International Journal of Analysis and Applications

From Probabilities to Workload: Calibrated ML and Decision-Curve Analysis for Passenger-Service Operations

Thanyaporn Sukdet, Warawut Narkbunnum*

Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand

**Corresponding author:* warawut.n@acc.msu.ac.th

ABSTRACT. Airline passenger-service operations increasingly rely on predictive decision support to triage large volumes of service cases under capacity constraints. However, many studies emphasize ranking performance while providing limited evidence on the reliability of probabilities and the operational feasibility of threshold-based deployment. This study examines whether calibrated probabilities from machine learning (ML) models can support actionable, workload-aware decisions in airline passenger service operations. We implement a leakage-controlled pipeline with frozen preprocessing and compare regularized logistic regression, random forests, and gradient-boosting trees. Predicted probabilities are calibrated on non-test data using isotonic regression. Evaluation on a held-out test set integrates discrimination, probability-quality metrics, and calibration diagnostics. Decision curve analysis (DCA) quantifies net benefit relative to treat-none and treat-all strategies across an operational threshold window (0.05–0.15). To translate decision policies into operational implications, thresholds are mapped to workload units, including alerts per day and expected counts of true and false positives per day, assuming a nominal daily volume of 1,200 cases. Segment audits assess robustness across cabin class, travel type, delay band, and distance band. After calibration, the selected gradient-boosting model yields positive net benefit throughout the threshold window and forms the upper envelope relative to baseline strategies. Workload translation indicates that thresholds in the 0.05–0.15 range preserve true-positive throughput while reducing false-positive alerts, enabling managers to tune intervention capacity with minimal loss of decision value. Segment audits indicate generally stable performance, while identifying localized conditions in which modest segment-specific thresholds or light post-alert triage may reduce unwarranted escalations. Overall, calibrated ML, combined with DCA and workload translation, provides operationally feasible, governance-ready support for passenger-service decision making, improving reliability and efficiency while allowing threshold policies to adapt to daily capacity and evolving operating conditions.

Received Mar. 23, 2026

2020 Mathematics Subject Classification. 62H30, 68T05, 62J02, 90B06, 62P30.

Key words and phrases. airline passenger-service operations; machine learning; probability calibration; decision-curve analysis; workload translation; capacity planning

https://doi.org/10.28924/2291-8639-24-2026-250

ISSN: 2291-8639

1. Introduction

Passenger-service operations are a critical component of the air-transport ecosystem, encompassing check-in, security screening, boarding, and in-flight service; performance at these touchpoints directly shapes perceived service quality, reliability, and even airport image [1]. Global demand has recovered strongly: industry statistical bulletins jointly released by ACI World and ICAO report that 2024 passenger volumes surpassed 2019 levels (≈9.5 billion; 104%), with continued growth expected thereafter [2]. Yet many terminals face persistent capacity and staffing constraints. Empirical and modeling studies of terminal processes repeatedly identify check-in, particularly security, as a bottleneck during peak demand, where queueing dynamics degrade throughput and service reliability [3], [4]. At the network level, delay analyses show that "reactionary" (knock-on) delays, often triggered by missed connections and terminal-side congestion, account for the largest share of delay minutes in Europe (≈ 46–48% in 2023), underscoring how passenger-flow disruptions propagate to flight operations [5]. Consistent with this, network science evidence demonstrates that passenger/crew connections are a dominant mechanism for systemic delay propagation, so a single checkpoint bottleneck can cascade into missed connections and widespread schedule disruption [6]. From an operations management perspective, security lane staffing and scheduling policies are often designed around maximum wait constraints (e.g., ≤45 minutes), yet achieving these targets during peak periods requires anticipatory control rather than reactive measures [7]. As airlines pursue ambitious efficiency and sustainability objectives while facing weather and geopolitical shocks, improving system reliability and resilience, therefore, calls for decision support that links predictive analytics to actionable staffing and queue-management interventions.

Much of the recent transportation research has focused on machine learning (ML) to predict demand, delays, and passenger satisfaction. Hybrid and ensemble methods, such as optimized random forests, consistently outperform traditional techniques in flight-delay prediction across various datasets, including recent deep and hybrid models [8], [9], [10]. Deep learning and modern ML pipelines have likewise achieved high predictive performance for airline passenger satisfaction in peer-reviewed settings [11], [12]. Telematics and kinematic features further enable risk stratification in mobility and safety applications, supporting targeted interventions [13], [14], [15]. However, practical use in passenger-service operations remains limited because probability outputs are rarely calibrated and are seldom translated into actionable workload or policy metrics. Reviews of classifier calibration emphasize that discrimination metrics (e.g., accuracy, AUC) are insufficient for decision support without evaluating probability reliability (e.g., Brier score, calibration curves, ECE) and post-hoc calibration [16]. In public-transport satisfaction research, non-linear effects of waiting and crowding are well documented, but calibrated risk estimates and explicit decision analyses are rarely reported [17], [18]. Likewise, many flight-delay studies foreground predictive accuracy

while giving limited attention to the reliability of delay probabilities or to threshold choices that connect predictions to staffing and resource allocation [8], [19]. Without calibration and decision-analytic evaluation, operators cannot identify probability cut-offs that yield positive net benefit relative to "treat-all" or "treat-none" strategies; decision-curve analysis offers a principled framework for translating predictions into policy [20], [21].

Another gap concerns stability across operational segments that are meaningful to airlines. Passenger expectations vary by cabin class, flight distance, delay bucket, and travel purpose; yet ML models are often trained and reported on aggregate data, masking segment-specific biases or instabilities. This raises concerns for fairness and governance: for example, deploying a miscalibrated model that underestimates dissatisfaction for economy passengers could exacerbate inequities. Recent transportation studies have begun to formalize and audit such issues using fairness definitions like equality of opportunity (equal true-positive rates) and equality of prediction accuracy (equal error rates) [22], [23]. Fairness-aware approaches adjust thresholds or add regularization to balance error rates across user groups, but comparable assessments remain rare in passenger-service operations [23], [24]. System-level resilience, therefore, requires decision-support tools that not only perform well on average but also deliver consistent reliability across segments and remain robust under distribution shifts such as post-pandemic travel patterns, an aim aligned with calibrated ML and segment-specific audits.

We quantify probability accuracy using the Brier score and the Murphy decomposition and assess miscalibration with expected calibration error (ECE), calibration-in-the-large, and calibration slope. To translate probabilistic accuracy into operational value, we employ decision curve analysis, in which net benefit captures the standardized trade-off between true-positive yield and false-positive burden at a given threshold.

Our study addresses these gaps by developing a calibrated ML decision-support pipeline for airline passenger service operations. The pipeline controls for data leakage; applies isotonic regression to calibrate probability outputs; and evaluates reliability using the Brier score, Murphy's decomposition (reliability, resolution, uncertainty), and expected calibration error (ECE) [25], [26], [27], [28], [29], [30]. Decision-curve analysis (DCA) then translates calibrated probabilities into net-benefit curves across policy thresholds by weighting true positives against the harms of false positives and plotting net benefit as a function of threshold probability [20], [21]. This framework identifies threshold ranges where model-guided interventions outperform default policies (treat-all vs. treat-none) and connects predictions to actionable policy windows. Finally, we translate net benefit into operational workloads: alerts per day and expected true and false positives per day to support staffing and resource planning.

This study makes four contributions. First, we present a leakage-controlled pipeline with isotonic calibration, ensuring probability fidelity for downstream decisions. Second, we evaluate decision utility via net benefit within an operationally justified threshold window (0.05–0.15),

directly linking model choice to staffing and queue management. Third, we translate model outputs into workload metrics (alerts/day, TP/FP per day) to support capacity planning. Fourth, we conduct segment-level audits (cabin class, distance, delay, travel type) to assess stability and inform segment-specific operating points, thereby improving reliability across heterogeneous traffic mixes.

We address three research questions: RQ1: Within the operational threshold window, do calibrated models achieve higher net benefit than treat-none and treat-all strategies, and across what sub-intervals is this advantage sustained? RQ2: How stable are net benefit and workload (TP/FP per day) across key operational segments, and can segment-specific thresholds reduce FP-burden without materially eroding TP-yield? RQ3: How do calibration properties (Brier/Murphy, ECE, calibration-in-the-large/slope) explain observed differences in net benefit and workload?

2. Literature review

2.1 ML for Passenger Operations

2.1.1 Safety risk prediction

Machine learning applications for road-user and passenger safety increasingly use high-frequency telematics to detect elevated crash risk and enable proactive actions [14], [31]. Using decision-adjusted risk models with kinematic data enhances the prioritization of high-risk drivers by focusing on the utility of interventions rather than just discrimination [15]. Current real-time systems combine sequence models with calibrated-confidence learning, ensuring predicted crash probabilities match observed data, a key factor for reliable alerts at scale [32]. In addition to point estimates, conformal prediction provides pre-crash injury-risk ranges with guaranteed coverage, offering auditable probabilistic intervals for operational decisions [33]. Throughout this research, reviews and design analyses highlight that explicit probability calibration, usually through post-hoc methods applied to gradient-boosted or deep models, is essential to avoid overconfident recommendations that could jeopardize safety [14].

