Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 232 (2024) 2478–2487

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Industry 4.0 and Smart Manufacturing

Performance Evaluation and Explainability of Last-Mile Delivery

Ângela F. Brochado[a,*], Eugénio M. Rocha[b,c], Emmanuel Addo[b], Samuel Silva[d,e]

[a] *Department of Economics, Management, Industrial Engineering and Tourism (DEGEIT), University of Aveiro, Portugal*
[b] *Department of Mathematics (DMat), University of Aveiro, Portugal*
[c] *Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal*
[d] *Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Portugal*
[e] *IEETA—Instituto de Engenharia Electrónica e Informática de Aveiro, Aveiro, Portugal*

Abstract

The demand for last-mile delivery (LMD) services worldwide increased following online sales growth, so better methods to assess efficiency issues are paramount. This work explores a data-driven approach to evaluate LMD services and inform logistics service providers about possible improvement directions. It uses multi-directional efficiency analysis to benchmark LMD services based on process variables, such as delivery time and service cost. Then, by fitting machine learning models and using explainability algorithms with new metrics, characterizes factors that influence LMD performance. Early discussions with experts show that the approach produces understandable and integrable results that generate valuable insights, e.g., regarding the impact of each variable on service quality informing the direction for further improvement action.

Keywords: Last-mile logistics; performance evaluation and improvement recommendations; multi-directional efficiency analysis (MEA); eXplainable Artificial Intelligence (XAI)

1. Introduction

In recent years, there has been a notable exposure to the vulnerability of supply chains and transportation networks, coinciding with a significant surge in demand for last-mile logistics services. The COVID-19 pandemic has posed a substantial threat to various aspects of modern life, prompting predominantly reactive rather than proactive operational responses. Concurrently, logistics networks have faced immense strain due to the escalating popularity of online shopping, highlighting several procedural inefficiencies along the way [12]. Consequently, customer expectations for a faster and cost-free delivery system with more frequent deliveries became more prominent [8]. According to a report made by Precedence Research [26], the global last-mile delivery transportation market size, valued in 2022 at

[*] Corresponding author.
E-mail address: filipabrochado@ua.pt

10.1016/j.procs.2024.02.067

USD 142.9 billion, is expected to hit around USD 297 billion by 2030. The increasing middle-class population in both developed and developing regions has also prompted this expansion. Therefore, the importance of forging partnerships between public or private entities for environmental sustainability, and technology innovation and integration has never been more imperative.

On the technological side, recent achievements in machine learning (ML) have triggered a wave of new artificial intelligence (AI) applications where significant advantages have been recognized across a wide range of fields. ML is here defined as a subset of AI that enables computers to autonomously analyze and learn from data, thereby enhancing their cognitive capabilities [2]. AI is perceived as a ”system that displays intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals” [1, p. 1]. One of the inherent limitations of numerous ML algorithms is their opaque nature, often referred to as a ”black box.” This implies that comprehending and fully understanding these algorithms, even for domain experts, is highly challenging [22], with implications on how these can be deployed outside the lab and strongly influencing how end-users may trust their outcomes. Given such complexities, by the end of 2021, the topic of explainable artificial intelligence (XAI) became a hot research theme [15]. The term encompasses a collection of processes and methodologies (e.g., SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic (LIME) are the most widely used [22]) that enable researchers, developers, and users to comprehend and trust the outcomes generated by ML, deep learning (DP) or neural network (NN) algorithms, also called *AI-powered decision-making* [16]. According to [12], the integration of AI into decision-making processes may provide enhanced real-time information about any business process with data. Jointly, XAI may work as a guide for decision-makers by giving recommendations for improvement (determining the most relevant features that explain a certain output – some examples in manufacturing are given in [7, 27]).

The new challenges and goals to embrace Logistics 4.0 [3] require harnessing the latest technologies that can take the most out of the existing data to support decision-making. However, the use of ML/AI applications for last-mile delivery (LMD) performance evaluation is still in its infancy, in fact, most applications are focused on anomaly detection, forecasting, and planning (see [14]). Solely the work of [29] has presented a blockchain-based approach to evaluate customer satisfaction in urban logistics. The authors used the Long Short-Term Memory algorithm with the following criteria: cost performance, information transparency, cargo damages, and on-time delivery rate. Additionally, to the authors' knowledge, no studies have been found in the literature using XAI techniques to help explain LMD performance modelled as an ML algorithm.

Before moving further in this paper, and for the sake of clarification, the literature mentions that the definition of LMD may suffer slight variations depending on the type of business undergone by the logistics company: business-to-business (B2B), business-to-consumer (B2C) or consumer-to-consumer (C2C) [25]. In this paper, the most general one suggested by Motavallian (2019) has been followed, and LMD is defined as: ”The last transportation of a consignment in a supply chain from the last dispatch point to the delivery point where the consignee receives it.” [23, p. 106].

The original main contribution of this paper is the development and application of a novel data-driven approach to evaluate and explain the performance of several LMD services, guiding logistics service providers towards higher efficiency service rates. The novelty of the approach may serve as a catalyst for future research on the topic and applications in other logistics scenarios.

In the next Section 2, the methodology is described in detail, and a real case study is explored in Section 3.

2. Methodology

To provide clarification, Fig. 1 summarizes the methodology employed in this study. The main scheme of this work entails (1) a benchmark of the delivery services (using MEA); (2) an identification of the best ML model (i.e. XGBoost and Random Forest) for fitting the benchmark rank in each destination zone, where the data was enriched with features not used in the benchmark (i.e., the so-called root cause variables); and (3) the use of explainability techniques (i.e. SHAP and LIME) to identify the impact of the root cause variables. For the sake of brevity, detailed explanations of XGBoost and LIME are omitted in favour of a more comprehensive look into the mathematical frameworks underlying MEA and SHAP, the latter specifically to address modifications introduced to the standard algorithm.

