

Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach

Antonio Panarese ^{1,*} , Giuseppina Settanni ¹ , Valeria Vitti ¹ and Angelo Galiano ¹ 

Dyrecta Lab, IT Research Laboratory, Via Vescovo Simplicio 45, 70014 Conversano, BA, Italy

* Correspondence: antonio.panarese@dyrecta.com

Abstract: Organizations engaged in business, regardless of the industry in which they operate, must be able to extract knowledge from the data available to them. Often the volume of customer and supplier data is so large, the use of advanced data mining algorithms is required. In particular, machine learning algorithms make it possible to build predictive models in order to forecast customer demand and, consequently, optimize the management of supplies and warehouse logistics. We base our analysis on the use of the XGBoost as a predictive model, since this is now considered to provide the more efficient implementation of gradient boosting, shown with a numerical comparison. Preliminary tests lead to the conclusion that the XGBoost regression model is more accurate in predicting future sales in terms of various error metrics, such as MSE (Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) and WAPE (Weighted Absolute Percentage Error). In particular, the improvement measured in tests using WAPE metric is in the range 15–20%.



Citation: Panarese, A.; Settanni, G.; Vitti, V.; Galiano, A. Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. *Appl. Sci.* **2022**, *12*, 11054. <https://doi.org/10.3390/app122111054>

Academic Editors: Alfonso Monaco and Loredana Bellantuono

Received: 30 September 2022

Accepted: 28 October 2022

Published: 31 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: sales forecasting; regression; gradient boosting; extreme gradient boosting (XGBoost); accuracy metrics

1. Introduction

The retail sector has transformed over time, integrating IT innovation to interact better with today's consumers increasingly connected to the Internet. Modern technology tools provide users with highly personalized and fluid experiences across every shopping channel. Companies by means of data analysis can offer a better customer experience, with faster, more reliable, services and better operational efficiency.

The digital transformation and expansion of online commerce allows companies to implement an omni-channel strategy. Therefore, the need arises to manage the various sales channels and various stores of a company. This research aims to investigate the applicability and performance of approaches based on boosting techniques in a case study related to the sales prediction in a multi-store or multi-channel context. In particular, two different implementations are designed, implemented and preliminarily tested. The output of the developed approach helps an e-commerce company to automate its internal processes and optimize workflows.

Most business companies have an online presence and have expanded online interfaces to offer additional services. Today, customers have the opportunity to browse an inventory, locate their purchases and collect them in the store. Companies are increasingly relying on data analysis for their operations, including product positioning, inventory optimization and forecasting of supply and demand.

The use of technological tools in the sales sector, through both traditional channels and e-commerce websites, favors the collection of Big Data. These data contain valuable information for companies, which can leverage them to develop effective business strategies. The large amount of data, their heterogeneity and the speed they are collected with,

force the use of Big Data Analytics methods in order to mine this information [1]. In particular, machine learning provides several methods for optimizing the supply chain and product-sourcing operations, for both marketing and customer loyalty. Starting from data collected on product sales and consumer behavior, machine learning can offer highly personalized recommendations on the purchase of products. For example, customers can receive personalized content and product suggestions that could catch the customer interest. The analysis of historical data using machine learning makes it possible to evaluate a variety of factors in order to optimize the choice for suppliers [2,3] and the product-delivery routes for greater logistical efficiency [4,5]. The automated customer segmentation is very useful for maximizing sales, as it allows companies to understand the specific needs of customers and to communicate with them in a personalized way. For example, clustering analysis permits the identification of the best customers, encouraging them to make new purchases through targeted email marketing campaigns.

Machine learning approaches are used in various technical and scientific applications for their ability to automate processes and improve performances. The research [6] presents a methodology, based on k-means clustering, which can be applied to automatically control the quality of various industrial production processes, such as turning, glazing and laser cutting. Authors [7] propose an approach that uses machine learning algorithms to process rock tunnel face images. The developed methodology is used to segment rock fractures and quantify the fracture traces. One study [8] discusses an automatic approach based on machine learning algorithms, such as a gradient boosting tree, for the automatic classification of rock traces. An added value of machine learning lies in its predictive nature. In fact, in addition to being able to detect the freshness of perishable goods [9], leakage in water-piping networks [10] and the wear of machinery [11], machine learning-based technologies are able to predict the propagation rate of the plume in underground water leakages [12], and customer behavior, future trends and market demand [13].

The paper is organized as follows. In this introductory section, the context in which the research is carried out and its objectives are briefly discussed, highlighting the impact of modern technological tools on sales. In the next section, the focus is on machine learning and its application to sales forecasting in different fields. In Section 3, the need to apply machine learning methods to sales prediction is highlighted. Furthermore, the motivations for the present research study and the contributions of the proposed predictive model are presented. With the aim of providing useful information for understanding the developed methodology, Section 4 describes in depth the algorithms used in the research, also providing some mathematical details. Then, the implementation of the implemented predictive models is discussed. Section 5 reports and interprets the experimental results of some preliminary tests. In particular, the different approaches are compared using different accuracy metrics. Finally, Section 6 draws the conclusions of the research and sets out future developments.

2. Application of Machine Learning to Sales Prediction

In recent years, machine learning has been applying more and more frequently to sales prediction, due to the different approaches it makes available based on the particular case study [14].

Machine learning allows the implementation of systems capable of automatically learning from data. The system processes historical data as input, in which it searches for repetitive models in order to extract patterns that allow it to predict future purchases and make better decisions. Therefore, the machine learning algorithm does not require any human programming as it automatically generates a model based on the data that it has analyzed. The main advantage of machine learning methods is the ability to process large amounts of data from various sources. However, the result obtained depends on the quality and quantity of the available data.

Machine learning algorithms need to be trained in order to provide reliable output. The main types of training approaches are supervised learning, reinforcement learning

and unsupervised learning [15]. In the case of supervised learning, a model is built by using a labeled training dataset. During training, the model processes input data where the output signals of the data are known. In this type of training, the model learns to generalize and will be able to match the correct output to new input data. Reinforcement learning seeks to develop a system that improves its performance based on interactions with the environment. In this case, the information about the current state of the environment does not come from a label, but the feedback is provided by a reward signal. During this interaction with the environment, the agent performs some actions that receive a reward (positive real value) or a penalty (negative real value). If the agent has approached the target after the action, the feedback is positive and the reinforcement function awards a reward. Conversely, when the agent moves away from the target after the action, the feedback is negative and a penalty is assigned. Through the interaction with the environment, the agent takes an exploratory approach during which it learns a series of actions that maximize the reward. Finally, in the case of unsupervised learning, the model cannot be trained on a set of data prepared with the correct corresponding output, but it must autonomously identify the differences or similarities between the inputs by identifying their main characteristics. Therefore, the algorithm must discover any existing relationships.

