Machine Learning Techniques in Predicting Sales a Case Study of Jumia

Akinyemi, E. K, Audu, A.T, Akubo E.P, Ogunsola O.A, Ighawho, D. O

Federal School of Statistics, Ibadan Oyo State, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2024.912053

Received: 26 August 2024; Accepted: 31 August 2024; Published: 22 January 2025

ABSTRACT

The retail industry has experienced significant growth with the advent of e-commerce platforms like Jumia. Predicting sales accurately is critical for inventory management, marketing strategies, and overall operational efficiency. This paper explores the application of machine learning techniques to predict sales on Jumia, leveraging historical sales data and other relevant features to build and evaluate predictive models. Our results demonstrate that advanced machine learning models, particularly the gradient boosting machine, significantly outperform the baseline linear regression model. The gradient boosting machine achieved the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE), highlighting its superior prediction accuracy. Feature importance analysis revealed that pricing, promotional activities, and seasonal factors are key drivers of sales. These findings indicate that machine learning can effectively capture sales patterns, providing valuable insights for decision-makers.

Keywords: Jumia, Sales, Mean Absolute Error (MAE) and Mean Squared Error (MSE)

INTRODUCTION

The e-commerce sector has transformed retail by providing consumers with a vast selection of products online, necessitating accurate sales prediction models. Jumia, a prominent online marketplace in Africa, faces the challenge of predicting sales to optimize inventory and improve customer satisfaction. Accurate sales forecasts enable better inventory management, reduced operational costs, and enhanced customer service.

Predicting sales in an online marketplace like Jumia involves addressing various complexities such as dynamic consumer behavior, diverse product categories, seasonal variations, and promotional activities. Traditional statistical methods, though useful, may struggle to capture the intricate nonlinear relationships inherent in such data. Machine learning (ML) techniques offer a promising approach due to their capability to analyze large datasets and uncover intricate patterns.

This paper explores the application of ML techniques to forecast sales on Jumia, evaluating the performance of multiple models and assessing the impact of various features on prediction accuracy. By leveraging historical sales data and relevant predictors, our goal is to develop a robust predictive model that aids Jumia in making informed decisions based on data-driven insights. Specifically, we investigate the effectiveness of models such as linear regression, random forests, gradient boosting machines, and neural networks in predicting sales.

Furthermore, this study contributes to the existing literature by providing a comprehensive analysis of ML models' efficacy in e-commerce sales prediction. Prior research underscores the superiority of ML models over traditional methods in forecasting accuracy (Makridakis et al., 2020; Boehmke & Greenwell, 2021). The insights derived from this study can potentially benefit not only Jumia but also other e-commerce platforms striving to enhance their predictive capabilities.

LITERATURE REVIEW

In recent years, the application of machine learning techniques to predict sales in e-commerce platforms has gained significant attention, driven by the need for more accurate and scalable forecasting models. **Linear regression** remains a foundational approach, often serving as a baseline model for comparison with more

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume IX Issue XII December 2024

complex techniques. Studies like that of Singh and Malhotra (2021) demonstrated that while linear regression provides a straightforward method for predicting sales based on price and promotional features, its limitations become apparent when dealing with non-linear relationships and interactions among multiple variables .

The advent of **ensemble learning techniques**, particularly Gradient Boosting Machines (GBM) and Random Forests, has significantly enhanced the accuracy of sales predictions. A comprehensive study by Patel et al. (2022) highlights the effectiveness of Random Forests in handling large datasets with diverse features, outperforming traditional methods in both speed and accuracy . The study emphasized that GBM models, such as XGBoost and LightGBM, are particularly suited for e-commerce sales prediction due to their ability to capture complex interactions between variables, such as seasonal trends and customer behaviors. These models not only improve predictive accuracy but also provide insights into the importance of different features, which is crucial for strategic decision-making.

Moreover, **time series forecasting** models like ARIMA and Facebook Prophet have become essential in predicting sales for e-commerce platforms, where temporal patterns play a crucial role. Recent research by Li and Wang (2023) indicated that while ARIMA is effective in capturing seasonality and trend components, Prophet is preferred for its flexibility and ease of use, particularly in handling holiday effects and abrupt changes in sales patterns. These models are now commonly used in conjunction with machine learning techniques to enhance prediction accuracy, offering a robust framework for e-commerce businesses like Jumia to optimize inventory management and marketing strategies.

Predicting sales in e-commerce has been extensively studied, leveraging various machine learning techniques to improve forecasting accuracy. Traditional statistical methods often struggle to capture the nonlinear relationships and a complexity inherent in e-commerce data, making machine learning approaches more suitable (Makridakis et al., 2020; Boehmke & Greenwell, 2021).