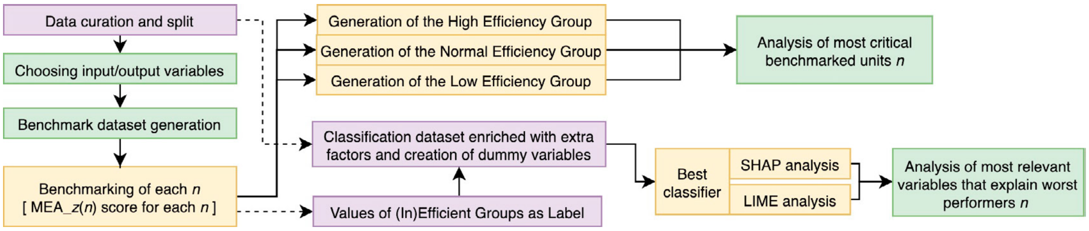
Data curation and split
Choosing input/output variables
Benchmark dataset generation
Benchmarking of each n
[MEA_z(n) score for each n]
Generation of the High Efficiency Group
Generation of the Normal Efficiency Group
Generation of the Low Efficiency Group
Analysis of most critical benchmarked units n
Classification dataset enriched with extra factors and creation of dummy variables
Values of (In)Efficient Groups as Label
Best classifier
SHAP analysis
LIME analysis
Analysis of most relevant variables that explain worst performers n

Fig. 1: Overall study methodology flow. Color legend: yellow boxes represent the steps where mathematical techniques or algorithms are employed to the dataset; purple boxes represent intermediate steps of data categorization or enrichment; green boxes represent decision-making steps.

2.1. Benchmarking using Multi-directional Efficiency Analysis (MEA)

Based on the Data Envelopment Analysis (DEA) methodology, Multi-directional Efficiency Analysis (MEA), proposed in [5], is a non-parametric approach from operational research used to evaluate the relative performance of a process or organization (in this paper we define it as a *benchmark unit*) based on a set of input and output variables. The advantage of MEA over DEA is that (1) it offers additional insights into the potential improvements precisely for each factor encompassed within the model, and (2) it computes an inefficiency metric to indicate which factors should be addressed to minimize performance inefficiencies. In the case of this paper, relative performance is assessed through the input-oriented version, i.e., benchmark units with higher ranks are those capable of minimizing inputs while outputs are suitably normalized and comparable. Henceforth, a detailed exposition of the mathematical foundation of MEA's algorithm and the accompanying notation is presented.

Let $n \in \mathbb{N}$ represent the index of a benchmark unit (i.e. a delivery service). For notation simplicity, $[m]$ defines the set $\{1, ..., m\}$ for some $m \in \mathbb{N}$. Therefore, any given n produces O outputs $y_o(n)$, where $o \in [O]$, using I inputs $x_i(n)$, where $i \in [I]$. Within this mathematical model, the initial $1 < D \leq I$ inputs are denoted as discretionary inputs, representing the variables that participate in the optimization process. Conversely, the non-discretionary inputs are variables that are inherently fixed and cannot be altered. Hence, the inputs vector is denoted by $x(n) \in \mathbb{R}^I$ and $y(n) \in \mathbb{R}^O$ defines the outputs vector for a given n.

MEA is now computed to benchmark the performance of all n. Considering the variable returns to scale (VRS) model for the efficiency measurement of benchmark units (see [6]), we define the set $\Lambda = \left\{\lambda \in \mathbb{R}^K \ : \ \sum_{n \in [K]} \lambda_n = 1\right\}$. For n running in $[K]$, d running in $[D]$, j running in $[J]$, and a fixed $\bar{n} \in [K]$, the MEA score of a certain observation $z(\bar{n}) = (x(\bar{n}), y(\bar{n}))$ is determined by solving the following linear optimization programs:

$$\begin{array}{c} \underline{\text{Problem } P^{\alpha}_{d}(z, \bar{n})} : \min \alpha_d(\bar{n}) \textit{ such that} \\ \sum_n \lambda_n x_d(n) \leq \alpha_d(\bar{n}), \\ \sum_n \lambda_n x_i(n) \leq x_i(\bar{n}),\ i \in [I], i \neq d, \\ \sum_n \lambda_n y_l(n) \leq y_l(\bar{n}),\ l \in [O], \end{array} \qquad \begin{array}{c} \underline{\text{Problem } P^{\beta}_{o}(z, \bar{n})} : \max \beta_o(\bar{n}) \textit{ such that} \\ \sum_n \lambda_n x_i(n) \leq x_i(\bar{n}),\ i \in [I], \\ \sum_n \lambda_n y_s(n) \leq \beta_o(\bar{n}), s \in [O], \\ \sum_n \lambda_n y_l(n) \leq y_l(\bar{n}), l \in [O], l \neq o, \end{array}$$

$$\begin{array}{c} \underline{\text{Problem } P^{\gamma}(\alpha, \beta, z, \bar{n})} : \max \gamma(\bar{n}) \textit{ such that} \\ \sum_n \lambda_n x_i(n) \leq x_i(\bar{n}) - \gamma(\bar{n})(x_i(\bar{n}) - \alpha^*_i(\bar{n})), i \in [D], \\ \sum_n \lambda_n x_i(n) \leq x_i(\bar{n}), i \in [I] \setminus \{d\}, \\ \sum_n \lambda_n y_l(n) \geq y_l(\bar{n}) + \gamma(\bar{n})(\beta^*_l(\bar{n}) - y_l(\bar{n})), l \in [O], \end{array}$$