Each machine learning forecasting method is characterized by different strengths and weaknesses, such as stability, responsiveness, the amount of data needed, the planning time horizon and computational efficiency. For this reason, sometimes a combination of multiple methods can achieve the best results. Authors [16] present a sales forecasting approach that combines different clustering techniques, such as a self-organizing map (SOM), a growing hierarchical self-organizing map (GHSOM) and k-means, with different machine learning methods, such as support vector regression (SVR) and extreme learning machine (ELM). The resulting models are discussed and compared for computer product sales forecasting. Authors [17] discuss the use of three different machine learning approaches, such as Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boosting (GB), to predict sales. The comparative study of these different approaches shows that the GB model achieves better performance and greater accuracy. In particular, Root Mean Square Error (RMSE), Mean Square Error (MSE) and Absolute Error (AE) are calculated for the various approaches.

In the field of sales prediction, methodologies based on approaches that implement boosting algorithms are of particular importance, due to the accuracy of the predictions [18–20]. In one study [21], an extreme gradient boosting (XGBoost) algorithm is used to implement a predictive model applied to the forecast of sales in the large-scale retail sector. The discussed method is tested on the prediction of various products and validated by comparing the predicted values with real data. Accuracy is measured by mean of MSE and RMSE errors.

The ability to predict the occurrence of various business problems helps companies to act early, so that they can limit loss, and continue to profit. One study [22] discusses two different machine learning approaches that have been developed in order to predict back orders of a product. In detail, Distributed Random Forest (DRF) and GB algorithms are used. Models based on these two approaches show similar performance in predicting early back orders of products. The comparison has been made by using the Receiver Operating Characteristics (ROC) curve, the classification accuracy and the confusion matrix.

The authors of the research [23] apply differently various machine learning approaches in order to forecast costs in the green construction industry. In particular, models based on XGBoost algorithms, deep neural networks (DNN) and random forest (RF) have been developed and compared. The XGBoost model has been proved more accurate than the other models in terms of different performance evaluation metrics, such as MAE (mean absolute error), MSE, MAPE (mean absolute percentage error) and R2 (coefficient of determination).

3. Contribution of the Model

The digital transformation, which has affected various aspects of human society, is bringing about changes in the sales sector. Technological innovation has led to the global success of e-commerce and multi-channel sales. Innovative tools provide new opportunities that companies active in the sector must seize to remain competitive on the market. In particular, modern methodologies based on machine learning algorithms make it possible to process the large amount of data that companies collect. The information extracted during data processing makes the company capable of devising winning market strategies and optimizing its processes. The prediction of sales is fundamental as it allows the optimization of warehouse management and processes related to the supply and sale of products. Various machine learning methodologies are spreading in modern research, but it is not yet possible to determine which is the best performing approach. Moreover, to the difficulty of identifying the appropriate model is added the awareness that there is no universally valid approach [13,14]. For this reason, it is necessary to carry out research in specific case studies, comparing different methodologies.

The research proposed by this paper fits into this context, comparing two different implementations of boosting technologies in an interesting specific case study. The main purpose of the paper is to provide a multi-store company with an accurate tool for sales forecasting. Despite the numerous studies that deal with forecasting sales, there is no specific study that addresses the problem of predicting sales in a multi-store e-commerce, in a multi-channel sale or in a chain of stores. This research seeks to fill this gap by implementing and preliminarily testing an agile and flexible method that can be applied to different stores. For example, in the case of a multi-store e-commerce, the developed algorithm can be trained on a different dataset for each store, thus creating a different predictive model for each store. In fact, the implemented XGBoost-based approach offers the possibility of tuning its hyperparameters on the specific dataset and creating a specific model for each prediction.

In the same way, it is possible to create a model using the historical data of two or more stores, in order to manage the sales of the corresponding sales channels or warehouses as a whole. This opportunity to train the model on a single store or on a selection of stores makes it flexible and adaptable to different situations. For example, in the case of a new product or an item that is rarely sold, there are not many data available, so it is convenient to train the model on a single training dataset containing the data of the various stores in order to have more precise information on the demand for the product. In the case of a new product, the time series of a similar product can also be used for training.

4. Methodology

In this paper we present a study related to an industrial project that aims to equip a company, active in the e-business sector, with innovative digital tools for the optimization of the stores management. Figure 1 shows the architecture of the system that has been designed to allow the company to automatically acquire electronic orders from the various company stores. The web service system permits the stores to interact with the headquarters. In fact, the web service software acquires the json files from the individual stores and deposits them in csv format in the Data Base. A technological platform processes the data collected using a predictive model based on machine learning algorithms. The company's headquarters, by logging into the platform, has the possibility to analyze the output obtained by the predictive algorithms on a dashboard including graphs and tables. The focus of this paper is about the implementation and testing of the machine learning model.

There are currently several methodologies for building predictive models. Decision trees (DT) are used to rank input data points or to predict values of output variables based on inputs supplied to the model [24–26]. Starting from DT, further performing techniques have been developed, such as the various boosting techniques.

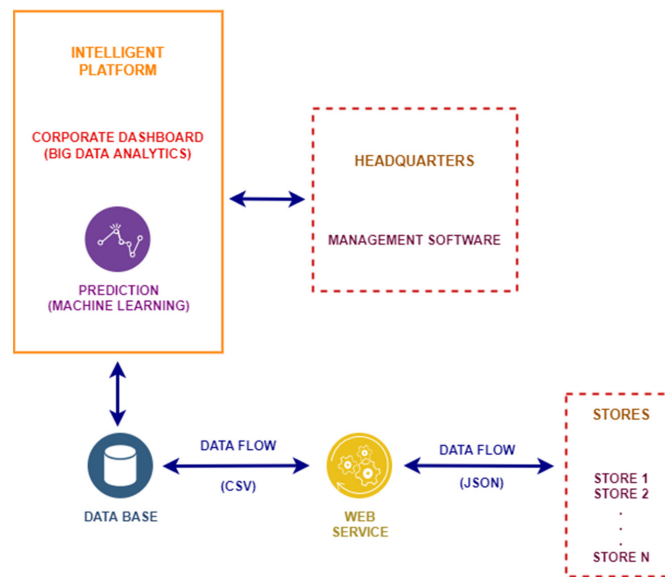


Figure 1. Functional scheme of the proposed intelligent system.

