ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Modeling and Application of E-Commerce Consumer Purchase Intention Prediction Using Multi-Source Data

Qiong'ao Mei[1,*], *Huiyuan Zhang*[2] *and Yu Chen*[1]

[1] Economics and Management, Huaibei Institute of Technology, Huaibei 235000, Anhui, China

[2] Management Engineering, Suzhou College of Information Technology, Suzhou 215000, Jiangsu, China

SUMMARY: *Facing the problems of lagging identification of in-session transactions, sparse positive class samples and dispersed evidence from multiple sources on e-commerce platforms, this paper constructs a multi-source data-driven prediction model of consumers' purchase intention, MSF-PIN.The study takes session as the basic object, and uniformly organizes behavioral logs, product attributes, evaluative sentiment, price promotions, and access contexts into 162,840 session-level samples, among which 24,912 purchase samples account for 15.30% of positive class. Among them, there are 24,912 purchase samples, accounting for 15.30% of positive categories. At the model level, behavioral sequence coding, cross-source gating fusion, and click-add-purchase multi-task learning are introduced to improve the ability to portray the strength of short-term interest paths and heterogeneous evidence. Experiments are conducted using temporal order slicing and deployment-oriented evaluation protocols. The results show that MSF-PIN has an AUC of 0.902, a PR-AUC of 0.523, an F1 of 0.587, a Logloss of 0.140, and an ECE of 0.032, which is overall better than the comparative models of LightGBM, CatBoost, DeepFM, AutoInt, DCN V2, and BST. Results obtained by checking each scenario one by one indicate that the model obtains an AUC of 0.913 in the scenario of returning visitors and a PR-AUC of 0.541 in the scenario of promotion. The deployment analysis further gives out that when we screen the top 10% high-intention sessions through predictive probability, the model can cover 48.9% of actual purchase samples, and the average inference delay is 34.9 ms for each 1000 sessions. The outcome illuminates that sample-level weight redistribution using many kinds of proof can promote the utilization degree of purchase intention forecast, thus give a directly gotten probability foundation for recommendation weight adjustment, coupon start, and customer service help.*

KEYWORDS: *e-commerce purchase intention prediction; multi-source data; gated fusion; behavior sequence modeling; deployment-oriented evaluation*

1 Introduction

The e-commerce platform's judgment of purchase intention occurs within a short window when the user remains on the page and the transaction has not yet been completed. In a session, a user may first go from the search results page to the product details page, then check reviews, switch specifications, compare prices, collect coupons, add to cart, and finally complete payment; or he or she may leave the platform after completing the same number of browsing actions. The

*mei714846508@126.com
https://doi.org/10.65102/is2026764
Published: 30 April, 2026

two types of sessions are similar in early behavior, but correspond to different operational actions. The former is suitable for triggering inventory reminders, limited-time benefits, customer service assistance, or recommendation position reinforcement, while the latter is prone to increase reach costs and user resentment if it is frequently intervened. Platforms need to identify high-intention users while the session is still in progress, and invest limited marketing resources in the visit process closer to the transaction. Therefore, the core value of purchase intention prediction is to provide a usable probabilistic basis for real-time recommendation, coupon placement, and membership operation, so that operational actions can be sorted and stratified before transactions occur.

The sample structure in the publicly available data shows the underlying difficulty of this task.The Online Shoppers Purchasing Intention Dataset, with sessions as predictors, contains 12,330 samples, of which 1,908 are purchasing samples and 10,422 are non-purchasing samples, and positive class sparsity is a direct feature of the task [1]. RetailRocket data records view, addtocart, and transaction events, with a raw event size of over 2.7 million events, but far fewer transaction events than browsing events, and a highly uneven distribution of samples in the behavioral funnel [2].The Coveo SIGIR 2021 E-Commerce Workshop Data Challenge further puts query, click, product text, image, and price metadata into the same task context, illustrating that e-commerce purchase intent has become difficult to explain by page visits or clicks alone [3]. Together, these data point to a modeling premise: purchase intention consists of a combination of behavioral paths, product conditions, evaluation feedback, price stimuli, and access context, and changes dynamically within a single session.

Existing research has gradually shifted from manual statistical features to deep interaction modeling. Early approaches typically used structured variables such as number of pages visited, length of stay, bounce rate, exit rate, page value, visitor type, and special date for categorization, which are easy to deploy and easy to interpret, but have limited expression of event sequences and product semantics. Review sentiment analysis studies have shown that evaluation text can complement factors that are difficult to reflect directly in behavioral logs, such as commodity quality, trust and perceived risk [4]; multimodal consumer behavior prediction studies have also shown that co-modeling commodity attributes, behavioral data, and external semantic information results in a more complete representation of the user's state [5]. In the direction of recommendation and ad click rate prediction, DeepFM jointly learns low-order and high-order feature interactions through factor decomposers and deep networks [6], DIN incorporates target product-related interests into click rate modeling [7], and BST uses Transformer to capture dependencies in e-commerce behavioral sequences [8]. These approaches demonstrate the importance of feature interactions and behavioral sequences, and also provide a transferable model basis for purchase intention prediction.

Conversion rate and multitask learning studies further reveal the funnel relationship in purchase intention prediction. There is a recursive relationship as well as jumps, delays, and interruptions as users move from clicking to adding to purchasing.ESMM jointly models clicks and conversions across the entire sample space, mitigating sample selection bias and sparsity in post-click conversion rate estimation [9]. Multi-tasking architectures such as MMoE, PLE, and others deal with correlations and conflicts between tasks through shared experts, task gating, or hierarchical extraction [10, 11]. For e-commerce sessions, clicks and adds cannot be directly equated to purchases, but they can provide antecedent signals for purchase intent. If purchase prediction only uses final transaction labels, the model will be affected by both positive class sparsity and delayed conversion; if the task relationship between clicks, add-ons and purchases can be reasonably utilized, the training process will obtain more stable supervised information.

These advances do not completely solve the problems of multi-source purchase intention prediction in real applications. First, most models still tend to splice behavioral statistics,

product attributes, review sentiment, and promotional variables into a unified feature vector. Splicing can increase input information, but it is difficult to account for the strength of evidence from different sources in a single session. Older visitors typically accumulate more complete browsing and up-selling paths, and newer visitors rely more on product price, review sentiment, and promotion exposure. Fixed fusion weakens this sample-level difference. Second, purchase intention is characterized by a distinct session order. A user returning to the same item consecutively, checking the specifications and then adding a purchase may have similar number of views and length of stay, but not the same purchase probability, as a user randomly jumping between multiple categories. Using only static aggregation variables compresses the recursive relationships in the event path. Third, while many offline experiments emphasize AUC or accuracy, platform deployment also needs to consider high-willingness population coverage, false-touch rate, probability calibration, and inference latency. If the model output probability lacks calibration, it is difficult to stably support coupon thresholds, customer service trigger thresholds, and recommendation position tuning thresholds, even if the sorting capability is better.

In the actual system, these deficiencies are further amplified. The recommendation link needs to complete candidate recall, sorting and reordering when the user refreshes the page, while the marketing link needs to decide whether to issue a coupon, what kind of coupon to issue, and whether to trigger the customer service portal synchronously based on the predicted probability. If the model only outputs the sorting score, it is difficult for the operation side to determine the reach threshold; if the model cannot explain the source of high scores, it is also difficult for the product, price and service teams to determine the direction of subsequent intervention. For multi-source purchase intention prediction, model structure, probability quality and deployment cost need to enter the research design simultaneously.

Based on the above gaps, this paper limits the research problem to session-level, multi-source data conditions for e-commerce consumer purchase intention prediction. This paper focuses on three specific types of questions: how multi-source fields can be organized into trainable session samples; how behavioral sequences, product attributes, evaluation sentiment, and marketing context can be dynamically fused at the sample level; and how prediction results can enter into operational decision-making through calibration, sub-scenario evaluation, and deployment threshold analysis. Around these issues, this paper constructs MSF-PIN models to map behavioral logs, product attributes, evaluative sentiment, price promotions, and access contexts into a unified session object, and generates purchase intention representations through in-source coding, behavioral sequence modeling, and cross-origin gating fusion, and introduces click-, add-, and purchase-related tasks to enable the final purchase prediction to utilize the antecedent behavioral signals.

