engineering proceedings

MDPI

Proceeding Paper

Leveraging Data Science for Impactful Logistics Automatons in the E-Commerce Sector †

Nabila Belhaj [1,*], Jérémy Patrix [2], Ouail Oulmakki [3] and Jérôme Verny [3]

[1] HighFi Lab, HighSys, 2 Rue Mozart, 92110 Clichy, France
[2] Opal Research, 4 Rampe Cauchoise, 76000 Rouen, France; jeremy.patrix@gmail.com
[3] NEOMA Business School, 1 Rue du Marechal Juin, 76130 Mont-Saint-Aignan, France; ouail.oulmakki@neoma-bs.fr (O.O.); jerome.verny@neoma-bs.fr (J.V.)
* Correspondence: nb.belhaj@gmail.com
† Presented at the 1st International Conference on Smart Management in Industrial and Logistics Engineering (SMILE 2025), 16–19 April 2025, Casablanca, Morocco.

Abstract: Automation technologies play a pivotal role in optimizing logistics operations within warehouse facilities. Retail companies invest in these technologies to maintain pace with the customers demands by increasing their production capacity while reducing their financial expenses. In this paper, we conduct a study on warehouse automation in the European e-commerce sector by analyzing historical data from three fulfillment centers. Accordingly, we explore diverse data science approaches applied to trained machine learning models to determine the automatons that have the greatest impact on financial costs. The purpose is to support supply chain managers in identifying the most profitable logistics automatons that merit consideration in future automation projects. The study offers a comprehensive analysis that encourages e-commerce companies to invest in tailored automation for future warehouse installations.

Keywords: data science; machine learning; automated warehouses; e-commerce; logistics 4.0; supply chain

check for updates

Academic Editors: Mustapha Hlyal, Abdessamad Ait El-Cadi, Kaoutar Kouzmi, Mariam Atwani and Kenza Izikki

Published: 13 June 2025

Citation: Belhaj, N.; Patrix, J.; Oulmakki, O.; Verny, J. Leveraging Data Science for Impactful Logistics Automatons in the E-Commerce Sector. *Eng. Proc.* **2025**, *97*, 31. https://doi.org/10.3390/engproc2025097031

Copyright: © 2025 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/).

1. Introduction

Nowadays, effective logistics operations are regarded as a strategic asset and a key factor in growth and success. E-commerce companies invest in warehouse automation projects to retain an advantage over their competitors [1]. Cost control is the top concern of companies, as well as their productivity [2]. However, multiple factors can prevent them from investing, such as the initial costs related to material acquisition, technology installation, and maintenance. Such projects require qualified expertise and the company may hesitate to invest in external experts. The integration of new technologies can disrupt existing optimized processes, requiring intensive training of employees and the optimization of their new process, creating skepticism regarding the return on investment (ROI), as too long to materialize [3].

Despite this reluctance, multiple e-commerce companies have adopted automation technologies for their warehouse operations for the sake of potential benefits [4]. Several studies [5] have been conducted on digital transformations in smart warehouses to investigate these concerns, each focusing on a disruptive technology, ranging from Cloud computing, Internet of Things (IoT), Big Data analysis, Digital Twins, to Robotics [6]. To name a few, ref. [7] studied the effect of using Robotics in logistics on cost accounting and productivity. Similarly, ref. [8] explored the impact of information and communication

technologies in automated warehouses, to assess their effectiveness in terms of costs and increased volumes.

Nowadays, companies venture further in the pursuit of performance, growth, and competitiveness by taking advantage of their data. They seek to draw insightful business strategies by leveraging data science principles to make informed decisions and optimize various aspects of their supply chain [9]. Data are widely acknowledged as a catalyst for improved decision making and increased profitability, as supported by [10]. Their research revealed that companies identifying as "data-driven" tend to outperform competitors in both financial and operational aspects. Recent studies have shown that nearly all organizations have invested in data to enhance profitability and mitigate the risks of missed opportunities [11]. A data-driven culture within an organization promotes the integration of data from various sources, to gain a holistic view of the supply chain. This implies using key performance indicators (KPIs) and analytics dashboards to monitor the adopted measures [12].

Our research work focused on the usage of data science techniques in the context of automated warehouses. The purpose was to draw insightful business conclusions for future warehouse automation projects. To this end, we determined the most impactful implemented automatons in terms of productivity (i.e., sales volumes) and cost reduction, in order to be considered for future warehouse implementations. To do so, we studied three warehouses of a European e-commerce company, each with a different degree of automation. By analyzing data across these warehousing facilities, we conducted data science methodologies and used various machine learning [13] techniques to determine the top logistics automatons that most impact financial costs. Our approach is intended to support companies in determining the most effective automatons to integrate in future automation projects.