4.1. Boosting Approaches

The first version of the boosting algorithm was Adaptive Boosting (AdaBoost), which builds a prediction model based on an ensemble of weak models [27]. In fact, a strong classifier $G(x)$ is constructed as a linear combination of weak classifiers in the following form:

$$G_N(x) = \sum_{i=1}^N \alpha_i g_i(x) = G_{N-1}(x) + \alpha_N g_N(x)$$

where x represents the input data, $g_i(x)$ is the prediction of the i -th weak classifiers and N denotes the number of iterations performed by the algorithm. In the generic i -th iteration, the corresponding weak learner is selected and the weight α_i is assigned in order to minimize the error [28], so finally, the output $\alpha_i g_i(x)$ is computed. The N parameter is established in the training step. Unfortunately, the AdaBoost is afflicted by noise and, consequently, by overfitting.

The Gradient Boosting (GB) algorithm is an evolution of the AdaBoost and is based on the calculation of the so-called residuals [29]. The residual is the difference between the real value y and the approximate value $\hat{y} = F(x)$ that is predicted [30]. The goal of the algorithm is to minimize the loss function $L(y, F(x))$, which can be defined through MSE in the following form:

$$L(y, F(x)) = \frac{1}{I} \sum_{i=1}^I (y_i - \hat{y}_i)^2$$

where I indicates the instance number.

The model parameters are I and the number M of trees. These parameters are varied in order to minimize the following pseudo-residual:

$$r_{im} = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$$

where $i = 1, \dots, I$ e $m = 1, \dots, M$.

The number M of trees, which coincides with the number of gradient boosting iterations in the model, is a regularization parameter. In fact, by increasing M it is possible to reduce the error during the training phase. On the other hand, a too-high M value causes overfitting, which worsens the model's ability to generalize [31,32].

XGBoost is a scalable technology that optimizes the boosting concept underlying the GB algorithm [33]. This efficient algorithm allows the implementation of a predictor with

excellent mathematical ability and with reduced computational costs [21,23]. XGBoost is effective and flexible due to the various hyperparameters [34]. As in the case of the GB algorithm, XGBoost is based on the iteration of N steps that allow the writing of the function $F_i^{(N)}$, which represents the prediction, in the following form:

$$F_i^{(N)} = \sum_{k=1}^N f_k(x_i) = F_i^{(N-1)} + f_N(x_i)$$

where x_i denotes the i -th input feature, $F_i^{(N-1)}$ denotes the prediction at the step $(N-1)$ and $f_N(x_i)$ denotes the learner at the step N .

The objective function, which provides an estimate of the model ability, is the sum of two contributions, that is, the training loss and the regularization term:

$$Obj^{(N)} = \sum_{i=1}^M L(y_i - \hat{y}_i) + \sum_{k=1}^N \sigma(f_k)$$

where L is the loss function and M is the number of the implemented observations. Here, $\sigma(f_i)$ expresses the complexity of the tree f_i . The regularization term controls the complexity of the model and takes the following form:

$$\sigma(f) = \gamma T + \frac{\lambda ||\omega||^2}{2}$$

where T is the number of the leaves of the tree and $\eta\omega$ is the output scores of the leaves, so that $||\omega||^2 = \sum_{j=1}^T \omega_j^2$. The value γ controls the minimum loss reduction necessary to further divide the leaf node and λ represents the regularization parameters. This regularization mechanism, which is not present in the standard GB algorithm, allows XGBoost to significantly reduce overfitting.

The generic unregularized XGBoost algorithm can be expressed through the pseudo code that is shown in Algorithm 1.

Algorithm 1 Pseudocode of the unregularized XGBoost algorithm

XGBoost Algorithm

Input: $\{(x_j, y_j)\}_{j=1}^M$ training set

Input: N tree (weak learners)

Input: $L(y, F(x))$ differentiable loss function

Input: α learning rate

Model initialization by using a constant value: $F_i^{(0)}(x) = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^M L(y_j, \theta)$

for $k = 1$ to N

1. Compute gradients and Hessians:

$$g_k(x_j) = \left[\frac{\partial L(y_j, f(x_j))}{\partial f(x_j)} \right]_{f(x)=f_{k-1}(x)}$$

$$h_k(x_j) = \left[\frac{\partial^2 L(y_j, f(x_j))}{\partial f(x_j)^2} \right]_{f(x)=f_{k-1}(x)}$$

2. Fit a tree by means of the training set $\left\{ \left(x_j, -\frac{g_k(x_j)}{h_k(x_j)} \right) \right\}_{j=1}^M$ by solving the following optimization problem:

$$\phi_k = \underset{\phi \in \Phi}{\operatorname{argmin}} \sum_{j=1}^M \frac{1}{2} h_k(x_j) \left[-\frac{g_k(x_j)}{h_k(x_j)} - \phi(x_j) \right]^2$$

$$f_k(x) = \alpha \phi_k(x)$$

3. Update the model:

$$F_i^{(k)}(x) = F_i^{(k-1)}(x) + f_k(x)$$

end

Output: $F_i^{(N)} = \sum_{k=0}^N f_k(x)$

XGBoost is a high-performance algorithm, in fact, it can be implemented on parallel architectures and its training is very fast [21,35–37]. Additionally, XGBoost has an embedded routine that optimizes the management of missing values.

4.2. Implementation of the Models

The boosting techniques that are based on an ensemble of weak predictive models allow the implementation of very accurate predictive models characterized by a small error from the comparison between real and predicted values [23,38].

In the present study, two different models based, respectively, on GB and XGBoost, are implemented and compared, with the aim of forecasting the daily orders of the various products in each branch of the company involved in the research project. The models have been trained using historical sales data.

The training process leads to the generation of a model created for a specific objective. The first step is the definition of the target that consists of a careful analysis of the input data, in which the input variables and the output target variable of the model are identified. In this phase, supervised training takes place; in fact, the algorithms process examples consisting of pairs of inputs and outputs. During this procedure, the relationship that links the input variables with the output variable is extrapolated, and a model is created capable of processing new system inputs by automatically generating forecasts on the output values. Item Code, Ordered Quantity, Customer Code and Branch Code have been chosen as input variables. Additionally, Date is the output variable, which will be linked to the predicted quantity.

