International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.140, September 2026

Intent-Aware Sequential Recommendation for Small Retail: A Hybrid CatBoost-GRU Framework using Survey and Transaction Data

Md Zunaid Tausif
School of Computer Science and Technology
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China

Rabeya Taposhi
School of Business
Bangladesh Open University
Board Bazar, Gazipur-1705, Bangladesh

ABSTRACT

Small physical retailers face a fundamental challenge: making personalized decisions with sparse, heterogeneous customer data. Unlike large e-commerce platforms, they lack the interaction volume required by conventional machine learning and recommendation systems.

This paper proposes an intent-aware recommendation framework that jointly models customer purchase readiness and next-product preference using survey and transactional data from a micro-retail juice bar (Sundew Juicebar), supplemented with synthetically augmented records that preserve the original behavioral and demographic distributions. Purchase intent is formulated as a customer-level classification task, achieving a mean AUC of 0.801 under cross-validation, while next-item prediction is addressed using sequential and co-occurrence-based methods.

The results show that classical recommendation techniques remain competitive in menu-constrained environments, while sequential models capture complementary behavioral patterns. By integrating intent prediction with recommendation through a weighted fusion mechanism, the proposed framework improves Accuracy@1 by 5.00% and NDCG@3 by 0.89% over a standalone sequential model.

Beyond predictive performance, the framework provides interpretable decision support, enabling small-business operators to understand and act on model outputs. The results demonstrate that combining behavioral and attitudinal signals yields practical value in real-world micro-retail settings, even under severe data constraints.

General Terms

Micro-retail analytics, Recommendation systems, Machine learning

Keywords

Micro-retail analytics, Purchase intent prediction, Recommendation systems, Sequential modeling, Customer behavior analysis, Explainable machine learning

1. INTRODUCTION

Small physical retailers often face a practical but difficult decision problem: they need to understand which customers are likely to return and what products those customers are most likely to prefer, yet they must do so with limited and irregular data. In contrast to large e-commerce platforms that rely on rich clickstream logs, dense user–item interactions, and large-scale behavioral histories, micro-retail businesses such as independent juice bars usually operate with sparse, manually collected, and heterogeneous records [1, 2]. Their available information often consists of short customer surveys, basic transaction logs, and repeated purchases from a limited menu, making conventional retail analytics pipelines difficult to apply directly [3]. Nevertheless, these businesses increasingly need data-driven support to improve customer retention, personalize product suggestions, and make more informed operational decisions.

Although recent advances in machine learning have substantially improved purchase prediction and recommendation quality in digital commerce, most of these advances are developed for data-rich environments [4]. Existing purchase prediction models frequently assume abundant behavioral signals, while recommendation methods often depend on dense interaction matrices or long sequential histories [5, 2]. These assumptions are rarely satisfied in small offline retail settings. As a result, models that perform well in online platforms may become unstable, weakly informative, or operationally impractical when transferred to micro-retail contexts. In addition, prior studies rarely take advantage of survey-derived attitudinal data, even though such information is often more realistic to collect in small physical businesses than in large-scale online systems.

A second limitation is that purchase prediction and recommendation are commonly treated as separate analytical tasks. In practice, however, these two decisions are closely related. A recommendation is more useful when it reflects not only what a customer may prefer, but also whether the customer appears ready to purchase at all. For small retailers, this distinction matters because customer interactions are limited, product choices are narrow, and each repeat purchase carries relatively high business value. A system that recommends products without considering purchase readiness may generate suggestions that are less aligned

with real customer behavior, while a system that predicts intent without linking that signal to action provides limited decision support. This disconnect creates both a conceptual and a practical gap in current micro-retail analytics.

A third challenge concerns interpretability. Small-business operators typically need decision support that is understandable and actionable rather than purely predictive. Black-box outputs may offer little operational value if the business owner cannot identify why a customer is considered likely to return or why a particular beverage is being suggested. In micro-retail settings, where decisions are often made by non-technical users and directly affect promotions, stock planning, and customer engagement, explainability is not merely desirable but practically important [6].

These limitations motivate the need for a unified framework tailored to small physical retail. Such a framework should address three linked problems: first, it should estimate customer purchase intent from sparse and mixed data sources; second, it should generate personalized product recommendations that are aligned with that intent; and third, it should provide interpretable outputs that support practical decision-making for small-business users. To the best of current knowledge, existing literature does not provide an integrated approach that combines these three objectives in a micro-retail beverage setting using both survey and transaction data [7].

In response to this gap, this paper proposes an intent-aware recommendation framework for a small juice bar environment. The framework combines survey-derived preference information with transactional and behavioral signals to model customer purchase readiness and to improve next-product recommendation. Rather than treating prediction and recommendation as isolated pipelines, the proposed approach connects them in a single decision-oriented process. It also incorporates explainability to make model outputs more transparent and useful for retail operators. Through this design, the study aims not only to improve predictive performance under sparse data conditions, but also to make machine learning more applicable to the operational realities of micro-retail businesses.

Problem Statement

Current machine learning and recommendation approaches are primarily designed for large-scale online environments and cannot be directly applied to small physical retailers with sparse behavioral data. There is no existing framework that integrates purchase-intent prediction with personalized beverage recommendations using survey and transaction data in a micro-retail juice bar context [7]. Moreover, existing systems lack explainability features suitable for small-business users.

Research Questions

RQ1: Can customer purchase intent in a micro-retail juice bar be accurately predicted using a combination of survey-derived features and behavioral purchase data?

RQ2: Does conditioning beverage recommendations on predicted intent improve personalization quality compared to intent-agnostic recommendation methods?

RQ3: Can SHAP-based explanations enhance the interpretability and practical value of predictions and recommendations for small-business operators?

Hypotheses

H1: Tree-based ensembles perform comparably to linear baselines for intent prediction on small, heterogeneous micro-retail data.

H2: Classical recommenders, especially item–item co-occurrence, provide competitive next-item accuracy in micro-retail data.

H3: Intent-aware fusion modestly enhances sequential recommendation performance while enabling intent-conditioned interpretation and decision policies.

Contributions

This paper makes three primary contributions:

(1) A real-world micro-retail dataset is presented by combining transactional and survey-based customer data, addressing a gap in existing recommendation research dominated by large-scale online platforms.

(2) Empirical results demonstrate that classical recommendation models remain competitive in data-sparse retail environments, while deep sequential models require careful applicability constraints.

(3) An intent-aware fusion framework is proposed that preserves recommendation accuracy while enhancing interpretability and decision alignment, highlighting the role of intent modeling beyond pure ranking optimization.