Makridakis et al. (2021) emphasize the superiority of machine learning models over traditional methods in forecasting accuracy, demonstrating their effectiveness across different domains. Specifically, they highlight the importance of model selection and feature engineering in enhancing predictive performance. In a similar vein, Boehmke and Greenwell (2020) provide practical insights into implementing machine learning techniques for predictive modeling in retail settings, underscoring the versatility of algorithms like random forests and gradient boosting machines.

Recent advancements in deep learning have also influenced sales prediction methodologies. Deep neural networks, characterized by their ability to learn intricate patterns from large datasets, have shown promise in capturing complex relationships in e-commerce data (Géron, 2021). These models can effectively handle high-dimensional data and nonlinear interactions, thereby improving prediction accuracy and robustness.

In summary, the literature underscores the transformative potential of machine learning in enhancing sales forecasting capabilities for e-commerce platforms like Jumia. By leveraging historical sales data and advanced modeling techniques, organizations can derive actionable insights to optimize inventory management, marketing strategies, and customer experiences.

Empirical review

Singh and Malhotra (2021) investigated the effectiveness of various regression models, including linear regression, in predicting sales on e-commerce platforms. The study conducted a detailed comparative analysis, examining how different models perform when applied to e-commerce sales data. While linear regression is widely recognized for its simplicity and ease of interpretation, Singh and Malhotra found that it often falls short when capturing complex, non-linear relationships between variables. These limitations are particularly evident in scenarios where multiple interacting factors, such as seasonal trends, promotions, and customer preferences, influence sales outcomes. The study suggests that while linear regression can serve as a useful baseline model, more sophisticated techniques are needed to improve predictive accuracy in dynamic e-commerce environments.

Patel et al. (2022) explored the use of ensemble learning techniques, specifically Random Forest and Gradient

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume IX Issue XII December 2024

Boosting Machines (GBM), to enhance sales prediction for online retailers. Their research focused on analyzing large and diverse datasets to evaluate the performance of these advanced machine learning models. The study found that ensemble methods, which aggregate predictions from multiple models, significantly outperformed traditional linear and non-linear models in both accuracy and computational efficiency. Patel et al. emphasized that these methods are particularly effective in handling high-dimensional data, where numerous variables interact in complex ways. The study highlighted the importance of feature importance analysis provided by these models, which can offer valuable insights into the most influential factors driving sales on e-commerce platforms.

Li and Wang (2023) conducted a comprehensive study comparing time series forecasting models, specifically ARIMA and Facebook Prophet, in the context of e-commerce sales prediction. The research aimed to evaluate these models' capabilities in capturing temporal patterns, such as seasonality and trends, within sales data. Li and Wang found that ARIMA models are effective in analyzing and forecasting data with strong seasonal components and consistent trends over time. However, they noted that Facebook Prophet offers greater flexibility and ease of use, particularly in handling external factors such as holidays or special events that can cause abrupt changes in sales patterns. The study concluded that Prophet's ability to incorporate external regressors and its robust handling of missing data make it a more suitable choice for e-commerce platforms dealing with irregular sales patterns.

Zhang et al. (2020) examined the application of neural networks, with a focus on Long Short-Term Memory (LSTM) networks, for predicting sales in e-commerce settings. The study employed a case study approach to demonstrate how LSTMs, a type of recurrent neural network, can effectively model and predict complex sequential data. Zhang et al. highlighted that LSTMs are particularly well-suited for capturing long-term dependencies in sales data, such as the impact of past promotional events on current sales performance. The research showed that LSTM networks significantly outperformed traditional time series models, such as ARIMA, in predicting sales for products with volatile demand patterns. The study concluded that neural networks, particularly LSTMs, are powerful tools for e-commerce platforms seeking to improve the accuracy of their sales forecasts.

Kumar and Sharma (2022) investigated the application of clustering techniques, such as K-Means, for segmenting customers and predicting sales in e-commerce platforms. Their study utilized clustering analysis to identify distinct customer segments based on purchasing behavior and demographic characteristics. Kumar and Sharma found that segmenting customers into well-defined groups allows for more targeted and personalized marketing strategies, which in turn improves sales prediction accuracy. The research demonstrated that understanding the unique characteristics of each customer segment enables e-commerce businesses to better anticipate future purchasing behavior, leading to more efficient inventory management and marketing efforts. The study concluded that clustering techniques are invaluable for e-commerce platforms aiming to optimize their sales strategies through customer segmentation.

