Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(11), 72-103
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

When the Answer Is Constructed: A Target-Provenance Audit of Machine-Learning Benchmarking and Explainable AI for Logistics Loading-Bay Provision

Shakir Adeyemi Tewogbade[1], Linda Oraegbunam[2], Rebecca Omowaiye[3], Badmus Ayodeji[4]

[1]Independent Researcher, Bradford, UK
[2]Independent Researcher, Leeds, USA
[3]Independent Researcher, Huddersfield, UK
[4]Independent Researcher, North Carolina, USA
Email: sesismic305@gmail.com

How to cite this paper: Tewogbade, S.A., Oraegbunam, L., Omowaiye, R. and Ayodeji, B. (2026) When the Answer Is Constructed: A Target-Provenance Audit of Machine-Learning Benchmarking and Explainable AI for Logistics Loading-Bay Provision. *Journal of Computer and Communications*, **14**, 72-103.
https://doi.org/10.4236/jcc.2026.1411004

Received: October 7, 2026
Accepted: November 8, 2026
Published: November 11, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Predictive accuracy is informative for infrastructure planning only when the target represents the quantity decision-makers intend to assess. We audit this requirement using an open inventory of 47,682 logistics sites in England and Wales whose heavy-goods-vehicle loading-bay counts were modelled from floor area and facility type. A reconstructed deterministic rule matches 47,678 records (99.992%). We reproduce an initial five-model prediction pipeline and examine feature ablation, spatially blocked validation, attribution, and residual-based flags. The reproduced pipeline achieves $R^2 = 0.9871$ and up to 98.7% of predictions within one bay. The decision tree using only floor area and category achieves $R^2 = 0.99985$. Spatial-regional features alone perform poorly under blocked validation, and four engineered features directly or indirectly incorporate the target. In the deduplicated audit configuration, features outside the reconstructed rule receive up to 33.7% of mean absolute SHAP attribution. This measures departure from the rule, not necessarily an error by the explainer. Out of the 685 baseline flags, 684 match the reconstructed rule exactly, while cross-model flag agreement is low (mean Jaccard 0.25). For simulated one-bay reductions at eligible sites, Random Forest recall is 60.0% with the full feature set and 98.8% with rule inputs alone. These perturbations assess detection of constructed deviations, not actual operational inadequacy. We propose a Target Provenance and Validity Protocol that separates target validity, predictive performance, and decision validity before model outputs are used in planning.

Keywords

Loading Bays, Freight Infrastructure Planning, Machine-Learning Benchmarking, Explainable AI, Target Validity, Target Leakage, Data Provenance, Spatial Validation

1. Introduction

Loading bays connect freight vehicles with the facilities they serve. Inadequate loading arrangements have a tendency to contribute to vehicle waiting. Also, constrained kerbside access can increase parking-search time, double parking, and disruption to other road users [1]-[4]. Clearly, studies involving kerbside loading show how access to loading space affects delivery performance. However, determining how many loading bays a warehouse or other logistics facility requires also depends on its operational demand regarding vehicle arrivals, loading times, and operating hours. In the United Kingdom, the Future of Freight strategy identifies freight as a strategic policy concern [5]. For facility-level assessment, credible benchmarks must account for site characteristics and loading demand.

Predictive modelling offers one possible source of comparative evidence. Open geocoded inventories provide information on logistics-site locations, sizes, and categories [6]. From a computational angle, studies in transport demonstrate the potential of tree ensembles for structured prediction tasks [7]. A model trained on independently observed provision has capabilities to estimate a conditional peer norm for a facility of a given size and type. Sites below that norm need to be selected for further investigation. Such a norm would describe prevailing provision, as it would not by itself establish adequate capacity. Post hoc methods like SHAP can explain how a fitted model uses its inputs [8] [9], supporting the scrutiny of operational decision tools [10] [11]. Critically, neither prediction nor explanation, however, validates the underlying measurement.

This study examines that measurement precondition. If recorded bay counts were produced by computational means rather than observed at each site, a model trained on those counts will mainly approximate the computational method. Its explanations describe the fitted surrogate, and its residuals compare two model outputs. Another problem that can be triggered is when the target signal enters the features or model-selection process, or when validation does not match the intended deployment setting [12]-[14]. The major concerns here are clear. They include how removing feature leakage or improving validation does not convert a generated target into an observed outcome. High predictive accuracy and plausible explanations can consequently coexist with an invalid interpretation of what the target measures.

We investigate this problem using an open inventory of logistics sites in England and Wales [6] [15]. The documentation of this inventory clearly identifies the bay field as modelled heavy-goods-vehicle (HGV) loading-bay provision.

Floor area and category were used to extrapolate from satellite-image counts for a small sample. The field is not an independently observed count for every site. This transparency makes the inventory useful for a methodological audit. The documented construction, supplemented by reconstruction of the recorded values, supplies a reference for target generation. It does not supply ground truth for actual provision or operational adequacy.

The first author initially implemented a reference prediction-and-benchmarking pipeline, named B0. The pipeline compares Poisson and negative binomial regressions with Random Forest, XGBoost, and LightGBM. It adds spatial and target-derived features, uses SHAP, and then flags positive prediction residuals as candidate under-provision. The audit pipeline was implemented to reproduce the feature specification and tree-model results in the initial pipeline B0. Also, the features, validation, and model configuration were varied based on the objectives of this study. The reproduced pipeline reaches $R^2 = 0.9871$, up to 98.7% of predictions within one bay, and 685 residual flags. These figures motivate the main question: does benchmarking establish evidence about actual loading-bay provision when the recorded target was constructed from the available predictors? Based on the question and other claims from the introduction, the aim of this study is to distinguish rule approximation from substantive evidence and to formulate checks that should precede the planning use of such benchmarks.

To achieve this aim, the study pursues three objectives:

1) To establish the construction of the recorded bay field and assess how predictive performance changes across feature families and validation designs.

2) To assess how attribution and prediction sensitivity vary for different model configurations, distinguishing fidelity to the fitted model from fidelity to the reconstructed target rule.

3) To assess rule compliance and the stability of flagged sites, and to measure the detection of simulated downward deviations from the rule at eligible sites.

The study makes four important contributions. First, it reconstructs a deterministic approximation to the target-generating rule that matches all but four of 47,682 records. Second, it contrasts rule-input models with spatial and target-derived feature sets, separating predictive fit from the validity of the quantity being predicted. Third, it evaluates feature attribution against a known construction mechanism on real site covariates, while recognizing that attribution to a feature outside that mechanism can faithfully describe the fitted model. Fourth, it examines residual-flag stability and detection of controlled perturbations. The entire analysis carried out in the study and the findings motivate the proposition of the eight-step Target Provenance and Validity Protocol (TPVP).

2. Background and Related Work

2.1. Loading Infrastructure as a Freight-Planning Problem

Recent urban-freight research shows that loading-space allocation and parking time influence delivery operations [16] [17]. Research involving driver behaviour

identifies trade-offs between actions like walking distance, parking-search time, cost, and enforcement considerations [4] [18]. Ghizzawi *et al.* [3] synthesise the literature around parking location, duration, inappropriate parking, and cruising. Modelling studies examine the implications of kerbside access and delivery strategies for traffic performance [2] [19], while facility-level evidence identifies operational and shipper characteristics that are associated with driver detention [1]. Clearly, provision is more than a count. It involves location, management, allocation, and competition for space [16] [20] [21].

Considering an individual logistics facility, adequacy concerns whether loading capacity meets operational demand. Floor area may be a useful proxy, but throughput, vehicle mix, operating hours, fulfillment arrangements, and network role influence the requirement for loading positions [22]. Reviews of freight planning and data infrastructure highlight the importance of better information for such assessments [23]-[25]. Open inventories can help address these information needs, provided analysts preserve the distinction between administrative attributes, estimated quantities, and independently verified infrastructure.

2.2. Statistical and Machine-Learning Approaches to Infrastructure Counts

Loading-bay counts are non-negative integers. Based on their empirical distribution, candidate models include Poisson regression, negative binomial regression, and zero-inflated or hurdle specifications [26] [27]. Transport studies illustrate the relevance of spatial dependence, measurement periods, and excess zeros to count modelling [28] [29]. These considerations concern model specification but require appropriate verification of what the count represents. A threshold-generated zero count does not establish a distinct behavioural process at logistics sites. Similarly, a large marginal variance-to-mean ratio does not establish conditional overdispersion.

Based on computational capabilities, machine learning models like Random Forests [30], XGBoost [31], and LightGBM [32] can accommodate nonlinear relationships and interactions without specifying a count-distribution likelihood. Their comparative performance depends on the data, tuning, and evaluation design. Methodological benchmarking also requires transparent design choices and sensitivity analysis [33]. Geographic features like density, DBSCAN clustering [34], and Local Moran statistics [35] must be assessed according to their information source and targeted prediction setting.

2.3. Leakage, Validation, and the Reproducibility of ML-Based Findings

Fundamentally, leakage occurs when a model uses information unavailable in the targeted prediction setting or when held-out outcomes influence model development [12] [14]. Direct outcome-derived predictors and test-set model selection are clear examples. Spatial dependence presents a related validation challenge.

Random splits may overstate performance for deployment in geographically distinct areas, making spatially blocked evaluation useful when that deployment is the objective [13] [36] [37]. Unsupervised preprocessing on a full inventory is a separate issue. It can be a better option for a fixed-inventory, transductive task but does not establish performance for previously unseen locations. A generated target introduces another problem. Prediction may recover the generating model rather than the intended real-world construct. Work on generated variables and imputation addresses uncertainty and inferential compatibility [38] [39], while data-centric research emphasizes provenance, labels, and data quality [40]-[42]. Shortcut learning offers a complementary account of high accuracy that comes from an unintended input-label relationship [36].

2.4. Explainable AI and the Problem of Faithfulness

Explainable AI supports scrutiny of models used in operational decisions [10] [11]. SHAP assigns feature contributions using a Shapley-value formulation, and TreeSHAP provides efficient computation for tree models [8] [9]. Notably, interpretation depends on the value function and treatment of feature dependence [43]-[45]. Correlated predictors can redistribute attribution, and unconditional permutation can produce unrealistic input combinations [46] [47]. Works from Slack *et al.* [48] and Adebayo *et al.* [49] demonstrate manipulation of explanations and the need for sanity checks. Rudin [50] advocates interpretable models for high-stakes decisions. From all claims, three properties should be separated. These include an explanation's faithfulness to the fitted model, the model's fidelity to the target-generating mechanism, and the target's validity for the domain construct. An explanation can satisfy the first while the other two are unsupported.

A documented deterministic target offers a useful testbed for this distinction. Its construction can be known even when actual real-world outcomes are not. These kinds of data combine realistic covariate distributions and geographic structure with an auditable label-generating rule. They can reveal which fitted dependencies extend beyond that rule, but they cannot establish the causal determinants of actual infrastructure or validate a planning standard.

2.5. Benchmarking as Decision Support

Benchmarking compares a unit with a defined reference and analyzes any divergence from the reference [51] [52]. A model-based peer benchmark estimates provision conditional on selected characteristics. Then, residuals identify units below or above that conditional norm. The interpretation depends on both the provenance of the outcome and the relevance of the comparison. For facility planning, even a benchmark trained on observed provision describes prevailing practice unless linked to independent evidence about operational adequacy. Data-driven loading-zone design provides related decision-support applications, but a network-design model should not be equated with a validated facility-level adequacy standard [16].

