






# Sales Forecasting Using Machine Learning Methods for Online Store

Nur Adlina Mohd Shahar<sup>1</sup>, Sofianita Mutalib<sup>1\*</sup> ,  
Shamimi A Halim<sup>1</sup> , William Ramdhan<sup>2</sup> 

<sup>1</sup> School of Computing Sciences, College of Computing, Informatics and Mathematics,  
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

<sup>2</sup> Universitas Royal, Jl. Prof.H.M. Yamin No.173, Kisaran Naga, Kec. Kota Kisaran Timur,  
Kabupaten Asahan, Sumatera Utara Indonesia.

ed2adlina@gmail.com, \*sofianita@uitm.edu.my,  
shamimi134@uitm.edu.my, williamramdhan@universitasroyal.ac.id

**Abstract.** Sales forecasting is a strategic activity that involves projecting future sales for goods or services in assisting businesses in making educated inventory decisions, increasing operational efficiency, and improving the overall supply chain. Leveraging machine learning and data analytics, sales forecasting can benefit significantly in making accurate sales projections from historical data. This paper highlights the Exploratory Data Analysis (EDA) and feature selection using Recursive Feature Elimination (RFE) process to identify relationships and key features affecting sales. This paper identified important patterns and relationships from retail store context hence, revealed key variables which affected sales which are holidays, promotions, assortments and competition. Hypotheses of multiple relationships manage to be further analysed to utilise data for further model development. This paper used the approach of XGBoost that is able to model sales with accuracy of 98.23%. The forecasted model further strengthens its results through benchmarking against evaluation metrics of Root Mean Squared Error (RMSE), Normalised Root Mean Squared Error (NRMSE), Kolmogorov Smirnov (KS) distance and Pearson Correlation Coefficient (PCC). The trend projections of each variable affecting store and product sales are visualised using a user-friendly dashboard for easy comprehension in extracting key takes from the extracted relationships. This analysis can benefit retail companies by offering a keen insight for better understanding of sales impact factors.

**Keywords:** Sales Forecasting, Exploratory Data Analysis, Retail, Supervised Methods.

## 1 Introduction

Forecasting has become an intrinsic and critical component of the business value chain for large retail organisations operating on a massive scale. Retail businesses frequently suffer sales drops and spikes; if these unexpected surges in sales are not detected, it would be disastrous for the company and inventory [1]. Furthermore, store sales are frequently influenced by various seasonality variables. If these tendencies,

© The Author(s) 2024

N. A. S. Abdullah et al. (eds.), *Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024)*, Advances in Computer Science Research 117,  
[https://doi.org/10.2991/978-94-6463-589-8\\_6](https://doi.org/10.2991/978-94-6463-589-8_6)

whether growth or reduction, are not detected in a timely manner, inventory may become scarce in the future. As a result, the need of assessing product data and effectively translating sales analytics to business has always been a key strength of successful conglomerates, particularly in the modern period. The retail industry, with its supply chain dynamics faces continuous challenges in accurately predicting consumer demand across diverse store environments and demand or price prediction models have to consider these variations in the selected variables [2]. Variations in consumer behaviour, regional preferences, seasonal influences, and market dynamics make traditional forecasting methods less effective in addressing the complexities inherent in diverse retail settings.

It creates the need for us to make analysis on sales factors on a bigger picture. In addition, by continually relying on these traditional methods, it leads to limited predictive capacities, resulting in mistakes in projecting future demand. Prior research on the traditional forecasting method highlighted that the model ARIMA was unable to handle non-linear patterns and dynamic changes as well as effectively capture seasonal variations in retail data [3]. Additionally, Mitra et al. [4] demonstrated that Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) models achieve higher accuracy than ARIMA by effectively managing non-linear relationships in complex retail datasets. Hence, this proposes development of a better sales forecasting model that allows retail stores looking to streamline inventory decision-making processes to improve business management. Moreover, the factors affecting the predictions need to be studied in developing the best model.

This paper aims to apply machine learning methods for sales forecasting model for online retail stores. The task remains challenging, due to different set of variables to be considered in the development. The remainder of this paper is organised as follows: Section 2 describes the past studies and presents the similar studies in text classification. Section 3 presents the research methodology for the study. The results and findings of the study are presented in section 4 and finally, section 5 concludes the paper.