The remainder of this article is structured as follows: In Section 2, we provide an overview of the used data science techniques. In Section 3, we introduce the European e-commerce company and the characteristics of its warehouses facilities. Afterwards, we present our data science methodology to determine the most impactful logistics automatons in Section 4. We also explore research work in the field of data science and digitization, focusing on their application in supply chain management in Section 5. Finally, we conclude with a summary of our main contributions and perspectives for future work in Section 6.

2. Background on the Used Data Science Techniques

Data science [14] is a multidisciplinary field that uses scientific methods, algorithms, processes, and systems to retrieve valuable insights and knowledge from data. It encompasses a combination of statistical analysis, Machine Learning (ML) [13], computer science, and domain-specific expertise for the purpose of interpreting complex patterns, trends, and relationships within data. Data science involves formulating hypotheses through collecting, exploring, and pre-processing data; applying analytical techniques; and iteratively refining models to uncover meaningful information. The ultimate objective of data science is to generate actionable insights, predictions, and knowledge that can inform decision-making processes across various domains. We explore some of the techniques in the following:

Pearson correlation analysis [15] is a statistical method that quantifies the strength and direction of a linear relationship between two continuous variables x and y. It measures the degree of change in the variable x affects the change in the variable y. The strength determines the closeness of the data points to a straight line, while the direction defines whether the variable x increases when the other variable y decreases, and inversely. The Pearson correlation coefficient measures this correlation between a dataset feature (i.e., variable) and the target variable that we seek to model and explain (e.g., average unit cost per item

sent in a warehouse). The Pearson coefficient is the degree ρ of linear relationship between two variables x and y, and is measured as follows:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sigma_x\sigma_y} \quad (1)$$

where $\overline{x}$ and $\overline{y}$ represent the sample means of the variables x and y, respectively. σ_x and σ_y are the standard deviations of the variables x and y, respectively, while n is the number of observations. A value that approaches 1 indicates a strong positive correlation, while a value that approaches -1 refers to a strong negative correlation. When the value is close to 0, this reveals a weak or a nonlinear correlation.

Regression Analysis [16] is a set of statistical methods intended to analyze the relationships between a dependent and continuous feature and one or several dependent features to provide actionable insights. These methods use knowledge about this relationship to draw generalization equations and then forecast for unobserved values of these independent features. If we have n independent features and a single dependent feature to model or predict, the regression equation, namely multiple or bivariate, would be $Y = \beta_n.X_n + \beta_{n-1}.X_{n-1} + \ldots\ \beta_2.X_2 + \beta_1.X_1 + \beta_0 + \epsilon$. Where Y is the dependent feature that we are trying to predict. $\beta_n, \beta_{n-1}\ \ldots\ , \beta_2, \beta_1$ are the coefficients of the independent features $X_n, X_{n-1}, \ \ldots\ , X_2, X_1$, respectively. These coefficients describe the amount of change that is brought to the dependent feature by the change that occurs in the independent features. β_0 is the Y-intercept, which is the value of Y when all the independent features equal zero. ϵ represents the error which captures the difference between the predicted and the observed values of the dependent feature. The solution for this equation would be to find the best curve that minimizes the squared error ϵ.

There are ML algorithms that solve this equation through creating regression models. ML algorithms automate the process of finding the optimal coefficients in a regression equation learned from a dataset during the training process. Some common machine learning algorithms for regression tasks include decision trees [17] that model data in a tree-like structure with nodes, branches, and leaves. Each node represents a decision based on a certain feature's value, while a branch denotes the outcome of that decision and a leaf represents the final prediction. Other ML algorithms such as ensemble techniques (e.g., Random Forest) [17] combine multiple models (i.e., learners) to create a better performing and robust model through aggregating their predictions. They build multiple decision trees during training and merge their predictions to provide more accurate and stable results. Boosting techniques [17] such as AdaBoost, XGBoost, and CatBoost are also used for regression problems. They function by combining the predictions of multiple weak learners to create a stronger and more accurate model. These techniques act by sequentially training these weak learners, with each subsequent learner targeting the mistakes of the previous learners. Deep Learning [17] can also provide a solution to such problems with Artificial Neural Networks (ANNs). These are composed of neurons that are organized in input, hidden, and output layers. The neurons are densely connected with weights that represent the strength of their connections. ANNs find applications in various domains, ranging from image and speech recognition, financial forecasting, natural language processing, medical diagnosis, and more. They have demonstrated a remarkable performance in tasks entailing complex patterns within data.

Mean Decrease in Accuracy (MDA) [18] is used to assess individual features in a predictive model. It operates by breaking the associations of each feature and the target feature. This is accomplished by randomly permuting the features and evaluating the subsequent increase in error in the machine learning model's performance (e.g., model's

accuracy). A drop in the model's performance is indicative of the importance of the removed feature for the prediction accuracy of the target feature. Since we are addressing a regression problem, MDA is translated into an $MDSE_i$ (Mean Decrease in Squared Error) for a feature i and it is given by $MDSE_i = MSE_{initial} - MSE_i$. Where $MSE_{initial}$ represents the initial MSE computed upon training the Random Forest model and describing its accuracy. MSE_i is the measured MSE upon removing the feature i. Note that the larger the $MDSE_i$, the more important the feature i is in decreasing the $MSE_{initial}$.