Two checks are very important in this discourse. First, analysts must distinguish deviations from observed peer provision, deviations from a constructed engineering reference, and evidence of inadequate loading capacity. These are different decision quantities. Second, flag lists depend on features, model family, tuning, rounding, and thresholds. Stability and sensitivity analyses are therefore needed before residuals are used to prioritise investigation. Controlled perturbations can assess whether a pipeline detects known departures from its reference [33].

2.6. Research Gap

The importance of data provenance, leakage control, appropriate validation, and careful interpretation of feature attribution is critically established in the literature. The unresolved question is how these checks interact when a planning benchmark is trained on a documented, nearly deterministic generated target. We examine three connected issues. The first issue is whether predictive success reflects the recovery of the target rule. The second is whether attribution and prediction sensitivity track that rule. The last is whether residual flags are stable and detect constructed downward deviations. The contribution lies in evaluating these issues together on a logistics-site inventory and translating the findings into a practical audit protocol.

3. Data and Target Provenance

3.1. The Logistics-Site Inventory

The data selected for this study are openly licensed logistics-site inventory compiled by de Saxe *et al.* [15]. Administrative property records were processed to obtain site locations, floor areas, and categories, while HGV loading-bay counts were subsequently modelled. The analysis file contains category, latitude, longitude, floor area in square metres (sqm), and recorded modelled bay count (bays). The audit code coerces numeric fields, excludes missing critical values, applies bounds of 100 - 500,000 m^2 and 0 - 200 bays, and restricts coordinates to 49˚ - 61˚N and −8˚ - 2˚ longitude. Only one of 47,683 records is excluded because latitude cannot be parsed, leaving 47,682 sites: 27,690 warehouses, 13,551 factories, and 6441 retail premises. The bounds are screening rules, not validated physical limits. Table 1 summarises the audit, and Figure 1 presents the geographic and marginal distributions.

Table 1. Data audit of the logistics-site inventory.

Property	Value	Implication
Records (raw/analyzed)	47,683/47,682	Only latitude coercion excludes a record; other screening bounds exclude none.
Categories	Warehouse 27,690; factory 13,551; retail 6441	Three-way classification only
Floor area	1000 - 310,773 m^2 (median 2332 m^2)	Left-truncated at 1000 m^2; heavy right tail (skewness 9.23)
Loading bays	0 - 120; mean 2.60; variance 60.10	Marginal dispersion; not a conditional model diagnosis

Continued

Sites ≤ 2000 m²	43.06%	Matches zero-bay share after rounding (43.06%).
Exact duplicate records	100	Can fall on both sides of a random split.
Unique coordinates (six-decimal groupings)	25,734 (64.5% of sites share a coordinate; max. 41 per point)	Distance-based features and random splits are affected.
Latitude range	50.09˚ - 55.79˚N	England and Wales only; the "≥55˚N" band (387 sites) is in northern England.
Spatial blocks (20 km)	420 occupied blocks	Unit of blocked cross-validation

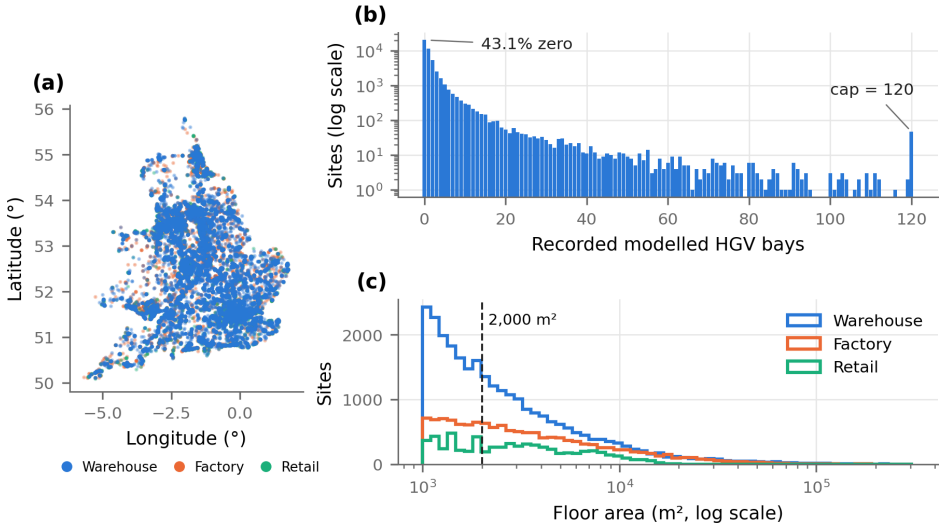


Figure 1. Data overview, regenerated from the source CSV. (a) Locations of 47,682 analyzed sites by category. (b) Distribution of recorded modeled HGV loading-bay counts on a logarithmic frequency scale; 43.1% record zero, and warehouses have a cap at 120. (c) Floor-area distributions by category on a logarithmic area scale; the dashed line marks 2000 m².

Four properties are necessary for interpretation. First, the minimum recorded floor area is 1000 m², so coverage does not represent all premises or all logistics activity. Second, coordinates rounded to six decimal places are shared by 64.49% of sites. The largest coordinate group contains 41 records. Co-location may reflect real shared sites or geocoding resolution, so these records should not automatically be treated as independent fine-scale locations. Third, the maximum latitude is 55.79˚N. The legacy feature labeled "scotland" for latitude ≥ 55˚N therefore denotes a northern latitude band within this England-and-Wales inventory. Fourth, the proportion at or below 2000 m² and the proportion recording zero bays both round to 43.06%. This distributional correspondence is a diagnostic clue.

3.2. Reconstructing the Data-Generating Process of the Target

de Saxe *et al.* [6] describe the estimation of HGV loading-bay counts from satellite imagery for 36 warehouses, 39 factories, and 22 retail premises. The documentation reports category-specific modelling, threshold assumptions, and upper limits. We reconstruct the operational mapping from the recorded values, distinguishing

this empirical reconstruction from the documented model description. For warehouses, the floor-area-per-bay parameter k is searched over 700 - 1000 m^2 in 0.25 m^2 increments. For factories, it is searched over 4000 - 9000 m^2 in 1 m^2 increments. For each category, candidate caps range from three below the recorded maximum to the maximum itself. Parameters maximise exact agreement on the complete analysed sample. The reconstructed mapping is:

$$bays_i = 0, \text{ if } sqm_i \le 2000;$$

$$bays_i = 1, \text{ if retail and } sqm_i > 2000;$$

$$bays_i = \min\{cap, 1 + \text{round}[(sqm_i - 2000)/k_c)\}, \text{ otherwise.} \quad (1)$$

Equation (1) is a threshold-anchored, rounded, and capped mapping. The recovered parameters $k = 823.5$ m^2 per additional warehouse bay, with cap 120, and $k = 6532$ m^2 per additional factory bay are presented in Table 2. They are shown with the scoring grid value at 6533 and cap 15. The equation reproduces 47,678 of 47,682 records (99.992%). The four discrepancies are two warehouses at exactly 2000 m^2 recording one bay, a warehouse at 32,058.49 m^2 recording 37 instead of 38 bays, and a factory at 96,742.68 m^2 recording 16 instead of the reconstructed cap of 15. These discrepancies reflect boundary, parameter, or implementation conventions. Figure 2 overlays the reconstructed mapping on the recorded target.

Table 2. Recovered imputation rule (Equation (1)) and exact-match rates.

Category	Sites	Threshold	Slope k (m^2 per bay)	Cap	Exact match
Warehouse	27,690	2000 m^2	823.5	120	99.989%
Factory	13,551	2000 m^2	6532 (6532 - 6533)	15	99.993%
Retail	6441	2000 m^2	— (one bay)	1	100.000%
All	47,682				99.992% (4 mismatches)

Note: Parameters maximize full-sample exact agreement over the stated grids and cap candidates. The factory interval denotes two equally scoring grid values, not a statistical confidence interval. NumPy rounding uses ties to the nearest even integer. The fitted rule is a retrospective reconstruction reference, not a model selected independently of held-out labels. The separate floor-area-and-category decision tree achieves $R^2 = 0.99985$ and 99.80% exact accuracy on a random 20% hold-out (unpruned scikit-learn DecisionTreeRegressor with default settings and random_state = 42; hold-out split seed 42; full specification in Section 3.2).

The reconstruction establishes that almost all variation in the recorded target is recoverable from floor area and category. A separate decision tree trained on those two original attributes achieves $R^2 = 0.99985$, exact-integer accuracy of 99.80%, and within-one-bay accuracy of 99.98% on a random 20% hold-out. The tree is a scikit-learn DecisionTreeRegressor with default settings and random_state = 42: squared-error criterion, best-split search, no maximum depth, a minimum of two samples to split a node, a minimum of one sample per leaf, all features considered at each split, and no cost-complexity pruning. Its two inputs are raw floor area in m^2 and category as an integer code (factory = 0, retail = 1, warehouse = 2), without

scaling. The hold-out is drawn with train_test_split (test_size = 0.2, random_state = 42), giving 38,145 training and 9537 test sites; this is the same partition as the baseline split. The fully grown tree has depth 9 and 131 leaves. Exact accuracy compares NumPy-rounded predictions with recorded counts, whereas within-one-bay accuracy uses unrounded predictions. The tree's accuracy confirms the ease of approximating the generated target, not the accuracy of the inventory against field counts. The concentration of zeros and large marginal variance also reflect the threshold, category mix, and floor-area distribution.

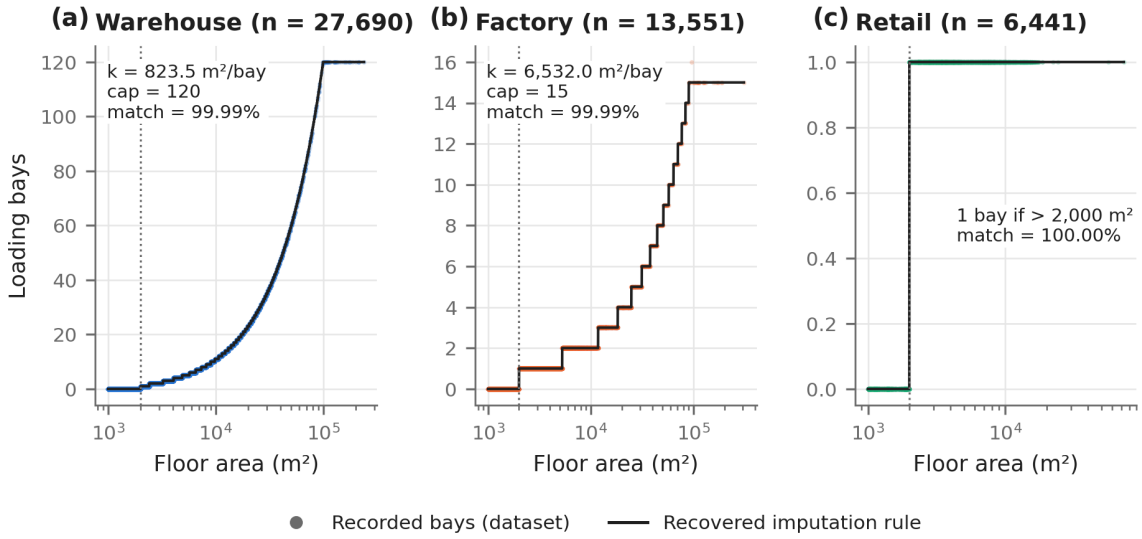


Figure 2. Recorded loading bays against floor area, by category, with the recovered imputation rule (Equation (1)) overlaid. Slopes k are expressed in m^2 of floor area per additional bay; "match" is the percentage of sites whose recorded count equals the rule's output exactly. The dotted line marks the 2000 m^2 threshold.

