Advanced Performance Analysis in Sport Footwear Sales Prediction Using Machine Learning

Ammar Al-Bazi[1], Qusay H. Al-Salami[2], Nauval Zulfikar[1], Zina Jerjees[3], Mahmood A. Ahmad[4]

[1]Operations and Information Management Department, Aston Business School, Aston University, UK

[2]Department of Business Administration, College of Administrative and Financial Sciences, Cihan University-Erbil, Kurdistan Region, Iraq

[3]Department of Computer Science and Creative Technologies, School of Computing and Creative Technologies University of the West of England, UK

[4]Dubai Business School, University of Dubai, UAE

***Abstract*—This research uses advanced machine learning techniques to analyse and predict sales trends in the sports footwear industry. It focuses on XGBoost, a state-of-the-art approach recognised for its predictive capabilities. The study aims to uncover critical performance drivers, enhance marketing strategies, and optimise sales efficiency. A thorough review of machine learning applications in sales forecasting within the footwear industry emphasises practical algorithms, innovative feature engineering, and external data integration to improve accuracy. Rigorous evaluation and comparison highlight XGBoost's superior performance over Random Forests and Gradient Boosting, surpassing traditional forecasting methods in predictive accuracy. The research delivers actionable insights into sales performance across regions, product categories, and channels, empowering strategic inventory management and marketing decisions. By showcasing the potential of advanced analytics, this study significantly contributes to retail sales forecasting and provides a robust framework for future research and industry applications.**

***Keywords* - Sales forecasting, Sports footwear industry, Machine learning, XGBoost, Predictive analytics**

I. INTRODUCTION

The sportswear industry is highly competitive and rapidly evolving, with substantial growth and transformation in recent years. Adidas aims to maintain market leadership in changing consumer preferences and dynamic market conditions as a key player. Accurate sales prediction and performance analysis are critical for strategic decision-making and sustaining a competitive edge (Kharfan & Chan, 2018, Massoudi et al., 2024).

Accurate demand prediction for new products is critical due to high investment levels and significant risks in competitive markets with evolving customer expectations and technological advancements. Traditional methods include market research, buyer surveys, expert opinions, diffusion models like Bass, and statistical models. Recently, machine learning has emerged as a superior option, especially when sample sales data on similar products exists. Machine learning offers faster, more precise predictions than human approaches, addressing challenges like sales loss from underestimation or costly surplus from overestimation (Hamoudia & Vanston, 2023).

Advances in big data and machine learning have revolutionised business operations, enabling valuable insights and data-driven decisions. Machine learning, a subset of AI, excels at identifying complex patterns in large datasets, offering precise predictions (Dhankhar et al., 2024, Sun, 2024). Applied extensively in retail and fashion, random forests and gradient boosting handle non-linear relationships, making them suitable for complex sales forecasting scenarios (Al-Salami et al., 2022). Integrating external data sources, such as social media trends and customer sentiment, further enhances predictive accuracy (Choi & Varian, 2012). Soltaninejad et al. (2024) introduced a sales forecasting algorithm leveraging AI-driven machine learning, combining time-series modelling, neural networks, random forests, and technical analysis. The approach improves forecast accuracy, optimising production, capital allocation, and business success (Abbas & Al-Salami, 2023; Mahmood et al, 2022).

This study applies XGBoost as one of the most promising machine learning approaches to predict Sports footwear products, focusing on Adidas shoe sales as an example, uncovering patterns, identifying key performance drivers, and developing predictive models to support inventory management and marketing strategies. It follows a comprehensive methodology, including data collection, pre-processing, exploratory analysis, feature engineering, model development, and evaluation, providing actionable insights for improved decision-making and sales performance. The model evaluation phase also uses other predictive models like Linear Regression, Random Forest, and Gradient Boosting. The model evaluation focuses on metrics like RMSE, MAE, and R^2, using k-fold cross-validation to ensure robustness, selecting the most reliable model for implementation.

II. LITERATURE REVIEW

In recent years, machine learning in retail and sales prediction has become increasingly popular, offering a new way out of the challenge of demand prediction in dynamic markets. Traditional methods, including statistical models and expert opinions, cannot handle the complexity of the modern retail environment, where a wide range of factors such as seasonality, promotions, and external data sources influence consumer behaviour. The ensemble and hybrid models of machine learning algorithms have been exceptionally promising in finding complex patterns and giving more accurate predictions (Nawkhass & Birdawod, 2017). This section will review applications of machine learning techniques for sales forecasting in the footwear and retail industries. This work includes but is not limited to Kharfan

DOI: http://doi.org/10.24086/icafs2025/paper.1582

and Chan (2018), who highlighted the effectiveness of machine learning in enhancing demand and sales of footwear products forecasting, proposing two frameworks including a general model using algorithms like regression trees, random forests, and neural networks, and a three-step model combining clustering and classification for similar product identification. In contrast, Dhankhar et al. (2024) integrated external data sources such as Google Trends and customer satisfaction indices with traditional sales data to predict Nike's sales. Employing ARIMA, exponential smoothing, and random forests, the study underscored random forests' superiority and the value of external data in predictive modelling.