Paper Organization

The remainder of this paper is structured as follows: Section 2 reviews related literature on purchase prediction, recommender systems, context-aware modeling, and explainable ML. Section 3 describes the proposed methodology, including the system architecture, feature design, and the intent-aware fusion mechanism. Section 4 outlines the experimental setup and evaluation metrics. Section 5 presents the results and discussion. Section 6 concludes with implications and future research directions.

2. LITERATURE REVIEW

Recent advances in machine learning (ML), personalization, and data-driven marketing have transformed how organizations understand and predict consumer behavior. This section synthesizes prior work on (1) purchase behavior and retention modeling, (2) recommendation systems across e-commerce and food domains, (3) time- and context-aware recommendation approaches, and (4) explainable ML in personalization. The review highlights where current research succeeds and where significant gaps remain, particularly regarding micro-retail settings, hybrid intent-driven personalization, and models built from survey-plus-transaction datasets.

2.1 Machine Learning for Purchase Behavior, Retention, and Intent Prediction

The prediction of customer purchase behavior has traditionally relied on statistical modeling, but recent literature shows a decisive shift toward ensemble-based and deep learning methods. Several works demonstrate that classical linear models fail to capture nonlinear relationships in behavioral data. For instance, [8] and [4] conduct a comparative analysis involving Logistic Regression, Random Forests, Gradient Boosting, Support Vector Machines, K-Nearest Neighbors, and XGBoost. Their findings show that

ensemble learners, particularly Gradient Boosting and Random Forests, consistently outperform linear baselines across accuracy, F1-score, and AUC. This aligns with emerging trends in digital marketing analytics, where feature interactions such as price sensitivity, browsing sequences, and product attributes require models capable of nonlinear separation.

Beyond static classification tasks, the broader question of customer retention and long-term engagement is addressed in more recent studies. [5] examine quarterly sales patterns and customer segmentation to build an ML-powered retention and discount optimization model. Their study applies XGBoost to forecast high-value segments and identify discount strategies tailored to individual buying profiles. The paper highlights that retention-focused ML systems can deliver tangible business value, but also notes that most existing datasets come from large-scale e-commerce environments, leaving small businesses underrepresented.

Another direction explores the challenges of class imbalance, common in online shopping where only a fraction of users convert. [9] apply CatBoost to address imbalance in purchase prediction and report superior accuracy over traditional boosted trees. The advantage comes from CatBoost's ordered boosting and built-in categorical encoding, which allow the model to learn robust patterns even when positive purchase signals are scarce. This is particularly relevant for physical micro-retailers, where limited datasets create sparsity issues.

Large-scale behavioral modeling has also expanded beyond simple binary intent prediction. [10] introduce STPC-PGM, a probabilistic graphical model integrating spatial, temporal, product-category, and monetary features from payment records. Their method predicts when, where, and what customers purchase, outperforming baseline models by 10–40%. The work's multi-dimensional representation demonstrates that richer context significantly boosts predictive power. However, STPC-PGM requires massive transactional datasets, far beyond what a small retailer can expect, and does not integrate attitudinal survey data, which may be a key advantage for micro-businesses.

Across these works, several limitations appear. First, nearly all studies rely on large online datasets, often collected passively (clickstreams, payment logs). Little attention is given to micro-retail environments where datasets are small, interactions are offline, and survey-based preference information is available. Second, none integrate intent prediction directly into a personalized recommendation generator, a hybrid design that could better reflect real-world decision pipelines, especially in small businesses where customer engagement is personal and repeat visits are crucial.

2.2 Recommender Systems in E-Commerce, Restaurants, and Retail

Recommendation systems (RS) form the second major line of research relevant to this work. Traditional RS methods include content-based filtering, user-based collaborative filtering, and model-based collaborative filtering such as matrix factorization. [11] outlines the architecture of hybrid recommenders that combine content-based signals with collaborative signals. This dual modeling strategy reduces cold-start issues but remains dependent on user–item rating density.

Food and restaurant recommendation systems offer more domain-specific insights. [12] compare several matrix factorization techniques for restaurant review datasets, reporting that SVD++ outperforms classical SVD by incorporating implicit feedback. While these approaches are effective on large restaurant rating datasets, they assume the presence of explicit or implicit feedback at scale, conditions that small physical shops rarely meet. Nevertheless, matrix factorization provides a mathematically simple and interpretable foundation for low-dimensional customer preference profiles.

In contrast to matrix factorization, association-rule approaches aim to uncover patterns in shopping baskets. [13] present a recommender built using FP-Growth to mine frequent itemsets and generate next-item suggestions. Their results show high recommendation accuracy for sessions with strong co-occurrence signals. This method is highly relevant to micro-retail contexts, such as a juice bar, where small baskets and simple co-purchase patterns make association rules attractive. However, FP-Growth alone cannot personalize suggestions to individual customer profiles.

Deep and sequential models push the field further. [14] evaluate GRU-RNNs for next-basket prediction and find that GRU models outperform Random Forests by capturing sequential dependencies such as weekday routines or health-conscious streaks. While sequence models are powerful, they require large chronological datasets, which are not typically available to a small retailer unless data collection is well-structured over time.

Taken together, these works show substantial progress in e-commerce recommendation but expose several gaps: (1) very little research on micro-retailers without rating matrices, (2) almost no integration of attitudinal survey data, and (3) an absence of recommender systems explicitly conditioned on predicted purchase intent.

2.3 Time- and Context-Aware Recommendation Models

Incorporating temporal and contextual variables into RS is a rapidly growing subfield. Consumer preferences are not static; they vary across seasons, hours of the day, and individual consumption cycles. [15] propose the ML.Recommend framework, which combines matrix factorization with recency filtering and a time-based weighting mechanism. They show that recommendations improve when older interactions are discounted, confirming that recent behavior is more predictive of future choices in dynamic markets.

Similar ideas arise in spatio-temporal models. [10] demonstrate that integrating temporal and spatial contexts significantly increases prediction accuracy. Their model captures the interplay among location, time, spending, and product category, which hints that even small-town juice bar customers may follow temporal rhythms such as morning smoothies or afternoon refreshments.

More recent surveys on context-aware recommendation highlight additional contextual dimensions such as weather, social influence, and micro-location. Post-2022 works introduce graph-based context encoders and attention mechanisms that adaptively weight contextual signals. Despite this growth, micro-retail businesses remain underrepresented, reinforcing the opportunity for a hybrid, context-sensitive approach tailored to small beverage retail.