4. Audit Design

4.1. Overview

The audit pipeline is set up in four preparatory stages and three experiments (**Figure 3**). Stage 0 establishes provenance, while reconstruction of the recorded target is implemented in Stage 1. The baseline pipeline is reproduced in Stage 2. Stage 3 maps feature sources and roles. Predictive fit is examined in E1, while attribution and sensitivity relative to the rule are checked in E2. E3 examines residual flags and controlled perturbations.

4.2. The Baseline Specification (B0)

B0 is an author-developed pipeline, not an independently sourced workflow. The first author implemented it in the initial experimentation phase of this study, as a separate notebook (Loading Bay Prediction Pipeline) that ingests and cleans the inventory, engineers the spatial and target-derived features, fits the five models, computes TreeSHAP explanations, and flags under-provisioned sites from positive residuals. It follows a workflow commonly used to benchmark infrastructure

counts with tree ensembles, spatial features, and SHAP, but it was not taken from any published study or deployed planning application. The audit therefore demonstrates a methodological risk that such a workflow carries, using the authors' own initial pipeline as the case, rather than evaluating an existing application. The audit notebook reproduces the feature specification and main settings of that pipeline. Spatial and target-derived features are calculated on the full inventory before splitting. LightGBM uses the test partition for early stopping, and the reproduced baseline uses a random 80/20 split. These choices are retained deliberately so that the audit examines the pipeline as originally specified. Floor areas are used in their raw form and also as log1p(sqm), category is used as an ordinal code and two dummy indicators. Six unsupervised spatial features describe 5 km site density, DBSCAN membership and cluster size, and Gaussian kernel-density intensity. Five latitude bands are used for regional encodings. Four features are created for bay counts: Local Moran's I, its quadrant, its permutation-significance indicator, and the regional median. Category-by-latitude-band site counts and mean floor areas complete the inputs. The original 24-column matrix repeats each of the two category-region aggregates, leaving 22 unique features. B0 retains those duplicate columns. F1 deduplicates them. B0 and F1 are consequently related but distinct configurations.

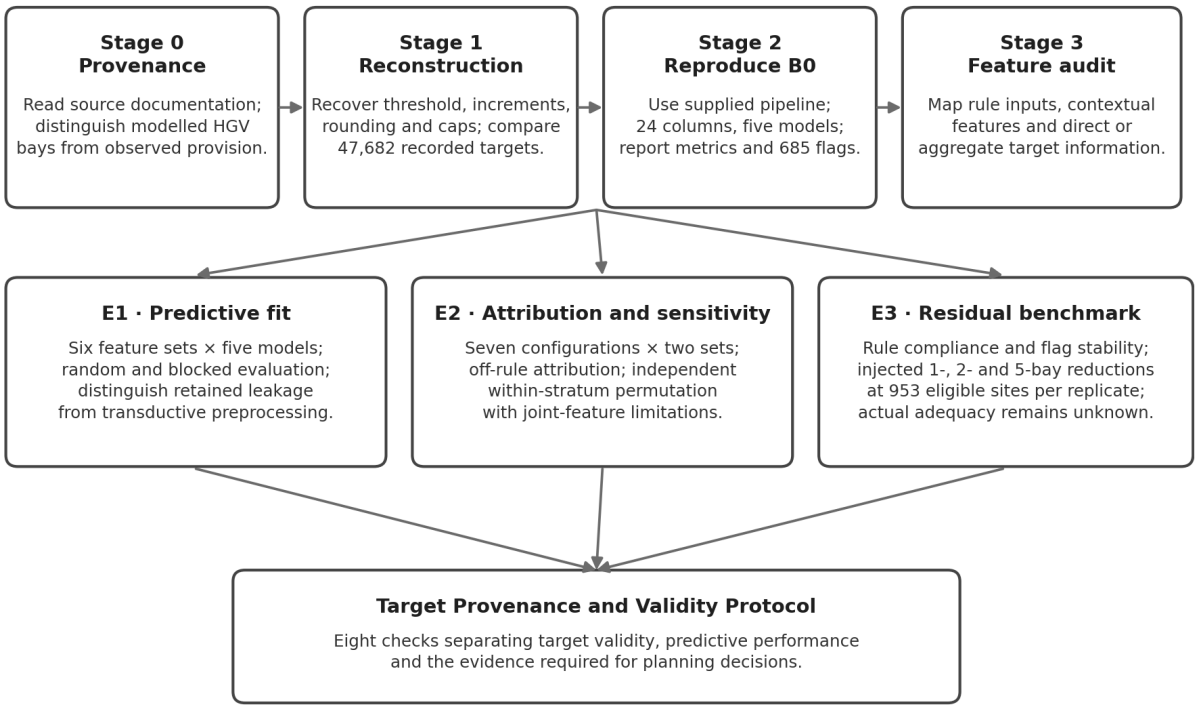


Figure 3. Revised audit framework. The preparatory stages establish documentary provenance, reconstruct the recorded rule, reproduce the author-built 24-column baseline (B0), and map feature sources. E1 - E3 examine predictive fit, rule-relative attribution and sensitivity, and residual-benchmark behaviour. Findings inform the Target Provenance and Validity Protocol.

The spatial and regional features are specified as follows. All distance-based fea-

tures use haversine distances with an Earth radius of 6371 km, computed on the full inventory with a ball-tree index. density_5 km counts other sites within 5 km of each site, excluding the site itself, and log_density_5 km is its log1p transform. DBSCAN uses a neighbourhood radius $\varepsilon = 3$ km (3/6371 radians on the haversine metric) and a minimum of 10 points per core neighbourhood; it identifies 450 clusters and labels 6.78% of sites as noise. in_cluster indicates non-noise membership, cluster_size is the number of sites in the site's cluster (zero for noise points), and log_cluster_size is its log1p transform. Kernel-density intensity uses SciPy's Gaussian KDE on longitude and latitude in decimal degrees with a scalar bandwidth factor of 0.05 applied to the sample covariance. This gives kernel standard deviations of approximately 0.060˚ of longitude and 0.054˚ of latitude (about 4.1 km east-west and 6.0 km north-south at 52.5˚N); the feature is log1p ($10^6 \times$ estimated density). Local Moran's I uses an eight-nearest-neighbour spatial weights matrix built on longitude-latitude coordinates in decimal degrees and row-standardised. Its significance indicator uses 99 conditional permutations (seed 42) and $p < 0.05$, and the quadrant follows the PySAL esda coding (1 = high-high, 2 = low-high, 3 = low-low, 4 = high-low). Because 64.5% of sites share a coordinate, ties among co-located neighbours are resolved by the libpysal nearest-neighbour implementation rather than by an explicit rule. The five latitude bands, with inclusive lower bounds, are ≥ 55˚N (labelled "Scotland"; 387 sites), 53.5˚ - 55˚N ("north_england"; 9870), 52.5˚ - 53.5˚N ("midlands"; 13,359), 51.5˚ - 52.5˚N ("south_east"; 14,682), and < 51.5˚N ("south_west"; 9384). These labels are legacy names for latitude ranges, not administrative regions, and "midlands" is the omitted reference category in the dummy encoding. The regional median and the two category-band aggregates are computed over these bands on the full inventory.

Table 3. Features in B0, grouped by source and role relative to the reconstructed target rule.

Family	Features	Role relative to the reconstructed rule
Floor area	sqm, log_sqm	Rule input
Category	category_enc, cat_retail, cat_warehouse	Rule input
Category × region proxy	cat_reg_count, cat_reg_mean_sqm (each entered twice in B0)	Contextual proxy; not an input of Equation (1)
Region	region_enc, reg_north_england, reg_scotland, reg_south_east, reg_south_west	Outside Equation (1)
Spatial (unsupervised)	density_5 km, log_density_5 km, in_cluster, cluster_size, log_cluster_size, log_kde_intensity	Outside Equation (1)
Target-derived	local_moran_I, moran_q, moran_sig, region_median_bays	Leakage (computed from the target on the full sample)

Note: Moran quadrant against the site's own bay count: quadrant 1 (n = 4395) and 4 (n = 5731) contain only sites with 3 - 120 bays; quadrants 2 (n = 7766) and 3 (n = 29,790) contain only sites with 0 - 2 bays.

Local Moran's I embeds the site's own label. For site i, it is proportional to $z_i \Sigma_j$

$w_{ij}\, z_j$, where z is the standardized bay count. The quadrant indicator also uses the sign of the site's centered outcome. In this dataset, it separates counts of 0 - 2 from counts of 3 - 120 exactly (Table 3). These inputs are inappropriate for predicting an unknown site count and can obscure anomalies when used for residual benchmarking. The full-inventory regional median also includes held-out outcomes, although its five-valued summary carries much less direct site information. The audit therefore distinguishes direct target encoding from aggregate target information instead of assuming all four features have equal effects.

The Poisson and negative binomial GLMs use an intercept, training-standardised features, and up to 300 iteratively reweighted least-squares iterations. The negative binomial dispersion parameter α is fixed at 1.0. Random Forest uses 300 trees, maximum depth 20, minimum leaf size 4, and max_features = $\sqrt{p}$. XGBoost uses 300 trees, depth 7, learning rate 0.05, row and column subsampling 0.8, γ = 0.1, α = 0.5, and λ = 1.0. LightGBM uses 300 trees, 63 leaves, learning rate 0.05, column subsampling 0.8, α = 0.3, and λ = 0.5. Its configured row-subsampling fraction is inactive with subsample_freq = 0. B0 early stopping has patience 30. The audit's GLM implementation clips predictions to 0 - 200. Both the baseline pipeline and the audit standardise all columns before fitting the GLMs. The fixed-α negative binomial fit is nevertheless sensitive to the software environment: identical code gave R^2 = 0.7631 in the baseline run and 0.7505 in the audit run, and XGBoost differs slightly for the same reason (0.9831 versus 0.9856). The reproduction therefore preserves the main tree results and flag count but is not numerically identical for every model.

4.3. Experiment 1: Accuracy Decomposition (O1)

E1 is in six feature sets (Table 4), five model families, and three validation designs. F0 contains the rule inputs and their encodings. F1 contains the 22 unique baseline features. F2 excludes the three Moran features, and F3 excludes the regional median. F4 isolates spatial and regional inputs, while F5 isolates target-derived inputs. This design assesses family-level predictive contributions.

The designs are a random 80/20 hold-out, shuffled five-fold cross-validation, and five-fold GroupKFold validation using 420 occupied spatial blocks. Blocks are approximately 20 km × 20 km. Longitude is converted using a fixed 52.5˚ reference latitude, and latitude is calculated using a spherical Earth approximation. Entire blocks are assigned to folds. This prevents co-located records from crossing a fold boundary. E1 LightGBM early stopping uses a random 10% subset of the training fold without refitting on all training observations. Other model settings are fixed. Geographic density, clustering, and category-region summaries are full-inventory features. F0 and F3 exclude target-derived predictors, but F3 is still a transductive, fixed-inventory evaluation. F1, F2, and F5 are set up to retain held-out target information. The full-sample reconstructed rule is scored on each fold only as a reconstruction reference. Metrics are R^2, MAE, RMSE, exact-integer accuracy, within-one-bay accuracy, and zero-bay accuracy.

Table 4. Feature sets used in Experiment 1 (E1).

