CatBoost-based context-aware purchase-decision prediction on tabular customer journey sequences

Phung Thai Kim [a,1], Ngan Lam Thi Bich [b,2,*], Trinh Le My [c,3], Duy Trinh Le [d,4]

[a] Department of Communications and Partnerships, University of Economics Ho Chi Minh City 700000, Viet Nam
[b] Product Department, VSC Science Technology Communications Group Company Limited, Ho Chi Minh City 700000, Viet Nam
[c] Faculty of Information Technology, UEH Mekong, University of Economics Ho Chi Minh City 700000, Viet Nam
[d] Department of Student Affairs, University of Economics Ho Chi Minh City 700000, Viet Nam
[1] phungthk@ueh.edu.vn; [2] nganlam.pm@gmail.com; [3] trinhlm@ueh.edu.vn; [4] duytl@ueh.edu.vn
* corresponding author

ARTICLE INFO

Article history
Received February 11, 2026
Revised March 22, 2026
Accepted April 23, 2026
Available online August 31, 2026

Keywords
Customer journey
Purchase-decision prediction
Tabular sequences
Context-aware
CatBoost

ABSTRACT

Building on the premise that real-time marketing decisions require models that are both accurate and resource-efficient, our study proposes a context-aware, tabular pipeline using CatBoost with ordered target encoding and greedy feature combinations to learn from customer journey sequences. In feature engineering, we derive period-level aggregates such as touch counts and mean dwell times across daily, weekly, or monthly intervals, compute channel-usage entropy to measure diversity, and quantify recency relative to the most recent interaction, capturing intensity and diversity of user behavior. These sequences are further enriched with context signals such as dwell time, device type, temporal gaps, and channel entropy, along with demographic attributes, to provide a rich representation of customer behavior. We apply the method to the Netherlands Travel dataset (May 2015–Oct 2016), trimming each journey to the first ten and last twenty touchpoints, engineering row-level and period-level aggregates, and collapsing them to the purchase level. CatBoost, a gradient boosting algorithm, leverages ordered target encoding and greedy feature combinations to capture nonlinear interactions while remaining interpretable and CPU-friendly. Stratified five-fold cross-validation yields an AUC of 0.9479, and tuning the F0.5 = 0.869 score produces a threshold of 0.723 with precision 0.897, recall 0.773, and accuracy 0.895. These results demonstrate that structured tabular representations of customer journeys, combined with tree-based learning, can achieve strong predictive performance while maintaining interpretability and computational efficiency, offering a practical modeling framework for decision-support systems in travel marketing.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Personalizing interventions and maximizing conversion in omnichannel environments requires models that not only infer purchase intent from multi-touch journeys (sequences of customer interactions across channels) but also remain practical to operate at scale [1], [2]. Journey modelling began with first-order Markov chains and shallow recurrent models, then advanced to long short-term memory (LSTM) for capturing longer dependencies [3], [4]. With the advent of Transformer architectures, deep learning models using self-attention mechanisms, sequence modelling improved further, enabling the learning of long-range, context-dependent relations [5], [6]. In parallel, recommender system studies show that incorporating context-of-use signals, such as dwell time, device, temporal gaps, and recency, improves discrimination beyond click sequences [7], [8]. As a result,

managerial thinking has shifted from optimizing isolated touchpoints to orchestrating end-to-end journeys across channels and devices [9], [10].

Sequential recommendation (SR) blends long-term preferences with short-horizon contextual dynamics and underpinning large-scale systems in e-commerce, social media, news, and advertising [11], [12]. Models such as SASRec and BERT4Rec, illustrate the transition from RNNs to attention-based architectures for intent modeling [13]. At the same time, peer-to-peer media, consumer reviews, social activity, and word of mouth now rival firm-controlled advertising as sources of brand information, reshaping the requirements for journey analytics [14]. This fragmentation necessitates more sophisticated journey analytics to deliver consistent cross-channel experiences [15], [16], while decision-making remains complicated by experience-based and credence attributes that make quality assessment difficult prior to purchase [17], [18].

While deep sequential recommendation models such as SASRec and BERT4Rec have demonstrated strong performance in large-scale recommendation settings, recent studies have also observed that gradient boosted tree models remain highly competitive when the available data contain heterogeneous contextual features and moderate sequence lengths. In such scenarios, tabular learners can effectively exploit rich categorical and temporal signals without requiring complex sequence architectures or extensive hyperparameter tuning.

Recent literature consistently reports measurable gains from incorporating contextual features, such as pacing and recency, into SR models compared to sequence-only baselines [19], [20]. Meanwhile, tabular learners continue to excel on heterogeneous data when provided with well-designed features and principled categorical handling, which refers to systematic methods for processing categorical data. This offers interpretability and operational economy when compared with deep sequence models [21], [22]. In this context, travel and tourism analytics have recognized cross-device and cross-touchpoint journey prediction as a practical benchmark, highlighting the value of device context and session dynamics for conversion modelling [23], [24]. Building on this, recent tourism-specific frameworks emphasize the necessity of dynamic adaptation to enhance personalized recommendations [25], ensuring that models capture the fluid nature of travel intent while addressing the complexity of multimodal behavioral signals [26]. Such adaptations are critical in high-stakes environments where real-time responsiveness is as vital as predictive accuracy.

Few studies systematically benchmark a CatBoost-style, ordered-time-series (TS) tabular pipeline against deep sequential recommendation (SR) baselines for travel journeys. leveraging journey-aware features such as recency, gaps, entropy, and boundary indicators [27], [28]. They also assess deployment-oriented metrics, such as PR-AUC and calibration, with CPU-bound computational footprints [29]. Despite advances in sequential recommendation using neural models like long short-term memory (LSTM) and Transformer architectures, three substantive gaps remain in the travel journey conversion literature:

Model Class Diversity: While most prior work privileges sequence-centric neural networks, especially in sequential recommendation [6], [11], [30]. There is limited benchmarking of tree ensemble methods, particularly CatBoost, optimized for high-cardinality categorical data in this context. This gap leaves unresolved the question of how tree ensembles compare to deep neural networks in terms of predictive efficacy and interpretability for travel journey conversion. [31]. While deep sequential recommendation models have demonstrated strong performance, the evaluation of machine learning versus deep learning remains a critical benchmark in high-stakes prediction tasks, such as credit risk or purchase propensity [32]. Furthermore, modern architectures have also sought to integrate traditional behavioral frameworks like RFM (Recency, Frequency, Monetary) into neural networks to improve segment-level classification and intent modeling [33].

Contextual Signal Exploitation: Existing research often relegates contextual variables such as dwell time and device information to secondary side information, rather than investigating their direct and interactive impacts on model performance [7], [8], [19]. The interplay between these contextual signals

and demographic segmentation remains insufficiently examined [34], hindering a deeper understanding of their influence on purchase decision-making.

Class Imbalance Mitigation: Marketing conversion datasets are intrinsically skewed. Although oversampling and cost-sensitive losses are well documented [35], few journey-level studies evaluate the joint impact of bootstrap up-sampling and algorithmic class weights on model robustness [36], [37], [38]. Beyond standard oversampling, advanced techniques such as elastic oversampling (EMO) with boundary skeletons have been proposed to improve model robustness on imbalanced data boundaries [26]. Additionally, hybrid optimization algorithms, such as those combining deep learning with swarm intelligence, offer new pathways for stabilizing predictions in skewed datasets [39].