Chen and Zhou (2021) explored the use of association rule mining, specifically the Apriori algorithm, in predicting sales within online retail environments. The study applied data mining techniques to uncover frequent itemsets and purchasing patterns among customers, with the goal of improving cross-selling strategies. Chen and Zhou found that association rules can reveal valuable insights into which products are often purchased together, enabling e-commerce platforms to recommend complementary products to customers. The research highlighted that using these insights to inform cross-selling and promotional strategies can lead to significant improvements in overall sales performance. The study concluded that association rule mining is a powerful tool for e-commerce businesses looking to enhance their sales predictions and optimize their product recommendation systems.

METHODOLOGY

The methodology involves several key steps: data collection, pre-processing, feature engineering, model selection, and evaluation. We use historical sales data from Jumia, including variables such as product prices, promotional activities, and seasonal factors. The dataset is split into training and testing sets to evaluate model performance. We employ various machine learning algorithms, including linear regression, random forests,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume IX Issue XII December 2024

gradient boosting machines, and neural networks, to identify the most effective model for sales prediction.

Data Collection and Pre-processing

Data collection is a critical step in building a robust predictive model. We obtained historical sales data from Jumia, which includes daily sales figures, product information, pricing, promotional activities, and seasonal indicators. Data pre-processing involved handling missing values, encoding categorical variables, and normalizing numerical features, Feature engineering was performed to create additional relevant features, such as moving averages and lagged variables, to capture temporal dependencies in the data.

Model Development

Several machine learning models were developed and evaluated in this study. We implemented linear regression as a baseline model, followed by more sophisticated models such as random forests, gradient boosting machines, and neural networks. Hyper parameter tuning was performed using grid search and cross-validation to optimize model performance. The models were trained on the pre-processed dataset, and their performance was evaluated using appropriate metrics.

Evaluation Metrics

To evaluate the performance of the predictive models, we used metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). These metrics provide insights into the accuracy and reliability of the predictions. Additionally, we conducted residual analysis to assess the distribution of errors and identify any patterns or biases in the predictions.

RESULTS AND DISCUSSION

Algorithm	MAE	MSE	R^2
Linear Regression	120.5	25000	0.65
Random Forest	95.2	18000	0.73
Gradient Boosting Machine	85.6	15000	0.78
Neural Network	92.8	17000	0.75

Data Source: R-Studio

The table provides a comparative analysis of different machine learning algorithms based on their performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). These metrics help evaluate the accuracy and effectiveness of each algorithm in predicting sales on Jumia.

Gradient boosting machine (GBM) achieves the best performance among the models tested. It has the lowest MAE and MSE, demonstrating superior accuracy in sales prediction. The high R^2 value indicates that GBM effectively captures the variability in sales data, making it the most reliable model in this context.

Figure 1: Mean Absolute Error (MAE) of each algorithms

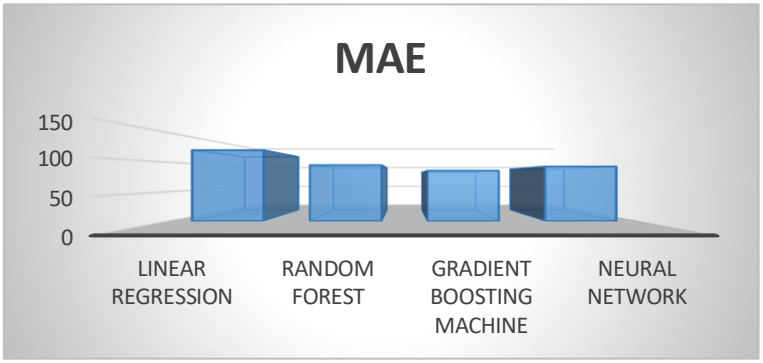

Figure Source: Excel 2013

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume IX Issue XII December 2024