2.1.2 Operational efficiency and delay management

In public transportation, open-data pipelines combined with machine learning can predict ferry delays and reveal operational factors such as headways, trip start times, and vessel identifiers, thereby supporting schedule reliability [34]. In aviation, hybrid methods that combine clustering with optimized ensemble learners have improved flight-delay forecasting, enabling earlier rescheduling and targeted resource deployment [8], [9]. Despite strong discrimination, many models fall short of converting probabilistic outputs into calibrated decision thresholds that guide operators when to intervene. Research on terminal queue management demonstrates that congestion at check-in and security significantly impacts throughput and reliability, and that lane-assignment rules and staffing based on explicit maximum-wait constraints can reduce

waiting times and improve flow [4], [7]. At the network level, empirical evidence shows that local bottlenecks create knock-on (reactionary) delays across connected flights, emphasizing the importance of linking predictions to upstream passenger-flow controls [6]. These findings underscore the operational value of predictive tools that not only forecast delays but also convert calibrated risk estimates into staffing and queue-management actions.

2.1.3 Passenger experience and demand forecasting

Evidence from public transport shows that non-linear machine-learning models can capture threshold effects in perceived service quality, e.g., steep drops in satisfaction once waiting times and crowding exceed certain levels, supporting experience-oriented service design [35]. Regarding demand, recent systematic reviews conclude that models combining socio-demographics, service attributes, and temporal patterns (e.g., lags/seasonality) yield more accurate ridership forecasts than traditional baselines, thereby enhancing inputs for scheduling and fleet management [36], [37]. However, recent domain reviews and surveys consistently report that uncertainty quantification, probability calibration, and fairness diagnostics are infrequently documented, which constrains the safe integration of forecasts into capacity planning and pricing [21], [23], [36], [37], [38], [39]. Decision-focused learning in ride-hailing exemplifies how aligning model training with downstream policy, here, subsidy allocation optimized according to platform objectives, can improve system-level outcomes, demonstrating a pathway to make demand and satisfaction models operationally meaningful in aviation [40].

2.1.4 Resilience and sustainability

Work on transport resilience increasingly investigates how predictive models support operations during disruptions by anticipating load, re-routing flows, and prioritizing scarce capacity. Graph-based spatio-temporal architectures have proved effective for forecasting traffic states under irregular conditions, improving short-horizon situational awareness that underpins resilient control policies [41], [42]. In parallel, fairness-aware demand modeling has begun to incorporate constraints that prevent systematic under-service of user groups or regions, linking predictive accuracy to equitable access and more sustainable resource allocation [23], [24]. Despite these advances, most studies still emphasize discrimination metrics and omit probability calibration or decision-analytic evaluation, reducing the operational value of forecasts for sustainable scheduling and recovery planning [20], [21], [25], [28].

In sum, ML research has improved predictive accuracy for many passenger-service tasks, yet calibrated probabilities and explicit decision rules remain uncommon [16], [50]. Absent calibration and decision-curve analysis, models are rarely audited across segments or translated into actionable thresholds, constraining their contribution to system safety, efficiency, reliability, and sustainability [20], [21], [23], [24], [29], [30].

2.2 Probability calibration and reliability

For model outputs to guide operations, predicted probabilities must agree with empirical frequencies; strictly proper scoring rules formalize this requirement, with the Brier score measuring the mean-squared gap between predictions and outcomes [43]. The classical Brier decomposition attributes error to reliability (calibration), resolution (separation of high- vs. low-risk cases), and uncertainty, diagnosing whether deficiencies stem from miscalibration or weak discrimination [27], [28].

Modern practice extends Brier analysis with reliability diagrams and scalar miscalibration summaries. Statistically principled calibration plots using optimal or smoothed binning reduce estimator bias and improve interpretability in applied settings [44], [45]. Empirical and survey evidence further shows that single-number metrics such as the Expected Calibration Error (ECE) are sensitive to binning choices and estimator specification, motivating careful reporting and, where relevant, class-specific variants [16]. In real-world use, the operative standard is simple: a predicted probability of 0.70 should be correct about 70% of the time in comparable cases; otherwise, thresholds and workload planning based on those probabilities may be unreliable [46], [47].

Post hoc calibration transforms scores into reliable probabilities. Platt scaling provides a parametric logistic mapping, whereas isotonic regression offers a monotone nonparametric alternative suitable for non-logistic distortions [48]. Recent journal surveys synthesize advances in modern machine learning (including deep and ensemble models) and recommend reporting both discrimination and calibration, often after post-hoc adjustment, as prerequisites for decision support [16].

When interval assurances are needed, conformal prediction provides distribution-free coverage guarantees for prediction sets/intervals, with recent expositions and applications showing how guaranteed-coverage risk bands can be audited in practice [33], [49]. Taken together, the Brier/Murphy framework, stable calibration diagnostics, and post-hoc calibration (plus conformal intervals, where appropriate) move ML from raw scores toward actionable, trustworthy probabilities for airline passenger service decisions.

2.3 Decision-Curve Analysis and Policy Thresholds

Conventional discrimination metrics (e.g., AUC) do not determine whether using a model improves decisions. Decision-curve analysis (DCA) evaluates net benefit across intervention thresholds by linking predicted probabilities to actions and comparing model-guided decisions with "treat-all" and "treat-none" strategies [50]. Recent methodological advances further extend DCA to quantify uncertainty and perform formal inference, including confidence intervals and hypothesis testing for net benefit, as well as a Bayesian formulation (bayesDCA) that supports posterior uncertainty statements about strategy dominance [51]. In DCA, the action threshold encodes the trade-off between false positives and false negatives; at each threshold, the net benefit equals the true-positive rate minus a penalty for false positives scaled by the threshold. Because

DCA requires only a threshold probability rather than full monetary utilities—and can be computed from standard validation data, it provides a practical bridge from predictive scores to policy choices [50], [51].

Within transportation, explicit decision-analytic evaluations remain uncommon, yet adjacent strands underscore the need for threshold-aware policy design. In ride-hailing, empirical evidence shows that incentive schemes structured by cut-offs, such as subsidies or bonus programs, materially influence drivers' labor supply and platform outcomes, with effects contingent on operational goals and constraints [52], [53]. Complementary studies further document that price- and bonus-like signals (e.g., surge or targeted incentives) alter driver relocation and market performance, reinforcing the case for formally selecting thresholds rather than relying on heuristics [54]. A complementary line of work aligns learning objectives with downstream allocation, demonstrating that decision-focused learning can optimize city-level subsidy allocation, a template for coupling prediction and policy that is directly compatible with DCA-style operating-point selection [40]. In parallel, a Q1 systematic review finds that in-vehicle telematics programs commonly rely on risk thresholds (e.g., speeding, harsh braking) to trigger feedback or incentives, reinforcing the need to select cut-offs based on operational value rather than heuristics [14].

Despite these advances, many transport ML applications still use thresholds heuristically (e.g., 0.50 or simple class-imbalance rules), leaving the operational value uncertain; DCA directly addresses this by quantifying where a model yields a positive net benefit relative to default strategies [50], [51].

We apply DCA to passenger-satisfaction risk within a predefined policy window of 0.05–0.15 dissatisfaction probability. The lower bound screens out overly aggressive operating points that would flood frontline operations, while the upper bound preserves feasible opportunities for proactive service recovery under realistic staffing levels. We report net-benefit gains within this window and translate chosen thresholds into daily workload metrics (alerts/day; expected true/false positives/day) to support decision-analytic justification by airline managers [50], [51].

2.4 Operational Stability and Governance

Implementing ML in passenger-service operations raises concerns about fairness, robustness, and governance. Distribution shifts and transferability are key issues: models trained in one region or time often lose accuracy when demand patterns or network conditions change [55], [56]. In transportation analytics, short-term traffic and safety studies show that adapting models across domains with transfer learning or domain adaptation improves out-of-domain performance relative to naïve deployment [57], [58], [59]. For passenger services, such shifts occur during seasonal peaks, post-pandemic travel patterns, or schedule and network changes; without ongoing monitoring and recalibration, risk estimates can systematically overestimate or underestimate dissatisfaction in new situations, reducing reliability and resilience [21], [60].

Governance thus requires explicit portability audits, such as external validation on new segments, calibration checks with updates, when necessary (like simple intercept/slope recalibration or non-parametric mapping), and documentation of any threshold changes to maintain fair and stable service levels across cabins, distance buckets, and delay categories [23].