Recent machine learning developments have revolutionised sales prediction by allowing the identification of complex patterns and reduced prediction errors. These developments include XGBoost, FB-Prophet, and ARIMA, which are techniques used in analysing sales data to show their capabilities of uncovering hidden risks and trends. For example, Malik et al. (2024) showed how these models can minimise inventory costs, improve planning, and enhance profitability in dynamic retail environments.

With the integration of external data sources, prediction models have become even more enriched, allowing for more accurate and context-aware predictions. Choi and Varian (2012) used Google Trends to show how machine learning models, such as Ridge and Bayesian Regression, can significantly improve retail sales prediction compared to traditional methods. Similarly, Xia et al. (2012) pointed out that ensemble techniques like random forests and gradient boosting outperformed other techniques by making plausibly effective use of customer attributes and product data. As Swaminathan and Venkitasubramony (2024) emphasised, hybrid models also play an important role in addressing uncertainties and achieving higher accuracy in volatile markets. Lourenço et al. (2024) combined the moving averages and XGBoost approaches to enhance the prediction for SMEs in the fashion sector. The approach enhanced the accuracy and supported better inventory management, sustainability, and brand equity.

The blending of qualitative and quantitative methodologies is emerging as another crucial ingredient of reliable prediction. In the same light, Gilliland (2010) supported integrating experts' judgment with advanced analytics to maintain agility in uncertainties. Second, the inclusion of advertising and promotional variables, as also corroborated by Holtrop et al. (2017), dramatically improves the predictability of the machine learning models toward yielding better strategic decisions.

Deep learning has opened a box of great potential to tackle the dynamic environment of retailing. Loureiro et al. (2018) investigated whether deep neural networks capture non-linear relationships in sales data more effectively and found that they outperformed traditional time-series models. Ahmed et al. (2010) compared different machine learning methods and concluded that models like neural networks and support vector machines are superior to traditional statistical methods if appropriately tuned. Athanasopoulos et al. (2023) also showed that machine learning performs better in time-series prediction for various industries.

Hybrid deep-learning models have been particularly effective in the sportswear industry. Xu et al. (2021) developed such a model by integrating MLP and CNN to predict sales for offline stores in China more accurately than traditional methods. Such models have supported production planning and inventory management in long-lifecycle markets.

Effective feature engineering and model selection are still crucial for the effectiveness of any prediction method. Bontempi et al. (2013) realised this and provided various feature extraction and model evaluation methods to enhance accuracy. On the same note, Makridakis et al. (2018) warned that serious tuning must be carefully done for machine learning to realise its full potential without overfitting. Goodfellow et al. (2016) provided foundational insights into neural network optimisation, reinforcing the importance of these techniques in predictive analytics.

The final insight has been into the increasing exploration of machine learning for optimising commercial strategies in the footwear industry. Zeynali and Albadvi (2023) illustrated how machine learning and revenue management are used to create effective promotional strategies that result in a considerable increase in revenue for brands such as Adidas and West Gear compared to traditional methods of offering discounts.

Despite significant advancements in machine learning for retail and sales prediction, its application to footwear sales prediction remains challenging and underexplored (Abed and Al-Salami, 2021). The above research has largely focused on broader retail sectors or specific industries such as apparel and sportswear, often neglecting the unique challenges of footwear sales prediction. Footwear, characterised by diverse styles, sizes, and seasonal demand variations, presents complexities that differ from other retail categories. While studies like those of Loureiro et al. (2018) and Lourenço et al. (2024) have highlighted the potential of machine learning in prediction for fashion and small-to-medium enterprises, few have delved deeply into the granular dynamics of footwear manufacturing and inventory management. Additionally, integrating external data sources, such as consumer sentiment or market trends, in predictive modelling for footwear production is scarce. This gap underscores the need for specialised research and an appropriate machine learning approach to optimise sales prediction accuracy and operational efficiency in the footwear industry, using advanced machine learning techniques to address its distinct requirements (Jameel and Al-Salami, 2023; Birdawod, 2022).

III. XGBOOST MACHINE LEARNING APPROACH

XGBoost stands for Extreme Gradient Boosting. This research is chosen as the main machine learning approach because of its great efficiency, flexibility, and ability to handle classification and regression problems equally. Unlike traditional algorithms, XGBoost is built on gradient boosting principles, incorporating advanced regularisation strategies, such as L1 and L2 penalties, optimisation techniques, and parallelised computations. These features make it particularly well-suited for analysing complex datasets with non-linear patterns and large volumes of data, such as retail sales data. XGBoost is unique in that it iteratively improves its predictions by addressing the residual errors of previous iterations, thus enabling it to capture intricate relationships between features and target variables. This capability ensures highly accurate predictions while avoiding overfitting, as demonstrated by Lourenço et al. (2024), who have shown the effectiveness

of XGBoost in enhancing sales prediction accuracy for SMEs in the fashion industry. Its computational efficiency and scalability further enhance its suitability for real-world applications, making it an ideal choice for predicting Adidas footwear sales and extracting actionable insights.
Another strong point of XGBoost is its computation efficiency. Using parallel process computation and optimised data structure, it can deal with voluminous data with complex computational calculations at a remarkable speed. Again, it contains built-in handling for missing values, which is practical in realistic applications. Its scalability and robustness make it an excellent fit for various prediction problems where interacting multiple variables modelling is extremely useful.
Therefore, XGBoost is used to predict sales performance because it outperforms others in handling non-linear data and features an important ranking with the assurance of producing good results even if the datasets are extensive and diverse. Its robust framework provides an excellent foundation to generate actionable insights, especially for industries like retail, which require accuracy in forecasts as inputs for decision-making and strategy formulation.