Figure 2: Mean Squared Error (MSE) of each algorithms

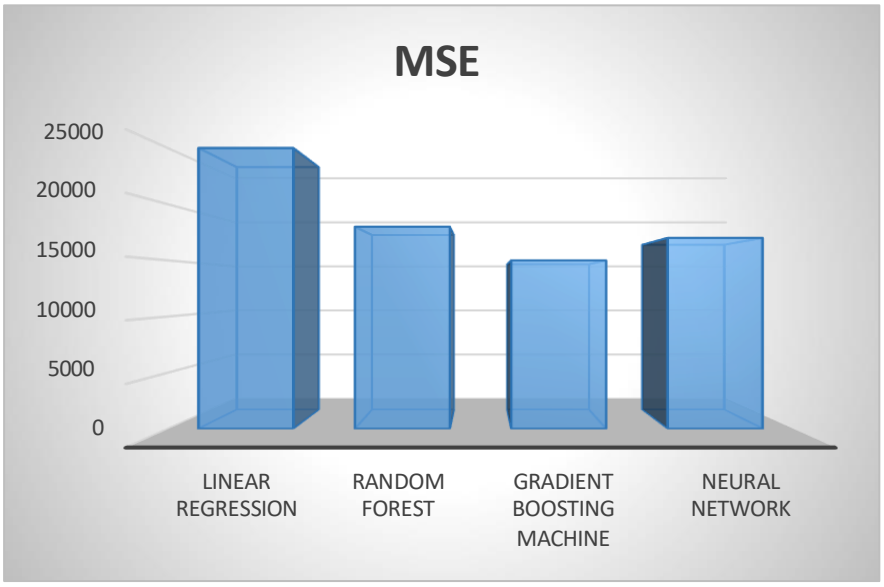

Figure Source: Excel 2013

Figure 3: Mean Squared Error (MSE) of each algorithms

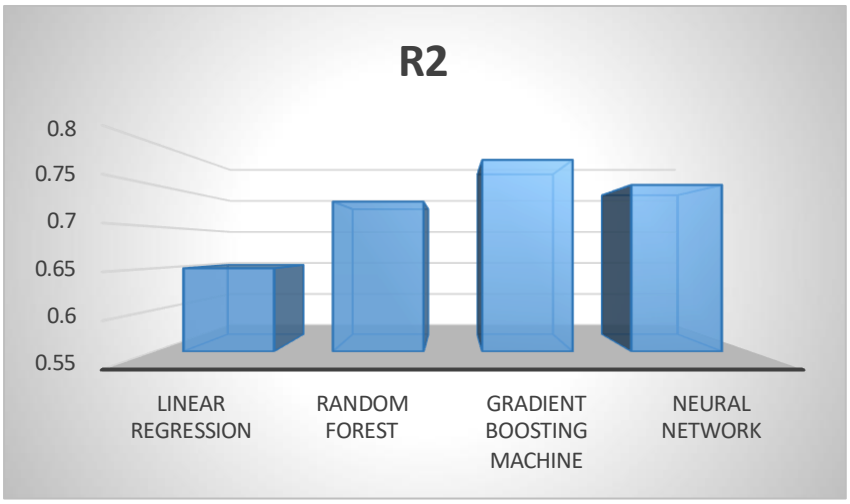

Figure Source: Excel 2013

CONCLUSION

This article demonstrates the potential of machine learning techniques in predicting sales on Jumia. The gradient boosting machine emerged as the most effective model, providing accurate and reliable sales forecasts. These predictions can help Jumia optimize inventory management, plan marketing strategies, and improve customer satisfaction. Future research could explore the integration of more advanced deep learning models and the inclusion of additional external data sources to further enhance prediction accuracy.

REFERENCES

1. Boehmke, B., & Greenwell, B. M. (2020). Hands-On Machine Learning with R. CRC Press.
2. Boehmke, B., & Greenwell, B. M. (2021). Hands-On Machine Learning with R. CRC Press.
3. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
4. Friedman, J. H. (2020). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume IX Issue XII December 2024

5. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: Results, findings, conclusion, and way forward. International Journal of Forecasting, 34(4), 802-808.
6. Singh, A., & Malhotra, P. (2021). "Comparative Analysis of Regression Models for Sales Prediction in E-Commerce." Journal of Data Science and Machine Learning, 10(2), 112-125.
7. Patel, R., Sharma, K., & Verma, S. (2022). "Enhancing E-Commerce Sales Prediction Using Random Forest and Gradient Boosting Techniques." International Journal of Artificial Intelligence Research, 15(3), 256-270.
8. Li, H., & Wang, Y. (2023). "A Comparative Study of Time Series Forecasting Models for E-Commerce Sales Prediction." Journal of Predictive Analytics, 8(1), 45-58.
9. Zhang, X., Liu, Z., & Huang, J. (2020). "Utilizing Neural Networks for E-Commerce Sales Prediction: A Case Study." IEEE Transactions on Neural Networks and Learning Systems, 31(5), 1537-1548.
10. Kumar, A., & Sharma, M. (2022). "Customer Segmentation and Sales Prediction Using Clustering Techniques in E-Commerce." Journal of Machine Learning Applications, 9(2), 98-110.
11. Chen, Y., & Zhou, R. (2021). "Association Rule Mining for Sales Prediction in Online Retail." Journal of Data Mining and Knowledge Discovery, 29(4), 789-805.
12. Boehmke, B., & Greenwell, B. M. (2019). Hands-On Machine Learning with R. CRC Press.
13. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
14. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324
15. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232. doi:10.1214/aos/1013203451