For each branch, a file containing historical data has been generated (sales point) and two customized regression models have been created, which will be invoked in the prediction phase of a given order so as to evaluate the predictive behavior with respect to real behavior. The regression predictive models, GB regressor (GBR) and XGBoost regressor (XGBR), are able to predict the quantity of a product that will be sold in each branch in the next 7 days, provided a sufficiently rich number of historical data are available. In case the dataset is poor, it is possible to use the Augmented Data technique already used in [21,39] to enrich insufficient data. In fact, in addition to the quality of the data, the quantity of data is also a factor that affects the method precision.

The available data have been used to build datasets for each branch. In order to prevent the overfitting, a cross-validation technique has been also applied, so that 80% of the data have been used to build the corresponding training dataset, and the remaining 20% has been adopted as a testing dataset useful for model validation. The data selection is random, by selecting 80% of each product sold in each branch. During the training, the model learns to correlate the various features and becomes able to generalize and, therefore, to predict the quantity of product sold when new data, different from the training data, are provided as input.

Python programming language has been used to implement the GBR and the XGBR algorithms. This language has been chosen because it provides very powerful open-source tools and libraries, such as scikit-learn and XGBoost. In particular, the XGBRegressor and GridSearchCV classes of scikit-learn have been used for implementing the XGBR model. GridSearchCV is a fundamental class for automatically obtaining an accurate prediction, as it allows an optimal choice of the hyperparameters to be made, which improves the computational efficiency and the prediction accuracy of the model [21,40]. The main tunable XGBoost parameters are learning_rate, which is denoted by eta, n_estimators, max_depth, colsample_bytree, gamma, min_child_weight, alpha and lambda [41–43]. Eta is the learning rate and affects how quickly the model fits the residual error. N_estimators indicates the number of decision trees of the model. Max_depth denotes the maximum depth of a tree. Colsample_bytree refers to the features fraction that is sampled for constructing each tree. Gamma states exactly the minimum loss reduction that is required to make a node split. Min_child_weight determines the minimum sum of weights of all observations that is required in a child. Alpha and lambda control the regularization on leaf weight, respec-

tively, by Lasso (L1 regularization) and Ridge (L2 regularization) techniques. Max_depth, min_child_weight and lambda are also used to control overfitting.

Algorithm 2 shows the algorithm useful for building a tree in the XGBoost regression model. Two tools are introduced, namely the similarity scores SS and the gain G . SS optimizes the tree growth, while G helps the various children to better split. Tree pruning is performed as a function of the difference between G and gamma, which is a tree complexity parameter set by the user. If this difference assumes a positive value, the tree is not pruned. On the other hand, if the difference is negative, pruning is performed and, again, gamma is subtracted from the next gain value way up the tree. Finally, the output value is computed for the remaining leaves.

Algorithm 2 Pseudocode for building the XGBoost tree for regression

XGBoost Regression Tree Algorithm

Compute the similarity score: $SS = \frac{(\text{Sum of residuals})^2}{\text{Number of residuals}} + \text{lambda}$

Compute the gain: $G = \text{Left tree } (SS) + \text{Right tree } (SS) - \text{Root } (SS)$

Prune the tree:

1. Compute $D = G - \text{gamma}$
2. while $D < 0$

prune and update the value of D

3. Compute output value for the remaining leaves: $\text{Output value} = \frac{\text{Sum of residuals}}{\text{Number of residuals}} + \text{lambda}$
-

5. Preliminary Testing

The model has been validated through tests conducted on the prediction of several products that catch company interest.

Figure 2a,b show the prediction of the sale of a given product in a specific branch performed, respectively, by GBR and XGBR. The prediction data (orange plot) are compared with the real values (blue plot).

The qualitative comparison of Figure 2a,b shows greater accuracy of the XGBoost model. The suitability of this last method is confirmed by the analysis carried out by the main accuracy metrics.

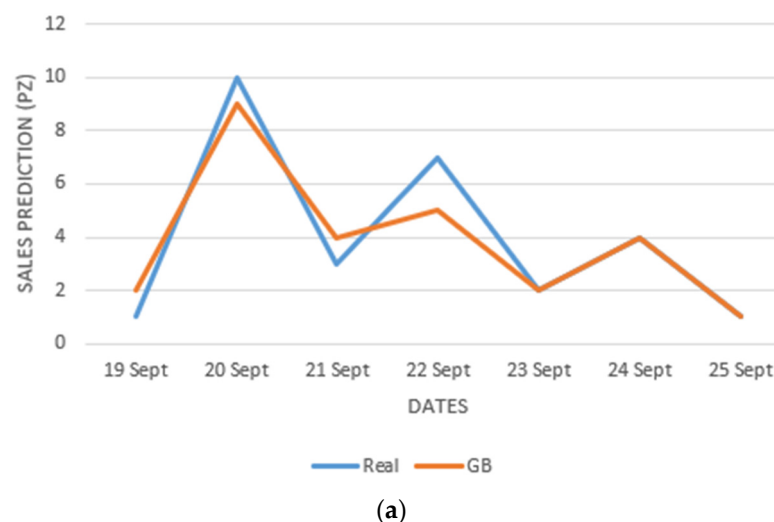


Figure 2. Cont.

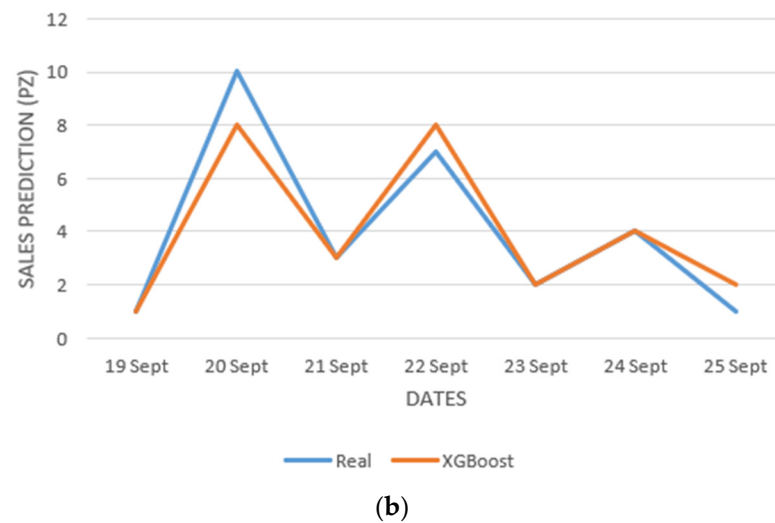


Figure 2. Predictive model test for a specific product and branch: (a) Gradient Boosting; (b) XGBoost.