A. Mathematical Background

XGBoost is a tree-based ensemble learning algorithm that extends the basic ideas of gradient boosting. Known for its regularisation and computational efficiency, it captures complex, non-linear relationships and handles feature-rich datasets effectively (Chen and Guestrin, 2016, Al-Bazi et al., 2023). At the core, Gradient Boosting is an iterative method that builds decision trees one by one, and each subsequent tree tries to improve the mistakes of the former ones. The final model is thus an ensemble of those trees; it makes predictions based on their combined outputs. XGBoost is particularly powerful because it optimises this process through regularisation, parallel computation, and an emphasis on computational efficiency.
The objective function of XGBoost is the most crucial factor in its performance. It incorporates a loss function, which measures the difference between the predicted and actual values, with a regularisation term that penalises model complexity. It is expressed as:

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here:

- $L(y_i, \hat{y}_i)$ is the loss function- commonly Mean Squared Error for regression tasks, which wants the model to predict as accurately as possible.
- $\Omega(f_k)$ is a regularization term penalizing the model for an overly complex model to avoid overfitting.
- f_k is the k-th decision tree in the model ensemble.

The regularisation term, $\Omega(fk)$, can include both $L1$ (lasso) and $L2$ (ridge) penalties, allowing XGBoost to control the complexity of the model by shrinking coefficients or completely excluding irrelevant features. This balances bias and variance, enabling the model to generalise unseen data effectively. A key feature of XGBoost is the ability to split nodes in each decision tree using a greedy algorithm. This splitting is driven by an objective function, which evaluates potential splits and how well they could contribute to reducing prediction errors. This approach ensures that each split minimises the residual error, leading to a robust model capable of handling non-linear relationships and interactions between variables.
Furthermore, XGBoost performance was optimised by hyper parameters. For instance, the maximum depth of a tree controls each tree's depth, hence balancing the model's complexity against overfitting. The learning rate defines each optimisation step, while the number of estimators defines the number of trees in the ensemble. Another important hyper parameter contributing to the robustness of the model is the subsample, which randomly selects a fraction of the data for each tree.
Beyond its mathematical underpinnings, XGBoost brings many engineering improvements. It supports parallel processing, which enables the efficient training of large datasets. It also has sparsity-aware algorithms that natively handle missing data, reducing the need for extensive data pre-processing. These enhancements make XGBoost not only accurate but also computationally efficient.
XGBoost has combined the power of gradient boosting with more sophisticated regularisation and optimisation techniques. Thus, it can handle a high degree of predictive accuracy and computational efficiency and, therefore, becomes very versatile in modelling complex datasets such as sales forecasts or other real-world problems.

B. Workflow Explanation

The workflow of the XGBoost approach will be presented in this section to provide an idea of how it works, highlighting some of its advanced prediction features. Fig. 1 below presents a workflow of how the XGBoost approach gradually enhances its predictions by constructing an ensemble of decision trees. Every step is specially designed to enhance the model's accuracy with a series of transformations from data preparation to the final prediction.

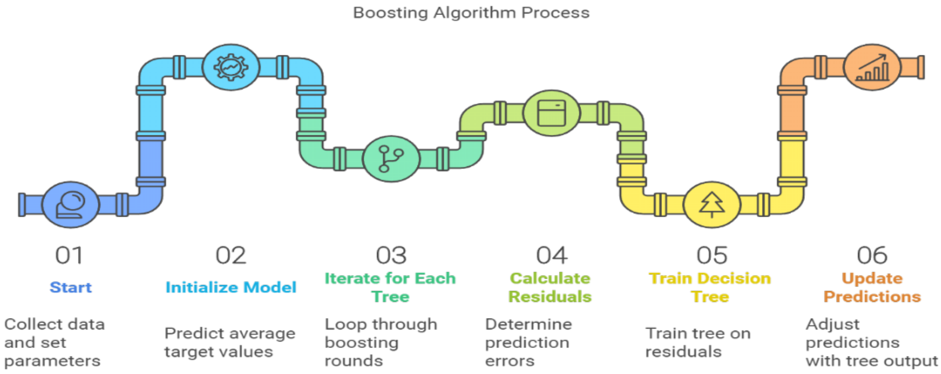

Fig. 1. XGBoost Workflow

DOI: http://doi.org/10.24086/icafs2025/paper.1582

First, it involves an input dataset with a set of features that are independent variables and target values which are dependent. After this, the whole dataset is partitioned into two main divisions: one for the model training process and the second one for testing its results. In this, the training subsection is used to formulate decision trees via residual errors developed by previously formed trees. Before this division occurs, the data goes through several pre-processing steps to get it into a form that will be as useful as possible for training: missing data handling, encoding of categorical variables, and feature normalising. Once the dataset is ready, the model enters its tree construction phase, building a sequence of decision trees. It calculates, at each step, the residuals from the previous tree and designs the new tree to correct these errors. This is an iterative approach in which the model keeps improving its predictions with every addition of a tree. The trees are grown based on an objective function that evaluates the best split at each node, aiming to reduce residual error and improve prediction accuracy. Regularisation is crucial in this step as it will avoid overfitting and ensure the model generalises new data well.

