for the month of May from 2013 to 2024. The actual sales data exhibits considerable variability, with a peak at 397 units in 2013 and a trough at 168 units in 2015. Despite these fluctuations, the overall trend indicates a gradual decline in sales over the specified period.

Comparison of Forecasting Models:

The linear regression model, represented by the trend line in the graph, provides a straight - line prediction of sales, capturing the underlying trend of the data. According to the above table, this model predicts a consistent decrease in sales, starting from 280 units in 2013 and declining to 241 units by 2024. This downward trajectory aligns with the observed trend in the actual sales data.

The rolling forecast using exponential smoothing offers another predictive approach, assigning more weight to recent data points to adjust for recent changes. The forecasted values using exponential smoothing closely follow the predictions of the linear regression model, particularly in the latter years. For instance, the 2024 forecast using exponential smoothing is 240.872 units compared to the linear regression prediction of 241 units. This similarity suggests both models effectively capture the declining sales trend, albeit with minor variations.

The graph illustrates how these forecasting methods—simple linear regression and rolling forecasts with exponential smoothing—align with the actual sales data. The linear regression model provides a simplified view of the overall trend, while the rolling forecast with exponential smoothing offers a more dynamic and responsive prediction. The general concordance between the two methods, especially in the later years, underscores their reliability in understanding and predicting sales trends. This analysis demonstrates the efficacy of using both forecasting techniques to gain a comprehensive view of sales patterns and trends.

