Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(4): 2092–2117.
DOI: 10.3934/era.2025092
Received: 10 January 2025
Revised: 26 March 2025
Accepted: 31 March 2025
Published: 14 April 2025

Research article

Striking a balance: navigating the trade-offs between predictive accuracy and interpretability in machine learning models

Miguel Arantes[1,*], **Wenceslao González-Manteiga**[2,3,*], **Javier Torres**[4] **and Alberto Pinto**[1,5]

[1] Departamento de Matemática, Faculdade de Ciências da Universidade do Porto, Porto 4785-999, Portugal

[2] Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, Santigo de Compostela, Spain

[3] CITMAga: Centro de Investigación e Tecnoloxía Matemática de Galicia, Santiago de Compostela, Spain

[4] Departamento de Matemática Aplicada, Universidad de Vigo, Vigo 28039, Spain

[5] LIAAD - INESC TEC, Porto 4785-999, Portugal

* **Correspondence:** Email: miguelarantes.as@gmail.com, aapinto@fc.up.pt.

Abstract: Sales forecasting is very important in retail management. It helps with decisions about inventory, staffing, and planning promotions. In this study, we looked at how to balance the accuracy of predictions with how easy it is to understand the machine learning models used in sales forecasting. We used public data from Rossmann stores to study various factors like promotions, holidays, and store features that affect daily sales. We compared a complex, highly accurate model (XGBoost) with simpler, easier-to-understand linear regression models. To find a middle ground, we created a hybrid model called *LR_XGBoost*. This model changes a linear regression model to match the predictions of XGBoost. The hybrid model keeps the strong predictive power of complex models but makes the results easier to understand, which is important for making decisions in retail. Our study shows that our hybrid model offers a good balance, providing reliable sales forecasts with more transparency than standard linear regression. This makes it a valuable tool for retail managers who need accurate forecasts and a clear understanding of what influences sales. The model's consistent performance across datasets also suggests it can be used in various retail settings to improve efficiency and help with strategic decisions.

Keywords: sales forecasting; retail management; machine learning models; predictive accuracy; interpretability; explainable artificial intelligence (AI); XGBoost; hybrid models

1. Introduction

In retail management, predicting sales is crucial for important tasks like managing inventory, setting staffing levels, and planning promotions. Accurate sales forecasts help make operations more efficient, reduce waste, and improve customer satisfaction [1, 2]. However, while it is important to have precise forecasts, it is just as important that the models used to make these predictions are easy to understand and provide useful information for people making decisions, like store managers [3, 4].

Retail environments are very complex, with many factors affecting sales patterns, such as promotions, competition, holidays, seasonal trends, and local market conditions [5]. Store managers often need to predict sales weeks in advance, and the accuracy of these forecasts can greatly affect how efficiently their stores run and how profitable they are [6].

There is also growing pressure for AI models to be transparent and accountable. For example, the European Union has introduced rules that require AI models to be clear, explainable, and free from bias. These rules show the need for models that not only give accurate predictions but also provide clear and understandable insights that decision-makers can trust [7].

In this study, we look at how to balance accuracy and interpretability in machine learning models for sales forecasting. We use available data from the Rossmann dataset as our main benchmark. We compare a complex model that is harder to interpret (XGBoost) with simpler linear regression models, including a hybrid model that we called *LR_XGBoost*, which we trained to replicate the predictions of XGBoost. This hybrid approach aims to combine the high accuracy of complex models with the interpretability of simpler ones.

Our goal of this study is to find a forecasting model that balances high accuracy with transparency, providing useful insights for decision-makers [3]. Achieving this balance is important because it helps store managers and other stakeholders make well-informed, data-driven decisions that improve efficiency and performance. By applying our approach to different retail contexts, we hope to offer a practical solution that can enhance decision-making across the industry.

2. Literature review

Balancing predictive accuracy with interpretability in machine learning models is a challenge that has been widely discussed across areas. Traditional models, such as linear regression (LR), have been valued for their simplicity, transparency, and ease of interpretation. These qualities are particularly important in industries like finance, healthcare, and retail, where decisions must be transparent and understandable to ensure trust, meet regulatory requirements, and communicate effectively with stakeholders [8].

However, as data becomes more complex and the demand for accuracy increases, more advanced models, like XGBoost and deep neural networks (DNNs), have become popular. These models often perform better in terms of accuracy because they can capture complex, non-linear relationships in the data. Despite their strengths, these models are often criticized for being hard to interpret, making them like ”black boxes” that are difficult to understand and explain [3]. This lack of transparency is a problem in fields where understanding the reasons behind a prediction is just as important as the prediction itself [9].

For example, in healthcare, DNNs have shown high accuracy in predicting diagnoses, but their lack of clear explanations can make it hard for doctors to trust or use them because they need to understand the factors behind a diagnosis [4]. Similarly, in financial services, complex models can make it harder to comply with regulations that require clear explanations for decisions, like those involved in credit scoring [5].

Studies, also verified in our study, show that while models like XGBoost can be very accurate, they may not always be the best choice when interpretability is important. This has led us to the development of hybrid approaches that try to combine the best features of both complex and simpler models—using the predictive power of advanced models while keeping the interpretability of models like LR. One approach is to train a simple, interpretable models to approximate the outputs of a more complex model, making the decision-making process clearer without losing too much accuracy [10].

The increasing complexity of machine learning models, such as XGBoost, has led to the need for techniques that enhance interpretability. Explainable artificial intelligence (XAI) aims to provide transparency, enabling stakeholders to understand and trust the predictions generated [11].

SHapley additive explanations (SHAP) is a widely used method to interpret complex models, calculating the contribution of each feature to a prediction based on Shapley value theory. SHAP has been successfully applied to tree-based models like XGBoost, offering insights into feature importance and interactions [12].

Researchers have applied XGBoost with XAI techniques to extract interpretable patterns from complex datasets, such as microfluidic drop coalescence and structural damping estimation, and identified critical features relevant to coalescence processes, enhancing the interpretability of the model's predictions [13].

In another study, an explainable AI model predicting the equivalent viscous damping ratio (EVDR) in dual frame–wall resilient systems was developed. Utilizing the XGBoost method on extensive datasets, they provided accurate predictions and valuable insights into the factors influencing EVDR, thereby enhancing the understanding and design of such structural systems [14].

The integration of XAI techniques with XGBoost improves interpretability and contributes to greater user confidence in model predictions.

In summary, even though advanced models like XGBoost improve predictive accuracy, the need for interpretability remains a key consideration, especially in fields where the stakes are high. Ongoing research continues to explore ways to bridge the gap between these two important aspects, ensuring that machine learning models are both accurate and easy to interpret, meeting the diverse needs of different industries [3,4].

2.1. The importance of interpretability

In applications where decisions have significant consequences, the importance of interpretability cannot be overstated. For instance, in credit scoring, the ability to explain decisions is critical for regulatory compliance and maintaining customer trust [5]. Similarly, in healthcare, the interpretability of models can significantly influence clinical decision-making, where understanding the basis of a prediction is as important as the prediction itself [2].

Linear regression continues to be widely used due to its simplicity and interpretability, enabling stakeholders to easily understand the relationships between variables [10]. This is particularly important in scenarios where transparency is required to validate a model's predictions and ensure they align with domain knowledge [4]. However, the simplicity of LR can also be a limitation, as it may not fully capture the complexities inherent in more intricate datasets, particularly those involving non-linear relationships or interactions between variables [1].

2.2. Challenges and improvements in post-hoc explanations

The growing use of complex models like XGBoost has spurred the development of post-hoc explanation techniques, such as LIME and SHAP, which aim to make these models more interpretable [9]. These methods provide insights into how models arrive at their predictions, adding a layer of transparency that is otherwise missing in black-box models. However, these methods are not without limitations. For example, LIME creates local surrogate models that approximate the predictions of the original model, but the stability of these approximations can be problematic. Small changes in input data can lead to different explanations, potentially resulting in inconsistent and misleading interpretations [15]. This variability raises concerns about the reliability of LIME in critical applications where consistent explanations are essential.