The model will then, after training, make the prediction on the test set using the aggregation of outputs from all trees. These are summed up to get a final prediction, with each tree making an improved prediction upon the previously made one. R^2, RMSE, and MAE, among other evaluation metrics assessing the accuracy of the regression model, can now be found.

The workflow also emphasises some key decision points, including hyper parameter tuning and the optional steps of model refinement, where further adjustments can be made based on the results from the evaluation phase. The above figure encapsulates the cyclical nature of XGBoost, where each iteration refines the model's understanding and accuracy. That is done until performance has attained an acceptable level or a certain stop criterion is met, such as the desired number of trees or error threshold. With this, the step-by-step nature of the approach ensures that XGBoost can deal with complex data and improve predictions without computational performance degradation.

C. Algorithm Explanation

The XGBoost algorithm details the steps for training and testing a model. It first involves data preparation, which involves dividing the dataset into features, X, and the target variable, y. The data must be divided into training and test subsets, preferably in an 80-20 split ratio, to ensure that most data is used for training the model while the rest is held for testing. This is also where other pre-processing steps are completed, like handling missing values and encoding categorical variables. Next, the model and hyper parameters are defined. An XGBoost Regressor model is initialised, and a grid of hyper parameters is specified. Key parameters include the maximum tree depth (max_depth), learning rate (learning_rate), number of trees (n_estimators), and the fraction of data to be used per tree (subsample). These parameters balance model complexity, learning speed, and prediction accuracy. Next, the algorithm does hyper parameter tuning with GridSearchCV, which means performing an exhaustive search and cross-validation. Five-fold cross-validation makes the model evaluate robustly and chooses the best configuration concerning the R^2 score. Once the optimal hyper parameters have been selected, the model can be trained on the training data, X_train, and y_train. The fitted model is then used to make predictions on X_test.

Finally, it computes evaluation metrics to check the performance of the model. It also calculates metrics like Root Mean Squared Error to measure the accuracy of the prediction and the R^2 score to measure the proportion of variance explained by the model. These metrics are printed out and analysed for the implementation of the algorithm. This algorithm has been structured well, making it clear and easy to apply to any regression.

Algorithm

Goal: Improve predictions step by step using patterns in the mistakes (errors).

1. **Start**:
   - Collect your data: Input features X and target values y.
   - Set parameters: Decide how many steps (trees) you want to build ($n\_estimators$) and how big each step should be (learning rate).
2. **Step 1: Initialise the Model**
   - Predict the average of all target values as the starting point:
   - $Initial\ Prediction = mean(y)$
   - Save this as the current prediction.
3. **Step 2: Iterate for Each Tree (Boosting Round)**
   - **For** $i = 1$ to $n\_estimators$:
     1. **Calculate Residuals**:
        - Find out how wrong the current predictions are:
          $Residuals = y - Current\ Predictions$
     2. **Train a Small Decision Tree:**
        - Use the input data X to train a simple tree that predicts these residuals.
     3. **Update Predictions:**
        - Adjust the current predictions using the tree's output:

$$New\ Predictions = Current\ Predictions + Learning\ Rate \times Tree\ Output$$

4. **Step 3: Final Model**
   - Combine all the small adjustments (trees) with the initial prediction to make the final prediction.

The above algorithm outlines in detail how to build a gradient-boosting model. This method makes iterative improvements in predictions by focusing on mistakes made in prior rounds. The process starts by making a simple guess, calculating the average of the target values, and providing the initial prediction. The algorithm computes the distance between the current predictions and the actual values at every round, called residuals. Then, it fits a small decision tree to learn from those residuals. It updates the predictions by adding the decision tree output with a fraction according to a learning rate to the current predictions. It repeats that several times and each step increase the model's precision. Lastly, all the predictions from the different iterations and the first guess come together to create the final model. This approach enables the model to learn incrementally and correct errors; hence, it becomes very effective in the case of complex datasets.

DOI: http://doi.org/10.24086/icafs2025/paper.1582

IV. RESULTS ANALYSIS AND DISCUSSION

Effective business decision-making depends significantly on precise sales prediction, especially for industries where demand fluctuations can significantly affect profitability, like retail. For a global brand like Adidas, accurate predictions of sales trends are imperative to optimise inventory management, enhance marketing strategies, and ensure operational efficiency. It enables the detection of complex patterns and relationships in sales data that might remain hidden when analysed through conventional methods using machine learning models. The sales dataset is collected from a secondary source of data called Kaggle.[1] It discusses the performance of five sophisticated machine learning models, namely, XGBoost, Random Forest, K-Nearest Neighbours (KNN), Linear Regression, and AdaBoost, in predicting Adidas's sales and operating profit. Each model is analysed on key metrics, such as R^2, RMSE and MAE are suitable for sales prediction. The results emphasise XGBoost as the most suitable model due to its accuracy and robustness. It can provide valuable insights that are useful in strategic planning and give Adidas an edge in competitive markets.