Service purchases progress in stages. Experience and credence attributes hinder pre-purchase evaluation and raise perceived risk. This drives increased information search [1], [40]. The services and marketing literature raise customer experience from isolated encounters to a journey-level construct. This journey spans cognitive, affective, sensory, and social responses across channels [9]. Multichannel and omnichannel scholarship emphasizes cross-touchpoint coordination rather than single-moment optimization in digital commerce [16], [38]. Methodologically, customer-journey mapping consolidates time-ordered interactions into auditable, data-driven artefacts for analysis, mining, and attribution [10], [41]. In practice, consideration sets are shaped by prior experience and brand exposure. Consumers then search, evaluate, and decide to purchase or abstain. This motivates models that respect temporal ordering and context [12], [42].

Customer journey methods involve mapping and analyzing customers' interactions with a service, offering, or brand [43]. The resulting customer journey maps typically highlight key touchpoints and areas for improvement. CJML makes a distinction between objective elements, which can be observed and logged (either manually or automatically), and subjective components, which encompass the highly personal customer experiences [43]. CJML provides a data-driven view that is not biased by people's perceptions of the actual process behavior [41].

Unlike impulsive or habitual purchases, service purchases typically require customers to proceed along different stages of the customer journey. Theoretically, this requirement stems from the experience and credibility qualities that typically dominate services and hinder customers' quality assessments before purchase [18]. During the pre-purchase stage, customers obtain and review brand information and evaluate the available brands; they next decide which service to purchase or whether to forgo purchase completely [17].

A customer journey map (CJM) is a collection of customer journeys, which is the type of data used in this research. Customer journeys capture the customer experience across multiple touchpoints over time. According to [41], the Customer is the focal actor experiencing the journey and consuming the firm's services relative to the firm's market: journey, mapping, goal, touchpoint, and timeline. A journey is the path of interactions, distinguished between the observed (actual) journey and the expected journey specified by stakeholders; mapping refers to the documentation of customer actions and observable events in the log. The goal is the customer's purpose for engaging, and it is useful for detecting detours versus optimal routes. A touchpoint is an interaction between the customer and the firm, and the Timeline is the temporal structure of the journey, with each touchpoint bearing a timestamp that allows strict ordering and duration analysis.

Sequential recommendation (SR) aims to predict the next interaction in a user's behavioral sequence by leveraging historical actions and temporal dynamics. Subsequent advances introduced long short-term memory (LSTM) networks and attention mechanisms culminating in Transformer-based architectures such as SASRec and BERT4Rec, which employ self-attention to integrate recent actions and long-term user preferences for making predictions [5], [6], [16]. Industrial-scale SR systems extend these architectures with real-time streaming inputs and auxiliary learning objectives to maintain throughput under production constraints [19]. However, empirical studies suggest that standard SR assumptions, such as the dominance of short-horizon intent may not generalize across domains where recency, diversity, and entropy influence decision-making differently [12].

This assessment motivates our exploration of context-aware tabular pipelines that systematically encode journey-aware features, while preserving interpretability and operational efficiency. Building on the preceding analysis, our approach seeks to bridge the gap by embedding boundary effects, temporal pacing, and cross-modal interactions within a CatBoost-based framework optimized for heterogeneous data. In doing so, we offer a practical alternative to computationally expensive deep SR architectures. This aligns with recent literature highlighting the importance of SHAP-driven interpretable forecasting to provide transparent managerial levers [44], particularly when analyzing mobile-based behaviors through explainable machine learning frameworks [45]. Such integration allows for optimizing customer retention budgets by aligning predictive accuracy with mathematical optimization [46], even when accounting for complex temporal promotional data and sales fluctuations [47].

Research objectives: (i) predictive performance: Evaluate whether a context-enriched CatBoost model can achieve AUC and precision-leaning F0.5 competitive with deep SR baselines on journey-level purchase prediction; (ii) operational applicability: Demonstrate that the pipeline is CPU-deployable, feature-store compatible, and threshold-tunable for precision-critical campaigns in omnichannel environments; (iii) interpretability & controls: Provide transparent insights boundary dynamics, pacing, device deviations and demographic interactions that yield transparent managerial levers and stable thresholding for campaign budgeting.

Although gradient boosted decision tree models and sequence-to-tabular transformations are widely used in applied machine learning, their systematic application to heterogeneous customer journey data remains relatively underexplored in marketing decision support contexts. In particular, customer journeys often consist of variable-length interaction sequences enriched with contextual attributes such as device type, traffic source, and temporal interaction patterns. These heterogeneous signals are not always naturally accommodated by traditional sequential recommendation architectures, which primarily focus on ordered event sequences

Expected contributions: (i) develop a tabular-sequence pipeline that consolidates variable-length touch streams into purchase-level vectors capturing boundary effects, temporal pacing, diversity and device–demographic cross-effects, compatible with ordered categorical encodings and greedy feature combinations; (ii) provide evidence that context-aware CatBoost delivers competitive discrimination on a real travel dataset while maintaining CPU efficiency suitable for production scoring; (iii) offer a deployable blueprint: data preparation, aggregation, re-balancing with class weights, stratified validation and precision-oriented threshold selection, aligned with journey-orchestration needs in fragmented media environments.

2. Method

2.1. Dataset

The study employs the publicly available Netherlands Travel Customer-Journey dataset covering May 2015 – October 2016. After cleansing (timestamp coercion, removal of out-of-range durations > 720 s) and sub-sampling (first 10 and last 20 interactions per journey), the corpus consists of 134,037 interaction rows across 25,632 purchase journeys. The binary field purchase_any indicates whether a journey culminated in a booking (1) or not (0).

We integrate four artifacts: LogFile Prepped (Excel), LogFile Prepped & Pivoted (Excel), TravelDataJourneys.csv (interaction-level), and TravelDataUsers.csv (user-level). Merges proceed by PurchaseID and UserID, then demographics are standardized. We remove redundant/noisy fields (Region2, Size of Manucipality, Highest education completed, purchase_own, touch point initiator, Number of Children, Age). Timestamps TIMESPSS are coerced to datetime64[ns]; Duration numeric and clipped to [1, 720] s per the data book. The target purchase_any flags 1 = booking, 0 = no booking.

Table 1. Variables (original dtypes as ingested).

No.	Variable	Data type	Brief description
1	UserID	int64	Unique anonymised user identifier
2	Gender	object	1 = Male, 2 = Female
3	Age	int64	Age in completed years
4	Age Range	object	Ordinal marketing age band (e.g., "18-25")
5	Region	object	SPSS _Regio5 code (1 Amsterdam/Rotterdam/DenHaag ... 5 South)
6	Size of Household	int64	Number of persons in household
7	Occupation	object	Employment status (10-level code)
8	Gross income €/year	object	Eight income brackets (1 < €12 900 – 7 > €79 900, and 8 unknown)
9	Number of Children	int64	Count of children in household
10	PurchaseID	int64	Unique identifier for a purchase journey
11	Date	datetime64[ns]	Timestamp of the touchpoint
12	Year	int64	Extracted calendar year
13	Month	object	Extracted calendar month
14	touchpoint type	object	Fine-grained channel label (see Table 2)
15	touch point category	object	Aggregated channel family (Website / App / Search)
16	touch point direction	object	Directionality of interaction (e.g., inbound vs outbound)
17	purchase_any	object	Target: 1 = booking, 0 = no booking
18	DEVICE_TYPE	object	MOBILE vs FIXED (desktop/laptop)
19	Duration	float64	Seconds spent on touchpoint (1–720 s)
20	Social class	float64	Ordinal socio-economic status (1 high ... 5 low)
21	Education (completed)	float64	Highest qualification (1 none ... 8 post-grad)
22	Lifestage	float64	Household life-stage segment (e.g., Young Singles)

Table 1 summarizes the 22 primary columns retained for modelling. We preserve the schema types to ensure correct treatment of categorical vs numerical fields. Table 2 shows 15 touchpoints from 1-10 and 12-16.

Table 2. Fifteen touch-point classes.

