



# Toward Reducing Unproductive Container Moves: Predicting Service Requirements and Dwell Times

Elena Villalobos

[villalobos\\_elena@tec.mx](mailto:villalobos_elena@tec.mx)

School of Government and Public  
Transformation, Tecnológico de  
Monterrey

Adolfo de Unánue T.

[unanue@tec.mx](mailto:unanue@tec.mx)

School of Government and Public  
Transformation, Tecnológico de  
Monterrey

Fernanda Sobrino

[fersobrinno@tec.mx](mailto:fersobrinno@tec.mx)

School of Government and Public  
Transformation, Tecnológico de  
Monterrey

David Aké

[david.akeuitz@tec.mx](mailto:david.akeuitz@tec.mx)

School of Government and Public  
Transformation, Tecnológico de  
Monterrey

Stephany Cisneros

[stephany.cisneros@tec.mx](mailto:stephany.cisneros@tec.mx)

School of Government and Public  
Transformation, Tecnológico de  
Monterrey

Jorge Lecona

Container Terminal Operations  
Veracruz, Mexico

Alejandra Matadamaz

Container Terminal Operations  
Veracruz, Mexico

School of Government and Public Transformation

Working Paper No. 31

Publication update: April 2026

---

# TOWARD REDUCING UNPRODUCTIVE CONTAINER MOVES: PREDICTING SERVICE REQUIREMENTS AND DWELL TIMES

---

 **Elena Villalobos**

Centro de Ciencia de Datos e Inteligencia Artificial  
School of Government and Public Transformation  
Tecnológico de Monterrey  
Mexico City, Mexico  
villalobos\_elena@tec.mx

 **Fernanda Sobrino**

Centro de Ciencia de Datos e Inteligencia Artificial  
School of Government and Public Transformation  
Tecnológico de Monterrey  
Mexico City, Mexico  
fersobrinno@tec.mx

**Stephany Cisneros**

Centro de Ciencia de Datos e Inteligencia Artificial  
School of Government and Public Transformation  
Tecnológico de Monterrey  
Mexico City, Mexico  
stephany.cisneros@tec.mx

**Alejandra Matadamaz**

Container Terminal Operations  
Veracruz, Mexico

 **Adolfo De Unánue T.**

Centro de Ciencia de Datos e Inteligencia Artificial  
School of Government and Public Transformation  
Tecnológico de Monterrey  
Mexico City, Mexico  
unanue@tec.mx

**David Aké**

Centro de Ciencia de Datos e Inteligencia Artificial  
School of Government and Public Transformation  
Tecnológico de Monterrey  
Mexico City, Mexico

**Jorge Lecona**

Container Terminal Operations  
Veracruz, Mexico

April 9, 2026

## ABSTRACT

This article presents the results of a data science study conducted at a container terminal, aimed at reducing unproductive container moves through the prediction of service requirements and container dwell times. We develop and evaluate machine learning models that leverage historical operational data to anticipate which containers will require pre-clearance handling services prior to cargo release and to estimate how long they are expected to remain in the terminal. As part of the data preparation process, we implement a classification system for cargo descriptions and perform deduplication of consignee records to improve data consistency and feature quality. These predictive capabilities provide valuable inputs for strategic planning and resource allocation in yard operations. Across multiple temporal validation periods, the proposed models consistently outperform existing rule-based heuristics and random baselines in precision and recall. These results demonstrate the practical value of predictive analytics for improving operational efficiency and supporting data-driven decision-making in container terminal logistics.

**Keywords** Machine Learning · Port Terminal · Data Product · Decision Support Systems · Mexico

## 1 Introduction

Container terminals play a central role in global supply chains, acting as critical nodes where maritime and inland transportation systems converge. Their efficiency directly affects vessel turnaround times, hinterland connectivity, and the overall performance of international trade networks [UNCTAD, 2025]. As global cargo volumes continue to grow, optimizing terminal operations has become increasingly important for maintaining competitiveness and ensuring service reliability.

Within container terminals, yard management is one of the most complex and consequential operational processes. Decisions regarding where and how containers are stacked influence equipment utilization, operational costs, and the number of unproductive moves—such as reshuffles (rehandling moves)—required to access containers buried beneath others [Gharehgozli et al., 2017]. Reducing these non-productive moves is essential, as they account for a significant share of total handling activity and represent substantial resource consumption, including fuel, labor time, and equipment wear.

Traditional approaches to yard planning rely on deterministic optimization models or static heuristics that assume known departure sequences. However, container behavior in practice is inherently stochastic: certain containers require pre-clearance processes prior to departure, consignee pickup behavior follows probabilistic rather than fixed schedules, and dwell times are influenced by complex interactions among cargo type, shipping line, origin port, and seasonal factors. These characteristics make the problem well-suited to machine learning [Bengio et al., 2018], which can exploit large volumes of historical operational data to learn predictive patterns that would be difficult to encode manually. Rather than optimizing a stacking plan given assumed departures, we reframe the problem as one of *prediction*.

This study develops a data-driven system designed to support yard planning by estimating two key pieces of information before container arrival: whether a container will require a pre-clearance handling service, and how long it is expected to remain in the terminal. Pre-clearance services refer to administrative procedures conducted by customs brokers prior to cargo release, which require the terminal to reposition the container for handling. By anticipating these service requirements and dwell times, yard planners can make more strategic stacking decisions, placing containers requiring service closer to designated areas, short-stay containers in more accessible locations, and long-stay containers in lower-priority zones, thereby reducing unnecessary reshuffles.

## 2 Literature review

Research on container terminal operations has evolved from rule-based heuristics to data-driven approaches that leverage machine learning to improve yard efficiency. This section reviews the literature, organized around three themes: container dwell-time prediction, stacking and relocation optimization, and the integration of predictive analytics in terminal operations.

### 2.1 Container Dwell Time Prediction

The prediction of container dwell time has received increasing attention as terminals seek to optimize yard space utilization. [Kourounioti et al., 2016] proposed one of the first systematic approaches using Artificial Neural Networks (ANNs) to predict import container dwell times, identifying discharge timing, port of origin, container dimensions, and cargo type as key determinants.

More recent studies have explored ensemble machine learning methods. [Yoon et al., 2023] compared six algorithms—including Random Forest, XGBoost, and LightGBM—for vessel dwell-time prediction at Busan Port, finding that all ML models outperformed the terminal’s operational reference. [Saini and Lerher, 2024] analyzed 2.8 million container records across fourteen ports, identifying factors such as free storage periods, transshipment status, and proximity to industrial hubs as significant predictors of dwell-time variation.

The need for model interpretability has led to research incorporating Explainable AI (XAI). [Lee et al., 2024] combined Process Mining with XAI techniques to identify key factors prolonging container dwell times, emphasizing the importance of transparency for operational adoption.

### 2.2 Container Stacking and Reshuffles

The container stacking problem is fundamental to yard efficiency. [Kap Hwan Kim, 1997] established the mathematical foundation for quantifying rehandle costs based on stacking configurations, demonstrating that stack height and departure sequence uncertainty are primary drivers of unproductive moves.

Various optimization approaches have been proposed to minimize reshuffling. [Chafik et al. \[2016\]](#) developed a Mixed Integer Programming (MIP) model comparing First-Come-First-Served and Best Fit Decrease heuristics. [Caserta et al. \[2011\]](#) applied metaheuristic methods to the block relocation problem, demonstrating scalability to realistic terminal sizes. [Borgman et al. \[2010\]](#) proposed online stacking rules for real-time decisions, highlighting that accurate departure time information significantly improves stacking outcomes.

The integration of prediction with optimization represents an emerging research direction. [Gaete G. et al. \[2017\]](#) proposed a decision support system that first predicts dwell times using Random Forest regression and then applies heuristics to minimize reshuffles—an architecture closely related to the present study.

### 2.3 Machine Learning in Maritime Operations

The broader adoption of machine learning in port operations has been documented in recent systematic reviews. [Jahangard et al. \[2025\]](#) analyzed 124 papers on predictive and prescriptive analytics in seaports, identifying a gap in research that combines predictive outputs with optimization models. [Heilig et al. \[2019\]](#) established a conceptual framework for data-driven terminal planning, emphasizing that analytics reduces uncertainties and identifies causes of operational inefficiencies.

Related prediction problems in port logistics demonstrate the versatility of ML approaches. [Xie et al. \[2025\]](#) applied machine learning to predict container exit terminals, incorporating unstructured cargo descriptions. [Saber et al. \[2025\]](#) developed hybrid ML models for vessel arrival prediction, demonstrating the value of combining multiple algorithmic approaches.