Set	Description	Features included	No. of features
F0	Floor area and category only (rule-input encodings)	sqm, log_sqm, category_enc, cat_retail, cat_warehouse	5
F1	All unique B0 features	F0 features: density_5 km, log_density_5 km, in_cluster, cluster_size, log_cluster_size, log_kde_intensity; region_enc, reg_north_england, reg_scotland, reg_south_east, reg_south_west; cat_reg_count, cat_reg_mean_sqm; local_moran_I, moran_q, moran_sig, region_median_bays	22
F2	F1 without the three Moran features	F1 minus local_moran_I, moran_q, moran_sig	19
F3	F1 without all four target-derived features	F1 minus local_moran_I, moran_q, moran_sig, region_median_bays	18
F4	Spatial and regional features only	density_5 km, log_density_5 km, in_cluster, cluster_size, log_cluster_size, log_kde_intensity, region_enc, reg_north_england, reg_scotland, reg_south_east, reg_south_west	11
F5	Target-derived features only	local_moran_I, moran_q, moran_sig, region_median_bays	4

4.4. Experiment 2: Attribution and Prediction Sensitivity Relative to the Rule (O2)

The reference mechanism in Equation (1) uses floor area and category. These are designated rule-input families, not empirically identified causes of actual loading-bay provision. Category-region counts and mean floor areas are contextual aggregates that can proxy category information. E2 fits seven configurations to each of F1 and F3 on the baseline random 80/20 partition: Random Forest with max_features $\in \{\sqrt{p}, 0.5, 1.0\}$, and XGBoost and LightGBM with column subsampling $\in \{0.8, 1.0\}$. E2 uses 300 trees without early stopping. TreeSHAP is computed for the first 1000 test rows; permutation importance uses the first 3000, five repeats, and the estimator's default R^2 score. SHAP values are computed with shap. TreeExplainer (SHAP 0.52.0), called on the fitted model alone with no background dataset and the default raw model output. Without background data, the explainer uses feature_perturbation = "tree_path_dependent", the path-dependent TreeSHAP algorithm. The expectation over absent features is taken along each tree's decision paths and weighted by the training-data cover stored in each node, so the implicit reference distribution is the 38,145-site training partition as recorded in the fitted trees. No separate interventional background sample is drawn, and the additivity check is disabled (check_additivity = False). The reported attribution shares are therefore path-dependent attributions. Interventional TreeSHAP with an explicit background sample could allocate attribution differently among correlated inputs, such as the several floor-area and category encodings. For continuity with the notebook, we retain the label spurious attribution share (SAS):

$$\mathrm{SAS} = \left[\sum_{j \notin C} I_j\right] \Big/ \left[\sum_j I_j\right] \tag{2}$$

where C denotes the encoded rule-input features.

For SHAP, I_j is the mean absolute attribution for feature j. For permutation importance, negative mean score reductions are set to zero before normalising. SAS based on permutation is therefore a share of positive score reductions, not a directly equivalent quantity to SHAP SAS. Headline SAS includes category-region proxies among features outside the rule. Strict SAS retains the same total denominator but counts only spatial, regional, and target-derived attribution in its numerator. Proxy shares are reported separately. SAS measures attribution outside the reconstructed rule. It does not establish inaccurate explanation of the fitted model or causal irrelevance in the real world. Top-five precision counts encoded rule-input columns among the five highest-ranked SHAP features. Because several columns encode the same underlying attribute, it is a representation-dependent diagnostic.

We also evaluate prediction sensitivity by independently permuting each feature outside the rule-input families within test strata sharing the same reconstructed rule output. This keeps each permuted feature's marginal distribution within an output stratum but breaks dependencies among features and with the fixed inputs. It can produce inconsistent combinations, such as a density value paired with an unrelated log-density. Consequently, the procedure is a diagnostic stress test. We report the mean absolute prediction change and the proportion whose rounded prediction changes. A predictor implementing Equation (1) exactly would be invariant, but sensitivity of a fitted model may reflect both off-rule dependence and extrapolation induced by the perturbation. Joint permutation of coherent feature groups or regeneration from valid inputs would be needed to separate these effects.

A supplementary rank-correlation diagnostic asks whether a fitted model relies more on inputs outside the rule near the rule's rounding steps, where floor area alone is least decisive about the recorded count. It uses the F3 LightGBM configuration with column subsampling 0.8 and its 1000 SHAP-explained test rows. Eligible observations are the warehouse and factory rows above 2000 m^2 in that sample (n = 508: 326 warehouses and 182 factories). Retail rows and rows at or below the threshold are excluded because Equation (1) has no rounding step there. The first variable is off-rule attribution: the sum of absolute SHAP values over all features outside the floor-area and category families. The second is the distance to the nearest rounding step, $|u - \lfloor u \rceil - 0.5|$, where $u = (\mathrm{sqm} - 2000)/k_c$ is floor area above the threshold in normalised bay units. The distance ranges from 0 at a rounding step to 0.5 midway between steps. Spearman's ρ between the two variables is reported with its two-sided p-value. Four capped warehouse rows are retained, although the cap makes their step distance immaterial.

4.5. Experiment 3: Validity, Stability, and Power of Residual Benchmarking (O3)

E3a reproduces the baseline Random Forest flags using the original 24-column

matrix. This includes duplicate aggregates. The model is trained on the baseline 80% partition and predicts all 47,682 sites. The resulting 685 flags combine training and test predictions. A site is flagged when its rounded predicted count exceeds its recorded count. We then assess how many flagged records match Equation (1). For a rule-compliant record, a positive residual represents a mismatch between the fitted surrogate and the reconstructed reference. It supplies no independent evidence of actual under-provision.

E3b evaluates flag stability using out-of-fold predictions from the fixed spatially blocked five-fold design. Three feature sets (F0, F1, and F3), three tree models, and five seeds (42, 7, 123, 2024, and 99) yield 45 flagged sets. Models use the E2 fitting functions without early stopping. Pairwise Jaccard similarity is $|A \cap B| / |A \cup B|$, with two empty sets assigned a similarity of 1. Comparisons across models hold the feature set and seed fixed. Also, comparisons across feature sets hold the model and seed fixed. Seed comparisons hold the feature set, model, and fold assignment fixed. These summaries distinguish algorithmic randomness from specification sensitivity.

E3c tests the detection of simulated downward deviations from the reconstructed rule. For each decrement $d \in \{1, 2, 5\}$, 953 sites are sampled without replacement, corresponding to 2% of the full inventory, from rule-compliant records with at least $d + 1$ recorded bays. Thus, the eligibility thresholds are 2, 3, and 6 bays, respectively. The eligible populations contain 15,656, 10,124, and 4836 records for $d = 1$, 2, and 5, so the 953 perturbed records are 6.1%, 9.4%, and 19.7% of the respective eligible populations, while remaining 2.0% of the full inventory in every case. Because residual flags concentrate among higher-count sites (Section 5.4), the rising contamination rate within the eligible population can raise precision at larger decrements independently of detection ability. Precision is therefore not directly comparable across decrement sizes. The zero-bay and one-bay records are never perturbed. Subtracting d leaves a positive integer count, and no clipping is applied to the contaminated target. Three replicates use seeds 42, 7, and 123. Because eligible populations differ by d, results across decrement sizes are not a same-site dose-response comparison. For F1, Moran features and regional medians are recomputed from the entire contaminated target, including held-out labels. This was carried out to replicate the outcome-dependent baseline. Random Forest and LightGBM are trained on contaminated training-fold outcomes and evaluated out of fold against the known perturbation indicators. A flag requires the rounded prediction to exceed the contaminated count by at least one bay. Recall and precision quantify detection of injected deviations. They do not validate the rule as an adequacy standard or establish detection of real operational shortages.

4.6. Reporting and Reproducibility

The audit pipeline uses fixed seeds for partitions, models, and perturbation replicates. The reported analyses were run on Google Colab with Python 3.13.15,

scikit-learn 1.6.1, XGBoost 3.4.1, LightGBM 4.6.0, and SHAP 0.52.0, as recorded in the results summary file that the notebook writes for each run. The GLMs use statsmodels, and the Local Moran statistics use libpysal and esda. E1 reports means across folds and exports fold-level metrics and R^2 standard deviations. The stability analysis holds fold assignment fixed across seeds. The summaries describe variation in the executed designs. GLM predictions are clipped to [0, 200], a numerical bound inherited from the cleaning specification rather than a verified physical maximum. All metrics use non-negative predictions; exact accuracy and flags use NumPy rounding. Within-one-bay accuracy uses the unrounded prediction. Negative binomial instability is reported as a limitation of this fixed specification, not evidence against the model family.

5. Results

5.1. The Baseline Specification: The Wrong Kind of Success

The reproduced baseline performance is shown in Table 5. Random Forest and LightGBM attain R^2 = 0.9871; within-one-bay accuracy is 98.1% and 98.7%, respectively. XGBoost reaches R^2 = 0.9856, Poisson regression 0.9648, and the fixed-α negative binomial specification 0.7505. The three tree models achieve 96.4% - 97.0% exact-integer accuracy and 99.86% - 100% zero-bay accuracy. These are strong results for reproducing the recorded generated target. They do not assess agreement with independently observed facility counts. The baseline Random Forest subsequently flags 685 sites for the full inventory.

Table 5. Test-set performance of the baseline specification (B0; random 80/20 split, n = 9537).

Model	R^2	MAE (bays)	RMSE (bays)	±1 bay	Exact	Zero-bay acc.
Poisson GLM	0.9648	0.684	1.491	86.8%	48.4%	65.9%
Negative binomial GLM	0.7505	1.286	3.970	78.1%	55.4%	89.7%
Random Forest	0.9871	0.123	0.904	98.1%	96.4%	99.9%
XGBoost	0.9856	0.089	0.955	98.6%	97.0%	100.0%
LightGBM	0.9871	0.098	0.902	98.7%	96.9%	100.0%
Sites flagged as under-provisioned (RF, full inventory)	685					

Note: ±1 bay = share of predictions within one bay of the recorded count; Exact = share of rounded predictions equal to the recorded count; Zero-bay acc. = exact accuracy on sites recording zero bays. LightGBM is early-stopped on the test set, as specified in B0.

5.2. Experiment 1: Accuracy Is Rule Recovery, and Spatial Context Adds Nothing

The summary of the outcome of E1 is presented in Table 6 and Figure 4. These comprise 330 scheduled learned-model fits: six feature sets × five models × (one hold-out split + five random folds + five blocked folds). From the setup, four findings emerged.