No.	Variable
1	Accommodations Website
2	Accommodations App
3	Accommodations Search
4	Information / Comparison Website
5	Information / Comparison App
6	Information / Comparison Search
7	Travel Agent Website Competitor
8	Travel Agent App Competitor
9	Travel Agent Search Competitor
10	Travel Agent Website Focus Brand
12	Travel Agent Search Focus Brand
13	Flight Tickets Website
14	Flight Tickets App
15	Flight Tickets Search
16	Generic Search

All identifiers remain hashed; joins use anonymized keys. The pipeline adheres to GDPR-aligned principles of data minimization and purpose limitation through aggregate, journey-level modelling.

2.2. Purchase-Decision Prediction workflow

Tourism analytics increasingly fuses click-stream with purchase logs to predict conversion and trip purpose, with cross-device identity and cross-touchpoint flows forming the operative unit of analysis [48]. In practice, device context, timing, and pacing introduce significant context heterogeneity across journeys, which must be modelled explicitly to recover segment-specific conversion dynamics [10]. For managerial use, transparent tabular learners, coupled with interpretable feature summaries, surface operational levers, such as boundary channels, pacing, and device tilt, that align with journey diagnostics and service design [42].

Feature families, including boundary indicators (first/last channel), temporal gaps, recency, and diversity/entropy, provide portable, data-efficient surrogates for short-horizon intent that complement side information in recommender pipelines [20], [49]. Evaluation in travel settings benefits from precision-oriented metrics, notably PR-AUC and lift@k, given imbalanced conversion labels and budget-constrained outreach [49], [50]. When journeys are scored for targeting, probability calibration (e.g., Brier score with reliability checks) stabilises thresholding against campaign economics and seasonal prior shifts [51]. Operationally, CPU-friendly gradient-boosted trees enable low-latency scoring, straightforward A/B experimentation, and lightweight MLOps on commodity infrastructure [14]. Common applications include next-best action, channel budget allocation, and journey-stage interventions that tie predictive signals to touchpoint orchestration and campaign governance.

Finally, process-aware monitoring (e.g., drift on recency, gaps in device features, as well as path bottleneck discovery) supports robust deployment and retraining policies in dynamic travel markets [39]. Taken together, these considerations strengthen the case for transparent, context-rich tabular models for travel journeys as an interpretable and deployable alternative to resource-intensive deep sequence stacks.

Fig. 1. shows how the context-aware tabular pipeline for purchase-decision prediction from customer journey logs works in this research. To contextualize the flow that follows, it is useful to briefly outline the major stages of the modelling pipeline. We begin by identifying and integrating relevant data sources and defining a consistent schema to ensure that all subsequent processing steps operate on a coherent foundation. The sequences are then trimmed to preserve temporal fidelity, and detailed row-level temporal context features are engineered to capture the dynamics of each customer journey. These row-level features are aggregated to the purchase level, producing a single representation per transaction. The pipeline next addresses class imbalance and splits the data to maintain representative distributions for training and evaluation. Finally, the learning algorithm is configured with appropriate hyperparameters, and performance metrics are selected to assess the model. This structured approach ensures that each stage builds upon the last, culminating in a robust, interpretable model.

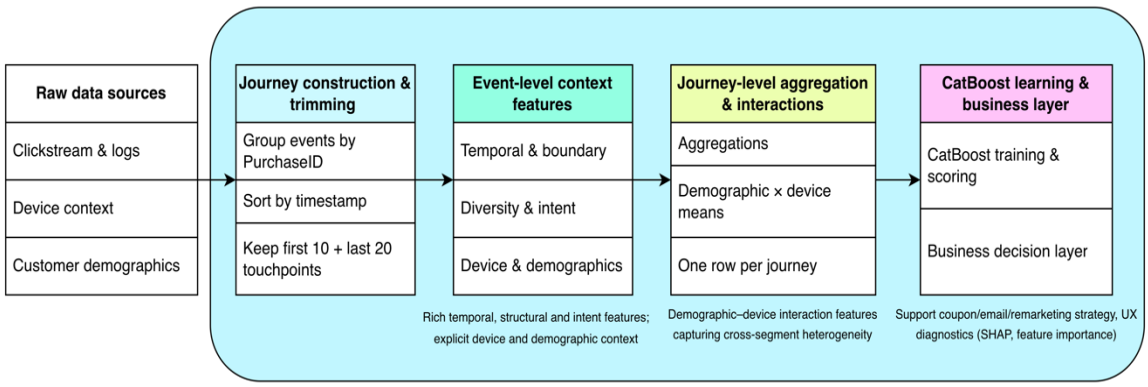


Fig. 1. Context-aware tabular pipeline for purchase-decision prediction from customer journey logs.

2.3. Tabular Customer Journey Sequences for heterogeneous journeys: GBDT and CatBoost

Scalable implementations such as XGBoost and LightGBM further improved efficiency via regularised objectives and histogram-based split finding, enabling fast training and inference on CPUs [52]. CatBoost advances the family with ordered boosting and ordered target statistics for categorical features to mitigate prediction shift and leakage, while building greedy categorical combinations to capture higher-order interactions without exploding dimensionality—well-suited to channel×device×stage effects observed in journeys [14], [37].

CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors. A decision tree recursively partitions the feature space Rm into disjoint regions Rj using binary split attributes a=1(xk>t), where xk is numeric or binary. Each leaf is assigned a value (regression) or class label (classification):

$$h(x) = \sum_{J=1}^{J} b_j 1_{(x \in Rj)} \quad (1)$$

CatBoost also exploits combinations of categorical features as additional features to capture high-order dependencies (e.g., joint information of user ID and ad topic). Because the number of combinations explodes, CatBoost builds them greedily during tree growth and converts them to target statistics (TS) on the fly.

Customer-journey signals that are most informative for purchase propensity tend to concentrate at the sequence boundaries: the head (initiation context such as source and entry channel) and the tail (late-stage intent and commitment cues), while mid-sequence events often exhibit higher redundancy or routine browsing [6], [16]. Retaining a fixed head–tail window therefore preserves the principal decision-relevant variance and reduces dimensionality prior to tabularisation, aligning with the empirically observed recency bias in sequential recommenders and the value of bidirectional context around the decision point [6], [16]. Head–tail explicitly centres the late-stage window and preserves entry conditions in a single, compact subsequence, which is particularly suitable when the deployment objective is precision-oriented scoring on CPU-bound infrastructure

For each journey j with chronologically ordered events $(e_1, \ldots e_{Tj})$, we keep

$$H_j = \{e_1, \ldots, e_{NFIRST}, T_j = \{e_{TJ-nlast+1,\ldots,} e_{Tj}\} \quad (2)$$

and define the retained set $R_j = H_j \cup T_j$ with $(N_{FIRST}, N_{LAST}) = (10,20)$ and a maximum retained length of 30. If $T_j \leq 30$ the journey is kept in full; if $T_j < N_{FIRST} + N_{LAST}$, the sets H_j and T_j overlap, and R_j is simply the union without duplication. This head–tail design explicitly captures initiation cues (e.g., acquisition channel, first device) and late-stage cues (e.g., last channel, compressed gaps), which are known to be discriminative in short-horizon intent modelling [6], [16].

Procedure: Temporal ordering: sort events strictly by timestamp within each PurchaseID; break ties deterministically (e.g., by event ID) and drop exact duplicates to ensure a well-defined order. Then windowing: compute H_j and T_j as above and materialise R_j. Integrity checks: record coverage statistics (min/median/mean/max retained length) and the overlap rate (share of journeys where $|H_j \cap T_j| > 0$). Next, pass-through constraints: enforce that all row-level features (3.3) are computed after trimming, so boundary indicators (e.g., channel_first, channel_last) and pacing metrics (gap_sec) refer to the retained subsequence and remain leakage-safe.