where $\lambda \in \Lambda$, $\alpha^*_d(\bar{n})$ and $\beta^*_o(\bar{n})$ are the optimal problem solutions of $P^{\alpha}_{d}(z, \bar{n})$ and $P^{\beta}_{o}(z, \bar{n})$, respectively. The ideal point of $(x(\bar{n}), y(\bar{n}))$ is given by the MEA output vector

$$\zeta(n) \doteq (\alpha^*_1(n), ..., \alpha^*_D(n), x_{D+1}(n), ..., x_I(n), \beta^*_1(n), ..., \beta^*_O(n)) \in \mathbb{R}^{I+O}. \tag{1}$$

Within this mathematical setting, the methodology for a specific observation $z(\bar{n}) = (x(\bar{n}), y(\bar{n}))$ consists of solving $(|D| + |O| + 1) \times K$ linear programs. Thus, the MEA score of each $n \in \mathcal{N}$ for a given dataset $z = \{z(n)\}_{n \in \mathcal{N}}$ is given by

$$MEA_z(n) = \frac{\frac{1}{\gamma^*(n)} - \frac{1}{D} \sum_{i \in [D]} \frac{x_i(n) - \alpha^*_i(n)}{x_i(n)}}{\frac{1}{\gamma^*(n)} + \frac{1}{O} \sum_{o \in [O]} \frac{\beta^*_o(n) - y_o(n)}{y_o(n)}}, \tag{2}$$

where $\alpha_i^*(n)$, $\beta_o^*(n)$ and $\gamma^*(n)$ represent the optimal solutions to the linear optimization problems $P_i^{\alpha}(z,n)$, $P_o^{\beta}(z,n)$ and $P^{\gamma}(z,n,\alpha^*,\beta^*)$, respectively. With the directional contribution of each input and output, the MEA score is then calculated. For the input $i \in [I]$, the contribution in the unit $z(\bar{n})$ is given by $\mathrm{mEff}_i(n)$, see equation (3), where $\chi_{[D]}$ defines the characteristics function of set $[D]$. Thus, $\chi_{[D]}(i) = 1$, if $i \in [D]$ and $\chi_{[D]}(i) = 0$ if $i \notin [D]$. For the outputs $o \in [O]$ the contribution is given by

$$\mathrm{mEff}_i(n) = \frac{x_i(n) - \gamma(n)(x_i(n) - \alpha_i^*(n))}{x_i(n)}\chi_{[D]}(i), \quad \text{and} \quad \mathrm{mEff}_o(n) = \frac{y_o(n)}{y_o(n) + \gamma(n)(\beta_o^*(n) - y_o(n))}. \tag{3}$$

As aforementioned, a distinguishable feature of MEA over DEA is that inefficiencies can be analyzed individually. Thus, the inefficiency index is here referred to determine the number of times each input was used inefficiently. Using the ideas in [6], for a given dataset $z = \{z(n)\}_{n\in[K]}$ the inefficiency index for each input index $i \in [I]$ and $n \in [K]$ is given by

$$\mathrm{mIneff}_i(n) = \frac{\sum_{n=1}^{N} \gamma(n)(x_i(n) - \alpha_i^*(n))}{\sum_{n=1}^{N} x_i(n)}. \tag{4}$$

After computing MEA, decision-makers are able to make a primary analysis of the most critical benchmark units n and derive initial improvement directions. Furthermore, referring to the scheme depicted in Fig. 1, the scores obtained from MEA are partitioned into different classes to be used in next steps.

For $p \in [0,1]$, let $\mu_p \in [0,1]$ be the p percentile of the set of MEA scores (e.g., $p = 0.5$ is the median value), then we use the following categorical score metrics for each benchmark unit n:

- **2-Classes Index** as the map $2\text{-CI}_p : \mathbb{N} \to \{0,1\}$ with two values:
  - Inefficient Group defined as $2\text{-CI}_p(n) = 0$, if $MEA_z(n) < \mu_p$;
  - Efficient Group defined as $2\text{-CI}_p(n) = 1$, if $MEA_z(n) \geq \mu_p$.
- **3-Classes Index** as the map $3\text{-CI} : \mathbb{N} \to \{\text{'L'}, \text{'N'}, \text{'H'}\}$ with three values:
  - Low Efficiency Group defined as $3\text{-CI}(n) = \text{'L'}$, if $MEA_z(n) < Q_1^*$;
  - Normal Efficiency Group defined as $3\text{-CI}(n) = \text{'N'}$, if $Q_1^* \leq MEA_z(n) \leq Q_3^*$;
  - High Efficiency Group defined as $3\text{-CI}(n) = \text{'H'}$, if $MEA_z(n) > Q_3^*$.

Here, the thresholds Q_1^* and Q_3^* are defined by $Q_1^* = \min\{\mu_{0.25}, 0.25\}$ and $Q_3^* = \max\{\mu_{0.75}, 0.75\}$, respectively.

In the subsequent step of the methodology, the 2-Classes Index is seen as the label of the machine learning classifier. The 3-Classes Index is more granular thus, it is relevant to understand the distribution of scores in each zone, allowing to identify critical zones with respect to the delivery efficiency, see subsection 4.1.