Focusing on the above tasks, this paper develops the following work. First, to establish the organizational caliber of multi-source e-commerce data to session-level samples, to incorporate behavioral events, product fields, evaluation sentiment, marketing context and transaction labels into the same prediction object, and to reduce the problem of variables from different sources being coarsely mixed before modeling. Second, a cross-source gated purchase intention prediction model is constructed to enable the model to adjust the sources of evidence for old visitors, new visitors, promotional sessions, and high-priced item sessions, and to preserve sequential information in the behavioral path. Third, a deployment-oriented evaluation design is used to examine probabilistic calibration, sub-scenario performance, high willingness population coverage, inference latency, and error sources in addition to overall performance so that model evaluation can correspond to recommendation, coupon, and customer service triggering scenarios in e-commerce platforms.

2 Methods

2.1 Multi-source data organization and session-level sample construction

In this paper, sessions are used as the basic prediction object. A session corresponds to a collection of behaviors around a product generated by the same visitor within consecutive visit windows, with labels determined by whether a purchase event occurs before the end of the session. The data organization first deals with object boundaries and then with field origins. online Shoppers data provides session-level variables such as page visits, dwell time, bounce rate, exit rate, page value, visitor type, and special dates; RetailRocket data provides view, addtocart, and transaction event sequences; Coveo Coveo data complements search queries, product text, images, and pricing metadata. The three types of public data are not cross-platform user spliced, and are uniformly mapped to the session-level field space for building the caliber of experiments for multi-source purchase intention prediction. The process of field alignment and label generation from multi-source data to session samples is shown in Figure 1.

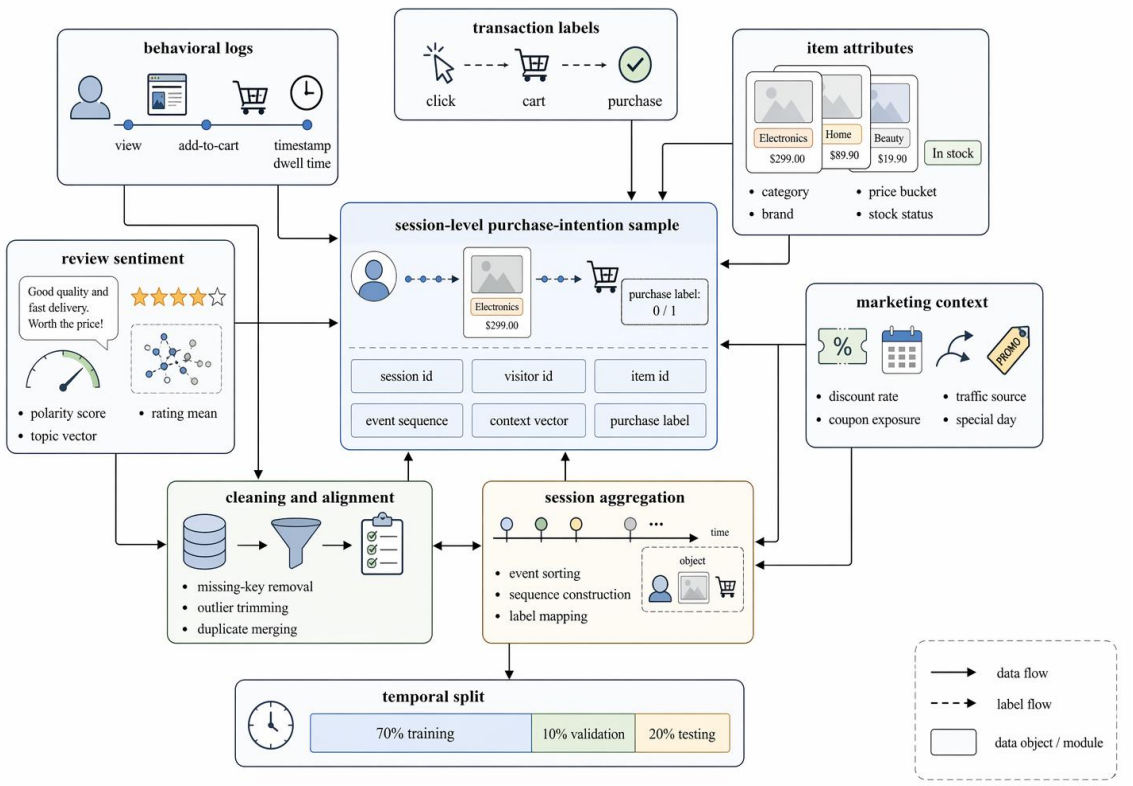

Figure 1: Multi-source data organization and session-level sample construction.

In Figure 1, behavior logs, product attributes, evaluation sentiment, marketing context and transaction tags enter the sample construction process as independent sources, respectively. Behavioral logs provide event sequence and dwell intensity, product attributes provide price, category, and inventory conditions, evaluation sentiment complements product word-of-mouth, marketing context records discounts, coupons, and access channels, and transaction tags are used to generate click, add, and purchase tasks. The cleaning module removes records with missing session id, item id, or timestamp before sample formation, and truncates visit duration, price, and discount rate by 1% and 99% bins. The session aggregation module sorts events by timestamp, preserves the original behavioral sequence, and generates the number of views, additions, repeat visits in the same category, dwell time on the product page, location of the

most recent additions, and the last visit action. The transaction labeling module maps in-session transaction or purchase events to purchase labels, addtocart events to add-purchase labels, and product detail page visits to click labels. Multiple source variables with modeling roles are shown in Table 1.

Table 1: Multi-source variables and modeling roles

Data Source	Main Fields	Model Representation	Modeling Function
Behavior Log	view, add to cart, timestamp, session length, dwell time	Behavior sequence embedding, time interval vector	Describes session path, interest intensity, and recent actions
Product Attributes	item id, category, brand, price bucket, stock status	Category embedding, standardized continuous variables	Describes product content, price tiers, and supply status
Review Text	review polarity, topic vector, rating mean, rating variance	Sentiment score, topic embedding	Supplements product reputation, quality perception, and risk signals
Marketing Context	discount rate, coupon exposure, traffic source, special day, weekend	Context embedding, continuous variables	Describes promotional stimuli, channel sources, and access environment
Transaction Labels	click, cart, purchase	Multi-task binary classification labels	Forms the prediction targets for clicks, add to cart, and purchases

Fields in Table 1 are not directly merged into a single wide table. Behavioral logging preserves the order of events, with item attributes, evaluation sentiment, and marketing context going into separate in-source encoders. This treatment preserves the semantic boundaries of the different sources and facilitates subsequent analysis of the model's sources of evidence for new visitors, returning visitors, promotional sessions, and high-priced item sessions. For category fields, this paper uses trainable embedding vectors; for continuous fields, outlier truncation and normalization are completed first; for evaluation texts, sentiment polarity, topic vectors and rating statistics are generated offline first, and only pre-calculated results are read in the online prediction stage, to avoid text encoding slowing down the inference chain. In-source characterization, as shown in equation (1).

$$e_i^{(m)} = \phi_m\left(x_i^{(m)}\right), \quad m \in \mathcal{M} \tag{1}$$

where $x_i^{(m)}$ denotes the original input of session i in class m data source, $\phi_m(\cdot)$ denotes the encoding function corresponding to this data source, $e_i^{(m)}$ denotes the encoded in-source representation, and $\mathcal{M}$ denotes the set of behavioral, commodity, evaluation, and contextual data sources. This formula corresponds to the in-mode representation in Table 1, which ensures that each type of data is encoded in its own semantic space before entering the cross-source fusion module.