## 2 Literature Review

Delving into the project's literature review in order to identify the gap, it begins by discussing the primary concept, which is supply chain, and then delves into seeing the link between supply chain and sales forecasting. In relation to the title, the application of sales forecasting is narrowed to the retail sector.

### 2.1 Supply Chain

Supply chain management is fundamentally controlling flow of products and services, to carry out supply chain operations and determining the most recent demands of the market [5]. The supply chain and sales forecasting are interconnected components of a dynamic system that covers a product's or service's whole lifespan. A properly optimised supply chain may anticipate demand variations by employing advanced sales forecasting tools, assuring timely manufacturing and distribution. This synergy

improves overall corporate responsiveness, lowers inventory costs, and increases customer satisfaction. Businesses that match their supply chain operations with realistic sales projections can achieve more efficiency and agility in a competitive market.

## 2.2 Sales Forecasting

Sales forecasting is an important part of corporate strategy since it provides valuable insights into inventory management, financial planning, and market positioning. Accurate sales predictions allow businesses to anticipate demand, manage resources efficiently, and reduce the risks associated with supply chain interruption. In today's increasingly data-driven corporate world, new forecasting approaches such as machine learning and ensemble techniques provide substantial benefits in accurately anticipating future sales patterns. If businesses properly comprehend the influence of factors affecting sales and demand, they will be able to make effective sales planning and management [6]. In recent research of sales forecasting methods done in the retail sector, XGBoost algorithm was one of the highlighted algorithms adopted on promotion, holiday, product and customer variables in a sales forecast of customer purchases for a large-scale retail company and achieved great accuracy for its forecasts [7].

## 2.3 Extreme Gradient Boosting

XGBoost is a gradient boosting-based parallel tree model. It uses the tree ensemble technique, with the advantage of scaling for a variety of situations while utilising fewer resources. The parallel and distributed processing in XGBoost speeds up model learning and model exploration. Many studies have shown the reliability of XGBoost. For instance, Auppakorn & Phumchusri [8] forecasted daily sales for variable-priced items using methods of XGBoost, time series methods and Linear Regression. Despite that, they concluded that XGBoost marginally outperforms other models, but because WAPE is only slightly more than the validation set, XGBoost is safe for real life retail utilisation. In similar context, Al Ali [9] tackled the problem by presenting demand forecasting for Rossmann Stores utilising regression (Linear Regression, Random Forest Regressor) and the XGBoost Algorithm, with findings showing that XGBoost produced more accurate forecasts than previous solutions. Linear regression analysis is used in statistics to understand the relationship between dependent and independent variables. This analysis shows the strength of the link between the two variables and if the independent variable significantly affects the dependent variable. So, if the XGBoost application was done, the outcome could be compared with linear regression.

## 3 Research Methodology

### 3.1 Data Acquisition and Pre-Processing

Based on the literatures, XGBoost algorithm is to be applied in our study for developing a model of sales forecasting for retail stores. The data collection leads to a relevant retail stores data from Kaggle, which is Rossmann Stores [10]. The Rossmann Stores' dataset has 236380 points of record. The information obtained from the datasets such as historical sales records, holiday, competition and promotional period are essential for this research. Additional factors will be included such as seasonality, state and weather data to improve the dataset's relevance. To handle missing values, outliers, and assure compatibility with ensemble learning techniques, the acquired data will be subjected to rigorous preparation, including cleaning and feature engineering. Next, the data will be presented in several charts for the process of exploratory data analysis (EDA). This was done using Python libraries primarily pandas, matplotlib and seaborn. For preparation on modelling, an 80:20 ratio of training and testing split on both datasets will be utilised to train and assess ensemble learning models, allowing for accurate demand forecasting over a wide range of retail locations. Furthermore, a feature selection process is done to choose which features to be analysed in forecasting demand from the two datasets involved. This section includes the use of Recursive Feature Elimination using decision trees to assess features that are significant in the modelling of demand for Rossmann Stores. The idea behind RFE is that by progressively deleting features not having high correlation to the target variable sales, the model's performance should either increase or remain stable with a smaller number of features. This helps to reduce overfitting, improve interpretability and accelerate training.