Hyperparameters and sensitivity: the pair (N_{FIRST}, N_{LAST}) represents a capacity–salience trade-off: increasing either raises representational capacity but may reintroduce redundancy; decreasing either risks losing salient context [6]. We recommend a one-shot sensitivity sweep (e.g., (5,10),(10,20),(15,15)) and report (i) OOF AUC and $F_{0.5}$, (ii) PR-curves near the operating region, and (iii) the stability of boundary features (e.g., consistency of channel_last) across settings [16].

The trimming step is linear in journey length and can be fused with sorting in a single pass; we persist the retained rows with a boolean mask (e.g., keep=1) to facilitate auditability and ablation.

Downstream aggregation then operates on R_j to produce purchase-level vectors with boundary, pacing, diversity and device-tilt information intact.

2.4. Encoding high-cardinality categoricals

Categorical variables were encoded using CatBoost's ordered target encoding scheme. This mechanism computes category statistics using only preceding observations within each training fold, thereby preventing information leakage from the validation data during the encoding process.

One-hot encoding is generally avoided for high-cardinality categorical because it inflates dimensionality and dilutes signal under sparsity in clickstream-like telemetry [53]. Naive target-statistic encodings can be problematic when a category is rare or has very few samples (i.e., imbalanced data), the calculated target-statistic can be unreliable and noisy, leading to overfitting [52]. Target-statistic encodings compress categories to outcome-linked statistics but require fold-wise or ordered computation to avoid leakage in non-CatBoost learners, a caveat we retain in ablations [14]. CatBoost operationalizes this safely via ordered TS and feature combinations, which are key for capturing cross-effects such as segment×device or channel×boundary interactions in travel [14].

This study applied two critical Catboost techniques to improve upon standard gradient boosting and target encoding. First, it implements ordered boosting, a permutation-driven alternative to the classic boosting algorithm. This approach trains on a sequence of data permutations to build models that avoid the "prediction shift" caused by target leakage, a problem common in other GBDT implementations [14]. Second, instead of requiring manual creation of feature interactions, CatBoost greedily generates feature combinations during tree construction. These new combinations are then treated as additional categorical features, allowing the model to capture complex interactions efficiently without a combinatorial explosion of features.

We start from the raw event log. Each row has a timestamp, so we break it into day, week, and month components, then compute rolling summaries per journey. For each PurchaseID, we keep track of the number of events per window and the average dwell time in seconds. On top of these basic counts and means, we build a richer set of temporal features. Each event carries its ordinal position in the journey, a flag indicating whether it occurs in the late stage, and the ratio of late touches for that journey. We also compute the gap in seconds between consecutive contacts, record which channel appears first and which appears last, measure channel diversity using Shannon entropy over the channel distribution, and add a recency measure in hours since the last interaction. Where the dataset allows, we construct an “intent ratio” that quantifies the proportion of touches that fall into high-intent categories such as search, email, or remarketing.

These row-level features are then collapsed into a fixed-length vector for each journey. The goal is to summaries a variable-length sequence into a single, information-dense record per PurchaseID without losing the timing signal. We therefore aggregate intensity and pacing through counts of contacts and inter-event intervals at daily, monthly and demographic levels, and through mean dwell times within demographic segments. Boundary effects are reduced to a small set of statistics, for example, the average late-touch ratio and the identifiers of the first and last channel used in the journey. Diversity is summarized either as the mean entropy across events or as the entropy of the final state; both choices avoid leakage because all events belong to the same closed journey. Recency is computed once per journey and passed as a scalar. Demographic attributes such as gender, age band, highest level of education, social class, life stage, and annual gross income (in EUR) are collected and subsequently integrated with device information.

To capture cross-sectional heterogeneity, we form explicit demographic–device interactions. At the row level, we compute statistics such as the mean duration by social class and device or by life stage and device (for example, dur_mean_socialclass_dev or touch_mean_lifestage_dev). We then average these features at the journey level, so that each demographic dimension contributes a single scalar per device type. This gives the model a way to learn segment-specific usage profiles, such as longer dwell times for certain income bands on mobile versus fixed devices. Boundary-channel information is kept compact by

storing the first and last observed channel as categorical attributes at the journey level, rather than expanding them into a wide one-hot matrix. If needed, we also compute the dominant channel and its share of touches.

All aggregations are deliberately restricted to events belonging to a single PurchaseID. This design prevents temporal leakage: the label for a journey never uses information from future journeys or from other users, and we do not inject global statistics such as per-user means into the test fold. In practice, the feature pipeline is implemented in pandas. We group by PurchaseID with a dictionary that maps each column to the corresponding aggregation function (for example, {'dwell_sec': ['mean', 'max', 'count'], 'gap_sec': ['mean']}), and we assert that the output has exactly one row per journey. High-cardinality categorical fields are cast to the category dtype to reduce memory usage, and intermediate aggregates are saved as Parquet files to ensure reproducibility and fast I/O. Missing values are filled with neutral values: zero for undefined ratios or entropy, and the median for continuous variables, which helps tree-based models avoid spurious splits. We do not apply scaling, as CatBoost and other gradient-boosted trees are invariant to monotonic transformations, and any winsorisation of durations is handled upstream by clipping dwell times to the [1, 720]-second range. After these steps, each journey is represented by a single wide row containing counts, means, maxima, device-specific dwell-time deviations, and demographic–device interactions.

Beyond pure sequence modelling, incorporating contextual features (e.g., dwell time, device type, temporal spacing, interaction entropy) has consistently improved discrimination over sequence-only baselines [7]. These features enrich the representation of user intent by capturing latent behavioral signals that sequences alone cannot convey, such as device, demographic, and interaction data, as well as pacing dynamics that signal purchase readiness. Despite these gains, deep neural SR models often integrate context in an ad hoc manner, treating it as auxiliary metadata rather than a core predictive dimension, thereby limiting interpretability and complicating deployment.

2.5. Class imbalance, operating points, and calibration

Journey-level labels in sequential recommendation systems often exhibit severe class imbalance, where positive engagement events (e.g., conversions, purchases) constitute a small fraction of the dataset compared to non-engagement instances. This imbalance can lead to biased estimators, causing models to overfit the majority class and fail to generalize for minority behaviors that drive critical business outcomes. To mitigate these issues, a combination of re-sampling strategies (oversampling of minority instances or undersampling of majority instances) and class-weighted loss functions has been shown to stabilize gradient updates while preserving minority signal integrity.

For deployment in probability-driven decision engines, accurate probability calibration is critical. Miscalibrated models can lead to suboptimal thresholding, resulting in over-targeting or under-targeting of user segments. Calibration techniques such as Platt scaling, isotonic regression, and temperature scaling help align predicted probabilities with empirical frequencies, thereby improving trustworthiness in operational settings [54]. The class prior is imbalanced (positives ≈ 33.3%, negatives ≈ 66.7%), which, if unaddressed, biases optimization toward the majority and degrades precision at decision thresholds. We apply bootstrap up-sampling of positives to achieve ≈1:2 (pos:neg), and use class weights {0:1, 1:2} in CatBoost to stabilize minority learning [18]. We adopt 5-fold StratifiedKFold with random_state=42.

To handle variable-length customer journeys while preserving the most informative interactions, a head–tail retention strategy is adopted. Specifically, the first 10 and last 20 touchpoints of each journey are retained. This design is motivated by behavioral observations in digital marketing, where early interactions often capture initial discovery intent, while late-stage interactions are more indicative of purchase readiness and conversion signals.

In this study, stratified K-fold cross-validation is applied at the journey level, where each observation represents a completed customer journey aggregated into a fixed-length feature vector. As a result, the prediction task is formulated as a classification problem over independent journey instances, rather than a sequential forecasting task over time-ordered events.