Fairness across segments. Passenger transportation systems serve heterogeneous users whose expectations and sensitivities vary by cabin class, travel purpose, and region. In predictive decision support, widely used fairness notions include equality of opportunity and equalized odds, which require comparable true-positive (and often false-positive) rates across protected groups, and accuracy-based notions that seek prediction performance independent of protected attributes [61]. Recent advances demonstrate that fairness can be enforced with minimal loss in overall accuracy through either in-processing penalties or post-processing adjustments, such as threshold optimization, to balance error rates across groups [62], [63]. In transportation applications, emerging work shows that fairness-regularized forecasting pipelines and post-deployment threshold calibration can narrow inter-group performance gaps while preserving system-level utility [23]. Platform studies in ride-hailing likewise indicate that fairness-aware matching and policy tuning can improve equity (e.g., income or service allocation) without materially degrading efficiency, supporting transportation-equity goals [64], [65]. For passenger-service operations, routine segment audits can reveal systematic under-prediction of dissatisfaction in specific cabins, distance bands, or delay types; segment-specific thresholds or recalibration should then be applied to equalize relevant metrics (e.g., recall) across groups and ensure equitable resource allocation [23], [63].

Transparency and accountability are central themes. Emerging governance scholarship highlights explainability and auditability as essential for trustworthy machine learning, advocating for documenting model logic, uncertainties, and accountability pathways [66]. In transportation studies, explainability has become increasingly common. For example, mode-choice and safety applications now include SHAP-based analyses to attribute feature contributions alongside predictive models [67]. In incident management, Kernel density estimation and Monte Carlo simulation were used to derive clearance-time distributions, accompanied by reliability assessments akin to calibration diagnostics, to operationalize transparent uncertainty quantification [68]. Similarly, queue-management evidence emphasizes the importance of real-time data and adaptive staffing; discrete-event simulation used in airport terminals serves as a decision tool for evaluating staffing levels and reducing wait times [69]. Human oversight remains crucial: model outputs should serve as decision aids, with clearly communicated uncertainties such as calibrated probabilities and, when appropriate, decision-curve–based net benefits and include explicit workload metrics like expected true/false alerts per operating threshold [50], [70]. Governance also involves life-cycle monitoring, tracking performance, and calibration over time and across passenger segments, initiating recalibration or

threshold adjustments when drift is detected, and maintaining an auditable record of these changes [71], [72].

In summary, converging strands across machine learning, calibration, decision analytics, and operational governance provide a foundation for calibrated decision support in passenger-service operations; however, an integration tailored to airline contexts remains underdeveloped. Our study addresses this gap by combining leakage-controlled modeling, isotonic calibration with Brier/Murphy/ECE diagnostics, decision-curve analysis, and segment-specific audits. By translating calibrated probabilities into net-benefit curves and workload tables, the approach aims to support safe, efficient, reliable, and resilient passenger services.

3. Methodology

3.1 Study aims and design

We developed and evaluated a probabilistic machine-learning system to support operational decision-making in airline passenger-service processes. The primary aim was to generate well-calibrated class probabilities that can be translated into actionable workload plans under realistic operating constraints. Beyond discrimination, the evaluation emphasizes probability quality and decision-analytic utility within a prespecified policy window of threshold probabilities, $pt \in [0.05, 0.15]$, reflecting the range typically discussed by operations managers for proactive interventions in queue control, staffing, and exception handling.

We adopted a supervised-learning design with disjoint training, validation, and test partitions. All preprocessing specifications and candidate models were defined on the training set; the validation set informed model and hyperparameter selection; and the held-out test set was reserved exclusively for final reporting, decision-curve analysis, and workload translation. To prevent data leakage, all learned preprocessing artifacts (e.g., encoders, imputers, scalers) were fit on training data only, frozen, and then applied to validation and test features without refitting. Random seeds, hashing rules for splits, and software versions are recorded to enable exact replication (see Appendix: Runbook and Research Data Statement).

3.2 Data and outcome

We analyzed routinely collected passenger-service records comprising demographic attributes, booking characteristics (e.g., cabin class, travel type), service ratings, operational timings (e.g., departure/arrival processing), and flight descriptors (e.g., distance band, delay indicators). The binary outcome was passenger satisfaction, with the positive class defined as "satisfied." The class prevalence was moderate, supporting both discrimination analysis and decision-analytic assessment within the policy window. Data quality controls encompassed schema validation, range checks, rare-level consolidation for high-cardinality categoricals, and predefined rules for missingness handling; full field definitions and quality checks are provided in the data dictionary and schema files (Appendix).

3.3 Feature engineering and operational segmentation

We constructed features from the raw inputs and derived operational segments to enable interpretable workload reporting. Segments comprised:

1. cabin class mapped to {economy, premium-economy, business};

2. travel type mapped to {business, leisure};

3. departure-delay band derived from departure delay in minutes using empirical quantiles (default tertiles, negatives truncated at zero), and

4. distance band derived from flight distance (short/medium/long by quantiles).

These segments are used only for stratified analyses and do not alter the primary model selection.

3.4 Modeling pipeline

Numeric features were standardized and missing values imputed using transformers fit on the training set and then frozen. Categorical variables were encoded with a consistent, training-fit vocabulary to prevent leakage. We evaluated regularized logistic regression, random forests, and gradient-boosting trees (LightGBM). For each candidate, a single pipeline concatenated the frozen preprocessing with the estimator. Training used only the training split; the validation split informed model family selection. Selection emphasized ranking performance (ROC-AUC) while reporting F1, accuracy, and the Brier score to capture both discrimination and probability quality.

3.5 Probability calibration

Because operational policies act on predicted probabilities, we assessed and, when necessary, calibrated outputs on non-test data using isotonic regression. Probability quality was quantified by the Brier score, expected calibration error, and reliability curves, complemented by the Murphy decomposition when informative. The calibrated probabilities were produced on the held-out test set for all downstream analyses.

3.6 Evaluation protocol

We reported discrimination (ROC-AUC and, where helpful, PR-AUC) and probability quality on the test set. Threshold-based summaries focused on three pre-specified operating points (threshold probabilities 0.05, 0.10, and 0.15), with confusion-matrix–derived metrics (precision, recall, specificity, F1). The Youden index computed over a fine grid yielded a high-specificity reference typical of screening contexts.

3.7 Decision-analytic and operational evaluation

Decision-curve analysis quantified net benefit relative to treat-none and treat-all strategies across a continuous threshold range, with emphasis on the 0.05 – 0.15 policy window. We also reported the difference in net benefit versus treat-none to aid interpretation and avoid scale compression outside the window. To link model guidance to staffing feasibility, thresholded probabilities were translated into daily operational workload alerts, true positives, and false

positives, assuming a nominal daily volume of 1,200 passenger-service cases (alternative volumes scale linearly). Workload was summarized overall and by operational segments (cabin class, travel type, delay band, distance band). Subgroup stability was assessed by comparing true- and false-positive rates across segments, with surface disparities that may warrant segment-specific thresholds, staffing adjustments, or targeted monitoring, while avoiding overinterpretation of very small cells.

3.8 Reproducibility

All steps, splitting, feature construction, preprocessing, model training, calibration, DCA, and workload translation, were scripted and versioned. Random seeds were fixed where applicable for deterministic behavior. The full run sequence and parameter settings are documented in an appendix runbook to facilitate exact replication.

4. Results

4.1 Data characteristics and class balance

We first describe the study cohorts to contextualize subsequent performance and decision-analytic results. Table 1 summarizes, by split (train, validation, test), the number of records, number of positive cases, positive-class prevalence, overall missingness, number of modeled features, and concise summaries of key covariates.

Table 1. Descriptive characteristics of the study cohorts by split (train, validation, test)

split	train	validation	test
n_rows	66,498	16,625	25,976
n_cols	28	28	28
n_positive	28,816	7,204	11,403
prevalence	0.43	0.43	0.44
missing_rate_overall	0	0	0
n_features	27	27	27
Age_mean	39.38	39.46	39.62
Age_sd	15.15	15.05	15.14
Arrival Delay in Minutes_mean	15.32	15.03	14.74
Arrival Delay in Minutes_sd	38.98	38.73	37.52
Baggage handling_mean	3.63	3.63	3.63
Baggage handling_sd	1.18	1.19	1.18
Checkin service_mean	3.30	3.30	3.31
Checkin service_sd	1.26	1.26	1.27
Cleanliness_mean	3.29	3.29	3.29
Cleanliness_sd	1.31	1.31	1.32

Across splits, the positive class is defined as satisfied. Cohorts show consistent class balance (train n = 66,498; validation n = 16,625; test n = 25,976), with a prevalence of around 43–44%, and no observed overall missingness (0%), supporting both discrimination analysis and decision-curve evaluation on the held-out test set. Numeric covariates are stable across splits (e.g., age ≈ 39.5 ± 15 years; arrival delay ≈ 15 ± 38–39 minutes). Service-quality items such as baggage handling, check-in service, and cleanliness cluster around neutral-to-positive central tendencies (~3.3 ± 1.2–1.3 on a 1–5 scale). Categorical factors used for segmentation (e.g., cabin class, travel type, distance bands, delay bands) are sufficiently represented across levels to allow subgroup analyses without small-cell concerns.