### 3.2 Modelling

XGBoost, or extreme gradient boosting, builds upon gradient boosting with possible enhancements. XGBoost enhances performance and can solve real-world problems with a limited number of resources [4] [11]. XGBoost is a gradient boosting-based parallel tree model. It uses the tree ensemble technique, with the advantage of scaling for a variety of situations while utilising fewer resources. The parallel and distributed processing in XGBoost speeds up model learning and model exploration. The XGBoost modelling structure with chosen parameters that worked best for achieving high accuracy are by setting the number of estimators as 200, maximum depth of 9, learning rate at 0.2, subsample set at 0.8 and the samples by tree of 0.9. The predictions are then stored in 'y\_pred\_xgb'.

### 3.3 Model Evaluation

The evaluation of the model and its performance are critical aspects of the system development process. The constructed models are carefully tested using validation datasets and indicators to guarantee that they can anticipate demand consistently across the two retail stores involved. This assessment compares the model's predictions to real-world historical data, measuring the model's performance using

measures such as Root Mean Squared Error (RMSE), Normalised Root Mean Squared Error (NRMSE), Kolmogorov Smirnov (KS) distance and Pearson Correlation Coefficient (PCC) [12]. These metrics will be used in comparing the performance of XGBoost and the baseline model, which is linear regression [12] [13].

Formula for RMSE:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_t)^2} \quad (1)$$

Formula for NRMSE:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (2)$$

Formula for KS Distance:

$$KS(\mu, \sigma) = \min \left\{ \frac{k}{n} - \Phi \left( \frac{x_k - \mu}{\sigma} \right), \Phi \left( \frac{x_k - \mu}{\sigma} \right) - \frac{k-1}{n} \right\} \quad (3)$$

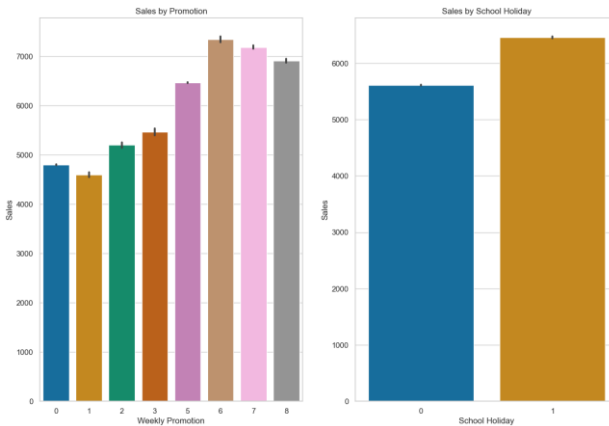
Formula for PCC:

$$p = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (4)$$

## 4 Results and Findings

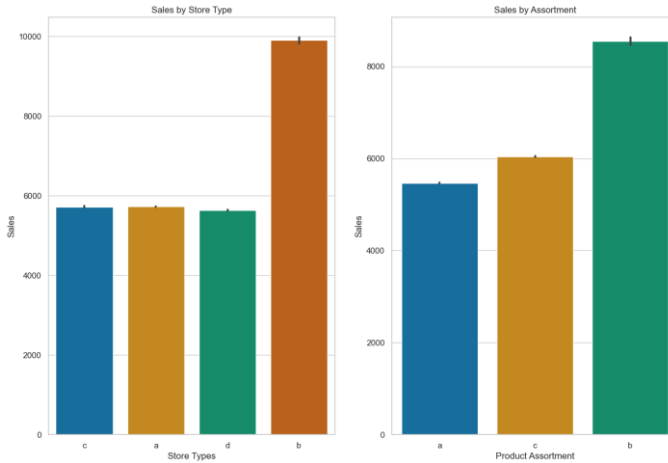
### 4.1 Outcome for Data Preprocessing

In the Rossmann Stores' dataset, there are missing values present as detected through the checking process using the function, `is.na()`. The missing values present in 'competition\_distance', 'competition\_open\_since\_month', 'competition\_open\_since\_year', 'promo2\_since\_week', 'promo2\_since\_year' and 'promo\_interval' variables are then filled in using the function `fill.na()` using Python language.



**Fig. 1.** Numerical Attributes Analysis

Figure 1 illustrates some of the numerical attributes using functions from the `matplotlib.pyplot` library. It can be seen promotion increases both sales and customers in all stores, especially during school holidays. In fact, stores that stay open during the school holidays enjoy more sales than on usual days. During school vacations, more stores are open than during state holidays and the sales increase around Christmas week and Easter week might be attributable to individuals purchasing more items over the holiday season. The promotions done from weeks six and seven are interrelated with the school holiday weeks, to show that Rossmann stores took advantage of the holidays to increase sales via promotion that is proved to be advantageous. Hence, numerical attributes concerning seasonality affect sales data.