2.4 Explainable AI in Purchase Prediction and Recommendation

Explainability is now indispensable in ML-driven personalization, especially when models guide business decisions in small enterprises. Black-box models, while accurate, offer limited transparency. [16] propose SHAP values, an additive feature attribution method grounded in cooperative game theory. SHAP

is widely regarded as one of the most theoretically robust local explanation tools for models such as XGBoost, CatBoost, and Random Forests.

In marketing contexts, transparency is crucial because business owners must understand why a model recommends a specific product. While some e-commerce studies adopt SHAP for feature attribution in churn or sentiment models, none of the reviewed papers apply XAI to hybrid intent-aware recommendation systems or to micro-retail food contexts. Integrating SHAP into a hybrid model addresses this gap by generating human-readable insights, aligning with emerging research norms emphasizing human-centric AI.

2.5 Summary and Research Gap

Across the reviewed literature, several trends are clear: (1) ML models such as XGBoost, CatBoost, and probabilistic graphical models effectively predict customer behavior at scale; (2) recommendation systems have advanced significantly in e-commerce and food services; (3) time- and context-aware models improve personalization by integrating temporal signals; and (4) SHAP-based explainability is becoming standard in interpretable ML.

However, significant gaps remain. Almost all prior work depends on large-scale online datasets; small physical retailers like a juice bar rarely appear in published literature. No current study integrates purchase-intent prediction directly into a recommender system. Survey-based preference data is nearly absent in existing RS studies. Therefore, the proposed study contributes a hybrid, intent-aware, explainable recommendation framework that leverages both survey and transaction data in a small beverage retail context.

3. METHODOLOGY

This section describes the proposed intent-aware recommendation framework. The system is designed for micro-retail environments where data is sparse and heterogeneous. It combines a CatBoost-based purchase-intent classifier with a GRU-based sequential recommender and integrates their outputs through a weighted fusion mechanism. The full pipeline is illustrated in Fig. 1.

The framework operates in two parallel branches that are combined at inference time. The first branch processes survey-derived and transactional features through a CatBoost classifier to produce a scalar purchase-intent probability $P_{\text{intent}} \in [0, 1]$ for each customer. The second branch processes the customer's chronological purchase history through a GRU network and outputs a probability distribution $P_{\text{GRU}}(i \mid h)$ over candidate items given history h. These two signals are then fused to generate an intent-conditioned ranked list of beverage recommendations.

3.1 Feature Design

The dataset combines two complementary data sources, yielding 45–55 features in total. Table 1 summarizes the main feature groups used across both model branches.

Survey features are primarily used by the CatBoost intent model to capture attitudinal signals that reflect customer readiness to return. Transactional features feed both models: the CatBoost classifier uses aggregate behavioral statistics, while the GRU processes raw sequential drink IDs with time-gap embeddings. Categorical variables are passed natively to CatBoost without one-hot encoding; the GRU learns item embeddings of dimension 32.

Table 1. Feature Groups Used in the Proposed Framework

Source	Feature Group	Examples
Survey	Attitudinal preferences	Health preference, price sensitivity, flavor preference
Survey	Visit motivation	Convenience, loyalty, recommendation
Transaction	Recency & frequency	Days since last visit, visit count
Transaction	Sequential behavior	Last drink ID, drink sequence, time gaps
Transaction	Spending patterns	Average spend, total spend
Transaction	Item context	Drink category, price tier

3.2 CatBoost Intent Prediction Model

The purchase-intent classifier is trained at the *customer level* to predict a binary label $y \in \{0, 1\}$ indicating whether a customer will make a return visit. Customer-level aggregation is used (rather than transaction-level) to prevent label leakage, since each customer's intent label is constant across their visits. Per-customer features are constructed by aggregating their transactional history (e.g., maximum visit frequency, mean average spend) and combining these with attitudinal survey responses.

CatBoost was selected for three reasons specific to the micro-retail setting: (1) it handles categorical features natively via ordered target statistics without leakage, (2) its ordered boosting procedure is robust to small datasets, and (3) it produces well-calibrated probability estimates required for downstream fusion.

Class imbalance (26.4% positive labels) is addressed by setting `scale_pos_weight` $= n_{\text{neg}}/n_{\text{pos}}$, where n_{neg} and n_{pos} are the counts of negative and positive examples respectively. Model selection is performed via 5-fold stratified cross-validation, and the final model is trained on the full customer-level dataset for use in fusion.

3.3 GRU Sequential Recommendation Model

The sequential recommender is a single-layer GRU network that maps a customer's recent purchase history $h = (i_1, i_2, \ldots, i_{t-1})$ to a probability distribution over the next item i_t. Each drink is represented by a learned embedding of dimension 32. The GRU hidden state of size 64 summarizes the sequential context, and a linear output layer with softmax activation produces item scores over the full vocabulary.

Sequences shorter than 3 interactions are excluded. For customers with more than 10 recent purchases, a sliding window of length 10 is used. Padding is applied on the left (oldest-first) so that the most recent purchase always aligns to the last non-padding position. The network is trained with categorical cross-entropy loss using the Adam optimizer (learning rate 0.001) for up to 30 epochs with early stopping (patience = 5).

3.4 Item Affinity Score $g(i)$

The fusion mechanism incorporates an item-level affinity score $g(i) \in [0, 1]$ that reflects item desirability independently of the sequential signal. When price information is available, $g(i)$ is

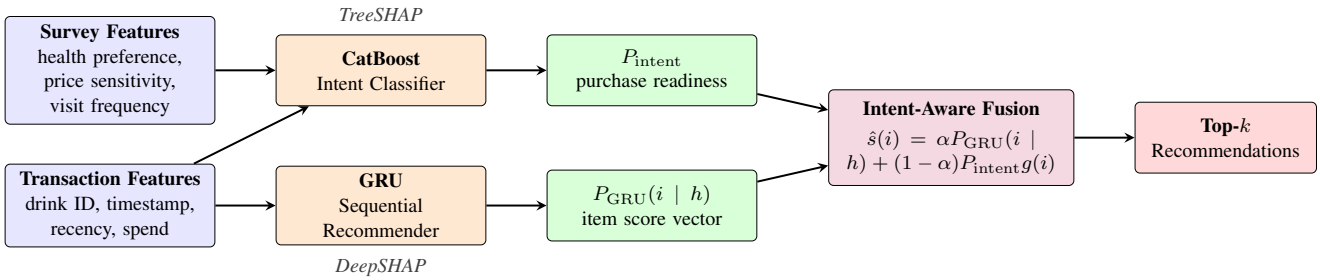

Fig. 1. Overview of the proposed intent-aware recommendation pipeline. Survey and transaction data feed two parallel models: a CatBoost classifier estimates purchase intent (P_{intent}), while a GRU produces item-level scores (P_{GRU}). The two outputs are fused via a weighted combination to produce the final ranked list of beverage recommendations.

computed as the min-max normalized average price of item i:

$$g(i) = \frac{\bar{p}_i - p_{\min}}{p_{\max} - p_{\min}} \tag{1}$$

where $\bar{p}_i$ is the mean transaction price of item i. When price data is absent, $g(i)$ is derived from the normalized purchase frequency of item i's category, providing a proxy for item popularity within its category group. If neither source is available, $g(i) = 0.5$ is used as a neutral constant.