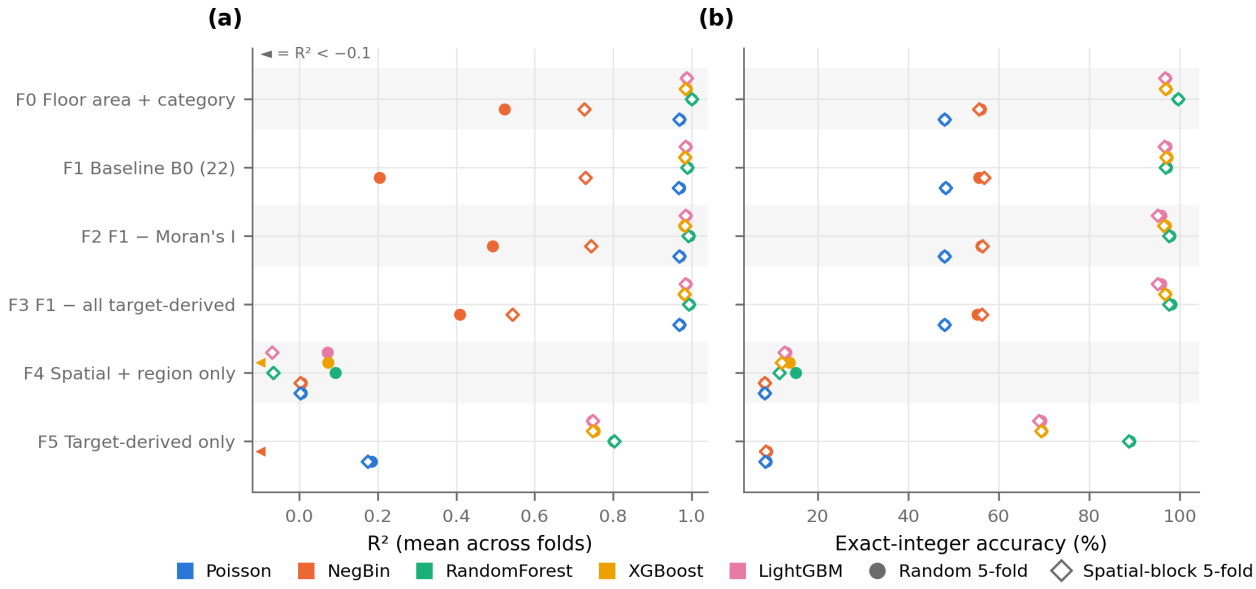


Figure 4. Predictive-fit audit (E1). Mean R^2 (a) and exact-integer accuracy (b) under random five-fold and spatially blocked five-fold evaluation. The plotted "F1 Baseline B0 (22)" label denotes the deduplicated audit feature set, not the original 24-column B0 matrix. The reconstructed rule, identified using all recorded outcomes, has approximately unit R^2 and 99.99% exact agreement; these are reconstruction diagnostics rather than independently validated predictive scores.

Table 6. Accuracy decomposition (E1): mean across folds under random (R) and spatially blocked (B) 5-fold cross-validation.

Feature set	RF R^2 (R)	RF R^2 (B)	RF exact (R)	RF exact (B)	LGBM R^2 (R)	LGBM R^2 (B)	Poisson R^2 (R)	Poisson R^2 (B)
F0 Floor area + category	0.9999	0.9998	99.65%	99.63%	0.9866	0.9871	0.9701	0.9688
F1 Deduplicated audit (22)	0.9903	0.9886	97.18%	96.88%	0.9856	0.9844	0.9694	0.9673
F2 F1 - Moran's I	0.9942	0.9917	97.93%	97.64%	0.9854	0.9837	0.9699	0.9685
F3 F1 - all target-derived	0.9946	0.9926	98.17%	97.65%	0.9854	0.9837	0.9699	0.9685
F4 Spatial + region only	0.0925	−0.0662	15.09%	11.52%	0.0723	−0.0688	0.0057	0.0027
F5 Target-derived only	0.8008	0.8029	88.97%	88.69%	0.7464	0.7484	0.1840	0.1745
Recovered rule (oracle)	1.0000	1.0000	99.99%	99.99%	—	—	—	—

Note: RF = Random Forest (max_features = $\sqrt{p}$); LGBM = LightGBM (early-stopped on a validation split of the training fold). XGBoost results follow LightGBM closely, and the negative binomial GLM is unstable (mean R^2 0.21 - 0.75 on F0 - F3); both are reported in the supplementary tables. Blocked design: 20 km grid cells.

First, rule-input features provide the strongest Random Forest fit. With F0, R^2 is 0.9999 under random cross-validation and 0.9998 under blocked validation. The exact accuracy is 99.65% and 99.63%, respectively. Adding the other 17 unique baseline features lowers random-fold exact accuracy to 97.18% and increases MAE from 0.0077 to 0.0944 bays. With max_features = $\sqrt{p}$, each F1 split considers four of the 22 columns after integer truncation. The probability of con-

sidering informative rule-input columns therefore changes with the feature representation. The result demonstrates specification sensitivity in this forest.

Second, spatial and regional features alone have weak predictive performance. F4 Random Forest reaches $R^2 = 0.0986$ on the random hold-out, but the three tree models have negative R^2 under blocked validation (approximately −0.07 to −0.12). Thus, their squared errors exceed the test-fold mean-reference error. This pattern is compatible with spatial composition supplying limited local predictive information that fails under geographic blocking. It is not evidence that geography is irrelevant to actual loading demand. Moreover, geographic features were constructed on the full inventory, so this is not a fully inductive test at unseen locations.

Third, target-derived inputs are predictive even without floor area or category. F5 Random Forest achieves $R^2 \approx 0.80$ and approximately 89% exact accuracy. This is expected because its inputs contain the outcome, including the site's own count through Moran statistics. When rule inputs are also present, leakage need not increase accuracy. Removing all four target-derived features improves F1 Random Forest random-fold R^2 from 0.9903 to 0.9946. Predictive ablation alone, therefore, cannot justify the absence of leakage. Its presence is established by feature construction, while its measured effect depends on the other available inputs.

Fourth, validation designs produce similarly high R^2 for models supplied with the rule inputs, but not identical scores. The E1 output includes an XGBoost F0 difference of approximately 0.0033 between random and blocked five-fold R^2 (0.9871 versus 0.9838). All the tree configurations nevertheless remain highly predictive of the constructed target. Blocking removes co-location across folds but cannot remove direct target leakage retained in F1, F2, and F5, or establish target validity.

The Poisson specification attains R^2 around 0.97 but approximately 48% exact-integer accuracy in the reproduced hold-out. A log-linear conditional mean does not readily recover the rounded, capped mapping. The fixed-α negative binomial model is unstable across feature sets and validation designs and performs particularly poorly on F5. These comparisons concern the specifications, which include redundant encodings and clipping. They do not establish the general superiority of tree models over appropriately tuned count regressions. Reporting R^2 with metrics like MAE and integer-based metrics makes the practical consequences of these differences clearer.

5.3. Experiment 2: Model Attribution Extends beyond the Reconstructed Rule

The attribution and prediction sensitivity for seven configurations on each of F1 and F3 are shown in Figure 5 and Table 7, respectively. Rule-input attribution and attribution to other feature families are distinguished without interpreting SHAP as a causal estimator.

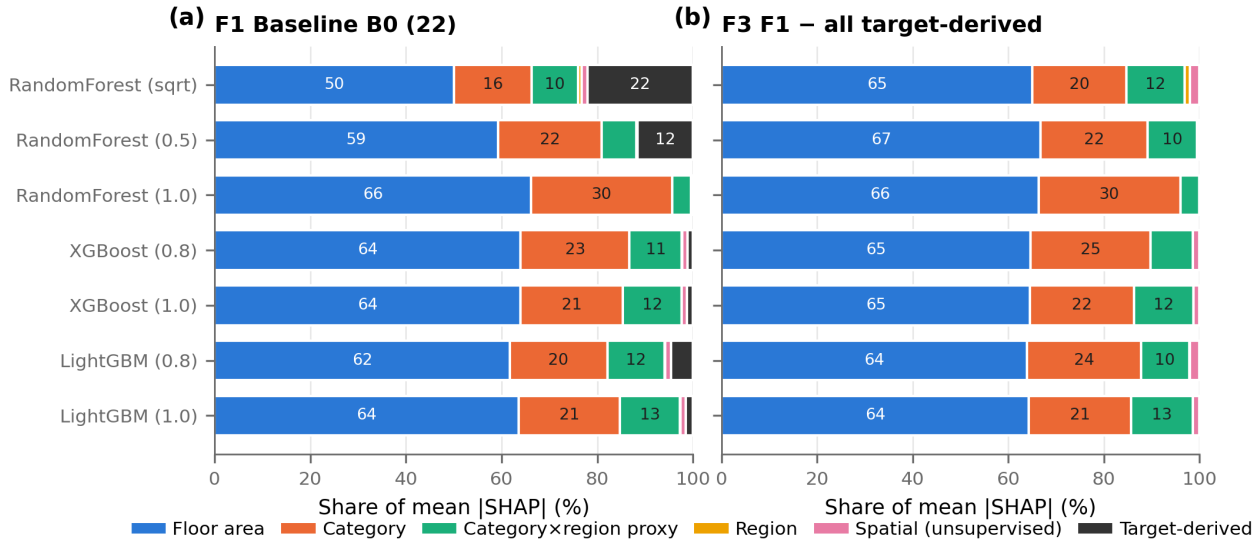


Figure 5. Shares of mean absolute SHAP attribution by feature family for seven configurations on F1 and F3. The figure's legacy "F1 Baseline B0 (22)" label denotes the deduplicated audit variant. Floor area and category are inputs to the reconstructed target rule; other families may be used by the fitted model without appearing in that rule.

Table 7. Attribution relative to the reconstructed rule and prediction sensitivity for seven configurations on F1 and F3.

Set	Configuration	Test R^2	Off-rule SHAP (%)	Strict SAS	Positive perm. SAS (%)	Top-5 prec.	Pred. changed
F1	RF, max_features = √p	0.9895	33.7%	24.0%	12.5%	0.6	14.0%
F1	RF, max_features = 0.5	0.9995	19.1%	11.8%	4.7%	0.8	6.8%
F1	RF, max_features = 1.0	0.9999	4.4%	0.3%	3.7%	0.8	1.1%
F1	XGBoost, colsample = 0.8	0.9853	13.2%	2.3%	19.8%	0.8	4.6%
F1	XGBoost, colsample = 1.0	0.9853	14.7%	2.3%	17.2%	0.4	5.3%
F1	LightGBM, colsample = 0.8	0.9857	17.9%	5.9%	19.4%	0.6	8.5%
F1	LightGBM, colsample = 1.0	0.9859	15.2%	2.6%	16.3%	0.4	6.2%
F3	RF, max_features = √p	0.9955	15.3%	3.2%	6.6%	0.8	13.5%
F3	RF, max_features = 0.5	0.9998	10.8%	0.5%	4.8%	0.8	8.5%
F3	RF, max_features = 1.0	0.9998	4.0%	0.0%	3.5%	0.8	1.2%
F3	XGBoost, colsample = 0.8	0.9827	10.3%	1.5%	11.2%	0.8	5.5%
F3	XGBoost, colsample = 1.0	0.9822	13.7%	1.3%	17.1%	0.4	4.9%
F3	LightGBM, colsample = 0.8	0.9827	12.3%	2.2%	11.5%	0.8	7.2%
F3	LightGBM, colsample = 1.0	0.9830	14.3%	1.5%	16.4%	0.4	6.0%

Note: SAS is the share of the mean absolute SHAP attribution outside encoded rule inputs; strict SAS omits category-region proxies from the numerator but retains the full denominator. Permutation SAS normalises positive mean R^2 reductions after clipping negative importances to zero. Top-five precision is representation-dependent because several columns encode the same rule input. Prediction change is measured under independent within-rule-output-stratum permutations, which may disrupt feature dependencies. The first row is the deduplicated F1 audit forest, not the original 24-column B0 fit.

In the F1 Random Forest configuration with max_features = √p, 33.7% of mean absolute SHAP attribution falls outside the encoded rule-input families. Target-derived features receive 22.1%, and category-region proxies receive approximately 9.7%; the quadrant indicator ranks third among individual columns (Figure 6(b)). Three of the five highest-ranked columns encode rule inputs. These numbers describe the fitted 22-feature model. They do not establish that SHAP misrepresents that model, nor do they estimate the contribution of geography to actual loading-bay provision.

Purely spatial features receive approximately 1.1% of attribution in the F1 forest and at most 2.1% across the 14 fits. The largest off-rule share comes from outcome-derived features and category-region proxies, rather than unsupervised geographic context. The audit therefore identifies the information path behind apparently plausible explanations instead of assuming that every engineered spatial feature materially drives the model.