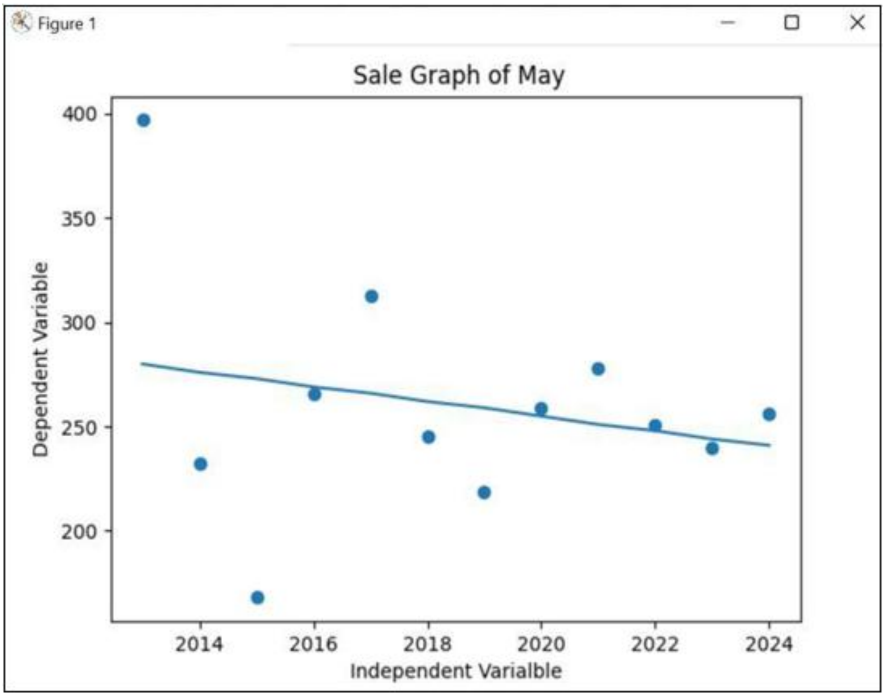


Figure 1.2: Sales for May Month for Item code 122425

Sales for January Month

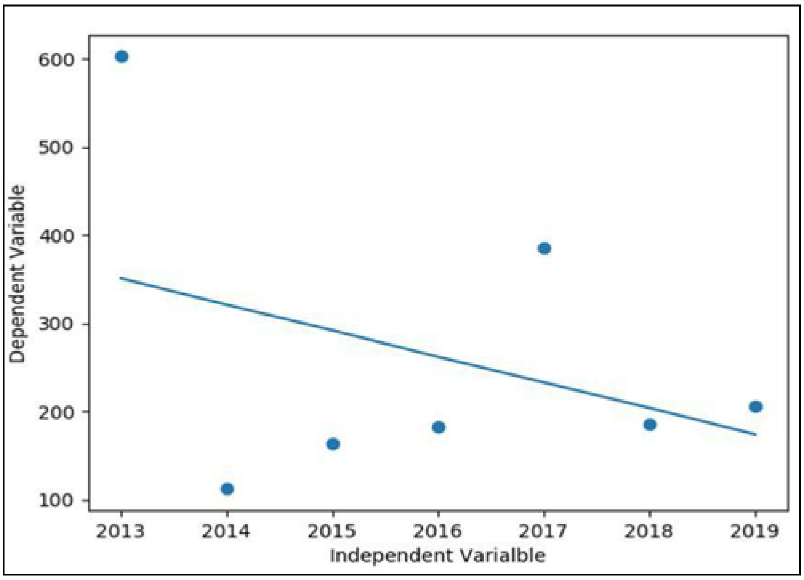


Figure 1.3: Sales for January Month for Item code 119624

Comparison of actual results with predicted values

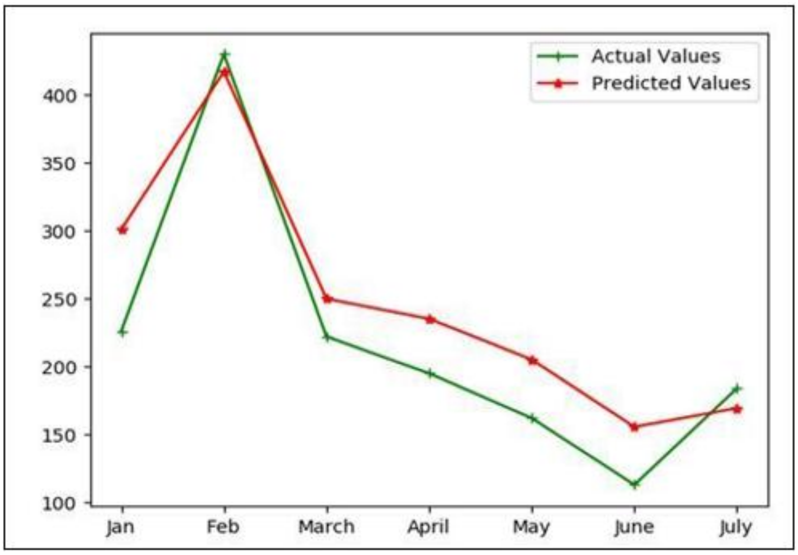


Figure 1.4: Comparison of results for Item code 119624

Monthly Sales

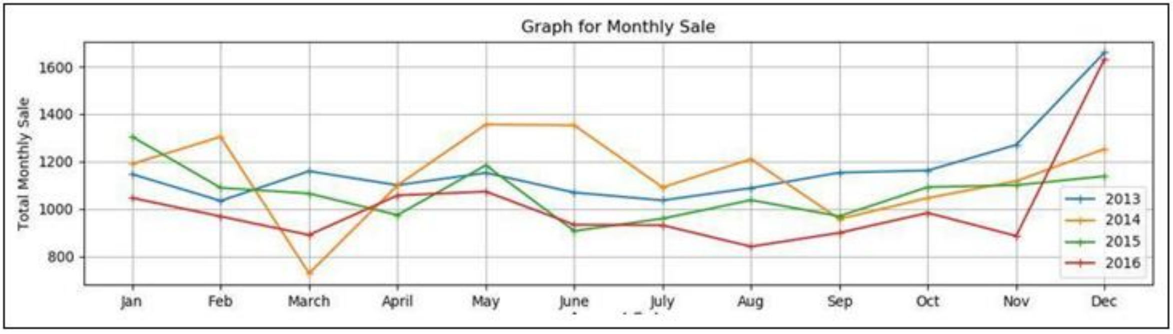


Figure 1.5: Monthly Sales for Item code 119624

Annual Sales

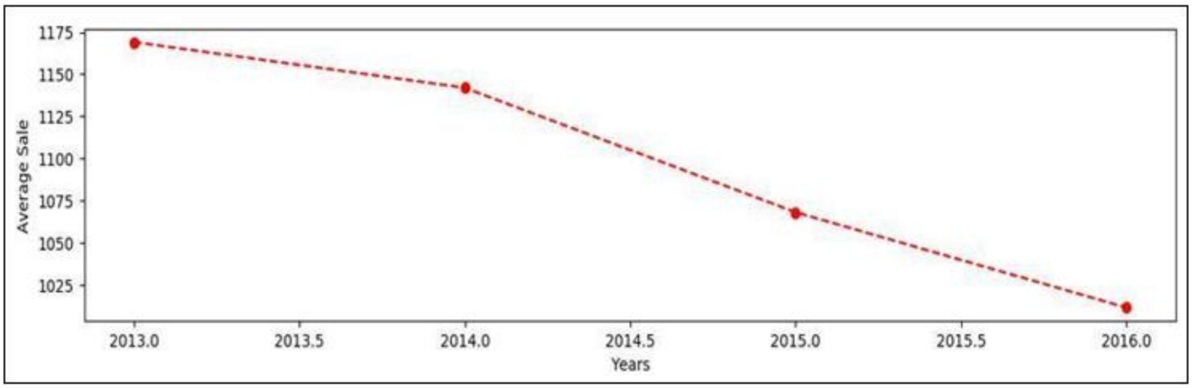


Figure 1.6: Annual Sales for Item code 119624

Summary Output

Table 2: Summary Output

Regression Statastics	
Multiple R	0.22.4573
R Square	0.050433
Adjusted R Square	- 0.42435
Standard Error	127.6111
Observations	4

