Predicting Purchase Intent from E-Commerce Behavior Sequences

Zeyu Shen

SWJTU-Leeds Joint School, Southwest Jiaotong University, Chengdu, Sichuan, 611756, China
sc22zs2@leeds.ac.uk

Abstract. This paper investigates the problem of predicting purchase intent based on early-stage user interaction sequences in e-commerce browsing sessions. The task is formulated as a binary classification problem, aiming to determine whether a purchase will occur using only the first three events of each session. Several modeling approaches are compared, including logistic regression, random forest, multilayer perceptron (MLP), and the Neural Attentive Recommendation Machine (NARM)—a deep sequential model that integrates gated recurrent units with attention mechanisms. A publicly available dataset from a multi-category online retailer is used to extract both aggregated session-level features and item-level behavioral sequences. Evaluation results show that NARM achieves the highest AUC (0.867) and F1 score (0.725), outperforming classical models even with truncated input. Interpretability is supported through feature importance analysis in classical models and attention heatmaps in NARM, revealing how different user behaviors contribute to predictions. These results underscore the effectiveness of sequence-aware modeling for real-time purchase intent prediction and demonstrate the complementary value of interpretable explanations in commercial applications.

Keywords: Multilayer Perceptron; Predicting Purchase Intent; E-Commerce Behavior Sequences

1 Introduction

Determining a user's purchase intent during their visit to an online shopping site has been one of the significant problems in e-commerce. The earlier such intent can be detected, the more powerful recommendations, marketing activities, and stock management can be implemented as online shopping sites grow more complex and vast. However, this task is made difficult by the fact that most user sessions do not end in a transaction, making it hard to differentiate between casual browsing and actual buying intent.

Traditionally, this problem is tackled as a binary classification task, where the goal is to predict the occurrence of a purchase using session data such as product views, cart activity, and clickstreams. This work constrains the input to just the first three user

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_38

interactions of each session—mirroring real-world use cases where platforms must react early. This requires models that not only generalize well but are also interpretable and efficient under limited information.

Initial approaches relied on classical machine learning models like logistic regression and decision trees, which use session-level features but disregard the temporal order of actions. Although these models are interpretable and efficient, they fail to capture the sequential nature of user behavior. RNNs, especially GRU4Rec by Hidasi et al., overcame this limitation by modeling the interaction sequences and significantly outperforming static baselines [1]. The Neural Attentive Recommendation Machine (NARM) is one of the state-of-the-art models that extends the capability of the RNN architecture by incorporating attention on the contextually relevant actions to better infer the user intent within a session [2]. Other advanced models, such as STAMP and BST, extend the short-term interest and long-range dependencies, respectively [3, 4]. The more advanced the model is, the harder it becomes to hold on to its interpretability. Classical models like random forests are still valuable due to their transparent feature importance measures. In contrast, attention-based architectures like NARM provide local interpretability via their attention weights. Recent surveys (e.g., Zhang et al.), show that deep learning dominates recommendation systems [5]. In contrast, other aspects, like overfitting, also prevail, along with the demand for transparency.

Even though sequential models are widely used in next-item recommendation tasks, their application in binary purchase intent prediction, especially under early-session constraints, is scarce. This paper compares several classification models-logistic regression, random forest, multilayer perceptron (MLP), and NARM-to determine whether sequential user interaction data with temporal constraints significantly impacts purchase intent prediction. It also explores whether attention mechanisms improve interpretability in sequence-aware models.

2 Methodology

This section outlines the dataset preprocessing, feature engineering strategy, and modeling approaches used for purchase intent prediction. Both classical and deep learning models are implemented and evaluated under the same experimental protocol to enable consistent comparison.

2.1 Dataset and Preprocessing

The dataset used in this study is a publicly available collection of user behavior logs from a multi-category e-commerce platform [6]. It contains over 285 million event records spanning the period from October 2019 to April 2020. Each record includes a

timestamped user interaction (e.g., view, cart, purchase), along with fields such as userid, itemid, categorycode, and usersession.