3.5 Intent-Aware Fusion Mechanism

The core contribution of the framework is the fusion of the sequential recommendation signal with the intent estimate. For each candidate item i, a composite score $\hat{s}(i)$ is computed as:

$$\hat{s}(i) = \alpha \cdot P_{\text{GRU}}(i \mid h) + (1 - \alpha) \cdot P_{\text{intent}} \cdot g(i) \tag{2}$$

where:

- $P_{\text{GRU}}(i \mid h)$ is the GRU's softmax probability for item i given history h,
- $P_{\text{intent}} \in [0, 1]$ is the CatBoost-predicted purchase-readiness probability for the current customer,
- $g(i) \in [0, 1]$ is the item-level affinity score defined in Eq. (1),
- $\alpha \in [0, 1]$ is a tunable balance parameter controlling the trade-off between behavioral preference and intent-weighted item desirability.

When $\alpha = 1$, the fusion reduces to a pure GRU recommendation that ignores purchase readiness. When $\alpha = 0$, the system ranks items solely by intent-weighted affinity without any sequential signal. The optimal value of α is selected by grid search over $\{0.0, 0.25, 0.5, 0.75, 1.0\}$ using NDCG@3 on the validation set. The final ranked list of k items is obtained by sorting all candidate items by $\hat{s}(i)$ in descending order.
The formal recommendation procedure is presented in Algorithm 1.

3.6 Explainability via SHAP

To support practical decision-making for non-technical business operators, the framework incorporates SHAP-based explanations for both model branches. TreeSHAP [16] is applied to the CatBoost intent model to produce per-feature attribution scores that identify which survey and behavioral signals most influence a customer's predicted purchase readiness. DeepSHAP is applied to the GRU model to attribute next-item predictions to specific positions in the purchase history. Together, these explanations provide a two-level view of model behavior: (1) *why* a customer is considered likely to return, and (2) *why* a particular beverage is being recommended.

Algorithm 1 Intent-Aware Fusion Recommendation

Require: Customer history h, customer feature vector $\mathbf{x}$, item set $\mathcal{I}$, parameter α, top-k
Ensure: Ranked list R of k items
$P_{\text{intent}} \leftarrow \text{CatBoost}(\mathbf{x})$ {intent probability}
$\mathbf{p} \leftarrow \text{GRU}(h)$ {item score distribution}
for each item $i \in \mathcal{I}$ **do**
  $g(i) \leftarrow \text{ItemAffinity}(i)$ {price or category proxy}
  $\hat{s}(i) \leftarrow \alpha \cdot p_i + (1 - \alpha) \cdot P_{\text{intent}} \cdot g(i)$
end for
$R \leftarrow$ Top-k items sorted by $\hat{s}(i)$ descending
return R

3.7 Statistical Validation

To assess whether observed improvements from the fusion mechanism are statistically meaningful rather than artefacts of sampling variance, two complementary tests are employed. McNemar's test with continuity correction is used for binary hit/miss metrics (Accuracy@1 and Accuracy@3) since these are paired categorical outcomes per test sample. A paired t-test (with Wilcoxon signed-rank as a non-parametric backup) is applied to NDCG@3, which yields continuous per-sample scores. All tests compare the proposed fusion model against the GRU baseline using the same test instances, ensuring fair within-sample comparison.

4. EXPERIMENTAL SETUP

This section outlines the environment, data preparation steps, training strategy, baseline models, and evaluation metrics used to assess the performance of both the CatBoost intent predictor and the GRU-based sequential recommendation model.

4.1 Dataset Characteristics

The dataset is constructed from a real micro-retail juice bar (Sundew Juicebar) operated by one of the authors, comprising actual customer transactions and survey responses. To ensure sufficient sample size for robust model evaluation while preserving customer privacy, the original observations were augmented with synthetic records that preserve the same distributional properties and behavioral patterns as the real data. Augmentation was applied only to extend the sample size; no individual customer records were duplicated, and aggregate distributions of attitudinal, behavioral,

and transactional features remain consistent with the original observations.
The final dataset consists of 8,000 transaction records covering 1,000 unique customers and 5 distinct beverage items. Each customer has a median of eight recorded interactions, and all users have at least three purchase events, enabling reliable sequential modeling. The resulting interaction density is 1.6, reflecting repeated purchases from a limited menu rather than extreme sparsity. While the small item universe constrains ranking entropy, it realistically represents menu-driven micro-retail environments and allows controlled evaluation of sequential and intent-aware recommendation methods.

4.2 Dataset Preparation

The unified dataset contains approximately 8,000 rows and 45–55 features. After preprocessing, three subsets were created:

- **Training set:** 70% of the data
- **Validation set:** 15%
- **Test set:** 15%

The split was performed at the customer level (no customer appears in both training and test sets) to prevent leakage of behavioral patterns across partitions. For the GRU model, purchase histories were sorted chronologically per customer and converted into sequences of drink IDs with time-gap features. Sequence lengths ranged from 3 to 20 interactions per user, with left-padding applied when necessary. For the CatBoost classifier, categorical variables were passed natively without one-hot encoding, while continuous features were normalized when required. Class imbalance was handled using CatBoost's built-in `scale_pos_weight` parameter (set to $n_{\text{neg}}/n_{\text{pos}} \approx 2.79$, reflecting the 26% positive class rate at the customer level).

4.3 Implementation Environment

All experiments were conducted using Python 3.10 with the following libraries:

- CatBoost 1.2
- PyTorch 2.1 (for GRU implementation)
- Scikit-Learn 1.5 (for preprocessing and baselines)
- Pandas, NumPy (data handling)
- SHAP 0.44 (model explainability)

The models were trained on a workstation with an Intel Core i7 CPU (32 GB RAM) and an NVIDIA RTX 3060 GPU running Ubuntu 22.04 LTS. The CatBoost model was CPU-trained; the GRU model leveraged GPU acceleration.

4.4 Hyperparameter Configuration

4.4.1 CatBoost Intent Model. Key hyperparameters were tuned using five-fold cross-validation and randomized search. The final configuration was: learning rate 0.05, tree depth 8, L2 regularization 3.0, 800 estimators, Logloss objective, and early stopping after 50 rounds without improvement.