Resampling policy: we adopt a journey-level bootstrap up-sampling of the minority class to achieve a target ratio of approximately 1:2 (pos:neg). Concretely, let N0 and N1 denote the counts of negative and positive journeys after aggregation. If N1 <⌊N0/2⌋, we resample positives with replacement to N1′=⌊N0/2⌋; otherwise, resampling is a no-op. The resampled positives are then concatenated with the original negatives and shuffled with a fixed seed. This mild re-balance preserves a realistic prior while enhancing minority signal without manufacturing synthetic neighbours.

Class weighting: within the learner, we assign class weights {w0,w1}={1,2}, increasing the penalty for misclassifying positives and complementing the resampling policy. This combination stabilises training and usually offers a favourable precision–recall trade-off without excessive false positives at deployment thresholds.

Cross-validation protocol: we employ Stratified 5-Fold cross-validation at the journey grain: each fold contains disjoint sets of PurchaseID values and preserves the class ratio. Shuffling with a fixed random_state ensures reproducibility. Early stopping is performed on each fold's validation split using a patience of 100 iterations. Optional stratification refinements include binning by sequence-length deciles or device-tilt quantiles to further equilibrate covariate distributions across folds.

Each observation in the dataset corresponds to a completed customer journey aggregated at the purchase-decision level. Consequently, the modeling task does not attempt to predict future interactions within an ongoing sequence but instead classifies the outcome of an observed journey instance. Cross-validation is therefore performed across independent journey identifiers rather than across individual interaction events. This design reduces the risk of temporal leakage between training and validation partitions.

Robustness checks. As sensitivity analyses, we (i) vary the resampling target around 1:2 (e.g., 2:3, 1:1) to confirm stability of precision and recall; (ii) evaluate alternative selection criteria (e.g., PR-AUC or F1) when recall is prioritised; and (iii) perform a temporal walk-forward split on a held-out subset to approximate deployment latencies.

2.6. Learner and hyper-parameters

CatBoostClassifier: iterations=1200, learning_rate=0.08, depth=8, l2_leaf_reg=3, loss_function='Logloss', eval_metric='AUC', class_weights={0:1,1:2}, early_stopping_rounds=100. Categorical columns are passed via cat_features. Ordered TS and greedy feature combinations are handled internally.

For modelling, we employ CatBoost, an ordered boosting algorithm that handles categorical variables natively, providing robust performance without extensive feature engineering. Model validation uses stratified five-fold cross-validation with a fixed random seed to ensure that each fold reflects the distribution of the full dataset and that results are reproducible. Hyperparameters were determined through a coarse grid search on a pilot subset, settling on 1,200 boosting iterations, a learning rate of 0.08, a tree depth of eight, and an L2 leaf regularization coefficient of three; early stopping after 100 rounds is incorporated to prevent overfitting. Categorical features are passed explicitly via the `cat_features` parameter so that the estimator can leverage its native categorical handling. Routine orchestration tasks, such as splitting data and calculating performance metrics, are handled by Scikit-learn's utility functions, avoiding the need to implement these elements from scratch and ensuring consistent experimental workflows.

We report out-of-fold (OOF) ROC-AUC and select the operating threshold by maximizing $F_{0,5}$ over t∈[0.01,0.99]:

$$F_\beta = ((1+\beta^2)PR)/(\beta^2 P + R), \beta = 0.5 \quad (3)$$

where P and R are precision and recall computed from $\hat{y}_j(t) = 1\{P_j \geq t\}$. The selected threshold t* is logged alongside the model to support consistent scoring, and a confusion profile (TP,FP,TN,FN) at t* is produced for budget and risk communication.

Serving and monitoring. The production artefacts consist of the trained model, the feature schema, and t^*. Post-deployment, we monitor (i) precision/recall at t^*, (ii) class-prior shift, and (iii) drift in key covariates (recency, gaps, device deviations). Thresholds may be recalibrated against expected-profit curves if incentive costs or class priors shift materially.

We then produce a classification report at the chosen threshold. (PR-AUC/AP, lift@k, and calibration are recommended for deployment).

2.7. Evaluation metrics used in journey prediction

For multi-class or binary classification, we use precision, recall, and F1-score. Let TP, FP, FN denote true positives, false positives, and false negatives.

$$Precision = TP/(TP + FP) \quad (4)$$

$$Recall = TP/(TP + FN) \quad (5)$$

$$F1 = (2 \times Precision \times Recall)/(Precision + Recall) \quad (6)$$

The F1-score is the harmonic mean; it penalizes low values of either component. We further adopt F0.5 to emphasize precision in high-cost campaign settings. Low F1-scores require inspecting the confusion matrix to identify whether precision or recall is the bottleneck.

3. Results and Discussion

3.1. Descriptives

Following data cleaning and trimming, the final analytical dataset contained 134 037 individual interactions drawn from 25,632 distinct user journeys. The length of these sequences ranged from eight to thirty events, with an average of 19.77 events per journey. The target variable exhibited a 33.3 % positive and 66.7 % negative distribution, indicating a moderate imbalance between positive and negative outcomes.

Table 3. Data descriptives.

Set	Class	Count	Share (%)
Before balancing	Negative (0)	17,088	85.6
Before balancing	Positive (1)	2,864	14.4
Before balancing	Total	19,952	100.0
After balancing	Negative (0)	17,088	66.7
After balancing	Positive (1)	8,544	33.3
After balancing	Total	25,632	100.0

Each row is a PurchaseID (journey) after trimming; columns include touch counts and mean dwell by date/week/month, pacing (gap_sec), boundary cues (channel_first, channel_last, late_touch_ratio), device dwell deviations (dur_dev_FIXED/MOBILE), demographics×device means, and global context (recency_hours, intent_ratio, entropy_cat), plus the target (booking = 1).

Table 3 shows the number and shares of the dataset; after balancing, the final table is used as input to CatBoost, with categorical columns supplied via cat_features and numerical features left in native scale.

3.2. Discriminative performance across folds

After rebalancing, the dataset contains 17 088 non-purchase journeys and 8 544 purchase journeys. CatBoost is trained with class weights of 1.0 for the negative class and 2.0 for the positive class to offset the remaining skew. Under five-fold stratified cross-validation, the model attains fold-level AUC scores of 0.9508, 0.9449, 0.9431, 0.9493, and 0.9512. The out-of-fold mean AUC is 0.9479, and the variation across folds is small, indicating a stable ranking of journeys by purchase propensity.

To provide a more comprehensive evaluation of model performance under class imbalance, additional diagnostic visualizations were generated, including receiver operating characteristic (ROC) curves, precision–recall (PR) curves, and normalized confusion matrices. Precision–recall curves are particularly informative for imbalanced classification tasks because they emphasize performance on the positive class.

These visualizations provide a clearer understanding of the model's discrimination capability across decision thresholds. Fig. 2 shows the out-of-fold receiver operating characteristic (ROC) curve. The model achieves an ROC-AUC of 0.9479, demonstrating strong discrimination between purchasing and non-purchasing journeys. The model achieves an average precision (AP) of 0.9107, indicating strong performance on the positive class under class imbalance (Fig. 3).

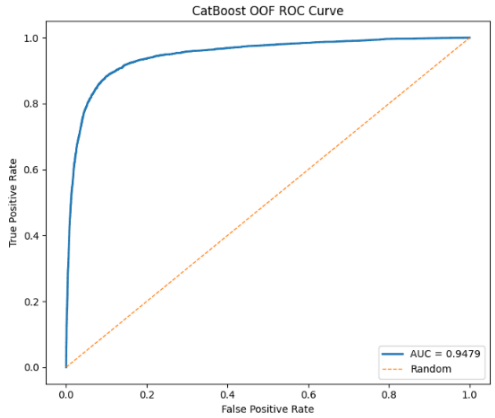


Fig. 2. CatBoost Out-of-fold ROC Curve.