The samples are divided in chronological order rather than randomly cut. All samples are sorted by session start time, with the first 70% used for training, the middle 10% for validation, and the last 20% for testing. The validation set is used for early stopping, category weight adjustment, temperature calibration, and Top-k threshold selection, while the test set retains the

original purchase percentage. This treatment is consistent with the platform's deployment conditions of using historical sessions to predict future sessions, and also reduces the risk of leaking the same user's later behavior to the training set. Session-level sample construction and cutoff statistics, as shown in Table 2.

Table 2: Session-level sample construction and split statistics

Data Scope	Sessions	Positive Purchases	Positive Ratio	Main Purpose
Full Session Samples	162,840	24,912	15.30%	Overall training and evaluation
Training Split	113,988	17,526	15.37%	Parameter learning
Validation Split	16,284	2,468	15.16%	Early stopping, thresholds, calibration
Test Split	32,568	4,918	15.10%	Final reporting
New Visitor Subset	41,306	4,902	11.87%	Cold start scenario
Returning Visitor Subset	69,521	12,114	17.43%	Historical behavior scenario
Promotion Subset	52,744	9,681	18.36%	Price stimulation scenario
High-Price Item Subset	37,908	4,716	12.44%	High average order value scenario

In Table 2, the proportion of purchase positive class for the complete sample is 15.30% and the proportion of purchase positive class for the test set is 15.10%, and the distribution of samples in the training, validation, and testing phases stays close to each other. The purchase proportion of the new visitor subset is 11.87%, which is lower than the 17.43% of the old visitor subset, indicating that historical behaviors have direct value for purchase judgments. The purchase proportion of the promotion subset is 18.36%, which is higher than the overall level; the purchase proportion of the high-priced goods subset is 12.44%, which is lower than the overall level. These differences provide the necessary basis for subsequent sub-scenario evaluations.

2.2 Multi-source gated purchase intention prediction model

The model structure of MSF-PIN is designed around session paths and cross-source evidence allocation. Session recommendation studies have illustrated that short-term behavioral sequences have a direct effect on preference judgments of anonymous or weakly historical users.DIN, DIEN, and BST improve user representations in terms of target item-related interests, interest evolution, and e-commerce behavioral sequence modeling, respectively; and SASRec and BERT4Rec have shown that self-attention structures can capture dependencies in sequences [12, 13]. In this paper, we compress sequence modeling into session-level purchase tasks, focusing on preserving the local paths between browsing, staying, adding, returning to the detail page, and promotional page visits.The cross-origin gating structure of MSF-PIN is shown in Fig. 2.

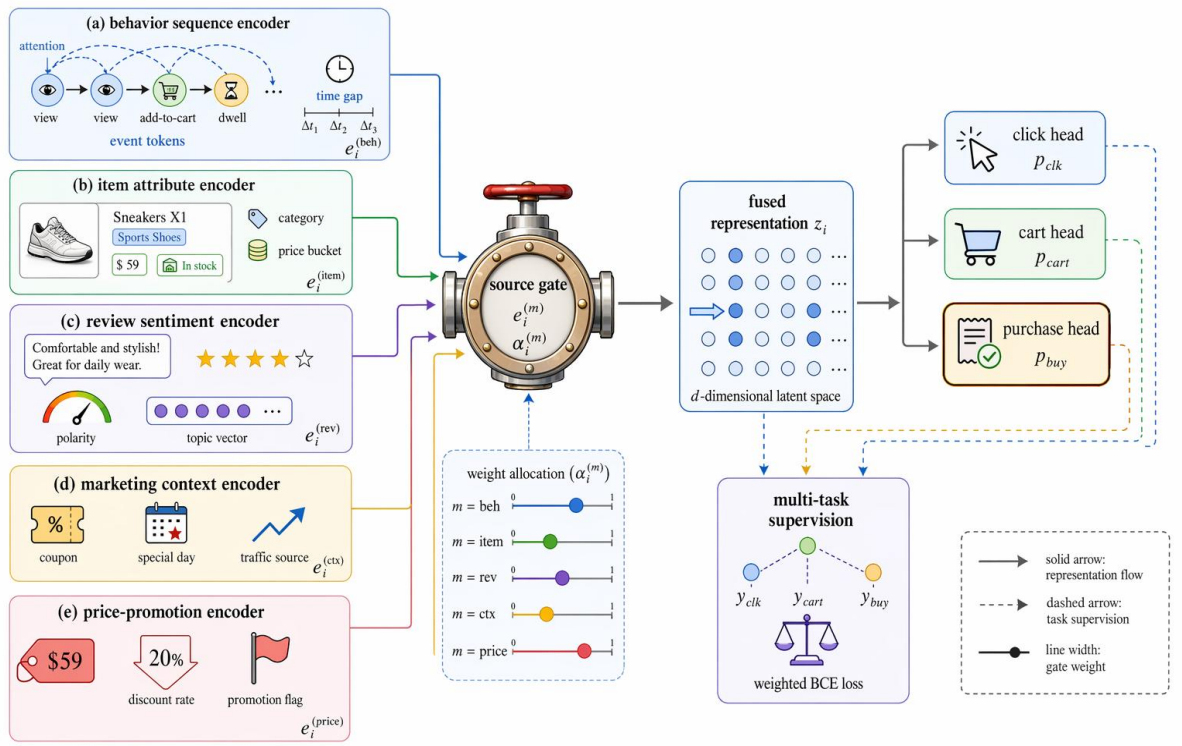

Figure 2: Cross-source gated fusion architecture of MSF-PIN.

In Fig. 2, the model consists of an in-source encoder, a behavioral sequence encoder, a source gate, a fusion representation, and a multitasking output header. The behavioral sequence encoder handles event variables such as view, addtocart, timestamp, and dwell time; the product attribute encoder handles category, brand, price bucket, and stock status; the rating sentiment encoder handles polarity score, topic The product attribute encoder handles category, brand, price bucket and stock status; the rating sentiment encoder handles polarity score, topic vector and rating statistics; and the marketing context encoder handles coupon exposure, traffic source, special day and discount rate. each encoder outputs an in-source representation, which is then fed into the source gate to compute the sample-level gating weights. The fused representations connect the click, cart, and purchase output headers, where purchase is the main task and click and cart provide antecedent behavioral supervision.

The behavioral sequence encoder models event types, locations, and time intervals separately. Event type embeddings are used to distinguish between browsing, adding and pre-purchase behaviors, location embeddings preserve the order of actions, and time interval embeddings record the waiting time between adjacent actions. The self-attention layer is used to determine which events in the session are more important for purchase intention. Successive paths to the same product detail page, adding a purchase after checking reviews, and returning to the price page after adding a purchase are usually closer to transaction readiness than a single browse. The specific session path characterization is shown in equation (2).

$$h_i^{(beh)} = Attn\left(u_{i,1} + r_{i,1} + t_{i,1}, \ldots, u_{i,L_i} + r_{i,L_i} + t_{i,L_i}\right) \tag{2}$$

where $u_{i,1}$ denotes the event type embedding of the lth event in the ith session, $r_{i,1}$ denotes the location embedding, $t_{i,1}$ denotes the time interval embedding, L_i denotes the event length of the session, $Attn(\cdot)$ denotes the self-attention coding function, and $h_i^{(beh)}$ denotes the

session path representation after encoding of the behavior sequence. This representation is used as an input to the behavior source when entering source-level fusion.