A. Predictions and Trends

The feature importance analysis of the XGBoost model highlights its predictive capabilities. Feature 1, with a score of 0.66158, stands out as the most influential, significantly outperforming other features. Features 2 and 0 follow with scores of 0.09627 and 0.08738, respectively, contributing meaningfully but not dominating. Features 6 and 3, with scores of 0.05830 and 0.05088, play moderate roles, while Features 4 (0.00184) and 8 (0.00297) have minimal impact. This indicates a concentration of importance in a few features, suggesting potential for feature selection or dimensionality reduction to enhance computational efficiency without sacrificing accuracy. See Fig. 2 for Adidas footwear monthly sales predictions.

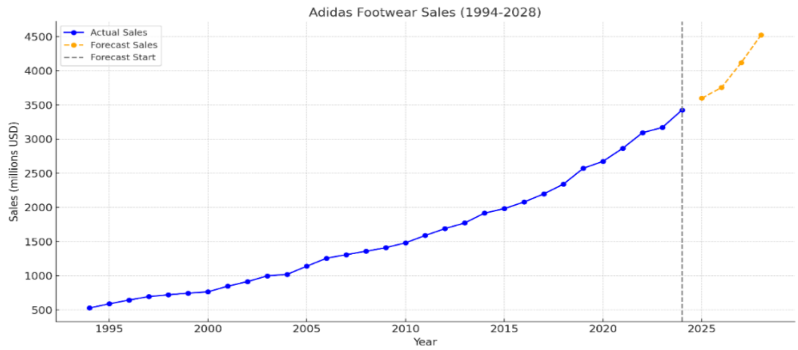

Fig. 2. Distribution of the prediction model feature importance scores

The sales trends for Adidas footwear from 1994 to 2028, as shown in the visualization, indicate a consistent upward trajectory driven by growth in market share, evolving consumer preferences, and strategic business decisions. Between 1994 and 2024, actual sales data exhibit a robust annual growth pattern, with notable fluctuations likely reflecting external factors such as market dynamics and economic conditions. From 2025 to 2028, forecasted sales maintain a positive trajectory, demonstrating continued expansion and market penetration. The forecast highlights a steady annual growth rate, albeit with reduced variability, suggesting a more predictable market environment or stable demand drivers. The transition from actual to forecast sales aligns with the observed historical growth trends, emphasizing the reliability of the forecasting model. These results underline Adidas's strong positioning in the global footwear market, with strategic opportunities for further innovations, regional market optimization, and inventory management improvements.

This analysis also aligns with the results of the prediction models, where XGBoost outperformed by utilising the most important features best. It gives an understanding that it is essential to prioritise key predictors, which would improve the model's performance and reduce the noise introduced by less impactful variables.

B. Performance Metrics

The performance of XGBoost in this study was stellar, as it outperformed all other tried models. It achieved the best R^2 value of 0.9963; that is, the model explained 99.63% of the variance in the sales data. This clearly shows the model's capability to capture the underlying patterns and trends of the data accurately. This high R^2 value shows the exceptional performance of XGBoost in modelling the complex relationships of various features in the data, such as seasonal patterns, product categories, and regional sales differences.

It was XGBoost that also recorded the lowest RMSE of 0.0072 and MAE of 0.015, both pointers to the model's exactitude in predicting sales. Both indicate that the difference between the predicted and actual sales values is minimal, pointing to the fact that predictions by XGBoost will be highly reliable and close to real-world outcomes.

One of the reasons that XGBoost can do so well is its capacity to handle complexity in dataset non-linear interactions. The algorithm makes a series of improvements to predictions by learning from errors made in previous steps, thus capturing subtle patterns in data that other models might be unable to spot. These then combine through iterative improvement aided by regularisation techniques to ensure the model performs well on new, unseen data, making it suitable for sales prediction in a dynamic retail environment such as Adidas footwear.

C. Comparison Study

Understanding consumer demand is crucial for appropriate inventory management, strategic marketing, and profitability for sports footwear businesses like Adidas. Due to the increasing competition within the retail sector, using advanced machine learning models for sales prediction has

[1] https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset

DOI: http://doi.org/10.24086/icafs2025/paper.1582

become highly relevant. The current study compares the performance of five machine learning models, namely XGBoost, Random Forest, K-Nearest Neighbours (KNN), Linear Regression, and AdaBoost, on the Adidas sales dataset. These models have been selected based on their theoretical strengths and practical applicability to prediction tasks, ranging from simple linear models to more complex ensemble and boosting techniques.

This comparison study aims to find the best model for predicting Adidas sales using historical data, considering price, discounts, regional variations, and seasonal trends. This study will compare the models using some key evaluation metrics. R^2, RMSE and MAE will provide actionable insights into which model can most accurately predict future sales and help guide strategic business decisions. The findings of this study are particularly relevant for businesses looking to enhance their predictive capabilities and optimise their supply chain management.