2.2. Best classification model for the benchmark scores of a zone

In order to understand how problem variables contribute to (in)efficient delivery services, we need to find a classification model that captures the correspondence between variables and MEA scores. Hence, the principal aim of this step is to construct the best 'black-box' map between problem variables and the values of the 2-Classes Index. For such purpose, several machine learning algorithms, e.g. using the Python package *scikit-learn*, can be used if the attained machine learning (ML) evaluation metrics are high enough. Then, root cause analysis can be achieved by XAI techniques, see the next step in the methodology scheme in Fig. 1.

In our case study, involving several delivery zones and after extensible hyperparameter optimization, the best models for each zone were obtained using the algorithms: Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), see [10]. A RF, a tree-based ML algorithm as an ensemble of multiple (simple) decision trees, was created in 1995 by Tin Kam Ho, then extended by Leo Breiman and Adele Cutler. Today, is one of the most known algorithms in ML. The XGBoost algorithm gained significant recognition in prominent Kaggle competitions and many use case applications due to its exceptional performance and rapid response in addressing classification and regression predictive modeling challenges, particularly for structured or tabular datasets. Recent examples of its effectiveness can be found in studies such as [18] and [13].

The ML models are not used as predictive algorithms, meaning that no train/test data split exists. The intention is to use the models as approximations to the true process behaviour, so all the data (the so-called *Classification dataset*) is used with the aim to obtain the highest possible ML evaluation metric. This dataset is composed by the label (the 2-Class Index) and the set of features: the MEA input variables, the MEA output variables, and all process variables that can be understood as factors for root cause analysis.

2.3. Performance Explainability using XAI techniques

For this work, two widely used techniques for XAI were employed, LIME and SHAP. Local Interpretable Model-agnostic Explanations (LIME) is an XAI technique supporting the interpretability of complex (black-box) models. The method is model-agnostic and works by approximating the behaviour of a model around a specific instance of interest, hence its local explanation ability. To this effect, it perturbs the instance by sampling nearby points and obtains predictions from the model. This enables LIME to build a simpler, interpretable model around the perturbed instances. By inspecting the coefficients of the interpretable model, LIME identifies the most influential features for a specific prediction, enabling an analysis to understand why a particular prediction was generated.

In what follows, we delve into more details about the SHAP technique, in order to clarify the version used and introduce three feature metrics that play a significant role in constructing the analytic graphs that allow a better interpretation of the root cause analysis of the LDM problem.

SHapley Additive exPlanation (SHAP) is a game theory approach that determines the ordinal contribution of individual features that influence a model's prediction f. It was first introduced by game theorist Lloyd Shapley in 1973 [24] mentioning that, in any coalitional game, the Shapley value is the average marginal contribution of the player to the overall marginal contribution of the set of players considering all possible permutations [28]. One of the most relevant issues in SHAP is the idea of unification of the attributive feature methods satisfying some properties: *Local Accuracy, Missingness, and Consistency* proposed in [21] and reviewed by Chen et. al in [9]. This approach has the advantage of working with marginal contributions in spite of conditional contributions as explained in [17].

For a given set of d players' indices $D = \{1, \dots, d\}$ and a coalitional game with associated value function $v : 2^D \to \mathbb{R}$, where 2^D denotes the power set of D, the (classical) SHAP value of a player i can be computed as

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} w_i(S)\ [v(S \cup \{i\}) - v(S)] \quad \text{with} \quad w_i(S) = \frac{|S|!(d - |S| - \{i\})!}{d!}, \tag{5}$$

where w_i is the weight of the coalition S, see [21, 19, 4, 9], also for variants of the classical method. Now, given a set of features $x \in \mathbb{R}^d$ (associated with the players' indices D) and a machine learning classifier $f : \mathbb{R}^d \to I \subset \mathbb{N}$, the key point is to construct a coalitional game value function $v \equiv v_{f,x}$, which can be chosen to represent various behaviors, including the model's loss for a single row or the global behavior of classifier on the training dataset [11]. Since this approach works with binary variables (dummy), the class of additive feature attribution methods is considered (see [21]), meaning that the explainer g is linear and verifies $g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i z'_i$, where $z' \in [0, 1]^d$, with values representing the absence or presence of features in the coalitional subset, respectively. The variable z' (the coalitional vector of missing values) is obtained from z (simplified input vector) as the result of a mapping h where $z' \approx x'$, $x = h_x(x')$, and $f(z) = f(h_x(z'))$. Notice that the explainer model g must satisfy the desirable properties of local accuracy, missingness, and consistency, as introduced in [20]:

- **Local accuracy:** The accuracy defines that for a given map function $x = h_x(x')$, we can also map between the simplified feature inputs z and simplified feature inputs of missing values z' so, then we can obtain an explainer model g which approximates f;
- **Missingness**: Missing values $x' = 0$ have no payoff value/marginal contribution in the coalitional game, so $x'_i = 0 \implies \phi_i = 0$;
- **Consistency**: Consistency states that if a model f is changed to another model f' the feature attribution assigned to a certain feature remains unchanged.

To compute the SHAP values, the value function is defined as $v_{f,x}(S) = g(h_x(z'))$. For each row index $r \in [R] \subset \mathbb{N}$ of a dataset with R rows, a SHAP value explaining the contribution of feature $i \in D$ can be computed as $\psi_i(x) = \phi_i(v_{f,x})$.

