

PREDICTING FUTURE E-COMMERCE PURCHASES FROM MULTI-VISIT SEQUENCES AND BEHAVIOURAL MICRO-INTERACTIONS

Marcin Gabryel^{1,2,*}, Milan Kocić², Aleksandra Gabryel²,
Marek Kisiel-Dorohonicki³, Zofia Patora-Wysocka⁴

¹*Department of Artificial Intelligence, Częstochowa University of Technology,
42-200 Częstochowa, Poland*

²*Spark Digitup, Plac Wolnica 13 lok. 10, 31-060 Kraków, Poland*

³*AGH University of Krakow, Poland*

⁴*Management Department, SAN University,
90-113 Łódź, Poland*

**E-mail: marcin.gabryel@pcz.pl*

Submitted: 12th January 2026; Accepted: 22nd May 2026

Abstract

This paper investigates future purchase prediction in e-commerce using sequences of user visits and behavioural micro-interactions. The study is based on anonymised production data from three e-commerce domains, comprising approximately 690,000 users and 2.61 million visits. User histories are modelled as temporally ordered sequences of visits, with each visit represented by 21 features. Six models are evaluated: three tabular approaches (SVM, XGBoost, CatBoost) and three sequential architectures (RNN, LSTM, GRU), each tested in two variants addressing class imbalance. The experimental setup employs a temporal train-test split, five random seeds, and bootstrap-based significance testing. The top-performing sequential models achieve an AUC-PR of approximately 0.379, significantly outperforming the strongest tabular baseline, XGBoost (AUC-PR $\approx$ 0.364). This advantage remains consistent across domain-specific evaluations as well as in a leave-one-shop-out setting. Further analysis shows that mouse movement and scrolling-related features provide the strongest predictive signal, whereas extending user histories beyond 40 visits yields only marginal performance gains.

Keywords: e-commerce, purchase prediction, clickstream data

1 Introduction

The ability to predict future purchasing behaviour is one of the central challenges in e-commerce analytics. Accurately identifying users

with an elevated likelihood of conversion enables not only more efficient allocation of advertising budgets, but also optimisation of product exposure, recommendations, discount strategies, and in-session interactions. In practice, models capable of

early prediction before a purchase is finalised are particularly valuable, as only then can the forecast be used operationally.

This problem is challenging for several reasons. First, e-commerce data are highly imbalanced: users who complete a purchase represent only a small fraction of those who browse. Second, customer behaviour is inherently sequential, where not only the number of interactions matters, but also their order, temporal spacing, shifts in activity intensity, and transitions between exploration, comparison, and decision-making phases. Third, real-world production datasets contain noise, duplicates, seasonal variation, and differences across commercial domains.

The literature identifies two main approaches to this problem. The first involves aggregating user behaviour into a tabular representation and training conventional machine learning models, particularly tree-based methods and boosting models [4, 15, 13]. The second approach employs sequential models, primarily RNNs, LSTMs, and GRUs, which aim to directly capture the temporal nature of clickstream data [3, 6, 14, 16, 18]. More recent studies combine these paradigms with additional techniques such as meta-classifiers, attention mechanisms, explainability methods, and richer representations of customer context [1, 6, 21].

This paper extends a previous conference version of the study [8]. The introduced changes include a transition from a simplified, static user representation to real-world variable-length visit sequences, a unified definition of the user throughout the data pipeline, a comparison of a broader set of models, the adoption of a temporal evaluation protocol, and an expanded analysis covering per-store results, a leave-one-shop-out experiment, feature ablation, and an investigation of the impact of sequence length.

The study focuses on four main research questions: (1) whether sequential models provide an advantage over strong tabular baselines in predicting future purchases, (2) what impact class imbalance handling strategies have on performance, (3) which groups of behavioural features provide the strongest predictive signal, and (4) to what extent trained models generalise across different e-commerce domains.

The core contributions of this work include: (1) the introduction of a production-realistic task definition with a 42-day observation window and a 14-day prediction window, (2) a comparative evaluation of six models and two class-balancing methods within a unified protocol, (3) demonstrating that a sequential representation provides a statistically significant advantage over the strongest tabular model, (4) a store-level decomposition of results, an analysis of cross-domain transfer, and the identification of the most informative feature groups, and (5) deployment-oriented insights regarding the trade-offs between performance, complexity, and robustness to distributional shifts.