The performance of the five models, namely XGBoost, Random Forest, K-Nearest Neighbours (KNN), Linear Regression, and AdaBoost, are evaluated on three important metrics, namely, R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), as shown in Fig. 3 and Table 1 below. These metrics give insight into the models' fitness, prediction accuracy, and generalisation of unseen data. Each of these metrics evaluated the different models and determined their effectiveness in predicting Adidas sales. The results obtained from these metrics showed that XGBoost outperformed all other models by a large margin.

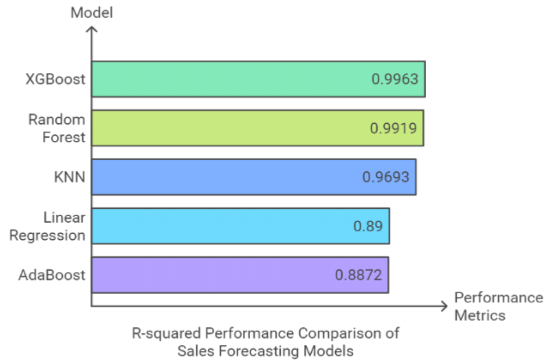

Fig. 3. Machine learning comparison for Sport footwear prediction

TABLE 1
Machine Learning Approaches Comparison Criteria

Model	R^2	RMSE	MAE
XGBoost	0.9963	0.0072	0.015
Random Forest	0.9919	0.0098	0.020
KNN	0.9693	0.0184	0.030
Linear Regression	0.89	0.1120	0.050

We employed RMSE and MAE to assess prediction accuracy, which measures our predictions' average magnitude of errors (Boehmke & Greenwell, 2019). Additionally, R^2 was used to assess the proportion of variance explained by the models, as suggested by Xia et al. (2012).

The best performance was observed for XGBoost, which yielded an R^2 of 0.9963, which means 99.63% of the variance in Adidas sales data was explained by it. That is, XGBoost captured most of the underlying patterns rather well. Moreover, with the lowest RMSE of 0.0072 and MAE of 0.015, this model resulted in predictions being very close to actual values. These results highlight the capability of XGBoost to model complex non-linear relationships and its robustness in handling feature-rich data. Its ability to avoid overfitting and deliver high predictive accuracy makes XGBoost especially suited for real-world sales prediction. However, even while not as accurate, Random Forest performed very well, with an R^2 of 0.9919, explaining 99.19% of the variance.

RMSE and MAE were slightly higher than for XGBoost, which are 0.0098 and 0.020, respectively, and were still low compared with other models. Random Forest will tend to generalise better thanks to its ensemble approach (combination of several decision trees). More importantly, it provides insights on feature importance, indicating factors like price, promotion, regional effects, and so on that significantly affect sales. K-Nearest Neighbours did moderately well, as evidenced by the R^2 value of 0.9693, thus accounting for 96.93% of the variance. Although still a strong result, it significantly lags both XGBoost and Random Forest in performance, thus proving that KNN has issues explaining complex relationships in data. This lower predictive power is reflected in a higher RMSE of 0.0184 and MAE of 0.030. The reliance of KNN on distance-based metrics makes it sensitive to noise. It struggles with big and dynamic datasets such as Adidas sales data, especially when prediction sharp increases due to promotions or other factors.

Linear Regression, as a baseline model, performed the weakest with an R^2 of 0.89, explaining just 89% of the variance. The RMSE (0.1120) and MAE (0.050) were significantly higher than the other models, indicating large deviations from actual sales values. Linear Regression's inability to capture non-linear relationships and complex feature interactions limits its performance in retail prediction, making it unsuitable for datasets with intricate patterns like Adidas's sales data.

AdaBoost is another ensemble method that had quite similar struggles to Linear Regression. With an R^2 of 0.8872, it explained only 88.72% of the variance, having an RMSE of 0.1150 and MAE of 0.055-the highest among all. While AdaBoost works quite effectively for simpler, linearly separable datasets, in the case of Adidas, this method could not capture the non-linear dynamic sales patterns, leading to poor predictive accuracy.

The precise performance comparison shows that XGBoost is the most effective model in predicting Adidas sales due to its high R^2 value and low RMSE and MAE. Random Forest is a good alternative, offering accuracy and interpretability, though not as precise as XGBoost. KNN, Linear Regression, and AdaBoost performed poorly, with higher error metrics and limited explanatory power; hence, they are unsuitable for complex prediction tasks like sales prediction.

Overall, XGBoost is the best model for making real-time sales forecasts, while Random Forest could be used instead when interpretability is a focus. The other models cannot handle the nature and dynamics of retail data.

D. Overall Discussion

Comparative analysis across the five machine learning models, namely XGBoost, Random Forest, KNN, Linear Regression, and AdaBoost, provides important pointers for their respective performances for predicting Adidas sales. Although all have their strengths, the efficacy of XGBoost concerning accuracy, scalability, and handling complex non-linear associations makes it the best option, which is consistent with findings from Xia et al. (2012), who highlighted XGBoost's superiority in handling complex data. Random Forest was a close second, offering strong performance with good interpretability, but it was slower in computation and slightly less precise than XGBoost (Breiman, 2001). KNN, while effective for simpler models, struggled with scalability, especially in dynamic datasets, as discussed by Cover & Hart (1967). AdaBoost was effective for some tasks but showed limitations in capturing non-linear interactions, similar to findings by Schapire (1990).