In detail, the metrics [44,45] used to evaluate the accuracy of the two developed models are the following:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

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

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\%$$

$$\text{WAPE} = \frac{1}{n} \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i|} 100\%$$

WAPE (weighted absolute percentage error) is very important in the calculation of the forecast error, in fact it is the absolute error averaged over the real requested quantity [46]. Therefore, compared to MAPE, WAPE allows the estimation in a more balanced way of the impact of the prediction on sales and profits [47,48]. MAPE is a good metric for measuring forecast errors, but it is not indicative in the case in which sales are intermittent or assume values close to zero. WAPE is a metric that overcomes this problem by weighting the error on total sales.

Figure 3 shows the values of the accuracy obtained for the test reported in Figure 2. The XGBoost model allows more accurate results to be obtained.

Table 1 summarizes the meaning and values of the main hyperparameters that were used during the run of Figure 2b. These values were automatically tuned by using Grid-SearchCV. Regarding the parameters of the GBR model, the following choices were made: $n_estimators = 500$, $max_depth = 4$, $min_samples_split = 2$, $learning_rate = 0.01$.

Table 1. Setting of hyperparameters for the developed XGBoost model.

Hyperparameter	Value	Meaning
Eta (learning_rate)	0.1	Shrinkage coefficient of each tree
N_estimators	215	Number of estimators
Max_depth	4	Maximum depth of the tree
Colsample_bytree	0.9	Subsample ratio of columns for each tree
Gamma	0	Minimum loss function reduction to make new tree-split
Min_child_weight	1	Minimum sum of weights in a child
Alpha	0	L1 regularization term on leaf weights
Lambda	1	L2 regularization term on leaf weights

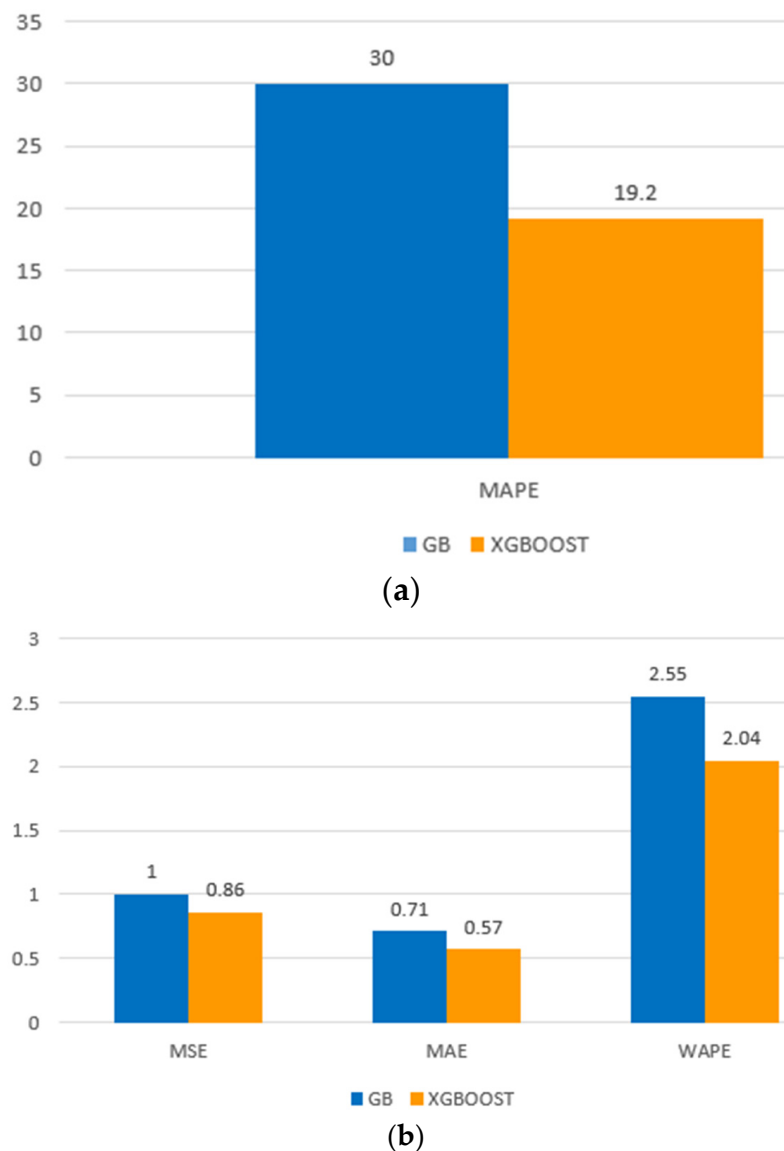


Figure 3. Accuracy metrics referred to the validation test of the GB and XGBoost models: MAPE (a); MSE, MAE and WAPE (b).

A second test was performed on a new product recently launched in two different stores, which stock up in the same warehouse, see Figure 4 for the test results. The predictions carried out by the XGBoost model are now less accurate than in the previous test, as the training dataset is less rich. The model fails to predict sales trends well. Figure 4c shows the results of the prediction performed with the XGBoost model trained on data collected from both stores. In this case, the prediction is more accurate. For completeness, Figure 5 also shows the comparison with the prediction reached using the GB model trained on the same assembled dataset. As in the test shown in Figure 2, the XGBoost model proves to be performing better than the GB model. In Figure 5, the accuracy metrics confirm the considerations made for this second test.