**Fig. 2.** Categorical Attributes Analysis

Meanwhile the categorical attributes in the dataset are state holiday, store types and assortment. Figure 2 shows some of the illustration of bar charts for store type sales and assortment sales by using Python functions, which is the `matplotlib.pyplot` library. Some of the insights gained is as shown, store B is the most popular and busy retail kind, as it is most likely situated in the city compared to stores A, C and D. Diving deeper into assortment sales, the extra product assortment sold out the most compared to basic and extended product assortment. This is most likely due to a specific product assortment strategy adopted by the majority of Rossmann stores in the dataset. Feature selection done using Recursive Feature Elimination (RFE) has revealed the list of features significant to forecasting sales which are; Store, Customers, Promo, Store Type, Assortment, Competition Distance, Competition Open Since Month, Competition Open Since Year, Competition Time Month, and Day of Week.

## 4.2 Exploratory Data Analysis Results

In this section of exploratory data analysis, several hypotheses have been tested and summarised to further analyse the data and get to know relations between features. We share three hypotheses that show a true conclusion and high relevancy.

- The first hypothesis has been done to indicate whether stores should sell more as months increase. Figure 3 further proved that the hypothesis is true because there is a trend of increasing sales over the months.

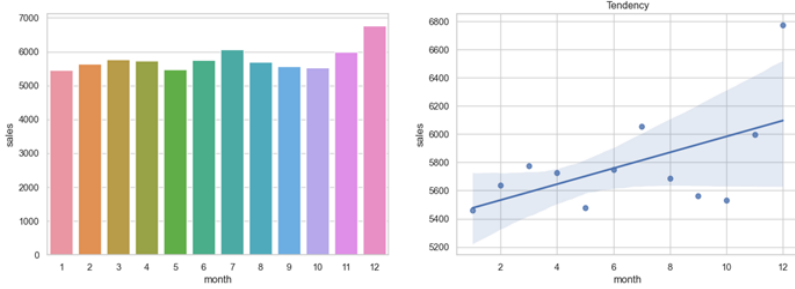


Fig. 3. Sales Trend Over the Months

- Another hypothesis was done as shown in figure 4 to see if stores should sell less on weekends, and this turned out to be true. This is because the total sales are lower on Saturdays and much lower on Sundays (as many stores do not open on Sundays). Considering the average sales on Sunday are high, this is due to the fact that stores open on Sunday only for a special occasion such as for holidays and promotions.

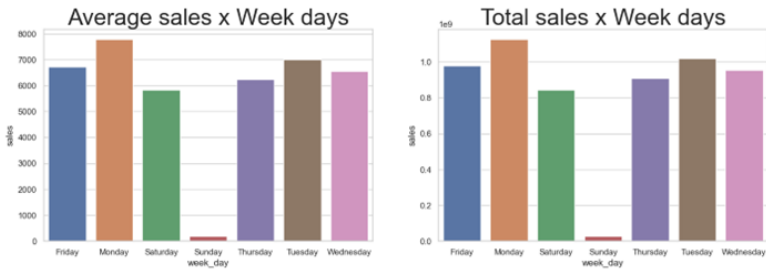
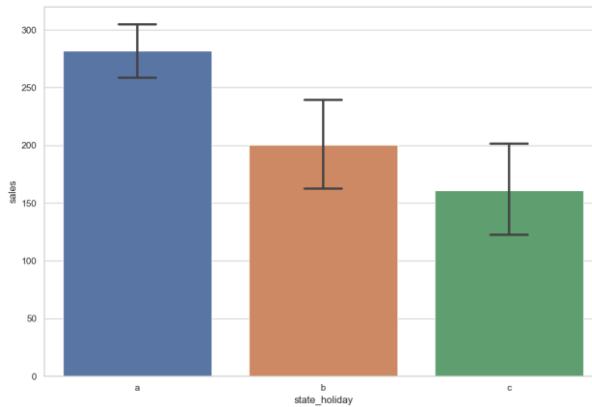


Fig. 4. Sales Distribution on Weekends and Weekdays

- Finally, the third hypothesis proves that stores open during Christmas holidays do sell more compared to other holidays. This is revealed as sales during the Christmas holiday are more expressive compared to sales on public holidays and normal days. In the figure 5 below, 'a' represents public holiday, 'b' represents Easter and 'c' represents Christmas.