SHAP values, derived from cooperative game theory, specifically the Shapley value concept, assign a contribution value to each feature for every prediction [16]. While SHAP provides a consistent approach to feature attribution, it is computationally demanding, especially for models with many features or large datasets. This computational load can limit its practicality in real-time applications or when working with large-scale models, where speed and efficiency are crucial [2, 17]. Moreover, although SHAP offers a theoretically robust method for determining feature importance, its complexity can obscure interpretability for non-expert users, leading to challenges in understanding and trusting model outputs [3].

Furthermore, relying solely on post-hoc explanations does not fully address the opaque nature of black-box models. These methods often result in fragmented explanations that do not provide a comprehensive understanding of the model's decision-making process, leading to potential misunderstandings or misinterpretations by end-users [18]. This piecemeal approach is particularly problematic in high-stakes fields like finance and healthcare, where decisions based on model predictions can have significant consequences [19]. For example, in credit scoring, where transparency and fairness are paramount, relying solely on post-hoc explanations may not provide the clarity required by regulators and stakeholders [5].

Recognizing these limitations, researchers have focused on improving interpretability by refining post-hoc methods or integrating interpretability directly into model design. Some approaches aim to enhance the stability and consistency of LIME and SHAP explanations, making them more reliable across different datasets and model configurations [20]. Other researchers advocate for developing inherently interpretable models that do not depend on post-hoc explanations but instead offer transparency through simpler, more understandable modeling techniques [2, 3]. For instance, linear regression models, when properly configured, can serve as interpretable alternatives that provide clear insights into the relationships between input features and predictions without sacrificing too much predictive power [10].

The ongoing advancements in this field highlight the importance of balancing predictive accuracy with interpretability. While complex models like XGBoost continue to lead in terms of accuracy, the trade-offs in terms of transparency and usability remain significant challenges. By refining post-hoc explanation methods and developing more interpretable models, the machine learning community can better meet the needs of stakeholders who require both high performance and transparency in their predictive models [3, 17].

2.3. Advancements in explainable AI

Recent progress in explainable AI (xAI) has focused on addressing the challenge of balancing interpretability with the demand for high predictive accuracy. One significant approach is the development of inherently interpretable models that are designed to be transparent from the outset, thereby eliminating the need for complex post-hoc explanations [3]. These models, such as traditional linear regression, decision trees, and generalized additive models, offer the advantage of being straightforward and easy to understand. However, they often struggle with the flexibility and predictive power needed to handle complex data patterns, limiting their effectiveness in tasks requiring high accuracy, such as deep learning applications [21].

To overcome the limitations of purely interpretable models, hybrid models have emerged as a promising solution. These models combine the strengths of both interpretable and complex models. For instance, training linear regression models on the outputs of more sophisticated models like XGBoost has been explored as a method to create models that maintain interpretability while achieving competitive accuracy [10]. This approach leverages the ability of complex models to capture intricate patterns in the data, while the linear regression component provides a transparent and interpretable framework for understanding these patterns. Such hybrid models have shown particular promise in applications like time series forecasting, where the inherent simplicity of linear models can be significantly enhanced by the deep insights provided by complex models [6, 22].

Another significant advancement in xAI is the growing use of counterfactual explanations. These explanations offer alternative scenarios that could lead to different outcomes, providing a clear and intuitive way to understand the decision-making process of complex models. Counterfactual explanations align well with the logic of linear models and can be applied to more sophisticated models, such as neural networks and ensemble methods, to make their predictions more accessible and understandable to non-experts [5, 19]. By offering clear "what-if" scenarios, counterfactual explanations help bridge the gap between complex model outputs and user interpretability, making them a powerful tool in both high-stakes fields like finance and healthcare as well as more general applications in retail and other industries [23].

The ongoing advancements in explainable AI reflect a broader trend toward making machine learning models not only more accurate but also more transparent and trustworthy. As xAI techniques continue to evolve, integrating interpretability into the modeling process from the ground up, along with developing effective hybrid models and counterfactual reasoning, is likely to play a key role in the future of AI, ensuring that these models are both powerful and understandable to all stakeholders [3, 17].

2.4. The case for linear regression guided by XGBoost

Given the challenges associated with black-box models and the limitations of post-hoc explanation techniques, there is growing interest in methods that enhance the interpretability of complex models. One effective strategy is to use linear regression models guided by the predictions of more sophisticated models like XGBoost [10]. This approach captures complex patterns while retaining transparency that is accessible to end-users, providing a balanced solution where both accuracy and interpretability are crucial [3].

Training a linear regression model on the outputs of an XGBoost model can create a surrogate model that mirrors the complex model's decisions in a more interpretable form. This method not only addresses the opacity of black-box models but also provides a simpler framework for decision-making, which is essential in fields like finance and healthcare where the ability to explain predictions is as important as the predictions [2].

In addition to linear regression, other hybrid models can also be explored. For instance, models that combine decision trees with XGBoost or that utilize ensemble techniques incorporating both interpretable and complex models offer promising directions for achieving a balance between accuracy and interpretability [24].

3. Methodology

Our methodology is carefully designed to evaluate the effectiveness of different machine learning models in predicting daily sales for Rossmann stores. The main goal is to find a balance between accuracy and interpretability, both of which are important for making informed decisions in retail. The models examined include XGBoost, which is known for its high predictive accuracy but lower interpretability, and more transparent models like LR and a modified linear regression model (*LR_XGBoost*) that aims to combine the accuracy of XGBoost with the clarity of linear regression.

Figure 1 outlines the overall methodology, which is detailed step-by-step below.

1) **Dataset split:** The original dataset is divided into two parts: train dataset and test dataset.
2) **Train XGBoost model:** The XGBoost model is trained using the train dataset. This model learns patterns in the data and makes predictions.
3) **Generate XGBoost predictions:** The trained XGBoost model produces predictions on the training data.
4) **Enhance the dataset:** These XGBoost predictions are added as new data to the train dataset, creating an enhanced dataset.
5) **Train linear regression model:** A LR model is trained on the enhanced dataset, which now includes XGBoost predictions.
6) **Generate predictions with LR:** The trained linear regression model makes predictions.
7) **Compare performance:** The test dataset is used to evaluate the models. The performance of LR, XGBoost, and the hybrid model (LR XGBoost) is compared using evaluation metrics.

Since our primary objective of this study is to demonstrate the methodology than optimize XGBoost through extensive hyperparameter tuning, a simple model configuration is used. To prevent overfitting, basic regularization techniques are applied. Early stopping is implemented by monitoring validation performance, ensuring that training halts when additional iterations do not yield significant improvements.

To evaluate the model's generalization capability, we compare its performance on the training (wmape of 11.16%) and test (10.18%) sets. The minimal variations in our evaluation metric values between the two datasets indicate that *lr xgboost* effectively captures relevant sales patterns without overfitting to noise or outliers. This confirms that the model maintains predictive accuracy while preserving interpretability.

Key methodological distinction: Unlike standard linear regression, which is trained directly on the original dataset, our approach first trains an XGBoost model. Its predictions are then incorporated as features in the enhanced dataset used to train the linear model. This enables the final model to benefit from XGBoost's predictive power while retaining the interpretability of a linear framework.

Figure 1 provides a visual overview of the methodology, showing the steps from data preparation to model evaluation. This flowchart offers a guide to the study's approach.

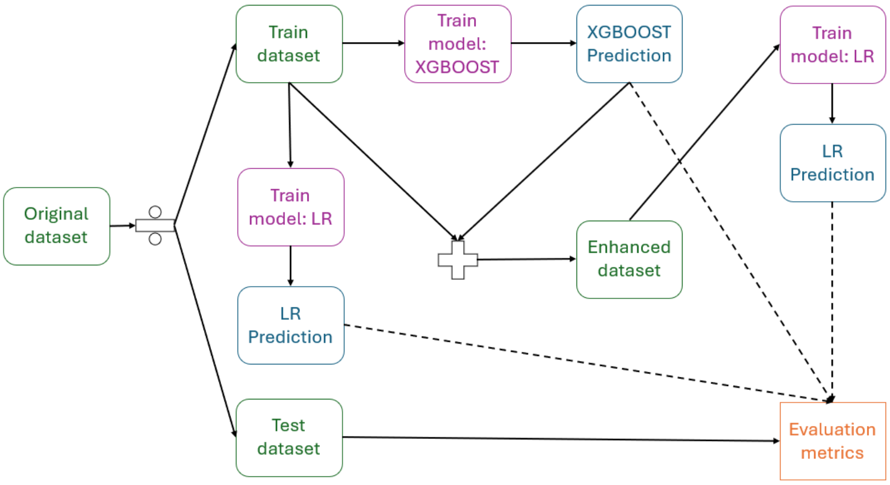

Figure 1. Methodology flowchart: Process of training and evaluating models.