4.2 Model selection on the validation split

Table 2 reports that models were trained on the training split and evaluated on the validation split, using the area under the ROC curve (AUC) as the primary selection criterion, with F1-score, accuracy, and Brier score as secondary diagnostics (higher is better, except for Brier). The candidates included calibrated gradient-boosted trees (LightGBM), random forest, and L2-regularized logistic regression. Results are summarized below.

Table 2. Validation performance by model (AUC as the primary criterion; F1, accuracy, and Brier as secondary diagnostics; higher is better except Brier).

model	AUC	F1	Acc	Brier
lightgbm	0.9952	0.9594	0.9654	0.0255
random_forest	0.9940	0.9559	0.9623	0.0289
logistic_l2	0.9285	0.8596	0.8805	0.0933

LightGBM achieved the highest AUC (0.9952), strong classification performance (F1 = 0.9594; accuracy = 0.9654), and the lowest Brier score (0.0255), indicating superior discrimination and probability quality. The random forest ranked second (AUC = 0.9940; F1 = 0.9559; accuracy = 0.9623; Brier = 0.0289), while logistic regression trailed (AUC = 0.9285; F1 = 0.8596; accuracy = 0.8805; Brier = 0.0933). Accordingly, LightGBM was selected as the reference model for held-out testing, calibration assessment, decision-curve analysis, and workload translation within the pre-specified policy window ($pt \in [0.05, 0.15]$).

4.3 Operating-point performance on the held-out test split

Using the selected calibrated LightGBM model, we evaluated thresholded performance on the held-out test split with emphasis on the pre-specified policy window ($pt \in [0.05, 0.15]$). Table 3 reports operating counts (TP, FP, TN, FN) and summary metrics (precision, recall, specificity, F1) at representative thresholds (0.05, 0.10, 0.15) and, for reference, at a high-confidence Youden point (~0.42).

Table 3. Operating-point performance on the held-out test split (LightGBM). Counts and metrics are given at pt = 0.05, 0.10, 0.15, and at a high-confidence point (~0.42).

threshold	0.05	0.10	0.15	0.42
N	25,976	25,976	25,976	25,976
TP	11,364	11,311	11,263	10,855
FP	2,555	1,843	1,450	399
TN	12,018	12,730	13,123	14,174
FN	39	92	140	548
prevalence	0.4390	0.4390	0.4390	0.4390
precision	0.8164	0.8599	0.8859	0.9645
recall	0.9966	0.9919	0.9877	0.9519
specificity	0.8247	0.8735	0.9005	0.9726
f1	0.8976	0.9212	0.9341	0.9582

At pt = 0.05, the system achieves very high recall (0.9966), precision of 0.8164, specificity of 0.8247, and F1 score of 0.8976, emphasizing sensitivity for proactive screening. Increasing the threshold improves precision and specificity with only a modest loss in sensitivity: at pt = 0.10, precision reaches 0.8599, recall is 0.9919, specificity is 0.8735, and F1 is 0.9212; at pt = 0.15, precision is 0.8859, recall is 0.9877, specificity is 0.9005, and F1 is 0.9341. These patterns align with operational use, where low-to-moderate thresholds identify the most at-risk cases while controlling false alerts. The Youden high-confidence point (pt ≈ 0.42) yields precision of 0.9645, recall of 0.9519, specificity of 0.9726, and F1 of 0.9582; although outside the screening window, it is suitable for confirmatory or resource-intensive actions. The observed positive-class prevalence on test data (0.4390; n = 25,976) aligns with Section 4.1 and supports discrimination and decision-analytic evaluation within this window.

4.4 Decision-curve analysis (held-out test)

We assessed operational utility using decision-curve analysis (DCA) over the prespecified policy window pt ∈ [0.05, 0.15]. The calibrated LightGBM model achieves a positive net benefit that is consistently above both reference strategies, treat-none (NB = 0) and treat-all, throughout this window, indicating value beyond naïve intervention policies.

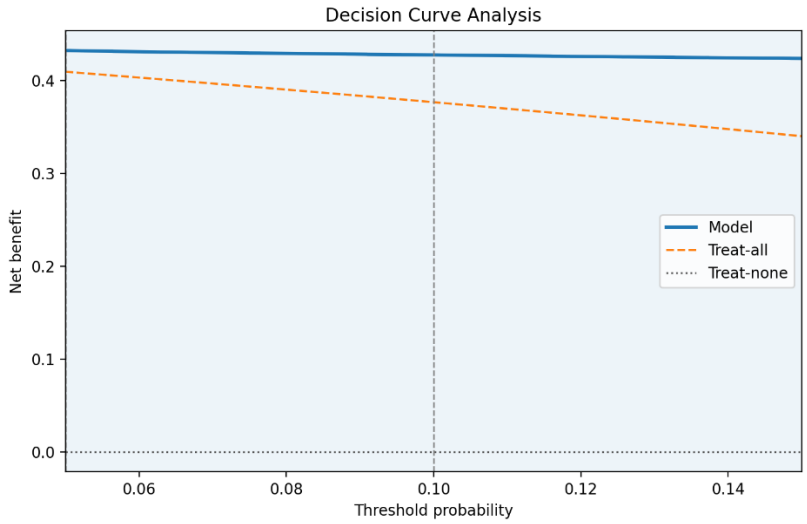

Figure 1. Decision-curve analysis on the held-out test split (LightGBM).

Model net benefit is compared with treat-none and treat-all across threshold probabilities; the shaded band denotes the policy window 0.05–0.15, and the vertical guide marks pt = 0.10.

To aid interpretation, we plot the incremental net benefit relative to the treat-none (ΔNB = NB_model − NB_none). Within pt ∈ [0.05, 0.15], ΔNB ranges approximately 0.424–0.433 with a gentle downward slope, implying that managers can choose thresholds within the window without materially eroding decision value. The preregistered operating point pt = 0.10 lies on the upper envelope of this useful region and is therefore used in Sections 4.5–4.6 for workload translation and segment-level audits.

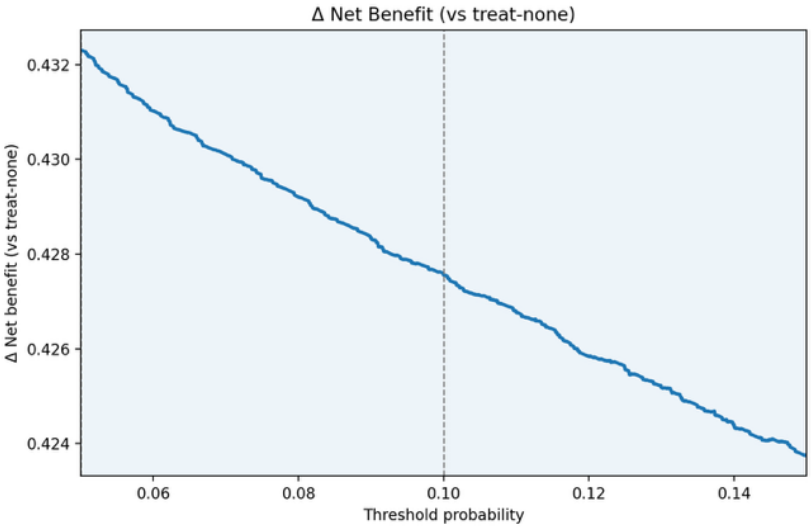

Figure 2. Incremental net benefit relative to treat-none (ΔNB) on the held-out test split; the vertical guide indicates pt = 0.10 and the shaded band denotes the policy window 0.05–0.15.

4.5 Workload translation and alert volumes

To link probabilistic outputs with day-to-day operations, decision thresholds were mapped to expected alert volumes on the held-out test split, focusing on the prespecified policy window pt ∈ [0.05, 0.15]. Within this window, total alerts per day decline monotonically as the threshold increases, whereas the true-positive yield remains nearly constant.

Table 4. Workload translation at representative operating points (alerts/day and decomposition into TP and FP).

threshold	0.05	0.10	0.15
TP	11,364	11,311	11,263
FP	2,555	1,843	1,450
TN	12,018	12,730	13,123
FN	39	92	140
alerts	13,919	13,154	12,713
alerts_per_day	643.0	607.7	587.3
TP_per_day	525.0	522.5	520.3
FP_per_day	118.0	85.1	67.0

At pt = 0.05, the system generates approximately 643.0 alerts/day, comprising about 525.0 true positives and 118.0 false positives. Increasing the threshold to pt = 0.10 reduces the alert load to roughly 607.7 per day (≈522.5 true positives; 85.1 false positives). At pt = 0.15, the alert volume is approximately 587.3 per day, with about 520.3 true positives and 67.0 false positives.

Aggregating across 0.05–0.15, the true-positive volume is stable (≈520–525 per day), while the false-positive volume decreases from roughly 118 to 67 per day, a reduction of about 43%. This pattern offers a tunable trade-off between workload intensity and unwarranted escalations. The preregistered operating point pt = 0.10 is therefore used in Section 4.6 for segment-level audits, as it preserves sensitivity while materially reducing the false-alert burden relative to pt = 0.05.