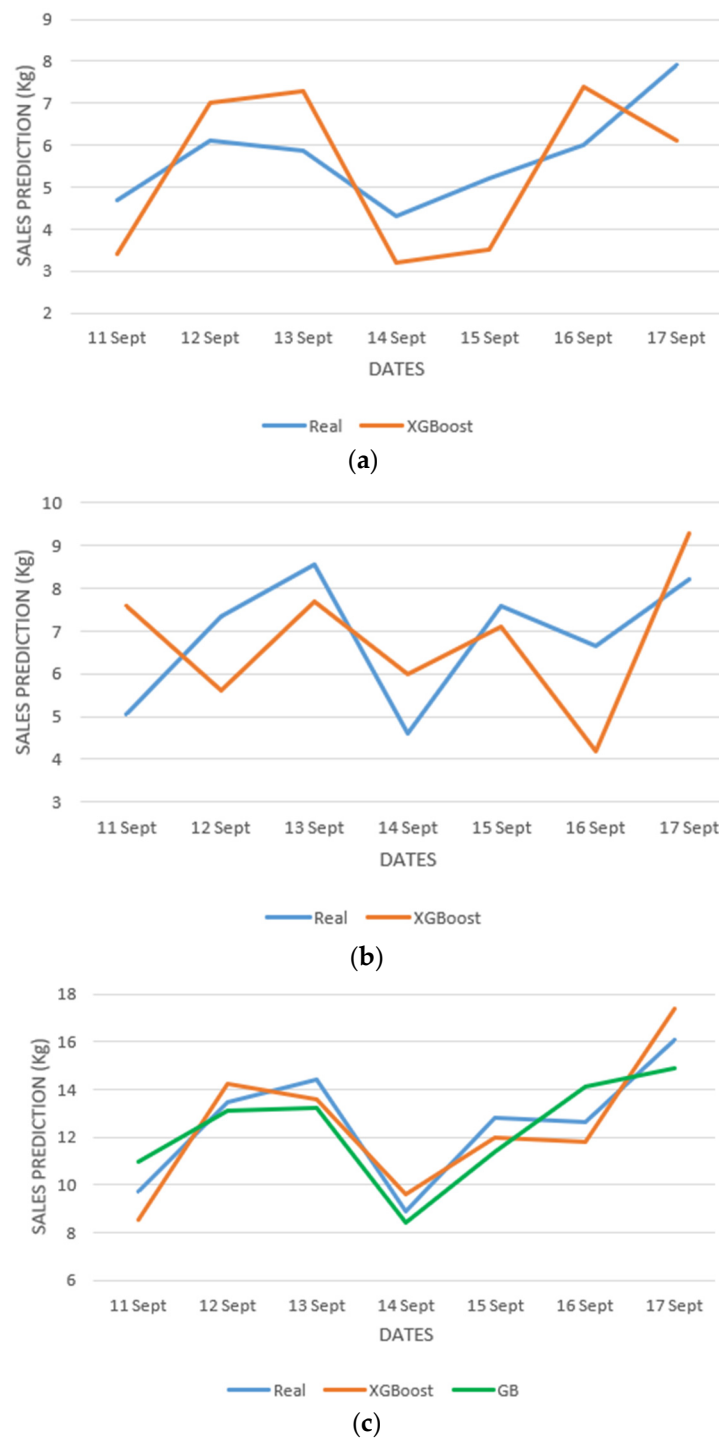
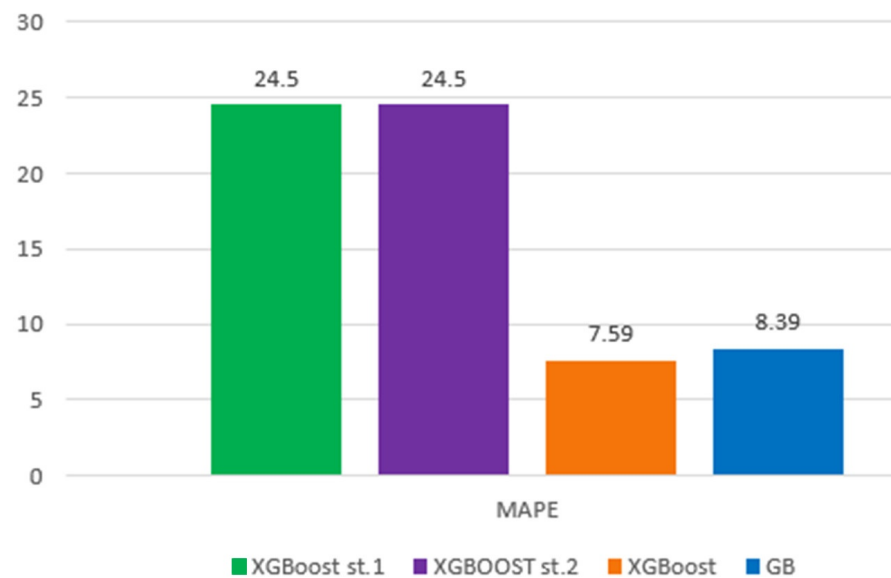
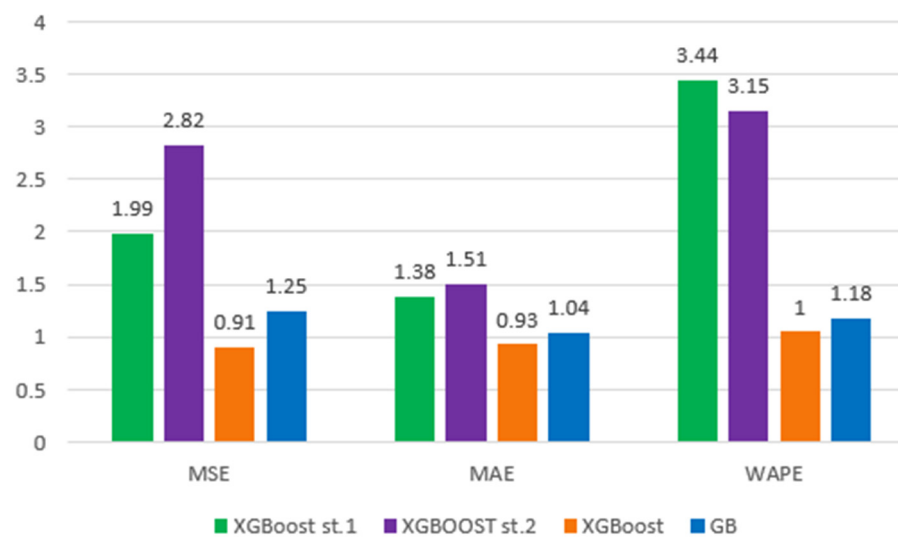


Figure 4. Predictive model test for a product just launched on the market and sold in two stores, indicated with st.1 and st.2: (a) XGBoost prediction for st.1; (b) XGBoost prediction for st.2; (c) comparison of XGBoost and GB prediction for the overall sales in st.1 and st.2.