2 Literature Review

Research on purchase prediction in e-commerce has evolved in tandem with the increasing availability of clickstream data and marketing analytics platforms. Early studies focused primarily on models leveraging aggregated features. Martínez et al. [15] proposed a non-contractual purchase prediction framework based on features derived from customer history and tree-based boosting. Chen et al. [4] addressed the class imbalance problem by integrating factorisation machines, random forests, and cost-sensitive learning. Kim et al. [13] demonstrated that combining customer attributes with browsing patterns improves model accuracy, aligning with the standard practice of combining demographic data, traffic source, and on-site activity. Gabryel et al. [10] proposed the use of conditional GAN networks for synthetic data generation to accelerate the development of models predicting users' likelihood of making a purchase.

In parallel, a research strand emerged focusing on the sequential nature of user behaviour. Sakar et al. [18] compared MLPs and LSTMs in the context of session-based data, showing that recurrent models effectively leverage event sequences and activity duration. Koehn et al. [14] demonstrated that deep learning models trained on clickstream data are competitive with conventional methods, particularly when behavioural representations are not oversimplified by manual aggregation. Requena et al. [16] investigated purchase intent prediction under limited browsing information, indicating that even

sparse signals can be informative when their structure is appropriately modelled.

An important line of research focuses on early purchase prediction. Esmeli et al. [7] showed that even initial user–platform interactions provide signals sufficient for the effective recognition of purchase intent. In subsequent work, the same authors extended their model to include contextual and loyalty-related features, emphasising the importance of interpreting current behaviour in the context of prior customer history. This is particularly relevant from a marketing perspective, as a model that only yields accurate predictions shortly before transaction completion has lower operational value than one capable of providing sufficiently early predictions.

The literature also explores more complex modelling frameworks. Ehsani and Hosseini [6] proposed an Oracle meta-classifier approach for clickstream data, demonstrating that selective model combination improves performance in environments characterised by heterogeneous behavioural patterns. Vaghari et al. [21] employed attention mechanisms for the dynamic modelling of user intent and product context. Zhang et al. [22] introduced a hybrid variant combining RNNs with Naive Bayes for large-scale data from the Ali Tianchi platform, arguing that the sequential nature of behaviour can be effectively integrated with simpler decision-making mechanisms.

The rationale for applying deep learning to tabular data remains a critical point of debate. Shwartz-Ziv and Armon [19], as well as Grinsztajn et al. [11], demonstrated that tree-based models frequently remain a robust baseline for typical tabular tasks, particularly in the presence of moderate data volumes and constrained tuning budgets. Bastos and Bernardes [1] went further, showing that gradient boosting models not only effectively predict conversion but also provide interpretable insights into the most important predictors. Consequently, any study proposing sequential models should evaluate them against strong tabular baselines rather than relying solely on simple classifiers.

In a broader research context, the systematic review by Chen et al. [5], encompassing 98 publications from 1995 to 2023, offers a comprehensive synthesis of the field. The authors identify the primary research strands as product recommen-

dations, marketing targeting, browsing behaviour analysis, and predictive methods based on purchase history and clickstream data. The review also emphasises the need for more rigorous model comparisons, greater attention to the quality of experimental protocols, and a stronger alignment between business value and technical model evaluation.

Building on the aforementioned literature, three primary research gaps can be identified that are addressed in this paper. First, there is a scarcity of experiments that concurrently evaluate strong tabular baselines against sequential models trained on real-world user visit sequences. Second, relatively few studies simultaneously analyse the impact of class imbalance handling strategies, provide a breakdown of results across sales domains, and evaluate cross-store transfer. Third, the literature frequently relies on session-level data or aggregated representations, whereas in practical deployment multi-visit user histories and behavioural micro-interactions, such as mouse movements, scrolling, keyboard input, and focus changes play a crucial role.

3 Data and Problem Definition

The study was conducted on anonymised production data sourced from a system monitoring user activity in online stores. The dataset comprises three sales domains: a specialist retailer (Domain A), a bookshop (Domain B), and an online pharmacy (Domain C). For each user, a sequence of visits was recorded, along with a set of behavioural statistics describing interactions with the website. The raw data were sourced from tables containing information on individual visits, as well as from tables with micro-interaction statistics computed for each visit.

A key step in data preprocessing was defining a stable user identifier. Rather than relying on a single session-based field, users were identified using a combination of an IP address and a device fingerprint hash [9]. This approach reduces the risk of counting the same user multiple times across visits, lowers the likelihood of incorrectly linking independent sessions, and makes the identification independent of cookies. The records were then ordered chronologically, deduplicated, and aggregated into user histories.

The task is formulated as a binary prediction of future purchase. For each user, an observation window covering 42 days of visit history was constructed, with the target label assigned a value of 1 only if a purchase occurred within the subsequent 14-day prediction window. This approach more accurately reflects real-world deployment conditions than the classification of individual sessions ending in conversion, as the model predicts a future event based on the user's cumulative interaction history with the website.