**Fig. 5.** Sales During Popular Holidays

### 4.3 Modelling Results

Table 1 displays the model evaluation results done on the XGBoost and Linear Regression model for forecasting the sale based on attributes; store type, customers, promo, assortment and competition, using RMSE, NRMSE, KS Distance and PCC tests. The linear regression is used as the baseline model.

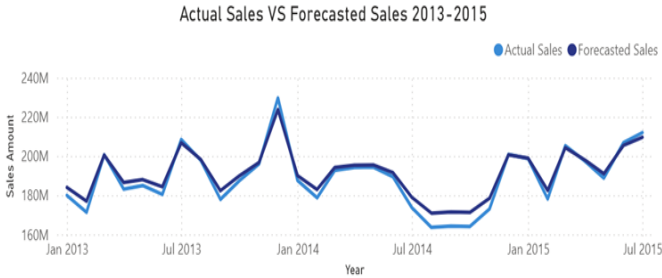
**Table 1.** Rossmann Stores Model Evaluation.

Model/Metric	RMSE	NRMSE	KS Distance	PCC
XGBoost	467.342883	0.023367	0.071100	0.991593
Linear Regression	1396.423712	0.069821	0.090955	0.922468

The PCC test shows great results where Rossmann stores' model achieved 0.991593. The results are close to 1, indicating a strong positive correlation between the variables trained in the two sets of data, hence showing good linearity between data. Meanwhile, KS distance value is 0.071100, is closer to 0 indicating that the variables are highly similar and correlating one another, hence how closely two sets of data's variables resemble each other in terms of their probability distribution. For RMSE evaluations done, Rossmann stores resulted in 467.342883, which does not necessarily indicate that there is overfitting. RMSE value is noted to be in the same unit as the forecasted values, this high value occurred due to the large sales value which are in the hundred thousand- and million-dollars range, thus this range of RMSE is acceptable. A difference of \$400 on a sale of \$100,000 represents a 0.4% error, which might be reasonable for long-term forecasting purposes. However, NRMSE is used to evaluate the model better than RMSE since it normalises the values of RMSE. The values achieved at 0.023367 suggested the almost perfect fit of the model as it has relatively low errors.

#### 4.4 Forecast Results

Sales forecasts have been compared with actual sales value from XGBoost. The predictions generated by all models on the test set have been transferred from Microsoft Visual Code into table form for better comparison and comprehension. The forecasts of the XGBoost model only have a slight deviation from actual sales. From these results, the forecasts obtained from XGBoost model is more accurate and showing higher proximity to actual sales than Linear Regression, as shown in Figure 6.



**Fig. 6.** Rossmann Stores Forecasted Sales Projection

Hence, with the model, forecasts for sales affected by the demand factors have been made with great accuracy. Thus, the forecasted sales are similar to the actual sales made to be applied to future forecasts ahead of the date intended. The store context forecasted sales from Rossmann Stores is shown below with the trend of sales projection. The forecasts for the time period from the year 2013 to 2015 are shown in figure 15. The trend here is that the actual sales mostly appear to be higher than the forecasted sales meaning that the forecasting model underestimated the actual sales a number of times. Moreover, since the model has been trained on seasonal data as well, the fluctuations indicate that the seasonality in the data followed a similar pattern of rising and falling throughout the year. This suggests that the forecasting model has captured some of the seasonal trends but did underestimate some of the sales volume. In a similar study on retail forecasting models, it was found that neither traditional models like ARIMA nor machine learning models such as LightGBM can effectively capture seasonality, while perfecting the nuances of sales volume during highly volatile periods. On another occasion, the incorporation of meta-learning, enhancing the model's predictive accuracy by proving that an ensemble approach can handle the complexity and seasonality in retail sales data, to anticipate sales volume [14] [15].

If sales volume is consistently underestimated, it might result in stockouts, in which the online store runs out of popular products during peak demand periods. This can lead to wasted sales opportunities and possibly push customers to rivals. Furthermore, for online retailers, particularly those that rely on just-in-time supply chains, underestimating sales volume can strain supplier relationships and cause

restocking delays, worsening the stockout issue. This can affect the entire supply chain, causing delays and higher expenses.