We use a dataset with historical sales records from 1115 Rossmann stores. This dataset includes factors like promotions, holidays, and other indicators, all of which we think are crucial for accurate sales predictions. Initially, the dataset is split into two parts: A training set and a test set. The training set is used to develop the models, while the test set (of the last 30 days) is used to evaluate the models' performance. This split is important to ensure that the models can generalize to new, unseen data rather than just memorizing the training data. For the purposes of this study, we perform this split three times: First using the first 30 days of the final 90 days, then the first 30 days of the final 60 days, and finally the last 30 days alone to validate the consistency of the results.

During the training phase, the XGBoost model is developed using the training dataset. XGBoost is chosen because of its robustness and efficiency in handling large datasets with complex relationships. Furthermore, a linear regression model is trained on the same data to serve as a baseline for comparison due to its simplicity and interpretability.

To bridge the gap between the interpretability of linear regression and the accuracy of XGBoost, the *LR_XGBoost* model is created. This involves retraining the linear regression model using the predictions from the XGBoost model as inputs, aiming to capture the complex patterns identified by XGBoost while maintaining the transparency of a linear model.

After training, the models (XGBoost, LR, and *LR_XGBoost*) are used to generate predictions on the test dataset. These predictions are compared to the actual sales data to assess their accuracy. The primary evaluation metric is the weighted mean absolute percentage error (WMAPE), which offers a weighted assessment that takes into account the size of sales. WMAPE is particularly suitable for this study because it provides a more nuanced measure of forecasting accuracy than other metrics. Besides accuracy, the interpretability of the models is also evaluated, especially how well the *LR_XGBoost* model balances actionable insights with high predictive accuracy.

Using this methodology, we aim to provide valuable insights into the trade-offs between accuracy and interpretability in machine learning models for sales forecasting. The use of the Rossmann dataset, an available resource, ensures that the findings can be replicated and serve as a benchmark for future research in this field.

3.1. Data collection, analysis, and pre-processing

The dataset used in this study consists of historical sales data from 1115 Rossmann stores, collected over several years. This dataset provides a detailed view of the factors influencing daily sales across store locations. Key features of the dataset include:

- **Store**: A unique identifier for each store.
- **DayOfWeek**: The day of the week, from 1 (Monday) to 7 (Sunday), capturing weekly patterns in consumer behavior.
- **Date**: The specific date for each record, enabling trend analysis over time.
- **Month**: The month of the year, capturing seasonal variations that affect sales trends.
- **Sales**: The target variable, representing total sales for a store on a given day.
- **Customers**: The number of customers visiting the store on a given day, which is closely related to sales.
- **Open**: A binary variable indicating whether the store was open (1) or closed (0) on a particular day.
- **Promo**: A binary indicator showing whether a store was running a promotion on a given day.
- **StateHoliday**: A binary variable indicating whether the day was a state holiday.
- **SchoolHoliday**: A binary variable indicating whether the day was a school holiday, typically affecting store traffic.

Understanding the distribution and characteristics of the sales data is crucial for effective model training and evaluation. In Table 1, key statistical properties of the sales data from the 1115 Rossmann stores are summarized to provide insights into the variability and trends within the dataset.

Table 1. Descriptive statistics of Rossmann store sales

Statistic	**Value**
Mean (Average Sales)	5773.819
Median (Middle Value)	5744
Minimum Sales	0
Maximum Sales	41551
Standard Deviation (SD)	3849.926

In Tables 2 and 3, promotions and holidays significantly impact sales, and their effects are analyzed to better understand these influences:

Table 2. Mean sales on promotion and non-promotion days.

Promotion Status	Mean Sales
Non-Promo Days	4406.051
Promo Days	7991.152

Table 3. Mean sales on state holidays vs. non-state holidays.

State Holiday Status	Mean Sales
Non-State Holiday	5947.484
State Holiday	258.1596

Mean sales on promo days vs. non-promo days: the average sales during promotions are notably higher (around 7991 units) compared to non-promotional days (about 4406 units), showing the strong effect of promotions on sales volumes.

Impact of state holidays: on state holidays, sales drop significantly to an average of 258 units, compared to 5947 units on non-state holidays. This sharp decline is likely due to store closures or reduced operating hours during public holidays.

In Table 4, sales distribution across different days of the week was also analyzed:

Table 4. Average sales by day of the week.

Day of Week	Average Sales
Monday	7809.0445
Tuesday	7005.2445
Wednesday	6555.8841
Thursday	6247.5759
Friday	6723.2743
Saturday	5847.5626
Sunday	204.1832

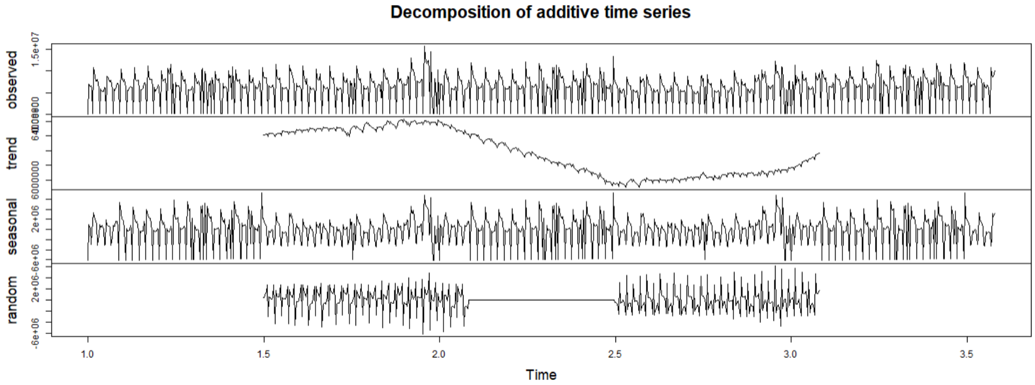

Figure 2. Decomposition of additive time series: observed, trend, seasonal, and random components.

Sales by day of the week: sales are highest on Mondays, averaging around 7809 units, and lowest on Sundays, with just 204 units on average. This variation likely reflects store operational hours and consumer shopping behavior, where Sunday operations are limited or stores may be closed.

To better understand the components of the time series data, a decomposition analysis was conducted, breaking down the observed sales data into trend, seasonal, and random components, as shown in Figure 2.

Trend component: the trend component shows a gradual increase in sales over time, possibly reflecting overall growth in store popularity or market expansion.

Seasonal component: the seasonal component captures the regular fluctuations in sales that occur weekly, likely influenced by consumer habits and promotional cycles.

Random component: the random component represents irregular, non-systematic variations in sales, which could be due to unexpected events or anomalies.

The data was carefully pre-processed to ensure it was suitable for modeling, involving the following steps:

- **Handling missing values**: Missing values are carefully filled in to maintain dataset integrity and avoid bias during model training.
- **Normalization of continuous variables**: Continuous variables such as "Sales" and "Customers" are normalized to bring them onto a similar scale, improving the performance of models like linear regression.
- **Encoding of categorical variables**: Categorical variables are converted into binary format using one-hot encoding. This process turns each category into a binary feature, enabling the models to process this information effectively.

After pre-processing, the dataset is split into training and testing sets. The training set is used to develop the models, covering most of the data to provide a strong foundation for learning. The testing set is reserved for evaluating model performance, ensuring that the results can be applied to new, unseen data.

3.2. Model training

The model training process in this study is designed to evaluate the high predictive accuracy of a complex, less interpretable model, and the interpretability of simpler, explainable models, like regression.