### 2.4 Research Gaps and Contributions

Despite substantial progress, several gaps remain in the literature. First, while dwell time prediction has received considerable attention, the prediction of *service requirements* from a terminal operator’s perspective remains largely unaddressed. Substantial research exists on customs risk assessment for enforcement purposes—including ML-based targeting systems that help authorities identify containers for inspection based on contraband or compliance risks [[Hillberry et al. \[2022\]](#), [Vijayakumar \[2025\]](#), [Peri \[2020\]](#)]. However, these models serve government agencies rather than terminal operators. Predicting which containers will require pre-clearance handling for yard planning purposes—where the goal is operational efficiency rather than enforcement targeting—has not been explored. Second, most studies address either prediction or optimization in isolation; integrated approaches that feed predictions directly into operational decisions are limited [[Jahangard et al. \[2025\]](#)]. Third, rigorous temporal validation using frameworks such as temporal cross-validation [see [Roberts et al. \[2017\]](#)] is rarely implemented, raising concerns about data leakage [see [Rayid Ghani et al. \[2020\]](#)] and deployment validity. Fourth, comparisons typically use random or dummy baselines rather than actual operational heuristics employed by terminals.

This study addresses these gaps by: (1) developing models for both service requirement and dwell-time prediction; (2) employing rigorous temporal cross-validation<sup>1</sup>; (3) comparing against operational baselines currently used in practice; and (4) providing empirical results from a Latin American container terminal, expanding the geographic scope of this research domain.

## 3 Problem description

One of the critical operational processes in container terminals is the placement of import containers in the yard area. Each day, hundreds of containers are unloaded from incoming vessels and distributed across different yard blocks after discharge. The movement of containers within the terminal—from the quay to storage areas and eventually to the gate—is referred to as a **container move**. To optimize the use of available space, terminals stack containers vertically, a practice known as *container stacking*.

The stacking of containers inevitably leads to situations in which containers scheduled for departure or service are located beneath others. Retrieving these buried containers requires temporarily removing those above them, a process referred to as *reshuffling* or *rehandling*. Terminal operators classify these operations as **unproductive moves** or **waste**, since they do not directly contribute to the logistical flow of cargo.

Unproductive moves represent one of the major sources of inefficiency in yard operations. In the terminal where this study was conducted, up to 75% of all container-handling moves were classified as unproductive. Of these, approximately 51% were associated with containers requiring a pre-clearance service (hereafter referred to as “service”),

<sup>1</sup>Employing the Triage framework [[Center For Data Science and Public Policy \[2025\]](#), [Rayid Ghani \[2024\]](#)]

i.e., an administrative process carried out by customs brokers prior to cargo release, which requires the terminal to perform additional handling operations to position the container for review<sup>2</sup>. These figures indicate that containers requiring pre-clearance handling are a particularly significant driver of operational inefficiency in this setting.

A promising strategy to mitigate these inefficiencies is to anticipate, *before a vessel is discharged*, which containers are likely to require a service and to estimate their expected dwell time. Early identification of containers likely to require pre-clearance services enables yard planners to position them closer to designated service areas, reducing the distance and number of handling operations required for subsequent retrieval. Similarly, predicting dwell time supports differentiated stacking: containers expected to leave the terminal sooner can be placed in more accessible positions, while long-stay containers can be stored in less critical sections of the yard without obstructing outbound flows.

By integrating predictive information on both service and dwell times, yard planners can make more strategic and proactive placement decisions. This approach has the potential to substantially reduce rehandling operations, improve equipment utilization, and optimize overall yard space management. To operationalize this predictive framework, the next section presents the data sources, operational context, and system architecture that support the development and evaluation of the proposed models.

## 4 Data, operational context and architecture

### 4.1 Data

The port terminal provided two databases: (1) container operations and (2) container movements within the terminal yard. Additionally, a publicly available database known as the HS Catalog (*Harmonized System*) was used. A brief description of each dataset is presented below.

**Operations:** The terminal generated a view<sup>3</sup> from its CTS production database.<sup>4</sup> To create this view, the terminal combined information from various tables referring to container attributes and date-related information. The original version of the view contains approximately two million containers (a similar number to the work of Saini and Lerher [2024]), covering the most recent several years of operations<sup>5</sup>. Some examples of the container attributes included in this view are: loading port, container dimensions and type, cargo description, consignee, vessel arrival date, container entry and exit dates, country of origin, among others.

**Movements:** The terminal also shared another view from its database containing the movements performed by cranes to relocate containers within the terminal yard. This dataset contains approximately 13 million rows. The data spans the same multi-year period and was used for exploratory analysis of container movements within the terminal.

**HS Catalog:** The HS Catalog (*Harmonized System*) is publicly available and serves as a numerical standardization method for the classification of traded goods. It is used to identify products for tariff and tax assessment and for the compilation of trade statistics. The catalog is updated every five years and serves as the foundation for international import and export classification systems. Specific six-digit codes are assigned to different categories and products<sup>6</sup>.

Access to operational data was subject to strict security and governance controls. All preprocessing and initial data transformations were conducted within the terminal's own computing infrastructure. Dedicated machines were provisioned with restricted permissions, allowing access only to the specific container attributes required for predictive modeling. Sensitive fields were excluded whenever possible and, when operationally necessary, were encrypted or anonymized by the terminal prior to researcher access. No direct access was granted to core production systems, and all analytical workflows operated on controlled extracts and derived views. These measures ensured compliance with the terminal's data protection policies while preserving the analytical value of the datasets.

---

<sup>2</sup>This process should not be confused with formal customs inspections conducted by authorities, which fall outside the scope of this study.

<sup>3</sup>Views are virtual tables formed by a *query* that are computed each time they are called.

<sup>4</sup>CTS stands for *Container Terminal System*, the production database used by the terminal.

<sup>5</sup>Exact calendar years are omitted at the request of the terminal port to preserve confidentiality. Temporal patterns are presented using relative time references and anonymized year labels (e.g., "20XX").

<sup>6</sup><https://www.trade.gov/harmonized-system-hs-codes>

## 4.2 Data Product Pipeline

To predict container service requirements and dwell times (DT), we developed a comprehensive data processing and enrichment pipeline. This workflow forms the technical backbone of the data product that supports predictive modeling for enhanced operational yard planning (see Figure 1).

The pipeline starts with the integration of heterogeneous **operational data** sources generated across different systems involved in container terminal operations. To ensure traceability and reproducibility, the data are organized into a staged schema architecture. First, all incoming datasets are ingested as-is into a schema referred to as RAW, which preserves the original structure, formats, and potential inconsistencies of the source systems.

In a second step, the data are transformed into a CLEAN schema, where variables are standardized, cast to their appropriate data types, and semantically interpreted according to their operational context. This stage resolves inconsistencies, harmonizes naming conventions, and establishes a coherent representation of the underlying processes.

For the modeling stage, a domain-driven data model is constructed in a schema named ONTOLOGY. This schema is organized around two main structures. The first corresponds to entities, which in this study are containers and their atemporal attributes—i.e., characteristics that remain invariant over time, such as weight, dimensions, or cargo type. The second structure captures events, representing all operational occurrences that affect entities over time. These events are explicitly associated with timestamps, enabling a temporal representation of container trajectories and operational states.

Subsequently, two record linkage processes are implemented to enrich the data model and ensure consistency across operational entities. The first focuses on the classification of merchandise descriptions into standardized cargo categories, while the second addresses the deduplication and consolidation of consignee identities. Both processes, briefly described below, are essential for generating stronger predictors and enabling reliable modeling.

**Merchandise:** The database includes a variable denoted as merchandise description, which is a free-text field provided by users to describe the contents of each container. These descriptions may consist of natural-language text or, in some cases, references to codes derived from the Harmonized System (HS) catalog. Due to this heterogeneity, the raw descriptions cannot be directly used as structured predictors.

To obtain a standardized representation of cargo type, merchandise classification is grounded on the Harmonized System (HS) nomenclature, which provides an internationally recognized taxonomy comprising 21 high-level sections and 97 detailed chapters. However, only approximately 25% of the containers include an explicit HS reference in their description. Consequently, an automated classification approach based on Natural Language Processing (NLP) is required for the remaining cases.

Specifically, a TF-IDF representation [Salton and Buckley, 1988] is employed to quantify the relevance of terms in each merchandise description relative to the HS catalog. TF-IDF vectors are constructed for HS chapters using their textual definitions and compared against container-level descriptions to assign the most relevant chapter. This mapping yields a standardized cargo category that is subsequently used as a predictive feature across all models.

Alternative embedding-based approaches, including Word2Vec [Mikolov et al., 2013], were evaluated but ultimately discarded due to higher computational cost and inferior classification performance in this setting. The selected TF-IDF-based approach was validated through manual inspection of a random sample of container descriptions and qualitative comparison between HS chapter definitions and the resulting classified descriptions. As a result of this process, standardized merchandise classification coverage increased to 88% of all containers.

**Consignee:** This information contains variations in spelling, abbreviations, punctuation, or formatting for the same underlying entity. These inconsistencies lead to artificial inflation in the number of unique consignees and introduce noise when the variable is used directly for analysis or modeling. To address this issue, a record linkage process is implemented to deduplicate and consolidate consignee identities.

The linkage procedure begins by generating candidate pairs of consignees and computing their string similarity using character-based trigrams. To reduce the computational burden associated with exhaustive pairwise comparisons, a blocking strategy is applied. Comparisons are restricted to consignees whose names share the same initial character, which substantially reduces the number of candidate pairs while preserving relevant matches.