Following user-level deduplication, the dataset comprises approximately 690,000 users and 2.61 million visits. As shown in Table 1, the specialist retailer is the largest domain, accounting for the majority of observations. In contrast, the online pharmacy represents the smallest domain and is characterised by the lowest positive class ratio, making it the most challenging in terms of prediction performance.

All data were processed in anonymised form. The study did not use order contents, customer personal data, or any information enabling direct identification. The analysis was restricted exclusively to behavioural and technical signals at both the session and user levels.

3.1 Features and Data Representation

Each visit was described using 21 numerical features grouped into four blocks. The first block comprises general visit-level features: visit depth, device type, traffic source, visit duration, and time elapsed since the previous visit. The second block covers mouse and scrolling micro-interactions, including the number of clicks, cursor path length, the number and uniqueness of movement points, the mouse trajectory bounding-box area, the number of scroll events, and scroll depth. The third block captures keyboard input, including the total number of key presses, such as backspace and alphanumeric characters. The fourth block describes focus switching between the browser window and tab, as well as the average delay between blur and focus events.

Based on the same data, two representations were constructed: tabular and sequential. In the tabular view, aggregated feature statistics were computed for each user and fed into SVM, XGBoost,

and CatBoost models. In the sequential view, each user history was represented as a temporally ordered sequence of visits with a maximum length of 40 steps. Shorter sequences were padded using masking, while longer ones were truncated to the most recent visits. This representation allows sequential models to capture behavioural changes over time, rather than relying on a single aggregated user profile.

3.2 Data Splitting and Evaluation Protocol

The data split was conducted in a strictly temporal manner. Users were sorted by the timestamp of their last observed visit. The most recent 15% of users were assigned to the test set, while the most recent 15% of the remaining 85% were used as the validation set. This approach more faithfully mirrors real-world deployment conditions than a conventional random split, as the model is evaluated on a subsequent cohort of users, accounting for natural seasonal drift and changes in marketing campaigns.

In parallel with the main experiment, three groups of additional analyses were conducted. First, the test results were disaggregated across the three retail domains to verify whether the performance advantage of a particular model family was not merely driven by the dominance of the largest store (see Section 5.3). Second, a leave-one-shop-out experiment was carried out, in which a model trained on data from two stores was evaluated on the third (see Section 5.4). Third, feature ablation studies (Table 6) and a sensitivity analysis with respect to the maximum sequence length (Table 7) were performed.

4 Methodology

This study compares six models selected to encompass both strong tabular baselines and standard sequential architectures used for modelling clickstream data. In the tabular setting, SVM with an RBF kernel, XGBoost, and CatBoost were used. In the sequential setting, three recurrent architectures were applied: a vanilla RNN, LSTM, and GRU. All models were evaluated under identical experimental conditions, using the same input data, temporal split, and performance metrics.

Table 1. Characteristics of the source data and the test set.

Domain	Total visits	Total users	Total purchases	Test users	Test purchases	Positive class ratio (test)
Domain A (specialist)	2,062,013	529,874	33,036	79,020	12,836	16.24%
Domain B (bookshop)	476,417	133,570	6,814	21,920	2,859	13.04%
Domain C (pharmacy)	74,834	27,287	567	2,670	170	6.37%
Total	2,613,264	690,731	40,417	103,610	15,865	15.31%

SVM was treated as a standard baseline, consistent with established research practice. Due to its quadratic training cost, it was trained on a deterministically sampled subset not exceeding 15,000 instances. XGBoost and CatBoost were selected as representatives of modern boosting frameworks, which are often among the strongest available models for tabular tasks. The sequential models shared a common architectural backbone, differing only in the type of recurrent cell employed.

4.1 Model Configuration

All sequential models were trained using the Adam optimizer ($lr = 5 \cdot 10^{-4}$), with a batch size of 768, for a maximum of 12 epochs. Early stopping was applied based on validation AUC-PR (*patience* = 4). To mitigate overfitting and training instability, dropout was used (0.30 between recurrent layers and 0.20 in the classification head), together with gradient clipping (max norm = 1.0).

The shared architecture for all three sequential models is structured as follows: an input sequence of shape (B, T, F), where B denotes the batch size, T the number of time steps (visits), and F = 21 the number of features. The sequence is processed using `pack_padded_sequence` operation with right-side padding. It is then passed through two bidirectional recurrent layers with a hidden size of 128, resulting in an output dimensionality of 256 per time step.