For each feature $i \in D$, we introduce the following metrics:

- **Negative impact of feature** i: Is the sum of all benchmark units' contributions with negative SHAP values, i.e., $\gamma_i^- = \sum_{r\in[R]} \min\{0, \psi_i(x^{(r)})\}$;
- **Positive impact of feature** i: Is the sum of all benchmark units' contributions with positive SHAP values, i.e., $\gamma_i^+ = \sum_{r\in[R]} \max\{0, \psi_i(x^{(r)})\}$;
- **Global impact of feature** i: Is the sum of all benchmark units' contributions, i.e., $\gamma_i = \sum_{r\in[R]} \psi_i(x^{(r)})$.

These metrics support an analytic graph analysis, which the authors believe better explains the root cause factors of the problem under study (see Fig. 3) when compared with standard SHAP plots.

3. Case study: Last-mile delivery in Portugal from Porto's logistics platform

The main application of the proposed approach was to explore a dataset provided by MAEIL, a company based in Portugal that develops transporter logistics management software and Enterprise Resource Planning (ERP) integration for small and medium-sized transportation companies. The dataset contained information from April 3 to May 3, 2023.

Due to the complexity of the delivery network and presentation limitations, two filters to the dataset were applied. Firstly, solely the deliveries with origin in the logistics platform of Porto have been presented (the second largest urban center in Portugal and with one of the highest LMD volumes - 400k instances in the dataset of the period of study). In this context, there are 47 different destination zones. Thus, a second filter was applied for destination zones and solely 4 zones have been included for the final analysis (those whose delivery volume represents more than 25% of the total of deliveries: *Zone1*, *Zone4*, *Zone7* and *Zone21*). Herein, each n used in the MEA algorithm is defined as the index n of the classification dataset associated with the tuple $(LMDserviceID, zoneID)$, where the first parameter represents the unique identifier of the LMD service and the second the unique identifier of the destination zone where the LMD was conducted (see Fig. 2). As mentioned earlier, MEA is characterized by its approach of perceiving all problems as ”black box” problems, where inputs are consumed and outputs are produced. For this case study, the input variables are the *Delivery Time* and the *Service Cost*, whilst the outputs are the *Delay Time* and *CO2 Emissions*.

The *Delivery Time* represents the difference between the time a *LMDserviceID* is available in the warehouse and the time it is ”concluded” in the ERP, meaning that the product was delivered to the end customer in the respective *zoneID*. The *Service Cost* represents an estimation regarding the total cost of the LMD service. The *Delay Time* of a service is defined as the difference between the *Delivery Time* of the service and the median value of the *Delivery Time* of the corresponding *zoneID* (during the above-mentioned period), where negative values are truncated to zero. Lastly, the *CO2 Emissions* is estimated as a combination of parameters, i.e., the fleet's initial investment, fuel, insurance, maintenance, and human resources-related costs, per km. Table 2 refers to the global characterization of these variables per *zoneID*. Please note, the best-ranked n from the application of the benchmarking algorithm, are those that consume fewer inputs and produce more outputs. This means that in order to maximize the outputs outlined in Fig. 2, both Delay Time and CO2 Emissions need to be converted into their complementary version. The complementary value of an entry j of the variable V is defined by $cV_j = max_n V_n - V_j$. So, if V is maximized, then the complementary cV is minimized, and the other way around. Table 2 shows the output variables in their complementary version, e.g., ***cDelay** Time*. Recall that MEA assumes (by construction) that efficient services minimize inputs and maximize outputs.

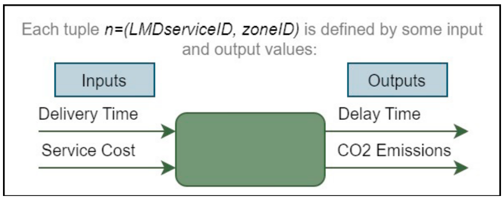
Each tuple n=(LMDserviceID, zoneID) is defined by some input and output values:
Inputs
Outputs
Delivery Time
Service Cost
Delay Time
CO2 Emissions

Fig. 2: Initial problem characterization of the case study.

Variable	zone1	zone4	zone7	zone21
Delivery Time (hours)	10.97 ± 1.39	15.89 ± 2.50	16.35 ± 2.79	15.37 ± 3.55
Service Cost (euros/package)	4.77 ± 9.37	1.74 ± 0.90	1.23 ± 0.46	1.44 ± 0.24
cDelay Time (hours)	5.55 ± 1.18	6.34 ± 2.19	6.55 ± 2.29	5.33 ± 3.03
cCO2 Emissions (grams)	111.78 ± 8.62	2.69 ± 0.49	0.71 ± 0.37	19.33 ± 0.15

Table 1: Global characterization of the variables used in the benchmark analysis, per *zone ID* (mean ± standard deviation values).

4. Results and discussion

4.1. Analysis of most critical zones of last-mile delivery

Following the methodology presented in Fig. 1, the primary result of this work is depicted in Table 2. Both *zone4* and *zone21* exhibit the lowest percentage of LMD services with class H. Nevertheless, it is noteworthy that *zone4* also possesses the lowest percentage of LMD services with the lowest MEA scores (class L). Consequently, despite the scarcity of high performers in this zone, the overall performance surpasses that of *zone21*. Thus, it can be inferred that attention from decision-makers should be directed towards investigating *zone21*. Moreover, recall that class L is found by using the first quartile threshold, so a 25% is expected for class L, meaning that *zone4* has fewer LMDservices in the low efficiency group, increasing the idea that it is a zone centered on the normal efficiency group (class N) with small variability. What seems interesting for further investigation (acquiring new data) is the fact that there is a low percentage of LMD services within the high efficiency group (class H), meaning that there is space to introduce actions that promote a significant improvement of the zones' performance. Using (3), the inefficient values $\text{mIneff}_i(n)$ and $\text{mIneff}_o(n)$ play a relevant role in identifying the variables and their amount of inefficiency, for each LMDservice n (data was not presented here for space reasons).