In terms of feature interactions, Wide & Deep, DeepFM, xDeepFM, AutoInt, DCN V2 and FiBiNET provide references for joint modeling of memory generalization, low-order and high-order interactions, explicitly compressed interactions, self-attentive feature interactions, cross-networks, and feature importance modeling, respectively [14-19]. dlrm and ncf provide common structural backgrounds for sparse category features and user-goods interactions [20-21]. and user-goods interactions provide a common structural context [20-21]. These methods are capable of learning complex field relationships, but most structures use fixed fusion for evidence from different sources. The weight of evidence in e-commerce sessions is not stable. Older visitors tend to rely on history paths and up-sell actions, while newer visitors rely more on product attributes, evaluation sentiment, and discount exposure; price variables are more sensitive in promotional sessions, and evaluations and after-sale information are more sensitive in high-priced item sessions.MSF-PIN uses source gating to assign sample-level weights to different data sources. The gating weights, as shown in equation (2).

$$\alpha_i^{(m)} = \frac{exp\left(q^\top tanh\left(W_m e_i^{(m)} + b_m\right)\right)}{\sum_{r\in\mathcal{M}} exp\left(q^\top tanh\left(W_r e_i^{(r)} + b_r\right)\right)} \tag{3}$$

where $\alpha_i^{(m)}$ denotes the gating weights of the class i data sources in session i, q is the learnable query vector, W_m and b_m are the transformation parameters of the class m data sources, and $e_i^{(m)}$ is the within-source representation.Softmax normalization allows the weights of the sources to be compared in the same session. The gating weights are not global constants but vary with samples. Cross-source fusion characterization, as shown in equation (3).

$$z_i = \sum_{m\in\mathcal{M}} \alpha_i^{(m)} e_i^{(m)}, \quad p_i^{(k)} = \sigma(w_k^\top z_i + b_k), \quad k \in \{clk, cart, buy\} \tag{4}$$

where z_i denotes the cross-source fusion representation of the ith session, $p_i^{(k)}$ denotes the prediction probability of task k, $\sigma(\cdot)$ is the Sigmoid function, w_k and b_k are the task output layer parameters, and clk, cart, and buy denote the click, add, and buy tasks, respectively. The fusion representation first absorbs sample-level weights from different sources before entering the multi-task prediction layer.

Purchase intention belongs to a funneled behavioral task. There are recursive relationships between click, add and buy, as well as jumps, delays and interruptions.ESMM jointly models clicks and conversions through the whole sample space to alleviate sample selection bias in CVR estimation [22]; MMoE, PLE and FDN improve multi-task recommendation modeling from the perspectives of task-relationship gating, shared-structure separation and negative migration control [23-25]. In this paper, click, cart and buy are set as related tasks, where buy is the main task, and cart and click are used for stabilizing the antecedent behavioral supervision.T2G-Former's organization of heterogeneous form feature relationships also illustrates that the relationships between complex fields need to be modeled explicitly, and cannot be completely relied upon for primitive splicing [26].

$$\mathcal{L} = \sum_{k\in\{\text{clk,cart,buy}\}} \lambda_k \left[-y_i^{(k)} \log p_i^{(k)} - \left(1 - y_i^{(k)}\right) \log\left(1 - p_i^{(k)}\right)\right] + \eta\|\Theta\|_2^2 \tag{5}$$

where $\mathcal{L}$ denotes the training objective, λ_k denotes the loss weight of task k, $y_i^{(k)}$ denotes the true label of the ith session on the task k, $p_i^{(k)}$ denotes the prediction probability,

η denotes the regularization coefficient, and Θ denotes all trainable parameters. During training, the purchase task is set to have the highest weight, followed by the add purchase task, and the click task is used to stabilize the early behavioral representation. For the purchase positive class sparse problem, the weight of positive class samples is increased in the training phase, and the original positive and negative ratios are maintained in the testing phase.

2.3 Experimental protocol, baselines and evaluation metrics

The experimental protocol is constrained with historical sessions predicting future sessions. The entire sample is sorted by session start time and then cut into training set, validation set and test set. The training set is used for parameter learning, the validation set is used for early stopping, threshold selection, and probability calibration, and the test set is used only for the final report. This protocol avoids future behavior leakage caused by random cuts and is closer to the actual usage of rolling training and online prediction in e-commerce platforms. The experimental setup, comparison model and evaluation interface, are shown in Fig. 3.

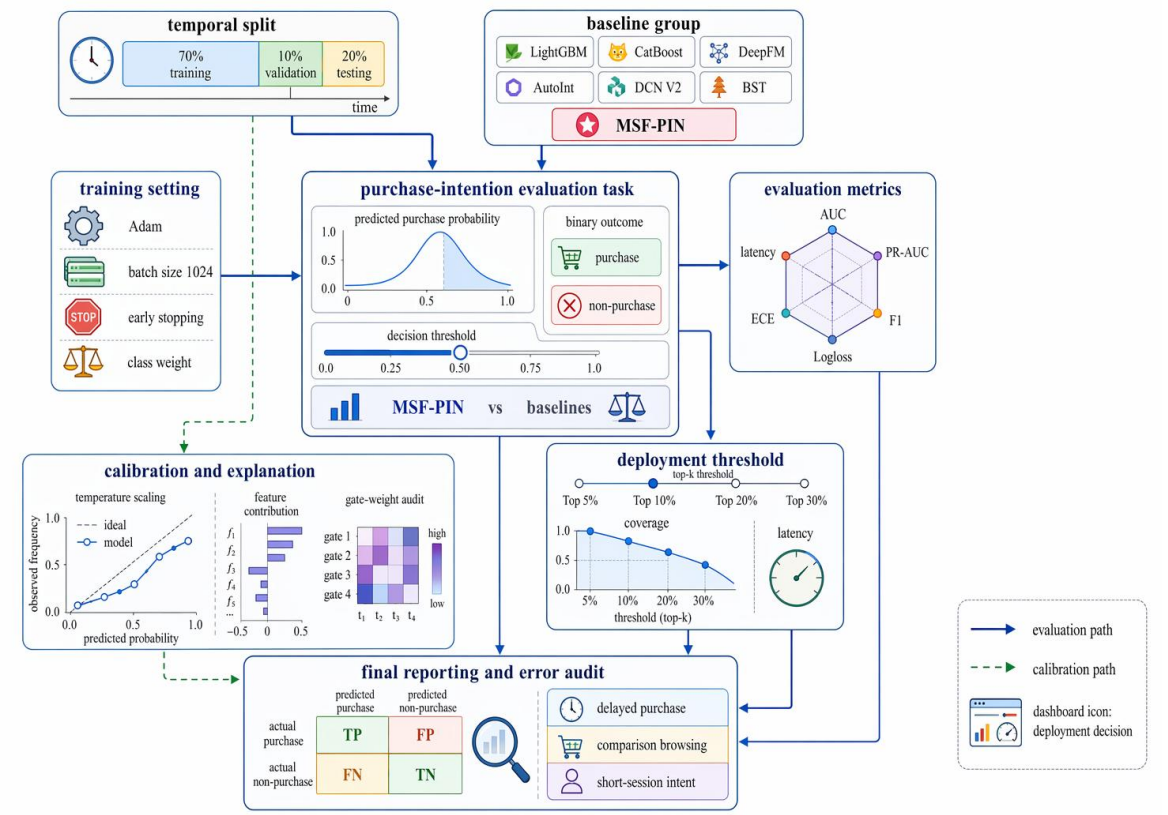

Figure 3: Experimental protocol and deployment-oriented evaluation interface.

In the Fig. 3, the temporal split carries out the control for training, validation and test objects; The baseline group carries out the same evaluation work together with MSF-PIN; training arrangement includes Adam, batch size, early stopping and category weight; evaluation index outputs AUC, PR-AUC, F1, Logloss, ECE and time delay; The work of calibration and explanation carries out reading on the probability of the validation set and temperature calibration, and therefore outputs feature contribution and gating. The metrics this method outputs include AUC, PR-AUC, F1, Logloss, ECE, and also latency; calibration and explanation reads the probability of validation set, carries out temperature calibration work, and outputs the check of feature contribution and gating weight; The deployment threshold makes

the transformation of the test set's predicted probability into Deployment threshold carries out the conversion of the predicted probability of the test set into strategies that reach Top 5%, Top 10%, Top 20% and Top 30%. The final report has in it overall performance, sub-scenario performance, ablation results, efficiency results and error sources.

The comparison models are divided into three groups. The first group is Logistic Regression, LightGBM and CatBoost. Logistic Regression is used to provide a linear classification baseline; LightGBM improves the efficiency of large-scale tree model training through gradient one-sided sampling and mutually exclusive feature bundling [27]; CatBoost reduces the risk of target leakage through ordered boosting and category feature processing to reduce the risk of target leakage [28]. The second group is Wide & Deep, DeepFM, xDeepFM, AutoInt and DCN V2 for comparing different feature interaction structures. The third group, BST, is used to test the usefulness of behavioral sequence modeling for purchase intention prediction. All models use the same training, validation and test sets, with purchase labels kept consistent and no negative sample downsampling in the test set.