SHapley Additive exPlanation [19] defines Shapley values and is derived from cooperative game theory [20] to estimate the importance of each feature in a model's predictions. The features are viewed as players that cooperate in a game, while the prediction task for an instance of the dataset is viewed as the game. The players cooperate in a coalition to gain payouts depending on their contribution to the total payout. More specifically, a Shapley value [21] represents the mean marginal contribution of a given feature value across all possible cooperations. A Shapley value $\phi_i(M)$ for a particular feature i is computed according to the following equation:

$$\phi_i(M) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [M(S \cup \{i\}) - M(S)] \quad (2)$$

S denotes the coalition of features, not including feature i. N is the set of all features. $M(S \cup \{i\})$ represents the model's prediction when the feature i is included in the coalition. $M(S)$ is the model's prediction when the feature i is excluded from the coalition. $|S|$ is the number of features in a coalition S, and $|N|$ is the total number of features in the dataset. The Shapley value basically computes how much the change in each feature value contributes to the change in a prediction compared to the average prediction. The payout represents the actual prediction for an instance of the dataset minus the average prediction computed for all the instances.

3. Case Study: European E-Commerce Warehouses

The e-commerce company that we studied specializes in exclusive discount sales across several European countries. Their clients have access to a private and flash sales platform. Sales events are provided in a wide range of categories, including fashion, technology, food, travel, jewelry, etc. The company is supported by third party logistics (3PL) and owns several warehouses across 10 European countries to process shipments and returns. More than 42 million parcels have been prepared and shipped either directly at the drop-shippers or in their warehouses. In the drop-shipper scenario, the latter proposes the sale on their website and then prepares and directly sends the orders. In their warehouse scenario, the orders are received before the sale begins and then prepared in the same warehouse. In the latter scenario, two types of sales are proposed within their warehouses, namely pre-reception and post-sales. In pre-reception sales, the products are already in their warehouse and the estimated delivery time is around 48–72 h. In post-sales, the products are still at the brand's (or supplier's) warehouse and are shipped upon the ending of the sales on the company's website. The estimated delivery time is approximately 3 weeks.

During recent years, this company has been developing and restructuring their manner of warehousing to keep pace with increasing demands [22]. To do so, its logistics have gone through significant changes by adopting several degrees of automation in its warehousing operations. In this paper, we studied and analyzed data coming from three of their automated distribution centers. Figure 1 depicts the 11 automatons deployed by the warehouses WH1, WH2 and WH3, and their names are shown in Table 1. The latter emphasizes the difference in automation of the three warehouses. Table 1 presents a summary of the implemented automatons, including automated and semi-automated automatons for

each warehouse. A green checkmark indicates that a given automaton is deployed within a warehouse, whereas a red cross signifies that it is not deployed. We notice that WH2 is equipped with more recent technologies compared to the others. WH2 is much more automated than WH1 and WH3, in this order. The table displays also their characteristics in terms of floor surface and their storage capacity measured by the number of location units. It also shows their production capacity in terms of sent items per week and their picking capacity, which is the number of scanned references.

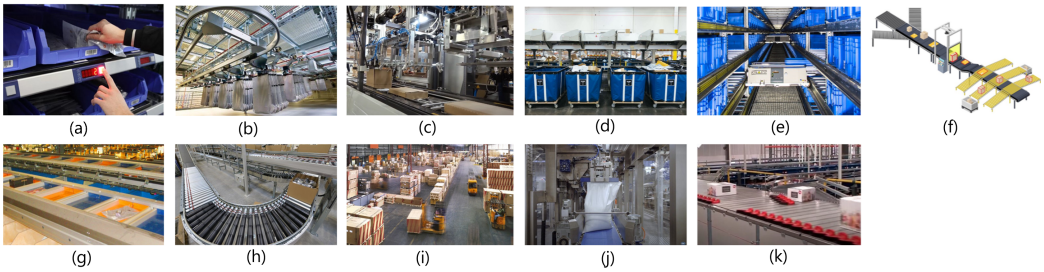


Figure 1. The 11 logistics automatons from our data study of the European warehouses.

Table 1. Comparison—Warehousing automatons and characteristics.