Efficient Groups	zone1		zone4		zone7		zone21	
L : 3-CI(n) $\in [0, Q_1^*[$	102	25.1%	490	15,5%	1098	25,0%	778	**25,0%**
N : 3-CI(n) $\in [Q_1^*, Q_3^*]$	300	73,9%	2664	84,4%	3274	74,5%	2330	74,9%
H : 3-CI(n) $\in]Q_3^*, 1]$	4	1,0%	4	0,1%	24	0,5%	4	**0,1%**
	406	100,0%	3158	100,0%	4396	100,0%	3112	100,0%

Table 2: Number and percentage of *LMDserviceIDs*, per *zoneID*, with MEA scores associated to classes L, N and H.

4.2. Analysis of most relevant variables that explain last-mile delivery services belonging to the efficient or inefficient groups

Before the application of LIME and SHAP, an extensive grid hyperparameter $F_1 - score$ optimization was performed for XGBoost, Random Forest, Support Vector Regression, and Artificial Neural Network. The best results were obtained for XGBoost (*zone7*) and Random Forest (the other zones), with $F1$-scores and accuracy scores in the range $[0.89, 0.94]$. Then, the application of LIME and SHAP was proceeded for each best model, without significant differences in the relative order of variable importance so, in what follows, only SHAP results are presented.

The first generated graph per zoneID was the well-known beeswarm (see Fig. 3**(a)**, Fig. 3**(c)**, Fig. 3**(e)**, Fig. 3**(g)**), one of the most common graphical results of SHAP, for those familiar with the approach. However, not satisfied with such plots, extra efforts were made to produce a slightly more extensive plot, capable of revealing in-depth information from SHAP, to specifically highlight the root causes of MEA scores (see Fig. 3**(b)**, Fig. 3**(d)**, Fig. 3**(f)**, Fig. 3**(h)**). These new bar plots are able to provide insights into the three metrics outlined in subsection 2.3. Looking at the red and blue bars, respectively, it is possible to assess the **positive impact** (γ_i^+) and **negative impact** (γ_i^-) in that a certain variable produces in the MEA score attributed to the *LMDserviceIDs* per *zoneID*. Additionally, within