(a)



(b)

Figure 5. Accuracy metrics referred to the second validation test: MAPE (a); MSE, MAE and WAPE (b).

6. Conclusions

A prototype platform is developed to address the difficult task of managing a trading company, which consists of several branches and warehouses, by means of automatic digital tools. The large amount of data collected, which needs to be processed, makes this task very challenging. The goal of this research is to develop and test two accurate and prediction models based on a gradient boosting algorithm. For this purpose, two different implementations are made, namely GB regressor and XGBoost regressor. Their comparison by mean of various accuracy metrics (MAE, MSE, MAPE, WAPE) highlights that XGBoost is the best performing algorithm. Preliminary tests are carried out for two specific products of a given point of sale. In the first test, the models are applied to the prediction of the sales of a product already on the market. In the second test, the behavior of the XGBoost model in predicting the sales of a product recently launched on the market is studied. In this case, better results are obtained by training the model on a dataset that includes data from two stores belonging to the same company and using the same warehouse. In the subsequent

research phase, the method will be tested on the prediction of the sale of products sold in small quantities. In this case, the AD technique will be used in order to increase accuracy.

Author Contributions: Conceptualization, A.P., G.S. and V.V.; methodology, A.P. and G.S.; software, V.V., A.P. and G.S.; validation, A.P., G.S. and A.G.; formal analysis, A.P. and G.S.; investigation, A.P. and G.S.; resources, A.G.; data curation, V.V., A.P. and G.S.; writing—original draft preparation, A.P.; writing—review and editing, A.G., G.S. and A.P.; visualization, A.P. and G.S.; supervision, G.S.; project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable, the study does not report any data.

Acknowledgments: The proposed work has been developed within the framework of the industry project titled: “Piattaforma avanzata di gestione ordini e di interfacciamento gestionale orientata all’analytics Big Data: W-ORD”. [Advanced order management and management interfacing platform oriented to Big Data analytics: W-ORD].

Conflicts of Interest: There are no conflicts of Interest.

References

- Boone, T.; Ganeshan, R.; Jain, A.; Sanders, N.R. Forecasting sales in the supply chain: Consumer analytics in the big data era. *Int. J. Forecast.* **2019**, *35*, 170–180. [\[CrossRef\]](#)
- Islam, S.; Amin, S.H.; Wardley, L.J. Machine learning and optimization models for supplier selection and order allocation planning. *Int. J. Prod. Econ.* **2021**, *242*, 108315. [\[CrossRef\]](#)
- Cavalcante, I.M.; Frazzon, E.M.; Forcellini, F.A.; Ivanov, D. A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *Int. J. Inf. Manag.* **2019**, *49*, 86–97. [\[CrossRef\]](#)
- Snoeck, A.; Merchán, D.; Winkenbach, M. Route learning: A machine learning-based approach to infer constrained customers in delivery routes. *Transp. Res. Procedia* **2020**, *46*, 229–236. [\[CrossRef\]](#)
- Tarapata, Z.; Nowicki, T.; Antkiewicz, R.; Dudzinski, J.; Janik, K. Data-Driven Machine Learning System for Optimization of Processes Supporting the Distribution of Goods and Services—A case study. *Procedia Manuf.* **2020**, *44*, 60–67. [\[CrossRef\]](#)
- Massaro, A.; Panarese, A.; Dipierro, G.; Cannella, E.; Galiano, A.; Vitti, V. Image Processing Segmentation applied on Defect Estimation in Production Processes. In Proceedings of the IEEE International Workshop on Metrology for Industry 4.0 & IoT, Rome, Italy, 3–5 June 2020; pp. 565–569. [\[CrossRef\]](#)
- Chen, J.; Zhou, M.; Huang, H.; Zhang, D.; Peng, Z. Automated extraction and evaluation of fracture trace maps from rock tunnel face images via deep learning. *Int. J. Rock Mech. Min. Sci.* **2021**, *142*, 104745. [\[CrossRef\]](#)
- Chen, J.; Huang, H.; Cohn, A.G.; Zhang, D.; Zhou, M. Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning. *Int. J. Min. Sci. Technol.* **2021**, *32*, 309–322. [\[CrossRef\]](#)
- Massaro, A.; Panarese, A.; Galiano, A. Infrared Thermography applied on Fresh Food Monitoring in Automated Alerting Systems. In Proceedings of the IEEE International Workshop on Metrology for Industry 4.0 & IoT, Rome, Italy, 3–5 June 2020; pp. 554–558. [\[CrossRef\]](#)
- Massaro, A.; Panarese, A.; Galiano, A. Technological Platform for Hydrogeological Risk Computation and Water Leakage Detection based on a Convolutional Neural Network. In Proceeding of the IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT), Rome, Italy, 7–9 June 2021; pp. 225–230. [\[CrossRef\]](#)
- Mateus, B.C.; Mendes, M.; Farinha, J.T.; Cardoso, A.M. Anticipating Future Behavior of an Industrial Press Using LSTM Networks. *Appl. Sci.* **2021**, *11*, 6101. [\[CrossRef\]](#)
- Massaro, A.; Panarese, A.; Selicato, S.; Galiano, A. CNN-LSTM Neural Network Applied for Thermal Infrared Underground Water Leakage. In Proceeding of the IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT), Rome, Italy, 7–9 June 2021; pp. 219–224. [\[CrossRef\]](#)
- Ensafi, Y.; Amin, S.H.; Zhang, G.; Shah, B. Time-series forecasting of seasonal items sales using machine learning—A comparative analysis. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100058. [\[CrossRef\]](#)
- Erjiang, E.; Yu, M.; Tian, X.; Tao, Y. Dynamic Model Selection Based on Demand Pattern Classification in Retail Sales Forecasting. *Mathematics* **2022**, *10*, 3179. [\[CrossRef\]](#)
- Raschka, S. *Python Machine Learning*; Packt Publishing Ltd.: Birmingham, UK, 2015.
- Chen, I.-F.; Lu, C.-J. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Comput. Applic.* **2017**, *28*, 2633–2647. [\[CrossRef\]](#)