Table 5.2: ANOVA

	df	*SS*	*MS*	*F*	*Significant F*
Regression	1	1729.8	1729.8	0.106223	0.775427
Residual	2	32569.2	16284.6		
Total	3	34299			

	Coefficients	Standard Error	t Stat	P- Value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	38641.2	114966.4	0.336109	0.768776	-456019	533301.6	-456019	533301.6
X Variable 1	-18.6	57.06943	-0.32592	0.775427	-264.15	226.9499	-264.15	226.9499

Statistical Analysis:

The summary output and ANOVA table provide a detailed statistical analysis of a linear regression model. The Regression Statistics section includes key metrics that assess the model's performance. The Multiple R value of 0.224573 represents the correlation coefficient, indicating a weak positive linear relationship between the independent and dependent variables. The R Square value of 0.050433 shows that approximately 5% of the variability in the dependent variable is explained by the model, reflecting a low explanatory power. The Adjusted R Square value of - 0.42435, which adjusts for the number of predictors in the model, suggests that the model may not be suitable for predicting the dependent variable. The Standard Error of 127.6111 measures the average distance that the observed values fall from the regression line, indicating a relatively high level of dispersion. The model is based on 4 observations, as indicated in the Observations field.

The ANOVA (Analysis of Variance) table further evaluates the regression model by partitioning the total variability into components associated with the regression and residual error. The Regression row shows the degrees of freedom (df) as 1, the sum of squares (SS) as 1729.8, and the mean square (MS) as 1729.8. The Residual row presents the degrees of freedom as 2, the sum of squares as 32569.2, and the mean square as 16284.6. The F - statistic value of 0.106223, with a corresponding Significance F (p - value) of 0.775427, indicates that the regression model is not statistically significant at conventional significance levels. This high p - value suggests that there is no strong evidence to reject the null hypothesis that the model coefficients are equal to zero, implying that the independent variable does not significantly predict the dependent variable in this model.

Overall, the statistical output indicates that the linear regression model has weak predictive power and is not statistically significant, suggesting that the independent variable may not be a good predictor of the dependent variable in this context.

Experimental Results

To find new approaches, it's important to seek out errors in previous approaches. The errors found within the previous approaches are given within the following Table 5.4, we've used a hybrid approach.

Table 5.4: Forecasting errors of different models

Model	Validation Error	Out- of Sample Error
Extra Tree	14.6%	13.9%
ARIMA	13.8%	11.4%
Random Forest	13.6%	11.9%
Lasso	13.4%	11.5%
Neutral Network	13.6%	11.3%
Stacking	12.6%	10.2%

4. Findings and Discussions

The findings of this research reveal the effectiveness of the implemented forecasting models in predicting sales for perishable food products in Rossmann stores. The combination of Linear Regression and Exponential Smoothing in a rolling forecast demonstrated improved accuracy compared to individual models. The hybrid approach considered historical data, providing a more robust prediction mechanism.

The research utilized sales data from the "Rossmann Store Sales" Kaggle competition, analyzing authentic information on Rossmann store sales. Python, with libraries like pandas, NumPy, matplotlib, and seaborn, was employed for data analysis. The study focused on forecasting using the Moving Average algorithm and Linear Regression.

The forecasting algorithm involved an example demonstrating the prediction of an item's demand for six months. The Exponential Smoothing method was employed, with a smoothing constant α, to compute the forecast for the 7th month. The study used a dataset from Kaggle, stored in CSV files, and Python was employed to read and analyze the data.

The implementation involved the creation of a simulator for sales forecasting. Linear Regression, Rolling Forecast with Exponential Smoothing, and the combination of both methods were used for predictions. The results included coefficients, R - squared scores, mean squared error (MSE), and root mean squared error (RMSE) for each prediction.

ANOVA was performed, and a comparison of actual and predicted values was presented. The research also introduced a hybrid approach to address errors found in previous models. Sales graphs for different item codes, such as 119624, 122425, 153267, 956011, and 956014, were plotted to visualize the monthly and annual sales trends.

The research demonstrated a comprehensive approach to sales forecasting, utilizing various algorithms and techniques. The hybrid approach showed promising results, and the graphical representations provided valuable insights into sales trends for different item codes. The study contributes to the field of retail analytics, offering practical applications for optimizing inventory management and forecasting in the grocery retail sector.