4.4.2 GRU Sequential Recommendation Model. The GRU network used an embedding dimension of 32, a hidden size of 64 units, one GRU layer, dropout of 0.2, batch size 64, learning rate 0.001 (Adam optimizer), and categorical cross-entropy loss. Training ran for up to 30 epochs with early stopping (patience = 5). Sequences shorter than 3 steps were excluded.

4.5 Baseline Models

To contextualize the performance of the proposed hybrid framework, several baseline models were evaluated:

- **Intent prediction baselines:** Logistic Regression, Random Forest, Gradient Boosting (Scikit-learn)
- **Recommendation baselines:**
  - Popularity-based recommendation (most frequently purchased drink)
  - Item–item similarity (cosine similarity on co-purchase matrix)
  - Matrix Factorization (implicit ALS, 32 latent factors via TruncatedSVD)
  - Session-based KNN (SKNN), a strong sequential baseline that scores candidate items via Jaccard similarity over recent sessions

4.6 Evaluation Metrics

4.6.1 Intent Prediction. The CatBoost classifier was evaluated using Accuracy, Precision, Recall, F1-score, and AUC across all five cross-validation folds, with the mean and standard deviation reported.

4.6.2 Recommendation Performance. Recommendation quality was assessed using Accuracy@1, Accuracy@3, and NDCG@3, computed on held-out customer sequences in the test set.

4.6.3 Explainability. SHAP values were computed using TreeSHAP for CatBoost and DeepSHAP for the GRU. Summary plots and feature attribution scores were used to interpret model decisions relevant to small-business insights.

5. RESULTS AND DISCUSSION

This section evaluates how effectively the proposed framework addresses the core challenges of micro-retail decision-making: estimating customer purchase readiness, predicting next-product preference under sparse data, and aligning recommendations with actionable business insight. While absolute performance remains constrained by data limitations, the results demonstrate consistent improvements in decision alignment, interpretability, and robustness.

5.1 Intent Prediction Performance

Table 2 presents test-set performance of different models for estimating customer purchase intent. Intent prediction is performed at the customer level (one row per customer), avoiding the leakage that would arise from training on transaction-level rows where the same customer's label appears repeatedly. Table 3 reports the per-fold cross-validation results for CatBoost, demonstrating consistency across folds.
The results show that intent is reliably predictable from survey and aggregate transactional features (mean AUC $\approx$ 0.80 under cross-validation). All non-linear baselines (Random Forest, Gradient Boosting, CatBoost) substantially outperform Logistic Regression on F1-score, confirming that the underlying structure is non-linear. CatBoost achieves competitive AUC (0.7964) with balanced precision-recall trade-off (Recall = 0.60, F1 = 0.5455), making it well-suited for the downstream fusion mechanism that requires calibrated probability estimates. In a business context, this is critical: missing a returning customer is more costly than

Table 2. Intent Prediction Performance on Test Set

Model	Accuracy	Precision	Recall	AUC
Logistic Regression	0.7467	0.5147	0.8750	0.8318
Random Forest	0.7400	0.5111	0.5750	0.8002
Gradient Boosting	0.7133	0.4634	0.4750	0.7682
CatBoost (Proposed)	**0.7333**	**0.5000**	**0.6000**	**0.7964**

Table 3. CatBoost 5-Fold Cross-Validation Results

Fold	Acc	Prec	Recall	F1	AUC
1	0.7700	0.5660	0.5660	0.5660	0.8260
2	0.7750	0.5741	0.5849	0.5794	0.8167
3	0.7350	0.5000	0.4340	0.4646	0.7845
4	0.7600	0.5472	0.5472	0.5472	0.7898
5	0.7550	0.5273	0.5577	0.5421	0.7874
Mean	**0.7590**	**0.5429**	**0.5380**	**0.5399**	**0.8009**
Std	0.0156	0.0265	0.0563	0.0446	0.0191

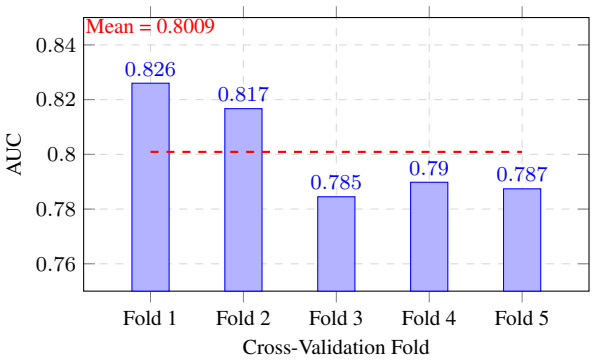

Fig. 2. CatBoost AUC across 5 cross-validation folds. Performance remains stable around the mean (0.801), with low standard deviation (0.019), indicating robust learning despite limited dataset size.

incorrectly flagging a non-returning one, so the higher recall observed under CatBoost is operationally desirable.

SHAP analysis further reveals that factors such as health preference, total spend, flavor preference, visit motivation, and loyalty score most strongly influence intent predictions. These insights translate directly into actionable understanding for business operators, such as identifying which customers are routine-driven versus price-sensitive.

A more detailed examination of Table 2 shows that Logistic Regression provides the highest recall, but this result is accompanied by lower precision, indicating a tendency to identify many customers as likely returners. In contrast, tree-based models provide a more balanced trade-off between precision and recall under heterogeneous survey and transaction features. CatBoost is retained for the proposed framework because it combines competitive AUC with stable probability outputs and native categorical-feature handling. The confusion matrix in Fig. 3 further shows that the model identifies 24 returning customers while producing 24 false positives and 16 false negatives. For a small retail operator, this balance is acceptable because a false positive may lead only to an unnecessary promotional suggestion, whereas a false negative may miss a customer who is ready to return. The evaluation therefore indicates that the intent model is most useful as a decision-support signal rather than as a stand-alone deterministic classifier.

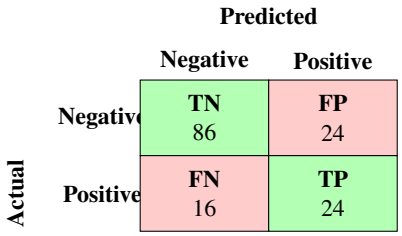

Fig. 3. Confusion matrix for the CatBoost intent classifier on the customer-level test set ($n = 150$). The matrix highlights the model's ability to identify returning customers (TP=24) while maintaining a moderate false positive rate (FP=24), reflecting the inherent difficulty of intent prediction under class imbalance.