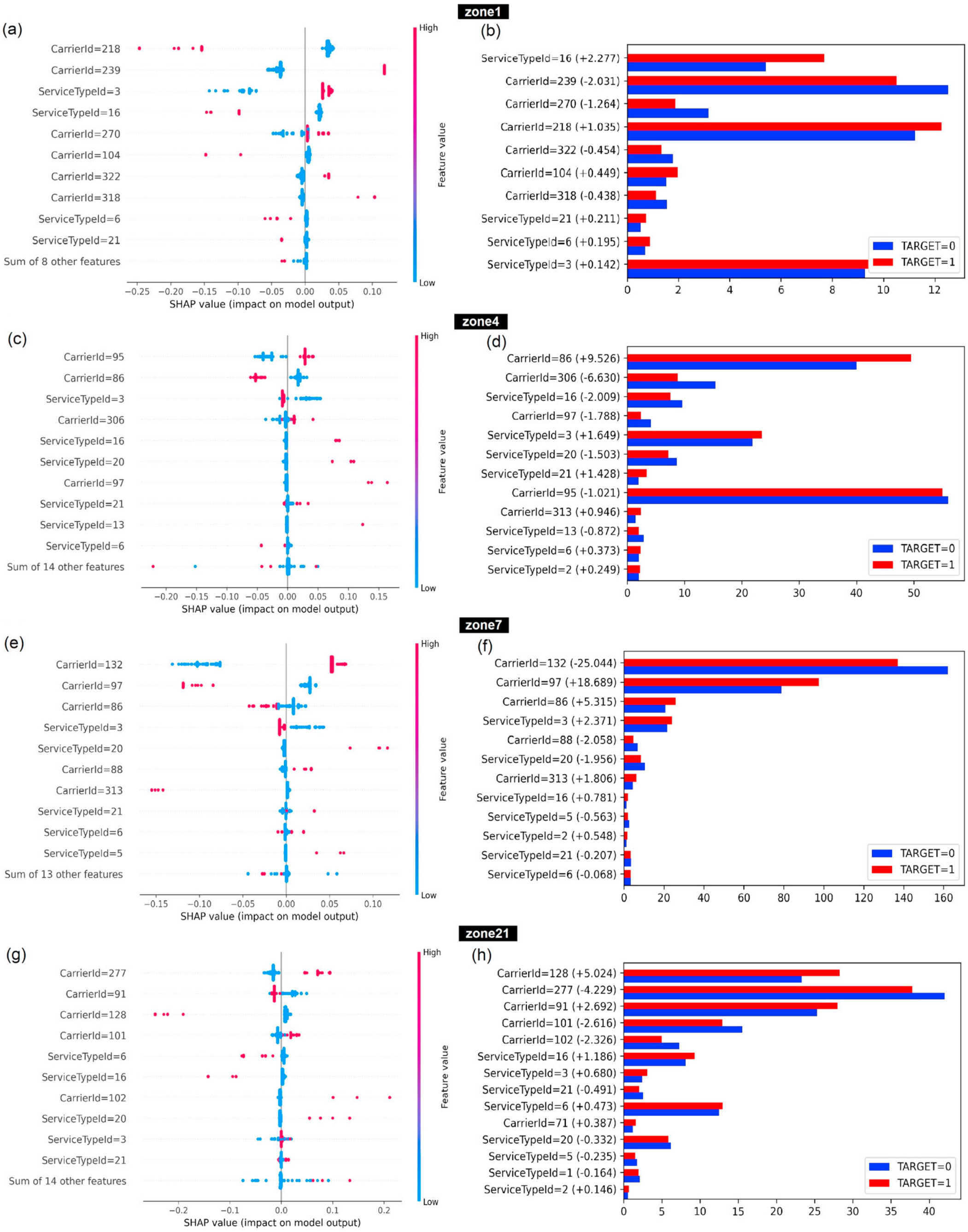

Fig. 3: SHAP results. Plots **(a), (c), (e)** and **(g)** - typical beeswarm graphs obtained with SHAP. Plots **(b), (d), (f)** and **(h)** - extended bar plots representing the variables' positive (red bar) or negative (blue bar) impact on the MEA score of an LMDserviceID per zoneID; and the difference between these impacts (value indicated within parentheses on the y-axis). Please note, the order of the y-axis variables is determined by the module of the value within parentheses.

parentheses on the y-axis, the **global impact** (γ_i) of a variable can be found. Such global metric aims to evaluate if the impact of a variable is leaning more towards a positive or negative effect in efficiency of the service .

An initial examination of the bar plots enables the identification of a group of *ServiceTypeIDs* and *CarrierIDs* for which implementing corrective measures within their LMD service would yield a positive impact on the performance for its corresponding *zoneID*. Nevertheless, a deeper analysis reveals that the *ServiceTypeID=3*, which, despite exhibiting the lowest global impact in the case of *zone1*, possesses the highest potential to substantially enhance the ranking of *serviceIDs* within all four zones. This observation bears substantial significance since, in a general analysis, this variable would likely not be deemed a priority for improvement.

Additionally, assessing the global impacts, in *zone4*, the *CarrierID=86* is pointed out as a good case practice, having the highest positive (+9.526) impact on LMD performance when compared to other carriers (note that the typical beeswarm graph in the left could not provide such insights directly). Contrarily, showing a big negative impact of -25.044, services made by *CarrierID=132* should be cautiously analysed and assessed to improve performance efficiency. When γ_i is significantly lower than the $\min\{\gamma_i^-, \gamma_i^+\}$, e.g. for *CarrierID=95* in *zone4*, such variable should be subject to monitoring and improvement actions since a small variation percentage produces a big impact on the global performance. These heuristic rules can be incorporated into an automatic recommendation system, connected to a digital twin processing logistic data in real-time and periodically triggering new recommendations.

5. Conclusions and future work

This work develops and applies a novel data-driven approach to evaluate and explain the performance of LMD services, guiding logistics service providers towards higher efficiency service rates through recommendations generated by new SHAP metrics, derived from ML fitting of a benchmark score.

The results obtained hold substantial relevance for this study. They enable precise identification of the most influential factors contributing to good and bad performance rankings, including those variables that unexpectedly have the capacity to improve the ranking of LMD services within the inefficient group. Such ”root cause” findings are highly valuable in guiding decision-makers towards targeted improvements. Furthermore, the authors believe that the performance assessment and explainability methodology developed and validated in this work can be applied to other practical cases in last-mile logistics if the problem definition and mathematical framework are strictly adhered to.

This is one of the first works of the NEXUS agenda incorporating more than 20 partners, which opens several future paths of application, robustness testing, interoperability with other solutions and mass deployment. One of the future activities of the research team is the creation of a Logistic Control Tower, which may benefit from solutions such as the one proposed in this work.

Author Contributions. *Conceptualization, methodology, and formal analysis*: AB, ER. *Data curation*: AB, ER, EA; *Python implementation*: ER, EA; *Methodology validation*: AB, ER, SS; *Manuscript writing*: AB, ER, SS.

Abbreviations: A. Brochado (AB), E. Rocha (ER), E. Addo (EA), S. Silva (SS). All authors have read and agreed to the published version of the manuscript.

Acknowledgements. The present study was developed in the scope of Agenda ”NEXUS Pacto de Inovação - Transição Verde e Digital para Transportes, Logística e Mobilidade” [C645112083-00000059], financed by PRR – Plano de Recuperação e Resiliência under the Next Generation EU from the European Union. A. Brochado holds a PhD grant in the scope of the Agenda NEXUS. E. Rocha was partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA), through the Portuguese Foundation for Science and Technology, project UIDB/04106/2020. S. Silva is partially supported by the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) funded by National Funds through the Foundation for Science and Technology, project UIDB/00127/2020.

Data Availability. The data used in this work is property of the project stakeholders so it’s not publicly available. For data requirements, please contact Gabriel Rosa (gabriel.rosa@maeil.pt). The authors are grateful for MAEIL’s collaboration and data provision.

References

[1] AI HLEG, 2019. A definition of AI: Main capabilities and scientific disciplines. Technical Report. European Comission. URL: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf.

[2] Alzubi, J., Nayyar, A., Kumar, A., 2018. Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series 1142. doi:10.1088/1742-6596/1142/1/012012.

[3] Arishi, A., Krishnan, K., Arishi, M., 2022. Machine learning approach for truck-drones based last-mile delivery in the era of industry 4.0. Engineering Applications of Artificial Intelligence 116, 105439. doi:10.1016/j.engappai.2022.105439.