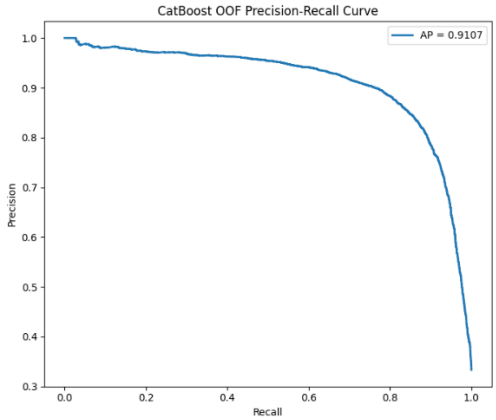


Fig. 3. CatBoost Out-of-fold Precision-Recall Curve.

The model achieves a lift of approximately 3 in the highest-ranked portion of the sample (Fig. 4), indicating that the top-ranked customer journeys contain substantially more purchasing instances than expected under random targeting.

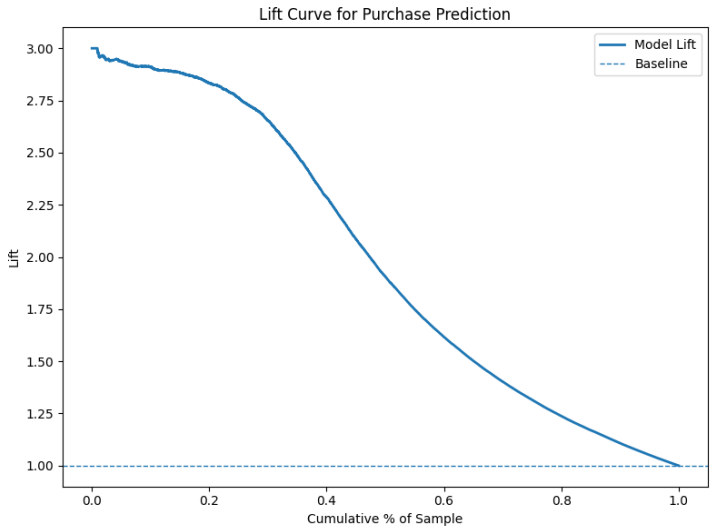


Fig. 4. Lift Curve for Purchase Prediction.

The results indicate that structured tabular representations can achieve strong predictive performance for customer journey classification tasks. Rather than replacing deep sequential models, this approach provides a practical alternative when datasets contain heterogeneous contextual attributes and when interpretability and computational efficiency are important considerations.

3.3. Threshold selection and precision-oriented operating point

After out-of-fold probabilities have been obtained, the ranking model is converted into a classifier by selecting a decision threshold. The threshold is chosen by maximizing the Fβ score, where P and R denote precision and recall. Setting β=0.5 places four times more weight on precision than on recall, which is appropriate when false positives translate into unnecessary outreach or costly incentives. Scanning the out-of-fold scores yields an optimal cut-off of 0.723. At this operating point, the model achieves a precision of 0.897 for positive predictions, a recall of 0.773, an F0.5 of 0.869, and an overall

accuracy of 0.895. The associated confusion matrix contains approximately 6,607 true positives, 761 false positives, 16,327 true negatives and 1,937 false negatives. In practical terms, about 7,368 journeys would be targeted, and only around 10.3% of contacted customers would be non-buyers, which is an attractive profile for campaigns that carry a non-trivial per-contact cost.

This precision-leaning setting exemplifies the cost-sensitive nature of threshold selection. When campaign incentives or call-center follow-ups are expensive, operating in a region that tightly controls false positives while still capturing the majority of genuine buyers offers a favorable trade-off. In contexts where missed opportunities are more costly than wasted outreach, the same framework can be re-optimized with a larger β to shift the balance towards recall, but the present results show that a precision-first stance already delivers strong business value.

Further inspection of the engineered features suggests that the model performance is largely driven by temporal dynamics and journey structure. Temporal descriptors such as recency_hours, gap_sec and late_touch_ratio capture the proximity of the interaction sequence to the final purchase event, reflecting the importance of the "final consideration" phase of the journey. Structural aspects of the path also play a role: entry and exit channels (channel_first, channel_last) and interaction diversity (entropy_channel) provide signals about how users navigate the journey before conversion. In addition, engineered demographic–device interaction terms (for example, dur_mean_socialclass_dev) capture heterogeneity across user segments, indicating that engagement patterns vary across social class, income, and device type. Overall, the predictive strength of the model appears to arise from the combination of temporal cadence, journey structure and user context rather than from any single feature alone.

From an operational standpoint, the same design lends itself to straightforward deployment. Training and scoring run comfortably on commodity CPUs, so the model can be integrated into standard marketing or CRM stacks without GPU infrastructure. The tabular representation and CatBoost's ordered encodings handle high-cardinality categorical variables in a compact way, which simplifies feature storage and monitoring. Because the features are interpretable, the system not only predicts but also suggests concrete levers: a high intent ratio, compressed inter-event gaps, or prolonged mobile dwell times can trigger next-best-action rules or remarketing flows, while SHAP explanations help product teams localize UX frictions along the journey.

Viewed through a managerial lens, three themes emerge. Targeting efficiency is achieved via the precision-oriented threshold, which curbs budget waste in retargeting and incentive programmes. Operational simplicity follows from relying on a CPU-bound, tabular pipeline rather than a deep sequential stack, reducing engineering overhead and easing MLOps integration [23], [35]. Finally, actionability arises from the model's interpretable feature set: patterns linking dwell-time variance, device type and social class indicate where adjustments to the user experience or to the media mix are most likely to lift conversion.

3.4. Limitations

Although the findings are encouraging, the study's scope is still narrow. We therefore outline a staged roadmap to strengthen temporal validity, broaden the baselines, and harden the pipeline for long-term, cross-market use.

First, temporal validity can be improved by replacing stratified K-fold with blocked or walk-forward cross-validation. These schemes respect the time order of events and reduce look-ahead bias in log data. Second, the set of baselines should be widened. Tuned XGBoost and LightGBM models can be benchmarked alongside lightweight sequential recommenders such as SASRec or BERT4Rec configured with reduced depth or sequence length. All models need to run under the same feature views and cost assumptions, and their training time, inference time, memory use, and model size should be reported to quantify how CPU-friendly they are compared with deep sequential recommenders [21], [23]. Third, evaluation ought to move beyond ROC-AUC and F0.5. Ranking-centric metrics such as PR-AUC (average precision), lift@k, and calibration measures like the Brier score and expected calibration error with temperature scaling provide a closer match to marketing use cases [54]. Statistical comparisons can

then rely on DeLong's test for correlated ROC-AUCs, while bootstrap confidence intervals for PR-AUC, lift@k, and threshold-based metrics help assess robustness [55].

Fourth, future work could incorporate systematic ablation studies that remove groups of features, such as recency indicators, inter-event gaps, entropy measures or boundary-related variables, in order to better understand the relative contribution of each feature block. These experiments could be complemented with post-hoc interpretability techniques. For instance, accumulated local effects may provide insights into global feature patterns, while SHAP-based approaches could further illuminate both local and global feature attributions. Fifth, generalisation needs to be tested across different seasonal periods and adjacent verticals, for example flights versus accommodations. If user, device or channel distributions drift, domain-adaptation techniques should be applied. Finally, pipeline hardening should include data-shift monitoring on key context features (such as dwell-time distributions and recency profiles), automated calibration checks, and threshold re-tuning driven by campaign economics so that the system remains reliable over time [27].