Automatons	**WH 1**	**WH 2**	**WH 3**
Put-to-the-light system for mono SKU sorting (Figure 1a)	✓	✗	✗
KNAPP pocket sorter system (Figure 1b)	✓	✗	✗
Automated packing with JIVARO (Figure 1c)	✓	✓	✓
Outbound transitic & sorting solution for bags (Figure 1d)	✓	✗	✗
OSR (Optical storage & Retrieval) Boxes (Figure 1e)	✓	✗	✗
Outbound transitic & sorting solution for parcels (Figure 1f)	✓	✗	✗
Bombay & Tilt-Tray item sorter (Figure 1g)	✗	✓	✗
SAVOYE Loop system for semi-automatic picking (Figure 1h)	✗	✓	✓
Transitic for parcel handling across the warehouse (Figure 1i)	✗	✓	✗
Automated bag packing & sorting machinery (Figure 1j)	✗	✓	✗
TGW parcel sorter in the shipping area (Figure 1k)	✗	✓	✓
Characteristics			
Floor surface in m^2	30,000	20,000	30,000
Storage capacity (location units)	4 k	No limit	4.8 k
Production capacity (items per week)	285–400 k	>400 k	200–350 k
Picking capacity (reference numbers)	No limit	No limit	No limit

4. Impact of Logistics Automatons: Correlation and Regression Analyses

We intend to evaluate the effects of the logistics automatons on the average unit cost (AUC) per sent item in a warehouse. To do so, we need to determine the most impactful automatons among the ones used for the cost reduction. Accordingly, we provide a general description of the dataset, and then we carry out a correlation and regression analysis. Additionally, we assess the importance of features within our dataset, as follows.

4.1. Dataset Description and Implementation Details

The dataset is composed of 2585 data points collected from the three distribution centers WH1, WH2, and WH3 during the year 2021. For each warehouse, we have the target (i.e., label) feature: average unit cost, which describes the average cost incurred per given good for their shipment to the end customer including consumable and non-consumable expenses (i.e., warehousing, shipping) [float] (i.e., dependent variable) and the following features (i.e., independent variables):

- Sector: name of the sector that gathers a specific type of handled product [string].
- Month: ID of the month during which a number of items are sent per sector [integer].
- Sent items: number of sent items in a sector within a given month [integer].
- Consumable costs: expenses related to the consumption of supplies and materials such as packaging, labels, pallets, etc. [float].
- Non-consumable costs: expenses that encompass rent, energy, automatons maintenance, human labor, etc. [float].
- "Put-to-the-light system for mono SKU sorting"; "KNAPP"; "JIVARO"; "Outbound transitic & sorting solution for parcels"; "Outbound transitic & sorting solution for bags"; "Bombay & Tilt-Tray"; "SAVOYE Loop"; "Transitic for parcel handling"; "Automated bag packing & sorting"; and "TGW Parcel Sorter": ID of each automaton (1 = used in a given warehouse, 0 = not used) [integer].

To manipulate the dataset, we made use of python libraries such as *numpy 1.21.5* [23] and *pandas 1.3.5* [24] that provided tools to manipulate, structure, clean, transform, and analyze the dataset. We used *matplotlib 3.5.3* [25] and *seaborn 0.12.2* [26] for statistical data visualization. We also used the *scikit-learn 1.0.2* library [27], which includes a variety of machine learning algorithms such as linear and logistic regression, decision trees, random forests, etc. We also used this to perform data pre-processing, exploratory data analysis, and train-test splitting, for building and evaluating the machine learning models. We created *xgboost 1.6.2 and catboost 1.2.2* regressors using their own libraries [28,29]. Similarly, we also investigated the *rfpimp 1.3.7* library [30] to analyze, determine, and visualize the importance of features in a random forest model. This helped to evaluate the contribution of each feature to the model. We also explored the *shap 0.41.0* library [31] to interpret the outcome of our models and help in gaining insights into black-box model predictions. We also used the *keras 2.10.0* API [32] integrated into *tensorflow 2.10.0* framework [33] to create and evaluate neural network models. The experiments were executed on a 64-bit windows PC, equipped with 16 GB of RAM and Intel core i7 with 2.30 GHz frequency.

4.2. Data Science Approach and Experimental Results

To determine the features most impacting the label, we followed a process involving several key steps. Figure 2 provides an overview of these steps, as follows:

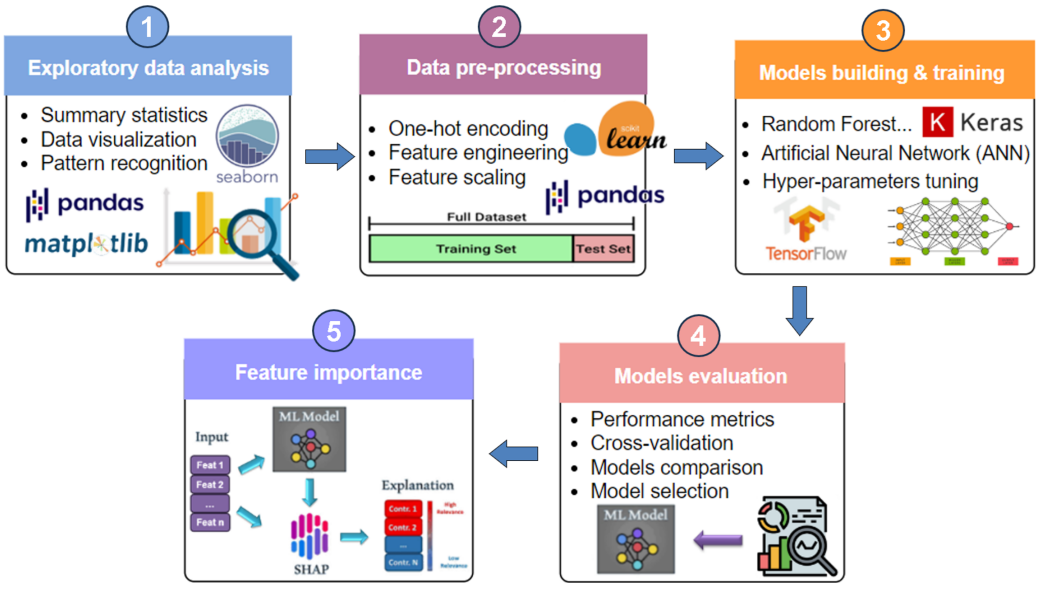
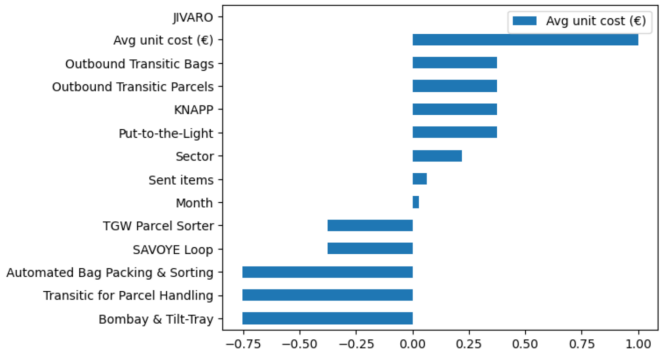


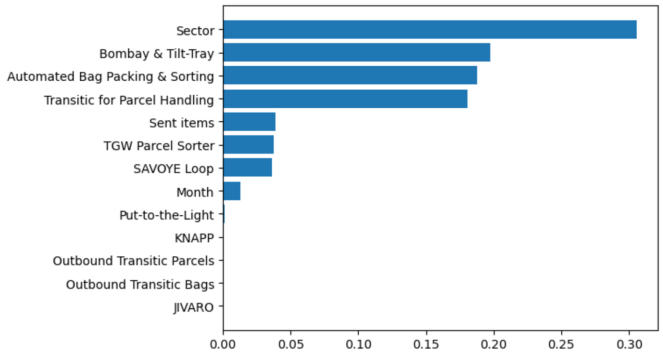
Figure 2. Overview of our data science methodology.

1. Exploratory data analysis: we started by cleaning the dataset, structuring its features, and making sure there were no missing values, in order to make statistical summaries of the data distribution. We also performed data visualization to determine tendencies or patterns that might indicate interesting relationships or insights. We performed a Pearson correlation [15] analysis to examine relationships between variables and identify which were strongly or weakly correlated. Accordingly, we looked for the automatons that had a strong negative correlation with the average unit cost, since we sought to reduce it. Figure 3a shows that these three automatons were (n°1) *Bombay & Tilt-Tray*, (n°2) *Transitic for Parcel Handling*, and (n°3) *Automated Bag Packing & Sorting*, represented with a correlation coefficient of -0.75. Hence, they were the most impactful automatons on the cost reduction. Needless to say, the *JIVARO* machine contributes significantly to time and cost reductions. Since it is used within all the warehouses, its impact was not considered in our analysis.
2. Data pre-processing: We made sure the data were completely clean for the next steps and contained solely the features that were relevant to our analysis (e.g., no irrelevant temporal data in our case, etc.). The features and the label were separated for the prediction, and then the data were split into two sets: training (80%) and test (20%) sets. We converted categorical variables into dummies with one-hot encoding for ML models, manipulating solely numerical variables, and then we normalized or standardized the data if needed.
3. Model building and training: The label to predict was a continuous numerical feature, which led us to create regression models that were able to map the features to the continuous target label. Thus, we created, trained, and launched predictions with the following regression models: LinearRegression, DecisionTreeRegressor, RandomForestRegressor, XGBRegressor, AdaBoostRegressor, CatBoostRegressor, and Artificial Neural Network (ANN) [17]. Afterwards, we set up the algorithm hyperparameters and launched the learning process on the training data.
4. Model evaluation: We launched the predictions on the test data and then we compared the predictions with the actual values. We selected Root Mean Squared Error (RMSE) as the performance metric to evaluate the models. We noticed that the RandomForestRegressor had the lowest RMSE and performed the best among the models, hence we selected it for our subsequent steps. As seen in Figure 3c, we illustrated the model's performance by comparing the predicted values with the true one. The blue diagonal line represents perfect predictions, while the red dots represent the actual predictions. Dots on the diagonal indicate correct predictions and scattered dots show prediction errors. We used this model to measure the Mean Decrease in Accuracy in Figure 3b, which depicts the decrease in the model's accuracy score whenever the related feature was removed. This demonstrates the feature's importance regarding the decrease in the average unit cost. It is logical to expect that this cost can be greatly impacted by the type of products being handled in each sector. In each sector, we have varying products requirements, handling complexities, and storage needs, which can affect the overall operations cost. Larger or heavier products may require specialized automatons or additional handling efforts, which can increase the cost per unit compared to smaller and lighter items within each sector. Figure 3b shows that this cost was also influenced by three logistics automatons: (n°1) *Bombay & Tilt-Tray*, (n°2) *Automated Bag Packing & Sorting*, and (n°3) *Transitic for Parcel Handling*.
5. Game theory value-based feature importance: We used the Shap [34] interpretation to grasp the importance of each feature. We computed the Shapley values to assess the predictions of the random forest model, as shown in Figure 3d. As expected, the sector feature took first place, followed by the three automatons: (n°1) *Bombay & Tilt-Tray*,