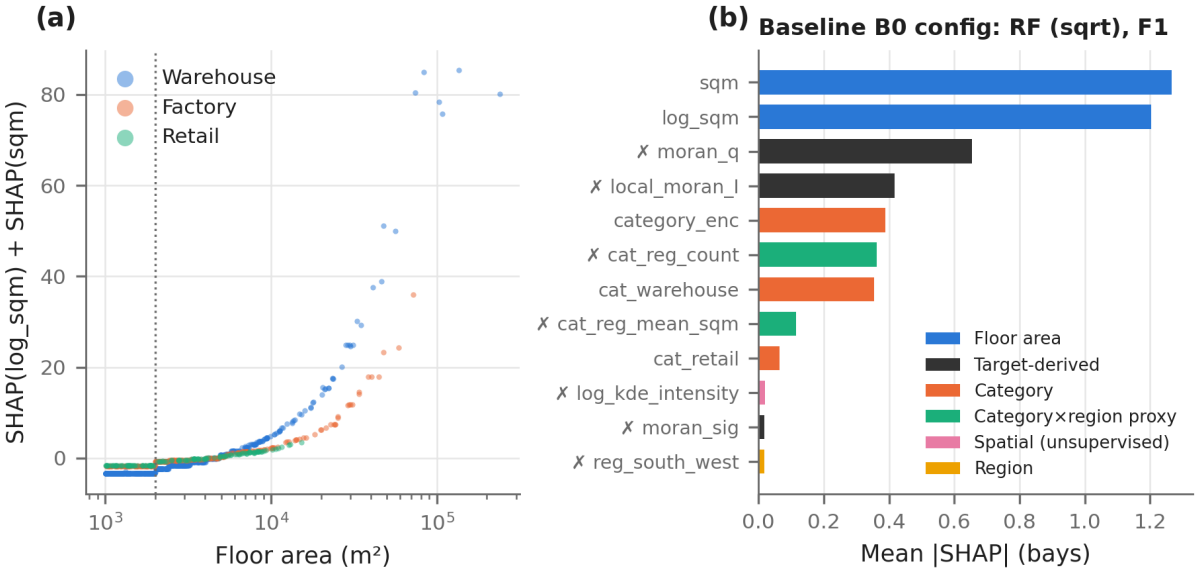


Figure 6. (a) Combined floor-area SHAP attribution for LightGBM on F3, coloured by category, broadly follows the threshold and category-specific slopes of the reconstructed rule. (b) Twelve highest mean absolute attributions for the deduplicated F1 Random Forest with max_features = √p. Features marked ✗ are outside the encoded rule-input families; this marker does not imply inaccurate attribution to the fitted model or demonstrated causal irrelevance in actual provision.

Three patterns qualify the attribution result. First, increasing Random Forest max_features from √p to 1.0 raises hold-out R^2 from 0.9895 to 0.9999 and reduces SHAP SAS from 33.7% to 4.4%; strict SAS falls from 24.0% to 0.3%. Model configuration therefore changes how closely fitted dependencies follow the rule. Second, SHAP SAS and permutation SAS differ. The √p forest has 33.7% versus 12.5%, while some boosted configurations have larger permutation shares. Because the measures use different scales, perturbations, and normalisations, their disagreement is not a direct ranking of explainer correctness. Third, the floor-area attribution broadly reflects the rule's threshold and category-specific shape (Fig-

ure 6(a)). The rank correlation between off-rule attribution and distance to the nearest rounding step (Section 4.4) is $\rho = -0.08$ (p = 0.07, n = 508). Its sign is consistent with slightly more off-rule attribution near rounding steps, but the association is negligible and not distinguishable from zero at the 5% level. It therefore provides no evidence that off-rule attribution concentrates at the rule's breakpoints.

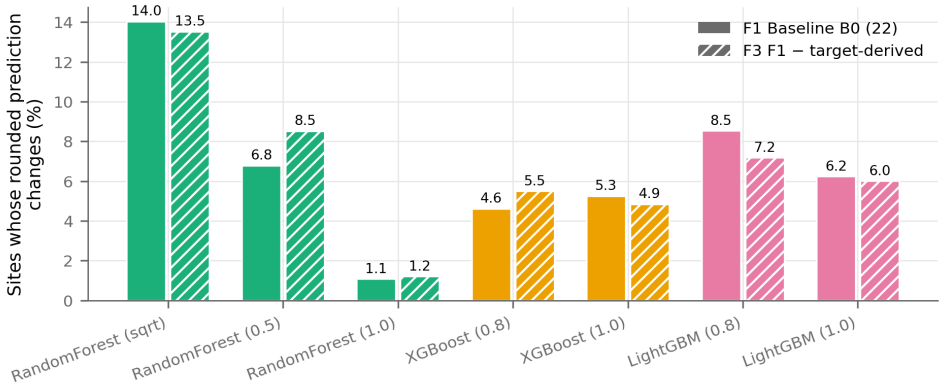


Figure 7. Prediction sensitivity under independent permutation of features outside the rule-input families within strata sharing the same reconstructed rule output. Bars report the proportion of rounded predictions that change. F1 is the 22-feature audit variant. An exact rule predictor would be invariant; changes in fitted predictions may also reflect inconsistent feature combinations introduced by permutation.

The stress test changes rounded predictions at 14.0% of sites for the F1 $\sqrt{p}$ forest and 13.5% for its F3 counterpart (Figure 7). Boosted-model values are approximately 4.6% - 8.5%, while Random Forest with max_features = 1.0 is close to 1%. These results are consistent with configuration-dependent reliance on inputs outside Equation (1). Because independent permutations can move records away from valid joint feature combinations, they do not isolate that reliance from extrapolation effects. With attribution and ablation, the test provides a diagnostic of rule fidelity. It does not establish a failure of SHAP or identify causal drivers in the real world.

5.4. Experiment 3: The Benchmark Flags Model Error, Not Under-Provision

The summary of E3 is shown in Table 8 and Figure 8, respectively. Out of the 685 baseline flags, 684 (99.85%) have recorded counts equal to Equation (1). Their positive residuals therefore represent surrogate error relative to the reconstructed rule. The discrepancy in the remaining record from the rule is not evidence of real under-provision either. One-bay residuals account for 522 flags (76.2%). Flagged sites have a median recorded count of 15 bays versus one across the inventory, with 623 warehouses, 57 factories, and five retail sites. The flag rate is 1.92% for the test partition and 1.32% for the training partition, consistent with their different fitted-prediction status.

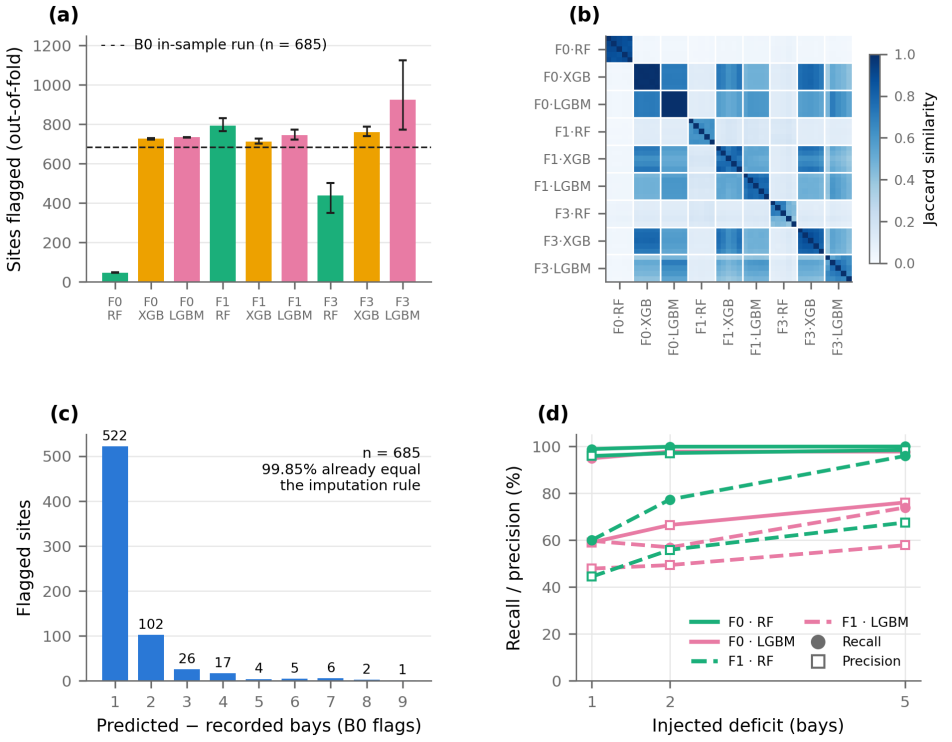


Figure 8. Benchmark validity (E3). (a) Sites flagged as under-provisioned under out-of-fold, spatially blocked prediction, by feature set and model; bars are means over five seeds, whiskers indicate the range; the dashed line marks the 685 flags of the B0 in-sample run. (b) Jaccard similarity between all 45 flagged sets (blocks of five seeds). (c) Size of the predicted deficit among the 685 B0 flags. (d) Recall (filled markers) and precision (open markers) for placebo deficits of 1, 2, and 5 bays injected into 2% of sites.

Table 8. Validity, stability, and detection power of residual-based benchmarking (E3).

Measure	Value
A. Baseline benchmark B0 (RF, 24 columns, mixed training/test prediction)	
Sites flagged as under-provisioned	685 (1.44%)
Flagged sites whose recorded count equals the rule	684 (99.85%)
Flags with a one-bay deficit	522 (76.2%)
Median recorded bays: flagged vs. all sites	15 vs 1
Flag rate: training vs. test sites	1.32% vs 1.92%
Flags by category (warehouse/factory/retail)	623/57/5
B. Stability (45 out-of-fold configurations)	
Flags per configuration (min-max)	46 (F0, RF) to 1124 (F3, LightGBM)
Mean Jaccard: same configuration, different seeds	0.79
Mean Jaccard: same feature set, different models	0.25
Mean Jaccard: same model, different feature set	0.45
Sites flagged by any/by all configurations	3036/6
Flagged sites complying with the rule (range)	97.8% - 99.9%

Continued

C. Placebo detection (recall/precision; deficits of 1, 2, 5 bays)	
F0, Random Forest	98.8/95.9; 99.8/97.1; 99.9/98.4
F0, LightGBM	94.9/59.0; 97.7/66.4; 97.7/76.0
F1, Random Forest	60.0/44.4; 77.3/55.8; 95.9/67.5
F1, LightGBM	59.7/47.8; 56.9/49.4; 73.9/57.8

Note: Each replicate subtracts exactly d bays from 953 rule-compliant records with at least d + 1 bays, leaving positive counts. Eligible populations contain 15,656, 10,124, and 4836 records for d = 1, 2, and 5, so the perturbed share is 6.1%, 9.4%, and 19.7% of each eligible population (2.0% of the full inventory). Recall and precision are means over three seeds, with no population uncertainty interval implied. F1 target-derived features are recomputed on the entire contaminated target, deliberately retaining held-out outcome information.

Out-of-fold flag counts range from 46 to 1124 across the 45 configurations. Mean Jaccard similarity is 0.79 across seeds for a fixed configuration, 0.25 across models at a fixed feature set and seed, and 0.45 across feature sets at a fixed model and seed. Overall, 3036 sites are flagged at least once and six are flagged by every configuration. Between 97.8% and 99.9% of flagged records comply with Equation (1). These results establish substantial specification sensitivity of the residual flags of the audit.