17. Cheriyan, S.; Ibrahim, S.; Mohanan, S.; Treesa, S. Intelligent Sales Prediction Using Machine Learning Techniques. In Proceedings of the 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, UK, 16–17 August 2018; pp. 53–58. [\[CrossRef\]](#)
18. Wisesa, O.; Adriansyah, A.; Khalaf, O.I. Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm. In Proceedings of the 2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP), Yogyakarta, Indonesia, 28–30 September 2020; pp. 101–106. [\[CrossRef\]](#)
19. Zhou, F.; Zhang, Q.; Sornette, D.; Jiang, L. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Appl. Soft Comput.* **2019**, *84*, 105747. [\[CrossRef\]](#)
20. Korolev, M.; Stanford University, Srandford, CA, USA; Ruegg, K.; Harvard University, Cambridge, MA, USA. Gradient Boosted Trees to Predict Store Sales. Personal communication, 2015.
21. Massaro, A.; Panarese, A.; Giannone, D.; Galiano, A. Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector. *Appl. Sci.* **2021**, *11*, 7793. [\[CrossRef\]](#)
22. Islam, S.; Amin, S.H. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. *J. Big Data* **2020**, *7*, 65. [\[CrossRef\]](#)
23. Alshboul, O.; Shehadeh, A.; Almasabha, G.; Almuflih, A.S. Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. *Sustainability* **2022**, *14*, 6651. [\[CrossRef\]](#)
24. Saiyin, X.; Hu, C.; Tan, D.; Liu, Y. Research on Apparel Sales Forecast Based on ID3 Decision Tree Algorithm. In Proceedings of the 3rd International Conference on Mechatronics and Industrial Informatics, Zhuhai, China, 30–31 October 2015; Atlantis Press: Amsterdam, The Netherlands; pp. 704–709.
25. Lytvynenko, T.I. Problem of data analysis and forecasting using decision trees method. *Probl. Program.* **2016**, *2–3*, 220–226. [\[CrossRef\]](#)
26. Johannes, R.; Alamsyah, A. Sales Prediction Model Using Classification Decision Tree Approach for Small Medium Enterprise Based on Indonesian E-Commerce Data. *arXiv* **2021**, arXiv:2103.03117.
27. Stamp, M. *Introduction to Machine Learning with Applications in Information Security*; Chapman & Hall: London, UK; CRC Press: Boca Raton, FL, USA, 2017.
28. Rojas, R. AdaBoost and the Super Bowl of Classifiers: A Tutorial Introduction to Adaptive Boosting. 2009. Available online: <https://www.inf.fu-berlin.de/inst/ag-ki/adaboost4.pdf> (accessed on 27 October 2022).
29. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
30. De’Ath, G. Boosted trees for ecological modeling and prediction. *Ecology* **2007**, *88*, 243–251. [\[CrossRef\]](#)
31. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 308–324. [\[CrossRef\]](#)
32. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurobot.* **2013**, *7*, 21. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
34. Liu, Q.; Zhang, M.; He, Y.; Zhang, L.; Zou, J.; Yan, Y.; Guo, Y. Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques. *J. Pers. Med.* **2022**, *12*, 905. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Lau, R.Y.K.; Zhang, W.; Xu, W. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Prod. Oper. Manag.* **2018**, *27*, 1775–1794. [\[CrossRef\]](#)
36. Zhang, L.; Bian, W.; Qu, W.; Tuo, L.; Wang, Y. Time series forecast of sales volume based on XGBoost. *J. Phys. Conf. Ser.* **2021**, *1873*, 012067. [\[CrossRef\]](#)
37. Panarese, A.; Bruno, D.; Tolia, P.; Ratynskaia, S.; Longo, S.; De Angelis, U. Molecular Dynamics Calculation of the Spectral Densities of Plasma Fluctuations. *J. Plasma Phys.* **2018**, *84*, 905840308. [\[CrossRef\]](#)
38. Hastie, T.; Tibshirani, R.; Friedman, J.H. Boosting and Additive Trees. In *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009; pp. 337–384, ISBN 978-0-387-84857-0.
39. Massaro, A.; Maritati, V.; Giannone, D.; Convertini, D.; Galiano, A. LSTM DSS Automatism and Dataset Optimization for Diabetes Prediction. *Appl. Sci.* **2019**, *9*, 3532. [\[CrossRef\]](#)
40. Qin, C.; Zhang, Y.; Bao, F.; Zhang, C.; Liu, P.; Liu, P. XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring. *Math. Probl. Eng.* **2021**, *2021*, 6655510. [\[CrossRef\]](#)
41. Ji, Q.; Zhang, S.; Duan, Q.; Gong, Y.; Li, Y.; Xie, X.; Bai, J.; Huang, C.; Zhao, X. Short- and Medium-Term Power Demand Forecasting with Multiple Factors Based on Multi-Model Fusion. *Mathematics* **2022**, *10*, 2148. [\[CrossRef\]](#)
42. Zhao, X.; Li, Q.; Xue, W.; Zhao, Y.; Zhao, H.; Guo, S. Research on Ultra-Short-Term Load Forecasting Based on Real-Time Electricity Price and Window-Based XGBoost Model. *Energies* **2022**, *15*, 7367. [\[CrossRef\]](#)
43. Yang, J.; Guan, J. A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm. *Information* **2022**, *13*, 475. [\[CrossRef\]](#)
44. Jierula, A.; Wang, S.; OH, T.-M.; Wang, P. Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. *Appl. Sci.* **2021**, *11*, 2314. [\[CrossRef\]](#)
45. Sani, U.S.; Malik, O.A.; Lai, D.T.C. Improving Path Loss Prediction Using Environmental Feature Extraction from Satellite Images: Hand-Crafted vs. Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 7685. [\[CrossRef\]](#)

-
46. Auppakorn, C.; Phumchusri, N. Daily Sales Forecasting for Variable-Priced Items in Retail Business, 2022, Association for Computing Machinery. In Proceedings of the 4th International Conference on Management Science and Industrial Engineering, Chiang Mai, Thailand, 28–30 April 2022; pp. 80–86. [\[CrossRef\]](#)
 47. Chase, C. *Demand-Driven Forecasting: A Structured Approach to Forecasting*; (Wiley and SAS Business Series); Wiley: Somerset, NJ, USA, 2013; pp. 83–84, 104–105, 113–115.
 48. Louhichi, K.; Jacquet, F.; Butault, J. Estimating input allocation from heterogeneous data sources: A comparison of alternative estimation approaches. *Agric. Econ. Rev.* **2012**, *13*, 91.