The following steps outline the model training approach, with more description in the next subsections:

- **XGBoost model training**: The XGBoost model is trained on the prepared dataset. Cross-validation is employed to fine-tune the model, preventing overfitting and ensuring robust performance on the test set.
- **Standard linear regression model training**: A standard linear regression model is trained using the same features as the XGBoost model, providing a baseline for comparison. The focus is on the model's interpretability, which enables a clear understanding of the relationship between input features and the target variable.
- **Hybrid model (*LR_XGBoost*) training**: The *LR_XGBoost* model is developed by retraining the linear regression model using the predictions from the XGBoost model as inputs. This method

aims to capture the patterns identified by XGBoost while maintaining the transparency of linear regression.

3.2.1. XGBoost model training

XGBoost, a method known for its efficiency and accuracy, is chosen as the primary non-interpretable model because of its ability to handle complex patterns in large datasets.

The XGBoost model is trained on the pre-processed dataset, focusing on optimizing its settings (hyperparameters) to achieve the best possible predictive accuracy. Important settings like the learning rate, maximum depth of trees, and the number of boosting rounds are fine-tuned using methods like cross-validation to prevent overfitting and ensure strong performance on unseen data [2, 5]. The choice of XGBoost aligns with recent advances in machine learning, where ensemble methods are often preferred for their ability to outperform single models, especially in complex, real-world datasets [9].

3.2.2. Linear regression models

In addition to XGBoost, two types of linear regression models are explored to evaluate their effectiveness in balancing accuracy and interpretability:

- **Standard LR**: The standard linear regression model serves as a baseline in this study. It is trained using the same features as the XGBoost model, offering a direct comparison in terms of predictive performance. Linear regression is widely used in fields where interpretability is crucial because it provides a clear understanding of the relationship between input features and the target variable. However, the simplicity of linear regression can limit its ability to capture complex, non-linear interactions within the data [3, 21].
- **Retrained LR with XGBoost (*LR_XGBoost*)**: To bridge the gap between the interpretability of linear regression and the predictive power of XGBoost, a retrained linear regression model, referred to as *LR_XGBoost*, is developed. This model is trained to mimic the predictions of the XGBoost model rather than the original target variable. By doing so, *LR_XGBoost* aims to replicate the results of XGBoost while maintaining the transparency of linear regression. This approach has been discussed in recent literature as a promising method to create models that offer both accuracy and transparency [10, 15].

The retraining process involves using the outputs of the XGBoost model as the dependent variable for the linear regression, effectively creating a model that translates the intricate, non-linear relationships captured by XGBoost into a more interpretable form. This method is particularly valuable in domains where stakeholders need to understand the reasoning behind predictions, such as retail management and financial forecasting [3]. Additionally, by fitting the linear regression model to the predictions of XGBoost, the study ensures that the *LR_XGBoost* model captures the same high-level patterns and interactions that contribute to XGBoost's accuracy, making it a robust and practical alternative for decision-makers who require both insight and performance from their models [4, 6].

Through this dual-model approach, we aim to demonstrate that it is possible to achieve a high level of predictive accuracy without sacrificing the interpretability necessary for effective decision-making in real-world applications. This methodology also enables a comparison of the trade-offs involved in choosing between models, providing a comprehensive evaluation of the strengths and limitations of each approach in the context of sales forecasting [2,21].

3.3. Evaluation metrics

The effectiveness of the models is evaluated using WMAPE, a metric specifically chosen for its ability to provide a nuanced assessment of predictive performance in the context of retail forecasting. WMAPE is particularly advantageous in scenarios like sales forecasting, where the goal is to minimize prediction errors relative to the actual sales values across a range of different store sizes and sales volumes.

The WMAPE is calculated as the ratio of the sum of absolute errors to the sum of actual values, with the result expressed as a percentage. This weighting mechanism ensures that the impact of errors is proportionate to the actual sales values, making WMAPE a more robust and meaningful measure of forecasting accuracy compared to simple mean absolute percentage error (MAPE). WMAPE is particularly effective in retail environments where sales data can vary significantly between stores and periods, helping to highlight the performance of the model in a way that accounts for this variability [6].

In summary, WMAPE is defined as:

$$\text{WMAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \tag{3.1}$$

where y_i represents the actual sales, and $\hat{y}_i$ represents the predicted sales. This metric has been shown to be highly effective in contexts where it is critical to balance the precision of predictions across varying scales, particularly in environments characterized by heterogeneous data [2,6].

In the context of this study, WMAPE was chosen over other metrics like root mean square error (RMSE) or mean absolute error (MAE) because it normalizes the errors, allowing for more meaningful comparisons across different stores with varying sales volumes. This normalization is critical when dealing with datasets where some stores might have significantly higher sales than others, which could otherwise skew the performance evaluation of the model [9].

The selection of WMAPE aligns with best practices in retail forecasting, where the accuracy of predictions directly impacts inventory management, staffing, and overall operational efficiency. Retailers need metrics that not only reflect the accuracy of predictions but also provide insights into the practical implications of forecasting errors. Large errors in predictions for high-volume stores, for instance, can have far more significant consequences than similar errors in low-volume stores. WMAPE's weighting mechanism addresses this by ensuring that the evaluation metric reflects the operational realities of retail management [22].

Moreover, the use of WMAPE in this study is supported by research that emphasizes the importance of using tailored evaluation metrics that align with the specific goals and characteristics of the application domain. In retail forecasting, where different products and stores can exhibit vastly different sales behaviors, WMAPE helps maintain a balanced assessment of model performance across the board.

By employing WMAPE, this study not only measures the predictive accuracy of the models but also ensures that the results are directly relevant to the practical needs of retail management, making the findings more actionable for decision-makers [2,3]. This alignment of evaluation metrics with practical objectives underscores our commitment to providing models that are accurate and useful in real-world applications.

Table 5 shows the top 10 and worst 10 stores based on their WMAPE scores using daily aggregation, comparing the performance of LR, *LR_XGBoost*, and XGBoost models.

Table 5. WMAPE comparison for daily aggregation - top 10 and worst 10 stores.

Store ID	**WMAPE LR**	**WMAPE *LR_XGBoost***	**WMAPE XGBoost**
Top 10 Stores			
192	75.38	37.44	11.15
274	53.47	28.60	8.27
263	34.72	13.82	9.61
540	42.91	22.96	11.27
512	36.58	16.92	6.75
797	33.52	14.96	6.29
663	50.06	31.54	6.62
708	45.00	28.90	11.91
1056	31.97	17.35	16.79
684	31.24	18.70	9.98
Worst 10 Stores			
798	7.64	7.58	6.61
185	10.30	10.23	8.78
280	6.65	6.59	6.26
781	10.21	10.15	9.38
624	7.61	7.55	7.31
928	9.04	8.99	8.22
503	7.64	7.61	7.26
717	9.60	9.59	8.84
446	8.76	8.75	8.09
610	9.24	9.23	8.44

Table 6 shows the top 10 and worst 10 stores based on their WMAPE scores using weekly aggregation, comparing the performance of LR, *LR_XGBoost*, and XGBoost models.

Table 5 presents the WMAPE for three different models—LR, the retrained linear regression model (*LR_XGBoost*), and XGBoost—across the top 10 best-performing stores and the worst 10 performing stores based on their predictive accuracy.

For the top-performing stores, *LR_XGBoost* consistently shows a substantial improvement in WMAPE compared to the standard Linear Regression model. For example, Store 192 shows a WMAPE reduction from 75.38% (LR) to 37.44% (*LR_XGBoost*). This significant reduction highlights the effectiveness of *LR_XGBoost* in improving predictive accuracy while maintaining interpretability. XGBoost demonstrates the lowest WMAPE values across all top 10 stores, indicating its superior accuracy. However, the gap between XGBoost and *LR_XGBoost* is relatively small, suggesting that *LR_XGBoost* can serve as a viable alternative when interpretability is crucial. The *LR_XGBoost* model performs significantly better than the standard Linear Regression model in this cases. For instance, Store 263 shows a reduction in WMAPE from 34.72% (LR) to 13.82% (*LR_XGBoost*), demonstrating the hybrid model's ability to capture more complex patterns while being interpretable.