The evaluation metrics cover ranking ability, positive class retrieval ability, probability quality, and deployment cost. auc measures the overall ranking ability, PR-AUC is suitable for positive class sparse scenarios, F1 reflects the balance between precision and recall at a fixed threshold, logloss measures the probability estimation error, ECE measures the calibration deviation between the predicted probability and the true purchase proportion, latency records the average inference time per 1,000 sessions, and latency records the average inference time per 1,000 sessions. Neural network probabilistic outputs may be overconfident; in this paper, we perform temperature calibration on the validation set and report the calibrated ECE on the test set [29]. Feature interpretation is performed using a joint analysis of local contributions and source gating weights, with local contributions referring to SHAP's idea of additive interpretation [30], and source gating weights used to determine which type of data sources the model relies more on in different scenarios.

3 Results and Discussion

3.1 Overall prediction performance and probability calibration

This section first examines the overall performance of the model on the full test set. Willingness-to-buy prediction faces the positive class sparsity problem, where accuracy is easily dominated by negative class samples, so the main results do not use accuracy as the core metric, but report both AUC, PR-AUC, F1, Logloss, ECE, and inference delay. Overall performance is used to answer whether the model has better ranking ability and probability quality, and the calibration results are used to determine whether the model can enter a threshold-triggered operational link. The sorting, positive class retrieval and probability calibration performance of the main model and the comparison model are shown in Fig. 4.

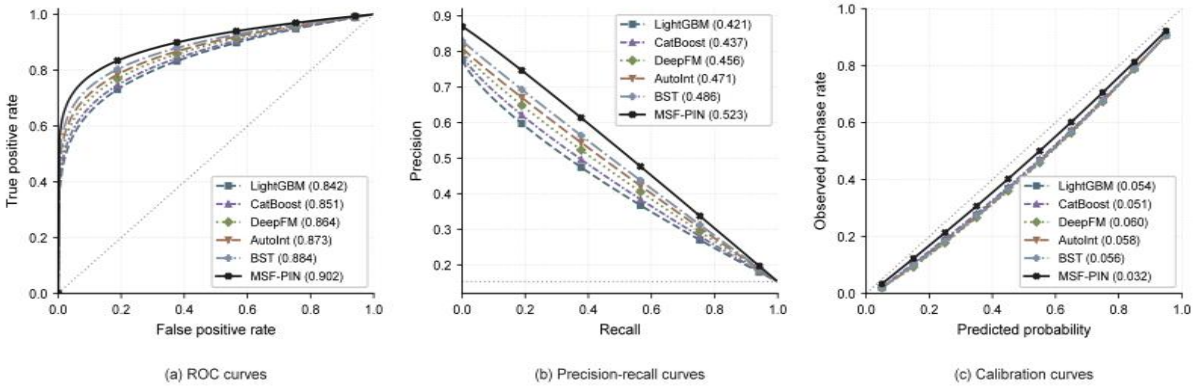

Figure 4: Overall ROC, precision-recall and calibration curves.

In Fig. 4, the ROC curve of MSF-PIN is overall located above each baseline model, the PR curve decreases more slowly in the high-precision interval, and the calibration curve is closer to the ideal diagonal. The overall prediction performance of different models, as shown in Table 3.

Table 3: Overall predictive performance of baseline models and MSF-PIN

Model	AUC	PR-AUC	F1	Logloss	ECE	Latency ms/1000 sessions
Logistic Regression	0.742	0.318	0.386	0.214	0.071	3.8
LightGBM	0.842	0.421	0.489	0.174	0.054	11.6
CatBoost	0.851	0.437	0.501	0.169	0.051	14.2
Wide & Deep	0.858	0.447	0.510	0.166	0.064	19.7
DeepFM	0.864	0.456	0.519	0.163	0.060	22.8
xDeepFM	0.868	0.462	0.526	0.161	0.059	28.5
AutoInt	0.873	0.471	0.532	0.158	0.058	31.4
DCN V2	0.879	0.479	0.541	0.155	0.055	29.8
BST	0.884	0.486	0.548	0.153	0.056	39.6
MSF-PIN	0.902	0.523	0.587	0.140	0.032	34.9

In Table 3, the AUC of MSF-PIN is 0.902, which is 0.018 higher than that of BST (0.884), 0.023 higher than that of DCN V2 (0.879), and 0.060 higher than that of LightGBM (0.842).The increase in AUC suggests that the model is able to more consistently prioritize purchased sessions over unpurchased sessions in the overall ordering.The PR-AUC is more sensitive to the positive class sparse task. The PR-AUC of MSF-PIN is 0.523, an improvement of 0.037 over BST, 0.052 over AutoInt, and 0.067 over DeepFM. This result suggests that behavioral sequence coding and cross-source gating are more effective for front-ranking recall of real purchase sessions.

The calibration score plots in Fig. 4 are consistent with the ECE results in Table 3. the ECE of MSF-PIN is 0.032, which is lower than CatBoost's 0.051, BST's 0.056, and DeepFM's 0.060. this result suggests that the model output probability is closer to the true purchase proportion. Scenarios such as coupon triggering, customer service assistance and inventory reminder rely on probability thresholds, and it is difficult for the operation side to set stable reach rules if the probability scale deviation is too large.The Logloss of MSF-PIN is 0.140, which is also lower than that of all the comparison models, suggesting that the model improves the probability estimation error outside of the sorting.

The differences between the models are consistent with their structural capabilities. The tree

model is more stable to structured variables, but cannot directly retain the event order; DeepFM, xDeepFM, and AutoInt are able to learn feature interactions, but have limited use of local paths between browsing, staying, and adding purchases; and BST introduces behavioral sequences that significantly outperform the general feature interaction model, but it does not have sufficient sample-level weight adjustment for evaluating sentiment, price promotions, and contextual variables. The advantage of MSF-PIN comes from two components: behavioral sequence encoding preserves session paths, and source gate redistributes different sources of evidence by sample state.

The latency outcomes do not reduce the deployment possibility of this model. The inference time delay of MSF-PIN is 34.9 ms each 1000 sessions, which is more low than that of BST at 39.6 ms each 1000 sessions and more high than that of DCN V2 at 29.8 ms each 1000 sessions. Taking into account the circumstance that MSF-PIN utilizes behavior sequences, evaluation sentiment, and marketing contexts at the same time, this latency can still be put into the model. Because MSF-PIN at the same time uses behavior sequences, evaluation feelings and marketing backgrounds, this time delay still lets people obtain real-time recommendation or similar real-time marketing connections. The situation that the evaluation text is not recoded in the online stage, but is put into the feature table by offline sentiment scores and topic vectors, is therefore the main cause for the controllable time delay.

3.2 Scenario-level performance, ablation and feature contribution

The full test set can only illustrate the average effect and cannot determine under which business conditions the model gains more. This section splits the samples by visitor type, promotion status, price tier, and session length, and analyzes the sources of model gains in combination with ablation heatmaps and feature contributions. The split-scenario results answer whether the model is effective only in some scenarios, the ablation results answer whether the core module actually produces an effect, and the feature contributions and gating weights are used to explain the sources of evidence for the model. The prediction differences under different visitor types, promotion statuses, price tiers, and session lengths are shown in Figure 5.