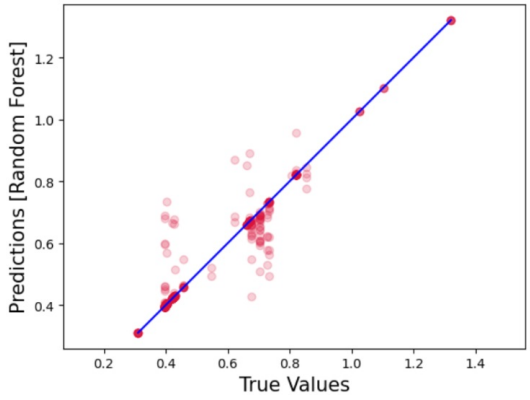
(n°2) *Automated Bag Packing & Sorting*, and (n°3) *Transitic for Parcel Handling*. These are the exact same logistics automatons that stood out before, which consolidates the previous results.



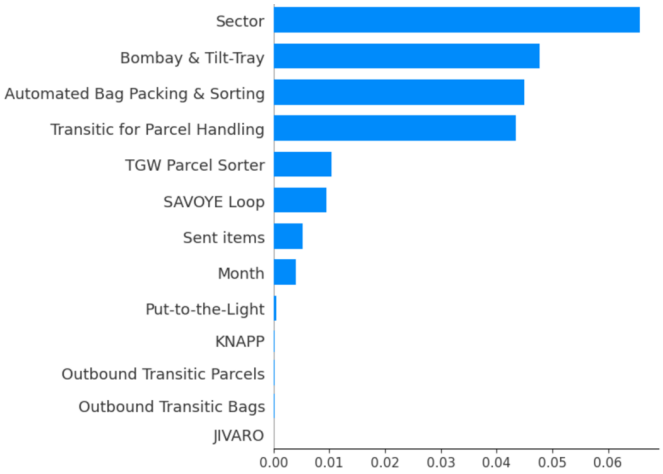
(**a**) Pearson correlation analysis



(**b**) MDA Feature importance—Random Forest



(**c**) Model performance: true (blue) vs. predicted (red) values



(**d**) Mean Shapley values—top three automatons

Figure 3. Experimentation results.

5. Related Work: Data Science for Strategic Decision Making in Supply Chain Contexts

The fusion of data science in many industrial sectors has become a crucial approach to enhance their efficiency. Many studies have shown its potential in helping to deduce insights from collected data, either to solve business problems or derive strategic decisions. Based on grounded analytical facts, it can play a pivotal role in shaping the future of digital transformations [35]. Hence, the companies have taken advantage of this by increasing the adoption of AI and ML techniques to enhance their operational practices and decision-making capabilities [36]. One can analyze massive datasets, discover patterns, and generate insights such as creating predictive maintenance models. Indeed, AI algorithms can reduce downtime and operational costs by forecasting failures and creating proactive maintenance strategies, as described in [37]. Studies such as [38] investigated the usage of data science for real-time inventory management for e-commerce companies to operate efficiently and found enhanced performance. Others focused on demonstrating an immediate relationship between the adopted AI techniques and the noted performance for the sake of competitiveness and profitability [39].

From this same perspective, other efforts have been directed to manage raw material inventory, to operate in a durable manner [9]. They faced challenges with late arrivals of raw materials affecting their “just in time” operations, causing late shipments and financial

losses. They addressed this issue by installing trackers on over 2000 trucks, which provided real-time location data. They also used AI to develop models that learned traffic patterns and predicted potential delays to understand their impact on the production schedules. This solution allowed better decision-making by creating revised production schedules, preventing line shutdowns, while fulfilling all orders. The AI-based model resulted in annual savings exceeding USD 40 million by eliminating production interruptions.

Another study [40] was conducted to analyze the impact of big data on supply chain management, to optimally manage the massive amount of data generated and shared among business and industrial entities. They investigated on-demand data analysis for a better control of inventory levels, to make reliable predictions about future demands [41]. Big data can also identify cost-effective routes for shipments by analyzing weather conditions and traffic patterns. It enables real-time tracking and monitoring of shipments to provide real-time adjustments whenever needed. It also helps in analyzing customer behavior, to better grasp their preferences and trends and then tailor the products and services to their needs.