For the worst-performing stores, the differences in WMAPE between the three models are much smaller. This suggests that for these stores, the models struggle to capture the underlying sales patterns effectively, leading to relatively similar levels of performance. For example, Store 798 shows WMAPE values of 7.64% (LR), 7.58% (*LR_XGBoost*), and 6.61% (XGBoost), indicating that even the more complex models are not significantly outperforming the simpler ones. In many cases, the WMAPE values for *LR_XGBoost* and LR are very close, indicating that the retraining process might not provide as much benefit for stores with inherently unpredictable sales patterns. This is evident in stores like 185 and 624, where the WMAPE difference between LR and *LR_XGBoost* is minimal. Despite the challenges, XGBoost tends to outperform the other models, though the margin is narrower compared to the top-performing stores. This consistent performance across both best and worst cases highlights XGBoost's robustness as a forecasting tool.

In summary, the *LR_XGBoost* model demonstrates significant improvements over standard Linear Regression, closing the gap between LR and XGBoost while maintaining interpretability. This makes it an attractive option for stores with more predictable sales patterns. While *LR_XGBoost* offers some improvement over LR in the worst-performing stores, the gains are smaller, indicating that these stores may require more complex or different modeling approaches to achieve significant accuracy improvements. Overall, the *LR_XGBoost* model provides a balanced solution that leverages the strengths of both linear regression and XGBoost, making it suitable for a wide range of stores, especially those where interpretability and actionable insights are critical.

We can evaluate the results of the top 10 and worst 10 stores in Table 6.

This table provides a clear view of how well the models performed across stores, identifying both the highest-performing and lowest-performing cases. This detailed evaluation helps in understanding the strengths and limitations of the models, especially in a retail environment where varying store characteristics can lead to significant differences in predictive accuracy.

In the top-performing stores, the *LR_XGBoost* model shows a substantial improvement in WMAPE compared to the standard Linear Regression model, demonstrating the effectiveness of combining interpretability with the predictive power of XGBoost. For instance, Store 192 shows a reduction in WMAPE from 75.38% (LR) to 35.21% (*LR_XGBoost*), nearly halving the error rate while benefiting from the linear model's interpretability. However, XGBoost consistently achieves the lowest WMAPE values, exemplifying its superior ability to handle the complexities in the data. For example, in Store 797, XGBoost achieves a WMAPE of just 3.38%, significantly lower than both *LR_XGBoost* (12.77%) and LR (33.52%).

The significant reduction in WMAPE achieved by *LR_XGBoost* compared to LR across the top 10 stores suggests that retraining a linear model to mimic XGBoost can lead to a considerable improvement in accuracy, making it a viable alternative for situations where interpretability is crucial. For example, Store 99 demonstrates a remarkable improvement, with WMAPE dropping from 23.81% (LR) to 9.49% (*LR_XGBoost*), bringing it closer to XGBoost's 4.27%. This indicates that *LR_XGBoost* effectively bridges the gap between the simple linear regression model and the more complex XGBoost, offering a balanced solution with enhanced predictive performance and interpretability.

Table 6. WMAPE comparison for weekly aggregation - top 10 and worst 10 stores.

Store ID	WMAPE LR	WMAPE *LR_XGBoost*	WMAPE XGBoost
Top 10 results			
192	75.38	35.21	8.18
274	53.47	28.60	2.72
263	34.72	11.73	6.45
540	42.91	21.26	7.50
797	33.52	12.77	3.38
512	36.58	16.92	4.68
663	50.06	31.04	3.87
1056	31.97	14.38	9.92
708	45.00	28.16	6.95
99	23.81	9.49	4.27
Worst 10 results			
809	2.82	2.71	2.23
582	12.86	12.76	12.22
426	5.30	5.22	4.94
312	3.03	2.95	2.71
474	3.96	3.89	3.80
328	3.49	3.42	3.16
475	7.44	7.38	6.91
448	4.23	4.19	4.13
672	6.33	6.29	5.85
216	11.92	11.91	11.56

For the worst-performing stores, the differences in WMAPE between the three models are much narrower. This trend suggests that for these stores, all models, including XGBoost, struggle to predict sales accurately, likely due to factors such as irregular sales patterns or external influences that are difficult to model. For instance, Store 582 shows WMAPE values of 12.86% (LR), 12.76% (*LR_XGBoost*), and 12.22% (XGBoost), indicating that even the sophisticated XGBoost model does not substantially outperform the simpler models.

In these cases, the *LR_XGBoost* model does not exhibit the same level of improvement seen in the top-performing stores, suggesting that the complexity added by XGBoost may not capture additional predictive power for stores with less predictable sales patterns. For example, Store 216 shows only a marginal improvement, with WMAPE slightly reducing from 11.92% (LR) to 11.91% (*LR_XGBoost*), while XGBoost achieves a WMAPE of 11.56%.

This consistent performance across both the best and worst cases indicates that while *LR_XGBoost* is highly effective in stores with more predictable sales patterns, it provides less benefit in scenarios where sales are inherently more volatile. However, XGBoost remains the most robust model, delivering the lowest WMAPE values across most stores, reinforcing its status as a powerful tool for sales forecasting, particularly when accuracy is prioritized over interpretability.

In summary, the *LR_XGBoost* model effectively narrows the gap between the interpretability of Linear Regression and the predictive accuracy of XGBoost in the best-performing stores, making it an attractive option for retail scenarios where both accuracy and transparency are necessary. However, for stores with more challenging sales patterns, the additional complexity of XGBoost may not provide a significant advantage, suggesting that alternative modeling approaches might be needed in such cases.

We analyze this with different cuts on time intervals, and the results were consistent, maintaining this consistency both weekly and daily, which suggests that we created a hybrid model with high accuracy and high interpretability.

3.4. Final model selection

The final model selection process involved a comprehensive evaluation of each trained model on the test set, focusing on the balance between predictive accuracy and interpretability—two critical factors for practical deployment in retail environments.

The XGBoost model demonstrated exceptional predictive accuracy, leveraging its ensemble learning approach to capture complex patterns and interactions within the data. However, despite its strong performance, XGBoost is inherently a black-box model, meaning that the decision-making process behind its predictions is not easily interpretable by stakeholders [3]. This lack of transparency can be a significant drawback in retail settings, where managers need to understand the rationale behind predictions to make informed decisions [15].

On the other hand, the standard LR model provided a high level of interpretability. As a simple, transparent model, LR enables stakeholders to easily comprehend how different features influence sales predictions. However, this model did not achieve the same level of accuracy as XGBoost, particularly in capturing the nonlinear relationships and interactions between features that are often present in complex retail datasets [9]. This trade-off between accuracy and interpretability is a common challenge in the application of linear models to complex datasets [21].

Finally, the retrained linear regression model (*LR_XGBoost*) was developed to address this trade-off. By training the LR model to fit the predictions made by the XGBoost model, *LR_XGBoost* was able to approximate the complex patterns learned by XGBoost while maintaining the inherent interpretability of linear regression [10]. This approach creates a model that mirrors the decision-making process of XGBoost but in a more accessible and interpretable form, making it particularly useful for operational deployment in retail environments where understanding the "why" behind predictions is crucial [11].

The *LR_XGBoost* model effectively bridges the gap between accuracy and interpretability. Its ability to provide accurate predictions while offering a transparent view of how those predictions are generated makes it a valuable tool for retail stores managers. This model enables managers to not only forecast sales more reliably but also to understand the underlying factors driving these forecasts, thus enhancing their decision-making capabilities [2].

Given its balanced performance, the *LR_XGBoost* model was selected as the final model for deployment. This decision was based on its ability to meet the dual criteria of high predictive accuracy and sufficient interpretability, making it a practical and effective tool for Rossmann store managers to integrate into their daily operations. The deployment of *LR_XGBoost* underscores the importance of adopting models that align with both the technical demands of accuracy and the practical need for transparency in retail management [3].

By selecting *LR_XGBoost*, this study highlights the potential of hybrid approaches that leverage the strengths of both complex and simple models, offering a roadmap for other retail organizations looking to enhance their forecasting capabilities without sacrificing interpretability [23].

The *LR_XGBoost* model improves explainability in the following ways:

- The final model remains a LR, meaning that stakeholders can interpret the coefficients in a straightforward manner.
- XGBoost is used only to generate new data points, not as a direct predictor. The features in the final regression have meaningful relationships with the target variable.
- Unlike pure XGBoost, where feature importance is hard to interpret, *LR_XGBoost* ensures that the relationship between each predictor and the target remains linear and explainable.