Candidate pairs with similarity scores above a threshold of 0.8<sup>7</sup> were retained and represented as edges in an undirected graph, where nodes correspond to consignee identifiers. A depth-first search (DFS) algorithm is then applied to identify connected components within the graph, effectively grouping the different textual representations of the same consignee across the dataset.

<sup>7</sup>The threshold was selected based on manual inspection of candidate pairs across different similarity levels.

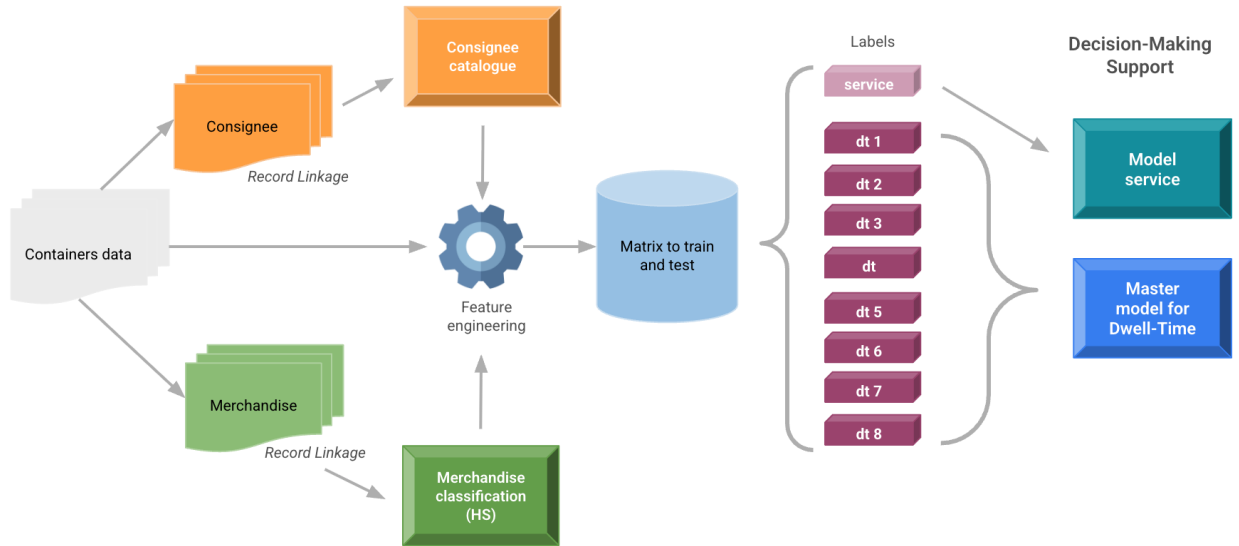


Figure 1: Data product pipeline.

This graph-based consolidation process results in a reduced and more consistent consignee catalog, where each connected component represents a unique underlying entity. By resolving duplicate identities and harmonizing consignee information, this procedure improves data consistency and reduces noise in downstream analyses. The resulting consolidated consignee identifier is subsequently used as a predictive feature in the models described in the following section.

Both standardized variables, the consolidated merchandise category and the deduplicated consignee identifier, are incorporated into the feature engineering stage. This stage is designed to derive explanatory variables that capture both operational dynamics and service-related patterns observed throughout the container lifecycle. The outcome is a structured **feature matrix** used for training and validating the machine learning models. This matrix is associated with **two labeling schemes**: one identifying whether a container will require service, and another specifying the expected dwell-time interval within the terminal.

Dwell time labels are further refined through a **decision rule** that assigns a unique final label to each container. Consequently, the data product outputs two ranked lists based on model scores. The first ranking orders containers according to their likelihood of requiring service, while the second ranking is derived from a decision rule applied to the dwell-time model scores, as described in the following sections.

## 5 Methodology

### 5.1 Analytical Formulation

This section describes the main elements used for model evaluation: the cohort and the labels. The cohort refers to the set of entities on which predictions are made, while the labels define the outcomes to be predicted.

**Cohort:** Containers that are within 24 hours of arriving at the terminal.

**Labels:** Nine distinct labels were defined, each determined 24 hours prior to the vessel's arrival at the terminal:

1. Which are the  $n$  containers that will leave the terminal in less than two days?
2. Which are the  $n$  containers that will leave the terminal on day  $m$ ?, where  $m \in \{2, 3, 4, 5, 6, 7\}$
3. Which are the  $n$  containers that will leave the terminal after more than seven days?
4. Which are the  $n$  containers that will receive service within the next seven days?

In earlier stages of the project, a more aggregated classification scheme was used: *less than two days*, *two to four days*, *four to seven days*, and *more than seven days*. However, a finer-grained, day-level labeling scheme was ultimately adopted to enhance predictive performance and generate more operationally actionable insights.

The choice to train separate binary classifiers for each dwell-time label—rather than a single multi-class model or regression approach—was motivated by three considerations. First, binary decomposition enables label-specific threshold optimization via ROC analysis, allowing operational trade-offs between precision and recall to be calibrated independently for each dwell time category based on their distinct operational costs. Second, the predictability of dwell time labels varies substantially: extreme categories (short and long stays) exhibit stronger predictive signals than intermediate durations, and binary classifiers allow each model to exploit the feature space differently without forcing shared representations. Third, this architecture provides operational flexibility, as terminals may prioritize accurate identification of specific categories (e.g., short-stay containers for accessible positioning) while tolerating lower precision on others.

### 5.1.1 Baseline Rate

Understanding the **baseline rate** is fundamental for assessing the difficulty of the predictive problem. Based on the most recent years of historical data, if a random set of containers is selected on any given day, only about 33% of them require at least one service.

Regarding dwell time, Table 1 presents the baseline distribution of containers according to the defined labels. For instance, *Dwell Time Label 1* corresponds to containers that exited the terminal in less than two days, whereas *Dwell Time Label 8* groups those that remained for more than seven days. Most labels account for between 8% and 12% of cases, except for the last category, which represents 28% of containers.

Table 1: Baserate dwell-time distribution over the study period.

Dwell Time Label	Cumulative Percentage	Label Percentage
1	8%	8%
2	18.36%	10%
3	30.60%	12%
4	42%	12%
5	54.07%	12%
6	64.27%	10%
7	72%	8%
8	>72%	28%

### 5.1.2 Baselines

The terminal currently employs simple operational heuristics to anticipate which containers will require pre-clearance services. These rules constitute the **current operational baseline**, and they represent the level of performance that our machine learning models aim to surpass.

**Operational Baseline 1.** This rule relies on the historical behavior of consignees. Consignees that have frequently required pre-clearance services in the past are assumed to have a higher probability of requiring service for future containers.

**Operational Baseline 2.** Under this rule, if a consignee has required pre-clearance services for more than 85% of its containers over the previous six months, it is assumed that upcoming containers associated with that consignee will also require service.

In contrast, there is currently no automated mechanism or established operational rule for predicting container dwell time. As a result, dwell time models are evaluated against a **random assignment baseline**, which serves as the reference performance they must improve upon.

## 5.2 Features

Feature selection was guided by two criteria: (i) variables had to be available prior to the prediction event, and (ii) they needed to capture meaningful operational information relevant to container behavior. A joint effort was undertaken to construct a detailed temporal mapping of all variables in the database, allowing us to identify the precise operational moment at which each variable was recorded. This temporal alignment was critical for the design of predictive experiments and for the correct implementation of Temporal Cross-Validation (see Subsection 5.3.2).

**Base variables.** Predictor variables were derived from information available prior to container arrival at the terminal. Table 2 summarizes the core variables that define the foundation of the feature space. Variables marked with an asterisk

(\*, \*\*) underwent additional *record linkage* processes (see Section 4.2) to standardize consignee entities and to assign structured classifications (chapter and section) to merchandise descriptions.

Table 2: Base variables available prior to container arrival.

Variable	
Net weight	Container dimension
Gross weight	Container type
Hazardous cargo indicator	Cargo type
Liner client (yes/no)	Shipping line
Consignee*	Shipping line route
Chapter and section**	

**Feature generation.** From these base variables, an expanded feature set was constructed through systematic transformations, aggregations, and temporal windowing strategies. These operations were designed to capture behavioral regularities, temporal dynamics, and cross-entity relationships among operational actors (e.g., consignees, shipping lines, and cargo chapters).

Feature engineering included rolling-window statistics, historical frequency measures, trend indicators, and ratio-based variables computed at different aggregation levels. The following examples illustrate representative feature patterns used in the construction process (non-exhaustive):

- Rolling service-frequency indicators, such as the number and proportion of containers requiring services for a given consignee over the past  $n$  weeks.
- Rolling arrival counts and rates by cargo type or chapter within fixed temporal windows (e.g., three-week windows).
- Temporal trend features derived from recent dwell time distributions associated with specific shipping lines or routes.

In practice, multiple variants of these feature patterns were generated across different entities, time windows, aggregation levels, and normalization schemes, resulting in a rich and structured predictor space.