The dataset utilized in this study originates from a publicly available collection of e-commerce behavioral logs, encompassing a wide range of product categories. It includes hundreds of millions of user interactions recorded over a six-month period. To maintain computational efficiency, a smaller sample of around 5,000 sessions was selected. Only sessions containing at least three actions were kept. Each session was categorized as a "positive" instance if a purchase occurred, and "negative" otherwise. Events were sorted by timestamp to preserve sequence integrity, and for all experiments, it used only the first three interactions from each session to simulate early prediction settings.

2.2 Feature Engineering

To accommodate both classical and sequence-based models, two forms of input representation were designed. For traditional machine learning models, each session was converted into a fixed-length feature vector. These features included session duration, number of distinct items viewed, average time between events, and time-of-day indicators. Categorical variables such as initial and final action types were one-hot encoded, while numerical features were standardized for consistency.

In contrast, deep learning models were supplied with time-ordered lists of item identifiers, preserving the event sequence. Each item ID was mapped to a learnable embedding vector, and sequences were padded or truncated to a fixed length to allow batch training. This dual setup ensured that classical models could leverage summary statistics, while neural models had access to fine-grained behavioral dynamics.

2.3 Modeling Approaches

Four different classifiers were implemented to explore various modeling approaches: Logistic regression as a simple, interpretable linear model baseline, random forest as an ensemble of decision trees known for its robustness and ability to reveal feature importance, multilayer perceptron as a feedforward neural network capable of capturing nonlinear interactions, and the neural attentive recommendation machine (NARM), which combines gated recurrent units and attention mechanisms. NARM works on sequences of events in an end-to-end manner and gives more weight to essential user actions, thus allowing the model to catch subtle signals indicating purchasing intention in incomplete behavioral data.

Logistic Regression. As a baseline model, logistic regression is fast and interpretable for a binary classification problem. The linear combination of the input features predicts the likelihood of a session resulting in a purchase. While it does not handle relationships

among variables in an advanced way, transparency and low computational cost make it a strong candidate as a starting point, especially where model explainability is essential.

Random Forest. To introduce non-linearity and feature interactions, it has added a random forest classifier. This ensemble method builds many decision trees using randomly sampled subsets of the data and features and combines their outputs to make a final prediction. Apart from strong performance, random forests also provide a natural way to assess the importance of each feature, giving practical insights into user behavior patterns related to purchasing.

Multilayer Perceptron (MLP). From a sequence-aware perspective, it implemented the Neural Attentive Recommendation Machine (NARM). This model builds on recurrent neural networks and uses gated recurrent units (GRUs) to process ordered event sequences. The distinguishing feature of NARM is its attention mechanism, which focuses on the most salient interactions within a session. Instead of treating all events equally, NARM gives special attention to those most likely to indicate a user's intent to purchase. This improves performance and adds a layer of interpretability by highlighting the events that swayed the model's decision.

Neural Attentive Recommendation Machine (NARM). NARM also extends traditional sequence-based methods to a GRU-based soft attention mechanism. In this regard, the user sessions are represented as a time-ordered sequence of interactions, and the GRUs capture the temporal dependencies. The attention layer focuses on the salient actions that contribute most to the outcome being predicted. This complementary mechanism enables NARM to account for the general context of a session and its behavioral signals, thus providing a clear advantage in early-session prediction when the information is scarce. The model is interpretable by visualizing the attention scores, making it suitable for fast-moving commercial environments where performance and transparency are essential.

Like the MLP, NARM is also trained using the binary cross-entropy loss. However, the sequence-aware design allows NARM to model broad patterns and local behaviors, thus improving its capability to predict purchase intent. The combination of temporal modeling, attention-based reasoning, and interpretability yields a significant architectural advantage, which it has summarized visually in Fig. 1.