Taken together, these directions underline several limitations of the current study. First, the empirical analysis focuses on a single modeling architecture and therefore does not provide a comprehensive comparison with alternative tree-based learners or deep sequential recommendation models. Second, the head–tail sequence retention strategy is fixed and may not fully capture informative interactions occurring in the middle segments of certain customer journeys. Third, the experiments rely on a single domain-specific dataset, which limits the ability to assess the robustness of the approach across different markets, seasonal dynamics, or customer behavior patterns. Addressing these limitations will be important for establishing the broader applicability of the proposed framework.

4. Conclusion

This study introduces a context-aware, tabular pipeline for purchase prediction using CatBoost with ordered target encoding. Achieving AUC = 0.9479 and $F_{0.5}$ = 0.869, this approach significantly outperforms traditional Markov Chains ($R^2 \approx 0.81$) [5] by overcoming "memoryless" [17] constraints and surpasses Deep Learning models like LSTM in predictive yield (77.3% vs. 62.33% recall) [4]. By "tabularizing sequences," the model captures complex non-linear interactions such as device interplay and journey boundaries more effectively than Recformer or SEM [42], without the high computational overhead. The findings prove that strategic contextualization (pacing, diversity, and boundary effects) allows tree-based learners to match or exceed the performance of latent sequence models. This establishes tabular boosting as a superior, high-discrimination alternative for behavioral modeling in unbalanced marketing datasets. A precision-first regime (89.7% Precision) minimizes wasted incentive spend compared to broader behavioral frameworks. The CPU-only scoring profile and compact feature store offer a faster time-to-value than GPU-dependent deep learning, enabling cost-effective, real-time A/B testing and scalable deployment across diverse digital channels. Future implementation should prioritize (i) production-grade temporal validation, (ii) decision-curve analysis for economic thresholding, (iii) drift monitoring of key contextual features, and (iv) automated dashboards to track precision/recall at business-critical cut-offs.

Declarations

Author contribution. All authors contributed equally to the conceptualization of the study, development of the methodology, data analysis, software implementation, and preparation of the manuscript.
Funding statement. This research received no external funding.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] Y. Chen and C. Prentice, "Integrating Artificial Intelligence and Customer Experience," *Australas. Mark. J.*, vol. 33, no. 2, pp. 141–153, May 2025, doi: 10.1177/14413582241252904.

[2] "AI elevates customer service in travel and transportation | Roland Berger." Accessed: Sep. 10, 2026. [Online]. Available: https://www.rolandberger.com/en/Insights/Publications/AI-elevates-customer-service-in-travel-and-transportation.html.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[4] M. Ćirić, B. Predić, D. Stojanović, and I. Ćirić, "Single and Multiple Separate LSTM Neural Networks for Multiple Output Feature Purchase Prediction," *Electronics*, vol. 12, no. 12, 2023, doi: 10.3390/electronics12122616.

[5] J. Zhang, "Deep Learning-Based Attribution Framework for Real-Time Budget Optimization in Cross-Channel Pharmaceutical Advertising: A Comparative Study of Traditional and Digital Channels," in *Proceedings of the 2025 International Conference on Software Engineering and Computer Applications*, New York, NY, USA: ACM, Jun. 2025, pp. 248–254. doi: 10.1145/3747912.3747951.

[6] J. Li *et al.*, "Text Is All You Need: Learning Language Representations for Sequential Recommendation," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2023, pp. 1258–1267. doi: 10.1145/3580305.3599519.

[7] Z. Cui, H. Wu, B. He, J. Cheng, and C. Ma, "Context Matters: Enhancing Sequential Recommendation with Context-aware Diffusion-based Contrastive Learning," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Oct. 2024, pp. 404–414. [Online]. Available: https://arxiv.org/html/2405.09369v5

[8] X. Zhou, R. Lumbantoruan, Y. Ren, L. Chen, X. Yang, and J. Shao, "Dynamic Bi-layer Graph Learning for Context-aware Sequential Recommendation," *ACM Trans. Recomm. Syst.*, vol. 2, no. 2, pp. 1–23, Jun. 2024, [Online]. Available: https://www.researchgate.net/publication/376379134

[9] N. L. S. d, E. Souza, M. T. Arante, L. F. Braghirolli, M. M. P. Mafia, and E. M. Frazzon, "Orchestrating a ship-from-store omnichannel operation using 4PL digital platform," *Int. J. Bus. Forecast. Mark. Intell.*, vol. 8, no. 1, p. 52, 2023, doi: 10.1504/IJBFMI.2023.127704.

[10] J. N. Sheth, V. Jain, and A. Ambika, "The growing importance of customer-centric support services for improving customer experience," *J. Bus. Res.*, vol. 164, no. September, p. 113943, Sep. 2023, doi: 10.1016/j.jbusres.2023.113943.

[11] M. Xu, F. Liu, and W. Xu, "A Survey on Sequential Recommendation," in *2019 6th International Conference on Information Science and Control Engineering (ICISCE)*, IEEE, Dec. 2019, pp. 106–111. doi: 10.1109/ICISCE48695.2019.00031.

[12] C. Wu, F. Wu, T. Qi, C. Li, and Y. Huang, "Is News Recommendation a Sequential Recommendation Task?," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2022, pp. 2382–2386. [Online]. Available: https://arxiv.org/pdf/2108.08984

[13] F. Sun *et al.*, "BERT4Rec," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Nov. 2019, pp. 1441–1450. [Online]. Available: https://arxiv.org/pdf/1904.06690.

[14] L. Theodorakopoulos and A. Theodoropoulou, "Leveraging Big Data Analytics for Understanding Consumer Behavior in Digital Marketing: A Systematic Review," *Hum. Behav. Emerg. Technol.*, vol. 2024, no. 1, p. 3641502, Jan. 2024, doi: 10.1155/2024/3641502.

[15] P. B. Pires, B. M. Perestrelo, and J. D. Santos, "Unpacking Customer Experience in Online Shopping: Effects on Satisfaction and Loyalty," *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, no. 3, p. 245, Sep. 2025, doi: 10.3390/jtaer20030245.

[16] T. S. G. Furquim, C. P. da Veiga, C. R. P. da Veiga, and W. V. da Silva, "The Different Phases of the Omnichannel Consumer Buying Journey: A Systematic Literature Review and Future Research Directions," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 1, pp. 79–104, Dec. 2022, doi: 10.3390/jtaer18010005.

[17] L. Amalia, Sertiningsih, F. Hayatuki, S. Muliani, and A. M. Junaidi, "Advantages and Disadvantages of Utilizing Markov Chains in a Digital Marketing Strategy to Predict E-Commerce Sales," *Sci. Technol. Sci. Soc.*, vol. 1, no. 3, pp. 43–61, Dec. 2024, doi: 10.59324/stss.2024.1(3).03.

[18] L. D. Hollebeek, S. Urbonavicius, V. Sigurdsson, R. Arvola, and M. K. Clark, "Customer Journey Value: A Conceptual Framework," *J. Creat. Value*, vol. 9, no. 1, pp. 8–26, May 2023, doi: 10.1177/23949643231157155.

[19] X. Xia *et al.*, "TransAct: Transformer-based Realtime User Action Model for Recommendation at Pinterest," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2023, pp. 5249–5259. doi: 10.1145/3580305.3599918.

[20] A. Muralidhar and Y. Lakkanna, "From Clicks to Conversions: Analysis of Traffic Sources in E-Commerce," *J. Media Manag.*, vol. 6, no. 12, p. 1, Dec. 2024, doi: 10.47363/JMM/2024(6)182.

[21] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?," in *Advances in Neural Information Processing Systems 35*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022, pp. 507–520. doi: 10.52202/068431-0037.

[22] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.

[23] R. Timmer, "Deep Learning versus Gradient Boosted Decision Trees: A Comparison on Noise Robustness," *Erasmus Univ. Rotterdam Erasmus Sch. Econ.*, no. July, 2023, [Online]. Available: https://thesis.eur.nl/pub/67870/Bachelor-Thesis-Ruben-Timmer.pdf.