Another industrial goods manufacturer aimed to minimize landfill waste across its operations [9], collaborating on a project spanning 100 manufacturing facilities and over 250 warehouses and distribution centers. The initiative successfully diverted over 96 million pounds of waste annually. Using data science, they were able to identify top-performing and under-performing sites. They exploited this analysis to standardize the performance of current sites. They were able to derive best practices and distribute them across all sites. Therefore, they only dedicated time to the sites that required substantial improvement. This effort not only contributed to environmental sustainability, but also resulted in significant annual savings of over USD 36 million through recycling and energy recovery initiatives.

6. Conclusions and Future Work

This paper analyzed data collected from three European warehouses to determine the most influential logistics automatons for cost reduction among those used. Our findings are intended to encourage companies to take advantage of their data, in order to invest in automation according to their specific needs. They are also intended to determine the most impactful automatons on cost reduction, to consider for future automation projects. The main objective was to showcase the integration of data science into warehouse logistics, to support supply chain managers in their strategic decisions.

It is worth noting that we conducted a technical benchmark of the installed automatons, but we decided to exclude it. The technical specifications on the vendors' websites were merely marketing arguments, which did not seem impartial. This prevented us from providing a truthful benchmark, which seemed one-sided and biased.

In our future work, we intend to conduct interviews on the automation before and after installation. By focusing on workforce adaptation, retraining needs, and potential resistance, we can provide a more holistic view of logistics automation by considering both technological benefits and social implications; ultimately offering a comprehensive analysis that addresses the well-being and adaptability of the workforce amidst increasing automation. This approach can help to derive truthful feedback on their advantages and drawbacks directly from their users, which seems a better alternative to a technical benchmark.

Furthermore, with the help of our industrial partner, we envisage collecting fine-grained data for a more profound analysis of their warehouse activities and finances. Consequently, the related variables need to be identified and the monitoring processes need to be implemented for a meaningful data collection. From the same perspective, we are seeking to expand our dataset to include companies from diverse industries and geographic regions, to enhance the validity and generalizability of our findings. We are

actively engaging with other industries to obtain additional datasets, aiming to provide a broader perspective on how data-driven approaches can optimize logistics automation across different sectors.

Author Contributions: Conceptualization, N.B.; methodology, N.B., J.P., O.O. and J.V.; software, N.B.; validation, J.P., O.O. and J.V.; formal analysis, N.B. and J.P.; investigation, N.B.; resources and data collection, O.O. and J.V.; data curation, N.B.; writing—original draft preparation, N.B.; writing—review and editing, N.B.; visualization, N.B. and J.P.; supervision, J.P., O.O. and J.V. 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: Data is unavailable due to privacy restrictions.

Conflicts of Interest: The authors declare that HighFi Lab and Opal research have no commercial conflict of interest.

References

1. Kawa, A.; Swiatowiec-Szczepanska, J. Logistics as a value in e-commerce and its influence on satisfaction in industries: A multilevel analysis. *J. Bus. Ind. Mark.* **2021**, *36*, 220–235. [CrossRef]
2. Cooke, J.A. Inventory velocity accelerates. *Logist. Manag.* **2002**, *42*, 8–33.
3. iCepts. *Optimizing Warehouse ROI: Automating Warehouse Processes for a Greater Return on Investment*; Technical Report; iCepts Technology Group, Inc.: Middletown, PA, USA, 2011.
4. Kumar, S.; Narkhede, B.E.; Jain, K. Revisiting the warehouse research through an evolutionary lens: A review from 1990 to 2019. *Int. J. Prod. Res.* **2021**, *59*, 3470–3492. [CrossRef]
5. Kembro, J.; Norrman, A. The transformation from manual to smart warehousing: An exploratory study with Swedish retailers. *Int. J. Logist. Manag.* **2022**, *33*, 107–135. [CrossRef]
6. van Geest, M.; Tekinerdogan, B.; Catal, C. Design of a reference architecture for developing smart warehouses in industry 4.0. *Comput. Ind.* **2021**, *124*, 103343. [CrossRef]
7. Brzezinski, L. Robotic Process Automation in Logistics—A Case Study of a Production Company. *Eur. Res. Stud. J.* **2022**, *XXV*, 307–315. [CrossRef]
8. Andiyappillai, N. An Analysis of the Impact of Automation on Supply Chain Performance in Logistics Companies. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1055*, 012055. [CrossRef]
9. Awaysheh, A. Leveraging Data Science to Enhance Your Supply Chain and Improve your Company's Performance. 2020. Available online: https://www.ascm.org/globalassets/ascm_website_assets/docs/leveraging-data-science-to-enhance-your-supply-chain-and-improve-your-companys-performance.pdf (accessed on 12 June 2025).
10. McAfee, A.; Brynjolfsson, E. Big Data: The Management Revolution. *Harvard Bus. Rev.* **2012**, *90*, 61–67.
11. Bentalha, B. Big-Data and Service Supply chain management: Challenges and opportunities. *Int. J. Bus. Technol. Stud. Res.* **2020**, *1*, 1–9. [CrossRef]
12. Chiaraviglio, A.; Grimaldi, S.; Zenezini, G.; Rafele, C. Overall Warehouse Effectiveness (OWE): A New Integrated Performance Indicator for Warehouse Operations. *Logistics* **2025**, *9*, 7. [CrossRef]
13. Sutton, R.S. *Learning to Predict by the Methods of Temporal Differences*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988; pp. 9–44.
14. Biernat, E.; Lutz, M. *Data Science: Fondamentaux et études de cas*; EYROLLES: Paris, France, 2015.
15. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 37–40.
16. Sarstedt, M.; Mooi, E. Regression Analysis. In *A Concise Guide to Market Research, Springer Texts in Business and Economics*; Springer: Berlin/Heidelberg, Germany, 2014.
17. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow. Concepts, Tools, and Techniques to Build Intelligent Systems*; O'REILLY: Springfield, MO, USA, 2023.
18. Bénard, C.; da Veiga, S.; Scornet, E. MDA for random forests: Inconsistency, and a practical solution via the Sobol-MDA. *arXiv* **2022**, arXiv:2102.13347.

19. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games II*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–317.
20. Peleg, B.; Sudhölter, P. *Introduction to the Theory of Cooperative Games*, 2nd ed.; Theory and Decision Library C; Springer: Berlin/Heidelberg, Germany, 2007; Volume 34.
21. Shapley Values—Github. 2025. Available online: https://christophm.github.io/interpretable-ml-book/shapley.html (accessed on 12 June 2025).
22. Baker, P.; Halim, Z. An exploration of warehouse automation implementations: Cost, service and flexibility issues. *Supply Chain Manag. Int. J.* **2007**, *12*, 129–138. [CrossRef]
23. NumPy. 2025. Available online: https://numpy.org/ (accessed on 12 June 2025).
24. Pandas. 2025. Available online: https://pandas.pydata.org/ (accessed on 12 June 2025).
25. Matplotlib: Visualization with Python. 2025. Available online: https://matplotlib.org/ (accessed on 12 June 2025).
26. Seaborn: Statistical Data Visualization. 2025. Available online: https://seaborn.pydata.org/ (accessed on 12 June 2025).
27. Scikit-Learn. 2025. Available online: https://scikit-learn.org/stable/ (accessed on 12 June 2025).
28. XGBoost Documentation. 2025. Available online: https://xgboost.readthedocs.io/en/stable/ (accessed on 12 June 2025).
29. CatBoost. 2025. Available online: https://catboost.ai/ (accessed on 12 June 2025).
30. Rfpimp. 2025. Available online: https://pypi.org/project/rfpimp/ (accessed on 12 June 2025).
31. SHAP Documentation. 2025. Available online: https://shap.readthedocs.io/en/latest/ (accessed on 12 June 2025).
32. Keras. 2025. Available online: https://www.tensorflow.org/guide/keras?hl=fr (accessed on 12 June 2025).
33. Tensorflow. 2025. Available online: https://www.tensorflow.org/ (accessed on 17 January 2024).
34. Shap—Github. 2025. Available online: https://github.com/slundberg/shap (accessed on 12 June 2025)
35. Rai, R.; Tiwari, M.K.; Ivanov, D.; Dolgui, A. Machine learning in manufacturing and industry 4.0 applications. *Int. J. Prod. Res.* **2021**, *59*, 4773–4778. [CrossRef]
36. Maheshwari, S.; Gautam, P.; Jaggi, C.K. Role of Big Data Analytics in supply chain management: Current trends and future perspectives. *Int. J. Prod. Res.* **2021**, *59*, 1875–1900. [CrossRef]
37. Cannas, V.G.; Ciano, M.P.; Saltalamacchia, M.; Secchi, R. Artificial intelligence in supply chain and operations management: A multiple case study research. *Int. J. Prod. Res.* **2024**, *62*, 3333–3360. [CrossRef]
38. Alsheyadi, A.; Baawain, A.; Shaukat, M.R. E-supply chain coordination and performance impacts: An empirical investigation. *Prod. Manuf. Res.* **2024**, *12*, 2379942. [CrossRef]
39. Nasiri, M.; Ukko, J.; Saunila, M.; Rantala, T. Managing the digital supply chain: The role of smart technologies. *Technovation* **2020**, *96–97*, 102121. [CrossRef]
40. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286. [CrossRef]
41. Lekic, M.; Rogic, K.; Boldizsar, A.; Zoldy, M.; Torok, A. Big Data in Logistics. *Period. Polytech. Transp. Eng.* **2021**, *49*, 60–65. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.