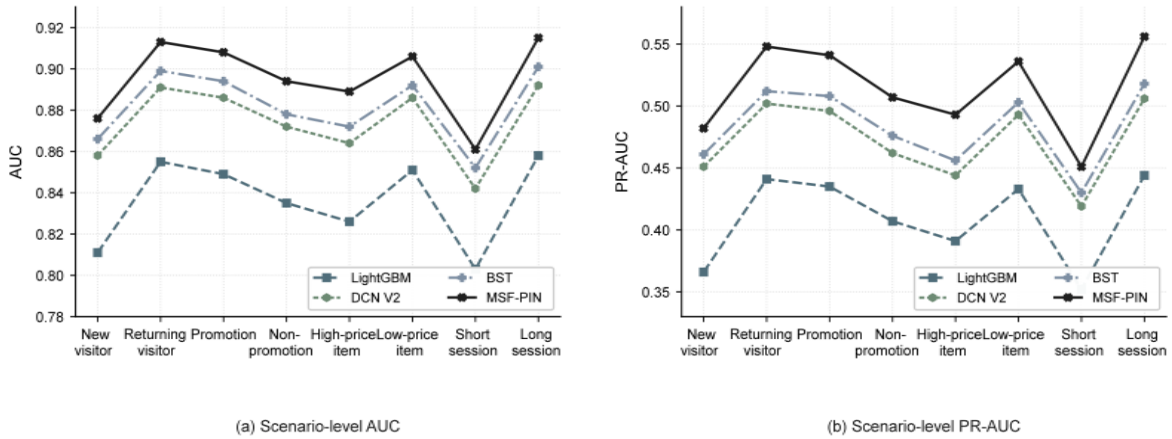

Figure 5: Scenario-level performance under visitor, promotion and price contexts.

In Figure 5, MSF-PIN has an AUC of 0.913 and a PR-AUC of 0.548 in the returning visitor scenario, and an AUC of 0.876 and a PR-AUC of 0.482 in the new visitor scenario. old visitors have a more complete session trajectory and a more stable behavioral pattern, and the browsing paths, add-purchase locations, and repeated visits in the same category can provide direct evidence for purchase judgments. Old visitors have more complete session trajectories and

more stable behavioral patterns, with browsing paths, purchase locations and repeated visits in the same category providing direct evidence for purchase judgments. New visitors lack historical behavior, and the model relies more on product attributes, evaluation sentiment and promotion context, so the ability of positive category retrieval is lower. The PR-AUC of MSF-PIN in the promotional scenario is 0.541, which is higher than that of 0.507 in the non-promotional scenario, suggesting that the discount rate, coupon exposure, and special date change the probability of purchase in the same behavioral pattern. The AUC in the high-price goods scenario is 0.889, which is lower than the 0.906 in the low-priced goods scenario. High-unit-price goods usually have a longer comparison cycle, and single-session labeling is easy to underestimate the subsequent purchase intention. The PR-AUC in the long-session scenario is 0.552, which is higher than that in the short-session scenario (0.468), indicating that the more complete the behavioral evidence is, the more obvious the value of sequence coding is. To determine the source of model gain, the core modules were removed item by item, as shown in Figure 6.

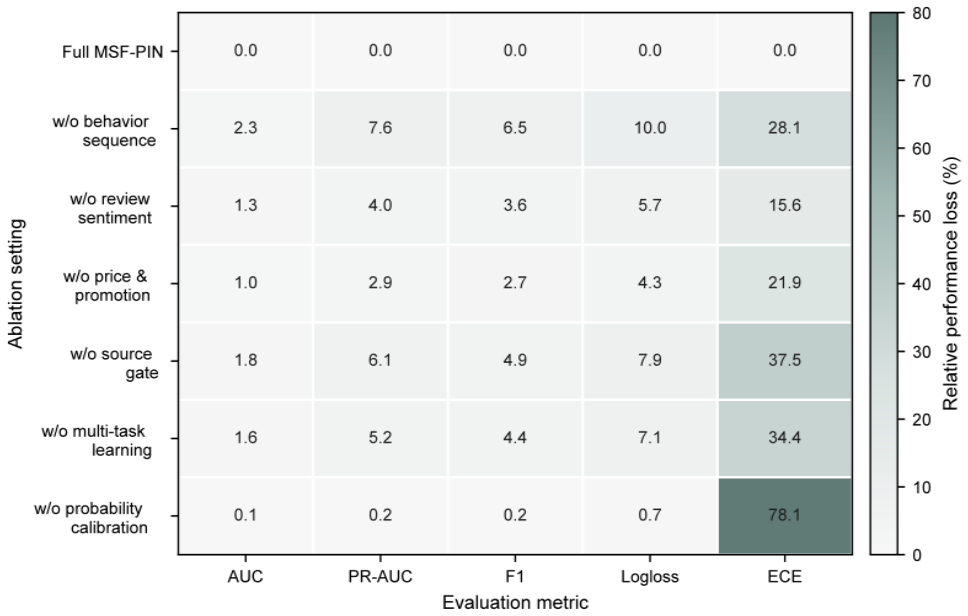

Figure 6: Ablation heat map with metric-wise normalization.

In Figure 6, removing the behavior sequence encoder decreases the AUC from 0.902 to 0.881 and the PR-AUC from 0.523 to 0.483, which are the largest decreases among all structural ablation settings. This result illustrates that the order in which views, stays, adds, and returns to detail pages occur cannot be completely replaced by static statistics. After removing source gate, AUC drops to 0.886, PR-AUC drops to 0.491, and ECE rises to 0.044. Fixed splicing allows weakly correlated sources to enter all samples, and the gating layer serves to reduce this noise. After removing the review sentiment features, PR-AUC decreases to 0.502; after removing the price and promotion context, PR-AUC decreases to 0.508. The review sentiment mainly affects the perception of product quality, while the price and promotion context mainly affects the conversion trigger. After removing multi-task learning, the PR-AUC decreases to 0.496, indicating that the click and add task provides effective prior supervision. After removing probability calibration, AUC stays at 0.901 and PR-AUC is 0.522, but ECE rises from 0.032 to 0.057, suggesting that the calibration module mainly affects the probability scale rather than the sorting ability. The feature contributions and source gating weights together reveal the sample-level decision basis of the model, as shown in Figure 7.

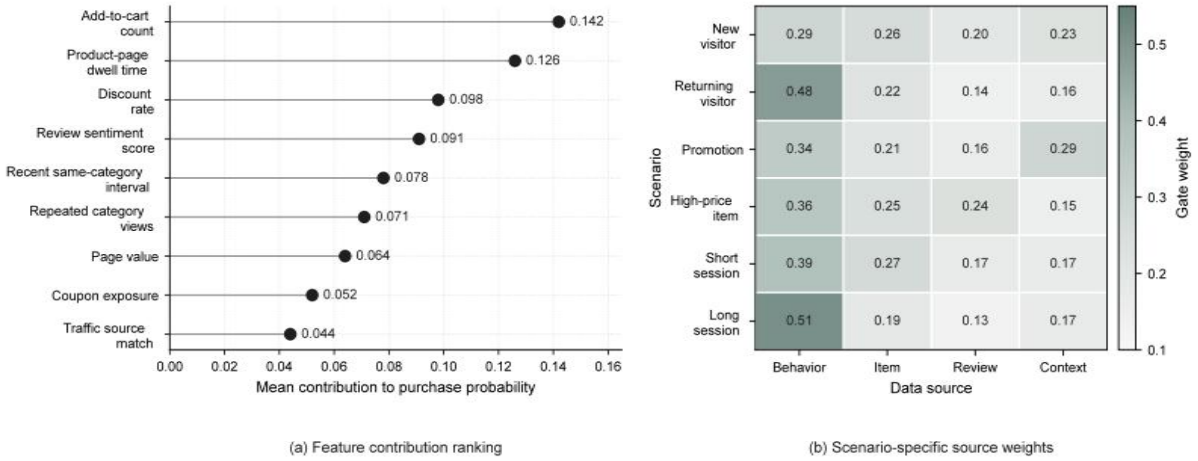

Figure 7: Feature contribution and cross-source gate-weight coupling.