**Feature taxonomy.** To maintain a structured and reproducible framework, all predictors were systematically categorized according to their nature and derivation process. Table 3 summarizes the taxonomy of generated features, which served as the foundation for model design and interpretability.

Table 3: Feature categories and descriptions.

Category	Description
Simple variables	Directly available from base container information, 24 hours prior to arrival.
Simple counts	Frequency of occurrence of specific variables within a defined time window.
Aggregated metrics	Ratios, proportions, and relationships between counts across temporal spans.
Differences	Temporal variations, percentage changes, and dispersion measures between periods.
Service	Frequency of services by consignee, chapter, or shipping line over time.
Dwell time	Descriptive statistics derived from historical container departure records.
Movements	Aggregated information on container movements by consignee and chapter.

**Data integrity.** To ensure predictive validity, strict controls were implemented to prevent *data leakage*. All features were computed exclusively from information available before the prediction point, guaranteeing that models had no access to future data. This approach ensures that model performance during training faithfully reflects behavior under operational deployment.

### 5.3 Model Governance

A total of more than 5,000 models were trained under different configurations to predict both service and dwell-time labels. Model training and data product generation were carried out using *Triage*, a predictive modeling framework developed by the *Data Science for Social Good* initiative [Center For Data Science and Public Policy, 2025]. Although originally designed for public policy applications, *Triage* is particularly suited for data science projects with a strong temporal structure, such as the present port terminal case study.

Model governance refers to the systematic oversight of decisions that directly influence the predictive modeling process. It ensures reproducibility, transparency, and methodological consistency throughout the experimental pipeline. Within this project, governance was exercised across three key dimensions: (i) the definition of predictive labels already described in Section 5.1, (ii) the selection and parameterization of algorithms, and (iii) the design of temporal configurations for training and validation.

#### 5.3.1 Algorithmic configuration

A diverse ensemble of algorithms was explored to assess both linear and non-linear modeling capacities, ranging from simple baselines to ensemble-based classifiers. Table 5 in the Appendix summarizes the principal algorithms and hyperparameters evaluated.

Model performance was evaluated against the operational baselines described in Section 5.1: consignee-based heuristics for service prediction and random assignment for dwell time.

#### 5.3.2 Temporal Structure in Model Training and Validation

Since the terminal data are inherently time-dependent, we implemented a **Temporal Cross-Validation (TCV)** strategy [Roberts et al., 2017] for model training and evaluation. This technique enables multiple experiments to be conducted under a chronological structure consistent with real-world operations, adapting to the specific nature of each label. In this approach, models are trained on historical data and evaluated on future periods, thereby preserving the temporal sequence and preventing information leakage (*data leakage*).

In the present project, a temporal structure is defined using data spanning three years of recent operations, which establishes the temporal bounds of the analysis. Within this period, models are trained using rolling historical windows of six months and subsequently evaluated over a one-month validation period, which serves as a proxy for real-world deployment intervals. This design allows the modeling framework to adapt to temporal trends and evolving operational conditions.

Prediction points are generated on a daily basis, mirroring the operational process in which containers arrive at the terminal each day. Outcome labels are defined using rolling time windows whose length depends on the target variable. Specifically, label occurrence windows range from 2 to 7 days, reflecting the difference between short-horizon service requirements and longer dwell-time dynamics.

Finally, all temporal configurations are updated through monthly model retraining. This retraining frequency balances the need to incorporate recent operational patterns while retaining sufficient historical information to ensure stable and reliable learning.

#### 5.3.3 Decision Rule for Dwell Time Label Assignment

Since the dwell-time prediction problem is decomposed into eight binary classification tasks, a decision rule is required to assign a single final label to each container. This rule is informed by the analysis of the *Receiver Operating Characteristic* (ROC) curves generated for each dwell-time model.

The ROC curve graphically represents the relationship between the *True Positive Rate* (TPR) and the *False Positive Rate* (FPR) across different classification thresholds. It serves as a standard tool to evaluate the discrimination capacity of binary classifiers.

For each dwell-time label, the ROC curve was computed, and the optimal point was identified—defined as the threshold that maximizes the difference between the true positive rate and the false positive rate (also known as the Youden index). This point was then used as the cutoff threshold to identify containers most likely to belong to each label.

Since a single container could surpass the optimal threshold for multiple labels, a rule was required to assign a unique dwell-time label to each container. The adopted decision rule is based on the *ranking of model scores* assigned to each container. Specifically, among the labels for which a container is classified as positive (i.e., surpassing the threshold), the label with the highest relative score ranking is selected.

The procedure is as follows:

1. For each label, containers with scores exceeding the optimal ROC-derived threshold are selected.
2. Within each label, containers are sorted in descending order of their scores and assigned a ranking position.
3. For containers appearing in multiple labels, the label corresponding to the best (highest) ranking is selected.

This approach systematically resolves label overlaps by selecting the classification where the model exhibits the greatest relative confidence.

## 6 Results

To evaluate model performance, two primary metrics were employed: **precision** and **recall**, both calculated over the temporal validation periods defined in the configuration. Precision measures the proportion of true positives among the containers predicted as positive by the model—of all containers the model identified as likely to require service, how many actually did so? Recall represents the proportion of actual positive cases correctly identified by the model—how many of the containers that indeed required service were detected by the model?

These metrics were adapted to the operational constraints associated with each predictive label. The following subsections present the results for *service prediction*, followed by those for *dwell time*, each with their specific analytical considerations.

Table 4 summarizes the key findings across both prediction tasks, comparing the best-performing models against their respective baselines.

Table 4: Summary of key results across prediction tasks.

Task	Best Model	Precision	Recall	vs. Baseline
Service ( $k = 300$ )	Random Forest	75%	100%	+50 pp
Dwell time <2 days	Random Forest	30–40%	80%	+50 pp
Dwell time 3–6 days	Random Forest	<25%	<40%	+15 pp
Dwell time >7 days	Random Forest	80%	90%	+40 pp

*Note:* Service baseline refers to operational heuristics; dwell-time baseline is random assignment. Metrics represent peak performance across validation periods. pp = percentage points.

### 6.1 Service

For service prediction, the model evaluation was grounded in the terminal’s daily operational capacity to assign containers to the service area. According to the terminal operations team, this capacity is approximately 300 containers per day. Accordingly, we set  $k = 300$  to compute the metrics  $\text{precision@}k$  and  $\text{recall@}k$ , defined as follows:

- $\text{precision@}k$ : proportion of correctly predicted containers among the top  $k$  containers identified by the model as requiring service.
- $\text{recall@}k$ : proportion of all containers that actually required service and were correctly identified within the top  $k$  prioritized cases.

The value of  $k$  can be adjusted to simulate different operational scenarios, as illustrated later in this section.

Figure 2 summarizes the performance of the main models selected from a broader set of evaluated approaches. While multiple models were evaluated, the figure highlights representative results that reflect the progressive analytical logic adopted in this study: starting from existing operational heuristics, extending them with simple machine learning methods, and ultimately deploying a fully data-driven model with enriched feature engineering.

The evaluation begins with the two operational heuristics currently used by the terminal to anticipate which containers will require service, represented in blue as *Baseline 1* and *Baseline 2*. As mentioned before, these rules rely on the historical behavior of consignees and constitute the terminal’s current decision-making approach. Across evaluation periods, these heuristics exhibit modest predictive performance: for  $k = 300$ , both achieve an average precision of approximately 25% and a recall close to 30%, the latter being comparable to the baseline rate. While these rules provide a useful starting point, the results indicate clear limitations in their ability to prioritize containers effectively.

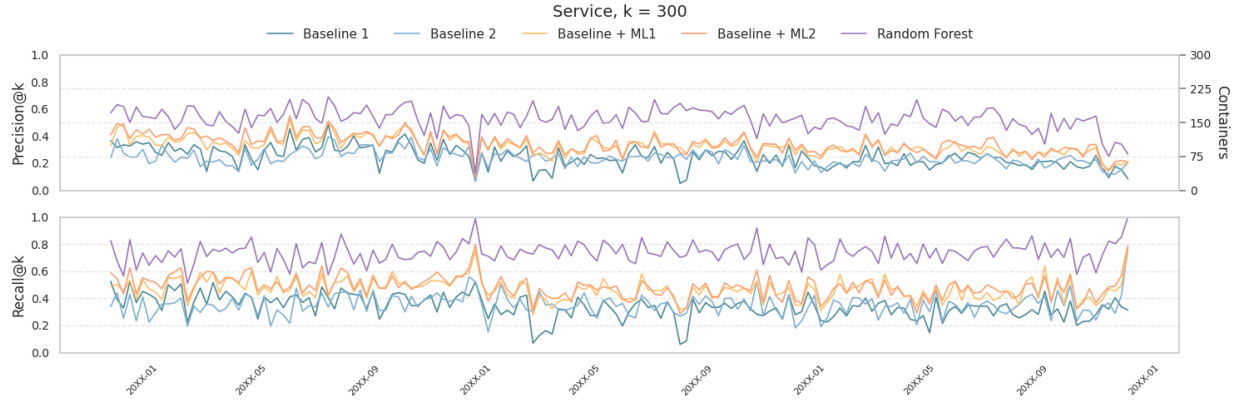


Figure 2: Weekly average precision and recall for the service label across the main evaluated models. The upper panel displays precision@k, while the lower one shows recall@k.