After unpacking, a dual aggregation scheme is applied: masked mean pooling and masked max pooling, whose outputs are concatenated, producing a 512-dimensional vector. The classification head consists of a linear layer (512–64) with ReLU activation and a dropout rate of 0.20, followed by a final linear layer (64–1) producing a logit. Masking ensures that padding positions do not affect pooling statistics. This shared architecture ensures that any performance differences can be attributed to the properties of the recurrent cell rather than differ-

ences in architectural design.

4.2 Class Imbalance and Preprocessing

Each model configuration was trained using two approaches to address class imbalance. In the class weighting approach, models were trained on the full training set, with class weights inversely proportional to class frequencies. For XGBoost, the `scale_pos_weight` parameter was used, while CatBoost employed `sample_weight`. For sequential models, a weighted loss function based on focal loss (Focal BCE, $\gamma = 2.0$) was applied, with an optional `pos_weight` parameter. In the downsampling approach, the majority class was randomly under-sampled to achieve a 1:1 class ratio prior to training.

The validation and test sets maintained their natural class distribution, regardless of the training approach used. This allowed for threshold tuning on a distribution that closely mirrors real-world conditions. Such an approach is particularly important when the objective is to maximise a specific performance metric or to calibrate the model according to the relative costs of false positive (FP) and false negative (FN) errors, rather than focusing solely on class separation.

Preprocessing involved standardisation or scaling tailored to each model family, missing value imputation, and masking of short sequences. Furthermore, random seeds were employed to control model initialisation, batch ordering, and sampling, thereby ensuring the reproducibility of the experiment.

4.3 Metrics, Thresholds, and Statistical Tests

The primary evaluation metric was AUC-PR (Area Under the Precision–Recall Curve), defined as the area under the curve connecting precision and recall values across successive decision thresholds. In tasks with a low prevalence of the posi-

Table 2. Compared models and key hyperparameters.

Model family	Model	Hyperparameters
Tabular	SVM	RBF kernel, probability=True, pipeline with RobustScaler, subset $\leq 15,000$
Tabular	XGBoost	n_estimators=600, max_depth=6, lr = 0.07, tree_method=hist, early stopping (patience = 30)
Tabular	CatBoost	iterations=600, depth=6, lr = 0.07, od_wait=30, eval_metric=PRAUC
Sequential	RNN	2 bidirectional layers, hidden = 128, dropout = 0.30, avg+max pooling, MLP head (512–64–1)
Sequential	LSTM	same as above, LSTM cell
Sequential	GRU	same as above, GRU cell

tive class, AUC-PR better reflects the practical utility of a model than accuracy or even AUC-ROC (Area Under the Receiver Operating Characteristic Curve), as it focuses on the precision–recall trade-off for the minority class. For a random classifier, the AUC-PR value corresponds to the positive class prevalence in the dataset (approximately 0.153 in the test set), providing a natural performance baseline.

Additionally, AUC-ROC was reported, which measures the model’s ability to discriminate between classes across the full range of decision thresholds, as well as the weighted F1-score (F1_weighted), defined as a class-frequency-weighted average of the harmonic mean of precision and recall, computed separately for the positive and negative classes. The decision threshold thr^* was selected on the validation set as the value maximising the F1-score via grid search over 0.10, 0.15, ..., 0.90. In practice, this threshold has important business implications, as it determines the trade-off between costly over-targeting and missing potential customers.

Each configuration was trained five times using random seeds 42–46, with the mean and standard deviation of the metrics reported. For AUC-PR and AUC-ROC, 95% bootstrap confidence intervals were computed on the test set (600 resamples). Differences between models were evaluated using paired bootstrap tests and McNemar’s test for binary classification tasks at the selected threshold. To control the risk of Type I error in multiple comparisons, p-values were adjusted using Holm’s method.

5 Experimental Results

5.1 Main Results on the Temporal Split

Table 3 presents a comprehensive summary of results for all 12 configurations (6 models $\times$ 2 imbalance handling methods) on the temporal split, averaged over five seeds. The configurations are ranked in descending order according to AUC-PR on the test set.

Sequential models achieve the highest AUC-PR values (see the upper part of Table 3). The top-performing LSTM configuration with downsampling yields a test set score of 0.3789 ± 0.0051 , while the RNN with downsampling and the GRU with class weighting achieve closely comparable results. This represents an improvement of approximately 0.015 in AUC-PR over the leading tabular model, XGBoost with downsampling (0.3636). While the magnitude of this improvement is not substantial, it is characterised by stability, reproducibility, and methodological significance. These findings indicate that when true visit sequences are used, the temporal signal is no longer merely apparent but provides measurable value (see Figure 1).