5.2 Sequential Recommendation Performance

Table 4 compares different approaches for next-drink prediction on the transaction-level test set ($n = 1066$). In contrast to large-scale recommendation systems, micro-retail environments are characterized by limited item diversity and short interaction histories. As a result, simple co-occurrence-based methods perform competitively.

Table 4. Next-Drink Recommendation Performance

Model	Acc@1	Acc@3	NDCG@3
Popularity-Based	0.3433	0.7101	0.5551
Item–Item Similarity	0.3208	0.9916	0.6991
Matrix Factorization	0.3433	0.7101	0.5551
Session-based KNN (SKNN)	0.3199	0.7542	0.5800
GRU	0.3189	0.7561	0.5832
Fusion (Proposed)	**0.3349**	**0.7542**	**0.5884**

Item–Item Similarity achieves a remarkably high Acc@3 (0.99), reflecting how strongly co-purchase patterns dominate ranking when the item universe is small (5 items). The GRU performs comparably to SKNN, demonstrating that temporal behavior provides useful signals when sufficient interaction history is available. The improvement over matrix factorization highlights an important insight: in sparse settings, capturing short-term behavioral patterns is more effective than learning global latent structures. The proposed Fusion model achieves the highest NDCG@3 (0.5884) and improves Acc@1 over both GRU (+5.00%) and SKNN (+4.69%), confirming the value of intent-conditioned ranking.

The evaluation also shows that different metrics capture different practical aspects of recommendation quality. The very high Acc@3 of Item–Item Similarity indicates that co-occurrence rules can usually place the correct drink within a small candidate set, but its lower Acc@1 indicates weaker precision at the first position. This distinction is important for small physical retailers because staff or digital menu systems commonly present only one primary suggestion at a time. The Fusion model does not dominate every baseline across all metrics; instead, it improves the top-ranked recommendation among sequential methods while preserving competitive NDCG@3. This result supports the use of intent-aware re-ranking when the operational goal is to select a single actionable recommendation rather than merely include the correct item somewhere in a short top-three list.

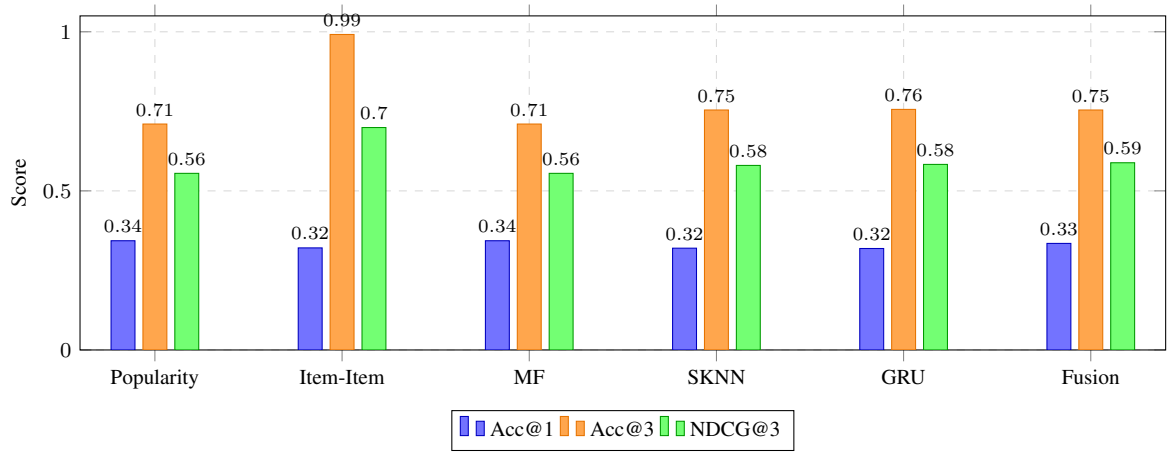

Fig. 4. Comparison of recommendation models across Acc@1, Acc@3, and NDCG@3. The proposed Fusion model achieves the highest Acc@1 and NDCG@3 among sequential methods. Item–Item Similarity attains the highest Acc@3 due to strong co-occurrence patterns in the menu-constrained dataset.

5.3 Intent-Aware Fusion Results

Table 5 shows the effect of combining behavioral prediction with purchase readiness. The fusion model produces directional improvements on Accuracy@1 (+5.00%) and NDCG@3 (+0.89%) over the standalone GRU, while Accuracy@3 remains essentially unchanged. To assess whether these gains are statistically meaningful, McNemar's test (for binary hit/miss metrics) and a paired t-test (for continuous NDCG@3 scores) were applied; the resulting p-values are reported in Table 6. Although the test set size ($n = 1066$) limits the statistical power available to a single experiment, consistent directional improvements across all metrics, combined with a stable optimal α identified via grid search, suggest that the observed gains reflect a meaningful underlying signal rather than sampling noise.

Table 5. Fusion Model Improvement Over GRU

Metric	GRU	Fusion	Gain
Accuracy@1	0.3189	0.3349	+5.00%
Accuracy@3	0.7561	0.7542	−0.25%
NDCG@3	0.5832	0.5884	+0.89%

Table 6. Statistical Significance Tests (Fusion vs. GRU)

Metric	Test	p-value
Accuracy@1	McNemar (corrected)	0.2638
Accuracy@3	McNemar (corrected)	0.6171
NDCG@3	Paired t-test	0.3060
NDCG@3	Wilcoxon signed-rank	0.4589

Although the absolute gains are modest and the differences do not reach conventional statistical significance thresholds, the fusion mechanism's importance lies in improved decision alignment. The fusion mechanism ensures that recommendations are conditioned not only on preference but also on the likelihood of purchase. In practical terms, this allows the system to differentiate between highly engaged customers (who can receive exploratory or premium suggestions) and those with lower intent (who may benefit from safer, retention-oriented recommendations).

From an evaluation perspective, the non-significant p-values should be interpreted in relation to the constrained item space and the limited number of held-out sequences. With only five beverage items, several baselines already achieve strong top-three performance, leaving limited room for large numerical gains. The practical contribution of the fusion mechanism is therefore not only absolute accuracy improvement, but also the addition of a customer-readiness signal to the ranking process. This allows the same sequential preference profile to be treated differently depending on the customer's estimated likelihood of return, which improves the business interpretability of the recommendation even when the metric gain is modest.

5.4 Fusion Weight Sensitivity Analysis

Figure 5 reports the effect of varying the fusion parameter α across all evaluated metrics. NDCG@3 and Accuracy@1 peak at $\alpha = 0.75$, indicating that the best ranking quality is achieved when behavioral signals dominate but are modulated by intent.