Building on these baselines, the next stage complements the operational rules with simple machine learning models, including decision trees and logistic regression, denoted as *Baseline + ML1* and *Baseline + ML2*. These hybrid approaches consistently improve performance, yielding gains of approximately 10–15 percentage points in both precision and recall relative to the rule-based baselines.

Finally, more sophisticated algorithms combined with specialized feature engineering and a comprehensive use of available data result in substantial performance improvements. In certain validation periods, the best-performing model achieves precision levels of up to 75% and recall values reaching 100%, indicating that all containers requiring service are correctly identified within the prioritized set. This model, depicted by the purple line in Figure 2, corresponds to a *Random Forest* classifier that integrates multiple dimensions of information, including historical service patterns, consignee characteristics, and recent behavioral indicators.

### 6.1.1 Effect of Available Space in the Service Area

Figure 3 illustrates how the best model would perform under a realistic operational scenario. On a typical day within the evaluation period, 1,726 containers arrived at the terminal, of which 347 required service. With  $k = 300$ , the top-performing model correctly identified 229 containers (66% precision) and captured 201 of the 347 that actually required service (58% recall).

This represents a substantial improvement—nearly 45 percentage points—over the best-performing combination of baseline and simple ML models, which achieved only 20% precision and 15% recall. A consistent pattern was observed across experiments: as  $k$  increases (i.e., more containers are prioritized), recall improves—since more true positives are captured—but precision decreases due to a higher number of false positives. This trade-off between precision and recall is crucial for defining operational strategies according to space availability in the service area.

## 6.2 Dwell Time

For dwell-time analysis, predictive models were trained for eight distinct labels, each corresponding to a specific duration that a container may remain in the terminal. This section reports results for three representative labels: containers with dwell times of less than two days, exactly five days, and more than seven days. The five-day label is selected as representative of the intermediate dwell-time categories, which include containers staying exactly three, four, six, and seven days.

Temporal evaluations for each label were conducted using an absolute  $k$  value corresponding to the historical number of containers that exhibited that specific dwell time. This value was computed by weighting the baseline rate of each label by the average number of containers arriving daily at the terminal (see Table 1). Although the results shown in the figures reflect this label-specific approach, it is important to note that the final dwell-time assignment for each container is not determined independently. As detailed in Subsection 5.3.3, this assignment is made through a decision rule that leverages optimal points derived from ROC curves to select the most appropriate label for each case.

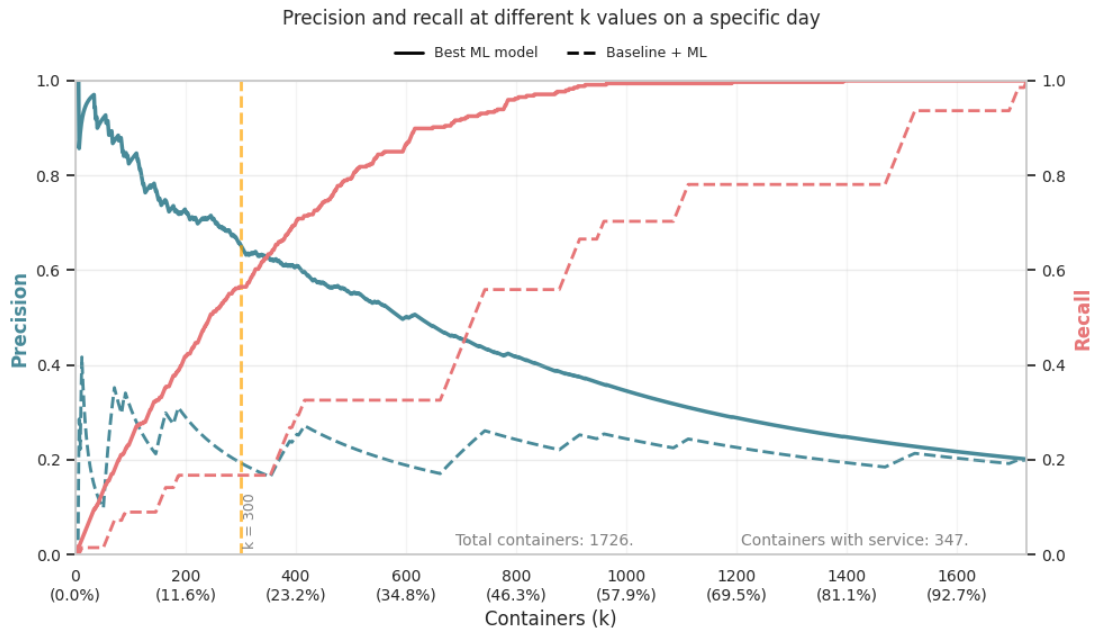


Figure 3: Precision and recall for different  $k$  values under varying service area capacities.

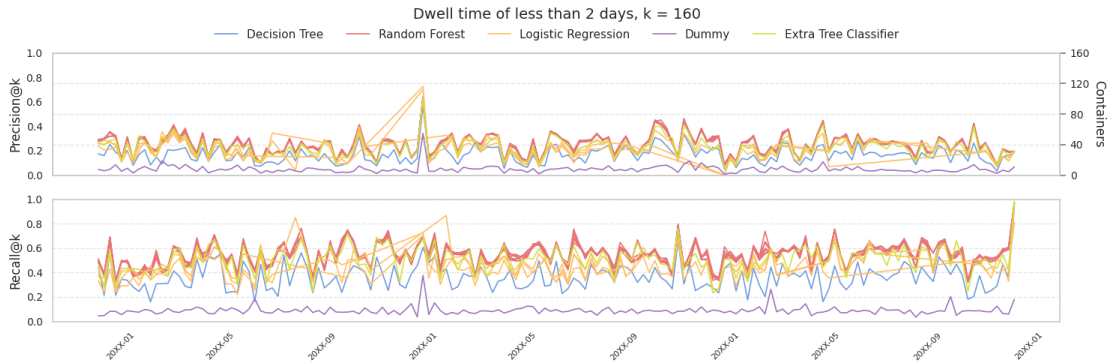


Figure 4: Weekly average precision and recall for dwell times of less than two days.

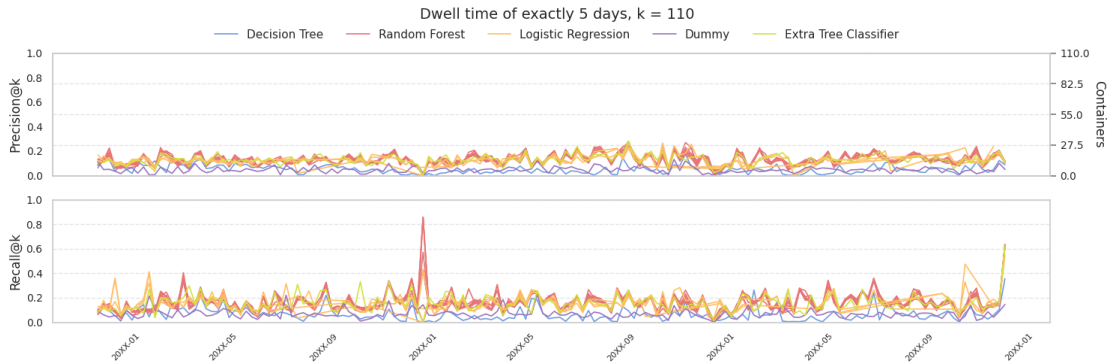


Figure 5: Weekly average precision and recall for containers with a dwell time of exactly five days. This label is used as a representative case for intermediate dwell-time categories (three, four, six, and seven days).

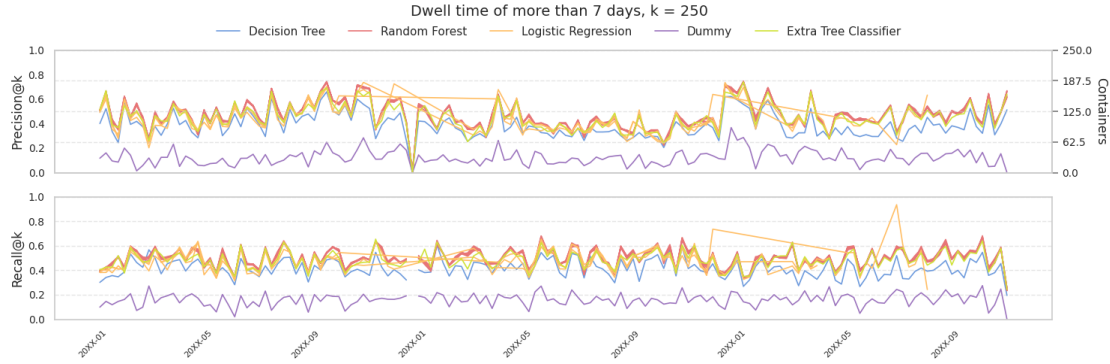


Figure 6: Weekly average precision and recall for dwell times longer than seven days.