It is worth noting that the ranking of models depends on the chosen evaluation metric. The RNN achieves the highest weighted F1-score (0.8116), although it slightly underperforms the LSTM in terms of AUC-PR. From a practical perspective, this does not constitute a contradiction, as AUC-PR evaluates model performance across the full range of decision thresholds, whereas F1 depends on a single decision threshold. This finding confirms that, in a deployment setting, it is important to distinguish between the model training stage and the decision policy calibration stage.

Among the tabular models, XGBoost clearly outperforms both CatBoost and SVM. CatBoost exhibits relatively high variance across random seeds (AUC-PR std ≈ 0.04), which may result from its internal feature and row sampling per tree.

5.2 Confidence Intervals and Pairwise Comparisons

Bootstrap 95% confidence intervals (CIs) for the top-performing sequential configurations and XGBoost indicate a clear separation between these model families (see Figure 2). For the LSTM with downsampling, the interval is [0.3719, 0.3878]; for the RNN with downsampling, [0.3730, 0.3889]; and for the GRU with class weighting, [0.3724, 0.3884]. The best XGBoost model yields an interval of [0.3554, 0.3711]. These results confirm that the advantage of sequential models is not driven by a single favourable seed, but persists under test set resampling.

Pairwise bootstrap comparisons and the McNemar test yield consistent conclusions: the top-performing sequential configurations significantly outperform SVM and CatBoost, and in several comparisons also outperform XGBoost after applying Holm's correction. From a practical standpoint, this indicates that the observed performance gain is meaningful. The differences between LSTM, RNN, and GRU are minor, suggesting that the choice of the sequential model family is more important than the choice of a specific recurrent cell variant.

5.3 Performance Breakdown by Store

Table 4 presents the AUC-PR values disaggregated across three sales domains for the top-performing configurations of both model families. The aim of this analysis was to verify whether the aggregate performance is not disproportionately influenced by Domain A (a specialist retailer), which accounts for 76% of the users in the test set.

The advantage of sequential models persists across all three sales domains (see Table 4 and Figure 2). In each domain, the top-performing sequential models consistently outperform the best tabular models. Furthermore, the performance variation across domains is moderate: for the leading sequential models, the gap between the highest- and lowest-performing domains does not exceed

approximately 0.03–0.04 in AUC-PR. These findings indicate that the advantage of sequential models is not an artefact of averaging, but a consistent pattern across the dataset.

Domain B (a bookshop) appears relatively more challenging than Domains A and C, even though its positive class prevalence is not the lowest. This may indicate more heterogeneous browsing paths or a higher proportion of exploratory behaviour that does not lead to purchase. In contrast, Domain C (an online pharmacy), despite having the smallest sample size, does not significantly impact the stability of the model ranking.