[4] Bhattacharya, A., 2022. Applied Machine Learning Explainability Techniques. Packt Publishing.

[5] Bogetoft, P., Hougaard, J.L., 1999. Efficiency Evaluations Based on Potential (Non-Proportional) Improvements. Journal of Productivity Analysis 12, 233–247. doi:10.1023/A:1007848222681.

[6] Bogetoft, P., Otto, L., 2011. Benchmarking with DEA, SFA, and R. volume 157 of *International Series in Operations Research Management Science*. Springer New York, New York, NY. doi:10.1007/978-1-4419-7961-2.

[7] Brochado, Â.F., Rocha, E.M., Pimentel, C., 2022. Understanding and predicting process performance variations of a balanced manufacturing line at bosch, in: Guarda, T., Portela, F., Augusto, M.F. (Eds.), Advanced Research in Technologies, Information, Innovation and Sustainability. Springer Nature Switzerland, Cham. chapter ARTIIS 202, pp. 357–371. URL: https://doi.org/10.1007/978-3-031-20319-0_27.

[8] Capgemini Research Institute, 2019. The last-mile delivery challenge. Technical Report. URL: https://www.capgemini.com/wp-content/uploads/2019/01/Report-Digital--Last-Mile-Delivery-Challenge1.pdf.

[9] Chen, H., Covert, I.C., Lundberg, S.M., Lee, S.I., 2023. Algorithms to estimate shapley value feature attributions. Nature Machine Intelligence 5, 590–601. doi:10.1038/s42256-023-00657-x.

[10] Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, California, USA. pp. 785–794. doi:10.1145/2939672.2939785.

[11] Covert, I., Lundberg, S., Lee, S.I., 2021. Explaining by removing: A unified framework for model explanation. Journal of Machine Learning Research 22, 1–90. URL: http://jmlr.org/papers/v22/20-1316.html.

[12] Demir, E., Syntetos, A., van Woensel, T., 2022. Last mile logistics: Research trends and needs. IMA Journal of Management Mathematics 33, 549–561. doi:10.1093/imaman/dpac006.

[13] Fang, Z.g., Yang, S.q., Lv, C.x., An, S.y., Wu, W., 2022. Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. BMJ Open 12. doi:10.1136/bmjopen-2021-056685.

[14] Giuffrida, N., Fajardo-Calderin, J., Masegosa, A.D., Werner, F., Steudter, M., Pilla, F., 2022. Optimization and Machine Learning Applied to Last-Mile Logistics: A Review. Sustainability 14, 5329. doi:10.3390/SU14095329.

[15] Google Trends, . Explainable artificial intelligence (2004-2023). https://tinyurl.com/googletrendsXAI. Accessed: 2023-06-26.

[16] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z., 2019. XAI—Explainable artificial intelligence. Science Robotics 4. doi:10.1126/SCIROBOTICS.AAY7120.

[17] Janzing, D., Minorics, L., Blöbaum, P., 2020. Feature relevance quantification in explainable ai: A causal problem, in: International Conference on artificial intelligence and statistics, PMLR. pp. 2907–2916. URL: https://proceedings.mlr.press/v108/janzing20a.html.

[18] Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., Geng, Y.a., 2022. Application of XGBoost algorithm in the optimization of pollutant concentration. Atmospheric Research 276, 106238. doi:https://doi.org/10.1016/j.atmosres.2022.106238.

[19] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence 2, 56–67. doi:10.1038/s42256-019-0138-9.

[20] Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint doi:arXiv:1802.03888.

[21] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4768–4777. doi:10.5555/3295222.3295230.

[22] Machlev, R., Heistrene, L., Perl, M., Levy, K.Y., Belikov, J., Mannor, S., Levron, Y., 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy and AI 9, 100169. doi:10.1016/J.EGYAI.2022.100169.

[23] Motavallian, J., 2019. Last mile delivery in the retail sector in an urban context. Ph.D. thesis. RMIT University. URL: https://core.ac.uk/download/pdf/237115181.pdf.

[24] Narahari, Y., . Game theory lecture notes by chapter 32. the shapley value. URL: https://gtl.csa.iisc.ac.in/gametheory/ln/web-cp5-shapley.pdf. Lecture delivered in October 2012. Indian Institute of Science, Bangalore, India.

[25] Olsson, J., Hellström, D., Pålsson, H., 2019. Framework of Last Mile Logistics Research: A Systematic Review of the Literature. Sustainability 11. doi:10.3390/su11247131.

[26] Precedence Research, 2022. Last Mile Delivery Transportation Market Size, Report 2022-2030. Technical Report. Precedence Research. URL: https://tinyurl.com/precedenceResearchLMDMarket.

[27] Rocha, E.M., Brochado, Â.F., Rato, B., Meneses, J., 2022. Benchmarking and Prediction of Entities Performance on Manufacturing Processes through MEA, Robust XGBoost and SHAP Analysis, in: 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–8. doi:10.1109/ETFA52439.2022.9921593.

[28] Rozemberczki, B., Watson, L., Bayer, P., Yang, H.T., Kiss, O., Nilsson, S., Sarkar, R., 2022. The Shapley Value in Machine Learning. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22) doi:10.48550/arXiv.2202.05594.

[29] Tian, Z., Zhong, R.Y., Vatankhah Barenji, A., Wang, Y.T., Li, Z., Rong, Y., 2020. A blockchain-based evaluation approach for customer delivery satisfaction in sustainable urban logistics. International Journal of Production Research 59, 2229–2249. doi:10.1080/00207543.2020.1809733.