## 5 Conclusion

This research paper managed to make sales forecasts in the context of retail stores, utilising an XGBoost approach that provides a fresh take on the problem compared to traditional methods of time series forecasting such as ARIMA. However, with great strength, limitations arose. The limitation to this research is the utilisation of only one modelling method that can be further enhanced for allowing the model to achieve greater results. Further enhancement can be done by incorporating external factors in creating sales forecasts. Future shaping can be done to see how forecasts can be used for dynamic pricing strategies or targeted promotions to influence customer behaviour and optimise overall sales. Moreover, this work can be improved by experimenting with Deep Learning techniques where the use of deep learning architectures like Long Short-Term Memory (LSTM) or Convolutional Neural Networks (CNN) can be carried out for handling complex seasonality or external factors. Moreover, a consideration in deployment phase for the forecasting model to be integrated into a real-world system can be explored via cloud deployment options or containerisation for scalability.

## Acknowledgement

The authors would like to express the gratitude to College of Computing, Informatics and Mathematics, and Research Management Center, Univesiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research fund and support.

## References

1. Ddfsfs
2. R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283–1318, Oct. 2022, doi: 10.1016/j.ijforecast.2019.06.004.
3. M. Z. Shahrel, S. Mutalib, and S. Abdul-Rahman, "PriceCop – Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform," *International Journal of Information Engineering and Electronic Business*, vol. 13, no. 1, pp. 1–14, Feb. 2021, doi: 10.5815/ijieeb.2021.01.01.
4. S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1394–1401, Dec. 2018, doi: 10.1109/icmla.2018.00227.
5. A. Mitra, A. Jain, A. Kishore, and P. Kumar, "A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach," *Operations Research Forum*, vol. 3, no. 4, Sep. 2022, doi: 10.1007/s43069-022-00166-4.

6. K. A. Mukhamedjanova, "CONCEPT OF SUPPLY CHAIN MANAGEMENT," *Journal of Critical Reviews*, vol. 7, no. 02, Jan. 2020, doi: 10.31838/jcr.07.02.139.
7. S. Punia and S. Shankar, "Predictive analytics for demand forecasting: A deep learning-based decision support system," *Knowledge-Based Systems*, vol. 258, p. 109956, Dec. 2022, doi: 10.1016/j.knosys.2022.109956.
8. A. Massaro, A. Panarese, D. Giannone, and A. Galiano, "Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector," *Applied Sciences*, vol. 11, no. 17, p. 7793, Aug. 2021, doi: 10.3390/app11177793.
9. A. Auppakorn and N. Phumchusri, "Daily Sales Forecasting for Variable-Priced Items in Retail Business. Proceedings of the 4th International Conference on Management Science and Industrial Engineering, 2022." doi: 10.1145/3535782.3535794.
10. M. A. Ali, "Retail Demand Forecasting," RIT Digital Institutional Repository. <https://repository.rit.edu/theses/11093/>
11. "Rossmann Store Sales | Kaggle." <https://www.kaggle.com/c/rossmann-store-sales>
12. T. Chen and C. Guestrin, XGBoost. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. doi: 10.1145/2939672.2939785.
13. A. Hagen, J. Strube, I. Haide, J. Kahn, S. Jackson and C. Hainje. Proposed high-dimensional Kolmogorov-Smirnov Distance. A Proposed High Dimensional Kolmogorov-Smirnov Distance, Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.
14. C. Sun, "Forecasting Retail Sales Via the Use of Stacking Model," in *Advances in economics, business and management research/Advances in Economics, Business and Management Research*, 2022, pp. 405–411. doi: 10.2991/978-94-6463-036-7\_59.
15. T. L. D. De Albuquerque Dallegrave, M. V. B. Da Silva, D. A. De Souza Neto, P. H. B. E. Sousa Junior, J. E. De Albuquerque Filho, and W. B. Santos, "Sales Forecast Optimization: Ensemble and Time Series Comparison," *Revista De Engenharia E Pesquisa Aplicada*, vol. 6, no. 5, pp. 110–119, Nov. 2021, doi: 10.25286/rep.v6i5.2153.
16. L. A. C. G. Andrade and C. B. Cunha, "Disaggregated retail forecasting: A gradient boosting approach," *Applied Soft Computing*, vol. 141, p. 110283, Jul. 2023, doi: 10.1016/j.asoc.2023.110283.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