In Figure 7, the contribution of the number of add-ons is 0.142, the dwell time on the product page is 0.126, the discount rate is 0.098, the evaluation sentiment score is 0.091, the interval between the most recent visits to the same category is 0.078, and the number of repeated views of the same category is 0.071. The first two items indicate that behavioral intensity is still the main evidence for purchase prediction; the discount rate and evaluation sentiment suggest that the same conversation path, price stimulus and product word-of-mouth change the final probability. The gating weights are consistent with the feature contributions: in the returning visitor scenario, the behavioral source weight is 0.48; in the new visitor scenario, the product source, evaluation source, and context source weights increase to 0.26, 0.20, and 0.23, respectively; in the promotion scenario, the context source weight increases to 0.29; in the high-price item scenario, the evaluation source weight increases to 0.29; and in the new visitor scenario, the evaluation source weight increases to 0.26, 0.20, and 0.23, respectively. item scenario, the evaluation source weight rises to 0.24. Instead of compressing the multi-source data into a fixed set of features, the model adapts the sources of evidence according to the session state.

Together, the sub-scenario, ablation, and interpretation results illustrate that MSF-PIN's enhancement comes primarily from session path retention, sample-level source weight assignment, and pre-task supervision. The model yields higher gains for old visitors, promotional sessions, and long sessions, and still has more room for error for new visitors, short sessions, and high-priced items. These results also inform the deployment thresholds and error auditing in the next section.

3.3 Deployment-oriented analysis, error sources and application boundary

Once the offline metrics are improved, it is also necessary to determine whether the model can enter the operational chain. Platforms are usually constrained by two types of constraints when reaching users: one is the percentage of people reached and the other is the online delay budget. Too low a reach ratio will miss some real buyers, and too high a reach ratio will increase the cost of mis-touch and disturbance. In this paper, we put reach ratio, delay budget and real purchase coverage into the same result interface and audit a typical error sample. The relationship between the real purchase coverage rate under different reach ratios and delay budgets is shown in Figure 8.

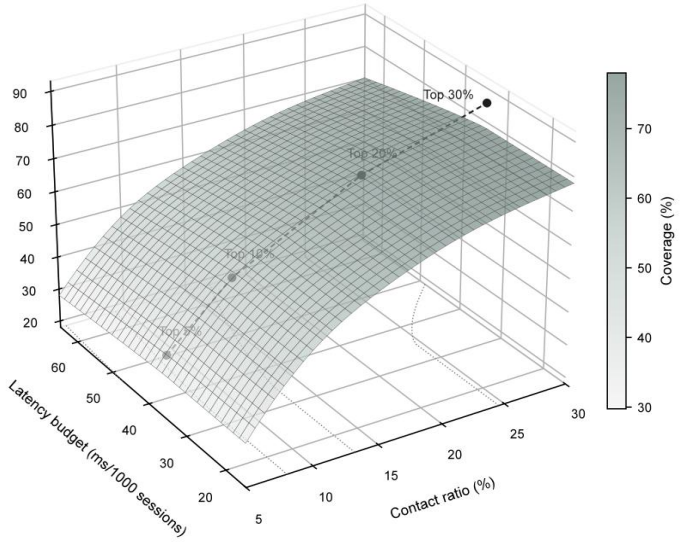

Figure 8: Three-dimensional deployment surface linking threshold, latency and coverage.

In Figure 8, when the reach ratio increases from Top 5% to Top 10%, the true purchase coverage rate increases from 31.6% to 48.9%; when the reach ratio increases to Top 20%, the coverage rate reaches 68.4%; and when the reach ratio continues to increase to Top 30%, the coverage rate reaches 79.7%, but the false positive rate rises to 22.9%. This change shows that high willingness prediction is more suitable for tiered use: Top 5% sessions can trigger inventory reminders, exclusive offers or customer service assistance; Top 10% to Top 20% sessions are suitable for recommendation position adjustment, collocation recommendations and light offers; sessions outside of the Top 20% should reduce the frequency of strong interventions, and be used more for content sorting and low-intrusive displays. The latency dimension in Fig. 7 shows that MSF-PIN around 34.9 ms/1000 sessions can cover the main deployment requirements. The latency increases significantly if the evaluation text is fed into the deep text encoder in real time, so the evaluation sentiment and topic vectors should be updated offline. Prices, inventory, coupon exposure, and special dates change more quickly and are suitable for maintenance via streaming feature tables. This deployment method preserves multi-source information while avoiding online links being slowed down by text encoding. To determine the source of errors, probability trajectories and error categories are audited for typical samples, as shown in Figure 9.

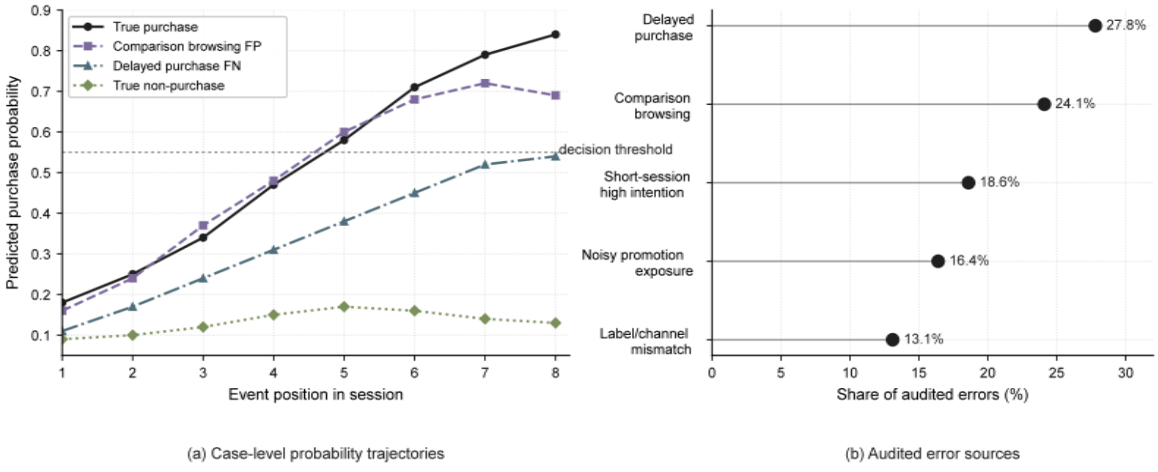

Figure 9: Case-level probability trajectories and audited error sources.

In Figure 9, the predicted probability of the true purchase sample gradually increases with event position and exceeds 0.70 after the 6th event, indicating that the model is able to identify steadily increasing evidence of purchase.The Comparison browsing FP sample rises to 0.72 in the mid-to-late session, but does not make a purchase in the end, due to the fact that the user repeatedly views the same product and stays for a long time, and the behavioral pattern is close to that of the true purchase sample. The Delayed purchase FN sample has a probability of 0.52 at the end of the current session, close to the trigger threshold, but the actual transaction occurs in a subsequent session and the current label fails to reflect the delayed conversion. The audit of error sources includes 27.8% delayed purchase, 24.1% comparative browsing, 18.6% short-session high intent, 16.4% noisy promotion exposure, and 13.1% label or channel mismatch. These errors suggest that session-level labeling still underestimates cross-session purchase intent.

Model outputs should be used as operational priority signals in business. If the high probability mainly comes from add-ons, repeat browsing and detail page stays, the page can prioritize the display of inventory, delivery and payment convenience information; if the high probability mainly comes from discount rates and special dates, coupons or limited-time reminders are more appropriate; if the high probability mainly comes from evaluation sentiment and rating stability, the page should strengthen after-sale protection and user reputation. Cold-start users, high-unit-price goods and cross-channel payment are still the main boundaries of the model. Subsequent experiments need to introduce cross-session attribution, long-term user state and uplift modeling to distinguish real purchase intention from pure browsing interest. Large-scale multimodal product characterization can further complement image, text and associated product information, and has subsequent extension value for complex e-commerce recommendation tasks [31].

4 Conclusion

This paper completes an integrated research from multi-source data organization, model construction to deployment and evaluation around e-commerce in-session purchase intention prediction. Based on 162,840 session samples and 24,912 purchase positive classes, this paper unifies behavior logs, product attributes, evaluation sentiment, price promotions, and access contexts into session-level objects, and constructs the MSF-PIN model accordingly.

(1) At the object organization level, this paper completes the unified mapping of multi-source fields to session samples, preserving the business boundaries between event sequences, product conditions, evaluative feedbacks and marketing stimuli. This treatment makes purchase intention prediction no longer limited to single clickstream statistics, and also provides a stable data caliber for subsequent sub-scenario analysis.