A key distinction from standard LR is that the training dataset is enhanced with knowledge extracted from XGBoost. This enables the model to improve accuracy, as can be validated in the next section, while keeping its predictions interpretable. The regression coefficients in *LR_XGBoost* represent the impact of features after incorporating the structured predictions of XGBoost, making them more robust than a traditional linear regression.

4. Results

In this section, we outline the findings from our analysis, focusing on how well the models predict sales across all the stores. We use the WMAPE as our primary metric to measure accuracy, as defined previously. WMAPE provides a clear and straightforward way to evaluate the models' performance.

We evaluate WMAPE, RMSE, MAE, and MAPE as potential metrics; however, WMAPE is the most suitable for this exercise due to its ability to account for variations in sales volume and provide a more meaningful comparison across different stores.

In particular, WMAPE is highly effective in retail forecasting, as it mitigates issues encountered with traditional MAPE when actual sales values are close to zero, preventing inflated percentage errors. Additionally, it provides a normalized error measure that is easily interpretable by business stakeholders. Compared to RMSE and MAE, WMAPE enables a more balanced evaluation, as RMSE tends to over-penalize outliers, while MAE does not scale relative to sales volume, making comparisons across stores less meaningful.

4.1. Predictive accuracy

Tables 7 and 8 show the predictive accuracy for the XGBoost model, the standard linear regression model, and the retrained linear regression model (*LR_XGBoost*) across all the stores.

Table 7. Rank-based comparison across models for the stores.

Model	**Rank Sum**	**Rank 1**	**Rank 2**	**Rank 3**
XGBoost	1865	675	310	130
LR_XGBoost	1950	305	785	25
Linear Regression	2875	135	20	960

These results indicate that *XGBoost* is generally the best-performing model, while linear regression typically shows the poorest performance.

- XGBoost emerges as the best model (with 675 times on first position), achieving top ranks in a majority of cases, which aligns with its reputation for high accuracy in complex datasets.

- *LR_XGBoost* provides a strong alternative, often ranking just behind XGBoost, indicating its effectiveness as a hybrid model that balances accuracy and interpretability and is only 25 times the worst model, compared to 130 for XGBoost and 960 times for LR.
- LR consistently underperforms compared to the other two models, highlighting its limitations in handling complex patterns within the data.

Table 8. WMAPE comparison across models for the stores.

Model	WMAPE Mean	WMAPE Max	WMAPE Min
XGBoost	9.21%	76.95%	4.12%
LR_XGBoost	10.18%	81.62%	3.99%
Linear Regression	11.83%	82.29%	4.08%

The results show that the *LR_XGBoost* model strikes a good balance between accuracy and being easy to interpret, making it a strong option for predicting retail sales.

- The *LR_XGBoost* model is more accurate than the standard linear regression model. For example, the WMAPE mean for *LR_XGBoost* is 10.18%, compared to 11.83% for the standard linear regression. This suggests that retraining the linear model with XGBoost predictions helps it capture more complex patterns, leading to better predictions while being easy to explain.
- The XGBoost is the most accurate overall, with a WMAPE mean of 9.21%, but *LR_XGBoost* is close behind with 10.18%. This small difference shows that *LR_XGBoost* can nearly match XGBoost's accuracy.
- The *LR_XGBoost* model provides a good middle ground between XGBoost's accuracy and linear regression's simplicity. While XGBoost is powerful, it's harder to understand because of its complexity. In contrast, *LR_XGBoost* keeps the linear structure, making it easier to explain, with only a slight drop in accuracy. This balance is important in real-world applications, where both understanding and accuracy matter.

In summary, the *LR_XGBoost* model is a strong choice for retail sales forecasting. It improves upon the standard linear regression model's accuracy while remaining easy to interpret, making it more useful for decision-makers. Given its similar performance to XGBoost, *LR_XGBoost* offers a good solution that meets both accuracy and interpretability needs.

4.2. Detailed analysis of results

Table 9 highlights the superior predictive accuracy of the XGBoost model over the standard linear regression model. This is shown by lower average and median WMAPE values, meaning XGBoost consistently performs better across most stores. Additionally, XGBoost's lower standard deviation in WMAPE values indicates that it delivers more consistent results across different stores, making it a reliable choice for retail forecasting.

However, the *LR_XGBoost* model shows significant improvement over the standard linear regression model. By using XGBoost's predictions to train the linear regression model, *LR_XGBoost* captures many of the complex patterns that XGBoost identifies while maintaining higher interpretability. This is evident in the WMAPE values, where *LR_XGBoost* achieves results very close to those of XGBoost.

For example, across the top 20 stores listed in the table, *LR_XGBoost* consistently shows a significant reduction in WMAPE compared to the standard linear regression model, with improvements ranging from around 10% to over 30%. This highlights the effectiveness of *LR_XGBoost* as a more understandable alternative that doesn't sacrifice much in terms of accuracy.

Table 9. Comparison of WMAPE between *LR_XGBoost*, linear regression, and XGBoost with significant improvements.

Store ID	lr_xgb_wmape	lr_wmape	xgb_wmape	dif_pp_lr	dif_pp_xgb
192	0.3744	0.7538	0.1115	0.3794	–0.2629
274	0.2860	0.5347	0.0827	0.2487	–0.2033
263	0.1382	0.3472	0.0961	0.2090	–0.0421
540	0.2296	0.4291	0.1127	0.1995	–0.1168
512	0.1692	0.3658	0.0675	0.1966	–0.1017
797	0.1496	0.3352	0.0629	0.1856	–0.0866
663	0.3154	0.5006	0.0662	0.1852	–0.2492
708	0.2890	0.4500	0.1191	0.1611	–0.1699
1056	0.1735	0.3197	0.1679	0.1462	–0.0056
684	0.1870	0.3124	0.0998	0.1254	–0.0872
676	0.1382	0.2635	0.0704	0.1253	–0.0678
132	0.1548	0.2790	0.0782	0.1243	–0.0766
99	0.1188	0.2381	0.0999	0.1192	–0.0189
41	0.0980	0.2168	0.0843	0.1188	–0.0137
815	0.2209	0.3385	0.0632	0.1176	–0.1577
671	0.1275	0.2383	0.0692	0.1108	–0.0583
305	0.1802	0.2909	0.0862	0.1106	–0.0941
500	0.1369	0.2463	0.0778	0.1094	–0.0590
457	0.1590	0.2650	0.0806	0.1060	–0.0784
353	0.1502	0.2526	0.0570	0.1024	–0.0933

When comparing the percentage point differences (*dif_pp_lr* and *dif_pp_xgb*) and percentage differences (*dif_perc_lr* and *dif_perc_xgb*) between *LR_XGBoost* and the other models, it is clear that *LR_XGBoost* offers a balanced trade-off. It improves significantly over the standard linear regression model, with only a slight decrease in accuracy compared to XGBoost. For instance, Store 192 shows a *dif_pp_lr* of 0.3794, indicating a 37.94% improvement over the standard linear regression, while the difference compared to XGBoost is minimal, with a *dif_pp_xgb* of just –0.2629, representing a slight decrease in accuracy of 26.29%.

These results strongly support the choice of *LR_XGBoost* as the best model for deployment. It effectively combines the high interpretability of linear regression with the accuracy of XGBoost. Given its substantial improvement over the standard linear regression and its nearly equal performance to XGBoost, *LR_XGBoost* is a practical and effective tool for retail sales forecasting. The model's ability to closely replicate XGBoost's predictive power while offering clearer, more understandable insights makes it especially valuable in situations where decision-makers need transparency to justify and understand predictions.

Overall, the *LR_XGBoost* model represents a significant advancement in developing machine learning models that are both accurate and interpretable, making it an excellent choice for retail applications where both factors are crucial.