The controlled perturbations assess detection of deviations from the rule. F0 Random Forest recall is 98.8% - 99.9% and precision is 95.9% - 98.4% across the three decrement sizes. F0 LightGBM recall is 94.9% - 97.7%, with precision of 59.0% - 76.0%. For one-bay decrements, F1 recall falls to 60.0% for Random Forest and 59.7% for LightGBM, with precision of 44.4% and 47.8%, respectively. This deterioration is consistent with contaminated outcome-derived inputs allowing models to reproduce the injected deviations. However, F0 and F1 also differ in other features and effective feature sampling, so the comparison does not isolate a causal effect of leakage alone. The decrement-specific populations and three-replicate design further limit comparisons across deficit sizes. The test demonstrates detection behaviour under the specified contamination scenario, not performance against field-verified shortages.

5.5. Consistency Checks and Scope of Robustness

The results support several consistency checks. Rule-input tree models retain high R^2 under random and blocked validation, but spatial-only tree performance deteriorates under blocking. Removing the regional median from F2 produces the same rounded LightGBM and Poisson metrics in Table 6. This is consistent with limited incremental information, but does not prove that those inputs are mathematically interchangeable. Off-rule SHAP attribution is evident under the strict definition excluding category-region proxies. Flag agreement is higher across seeds than across model families, and high rule compliance holds for out-of-fold flags as well as baseline flags. These checks strengthen the core provenance con-

clusion within this case.

6. Discussion

6.1. What the Models Learned, and What They Did Not

The main finding is that strong predictive performance can coexist with a mismatch between the modelled target and the phenomenon decision-makers intend to assess. The baseline achieves R^2 near 0.99 and up to 98.7% within-one-bay accuracy (**Table 5**), but the recorded outcome is almost entirely reproduced by a floor-area-and-category rule. The results establish a successful approximation of that constructed field. They provide no independent evidence that the models recover actual loading-bay counts. The failure occurs when this performance is interpreted as evidence of observed provision.

The distinction concerns analytical use and not flaws in the source inventory. de Saxe *et al.* [6] [15] explicitly describe the bay field as modelled and designed the inventory to support system-level freight analysis. Such a field can be appropriate for scenario modelling when its assumptions and uncertainty remain explicit. Using it to infer actual counts or site-level inadequacy creates a different evidential demand. This case illustrates why provenance documentation must be incorporated into task definition and interpretation, rather than cited only as a data-source description [40].

The geographic results are likewise conditional on the constructed target. Equation (1) has no geographic input, and the selected spatial feature family provides little transferable predictive fit under the executed blocked design. Contextual additions also change Random Forest feature sampling and residual behavior. These findings justify testing feature families against a simple reference rather than assuming that more spatial detail improves a benchmark. They do not imply that location is unimportant for actual freight operations, or that every geographic feature requires a causal interpretation to be useful for prediction.

6.2. Why Explanation Plausibility Does Not Establish Domain Validity

The attribution analysis clarifies the role of XAI in operational governance [10] [11]. SHAP explains a fitted model under a chosen attribution formulation. It does not independently verify the target or recover causal effects. When a model uses outcome-derived or contextual inputs, attribution to those inputs can be faithful to the model while extending beyond the rule that generated its training labels. The 33.7% share therefore quantifies off-rule attribution in one configuration, not a general SHAP failure rate. The distinction between model behaviour and domain interpretation is consistent with the dependence-sensitive attribution literature [43] [44].

Changing the forest's feature-sampling setting substantially improves both predictive fit and concentration of attribution on rule inputs. This demonstrates that interpretation depends on model design as well as on the explanation method. In

this dataset, the explicit reconstructed mapping provides the clearest account of target generation. It should not be treated as a validated physical model because it is transparent or highly predictive.

Differences between SHAP and permutation importance reinforce the need to specify what each measure evaluates [45]-[47]. They should not be presented as interchangeable confirmations of causality. For planning applications, a useful reporting standard is to disclose feature construction, attribution assumptions, sample selection, and sensitivity to coherent perturbations. Agreement between explanation methods may support consistency, but agreement is neither necessary for their distinct estimands to be valid nor sufficient to establish the target's substantive validity.

Documented generated targets can nevertheless support methodological research. They combine real covariates with an auditable construction reference. This allows researchers to examine rule approximation and attribution under realistic correlations. Such testbeds should be explicitly labeled as target-construction audits. Evaluation against the known rule and validation against actual field outcomes answer different questions and should remain separate.

6.3. Implications for Loading-Bay Planning and Benchmarking

For planning practice, there are three implications. First, the inventory is a useful source of processed site attributes and modelled HGV bay estimates. However, its bay field should not be treated as independently observed provision. Second, residuals from an ML surrogate of that field identify discrepancies from a constructed reference rather than operational shortages. Even with independently observed counts, a peer norm would describe typical provision unless linked to demand and service requirements. Third, any residual-based screening tool should disclose threshold sensitivity, out-of-fold flag stability, and the consequences of false positives and missed cases. Controlled perturbations can assess detection of specified deviations, but field verification is needed before flags support site-level planning decisions.

The reconstructed parameters also show how engineering assumptions can be inherited by downstream models. Above 2000 m^2, warehouse increments occur at approximately 823.5 m^2 per additional bay and factory increments at approximately 6532 m^2, with caps; larger retail records receive one bay. The imagery sample comprises 97 sites in total, with modeling based on selected observations and category-specific assumptions [6]. These choices were made for the inventory's modeling purpose. An ML surrogate cannot turn them into representative empirical evidence of adequate capacity. A defensible adequacy benchmark would require verified bay counts linked to loading demand, throughput, vehicle mix, and operating hours, with uncertainty and sampling coverage reported.

6.4. Theoretical Implications: Targets as Instruments

The conceptual contribution is to treat target construction as a measurement pro-

cedure that defines the learning task. Here, the targeted construct is physical HGV loading-bay provision, but the available label is an estimate produced from floor area and category. Training on that estimate evaluates approximation of the measurement procedure. Literature on generated variables and imputation highlights the need to account for construction assumptions and uncertainty [38] [39]. It does not establish that every generated label necessarily biases every predictive analysis. Suitability depends on whether the intended task is surrogate modelling, scenario analysis, estimation of actual provision, or operational decision support.

This interpretation connects target validity with shortcut learning and leakage audits while keeping the concepts distinct [12] [53]. Rule inputs can predict a generated label without leaking an unknown outcome, but the result may still be misinterpreted as evidence about actual provision. Conversely, target-derived features constitute an additional information problem even when rule inputs already determine the label. Target provenance, feature availability, and validation independence should therefore be audited separately. The proposed protocol makes these dependencies explicit rather than treating high accuracy as a substitute for measurement validity.

6.5. A Target Provenance and Validity Protocol

Table 9 presents the Target Provenance and Validity Protocol for planning benchmarks. It is an evidence-informed checklist derived from this case, not a prospectively validated standard. Provenance and distributional inspection should precede predictive modelling; a simple surrogate test can then reveal whether recorded outcomes are nearly recoverable from a documented rule. The remaining steps examine feature availability, deployment-aligned validation, incremental predictive value, interpretation, and decision consequences. A failed check requires clarification of the permissible analytical use instead of automatic rejection of every modelled target.

6.6. Limitations

There are a few limitations that define the scope of the findings. F3 excludes target-derived features but still uses transductive preprocessing. With this, the blocked folds are approximate and unbuffered. The independent within-stratum permutations disrupt joint feature relationships, so their sensitivity estimates cannot isolate off-rule dependence from extrapolation. Also, TreeSHAP settings rely on the installed implementation, additivity checking is disabled, and attribution samples were the first selected test rows rather than repeated samples. Lastly, no independently verified site counts, demand data, or operational adequacy assessments are available.

6.7. Future Research

Future work should prioritise independent measurement. A representative sample of verified HGV loading-bay counts, linked to throughput, vehicle mix, and

service requirements, would permit evaluation of both the inventory's imputation rule and any operational benchmark. Methodological extensions should reconstruct parameters within training folds when estimating generalisation. Inductive geographic features should be derived from training data. Repeated attribution sampling, explicit SHAP background settings, and more deficit-injection replicates would improve uncertainty assessment. Importantly, prospective applications of TPVP should test whether it changes modelling choices and decisions.

Table 9. Target provenance and validity protocol (TPVP) for ML benchmarking in planning.

#	Check	Question and procedure	Failure signal in this study
1	Provenance	Read the data documentation: is the target observed, modeled or imputed, from which inputs, and on what sample?	Bays imputed from floor area and category using 97 hand-counted sites.
2	Surrogate test	Fit a simple surrogate on documented construction inputs; distinguish full-sample reconstruction from held-out evaluation.	Two-input tree: hold-out $R^2 = 0.99985$; 99.80% exact; rule reconstruction uses the full sample.
3	Distributional sanity	Do zeros, heaps, caps, or steps in the target coincide with thresholds in a covariate?	Zero-bay share = share ≤ 2000 m^2 (43.06%); spike at cap 120
4	Leakage map	Is any feature computed from the target, or is any pre-processing or model selection done with test data?	Moran's I, quadrant, significance, regional median; test-set early stopping
5	Validation independence	Match validation to deployment; use group or spatial blocking and training-only preprocessing where required. Blocking does not remove target-derived leakage.	64.5% share six-decimal coordinates; spatial-only tree $R^2 < 0$ when blocked; preprocessing remains full-inventory.
6	Ablation	Does each feature family improve deployment-aligned performance over a simple rule-input baseline?	Spatial features add nothing; F0 is most accurate
7	Attribution and rule fidelity	State the attribution estimand and settings; distinguish model faithfulness from rule fidelity; use coherent perturbations and sensitivity checks.	Off-rule SHAP shares up to 33.7%; independent-permutation sensitivity is 14.0%, subject to joint-feature disruption.
8	Benchmark validity	Report flag stability, threshold sensitivity, and controlled-perturbation detection; verify actual adequacy independently.	Cross-model Jaccard 0.25; simulated one-bay recall 60% (F1); actual adequacy unobserved.

7. Conclusions

This audit shows how a highly accurate loading-bay predictor can reproduce a constructed target without establishing actual site provision. A thresholded, rounded, and capped rule matches 99.992% of recorded counts, and a simple model using its inputs outperforms the reproduced spatially engineered baseline. Outcome-derived inputs create direct information leakage. Other contextual inputs influence fitted dependencies and residual flags without appearing in the reconstructed rule. Out of the 685 baseline flags, 684 comply exactly with that rule. Flag lists vary across specifications, and detection of simulated downward deviations is weaker with the full feature set than with rule inputs alone. These findings concern target approximation and constructed perturbations, not field-verified loading shortages.

The central implication is that target validity must precede predictive benchmarking, and decision validity must follow it. Accuracy evaluates agreement with the supplied label, while attribution describes a fitted model. Neither independently validates what the label measures or whether a residual warrants action. The Target Provenance and Validity Protocol provides a structured way to make those distinctions explicit. Used within its stated limits, it can help planning analysts distinguish useful surrogate and scenario models from claims that require independently observed infrastructure and operational evidence.

Data Availability

The logistics-site inventory is openly available from the University of Cambridge Apollo repository [15]. The reported analyses use its five-field final_dataset.csv file. No independent site-level observations of actual bay counts were collected for this audit.

Code Availability

The analysis is documented in LoadingBay_GroundTruth_Audit_RUN. ipynb, with saved execution outputs. Loading_Bay_Prediction_Pipeline. ipynb supplies the baseline specification. The audit notebook exports tables, figures, and a results summary when run with the source CSV. The saved outputs support the numerical results reported here; exact regeneration also requires the data file and a compatible software environment. Public archival availability of the complete reproducibility package should be established before publication.