4.6 Segment-level workload and stability

Segmented workloads were evaluated at the prespecified operating point pt = 0.10 using the operational segments defined in Section 3.3: cabin class (economy, premium-economy, business), distance band (short/medium/long), delay band (low/mid/high), and travel type (business, leisure). Figure. 3 summarizes alerts per day by cabin × distance × travel; the departure-delay bands are marginalized to show throughput patterns across the highest-leverage segments while preserving consistency with the segmentation schema in Section 3.3.

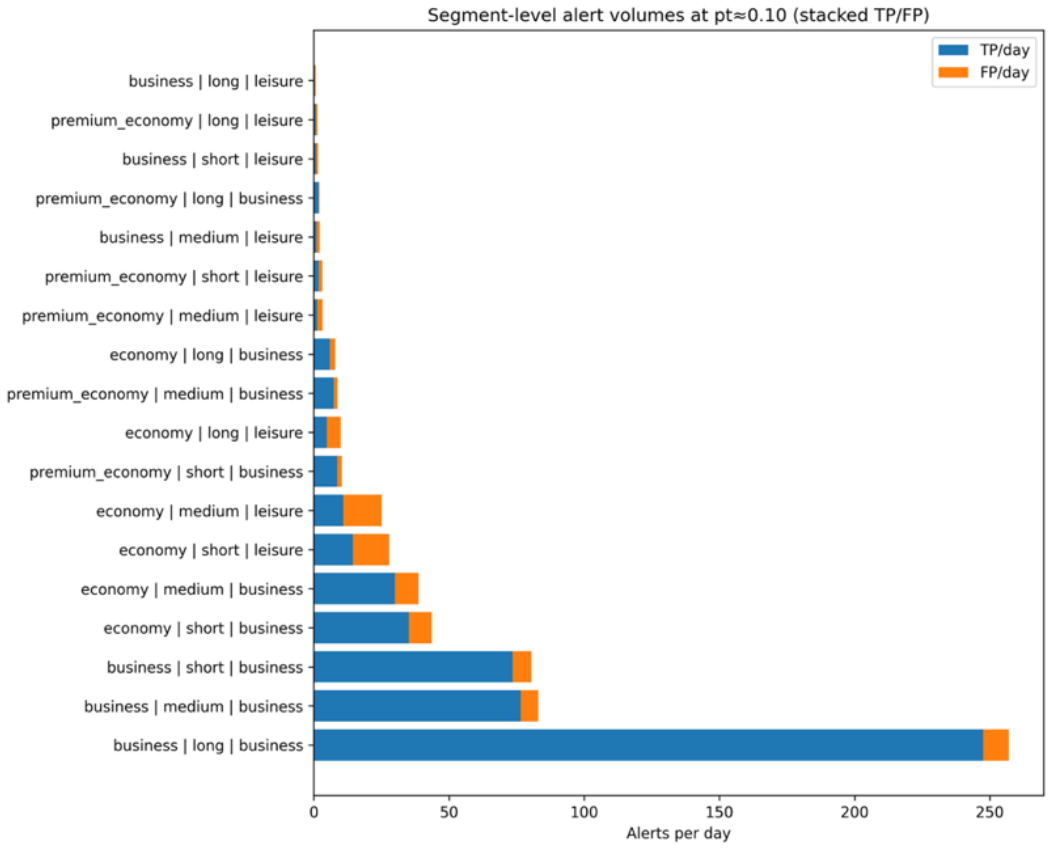

Figure 3. Segment-level alert volumes at pt ≈ 0.10. Bars show alerts per day by cabin class, distance band, and travel type (departure-delay bands marginalized); stacked components indicate true-positive and false-positive volumes.

Three observations are salient. First, business-cabin segments concentrate the largest throughputs with consistently high precision. Long-haul business travel generates the highest workload (≈ 250 alerts/day), with most alerts being true positives, indicating efficient alerting where volumes are greatest. Medium- and short-haul business segments contribute substantially fewer alerts (≈ 80–85 alerts/day) with similarly modest false-positive shares. Second, economy-class performance varies with distance and purpose. Economy–short (business) shows comparatively efficient alerting (≈ 40–45 alerts/day), whereas economy–medium (leisure) produces lower throughput with a higher false-positive share, suggesting candidates for threshold refinement or post-alert triage. Economy–long segments show low absolute volumes with mixed precision. Third, premium-economy segments are generally low-volume but heterogeneous: some business segments remain precise, whereas several leisure segments show elevated false-positive shares, warranting targeted governance (e.g., segment-specific thresholds or recalibration audits). Overall, pt ≈ 0.10 preserves stable true-positive throughput in the highest-volume segments while revealing pockets, notably economy–medium (leisure) and selected

premium-economy (leisure), where modest adjustments could reduce unwarranted escalations without materially diminishing coverage.

4.7 Calibration quality

We calibrated predicted probabilities using isotonic regression fit on the validation split, then applied the fixed calibrator to the held-out test split (no refitting). Probability quality was assessed with reliability curves based on ten uniform bins with shared edges for both splits, the expected calibration error (ECE; L1), and the Brier score. Table 5 reports ECE and Brier scores pre- and post-calibration.

Table 5. Calibration metrics (validation and held-out test).

split	n_bins	ece_pre_L1	ece_post_L1	delta_ece	brier_pre	brier_post
validation	10	0.0066	0.0000	-0.0066	n/a	n/a
test	10	0.0043	0.0037	-0.0006	0.0257	0.0258

With uniform bins, the post-calibration reliability curves for validation and test closely track the identity line across the full probability range. ECE decreases after calibration on both splits, indicating closer agreement between predicted probabilities and observed frequencies. On the test split, the Brier score remains essentially unchanged, consistent with calibration primarily improving probability alignment rather than altering discrimination. Empirical cumulative distribution functions (ECDFs) further show that the predicted probabilities are concentrated near 0 and 1, which explains why calibration yields modest but systematic gains in loss-based metrics.

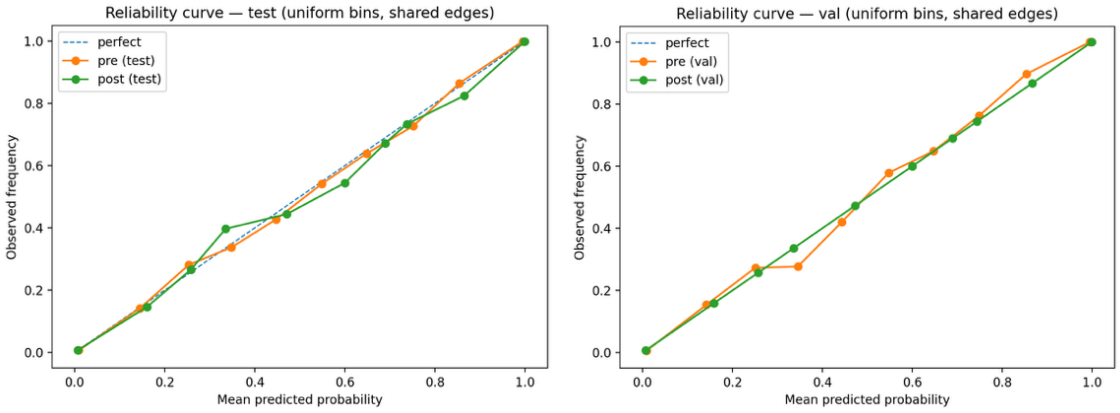

Figure 4. Reliability curves with uniform bins on validation and test after isotonic calibration; dashed line indicates perfect calibration.

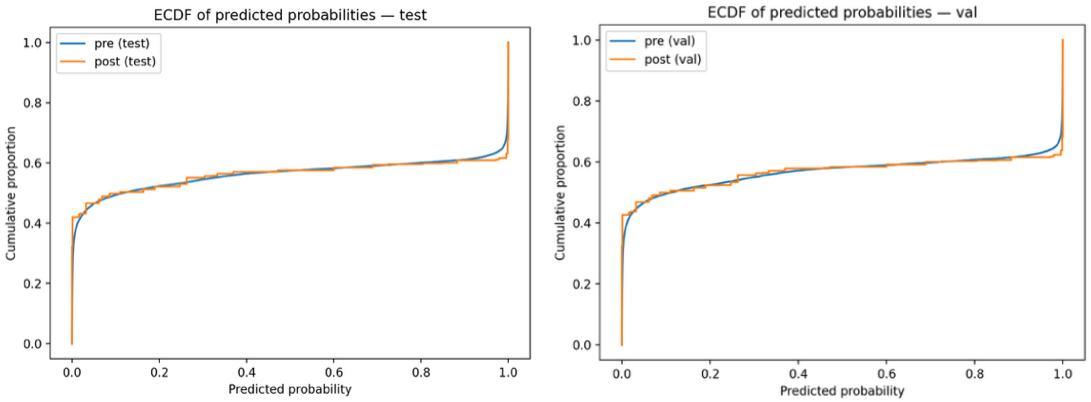

Figure 5. Empirical cumulative distribution functions of predicted probabilities before and after calibration on validation and test.