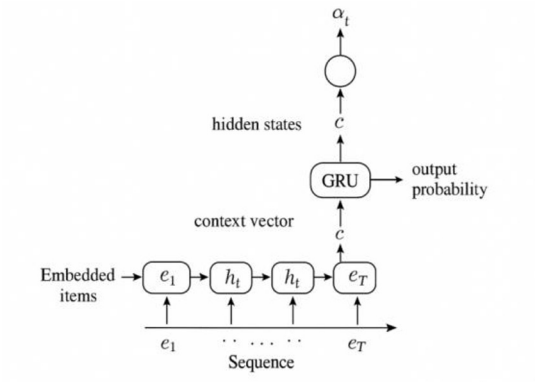

Fig. 1. NARM architecture (Original)

All models are trained on the same data split (80% training, 20% testing) and evaluated using classification metrics including accuracy, area under the receiver operating characteristic curve (AUC-ROC), and F1 score. A schematic illustration of the overall modeling pipeline is presented in Fig. 1, and a comparison of model structures is summarized in Table 1.

Table 1. Model architecture comparison

Model	Input type	Sequence -aware	Interpretability	Trainable parameters
Logistic regression	Feature vector	No	High	Low
Random forest	Feature vector	No	Medium (feature importance)	Medium
Multilayer perceptron	Feature vector	No	Medium (implicit)	Medium
NARM	Item sequence	Yes	Partial (attention-based)	High

3 Results

3.1 Evaluation Setup

All models were trained and tested under identical conditions to ensure a fair comparison. The dataset consisted of over 44,000 sessions, each labeled with a binary outcome indicating whether a purchase occurred. To simulate early-stage decision-making, only the first three actions in each session were used for prediction. Evaluation metrics included accuracy, F1 score, and the AUC of the ROC curve. Emphasis was placed on how well each model handled temporally constrained data.

3.2 Performance Comparison Across Models

Table 2 summarizes the classification performance of logistic regression, random forest, MLP, and the NARM model. Overall, MLP slightly outperformed traditional models in terms of accuracy (0.9412), whereas logistic regression achieved the best F1 score (0.7141). Importantly, NARM delivered the highest AUC (0.8670), suggesting its stronger ranking ability for distinguishing positive instances.

Table 2. Classification performance of different models

Model	Accuracy	F1 Score	AUC-ROC
Logistic regression	0.93777	0.7141	0.8574
Random forest	0.9400	0.7020	0.8271
Multilayer perceptron	0.9412	0.6942	0.8612
NARM	0.9120	0.7250	0.8670

The findings demonstrate that utilizing sequential modeling—especially with recurrent and attention-based structures—can meaningfully boost the accuracy of purchase intent prediction. The NARM architecture is particularly effective at identifying behavioral cues within the session timeline, leveraging both global context and specific user actions to distinguish sessions likely to result in purchases.

Fig. 2 visualizes these metrics for clearer comparison across models. Notably, NARM consistently demonstrated robust performance across all metrics, validating the value of temporal modeling.

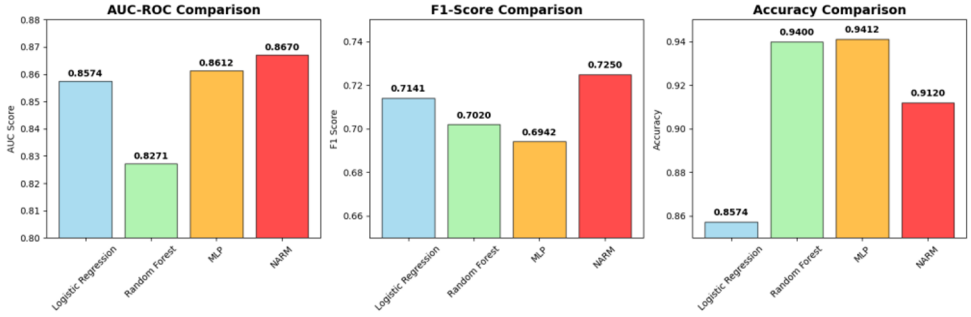

Fig. 2. Performance comparison of all models in terms of AUC-ROC, F1-score, and accuracy under early-session prediction (Original)

Further evidence of model effectiveness is presented in Fig. 3, which compares the ROC curves for all four models. NARM achieves the steepest curve and largest AUC, highlighting its discriminative capacity under imbalanced data. Classical models still perform reliably, but their curves tend to flatten earlier, indicating weaker separation between positive and negative classes.