5.4 Cross-Store Generalisation (Leave-One-Shop-Out)

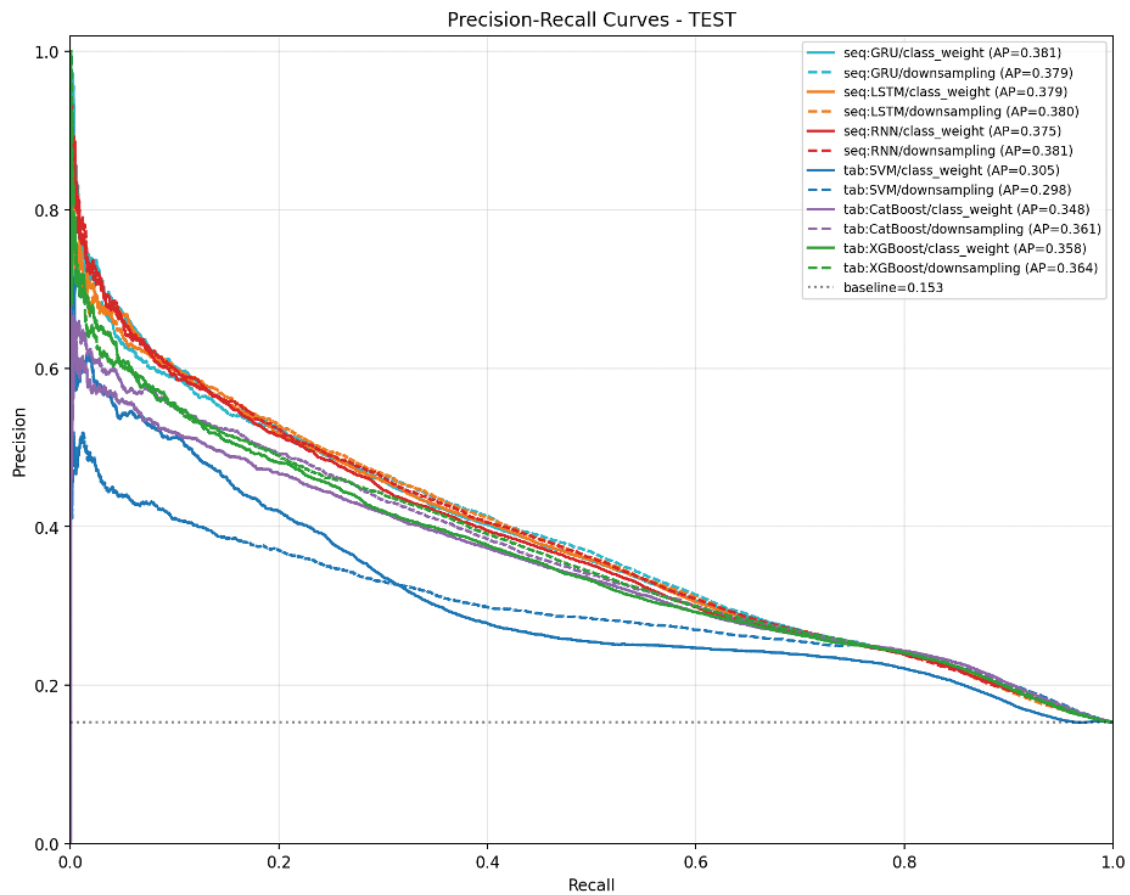
The leave-one-shop-out experiment evaluates whether a model trained on data from two stores can effectively classify users from a third, previously unseen store. Table 5 presents the mean AUC-PR values for the key configurations, with separate columns for each held-out domain alongside the overall average across the three scenarios.

The results of the leave-one-shop-out experiment (Table 5) are markedly lower than those achieved in the in-distribution temporal split setting. The drop in mean AUC-PR of approximately 0.13–0.15 represents an expected consequence of transferring models across domains characterised by differing assortments, traffic structures, and purchasing dynamics. Despite this decline, sequential models still exceed the baseline and maintain their leading positions in the ranking.

The most significant challenges in post-transfer prediction are found in Domain C, which can be attributed to its smallest sample size and the lowest baseline positive class prevalence. A model trained on the two smaller domains is capable of achieving satisfactory performance on Domain A, indicating that transfer is not restricted solely to the direction from the largest store towards smaller ones. This suggests that for a new sales domain, bootstrapping a model using data from existing stores is a viable strategy, although a decrease in performance compared to in-domain training should be expected.

Table 3. Results of all configurations on the temporal split (5 seeds, mean $\pm$ std).

Model family	Model	Imbalance handling	thr*	Validation AUC-PR	Test AUC-PR	Test AUC-ROC	Test F1w
Sequential	LSTM	downsampling	0.61	0.2644 $\pm$ 0.0019	0.3789 $\pm$ 0.0051	0.7414 $\pm$ 0.0020	0.7996 $\pm$ 0.0073
Sequential	RNN	downsampling	0.62	0.2647 $\pm$ 0.0026	0.3785 $\pm$ 0.0041	0.7410 $\pm$ 0.0020	0.8116 $\pm$ 0.0085
Sequential	GRU	class_weight	0.61	0.2682 $\pm$ 0.0012	0.3784 $\pm$ 0.0035	0.7415 $\pm$ 0.0015	0.8098 $\pm$ 0.0056
Sequential	GRU	downsampling	0.61	0.2648 $\pm$ 0.0012	0.3773 $\pm$ 0.0019	0.7429 $\pm$ 0.0019	0.8056 $\pm$ 0.0086
Sequential	LSTM	class_weight	0.61	0.2667 $\pm$ 0.0006	0.3768 $\pm$ 0.0045	0.7414 $\pm$ 0.0022	0.8073 $\pm$ 0.0074
Sequential	RNN	class_weight	0.61	0.2649 $\pm$ 0.0018	0.3724 $\pm$ 0.0031	0.7378 $\pm$ 0.0016	0.8078 $\pm$ 0.0079
Tabular	XGBoost	downsampling	0.75	0.2576 $\pm$ 0.0000	0.3636 $\pm$ 0.0000	0.7405 $\pm$ 0.0000	0.7973 $\pm$ 0.0000
Tabular	XGBoost	class_weight	0.75	0.2556 $\pm$ 0.0000	0.3575 $\pm$ 0.0000	0.7354 $\pm$ 0.0000	0.7983 $\pm$ 0.0000
Tabular	CatBoost	downsampling	0.67	0.2348 $\pm$ 0.0309	0.3382 $\pm$ 0.0404	0.7331 $\pm$ 0.0162	0.7651 $\pm$ 0.0614
Tabular	CatBoost	class_weight	0.65	0.2334 $\pm$ 0.0211	0.3341 $\pm$ 0.0271	0.7346 $\pm$ 0.0071	0.7697 $\pm$ 0.0701
Tabular	SVM	downsampling	0.76	0.2088 $\pm$ 0.0076	0.2960 $\pm$ 0.0087	0.7161 $\pm$ 0.0034	0.7679 $\pm$ 0.0137
Tabular	SVM	class_weight	0.10	0.2137 $\pm$ 0.0073	0.2908 $\pm$ 0.0113	0.6721 $\pm$ 0.0117	0.7922 $\pm$ 0.0119