Author Contributions

Conceptualization, Tewogbade, S.A.; Methodology, Oraegbunam, L.; Software, Tewogbade, S.A.; Validation, Omowaiye, R.; Formal analysis, Ayodeji, B.; Investigation, Tewogbade, S.A.; Data curation, Oraegbunam, L.; Writing—Omowaiye, R.; Writing—review and editing, Tewogbade, S.A.; Visualization, Oraegbunam, L.; Supervision, Tewogbade, S.A.

All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cheng, F., Han, C. and Tomasi, S. (2025) An Empirical Study of Truck Driver Detention Time: A Shipper's Perspective. *International Journal of Physical Distribution & Logistics Management*, **55**, 154-175. https://doi.org/10.1108/ijpdlm-05-2024-0188

[2] Burns, A.J., Michalek, J.J. and Samaras, C. (2024) Estimating the Potential for Optimized Curb Management to Reduce Delivery Vehicle Double Parking, Traffic Congestion and Energy Consumption. *Transportation Research Part E*: *Logistics and Transportation Review*, **187**, Article ID: 103574.

https://doi.org/10.1016/j.tre.2024.103574

[3] Ghizzawi, F., Galal, A. and Roorda, M.J. (2024) Modelling Parking Behaviour of Commercial Vehicles: A Scoping Review. *Transport Reviews*, **44**, 743-765. https://doi.org/10.1080/01441647.2024.2305202

[4] Amaya, J., Encarnación, T. and Delgado-Lindeman, M. (2023) Understanding Delivery Drivers' Parking Preferences in Urban Freight Operations. *Transportation Research Part A*: *Policy and Practice*, **176**, Article ID: 103823. https://doi.org/10.1016/j.tra.2023.103823

[5] Department for Transport (2022) Future of Freight: A Long-Term Plan. HM Government.

[6] de Saxe, C., Ainalis, D. and Cebon, D. (2024) A Dataset of Logistics Sites in England and Wales: Location, Size, Type and Loading Bays. *Data in Brief*, **54**, Article ID: 110399. https://doi.org/10.1016/j.dib.2024.110399

[7] Wang, S., Mo, B., Zheng, Y., Hess, S. and Zhao, J. (2024) Comparing Hundreds of Machine Learning and Discrete Choice Models for Travel Demand Modeling: An Empirical Benchmark. *Transportation Research Part B*: *Methodological*, **190**, Article ID: 103061. https://doi.org/10.1016/j.trb.2024.103061

[8] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[9] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., 4765-4774.

[10] Olan, F., Spanaki, K., Ahmed, W. and Zhao, G. (2025) Enabling Explainable Artificial Intelligence Capabilities in Supply Chain Decision Support Making. *Production Planning & Control*, **36**, 808-819. https://doi.org/10.1080/09537287.2024.2313514

[11] De Bock, K.W., Coussement, K., Caigny, A.D., Słowiński, R., Baesens, B., Boute, R.N., *et al.* (2024) Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda. *European Journal of Operational Research*, **317**, 249-272. https://doi.org/10.1016/j.ejor.2023.09.026

[12] Kapoor, S. and Narayanan, A. (2023) Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns*, **4**, Article ID: 100804. https://doi.org/10.1016/j.patter.2023.100804

[13] Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., *et al.* (2017) Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*, **40**, 913-929. https://doi.org/10.1111/ecog.02881

[14] Kaufman, S., Rosset, S., Perlich, C. and Stitelman, O. (2012) Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data*, **6**, Article No. 15. https://doi.org/10.1145/2382577.2382579

[15] de Saxe, C., Ainalis, D. and Cebon, D. (2023) Research Data Supporting "Logistics Sites in England and Wales". Apollo-University of Cambridge Repository.

[16] Castrellon, J.P. and Sanchez-Diaz, I. (2024) Effects of Freight Curbside Management on Sustainable Cities: Evidence and Paths Forward. *Transportation Research Part D*: *Transport and Environment*, **130**, Article ID: 104165. https://doi.org/10.1016/j.trd.2024.104165

[17] Reed, S., Campbell, A.M. and Thomas, B.W. (2024) Does Parking Matter? The Impact of Parking Time on Last-Mile Delivery Optimization. *Transportation Research Part*

E: *Logistics and Transportation Review*, **181**, Article ID: 103391. https://doi.org/10.1016/j.tre.2023.103391

[18] Crastes dit Sourd, R., Gastineau, P., Beziat, A. and Koning, M. (2025) Parking Preferences of Delivery Drivers in the Paris Region: Understanding the Role of Anticipation Using Hybrid Choice Models. *Transportation Research Part E*: *Logistics and Transportation Review*, **204**, Article ID: 104420. https://doi.org/10.1016/j.tre.2025.104420

[19] Muriel, J.E., Zhang, L., Fransoo, J.C. and Perez-Franco, R. (2022) Assessing the Impacts of Last Mile Delivery Strategies on Delivery Vehicles and Traffic Network Performance. *Transportation Research Part C*: *Emerging Technologies*, **144**, Article ID: 103915. https://doi.org/10.1016/j.trc.2022.103915

[20] Gregg, K. and Maisel, J. (2025) Cutting across the Curb—A Review of Recent Developments in Municipal Curb Management Policy in America and Canada. *Transport Policy*, **164**, 196-205. https://doi.org/10.1016/j.tranpol.2025.01.031

[21] Castrellon, J.P., Sanchez-Diaz, I. and Gil, J. (2024) Smart Loading Zones. A Data Analytics Approach for Loading Zones Network Design. *Transportation Research Interdisciplinary Perspectives*, **24**, Article ID: 101034. https://doi.org/10.1016/j.trip.2024.101034

[22] Transport for the South East (2025) Warehousing Provision Study. Transport for the South East. https://transportforthesoutheast.org.uk/app/uploads/2025/04/TfSE-Warehousing-Provision-Study-March-2025.pdf

[23] Maxner, T., Dalla Chiara, G. and Goodchild, A. (2025) The State of Sustainable Urban Last-Mile Freight Planning in the United States. *Journal of the American Planning Association*, **91**, 88-101. https://doi.org/10.1080/01944363.2024.2324096

[24] Pinheiro, C.D.P., Gonzalez-Feliu, J., Bertoncini, B.V., Rodrigues, B.A.M. and Loureiro, C.F.G. (2025) Defining and Implementing Freight Accessibility in Cities: Unification of Applications and Directions via a Systematic Literature Review. *Cities*, **167**, Article ID: 106271. https://doi.org/10.1016/j.cities.2025.106271

[25] Rivera-Gonzalez, C., Feng, Z., Klimczak, J., Bayanouni, H., Carr, K. and Roorda, M. (2025) A Freight Data Repository as Foundational Pillar for Urban Freight Research. *Research in Transportation Business & Management*, **61**, Article ID: 101417. https://doi.org/10.1016/j.rtbm.2025.101417

[26] Cameron, A.C. and Trivedi, P.K. (2013) Regression Analysis of Count Data. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9781139013567

[27] Hilbe, J.M. (2011) Negative Binomial Regression. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511973420

[28] Middela, M.S. and Ramadurai, G. (2024) Effect of the Measurement Period and Spatial Dependence on the Accuracy of Urban Freight Trip Generation Models. *Transportation Research Part A*: *Policy and Practice*, **179**, Article ID: 103884. https://doi.org/10.1016/j.tra.2023.103884

[29] Dzinyela, R., Shirazi, M., Das, S. and Lord, D. (2024) The Negative Binomial-Lindley Model with Time-Dependent Parameters: Accounting for Temporal Variations and Excess Zero Observations in Crash Data. *Accident Analysis & Prevention*, **207**, Article ID: 107711. https://doi.org/10.1016/j.aap.2024.107711

[30] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[31] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Pro-*

ceedings of the 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., 3146-3154.

[33] Van Mechelen, I., Boulesteix, A., Dangl, R., Dean, N., Hennig, C., Leisch, F., *et al.* (2023) A White Paper on Good Research Practices in Benchmarking: The Case of Cluster Analysis. *WIREs Data Mining and Knowledge Discovery*, **13**, e1511. https://doi.org/10.1002/widm.1511

[34] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (*KDD*-96), Portland, 2-4 August 1996, 226-231.

[35] Anselin, L. (1995) Local Indicators of Spatial Association—LISA. *Geographical Analysis*, **27**, 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[36] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., *et al.* (2020) Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nature Communications*, **11**, Article No. 4540. https://doi.org/10.1038/s41467-020-18321-y

[37] Meyer, H., Reudenbach, C., Hengl, T., Katurji, M. and Nauss, T. (2018) Improving Performance of Spatio-Temporal Machine Learning Models Using Forward Feature Selection and Target-Oriented Validation. *Environmental Modelling & Software*, **101**, 1-9. https://doi.org/10.1016/j.envsoft.2017.12.001

[38] Meng, X.-L. (1994) Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**, 538-558. https://doi.org/10.1214/ss/1177010269

[39] Pagan, A. (1984) Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, **25**, 221-247. https://doi.org/10.2307/2648877

[40] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., *et al.* (2021) Datasheets for Datasets. *Communications of the ACM*, **64**, 86-92. https://doi.org/10.1145/3458723

[41] Northcutt, C.G., Athalye, A. and Mueller, J. (2021) Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749.

[42] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M. (2021). "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. *Proceedings of the* 2021 *CHI Conference on Human Factors in Computing Systems*, Yokohama, 8-13 May 2021, Article No. 39. https://doi.org/10.1145/3411764.3445518

[43] Aas, K., Jullum, M. and Løland, A. (2021) Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *Artificial Intelligence*, **298**, Article ID: 103502. https://doi.org/10.1016/j.artint.2021.103502

[44] Chen, H., Janizek, J.D., Lundberg, S. and Lee, S.-I. (2020) True to the Model or True to the Data? https://arxiv.org/pdf/2006.16234

[45] Kumar, I.E., Venkatasubramanian, S., Scheidegger, C. and Friedler, S. (2020) Problems with Shapley-Value-Based Explanations as Feature Importance Measures. *Pro-*

ceedings of the 37*th International Conference on Machine Learning*, *PMLR* 119, 13-18 July 2020, 5491-5500.

[46] Hooker, G., Mentch, L. and Zhou, S. (2021) Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or There Is No Free Variable Importance. *Statistics and Computing*, **31**, Article No. 82. https://doi.org/10.1007/s11222-021-10057-z

[47] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, **9**, Article No. 307. https://doi.org/10.1186/1471-2105-9-307

[48] Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020) Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, 7-8 February 2020, 180-186. https://doi.org/10.1145/3375627.3375830

[49] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B. (2018) Sanity Checks for Saliency Maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 31, Curran Associates, Inc., 9505-9515.

[50] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, **1**, 206-215. https://doi.org/10.1038/s42256-019-0048-x

[51] Bayraktar, E., Zaim, S., Gudelek, M. and Golgeci, I. (2025) Global Logistics Efficiency after COVID-19: A Cross-Country Analysis Using the Logistics Performance Index. *International Journal of Logistics Research and Applications*, **29**, 1275-1306. https://doi.org/10.1080/13675567.2025.2540863

[52] de Koning, I., Kassahun, A. and Tekinerdogan, B. (2024) Benchmarking Circularity in Supply Chains: A Systematic Literature Review. *Journal of Environmental Management*, **366**, Article ID: 121676. https://doi.org/10.1016/j.jenvman.2024.121676

[53] Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., *et al.* (2020) Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, **2**, 665-673. https://doi.org/10.1038/s42256-020-00257-z