5. Discussion

This study examined whether calibrated machine-learning (ML) models can provide decision-making value for airline passenger service operations within a predefined screening window (threshold probabilities 0.05–0.15). Three key findings emerge. First, after isotonic calibration, the chosen gradient-boosting model consistently achieved positive net benefit across the policy window on the test set and outperformed the treat-none and treat-all strategies. This suggests that probability reliability, not just ranking performance, is enough to support proactive service-recovery screening without generating excessive false alerts. Second, converting net benefit into alerts per day and true and false positives per day revealed a notable trade-off: moving from 0.05 to 0.15 maintained nearly the same true-positive throughput while reducing the false-positive workload by about 40%. This tunable aspect is actionable for queue and staffing managers facing daily constraints. Third, segment-level audits showed stable coverage in high-volume business-cabin flows but identified pockets—particularly economy-medium leisure and some premium-economy leisure segments—where false-positive shares were relatively high, indicating the value of segment-specific operating points or post-alert triage.

Taken together, these results demonstrate that calibrated probabilities help close the gap between predictive modeling and operational control. In many transportation applications, thresholds are still chosen heuristically (e.g., 0.50), which can obscure the implications for workload and utility. In contrast, decision-curve analysis (DCA) explicitly connects threshold choice to the relative costs of false positives versus false negatives and shows where model-guided actions provide more benefit than harm. When combined with workload translation, managers can select thresholds that meet staffing limits while maintaining sensitivity to at-risk passengers. In this study, the nearly flat net-benefit curve within 0.05–0.15 suggests a forgiving operating region: small threshold adjustments can handle daily capacity fluctuations with minimal impact on decision value.

The calibration results also clarify why the decision value materialized. Reliability curves and small post-calibration ECE on validation and test indicate that predicted risks aligned with empirical frequencies across most of the probability range. Under such conditions, the Murphy decomposition and Brier score signal that residual loss is dominated by irreducible uncertainty rather than systematic miscalibration. The operational implication is straightforward: when a predicted dissatisfaction risk is, say, 10%, roughly one in ten similar future cases will indeed be dissatisfied. This frequency-faithful property is precisely what threshold-based workflows (screen-or-not, escalate-or-not) require planning capacity with confidence.

Segmented workload analysis advances governance. Passenger services are heterogeneous by cabin, distance, delay conditions, and travel purpose. A single global threshold can mask unequal error burdens, potentially leading to perceived inequities (e.g., more frequent unwarranted escalations for some segments). The audit here suggests two governance levers: (i) modest segment-specific thresholds in identified pockets to cap false-positive shares without materially eroding true-positive yield; and (ii) lightweight post-alert triage rules (e.g., secondary checks using readily available operational cues) to filter discretionary escalations. Because the model's overall net benefit was stable across the window, these levers can be deployed without destabilizing system-level performance.

From a transport-systems perspective, the contribution is twofold. Methodologically, the study demonstrates how leakage-controlled training, post-hoc isotonic calibration, decision-curve analysis, and workload translation form an integrated evaluation that goes beyond discrimination. Operationally, the pipeline ties model outputs to staffing and queue management through interpretable units (alerts/day, TP/FP/day) and supports resilience, as thresholds can be revisited when demand, crew availability, or disruption patterns shift. The framework also interfaces naturally with safety and reliability objectives: positive net benefit within the chosen window implies fewer missed opportunities for early service recovery at a controlled false-alert burden.

Several limitations merit discussion. First, the analysis used one institution's historical records with a fixed nominal daily volume for workload conversion; while volumes scale linearly, external validation on other routes, seasons, and airports is necessary before broad deployment. Second, the cost ratio implicit in the DCA threshold was not monetized; the 0.05–0.15 window reflects operational practice rather than an explicit utility model. Future work could elicit context-specific cost/benefit parameters and compare DCA-selected thresholds with those from cost-minimizing decision analysis. Third, calibration was performed once (validation-trained isotonic mapping applied to test). In production, drift monitoring with periodic recalibration (e.g., intercept/slope updates or re-fitting isotonic on fresh non-test data) is recommended, together with auditable threshold logs. Fourth, segment-level findings, while operationally informative,

may reflect varying sample sizes; intervals around segment metrics should be tracked, and very small cells should be pooled. Finally, we did not evaluate uncertainty for net-benefit curves (e.g., confidence bands); adding inference for DCA would quantify the stability of strategy dominance over time.

Two extensions appear especially promising. The first is adaptive thresholding: coupling the near-flat net-benefit region with daily capacity targets to compute thresholds that maintain a desired alert budget (e.g., via a small feedback controller using recent alert counts). The second is governance dashboards: embedding reliability plots, DCA, and segment workload panels into an operations console so that managers can jointly review calibration, decision value, and workload, facilitating transparent, data-driven threshold updates and proactive staffing plans.

In summary, calibrated ML, combined with decision-analytic evaluation, provided actionable, governance-ready decision support for passenger service operations. The approach balanced sensitivity to at-risk passengers with manageable false-alert workload, identified segment pockets for refinement, and offered a practical path to sustained, resilient queue and staffing control.

6. Conclusions

This study demonstrates that well-calibrated ML models can deliver consistent decision value for airline passenger-service operations within an operationally justified screening window. On a held-out test set, the calibrated gradient-boosting model achieved positive net benefit across 0.05–0.15 and, when translated to workload, maintained stable true-positive throughput while reducing false-positive alerts as thresholds increased. Segment-level audits highlighted where modest segment-specific thresholds or post-alert triage can curb unwarranted escalations without materially sacrificing coverage.

Contributions. (i) An integrated, leakage-controlled pipeline linking calibration (Brier/Murphy/ECE) to decision-curve analysis and workload translation; (ii) an operational recipe to choose thresholds in interpretable units (alerts/day, TP/FP/day) for queue and staffing decisions; and (iii) a governance lens via segment audits to promote fair and stable service levels.

Limitations and impacts. Results derive from one institutional dataset and a fixed nominal daily volume; cost ratios were not monetized, and DCA inference was not reported. Despite these limits, the framework advances safety, efficiency, reliability, and resilience by enabling proactive, auditable service-recovery screening with controllable workload.

Future work. Priorities include external validation across airports/seasons, drift-aware recalibration and adaptive thresholding tied to daily capacity, monetized utility analyses, and governance dashboards that integrate calibration, DCA, and workload panels for routine managerial review.

Abbreviations

The following abbreviations are used in this manuscript:

ACI	Airports Council International
AI	Artificial Intelligence
AUC	Area Under the Curve
DCA	Decision Curve Analysis
ECE	Expected Calibration Error
F1	F1-score (harmonic mean of precision and recall)
FN	False Negative
FP	False Positive
ICAO	International Civil Aviation Organization
IRB	Institutional Review Board
ML	Machine Learning
NB	Net Benefit
PR-AUC	Area Under the Precision–Recall Curve
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the ROC Curve
SD	Standard Deviation
SHAP	Shapley Additive Explanations
TN	True Negative
TP	True Positive

Data Availability and Third-Party Terms: This study uses de-identified secondary records on airline passenger service operations (customer experience and operational context). Owing to third-party licensing and confidentiality constraints, the raw data cannot be redistributed. Our Zenodo deposit includes retrieval scripts and exact commands to obtain the original data from its source (or, where permitted, a functionally equivalent public release) and to reproduce all derived artifacts (processed tables, figures, and model outputs). Any use of third-party data must comply with the original licenses and terms.

Ethics Approval and Consent to Participate: This study used publicly available, de-identified secondary data. According to institutional policy, it does not constitute human-subjects research and does not require IRB approval or informed consent.

Acknowledgments: This paper was financially supported by Mahasarakham Business School, Mahasarakham University, Thailand.

Author Contributions: Thanyaporn Sukdet (TS): Conceptualization; Methodology; Software; Formal analysis; Data curation; Visualization; Writing-original draft.

Warawut Narkbunnum (WN): Conceptualization; Methodology; Validation; Supervision; Project administration; Funding acquisition; Resources; Writing-review and editing.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M.H. Kim, J.W. Park, Y.J. Choi, A Study on the Effects of Waiting Time for Airport Security Screening Service on Passengers' Emotional Responses and Airport Image, Sustainability 12 (2020), 10634. https://doi.org/10.3390/su122410634.

[2] ACI World, Joint ACI World-ICAO Passenger Traffic Report, Trends, and Outlook, Advisory Bulletin, Airports Council International, Montreal, 28 January 2025. https://aci.aero/2025/01/28/joint-aci-world-icao-passenger-traffic-report-trends-and-outlook/.