Explanation of variables:

- **Store ID**: The unique identifier for each store in the dataset.
- **lr_xgb_wmape**: WMAPE for the retrained linear regression model (*LR_XGBoost*). This model combines the interpretability of linear regression with the accuracy of XGBoost.
- **lr_wmape**: The WMAPE for the standard linear regression model. This is used as a baseline for comparison.
- **xgb_wmape**: The WMAPE for the XGBoost model, known for its high accuracy in capturing complex patterns in the data.
- **dif_pp_lr**: The difference in percentage points between the WMAPE of the *LR_XGBoost* model and the standard linear regression model. A positive value indicates that *LR_XGBoost* is more accurate than LR.
- **dif_pp_xgb**: The difference in percentage points between the WMAPE of the *LR_XGBoost* model and the XGBoost model. A negative value indicates that *LR_XGBoost* is slightly less accurate than XGBoost, but the difference is small.

4.3. Understanding WMAPE

The WMAPE is a key metric for evaluating how well forecasting models perform in retail, especially when applied to a wide range of stores like those in the Rossmann chain. WMAPE is particularly useful because it adjusts prediction errors based on the actual sales values, making it an ideal metric for comparing models across stores with different sales volumes.

The *LR_XGBoost* model shows strong performance in balancing predictive accuracy with interpretability. As shown in the WMAPE comparisons, *LR_XGBoost* significantly outperforms the standard linear regression model by reducing the average WMAPE, bringing it closer to the performance of the XGBoost model. This is particularly noticeable in stores where *LR_XGBoost* achieves a substantial reduction in WMAPE compared to the standard LR model, highlighting its ability to capture complex sales patterns while maintaining the simplicity and transparency needed for practical use.

One of the key benefits of the *LR_XGBoost* model is that it retains much of the predictive accuracy of the XGBoost model, as evidenced by the small differences in WMAPE between the two models. This suggests that while *LR_XGBoost* is simpler, it does not significantly compromise accuracy. Instead, it offers a more interpretable alternative that can be more easily understood and validated by store managers. This interpretability is crucial in a retail setting, where decisions must be transparent and justifiable to a wide range of stakeholders.

Moreover, the reduction in WMAPE observed with *LR_XGBoost* across multiple stores indicates that this model can generalize well across different retail environments, making it a versatile tool for sales forecasting. By achieving a balance between accuracy and interpretability, *LR_XGBoost* emerges as a strong candidate for deployment in Rossmann stores and potentially in other retail chains with similar forecasting needs. This balance ensures that while predictive power is maintained, the models remain actionable and accessible to those who rely on them for day-to-day decision-making.

5. Discussion

This study provides important insights into finding the right balance between accuracy and interpretability in machine learning models used for sales forecasting in the retail sector. By comparing a highly accurate but less transparent model like XGBoost with more interpretable models like linear regression and a retrained linear regression model, we can draw several key conclusions that are valuable for both machine learning experts and retail professionals.

5.1. *Predictive accuracy vs. interpretability*

As expected, the XGBoost model delivered the highest accuracy by effectively capturing complex patterns in the sales data. This level of accuracy is especially useful in retail, where even small improvements in forecasts can lead to significant benefits, such as better inventory management and more effective staffing decisions [6]. However, the main downside of XGBoost is its lack of transparency, making it hard for decision-makers to fully trust or understand the predictions without knowing the factors driving them. This lack of clarity can be a major issue in situations where interpretability is important, such as in regulated environments or when explaining results to non-technical stakeholders [3].

On the other hand, while the standard linear regression model is easy to interpret, it doesn't perform as well in terms of accuracy. This limitation is well-known, as simpler models often struggle to capture the complexities and non-linearities found in real-world data [21]. In the fast-paced and complex world of retail, these limitations can lead to less effective decisions, potentially impacting everything from inventory levels to promotional strategies [2].

The retrained linear regression model (*LR_XGBoost*) offers a promising middle ground. By mimicking the predictions made by XGBoost, it retains much of XGBoost's accuracy while keeping the transparency that comes with linear regression. This hybrid approach allows the model to deliver high accuracy while also producing outputs that are easier for retail managers to understand [10]. This ability to generate accurate forecasts while providing clear insights into the factors that influence these predictions makes the model highly useful for decision-makers who need to justify their actions based on model predictions [3].

5.2. *Interpretability of model coefficients*

A potential concern regarding the *LR_XGBoost* approach is whether substituting the original target variable with XGBoost predictions compromises the interpretability of the resulting linear model. Given that XGBoost captures complex non-linear patterns, it is relevant to assess whether the resulting coefficients still reflect meaningful relationships with the original input features.

Nonetheless, the interpretability of *LR_XGBoostt* remains preserved for the following reasons:

- The final model remains a LR, meaning that its coefficients provide direct interpretability.
- The XGBoost predictions are not treated as additional features but as new training data, ensuring that LR is trained on meaningful patterns rather than raw, uninterpretable transformations.
- Since LR is learning from the structured predictions of XGBoost, its coefficients now represent the linear relationships within these refined predictions, rather than raw input features.

This approach enables *lr xgboost* to retain the transparency of LR while benefiting from the predictive power of XGBoost. Rather than modeling the raw features directly, the linear regression component is fitted on structured predictions derived from XGBoost, thereby enhancing the clarity and robustness of the inferred relationships.

If the coefficients of *lr xgboost* significantly differ from those of a standard LR model, it would indicate that XGBoost is capturing highly non-linear patterns. However, as long as the linear relationships remain stable, the model provides an interpretable summary of the predictive process.

5.3. Practical implications for retail forecasting

For retailers like Rossmann, accurately forecasting sales while understanding the factors driving these forecasts is crucial. The *LR_XGBoost* model offers a balanced solution that meets both needs, making it a strong candidate for use in retail settings. The model's predictions can help improve various aspects of retail operations, such as inventory management, staffing, and promotions, while also providing clear, actionable reasons behind these predictions.

This study also highlights the broader importance of using models that are explainable, especially in contexts where transparency is key to gaining the trust of stakeholders. While XGBoost's high accuracy is undeniable, its lack of interpretability can limit its usefulness in situations where decisions need to be explained to a diverse group of people. The *LR_XGBoost* model addresses this issue by offering a middle ground where high accuracy is achieved without sacrificing transparency [3,4].

5.4. Transparency and stakeholder trust

One of the key advantages of the *LR_XGBoost* model in the retail industry is its ability to balance predictive accuracy with transparency. Stakeholders, such as retail managers and decision-makers, often rely on forecasting models to plan inventory, pricing, and promotions. However, complex machine learning models, such as XGBoost, are difficult to interpret, making it challenging to justify their predictions in real-world business scenarios.

The *LR_XGBoost* model enhances stakeholder confidence for several reasons:

- The final model remains a Linear Regression, which enables decision-makers to see clear coefficients associated with sales-driving factors.
- Even though the training data incorporates XGBoost predictions, the decision-making process follows a simple and explainable regression framework.
- The structure of the model aligns with traditional forecasting techniques, making it easier for retail teams to adopt and integrate into business workflows.

To evaluate stakeholder confidence, the following criteria can be considered:

- **Coefficient stability:** If the coefficients remain consistent over different time periods, stakeholders can trust that the model is not making arbitrary decisions.
- **Interpretability of predictions:** The ability to explain why sales are expected to increase or decrease in response to key variables.
- **Comparison with traditional models:** If *LR_XGBoost* provides similar interpretability to standard regression models but with improved accuracy, it reinforces its reliability.

By preserving a transparent model structure and incorporating XGBoost's predictive capabilities, *LR_XGBoost* facilitates informed and justifiable decision-making by stakeholders.

5.5. Limitations and future work

Although the *LR_XGBoost* model provided a balanced solution in this study, it is important to acknowledge certain limitations. Retraining the linear regression model to follow XGBoost's predictions, while effective, may introduce biases that could affect how well the model performs with new data or in different retail settings [9]. There is also a risk of overfitting, especially if the training data does not fully represent future conditions, which could lead to less accurate predictions [21].

While the manuscript partially demonstrates the superior predictive performance of LR-XGBoost—particularly through the extensive sales data and the results in Table 7—its generalizability remains limited. The conclusions drawn are specific to the Rossmann dataset and may not extend to other retail settings or industries characterized by different sales dynamics and feature distributions. Consequently, while the model performs well in this context, its applicability to broader domains requires additional empirical validation. These limitations should be carefully considered when interpreting the results and assessing the model's potential for deployment in other environments.