### 6.2.1 Predictive Performance Across Dwell Time Labels

Figures 4, 5, and 6 illustrate predictive performance across selected dwell-time labels, revealing a clear pattern: model performance is strongest at the temporal extremes. Labels corresponding to containers that remain in the terminal for less than two days or more than seven days exhibit substantially higher precision and recall than intermediate dwell time categories. This suggests that operational behaviors associated with very short or very long stays are more regular and therefore easier for predictive models to capture.

The highest performance is observed for the label representing dwell times longer than seven days (Figure 6). For this category, several models achieve precision levels approaching 80% and recall values exceeding 90%, outperforming the random baseline by more than 40 percentage points. These results indicate that containers associated with prolonged stays follow more consistent operational patterns, making them particularly amenable to prediction.

For containers with dwell times of less than two days (Figure 4), the best-performing models achieve precision levels between 30% and 40% and recall values of up to 80%, outperforming the random baseline by approximately 50 percentage points. This result highlights the usefulness of predictive modeling for anticipating rapid departures and supporting the allocation of easily accessible yard positions.

In contrast, labels representing intermediate dwell durations—ranging from two to six days—exhibit more modest performance. For the representative intermediate case corresponding to a dwell time of five days (Figure 5), predictive performance remains moderate. Most models, including *ExtraTreesClassifier*, *Random Forest*, and *Logistic Regression*, achieve precision below 25% and recall rarely exceeding 40%. Although all models consistently outperform the random (*dummy*) baseline, the limited gains suggest higher operational variability within these temporal ranges, where multiple factors jointly influence mid-range dwell times and hinder consistent pattern identification.

Overall, the results demonstrate that dwell-time prediction is most effective at the extremes of the temporal spectrum, while intermediate durations remain more challenging. From an operational perspective, this finding presents a clear opportunity: containers predicted to remain in the terminal for extended periods can be proactively assigned to lower-priority yard locations where early retrieval is unlikely, thereby improving spatial efficiency and reducing unnecessary container movements.

### 6.2.2 Average Predictive Performance Over Time

Figure 7 shows the average precision and recall across all dwell time labels. This analysis was performed by selecting the best-performing model for each label, calculating its daily precision and recall, and then averaging these values at a weekly level.

Results were compared against a non-predictive benchmark, representing a purely random container assignment strategy. Across more than three years of validation using real operational data, the developed models consistently outperformed the random baseline in both precision and recall. Precision remained above 20% in most periods, while the random strategy barely reached 10%. Recall doubled or even tripled the levels achieved under random assignment.

These findings confirm that dwell-time prediction constitutes a practical decision-support tool for terminal operations. Rather than reacting as containers accumulate in the yard, operators can proactively plan and allocate strategic positions from the outset, thereby reducing unnecessary movements and optimizing spatial utilization.

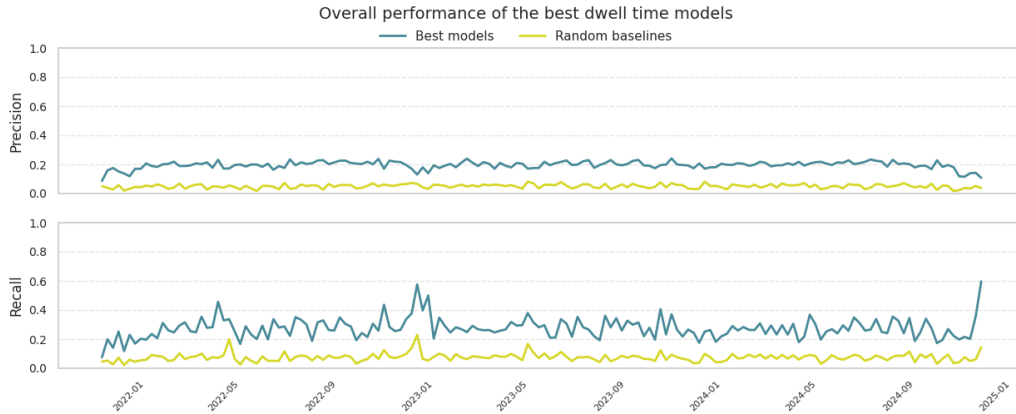


Figure 7: Average precision and recall of the best-performing models across dwell-time labels.

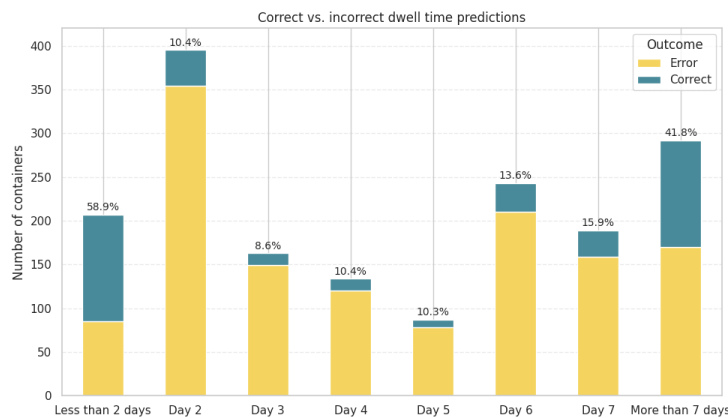


Figure 8: Correct and incorrect classifications by dwell-time category on a typical operational day.

### 6.2.3 Operational Performance of Dwell Time Assignment

Applying the decision rule described in Section 5.3.3, each container is assigned a unique dwell-time label. Figure 8 presents the distribution of correct and incorrect predictions relative to the actual dwell times of containers on a typical operational day. This visualization highlights the dwell-time ranges where models perform best and those where misclassifications are concentrated.

Two strategic findings emerge from this analysis:

- **Short-stay containers:** Models perform best at the extremes of operational behavior. For containers with dwell times shorter than two days, precision exceeds 58%, representing a direct opportunity to assign these containers to easily accessible yard positions upon arrival.
- **Long-stay containers:** For containers staying more than seven days, the model correctly classifies approximately 42% of cases, enabling early allocation to low-priority yard areas where prompt retrieval is not required.

These two groups—rapid-turnaround and long-stay containers—represent the greatest operational gains, as accurate identification significantly reduces unnecessary container movements.

On the other hand, **intermediate dwell times** (particularly between days 3 and 6) reveal areas for improvement. Precision declines and errors cluster in this range, likely due to greater operational variability and the heterogeneous factors influencing mid-range dwell durations. These results suggest that incorporating additional variables or modeling approaches may be necessary to better capture these intermediate behaviors.

## 7 Discussion

### 7.1 Interpretation of Results in Context

The predictive performance observed in this study aligns with and extends findings from prior research on container dwell-time prediction. [Kourouniotti et al. \[2016\]](#) reported that ANNs could effectively classify containers by dwell time using port of origin, container characteristics, and temporal features. Our results confirm these determinants while demonstrating that ensemble methods—particularly Random Forest—achieve superior performance when combined with consignee-level behavioral features and rigorous temporal validation.

The finding that extreme dwell-time categories (less than two days and more than seven days) are more predictable than intermediate durations reflects a pattern consistent with operational reality. Short-stay containers typically correspond to established logistics chains with reliable pickup schedules, while long-stay containers often involve pre-clearance holds or storage arrangements that follow identifiable patterns. Intermediate durations, by contrast, are subject to greater variability from factors outside the terminal’s visibility—such as consignee decisions, documentation delays, or transportation availability.

### 7.2 Comparison with existing approaches

The comparison against operational baselines rather than random classifiers distinguishes this study from much of the existing literature. While [Yoon et al. \[2023\]](#) and [Gaete G. et al. \[2017\]](#) demonstrated improvements over dummy classifiers, our evaluation against the terminal’s actual decision heuristics provides a more operationally meaningful benchmark. The 45 percentage point improvement in service prediction over combined baseline-plus-simple-ML approaches represents a substantive gain with direct implications for daily operations.

While container dwell time, yard congestion, and even customs-related predictions have received considerable attention in the literature, the specific problem of predicting which containers will require pre-clearance service from the terminal’s operational perspective has received limited attention. This gap is consequential: given that 51% of unproductive moves in this terminal are associated with service-requiring containers, accurate prediction of this label represents a direct lever for reducing operational inefficiency. Our model addresses this gap by learning from the terminal’s historical operational data to anticipate service requirements in advance, enabling yard planners to incorporate these predictions into their decisions before containers are positioned in the yard.

Our treatment of cargo descriptions also relates to recent work by [Xie et al. \[2025\]](#), who incorporated unstructured cargo text as features for predicting container exit terminals. While their approach used raw text processing techniques, we adopted a structured alternative: mapping free-text cargo descriptions to standardized HS (Harmonized System) chapters and sections through classification and record linkage. This structured representation reduces dimensionality while preserving semantically meaningful cargo categories, and enables the construction of interpretable behavioral features (e.g., historical service rates by cargo chapter) that would be difficult to derive from raw text embeddings.