[3] K. Abdulaziz Alnowibet, A. Khireldin, M. Abdelawwad, A. Wagdy Mohamed, Airport Terminal Building Capacity Evaluation Using Queuing System, Alexandria Eng. J. 61 (2022), 10109–10118. https://doi.org/10.1016/j.aej.2022.03.055.

[4] Z.A. Marshall, J.H. Mott, A.J. Gottwald, C.A. Patrick, L.R.I. Dy, Expediting Airport Security Queues Through Advanced Lane Assignment, J. Transp. Secur. 15 (2022), 245–262. https://doi.org/10.1007/s12198-022-00247-9.

[5] EUROCONTROL, All-Causes Delays to Air Transport in Europe – Annual 2023, CODA Digest, EUROCONTROL, 16 December 2024. https://www.eurocontrol.int/publication/all-causes-delays-air-transport-europe-annual-2023.

[6] P. Fleurquin, J.J. Ramasco, V.M. Eguiluz, Systemic Delay Propagation in the US Airport Network, Sci. Rep. 3 (2013), 1159. https://doi.org/10.1038/srep01159.

[7] A. Brun, E. Feron, S. Alam, D. Delahaye, Schedule Optimization and Staff Allocation for Airport Security Checkpoints Using Guided Simulated Annealing and Integer Linear Programming, J. Air Transp. Manag. 124 (2025), 102746. https://doi.org/10.1016/j.jairtraman.2025.102746.

[8] M. Dai, A Hybrid Machine Learning-Based Model for Predicting Flight Delay through Aviation Big Data, Sci. Rep. 14 (2024), 4603. https://doi.org/10.1038/s41598-024-55217-z.

[9] Z. Guo, B. Yu, M. Hao, W. Wang, Y. Jiang, F. Zong, A Novel Hybrid Method for Flight Departure Delay Prediction Using Random Forest Regression and Maximal Information Coefficient, Aerosp. Sci. Technol. 116 (2021), 106822. https://doi.org/10.1016/j.ast.2021.106822.

[10] D. Bala Bisandu, I. Moulitsas, Prediction of Flight Delay Using Deep Operator Network with Gradient-Mayfly Optimisation Algorithm, Expert Syst. Appl. 247 (2024), 123306. https://doi.org/10.1016/j.eswa.2024.123306.

[11] R. Murugesan, R. A P, N. N, R. Balanathan, Forecasting Airline Passengers' Satisfaction Based on Sentiments and Ratings: An Application of VADER and Machine Learning Techniques, J. Air Transp. Manag. 120 (2024), 102668. https://doi.org/10.1016/j.jairtraman.2024.102668.

[12] H. Mirzahossein, S. Rezashoar, Feature Importance Analysis of Optimized Machine Learning Modeling for Predicting Customers Satisfaction at the United States Airlines, Mach. Learn. Appl. 22 (2025), 100734. https://doi.org/10.1016/j.mlwa.2025.100734.

[13] S. Moosavi, R. Ramnath, Context-Aware Driver Risk Prediction with Telematics Data, Accid. Anal. Prev. 192 (2023), 107269. https://doi.org/10.1016/j.aap.2023.107269.

[14] J. Boylan, D. Meyer, W.S. Chen, A Systematic Review of the Use of In-Vehicle Telematics in Monitoring Driving Behaviours, Accid. Anal. Prev. 199 (2024), 107519. https://doi.org/10.1016/j.aap.2024.107519.

[15] H. Mao, F. Guo, X. Deng, Z.R. Doerzaph, Decision-Adjusted Driver Risk Predictive Models Using Kinematics Information, Accid. Anal. Prev. 156 (2021), 106088. https://doi.org/10.1016/j.aap.2021.106088.

[16] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, P. Flach, Classifier Calibration: A Survey on How to Assess and Improve Predicted Class Probabilities, Mach. Learn. 112 (2023), 3211–3260. https://doi.org/10.1007/s10994-023-06336-7.

[17] M. Börjesson, I. Rubensson, Satisfaction with Crowding and Other Attributes in Public Transport, Transp. Policy 79 (2019), 213–222. https://doi.org/10.1016/j.tranpol.2019.05.010.

[18] J. Soza-Parra, S. Raveau, J.C. Muñoz, O. Cats, The Underlying Effect of Public Transport Reliability on Users' Satisfaction, Transp. Res. Part A Policy Pract. 126 (2019), 83–93. https://doi.org/10.1016/j.tra.2019.06.004.

[19] M. Alfarhood, R. Alotaibi, B. Abdulrahim, A. Einieh, M. Almousa, A. Alkhanifer, Predicting Flight Delays with Machine Learning: A Case Study from Saudi Arabian Airlines, Int. J. Aerosp. Eng. 2024 (2024), 3385463. https://doi.org/10.1155/2024/3385463.

[20] A.J. Vickers, E.B. Elkin, Decision Curve Analysis: A Novel Method for Evaluating Prediction Models, Med. Decis. Mak. 26 (2006), 565–574. https://doi.org/10.1177/0272989X06295361.

[21] A.J. Vickers, B. van Calster, E.W. Steyerberg, A Simple, Step-by-Step Guide to Interpreting Decision Curve Analysis, Diagn. Progn. Res. 3 (2019), 18. https://doi.org/10.1186/s41512-019-0064-7.

[22] M. Hardt, E. Price, N. Srebro, Equality of Opportunity in Supervised Learning, in: Advances in Neural Information Processing Systems 29 (NIPS 2016), Curran Associates, 2016, pp. 3315–3323. https://doi.org/10.48550/arXiv.1610.02413.

[23] X. Zhang, Q. Ke, X. Zhao, Travel Demand Forecasting: A Fair AI Approach, IEEE Trans. Intell. Transp. Syst. 25 (2024), 14611–14627. https://doi.org/10.1109/TITS.2024.3395061.

[24] Y. Zheng, S. Wang, J. Zhao, Equality of Opportunity in Travel Behavior Prediction with Deep Neural Networks and Discrete Choice Models, Transp. Res. Part C Emerg. Technol. 132 (2021), 103410. https://doi.org/10.1016/j.trc.2021.103410.

[25] G.W. Brier, Verification of Forecasts Expressed in Terms of Probability, Mon. Weather Rev. 78 (1950), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[26] D.B. Stephenson, C.A.S. Coelho, I.T. Jolliffe, Two Extra Components in the Brier Score Decomposition, Weather Forecast. 23 (2008), 752–757. https://doi.org/10.1175/2007WAF2006116.1.

[27] S. Siegert, Simplifying and Generalising Murphy's Brier Score Decomposition, Q. J. R. Meteorol. Soc. 143 (2017), 1178–1183. https://doi.org/10.1002/qj.2985.

[28] A.H. Murphy, A New Vector Partition of the Probability Score, J. Appl. Meteorol. 12 (1973), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

[29] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html.

[30] A. Niculescu-Mizil, R. Caruana, Predicting Good Probabilities with Supervised Learning, in: Proceedings of the 22nd International Conference on Machine Learning – ICML '05, ACM Press, New York, 2005, pp. 625–632. https://doi.org/10.1145/1102351.1102430.

[31] G. Gao, S. Meng, M.V. Wüthrich, What Can We Learn from Telematics Car Driving Data: A Survey, Insur. Math. Econ. 104 (2022), 185–199. https://doi.org/10.1016/j.insmatheco.2022.02.004.

[32] M.R. Islam, D. Wang, M. Abdel-Aty, Calibrated Confidence Learning for Large-Scale Real-Time Crash and Severity Prediction, npj Sustain. Mobil. Transp. 1 (2024), 1. https://doi.org/10.1038/s44333-024-00001-9.

[33] J. Wei, Y. Miyazaki, F. Sato, Pre-Crash Injury Risk Prediction with Guaranteed Confidence Level: A Conformal and Interpretable Framework, Traffic Inj. Prev. 27 (2026), 583–593. https://doi.org/10.1080/15389588.2025.2538725.

[34] M. Sarhani, A. Nourmohammadzadeh, S. Voß, M. El Amrani, Predicting and Analyzing Ferry Transit Delays Using Open Data and Machine Learning, J. Public Transp. 27 (2025), 100124. https://doi.org/10.1016/j.jpubtr.2025.100124.

[35] E. Ruiz, W.F. Yushimito, L. Aburto, R. de la Cruz, Predicting Passenger Satisfaction in Public Transportation Using Machine Learning Models, Transp. Res. Part A Policy Pract. 181 (2024), 103995. https://doi.org/10.1016/j.tra.2024.103995.

[36] F. Rocco di Torrepadula, E.V. Napolitano, S. Di Martino, N. Mazzocca, Machine Learning for Public Transportation Demand Prediction: A Systematic Literature Review, Eng. Appl. Artif. Intell. 137 (2024), 109166. https://doi.org/10.1016/j.engappai.2024.109166.

[37] M. Attioui, M. Lahby, Congestion Forecasting Using Machine Learning Techniques: A Systematic Review, Future Transp. 5 (2025), 76. https://doi.org/10.3390/futuretransp5030076.