**Figure 1.** Precision–recall curves for 12 configurations on the test set. Solid line – class_weight; dashed line – downsampling. Horizontal line shows the positive class prevalence baseline (≈ 0.153).

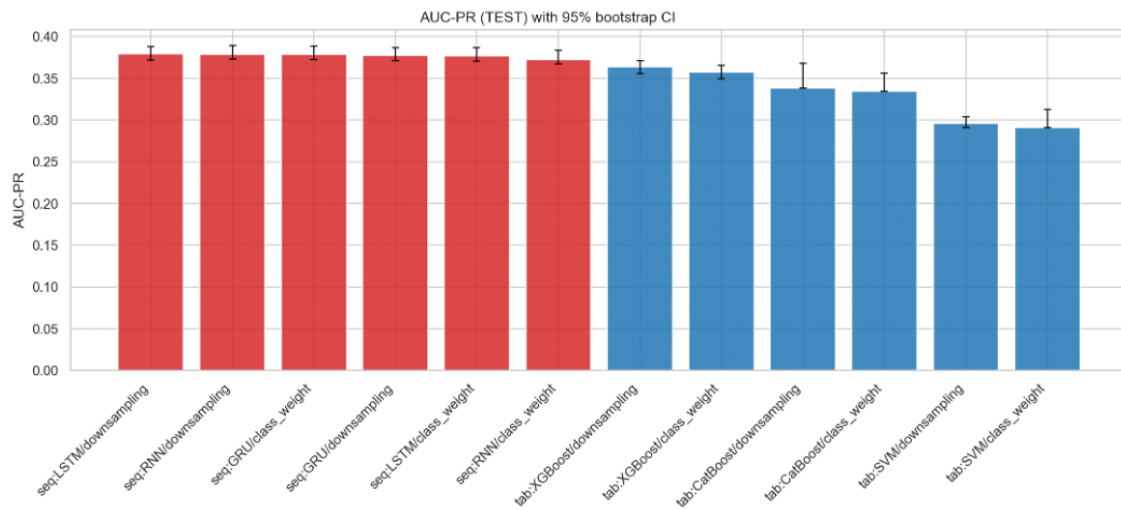


Figure 2. AUC-PR point estimates for all 12 configurations on the test set, with 95% bootstrap confidence intervals ($N = 600$). Sequential models form a distinct top group.

Table 4. Test set AUC-PR by sales domain.

Model	Imbalance handling	Domain A (specialist)	Domain B (bookshop)	Domain C (pharmacy)
GRU	class_weight	0.3856	0.3559	0.3834
LSTM	downsampling	0.3847	0.3568	0.3783
RNN	class_weight	0.3806	0.3488	0.3842
XGBoost	downsampling	0.3697	0.3423	0.3665
CatBoost	downsampling	0.3659	0.3373	0.3617
SVM	downsampling	0.3068	0.2570	0.2890

Table 5. Leave-one-shop-out AUC-PR performance (training on two domains, testing on a held-out domain).

Model	Imbalance handling	Domain C (held-out)	Domain A (held-out)	Domain B (held-out)	Mean AUC-PR
GRU	class_weight	0.2187	0.2586	0.2248	0.2340
LSTM	class_weight	0.2152	0.2586	0.2246	0.2328
RNN	class_weight	0.2088	0.2561	0.2243	0.2297
XGBoost	class_weight	0.2071	0.2457	0.2285	0.2271
CatBoost	class_weight	0.2068	0.2360	0.2302	0.2243
SVM	class_weight	0.1317	0.1809	0.1472	0.1533

5.5 Feature Ablation Study

Table 6 presents the ablation results for three feature groups: the baseline feature set (5 features), the extended set augmented with mouse movement and scrolling features (12 features), and the full feature set (21 features).