What stands out with XGBoost is the built-in regularisation techniques that include L1 and L2 regularisation, preventing overfitting even when working with complex datasets. This is particularly essential in sales prediction, as the outcomes of the interactions among many variables are non-linear. It also nails the linear and non-linear relationships of variables, making it very adaptive for such datasets. Another advantage is its scalability: it can maintain high performance as the dataset size increases, which is suitable for real-world applications such as retail prediction. Feature importance rankings from the model are also very informative, helping decision-makers understand the most influential factors on sales and allowing them to optimise their strategy accordingly. Random Forest comes a close second after XGBoost on performance, turning in a strong performance with an R^2 of 0.9919 and an RMSE of 0.0098. While it is not as precise as XGBoost, Random Forest has advantages in interpretability and robustness. Because it averages predictions over many decision trees, the ensemble nature of Random Forest reduces overfitting and improves generalisability. It also provides feature importance rankings, which allow businesses to identify key predictors, such as discounts or product categories. While less accurate than XGBoost, Random Forest remains a good option for businesses prioritising transparency and interpretability.

While KNN works fine with simpler data, it is not fit for the high-dimensional and dynamic nature of data involved in sales prediction. It achieved a fairly decent R^2 of 0.9693 but relies heavily on local patterns and distance metrics, making it relatively inefficient for capturing complex sales trends. Its performance was significantly weak at capturing sudden sales spikes or irregular patterns in sales, making it unfit for large-scale retail prediction. Linear Regression served as a baseline but was relatively poor for more complex, non-linear patterns. It resulted in an R^2 of 0.89 and an RMSE of 0.112 indicates that it cannot capture the intricate interactions driving the sales. Although easy to implement and interpret, linear regression cannot handle non-linearity and thus is limited to higher-order prediction tasks such as sales prediction at Adidas. The AdaBoost model performed no better, returning an R^2 value of 0.8872 and an RMSE of 0.115. Though a very powerful algorithm in many other cases, AdaBoost relies on weak learners internally and is sensitive to noise; thus, it performed poorer in modelling the complex interaction within the Adidas sales dataset.

Therefore, XGBoost is the most reliable model for predicting Adidas sales, considering its high accuracy, scalability, and handling of complex relationships. Random Forest presents a substantial alternative, especially when interpretability is required, while KNN, Linear Regression, and AdaBoost performed poorly due to their inability to model non-linear patterns. For future work, exploring hybrid models that combine the strengths of XGBoost and Random Forest could further optimise performance and interpretability.

V. CONCLUSION

This research analysis proves that advanced machine learning models are important for the prediction of sales and operating profit for Adidas footwear sales. Among the methods tested, XGBoost, Random Forest, and K-Nearest Neighbours emerged as the top performers, each offering unique advantages. XGBoost, with its robust gradient boosting framework, has shown the highest predictive accuracy for both sales and profit by competently handling non-linear interactions and employing regularisation techniques that reduce overfitting to a minimum, thus becoming ideal for large-scale datasets. Although it is computationally intensive, it provides sharp insight into sales patterns and key performance drives, which can be valuable in making decisions related to inventory management and marketing strategies. By ensemble learning, Random Forest showed good accuracy and robustness against overfitting and the interpretability of feature importance rankings. It would be beneficial in handling diverse data types, such as sales channels and geographical information, and balancing computational efficiency with quality predictions. While not as accurate as XGBoost or Random Forest, KNN did perform well in capturing regional and product-specific trends and would be suited for smaller-scale datasets or the targeted analysis of niche markets and seasonal fluctuations. Though generally problematic for high-dimensional data, the reliance of KNN on distance metrics remains an asset due to its simplicity and ability to adapt to local patterns. While XGBoost is the most accurate and scalable, Random Forest provides a

DOI: http://doi.org/10.24086/icafs2025/paper.1582

balance between interpretability and accuracy, and KNN is effective in targeted analysis; all of these are complementary strengths that Adidas can capitalise on to enhance sales prediction and strategic decision-making in the competitive sportswear market.

REFERENCE

Abbas, I. T., & Al-Salami, Q. H. (2023). Inverted Generational Distance Bat Algorithm for Many-Objective Optimization Problems. http://doi.org/10.24086/ICAFS2023/paper.905

Abed, F., & Al-Salami, Q. H. (2021). Calculate the best slope angle of photovoltaic panels theoretically in all cities in Turkey. *International Journal of Environmental Science and Technology*, 1-16.

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*, *29*(5-6), 594–621. https://doi.org/10.1080/07474938.2010.481556

Al-Bazi, A., Al-Salami, Q., Abbas, A., & Jerjees, Z. (2023, July). A System Based on Fuzzy Logic to Manage Operations in Container Yards. In *4th International Conference on Administrative & Financial Sciences* (pp. 88-95). Cihan University-Erbil.

Al Bazi, A., Tobechukwub, O., & Alsalami, Q. (2014). Improving Performance of Manufacturing Systems Using Simulation Technology (Powder Coating System as a Case Study). In Cihan University, *First International Scientific Conference*.