[24] G. Zabërgja, A. Kadra, C. M. M. Frey, and J. Grabocka, "Tabular Data: Is Deep Learning all you need?," vol. 2, pp. 1–55, Oct. 2025, [Online]. Available: http://arxiv.org/abs/2402.03970

[25] M. Badouch and M. Boutaounte, "TAHRF: enhancing personalized tourism recommendations with dynamic adaptation," *IAES Int. J. Artif. Intell.*, vol. 15, no. 1, pp. 374–382, Feb. 2026, doi: 10.11591/ijai.v15.i1.pp374-382.

[26] W. Zheng, J. Li, H. Pan, and X. Fan, "EMO: Elastic oversampling with boundary skeletons from minimal enclosing balls," *Neurocomputing*, vol. 672, no. April, p. 132818, Apr. 2026, doi: 10.1016/j.neucom.2026.132818.

[27] A. Tanaka, K. Sato, and Y. Nakamura, "Real-Time Modeling of Customer Journey in Multichannel Platforms," *J. Electron. Commer. Manag.*, vol. 1, no. 1, pp. 65–84, 2025, [Online]. Available: https://www.joecm.ir/article_222685_87a12dd210117584dd7642463cc6dece.pdf

[28] R. Du, D. He, Y. Song, Z. Ding, S. Chan, and X. Li, "Privacy-Preserving Short-Term Travel Demand Forecasting Based on Federated Learning," *IEEE Trans. Veh. Technol.*, vol. 74, no. 7, pp. 11436–11449, Jul. 2025, doi: 10.1109/TVT.2025.3547823.

[29] S. Chen *et al.*, "TabDeco: A Comprehensive Contrastive Framework for Decoupled Representations in Tabular Data," vol. 1, pp. 1–13, 2024, [Online]. Available: https://arxiv.org/abs/2411.11148v1

[30] A. V. Petrov and C. Macdonald, "Generative Sequential Recommendation with GPTRec," vol. 1, pp. 1–10, 2023, [Online]. Available: http://arxiv.org/abs/2306.11114

[31] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[32] I. Mapfumo and T. Shongwe, "Performance Evaluation of Machine Learning and Deep Learning Models for Credit Risk Prediction," *J. Risk Financ. Manag.*, vol. 19, no. 3, p. 210, Mar. 2026, doi: 10.3390/jrfm19030210.

[33] K. F. Balbal and D. Birant, "RFM-Net: A Convolutional Neural Network for Customer Segment Classification," *Appl. Sci.*, vol. 16, no. 5, p. 2223, Feb. 2026, doi: 10.3390/app16052223.

[34] M. Flórez, E. Carrillo, F. Mendes, and J. Carreño, "A Context-Aware Tourism Recommender System Using a Hybrid Method Combining Deep Learning and Ontology-Based Knowledge," *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, no. 3, p. 194, Aug. 2025, doi: 10.3390/jtaer20030194.

[35] R. A. Ruslan and N. Arbaiy, "Handling Imbalanced Datasets in Machine Learning: Challenges, Approaches, and Best Practices," *J. Appl. Sci. Technol. Comput.*, vol. 1, no. 2, pp. 1–17, Nov. 2024, doi: 10.30880/jastec.2024.01.02.003.

[36] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[37] G. Li, Y. Wu, Y. Bai, and W. Zhang, "ReMAHA–CatBoost: Addressing Imbalanced Data in Traffic Accident Prediction Tasks," *Appl. Sci.*, vol. 13, no. 24, p. 13123, Dec. 2023, doi: 10.3390/app132413123.

[38] H. Ribeiro, B. Barbosa, A. C. Moreira, and R. Rodrigues, "Customer Experience, Loyalty, and Churn in Bundled Telecommunications Services," *Sage Open*, vol. 14, no. 2, p. 21582440241245190, Apr. 2024, doi: 10.1177/21582440241245191.

[39] A. A. Alhussan, M. Metwally, and S. K. Towfek, "Predicting CO2 Emissions with Advanced Deep Learning Models and a Hybrid Greylag Goose Optimization Algorithm," *Mathematics*, vol. 13, no. 9, p. 1481, Apr. 2025, doi: 10.3390/math13091481.

[40] K. N. Lemon and P. C. Verhoef, "Understanding Customer Experience Throughout the Customer Journey," *J. Mark.*, vol. 80, no. 6, pp. 69–96, Nov. 2016, doi: 10.1509/jm.15.0420.

[41] G. Bernard and P. Andritsos, "A process mining based model for customer journey mapping," *CEUR Workshop Proc.*, vol. 1848, pp. 49–56, 2017, [Online]. Available: https://ceur-ws.org/Vol-1848/CAiSE2017_Forum_Paper7.pdf.

[42] R. S. Dixit, S. L. Choudhary, and N. Govil, "Analyzing the impact of artificial intelligence on the online purchase decision-making process through the lens of the UTAUT 2 model," *Discov. Comput.*, vol. 28, no. 1, p. 88, May 2025, doi: 10.1007/s10791-025-09575-5.

[43] L. Reinkemeyer, *Process Mining in Action*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-40172-6.

[44] L. Zhang and L. Jiang, "Game-theoretic SHAP-driven interpretable forecasting of air cargo demand using Bayesian-optimized random forests," *Front. Phys.*, vol. 13, Oct. 2025, doi: 10.3389/fphy.2025.1705687.

[45] M. Lee, I. Choi, and W.-C. Kim, "Predicting Mobile Payment Behavior Through Explainable Machine Learning and Application Usage Analysis," *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, no. 2, p. 117, May 2025, doi: 10.3390/jtaer20020117.

[46] A. Prashanthan, "An Integrated Framework for Optimizing Customer Retention Budget using Clustering, Classification, and Mathematical Optimization," *J. Comput. Theor. Appl.*, vol. 3, no. 1, pp. 45–63, Jul. 2025, doi: 10.62411/jcta.13194.

[47] M. Teixeira, J. M. Oliveira, and P. Ramos, "Enhancing Hierarchical Sales Forecasting with Promotional Data: A Comparative Study Using ARIMA and Deep Neural Networks," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 2659–2687, Nov. 2024, doi: 10.3390/make6040128.

[48] E. Neufeld, "Cross-Device and Cross-Channel Identity Measurement Issues and Guidelines," *J. Advert. Res.*, vol. 57, no. 1, pp. 109–118, Mar. 2017, doi: 10.2501/JAR-2017-008.

[49] F. M. Talaat, A. Aljadani, B. Alharthi, M. A. Farsi, M. Badawy, and M. Elhosseini, "A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing," *Mathematics*, vol. 11, no. 18, p. 3930, Sep. 2023, doi: 10.3390/math11183930.

[50] Q. Ai, X. Wang, and M. Bendersky, "Metric-agnostic Ranking Optimization," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '23. New York, NY, USA: ACM, Jul. 2023, pp. 2669–2680. doi: 10.1145/3539618.3591935.

[51] N. S. Dahake, P. Suresh Dahake, and A. Gajghate, "Beyond the Hype: A Critical Evaluation of Predictive Analytics Accuracy in High-Stakes Business Decisions," in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/IDICAIEI61867.2024.10842857.

[52] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017, [Online]. Available:

https://proceedings.neurips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

[53] O. Chapelle, E. Manavoglu, and R. Rosales, "Simple and Scalable Response Prediction for Display Advertising," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, pp. 1–34, Jan. 2015, doi: 10.1145/2532128.

[54] Hall of Franklin Institute, "Monthly weather review," *J. Franklin Inst.*, vol. 138, no. 6, pp. 1–8, Dec. 1894, doi: 10.1016/0016-0032(94)90228-3.

[55] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.