### 7.3 Methodological Considerations

The use of temporal cross-validation [see [Roberts et al. \[2017\]](#)] addresses a common methodological pitfall in port logistics studies. Many predictive studies employ conventional cross-validation, which can lead to data leakage when temporal dependencies exist in the data. By training exclusively on past data and evaluating on future periods, our approach ensures that reported performance metrics reflect realistic deployment conditions.

The decision rule combining eight binary classifiers through ROC-derived thresholds and score ranking offers a practical solution to the multi-class dwell-time prediction problem. While a single multi-class classifier might appear more elegant, the binary decomposition allows for label-specific optimization and provides interpretable confidence measures for each prediction. The Jaccard similarity analysis (see Appendix [B](#)) confirms that different labels identify largely distinct container subsets, validating this approach.

### 7.4 Limitations

Several limitations should be acknowledged. First, the models rely on information available at the time of vessel arrival, excluding factors that may emerge during the container’s stay—such as documentation issues or consignee requests for extended storage.

Second, this study was conducted at a single terminal in Mexico. While the methodology is transferable, the specific feature importance rankings and optimal hyperparameters may vary across terminals with different operational characteristics, cargo mixes, or regulatory contexts. Notably, the two prediction tasks may exhibit different generalization

patterns. Service prediction depends on pre-clearance processes and operational practices that vary across contexts, suggesting that models trained in one setting may require significant recalibration elsewhere. Dwell-time prediction, by contrast, captures logistics chain dynamics—short-stay containers reflecting efficient pickup operations and long-stay containers reflecting storage arrangements or holds—that likely share structural similarities across terminals, potentially enabling better cross-terminal transfer of learned patterns for extreme dwell-time categories.

Finally, this study validates predictive performance rather than directly measuring operational outcomes such as reshuffle reduction. Translating predictive accuracy into yard efficiency gains requires operational deployment or simulation studies with stacking position and retrieval sequence data not available in the current study.

## 7.5 Indicative Estimates of Potential Reshuffle Reduction

The potential operational impact of the predictive models can be estimated through a framework that connects prediction quality to reshuffle reduction.

### 7.5.1 Aggregate Impact Estimate

Terminal statistics indicate that 75% of container moves are classified as unproductive, and 51% of these involve containers requiring service. Let  $\alpha = 0.75$  denote the unproductive move ratio and  $\beta = 0.51$  the fraction attributable to service-requiring containers. Let  $\delta_s$  represent the fraction of service-related reshuffles that could be avoided through prediction-informed placement, and  $\delta_d$  the corresponding fraction for dwell-time-informed stacking of non-service containers.

The reduction in total handling operations is bounded by:

$$\Delta_{\text{total}} = \alpha \cdot \beta \cdot \delta_s + \alpha \cdot (1 - \beta) \cdot \delta_d \quad (1)$$

For service prediction, observed gains of approximately 50 percentage points in recall over the operational baseline suggest  $\delta_s \in [0.25, 0.40]$  under reasonable stacking policies. For dwell-time prediction, the strong performance at temporal extremes—precision exceeding 58% for short-stay and 42% for long-stay containers, which together represent approximately 36% of all containers—suggests  $\delta_d \in [0.10, 0.20]$ . Under these assumptions, the estimated reduction in total handling operations lies between **13% and 23%**, depending on stacking policy sophistication.

### 7.5.2 Disaggregated Analysis by Container Category

The impact is not uniformly distributed across container categories:

**Service containers** (33% of arrivals, 51% of unproductive moves) have outsized operational impact because service events create *discontinuities* in the container trajectory—the container must be extracted and moved to the service area regardless of its yard position. The model’s recall of up to 100% at  $k = 300$  suggests that under sufficient service area capacity, nearly all service-related reshuffles could be addressed through prediction-informed placement.

**Short-stay containers** (<2 days, ~8% of arrivals) have the highest per-unit reshuffle potential because they depart before most containers stacked around them. With model precision exceeding 58%, prediction-informed placement in accessible positions directly eliminates these high-cost reshuffles.

**Long-stay containers** (>7 days, ~28% of arrivals) present the inverse opportunity: correctly identified long-stay containers can be placed in deep storage positions where they will not obstruct earlier departures.

**Intermediate-stay containers** (2–7 days, ~64% of arrivals) contribute less marginal value per prediction due to lower prediction quality and smaller operational cost of adjacent-day misclassification. Nevertheless, even coarse classification provides stacking information superior to random assignment.

### 7.5.3 Stacking Policy Sensitivity and Asymmetric Costs

The magnitude of reshuffle reduction depends on the stacking policy that consumes the predictions. Simple zone segregation—routing containers to different yard areas by predicted category—requires minimal system integration and captures 10–15% reduction in total moves. More sophisticated within-block ordering, placing predicted earlier-departure containers atop later-departure containers, yields 15–25% reduction but requires integration with yard management systems.

Prediction errors carry asymmetric operational costs. Misclassifying a service container as non-service (false negative) incurs high cost: the container is buried in a standard block and must be excavated for service. The reverse error merely wastes premium service-area space. Similarly, misclassifying a short-stay container as long-stay generates multiple reshuffles at early departure, whereas the inverse error wastes accessible positions but causes no cascade. This asymmetry suggests that for service prediction, higher recall should be prioritized over higher precision, consistent with the operational capacity constraint ( $k = 300$ ) employed in this study.

#### 7.5.4 Validation Framework

While the analytical estimates above provide indicative bounds, definitive quantification requires simulation using historical operational data. The movements dataset (13 million records) enables reconstruction of actual stacking configurations and retrieval reshuffles. A counterfactual simulation comparing historical placements against prediction-informed placements—under physical constraints and using the model’s actual predictions including errors—would quantify achievable reshuffle reduction. Such simulation, together with perfect-information upper bounds establishing theoretical maxima, is the subject of ongoing work.

#### 7.6 Theoretical Contributions

This study contributes to the growing literature on prediction-optimization integration in port logistics [Jahangard et al., 2025]. By demonstrating that predictive outputs can meaningfully inform stacking decisions, we provide empirical support for data-driven approaches to yard management. The combination of service prediction and dwell-time estimation within a unified framework offers a more comprehensive decision-support capability than either prediction alone.

### 8 Conclusion

This study developed and validated a machine learning framework for predicting container service requirements and dwell times at a Mexican port terminal. The framework substantially outperforms current operational heuristics: service prediction achieved double the precision and recall of baseline approaches, while dwell-time prediction demonstrated strong accuracy for operationally critical categories—short-stay containers requiring accessible placement and long-stay containers suitable for storage areas.

The methodology employs rigorous temporal cross-validation, ensuring that reported metrics reflect realistic deployment conditions. The combination of service and dwell-time predictions within a unified framework provides comprehensive decision support for yard planners, enabling informed stacking decisions at the time of container arrival.

While predictive performance is validated, translating these gains into measured reshuffle reductions requires operational deployment or simulation studies. Several directions merit further investigation. First, operational validation through simulation studies would strengthen the connection between predictive accuracy and yard efficiency. Using historical retrieval sequences, a backtest could compare reshuffle counts under current stacking policies versus dwell-informed placement rules [cf. Gaete G. et al., 2017]. Such analysis requires detailed stacking position and retrieval sequence data, which cannot be fully reconstructed from the available movement records in the current study, but would provide direct quantification of move reductions achievable through prediction-informed placement.

Second, the model could be extended to incorporate real-time information updates as containers progress through their dwell period, potentially improving predictions for intermediate dwell-time categories. Third, integration with vessel scheduling systems could enable proactive yard reorganization before peak retrieval periods. Finally, extending the framework to other terminal operations—such as equipment allocation and gate scheduling—represents a natural progression toward comprehensive data-driven terminal management.

### Acknowledgments

The authors thank Zaid Hernández Solano, Carlos Eduardo Olvera Azuara, and Roberto Villarreal Ramírez for their contributions to data processing and analysis during some stages of this project.

### Author Contributions

**Elena Villalobos:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Adolfo De Unánue:** Conceptualization, Methodology,

Supervision, Software, Formal analysis, Investigation, Writing – Review & Editing. **Fernanda Sobrino**: Investigation, Writing – Review & Editing. **David Aké**: Project administration. **Stephany Cisneros**: Investigation. **Jorge Lecona**: Resources, Project administration, Writing – Review & Editing. **Alejandra Matadamaz**: Domain expertise, Writing – Review & Editing.

### Declaration of Competing Interests

Jorge Lecona and Alejandra Matadamaz are employees of the container terminal where this study was conducted. The study design, analysis, and interpretation were conducted independently by the academic authors.

### A Table of evaluated algorithms and hyperparameters.

Table 5: Summary of evaluated algorithms and hyperparameter configurations.