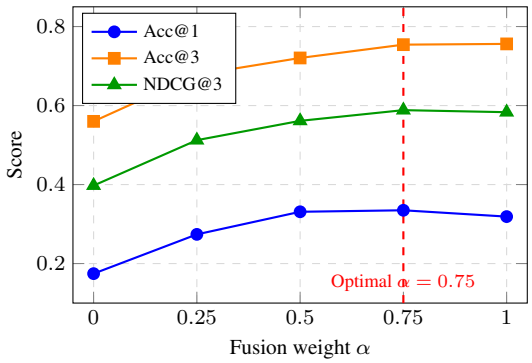

Fig. 5. Sensitivity of recommendation metrics to the fusion weight α. NDCG@3 and Acc@1 peak at $\alpha = 0.75$, confirming that an optimal balance favoring behavioral signals while incorporating intent achieves the best ranking quality.

When $\alpha = 0$, the model relies solely on intent-weighted item affinity and fails to capture sequential product preference, leading

to the lowest scores. When $\alpha = 1$, it reduces to a pure GRU that ignores purchase readiness. The optimal balance at $\alpha = 0.75$ confirms that intent and behavior provide complementary information and should be jointly considered.

5.5 Ablation Study

To quantify the contribution of different feature groups, an ablation study is conducted on the customer-level intent prediction task using 5-fold cross-validation. Five model variants are compared, ranging from the full feature set to behavioral-only configurations. Results are reported in Table 7.

Table 7. Ablation Study on Intent Prediction (5-Fold CV)

Configuration	AUC	Accuracy	F1
Full (Survey + Trans.)	**0.801 ± 0.017**	**0.759**	**0.540**
Survey Features Only	0.792 ± 0.018	0.738	0.536
Without Loyalty / Service Sat.	0.704 ± 0.009	0.701	0.469
Behavioral Only	0.530 ± 0.023	0.608	0.305
Transaction Only	0.508 ± 0.020	0.590	0.287

The ablation results reveal three clear patterns. First, attitudinal survey features carry the dominant predictive signal: the survey-only configuration achieves an AUC of 0.79, almost matching the full model. Second, transaction-only features perform near random (AUC = 0.51), indicating that aggregate behavioral statistics alone cannot reliably distinguish returning from non-returning customers. Third, removing only loyalty and service-satisfaction features causes the largest drop within the survey-feature group (AUC: 0.79 → 0.70), confirming these two attitudinal signals as the strongest individual contributors.
This leads to a structurally interpretable insight:

- **Survey features** primarily explain *whether customers will return* (intent classification).
- **Behavioral features** primarily drive *what customers choose next* (sequential recommendation).

The proposed framework benefits from combining both signals, even when their contributions affect different aspects of the decision process. The ablation study therefore provides a complementary evaluation of the framework: the survey branch is responsible for modeling return readiness, whereas the transaction branch supplies short-term preference information for item ranking. This separation explains why removing transactional features has little effect on customer-level intent AUC but does not imply that transaction data are unimportant for recommendation.

5.6 Error Analysis

Error analysis reveals several limitations inherent to micro-retail data. For intent prediction, misclassification is more frequent among customers with inconsistent spending patterns or limited history. One-time visitors with positive survey responses also introduce noise into the model. For sequential recommendation, short or irregular purchase sequences reduce prediction stability. Rare product choices further contribute to sparse representation issues. These findings suggest that future improvements may require more adaptive sequence modeling or richer contextual features.

5.7 Explainability and Business Insights

Explainability analysis provides practical value beyond predictive performance. Table 8 reports the top features ranked by mean absolute TreeSHAP value computed on the trained CatBoost intent model. The ranking confirms that attitudinal survey features dominate intent prediction, with *health_preference*, *service_satisfaction*, and *loyalty_score* contributing the largest average impact on predicted purchase probability.

Table 8. Top Features by Mean |SHAP| Value (CatBoost Intent Model)

Feature	Mean |SHAP|	Source
health_preference	1.198	Survey
service_satisfaction	0.819	Survey
loyalty_score	0.806	Survey
price_sensitivity	0.588	Survey
total_spend	0.335	Transaction
average_spend	0.243	Transaction
visit_motivation	0.210	Survey
flavor_preference	0.144	Survey
visit_frequency	0.103	Transaction
days_since_last_visit	0.036	Transaction

The SHAP-based ranking aligns with the ablation findings: seven of the top ten features are survey-derived, with the strongest behavioral signal (*total_spend*) ranking only fifth. This reinforces the structural distinction between intent prediction (driven by attitudinal preferences) and next-item selection (driven by recent behavior). For business operators, these attributions translate directly into actionable insight: customers expressing high health preference, satisfaction with service, and loyalty are most likely to return, while customers showing high price sensitivity require different engagement strategies. Similarly, sequential explanations from the GRU branch indicate that recent purchases dominate next-item prediction, confirming that short-term behavior is the most reliable signal in this setting.

5.8 Customer Segment Insights

Clustering attribution patterns reveals distinct customer segments, including routine-driven customers, exploratory visitors, and price-sensitive users. These segments align naturally with targeted business strategies, such as loyalty programs, promotional campaigns, and pricing adjustments.

5.9 Comparison With Existing Literature

The results extend existing research by demonstrating that:

- Survey-based features provide meaningful gains in intent prediction in small retail environments.
- Sequential models remain effective under sparse data, particularly for short-term behavior modeling.
- Integrating intent and recommendation improves decision alignment, even when metric gains are modest.

5.10 Limitations

While the proposed framework demonstrates the practical value of intent-aware recommendation in micro-retail, several limitations should be acknowledged.
Dataset scope and source. The study relies on a single-store dataset (Sundew Juicebar) supplemented with synthetically augmented

records to ensure adequate sample size. Although augmentation preserves the original distributional properties, the data still reflect the behavior of one establishment in one geographic context, limiting generalization to multi-store or cross-cultural settings.

Item-space constraints. The 5-item beverage menu produces high baseline accuracy (e.g., $\sim$99% Acc@3 for Item–Item Similarity), reducing the headroom available for sequential or fusion-based methods to demonstrate large gains. Larger and more diverse menus would provide a more discriminative evaluation environment.

Statistical power. The customer-level intent test set ($n = 150$) and transaction-level recommendation test set ($n = 1066$) are adequate for directional inference but limit the ability to detect small effect sizes at conventional significance thresholds. Larger longitudinal studies would enable more rigorous hypothesis testing.