(2) At the method and result level, MSF-PIN improves the purchase session recognition capability through behavioral sequence coding, cross-source gating and multi-task supervision. In the experiment, the model AUC is 0.902, PR-AUC is 0.523, and ECE is 0.032; Top 10% high intention sessions cover 48.9% of the real purchase samples. The results suggest that sample-level evidence allocation and antecedent behavioral supervision have a direct effect on purchase prediction under positive class sparsity.

(3) On the application boundary, the current scheme is still affected by cross-session delayed purchases, insufficient information about cold-start visitors, and long decision cycles for high-priced items. Follow-up work can combine cross-session attribution, long-term user state modeling and uplift evaluation to further improve the stability and strategy transformation efficiency of the model in complex e-commerce scenarios.

About the Author

Dr. Mei Qiongao conducted research on consumer experience mechanisms in offline service scenarios during his doctoral studies. His research was published in Industrial Revitalization Research (2024, Vol. 9, No. 3, pp. 27-39). Taking Sanya Atlantis as an empirical case, he constructed a chain model of "service scene – brand image/perceived value – satisfaction", and revealed the driving mechanism of the physical environment and emotional interaction of high-end resorts on consumers' behavioral intentions. Integrating service-dominant logic and consumer psychological assessment theory, he proposed a three-stage assessment framework for tourism experience value in the digital era.

Currently, his research extends to the online consumption field, focusing on data-driven consumer behavior modeling. He integrates multi-source heterogeneous data to construct a prediction model of consumers' purchase intention in e-commerce scenarios, explores the internal correlation between online and offline consumption behaviors, and provides theoretical and methodological support for precision marketing and the optimization of intelligent recommendation systems. mei714846508@126.com

Dr. Huiyuan Zhang holds a Ph.D. in Hotel and Tourism Management from Honam University, South Korea. Her research program focuses on three interconnected domains: (1) risk perception dynamics in senior tourism, (2) social network service (SNS) applications in tourism marketing, and (3) psychological mechanisms underlying travel consumption decisions.

Her seminal publication in Tourism Research (2021, 46(4), 297-316) pioneered the conceptualization of SNS engagement patterns as predictors of corporate trust formation, particularly examining Chinese social media users' behavioral trajectories. Subsequent work published in 2022 elucidated the mediating role of destination attributes in Chinese seniors' travel planning processes under risk scenarios.

Dr. Zhang's theoretical frameworks have been recognized for advancing the interdisciplinary understanding of digital-era tourism behavior, particularly through the integration of consumer psychology principles with risk management models. hyewon86@163.com

Chen Yu, Lecturer at the School of Economics and Management, with research interests in market research and consumer behavior. He serves as an editorial board member for two academic journals and actively participates in peer review and academic community development. A research proposal he authored won the Second Prize from the Anhui Provincial Committee of the Chinese People's Political Consultative Conference (CPPCC). His research focuses on empirical investigation and consumer decision-making mechanisms. He leads two Anhui Provincial Quality Engineering Projects and has published two academic papers. He is dedicated to integrating survey methods with behavioral analysis and continuously explores practical issues in marketing through data-driven research approaches. chenyu@hblgxy.edu.com

References

[1] Sakar, C. O., Polat, S. O., Katircioglu, M., et al. (2019). Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31, 6893-6908.

[2] RetailRocket. (2017). RetailRocket recommender system dataset. Kaggle Dataset.

[3] Tagliabue, J., Greco, C., Roy, J.-F., et al. (2021). SIGIR 2021 E-Commerce Workshop

Data Challenge. arXiv, arXiv:2104.09423.

[4] Ma, X., Li, Y., & Asif, M. (2024). E-commerce review sentiment analysis and purchase intention prediction based on deep learning technology. Journal of Organizational and End User Computing, 36(1), Article 335122.

[5] Zhang, X., & Guo, C. (2024). Research on multimodal prediction of e-commerce customer satisfaction driven by big data. Applied Sciences, 14(18), 8181.

[6] Ding, W., Wang, W., Kwok, S. H. D., et al. (2024). IntentionQA: A benchmark for evaluating purchase intention comprehension abilities of language models in e-commerce. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 2247-2266).

[7] Hidasi, B., Karatzoglou, A., Baltrunas, L., et al. (2016). Session-based recommendations with recurrent neural networks. In International Conference on Learning Representations Workshop. arXiv, arXiv:1511.06939.

[8] Wang, S., Cao, L., Wang, Y., et al. (2021). A survey on session-based recommender systems. ACM Computing Surveys, 54(7), Article 154.

[9] Zhou, G., Song, C., Zhu, X., et al. (2018). Deep interest network for click-through rate prediction. In Proceedings of KDD (pp. 1059-1068).

[10] Zhou, G., Mou, N., Fan, Y., et al. (2019). Deep interest evolution network for click-through rate prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 5941-5948.

[11] Chen, Q., Zhao, H., Li, W., et al. (2019). Behavior sequence transformer for e-commerce recommendation in Alibaba. arXiv, arXiv:1905.06874.

[12] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. In IEEE International Conference on Data Mining (pp. 197-206).

[13] Sun, F., Liu, J., Wu, J., et al. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of CIKM (pp. 1441-1450).

[14] Cheng, H.-T., Koc, L., Harmsen, J., et al. (2016). Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (pp. 7-10).

[15] Guo, H., Tang, R., Ye, Y., et al. (2017). DeepFM: A factorization-machine based neural network for CTR prediction. In Proceedings of IJCAI (pp. 1725-1731).

[16] Lian, J., Zhou, X., Zhang, F., et al. (2018). xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In Proceedings of KDD (pp. 1754-1763).

[17] Song, W., Shi, C., Xiao, Z., et al. (2019). AutoInt: Automatic feature interaction learning via self-attentive neural networks. In Proceedings of CIKM (pp. 1161-1170).

[18] Wang, R., Shivanna, R., Cheng, D., et al. (2021). DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In Proceedings of WWW (pp. 1785-1795).

[19] Huang, T., Zhang, Z., & Zhang, J. (2019). FiBiNET: Combining feature importance and bilinear feature interaction for click-through rate prediction. In Proceedings of RecSys (pp. 169-177).

[20] Naumov, M., Mudigere, D., Shi, H.-J. M., et al. (2019). Deep learning recommendation model for personalization and recommendation systems. arXiv, arXiv:1906.00091.

[21] He, X., Liao, L., Zhang, H., et al. (2017). Neural collaborative filtering. In Proceedings of WWW (pp. 173-182).

[22] Ma, X., Zhao, L., Huang, G., et al. (2018). Entire space multi-task model: An effective approach for estimating post-click conversion rate. In Proceedings of SIGIR (pp. 1137-1140).

[23] Ma, J., Zhao, Z., Yi, X., et al. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of KDD (pp. 1930-1939).

[24] Tang, H., Liu, J., Zhao, M., et al. (2020). Progressive layered extraction (PLE): A novel multi-task learning model for personalized recommendations. In Proceedings of RecSys (pp. 269-278).

[25] Zhou, J., Yu, Q., Luo, C., et al. (2023). Feature decomposition for reducing negative transfer: A novel multi-task learning method for recommender system. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 16390-16391.

[26] Yan, J., Chen, J., Wu, Y., et al. (2023). T2G-Former: Organizing tabular features into relation graphs promotes heterogeneous feature interaction. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 10720-10728.

[27] Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3146-3154).

[28] Prokhorenkova, L., Gusev, G., Vorobev, A., et al. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (Vol. 31, pp. 6638-6648).

[29] Guo, C., Pleiss, G., Sun, Y., et al. (2017). On calibration of modern neural networks. In Proceedings of ICML (Vol. 70, pp. 1321-1330).

[30] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774).

[31] Tiady, S., Jain, A., Sanny, D. R., et al. (2024). MERLIN: Multimodal & multilingual embedding for recommendations at large-scale via item associations. In Proceedings of CIKM (pp. 2315-2324).