Algorithm	Hyperparameter	Values tested
DecisionTreeClassifier	criterion	gini
	max_depth	5, 10, 50
	min_samples_split	10, 50, 100
RandomForestClassifier	n_estimators	200, 300
	criterion	gini
	max_depth	5, 10
	max_features	sqrt
	min_samples_split	10, 50
ScaledLogisticRegression	penalty	11, 12
	C	0.0001, 0.01, 0.1, 1.0
ExtraTreesClassifier	n_estimators	500
	criterion	gini
	max_depth	5, 10
	max_features	sqrt
	min_samples_split	50, 100

### B Similarity Analysis Across Dwell Time Labels

For each container, the model generates predictions across the eight possible dwell-time labels. To determine whether models prioritize the same containers across different labels or identify distinct subsets, we computed the pairwise similarity between the sets of containers prioritized by each label using the Jaccard similarity coefficient, which measures the proportion of shared elements between two sets. A value close to 1 indicates high overlap (the same containers appear in both labels), whereas a value close to 0 indicates distinct subsets.

Figure 9 shows the Jaccard similarity matrix across all combinations of dwell-time labels. Higher similarity values are concentrated between adjacent dwell-time categories—e.g., between the two-day label ( $\tau e2$ ) and the three-day label ( $\tau e3$ ) (0.38), or between the less-than-two-days label ( $\tau e1$ ) and the two-day label ( $\tau e2$ ) (0.35)—indicating partial overlap in prioritized containers for these models. However, for most other combinations, similarity is substantially lower, suggesting that the models identify largely distinct container subsets. This provides evidence that the models are learning differentiated and label-specific patterns rather than replicating identical predictions across multiple labels.

### References

UNCTAD. *Review of Maritime Transport 2025: Staying the Course in Turbulent Waters*. United Nations Publications, Geneva, 2025. ISBN 978-92-1-113096-5. URL <https://unctad.org/publication/review-maritime-transport-2025>.

Amir Gharehgozli, Joan P. Mileski, and Okan Duru. Heuristic estimation of container stacking and reshuffling operations under the containership delay factor and mega-ship challenge. *Maritime Policy & Management*, 44(3):373–391, April 2017. ISSN 0308-8839, 1464-5254. doi:[10.1080/03088839.2017.1295328](https://doi.org/10.1080/03088839.2017.1295328).

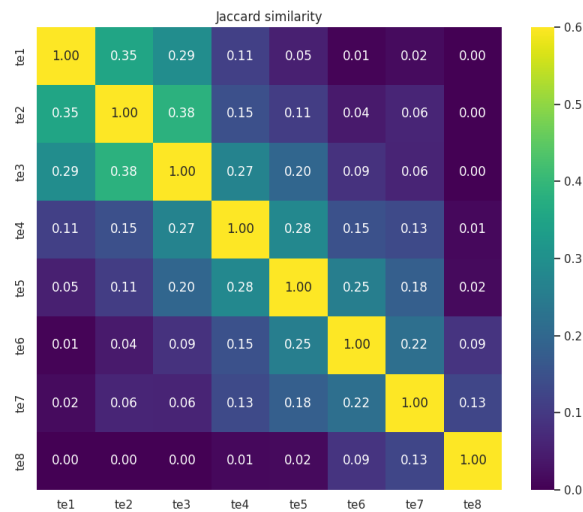


Figure 9: Jaccard similarity matrix across dwell-time labels.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine Learning for Combinatorial Optimization: A Methodological Tour d’Horizon. *arXiv:1811.06128 [cs, stat]*, November 2018.

Ioanna Kourounioti, Amalia Polydoropoulou, and Christos Tsiklidis. Development of Models Predicting Dwell Time of Import Containers in Port Container Terminals – An Artificial Neural Networks Application. *Transportation Research Procedia*, 14:243–252, 2016. ISSN 23521465. doi:[10.1016/j.trpro.2016.05.061](https://doi.org/10.1016/j.trpro.2016.05.061).

Jeong-Hyun Yoon, Se-Won Kim, Ji-Sung Jo, and Ju-Mi Park. A Comparative Study of Machine Learning Models for Predicting Vessel Dwell Time Estimation at a Terminal in the Busan New Port. *Journal of Marine Science and Engineering*, 11(10):1846, September 2023. ISSN 2077-1312. doi:[10.3390/jmse11101846](https://doi.org/10.3390/jmse11101846).

Mohan Saini and Tone Lerher. ASSESSING THE FACTORS IMPACTING SHIPPING CONTAINER DWELL TIME: A MULTI-PORT OPTIMIZATION STUDY. *Business: Theory and Practice*, 25(1):51–60, February 2024. ISSN 1648-0627, 1822-4202. doi:[10.3846/btp.2024.19205](https://doi.org/10.3846/btp.2024.19205).

Yongjae Lee, Kikun Park, Hyunjae Lee, Jongpyo Son, Seonhwan Kim, and Hyerim Bae. Identifying key factors influencing import container dwell time using eXplainable Artificial Intelligence. *Maritime Transport Research*, 7: 100116, December 2024. ISSN 2666822X. doi:[10.1016/j.martra.2024.100116](https://doi.org/10.1016/j.martra.2024.100116).

Kap Hwan Kim. Evaluation of the number of rehandles in container yards. *Computers & Industrial Engineering*, 32(4): 701–711, September 1997. ISSN 03608352. doi:[10.1016/S0360-8352\(97\)00024-7](https://doi.org/10.1016/S0360-8352(97)00024-7).

Razouk Chafik, Y. Benadada, and J. Boukachour. Stacking policy for solving the container stacking problem at a containers terminal. 2016.

Marco Caserta, Stefan Voß, and Moshe Sniedovich. Applying the corridor method to a blocks relocation problem. *OR Spectrum*, 33(4):915–929, October 2011. ISSN 0171-6468, 1436-6304. doi:[10.1007/s00291-009-0176-5](https://doi.org/10.1007/s00291-009-0176-5).

Bram Borgman, Eelco Van Asperen, and Rommert Dekker. Online rules for container stacking. *OR Spectrum*, 32(3): 687–716, July 2010. ISSN 0171-6468, 1436-6304. doi:[10.1007/s00291-010-0205-4](https://doi.org/10.1007/s00291-010-0205-4).

Myriam Gaete G., Marcela C. González-Araya, Rosa G. González-Ramírez, and César Astudillo H. A Dwell Time-based Container Positioning Decision Support System at a Port Terminal:. In *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems*, pages 128–139, Porto, Portugal, 2017. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-218-9. doi:[10.5220/0006193001280139](https://doi.org/10.5220/0006193001280139).

Mahdi Jahangard, Ying Xie, and Yuanjun Feng. Leveraging machine learning and optimization models for enhanced seaport efficiency. *Maritime Economics & Logistics*, February 2025. ISSN 1479-2931, 1479-294X. doi:[10.1057/s41278-024-00309-w](https://doi.org/10.1057/s41278-024-00309-w).

Leonard Heilig, Robert Stahlbock, and Stefan Voß. From Digitalization to Data-Driven Decision Making in Container Terminals, April 2019.

- Ying Xie, Dong-Ping Song, Jingxin Dong, and Yuanjun Feng. Predicting out-terminals for imported containers at seaports using machine learning: Incorporating unstructured data and measuring operational costs due to misclassifications. *Transportation Research Part E: Logistics and Transportation Review*, 202:104331, October 2025. ISSN 13665545. doi:[10.1016/j.tre.2025.104331](https://doi.org/10.1016/j.tre.2025.104331).
- Sunny Md. Saber, Kya Zaw Thowai, Muhammad Asifur Rahman, Md. Mehedi Hassan, A.B.M. Mainul Bari, and Asif Raihan. High-accuracy prediction of vessels' estimated time of arrival in seaports: A hybrid machine learning approach. *Maritime Transport Research*, 8:100133, June 2025. ISSN 2666822X. doi:[10.1016/j.martra.2025.100133](https://doi.org/10.1016/j.martra.2025.100133).
- Russell Hillberry, Bilgehan Karabay, and Shawn W. Tan. Risk management in border inspection. *Journal of Development Economics*, 154:102748, January 2022. ISSN 0304-3878. doi:[10.1016/j.jdeveco.2021.102748](https://doi.org/10.1016/j.jdeveco.2021.102748).
- Sruti Vijayakumar. Technology-centric and Data-Driven Customs Risk Management for Supply Chain Security. *World Customs Journal*, 19(1):38–63, April 2025. doi:[10.55596/001c.131745](https://doi.org/10.55596/001c.131745).
- Perspective on risk management systems for Customs administrations. <https://mag.wcoomd.org/magazine/wco-news-90/perspective-risk-management-systems/>, 2020.
- David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, August 2017. ISSN 09067590. doi:[10.1111/ecog.02881](https://doi.org/10.1111/ecog.02881).
- Rayid Ghani, Joe Walsh, and Joan Wang. Top 10 ways your Machine Learning models may have leakage. <https://www.dssgfellowship.org/2020/01/23/top-10-ways-your-machine-learning-models-may-have-leakage/>, 2020.
- Center For Data Science and Public Policy. Triage: General Purpose Risk Modeling and Prediction Toolkit for Policy and Social Good Problems, November 2025.
- Rayid Ghani. Triage: ML/Data Science Toolkit for Social Good and Public Policy Problems. <https://www.datasciencepublicpolicy.org/our-work/tools-guides/triage/>, 2024.
- Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi:[10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013.