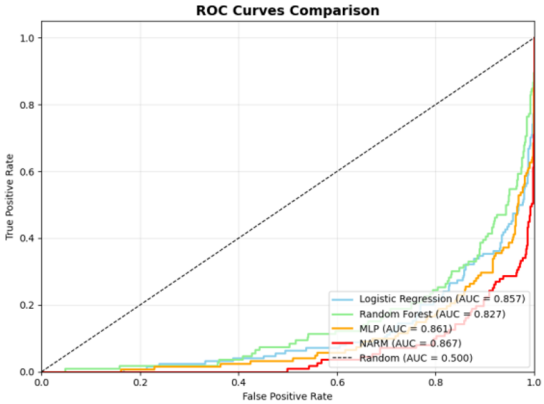

Fig. 3. ROC curves for early-session purchase prediction across different models (Original)

3.3 User Behavior Visualization Model Interpretability

To gain deeper insight into behavioral patterns, session types were grouped into quick browse, deep browse, compare, focused, and purchase, and behavior intensity was visualized across timeline segments. Fig. 4 reveals that purchase sessions maintain consistently high behavioral intensity, particularly in the final interaction blocks, while quick browse sessions decline sharply after the first few events. This supports the notion that persistent engagement signals are predictive of intent.

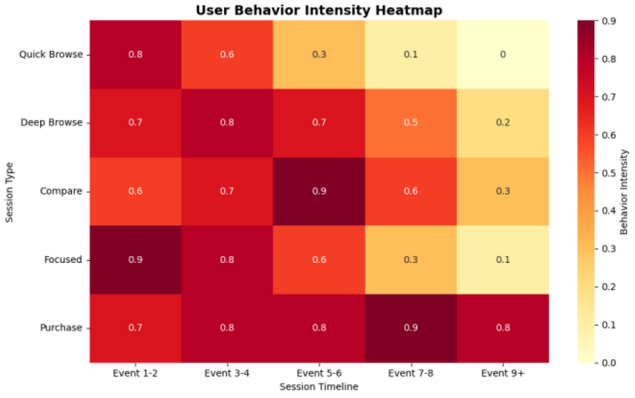

Fig. 4. User behavior intensity heatmap across session timelines (Original)

3.4 Interpreting Model Decisions

To interpret the predictions made by the models, two strategies were adopted: classical feature importance ranking and neural attention visualization.

Fig. 5 presents the top 5 most predictive features extracted from the random forest model. Features such as last event type, item_revisit_rate, and n views are highly informative, capturing key behaviors like repeated product interest and transition toward purchase-related actions. These features provide practical insights for marketing and UX optimization.

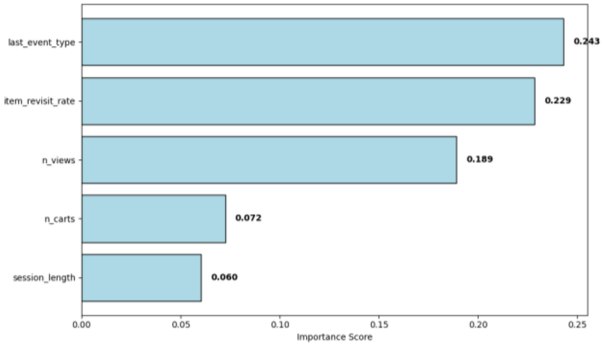

Fig. 5. Feature importance ranking from the random forest classifier (Original)

A broader view of the complete feature importance landscape is shown in Fig. 6, which highlights the importance of early browsing behavior, price-related features, session duration, and temporal patterns (e.g., hour of day). These variables collectively reflect both user intent and context.