Future research should explore how well this approach works in other fields beyond retail. Testing the *LR_XGBoost* model on different types of data, such as time-series data from other industries, could provide insights into how widely this method can be applied. Additionally, refining this approach could involve experimenting with other hybrid methods, such as combining linear regression with different ensemble models, to further improve the balance between accuracy and interpretability [2].

To ensure successful adaptation to different domains, the following adjustments may be necessary:

- Feature engineering must be adjusted to fit industry-specific characteristics (e.g., patient records in healthcare, financial ratios in banking).
- Different evaluation metrics may be more appropriate depending on the use case (e.g., F1-score for fraud detection instead of WMAPE).
- Regulatory requirements must be considered, particularly in sectors like finance and healthcare, where interpretability is legally mandated.

Future research encompassing these applications would provide valuable insight into the generalizability of *LR_XGBoost* and contribute to validating its robustness across diverse domains.

Overall, this study contributes to the ongoing discussion on how to balance accuracy and interpretability in machine learning, especially in areas where transparency is just as important as performance. By showing the potential of the *LR_XGBoost* model, this research offers a practical framework that can be adapted and used in various real-world situations where both accurate predictions and clear insights are needed.

6. Conclusions

We focused on finding a balance between accuracy and interpretability in predicting daily sales for Rossmann stores in Germany. We looked at three different models: The advanced XGBoost, a traditional linear regression model, and a hybrid model (*LR_XGBoost*) that retrains linear regression using predictions from XGBoost. Our main goal was to see if a simpler, more transparent model could achieve nearly the same accuracy as a complex model like XGBoost while being easier to understand.

Preliminary experiments conducted on a similar retail dataset and on a distinct medical dataset yielded comparable results, suggesting that the methodology may hold promise beyond the Rossmann case.

The results were clear. XGBoost was the most accurate, achieving the best WMAPE scores. However, this accuracy comes with a downside—it lacks transparency. In many practical cases, especially in retail, it is important to understand the reasons behind predictions, not just the predictions [3, 4].

On the other hand, the *LR_XGBoost* model did a great job of balancing accuracy with interpretability. By retraining the linear regression model using outputs from XGBoost, we achieved WMAPE scores that were almost as good as XGBoost's. Importantly, this was done without losing the simplicity and clarity that come with linear regression. In many stores, *LR_XGBoost* not only held its own but even outperformed the standard linear regression model. In some cases, it reduced WMAPE by more than 37 percentage points compared to the traditional linear model, showing its ability to capture complex patterns in the data while being easy to use [10].

These findings suggest that *LR_XGBoost* is not just a theoretical success but also a practical one, especially in retail where decision-makers need to understand and trust predictions. Its ability to match XGBoost's accuracy while offering greater interpretability makes it a valuable tool for Rossmann store managers. With reliable and understandable forecasts, they can make better-informed decisions.

In conclusion, the retrained linear regression model is a promising solution for Rossmann and possibly other retailers. It effectively combines the predictive power of XGBoost with the transparency of simpler models. This combination not only improves the accuracy of sales forecasts but also ensures that the insights are actionable and easy to understand, leading to better resource allocation and operational efficiency.

Looking ahead, there is room to refine this approach further. Future research could explore different training techniques or test the *LR_XGBoost* model in various retail settings. Expanding this method to other sectors might also show its wider applicability. Additionally, exploring hybrid models that combine the strengths of different approaches could open up new ways to achieve both high accuracy and interpretability in machine learning.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

The authors would like to express their sincere gratitude to the reviewers for their thoughtful and constructive feedback. This work is partially financed by National Funds through the Portuguese funding agency, FCT-Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

Conflict of interest

The authors declare that there are no conflicts of interest. Alberto Pinto an editorial board member for Electronic Research Archive and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

1. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735
2. R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, *IEEE Access*, **8** (2020), 40200–40216. https://doi.org/10.1109/ACCESS.2020.2976199
3. F. Giannotti, F. Naretto, F, Bodria, Explainable for trustworthy AI, in *Human-Centered Artificial Intelligence*, (eds. M. Chetouani, V. Dignum, P. Lukowicz and C. Sierra), Springer, Lecture Notes in Computer Science, (2023), 175–195. https://doi.org/10.1007/978-3-031-24349-3_10
4. A. Barredo-Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion*, **58** (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
5. X. Dastile, T. Celik, Making deep learning-based predictions for credit scoring interpretable, *IEEE Access*, **9** (2021), 50426–50440. http://dx.doi.org/10.1109/ACCESS.2021.3068854
6. S. Taylor, B. Letham, Forecasting at scale, *Am. Stat.*, **72** (2018), 37–45. https://doi.org/10.1109/ITSC.2019.8916985
7. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, European Commission, 2021.
8. L. Bohlen, J. Rosenberger, P. Zschech, M. Kraus, Leveraging interpretable machine learning in intensive care, *Ann. Oper. Res.*, (2024), 1–40. http://dx.doi.org/10.1007/s10479-024-06226-8
9. D. Alvarez-Melis, T. Jaakkola, On the robustness of interpretability methods, preprint, arXiv:1806.08049. https://doi.org/10.48550/arXiv.1806.08049
10. Y. Yang, M. Wu, Explainable machine learning for improving logistic regression models, in *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, (2021), 1–6. https://doi.org/10.1109/INDIN45523.2021.9557392
11. D. Gunning, D. W. Aha, DARPA's explainable artificial intelligence (XAI) program, *AI Mag.*, **40** (2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850
12. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, et al., From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.*, **2** (2020), 56–67. https://doi.org/10.1007/s42979-021-00815-1
13. J. Hu, K. Zhu, S. Cheng, N. M. Kovalchuk, A. Soulsby, M. J. H. Simmons, et al., Explainable AI models for predicting drop coalescence in microfluidics device, *Chem. Eng. J.*, **481** (2024), 148465. https://doi.org/10.1016/j.cej.2023.148465
14. C. Xie, J. Hu, G. Vasdravellis, X. Wang, S. Cheng, Explainable AI model for predicting equivalent viscous damping in dual frame–wall resilient system, *J. Build. Eng.*, **96** (2024), 110564. https://doi.org/10.1016/j.jobe.2024.110564

15. A. Barredo-Arrieta, I. Lana, J. Del Ser, What lies beneath: A note on the explainability of black-box machine learning models for road traffic forecasting, in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, (2019), 2232–2237. https://doi.org/10.1109/ITSC.2019.8916985

16. L. S. Shapley, A value for n-person games, in *Contributions to the Theory of Games, Volume II*, (eds. Harold W. Kuhn and Albert William Tucker), Princeton University Press, (1953), 307–318. https://doi.org/10.1515/9781400881970-018

17. A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, R. S. Amant, Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives, *IEEE Trans. Artif. Intell.*, **3** (2022), 852–866. https://doi.org/10.1109/TAI.2021.3133846

18. C. Bove, T. Laugel, M. J. Lesot, C. Tijus, M. Detyniecki, Why do explanations fail? A typology and discussion on failures in XAI, preprint, arXiv:2405.13474. https://doi.org/10.48550/arXiv.2405.13474

19. S. Wachter, B. Mittelstadt, C. Russel, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, preprint, arXiv:1711.00399. http://dx.doi.org/10.48550/arXiv.1711.00399

20. I. H. Sarke, Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions, *SN Comput. Sci.*, **2** (2021), 420. https://doi.org/10.1007/s42979-021-00815-1

21. Z. C. Lipton, The Mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, preprint, arXiv:1606.03490. https://doi.org/10.48550/arXiv.1606.03490

22. G. Van Houdt, C. Mosquera, G. Nápoles, A review on the long short-term memory model, *Artif. Intell. Rev.*, **53** (2020), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

23. P. Biecek, M. Chlebus, J. Gajda, A. Gosiewska, A. Kozak, D. Ogonowski, et al., Enabling machine learning algorithms for credit scoring, preprint, preprint, arXiv:2104.06735.

24. C. Zhang, P. Hoes, S. Wang, Y. Zhao, Intrinsically interpretable machine learning-based building energy load prediction method with high accuracy and strong interpretability, *Energy Built Environ.*, **5** (2024). https://doi.org/10.1016/j.enbenv.2024.08.006