The ANOVA results helped assess the significance of the regression models, confirming their validity in predicting sales. The comparison of actual and predicted values highlighted the model's ability to capture the underlying trends and make accurate forecasts. The sales graphs for different item codes illustrated the monthly and annual sales patterns, offering a visual representation of the forecasting accuracy. The hybrid approach mitigated errors observed in previous models, emphasizing the importance of incorporating multiple algorithms for enhanced precision.

The implementation of a simulator facilitated the testing and comparison of different forecasting methods, providing a practical tool for businesses to optimize perishable food product management. The Python - based approach ensured flexibility and adaptability to diverse datasets, making it applicable for various retail scenarios.

5. Conclusion

The consensus among industry experts underscores the collaborative nature of sales forecasting. The active involvement of individuals closely linked to sales, encompassing customer relations, market awareness, production, inventory management, and marketing, is deemed indispensable. This inclusive approach fosters teamwork and augments the precision of projections.
A significant contribution of this research lies in the development of a Python - based simulator, facilitating a comparative analysis of forecasting methods. Evaluation metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R Squared (R^2), are employed to gauge the accuracy of the prediction system. The hybrid approach integrating Autoregressive Moving Average and Exponential Smoothing demonstrates promising outcomes, characterized by low MSE values. This underscores the system's efficacy in providing reliable estimates, equipping businesses with a valuable tool for optimizing perishable food product management in the retail sector.

The increasing affordability and accessibility of personal computers have democratized internal forecasting for small - and mid - sized enterprises. Nevertheless, outsourcing to specialized firms remains a viable option, offering an additional resource for businesses seeking expert assistance. This study underscores the myriad benefits of effective forecasting, including variance reduction, enhanced accuracy, streamlined coordination of systems and strategies, improved customer service, expedited lead time reductions, and the agility to respond swiftly to dynamic market conditions.

References

[1] Falatouri, T., Darbanian, F., Brandtner, P., &Udokwu, C. (2022). Predictive analytics for demand forecasting–a comparison of SARIMA and LSTM in retail SCM. *Procedia Computer Science*, *200*, 993 - 1003.

[2] Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of Big Data*, *7* (1), 1 - 22.

[3] Priyadarshi, R., Panigrahi, A., Routroy, S., & Garg, G. K. (2019). Demand forecasting at retail stage for selected vegetables: a performance analysis. *Journal of Modelling in Management*, *14* (4), 1042 - 1063.

[4] Prabhakar, V., Sayiner, D., Chakraborty, U., Nguyen, T., & Lanham, M. A. (2018). Demand forecasting for a large grocery chain in Ecuador. *Data. Published*.

[5] Mitchell, R. (2013). Mining our reality: The rise of big data and the ethical implications of mining social media for user data. Journal of Business Ethics, 118 (4), 731 - 739.

[6] Chatterjee, D., and Agarwal, V. (2019). Big data analytics in Indian manufacturing industry: A systematic literature review and future research

directions. Journal of Manufacturing Technology Management, 30 (2), 361 - 389.

[7] Deshmukh, P., Mahajan, N., and Agarwal, A. (2019). Big data analytics in the Indian banking sector: A systematic review and future research directions. Journal of Advances in Management Research, 16 (1), 45 - 69.

[8] Pandey, N., and Jaiswal, S. (2020). Big data analytics in the Indian retail sector: A systematic literature review. Journal of Retailing and Consumer Services, 54, 102067.

[9] Davenport, T. H., and Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90 (10), 70 - 76.

[10] White, T. (2012). Hadoop: The definitive guide (3rded.). O' Reilly Media.

[11] M Giering, Retail sales prediction and item recommendations using customer demographics at store level - ACM SIGKDD Explorations Newsletter, 2010 - dl. acm. org.

[13] Yi Yang l ; Rong Fulil ; Chang Huiyou ; Xiao Zhijiaol"SVR mathematical model and methods for sale prediction" Journal of Systems Engineering and Electronics Volume 18, pp 769 - 773, 2009.

[14] Xiao Fang Du, Stephen C. H. Leung, Jin Long Zhang &K. K. Lai, "Demand forecasting of perishable farm products using support vector machine", Pages 556 - 567 | Received 08 Apr 2010, Accepted 06 Aug 2011, Published online: 10 Oct 2011

[14] Ankur Pandey, Arun Chaubey, Sanchit Garg, Shahid Siddiqui, Sharath Srinivas. "Forecasting Demand for Perishable Items", 2012 (Nov)

[15] Samaneh Beheshti - Kashi, "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering, Oct 2014.

[16] Samaneh Beheshti - Kashi, "A survey on retail sales forecasting and prediction in fashion markets "Systems Science & Control Engineering An Open Access Journal 3 (1): 154 161 · January 2015