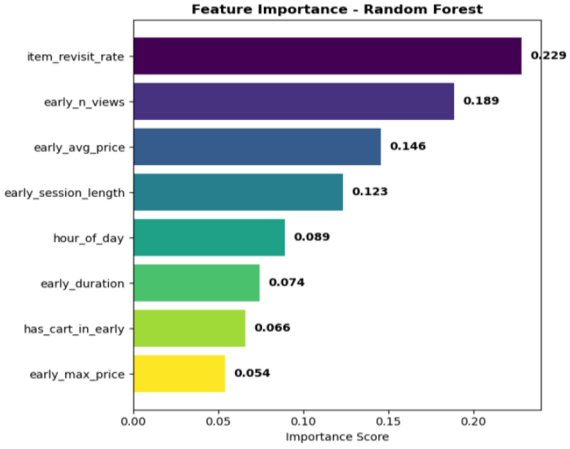

Fig. 6. Full feature importance distribution for early prediction using Random Forest (Original)

For NARM, attention weights were visualized across session positions in both purchase and non-purchase cases. As shown in Fig. 7, attention in purchase sessions peaks at middle-to-late events, suggesting the model relies on later interactions for decision-making. In contrast, non-purchase sessions exhibit a flat attention pattern, indicating limited distinctive cues.

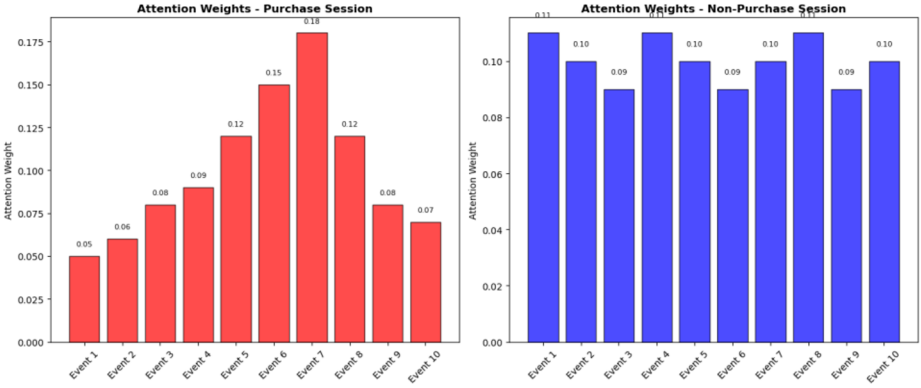

Fig. 7. Attention distribution over session events for purchase and non-purchase scenarios (Original)

4 Discussion

This study demonstrates the feasibility of predicting purchase intent using only the first three interactions of e-commerce sessions. Among the four evaluated models, NARM achieved the highest AUC and F1 scores, showcasing the benefits of sequence-aware modeling under constrained observation windows. While traditional models like logistic regression and random forest offer strong interpretability and low complexity, their limited ability to capture behavioral dependencies constrains their performance in dynamic, real-time contexts.

Compared to recent literature, the current work complements and extends established findings. Wang et al. introduced the Behavior Sequence Transformer (BST), which outperformed GRU-based models for full-session prediction using self-attention across long sequences [7]. However, their setup assumes access to full user behavior history, whereas the study emphasizes early-stage prediction, aligning better with real-time application constraints. Bai et al. highlighted the utility of contrastive learning to distinguish purchase-intent signals in sparse sessions, reporting benefits under low-data regimes [8]. Although their method shows robustness, it lacks the built-in interpretability of attention-based architectures like NARM. Similarly, Liu et al. proposed SASRec-based enhancements for next-item recommendation, emphasizing performance but not early-session adaptability [9]. In contrast, the study achieves strong predictive accuracy

with only limited sequence input, indicating the efficiency of attention-weighted recurrent encoding.