[38] J. Liang, J. Ke, S. Feng, H. Yang, Machine Learning for On-Demand Ride Services: A Survey, Artif. Intell. Transp. 2 (2025), 100008. https://doi.org/10.1016/j.ait.2025.100008.

[39] T. Peng, J. Gao, O. Cats, Uncertainty-Aware Probabilistic Travel Demand Prediction for Mobility-on-Demand Services, Transp. Res. Part C Emerg. Technol. 181 (2025), 105383. https://doi.org/10.1016/j.trc.2025.105383.

[40] J. Yang, L. Chen, Z. Su, W. Ma, Z. Zou, K. An, Decision-Focused Learning for Optimal Subsidy Allocation in Ride-Hailing Services, Transp. Res. Part C Emerg. Technol. 180 (2025), 105301. https://doi.org/10.1016/j.trc.2025.105301.

[41] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, in: 6th International Conference on Learning Representations (ICLR 2018), 2018. https://doi.org/10.48550/arXiv.1707.01926.

[42] B. Yu, H. Yin, Z. Zhu, Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, California, 2018, pp. 3634–3640. https://doi.org/10.24963/ijcai.2018/505.

[43] T. Gneiting, A.E. Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc. 102 (2007), 359–378. https://doi.org/10.1198/016214506000001437.

[44] J. Błasiok, P. Nakkiran, Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing, in: 12th International Conference on Learning Representations (ICLR 2024), 2024. https://doi.org/10.48550/arXiv.2309.12236.

[45] Z. Jiang, A. Liu, B. Van Durme, Addressing the Binning Problem in Calibration Assessment through Scalar Annotations, Trans. Assoc. Comput. Linguist. 12 (2024), 120–136. https://doi.org/10.1162/tacl_a_00636.

[46] G.S. Collins, P. Dhiman, J. Ma, M.M. Schlussel, L. Archer, B. Van Calster, F.E. Harrell, G.P. Martin, K.G.M. Moons, M. van Smeden, M. Sperrin, G.S. Bullock, R.D. Riley, Evaluation of Clinical Prediction Models (Part 1): From Development to External Validation, BMJ 384 (2024), e074819. https://doi.org/10.1136/bmj-2023-074819.

[47] B. Van Calster, E.W. Steyerberg, L. Wynants, M. van Smeden, There Is No Such Thing as a Validated Prediction Model, BMC Med. 21 (2023), 70. https://doi.org/10.1186/s12916-023-02779-w.

[48] J. Bröcker, L.A. Smith, Increasing the Reliability of Reliability Diagrams, Weather Forecast. 22 (2007), 651–661. https://doi.org/10.1175/WAF993.1.

[49] A.N. Angelopoulos, S. Bates, Conformal Prediction: A Gentle Introduction, Found. Trends Mach. Learn. 16 (2023), 494–591. https://doi.org/10.1561/2200000101.

[50] A.J. Vickers, B. Van Claster, L. Wynants, E.W. Steyerberg, Decision Curve Analysis: Confidence Intervals and Hypothesis Testing for Net Benefit, Diagn. Progn. Res. 7 (2023), 11. https://doi.org/10.1186/s41512-023-00148-y.

[51] G.N.F. Cruz, K. Korthauer, Bayesian Decision Curve Analysis with bayesDCA, Stat. Med. 43 (2024), 6042–6058. https://doi.org/10.1002/sim.10277.

[52] X. Chen, S. Bai, Y. Wei, H. Jiang, How Income Satisfaction Impacts Driver Engagement Dynamics in Ride-Hailing Services, Transp. Res. Part C Emerg. Technol. 157 (2023), 104418. https://doi.org/10.1016/j.trc.2023.104418.

[53] N. Xie, W. Tang, J. Zhu, J. Li, X. Chen, Understanding Causal Effects of Ride-Sourcing Subsidy: A Novel Generative Adversarial Networks Approach, Transp. Res. Part C Emerg. Technol. 157 (2023), 104371. https://doi.org/10.1016/j.trc.2023.104371.

[54] P. Ashkrof, G.H. de Almeida Correia, O. Cats, B. van Arem, On the Relocation Behavior of Ride-Sourcing Drivers, Transp. Lett. 16 (2024), 330–337. https://doi.org/10.1080/19427867.2023.2192581.

[55] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera, A Unifying View on Dataset Shift in Classification, Pattern Recognit. 45 (2012), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019.

[56] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A Survey on Concept Drift Adaptation, ACM Comput. Surv. 46 (2014), 1–37. https://doi.org/10.1145/2523813.

[57] Q. Hou, Y. Yang, J. Liang, X. Huo, J. Leng, A Deep Transfer Learning Approach for Real-Time Traffic Conflict Prediction with Trajectory Data, Accid. Anal. Prev. 214 (2025), 107966. https://doi.org/10.1016/j.aap.2025.107966.

[58] X. Zou, E. Chung, Traffic Prediction via Clustering and Deep Transfer Learning with Limited Data, Comput.-Aided Civ. Infrastruct. Eng. 39 (2024), 2683–2700. https://doi.org/10.1111/mice.13207.

[59] J. Shao, S. Li, K. Zhang, A. Wang, M. Li, Cross-City Traffic Prediction Based on Deep Domain Adaptive Transfer Learning, Transp. Res. Part C Emerg. Technol. 176 (2025), 105152. https://doi.org/10.1016/j.trc.2025.105152.

[60] E.W. Steyerberg, Y. Vergouwe, Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation, Eur. Heart J. 35 (2014), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207.

[61] J. Xu, Y. Xiao, W.H. Wang, Y. Ning, E.A. Shenkman, J. Bian, F. Wang, Algorithmic Fairness in Computational Medicine, EBioMedicine 84 (2022), 104250. https://doi.org/10.1016/j.ebiom.2022.104250.

[62] B. Bharti, P. Yi, J. Sulam, Estimating and Controlling for Equalized Odds via Sensitive Attribute Predictors, Adv. Neural Inf. Process. Syst. 36 (2023), 37173–37192. https://pmc.ncbi.nlm.nih.gov/articles/PMC11167624/.

[63] E. Small, K. Sokol, D. Manning, F.D. Salim, J. Chan, Equalised Odds Is Not Equal Individual Odds: Post-Processing for Group and Individual Fairness, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), ACM, New York, 2024, pp. 1559–1578. https://doi.org/10.1145/3630106.3658989.

[64] Y. Liang, Fairness-Aware Dynamic Ride-Hailing Matching Based on Reinforcement Learning, Electronics 13 (2024), 775. https://doi.org/10.3390/electronics13040775.

[65] T.W. Sanchez, Y. Qian, X. Yan, AI Applications in Transportation and Equity: A Survey of U.S. Transportation Professionals, Future Transp. 4 (2024), 1161–1176. https://doi.org/10.3390/futuretransp4040056.

[66] B.C. Cheong, Transparency and Accountability in AI Systems: Safeguarding Wellbeing in the Age of Algorithmic Decision-Making, Front. Hum. Dyn. 6 (2024), 1421273. https://doi.org/10.3389/fhumd.2024.1421273.

[67] K. Hamad, E. Alotaibi, W. Zeiada, G. Al-Khateeb, S. Abu Dabous, M. Omar, B.R.K. Mantha, M.G. Arab, T. Merabtene, Explainable Artificial Intelligence Visions on Incident Duration Using eXtreme Gradient Boosting and SHapley Additive exPlanations, Multimodal Transp. 4 (2025), 100209. https://doi.org/10.1016/j.multra.2025.100209.

[68] L. Obaid, K. Hamad, S. Barakat, Incident Duration Reliability Assessment Using Monte-Carlo Simulation and Kernel Density Estimation of Machine Learning-Based Models, Int. J. Transp. Sci. Technol. 20 (2025), 157–177. https://doi.org/10.1016/j.ijtst.2024.11.005.

[69] C. Oprea, M. Rosca, E. Rosca, I. Costea, A. Ilie, O. Dinu, A. Ruscă, Analyzing Passenger Flows in an Airport Terminal: A Discrete Simulation Model, Computation 12 (2024), 223. https://doi.org/10.3390/computation12110223.

[70] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, G. Salanti, Developing Clinical Prediction Models: A Step-by-Step Guide, BMJ 386 (2024), e078276. https://doi.org/10.1136/bmj-2023-078276.

[71] T. Dong, S. Sinha, B. Zhai, D. Fudulu, J. Chan, P. Narayan, A. Judge, M. Caputo, A. Dimagli, U. Benedetto, G.D. Angelini, Performance Drift in Machine Learning Models for Cardiac Surgery Risk Prediction: Retrospective Analysis, JMIRx Med 5 (2024), e45973. https://doi.org/10.2196/45973.

[72] F. Hinder, V. Vaquet, B. Hammer, One or Two Things We Know About Concept Drift—A Survey on Monitoring in Evolving Environments. Part A: Detecting Concept Drift, Front. Artif. Intell. 7 (2024), 1330257. https://doi.org/10.3389/frai.2024.1330257.