Al-Salami, Q. H., Saleh, R. K., & Al-Bazi, A. F. (2022). An efficient inventory model-based GA for food deterioration products in the tourism industry. *Pesquisa Operacional*, 42, e257747.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2023). Forecast reconciliation: A review. *International Journal of Forecasting*.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32. https://doi.org/10.1023/a:1010933404324.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Chapman and Hall/CRC. 1st Edition. https://doi.org/10.1201/9780367816377

Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2, 62-77. https://doi.org/10.1007/978-3-642-36318-4_3

Birdawod, H. Q. (2022). Using factor analysis to determine the most important factors affecting student absenteeism at Cihan University-Erbil. *Cihan University-Erbil Scientific Journal*, *6*(2), 1-8.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic record*, *88*, 2-9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27. https://doi.org/10.1109/tit.1967.1053964.

Dhankhar, S., Dhankhar, N., Sandhu, V., & Mehla, S. (2024). Forecasting Electric Vehicle Sales with ARIMA and Exponential Smoothing Method: The Case of India. *Transportation in Developing Economies*, *10*(2), 32. https://doi.org/10.1007/s40890-024-00216-y

Gilliland, M. (2010). *The business forecasting deal: Exposing myths, eliminating bad practices, providing practical solutions*. John Wiley & Sons.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*; MIT Press: Cambridge, MA, USA. The Korean Society of Medical Informatics. ISBN: 978-0262035613

Hamoudia, M., & Vanston, L. (2023). Machine Learning for New Product Forecasting. In *Forecasting with Artificial Intelligence: Theory and Applications* (pp. 77-104). Cham: Springer Nature Switzerland.

Hassan, A. M., Alomari, M. A., Al-Salami, Q. H., Alqurashi, F., Flayyih, M. A., & Sadeq, A. M. (2024). Numerical analysis of MHD combined convection for enhanced CPU cooling in NEPCM-filled a trapezoidal cavity. *International Communications in Heat and Mass Transfer*, *159*, 108343.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing*, *34*(1), 154–172. https://doi.org/10.1016/j.ijresmar.2016.06.001

Jameel, A. M., & Al-Salami, Q. H. (2023). Principal Component Analysis Technique for Finding the Best Applicant for a Job. *Cihan University-Erbil Journal of Humanities and Social Sciences*, *7*(1), 121-125.

Kharfan, M., & Chan, V.W. (2018). Forecasting Seasonal Footwear Demand Using Machine Learning.http://hdl.handle.net/1721.1/117612

Lourenço, C. E. E., Januario, L. F., & dos Santos, V. M. (2024). Advanced Sales Prediction: Harnessing Machine Learning for Dynamic Fashion Insights.

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81–93. https://doi.org/10.1016/j.dss.2018.08.010

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Malik, S., Khan, M., Abid, M. K., & Aslam, N. (2024). Sales Forecasting Using Machine Learning Algorithm in the Retail Sector. *Journal of Computing & Biomedical Informatics*, *6*(02), 282-294. hCps://doi.org/10.56979/602/2024

Massoudi, A. H., Fatah, S. J., & Jami, M. S. (2024). The role of artificial intelligence application in strategic marketing decision-making process. *Cihan University-Erbil Journal of Humanities and Social Sciences*, 8(1), 34-39.

Mahmood, N. H., Kadir, D. H., & Birdawod, H. Q. (2022). The full factorial design approach to determine the attitude of university lecturers towards e-learning and online teaching due to the covid-19 pandemic. *Cihan University-Erbil Scientific Journal*, *6*(1), 20-25.

Nawkhass, M. A., & Birdawod, H. Q. (2017). Transformed and solving multi-objective linear programming problems to single-objective by using correlation technique. *Cihan International Journal of Social Science*, *1*(1), 30-36.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*, 197-227. https://doi.org/10.1007/BF00116037

Soltaninejad, M., Aghazadeh, R., Shaghaghi, S., & Zarei, M. (2024). Using Machine Learning Techniques to Forecast Mehram Company's Sales: A Case Study. *Journal of Business and Management Studies*, *6*(2), 42-53. https://doi.org/10.32996/jbms.2024.6.2.4

Sun, L. (2024). Management Research of Big Data Technology in Financial Decision-Making of Enterprise Cloud Accounting. *Journal of Information & Knowledge Management*, 23(01), 2350067.

Swaminathan, K., & Venkitasubramony, R. (2024). Demand forecasting for fashion products: A systematic review. *International Journal of Forecasting*, *40*(1), 247-267. https://doi.org/10.1016/j.ijforecast.2023.02.005

Xia, M., Zhang, Y., Weng, L., & Ye, X. (2012). Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. Knowledge-Based Systems, 36, 253-259. https://doi.org/10.1016/j.knosys.2012.07.002

Xu, J., Zhou, Y., Zhang, L., Wang, J., & Lefloch, D. (2021). Sportswear retailing forecast model based on the combination of multi-layer perceptron and convolutional neural network. Textile Research Journal, 91(23-24), 2980-2994. https://doi.org/10.1177/00405175211020518

Zeynali, M. A., & Albadvi, A. (2023). Data-Driven Trade Promotion Optimization for Revenue Management in the Footwear Industry: A Case Study of Adidas and WestGear. 9th International Conference on Industrial and Systems Engineering.

DOI: http://doi.org/10.24086/icafs2025/paper.1582