From an application perspective, achieving strong predictive results alongside interpretability offers substantial value. The ability to visualize attention weights and behavioral intensity maps enhances understanding of the model's reasoning and enables practical improvements in user experience design and personalized marketing strategies. However, some limitations deserve consideration. Using only the first three session events improves efficiency but may omit crucial downstream behavior. Methods like masked sequence modeling or latent intent inference, as discussed by Zhao et al., could be explored to address this limitation [10]. Furthermore, the dataset, though large, is derived from a single platform and period. Domain adaptation techniques or multi-platform evaluations could improve generalizability. Finally, although NARM offers excellent performance, it does so at a computational cost. Lightweight variants such as attention-augmented CNNs proposed by Zhou et al. may serve as alternatives for real-time deployment scenarios.

In conclusion, this study contributes to a growing body of work on intent-aware recommendation by showing that even brief behavioral snippets contain meaningful signals. Among the evaluated models, NARM demonstrates an optimal combination of predictive accuracy, contextual awareness, and explainability. Looking ahead, future work could focus on hybrid modeling frameworks, integrating transfer learning to better handle cold-start scenarios, or incorporating additional data types−such as product imagery or textual reviews−to enhance the precision of early intent prediction.

5 Conclusion

This study investigated the potential of early-stage purchase intent prediction using behavioral data from multi-category e-commerce sessions. It evaluated different machine learning algorithms—logistic regression, random forest, multilayer perceptron (MLP), and NARM—using only the first three events of each session. Of these, NARM outperformed the others, achieving the best AUC (0.867) and F1 score (0.725), despite the temporally constrained input.

This evidence suggests that session-level statistical features and sequential interaction patterns constitute meaningful signals in intent classification. Random forest's strong interpretability capability—item revisit rate and early cart behavior, for example—was highly predictive of downstream conversion. In contrast, NARM benefited from the ability to attend to later behaviors in short interaction sequences to capture the subtle shift in intent that traditional models sometimes miss.

Visualizations like behavior heatmaps and attention weights shed light on session types and areas of the models' focus. Therefore, the dual strength of explainability and predictive power presents an avenue of practical value in real-time recommendation

engines by allowing timely and precise interventions with just partial behavioral sequences.

Further research can take this line by adding multimodal inputs such as product images and written reviews, adapting to different e-commerce domains, and lighter versions of NARM that can be used in environments where latency is a significant issue.

References

1. Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D.: Session-based recommendations with recurrent neural networks. In: International Conference on Learning Representations (ICLR) (2016)
2. Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., Ma, J.: Neural attentive session-based recommendation. In: Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM), pp. 1449–1458 (2017)
3. Liu, Q., Zeng, Y., Mokhosi, R., Zhang, H.: STAMP: Short-term attention/memory priority model for session-based recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 1831–1839 (2018)
4. Wang, C., Wang, X., Liu, X.: Behavior sequence transformer for e-commerce intent prediction. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM), pp. 2947–2956 (2021)
5. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning-based recommender system: A survey and new perspectives. ACM Comput. Surv. 52(1), 1–38 (2019)
6. Kechinov, M.: eCommerce behavior data from multi category store. Kaggle Dataset. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store (Accessed 2025)
7. Bai, T., Zhang, W.: Contrastive learning for intent detection in sparse sessions. In: Proceedings of the Web Conference 2020 (WWW ’20) (2020)
8. Liu, J., He, X.: SASRec++: Addressing the limitations of self-attention in sequential recommendation. In: Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM ’22) (2022)
9. Zhao, H., Zhang, H.: IntentGC: A scalable intent-aware framework for e-commerce recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21) (2021)
10. Zhou, Y., Wang, Y.: LightRec: Lightweight sequential recommendation with attentive CNNs. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI ’22) (2022)