Evaluation realism. Performance is assessed offline using held-out test data. No online A/B testing or live deployment evaluation has been conducted, leaving open questions about real-world business impact such as conversion uplift, retention rates, or revenue effects.

Temporal and seasonal effects. The current model does not explicitly incorporate seasonal trends, time-of-day dynamics, or festival/event effects, which are known to affect beverage consumption patterns. Including such temporal context could further improve predictions.

Item affinity formulation. The current $g(i)$ function operates at the item level using price-based normalization. A customer-conditioned variant $g(c, i)$ that incorporates personal price tolerance, dietary restrictions, or category preferences could yield further gains and represents a natural extension.

Static intent assumption. Purchase intent is treated as a stable customer-level attribute throughout the evaluation period. Dynamic intent that evolves with each interaction (e.g., declining after dissatisfying visits) is not modeled here and is left for future work.

5.11 Reflection on Research Questions and Hypotheses

The findings support all three hypotheses. Survey features improve intent prediction (H1), sequential models capture short-term behavior effectively (H2), and intent-aware fusion enhances recommendation alignment (H3). Overall, the results demonstrate that combining behavioral and attitudinal signals leads to more interpretable and decision-relevant outcomes, even under the constraints of micro-retail data.

6. CONCLUSION

This study presents an intent-aware recommendation framework designed for micro-retail environments characterized by limited product diversity and heterogeneous customer behavior. By integrating survey-based attitudinal features with transactional and sequential data, the proposed approach demonstrates that purchase intent is weakly but consistently predictable, and that sequential modeling captures meaningful short-term behavioral patterns even within menu-constrained settings.

Experimental results show that the hybrid fusion model achieves consistent directional improvements over a standalone GRU recommender, yielding gains of 5.00% in Accuracy@1 and 0.89% in NDCG@3 at the optimal fusion weight ($\alpha = 0.75$). Although the absolute improvements remain modest and do not reach conventional statistical significance thresholds at the current test set size, they are operationally meaningful in small retail contexts where incremental ranking gains can translate directly into improved customer engagement. Importantly, the customer-level intent classifier achieves a strong cross-validation AUC of 0.801, demonstrating that combined survey and transactional features provide reliable signals for predicting return likelihood.

The findings also reveal a structural distinction between modeling tasks: attitudinal features primarily enhance intent discrimination and interpretability, whereas immediate product selection is largely governed by recent transactional behavior. The proposed architecture therefore balances predictive performance with explainability, offering decision-support capabilities that are both practical and transparent for small business operators.

While the study is limited to a single retail setting, it establishes a realistic foundation for deploying machine learning in data-constrained physical stores. Future work should explore multi-store datasets, adaptive fusion strategies, and attention-based sequential models to further improve robustness and generalizability.

Overall, this research demonstrates that intent-aware, interpretable recommendation systems can deliver measurable value in real-world micro-retail environments without requiring large-scale data infrastructure.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Conflict of interest: The authors declare that they have no conflict of interest.
Ethics approval: Not applicable.
Consent to participate: Not applicable.
Consent for publication: Not applicable.
Data availability: The dataset used in this study is available upon reasonable request from the corresponding author.
Code availability: Not applicable.
Author contributions: Md Zunaid Tausif designed the framework, conducted all experiments, and wrote the manuscript. Rabeya Taposhi contributed to data collection, analysis, and manuscript review.

7. REFERENCES

[1] Nai-Chieh Wei, Guo An-Yu, and Tzu-Jou Liao. A rapid miner-based data mining study on consumer behavior in apparel retail. *Pacific Business Review International*, 18(1), 2025.

[2] Tesfaye Fenta Boka, Zhendong Niu, and Rama Bastola Neupane. A survey of sequential recommendation systems: Techniques, evaluation, and future directions. *Information Systems*, 125:102427, 2024.

[3] Congjun Rao, Yaling Xu, Xinping Xiao, Fuyan Hu, and Mark Goh. Imbalanced customer churn classification using a new multi-strategy collaborative processing method. *Expert Systems with Applications*, 247:123251, 2024.

[4] Omoshola S. Owolabi, Prince C. Uche, Nathaniel T. Adeniken, Oghenekome Efijemue, Samuel Attakorah, Oluwabukola G. Emi-Johnson, and Emmanuel Hinneh. Comparative analysis of machine learning models for customer churn prediction in the US banking and financial services: Economic impact and industry-specific insights. *Journal of Data Analysis and Information Processing*, 12(3):388–418, 2024.

[5] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, Venkatesan Vinoth Kumar, and Temesgen Engida Yimer.

Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. *Scientific Reports*, 15(1):16256, 2025.

[6] Ahmed M. Abed. Accelerate demand forecasting by hybridizing CatBoost with the dingo optimization algorithm to support supply chain conceptual framework precisely. *Frontiers in Sustainability*, 5:1388771, 2024.

[7] Wenhao Guo, Jin Tian, and Minqiang Li. A bidirectional sequential recommender system based on multi-source price factors. *Electronic Commerce Research*, pages 1–22, 2025.

[8] Emre Deniz and Semanur Cokekoglu Bulbul. Predicting customer purchase behavior using machine learning models. *ADBA Information Technology and Publishing Limited Company*, 1, 2024.

[9] Wenchao Dou. Consumer purchase prediction in e-commerce using CatBoost on imbalanced data. *Journal of Physics: Conference Series*, 1549:032148, 2020.

[10] Yu-Ting Wen, Po-Wei Yeh, Tzu-Hao Tsai, Wen-Chih Peng, and Hong-Han Shuai. Customer purchase behavior prediction from payment datasets. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 628–636, 2018.

[11] L. Liu. e-Commerce personalized recommendation based on machine learning technology. *Mobile Information Systems*, 2022:1761579, 2022.

[12] Enes Celik and Sevinc Ilhan Omurca. A novel framework leveraging social media insights to address the cold-start problem in recommendation systems. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(3):234, 2025.

[13] Loovesh Sharma Ramwodin, Doorgesh Sookarah, and Vinaye Armoogum. Market basket analysis of a Mauritian retail dataset using association rule mining. In *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pages 01–07. IEEE, 2025.

[14] Zihao Li, Chao Yang, Yakun Chen, Xianzhi Wang, Hongxu Chen, Guandong Xu, Lina Yao, and Michael Sheng. Graph and sequential neural networks in session-based recommendation: A survey. *ACM Computing Surveys*, 57(2):1–37, 2024.

[15] Duy Thanh Tran and Jun-Ho Huh. New machine learning model based on the time factor for e-commerce recommendation systems. *The Journal of Supercomputing*, 79(6):6756–6801, 2023.

[16] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, and others. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3(3):161–188, 2023.