The main conclusion from the ablation study (Table 6) is the high predictive value of features related to mouse movement and scrolling. For most models, extending the baseline feature set with these variables yields an improvement in AUC-PR of approximately 0.01 to 0.03. This indicates that micro-interactions constitute a meaningful signal of purchasing behaviour, rather than merely “interface noise”.

Keyboard and focus-related features contribute much less to the prediction process. In some configurations, the improvement is marginal, while in others a slight decrease in performance occurs. This may stem from a lower frequency of such activities, their lower stability, and their stronger dependence on page layout rather than on user purchasing intent. From a feature engineering perspective, this provides an important guideline: when operating under a limited computational budget or restricted data availability, it is advisable to prioritise signals related to mouse movement and scrolling.

5.6 Impact of Sequence Length

Table 7 presents changes in AUC-PR for sequential models with maximum user history lengths of 10, 20, 40, and 80 visits.

For all sequential model families, increasing the user history length from 10 to 40 visits leads to a clear improvement in performance (see Table 7). This finding confirms that very short user sequences are insufficient to capture stable behavioural patterns. Conversely, the gain observed between 40 and 80 visits is marginal and, in several configurations, practically negligible. This suggests that a sequence length of 40 steps represents an optimal trade-off between performance and computational cost.

It is worth noting that the overall ranking of model families remains consistent regardless of sequence length. Both GRU and LSTM models maintain comparably high predictive performance, while

the vanilla RNN remains slightly less stable. These findings suggest that the advantage of gated architectures stems not only from their ability to handle long sequences, but also from their enhanced ability to model temporally extended behavioural patterns.

6 Conclusion

This paper investigates the prediction of future e-commerce customer purchases using real-world visit sequences and behavioural micro-interactions. In contrast to approaches based on simplified user representations, the study employs a comprehensive data pipeline, real-world multi-visit sequences, a temporal evaluation protocol, and a broad set of models and complementary analyses.

The primary finding of this study is unequivocal: for the analysed task, sequential models demonstrate a sustained advantage over tabular models. The top-performing LSTM, RNN, and GRU configurations achieve an AUC-PR of approximately 0.379, whereas the strongest tabular model, XGBoost, yields a score of roughly 0.364. This advantage persists across individual sales domains and remains evident in the cross-domain experiment, although inter-store transfer results in a noticeable decline in predictive performance.

In-depth analyses reveal that mouse movement and scrolling features provide the highest predictive power, whereas keyboard and focus-related signals are of limited value. Extending user history improves performance up to approximately 40 visits, beyond which further increases yield only marginal gains. These findings carry both theoretical and practical implications, as they identify the most valuable data sources and demonstrate how to optimise the trade-off between computational cost and predictive performance.

Future work should extend these comparisons to include transformer architectures [12][20], domain adaptation procedures, and probability calibration with respect to specific cost functions. Bilski et al. [2] proposed the FSGQR algorithm as a scalable alternative to the Adam optimizer for neural network training, which could be leveraged to accelerate the training of sequential models and improve their generalisation on imbalanced e-commerce datasets. Furthermore, Romaszewski et

Table 6. Feature ablation results – AUC-PR on the test set.

Model	Imbalance handling	Baseline feature set (5)	Extended feature set (12)	Full feature set (21)	Conclusion
LSTM	downsampling	0.3654	0.3864	0.3789	best performance after adding mouse/scroll features
RNN	downsampling	0.3627	0.3871	0.3785	full feature set slightly degrades performance
GRU	class_weight	0.3653	0.3783	0.3784	keyboard/focus features are neutral
XGBoost	downsampling	0.3487	0.3557	0.3636	mouse/scroll signals are beneficial
CatBoost	downsampling	0.3168	0.3484	0.3382	mouse/scroll signals are highly predictive
SVM	downsampling	0.2851	0.3076	0.2960	full feature set does not improve performance

Table 7. Impact of maximum sequence length on AUC-PR of sequential models.

Model	Imbalance handling	10 visits	20 visits	40 visits	80 visits
GRU	class_weight	0.3506	0.3706	0.3770	0.3857
GRU	downsampling	0.3524	0.3713	0.3796	0.3803
LSTM	class_weight	0.3466	0.3641	0.3724	0.3742
LSTM	downsampling	0.3530	0.3731	0.3809	0.3826
RNN	class_weight	0.3545	0.3699	0.3754	0.3809
RNN	downsampling	0.3450	0.3639	0.3721	0.3716

al. [17] analysed Random Forest-based models in the context of numerical data processing, confirming their effectiveness on tabular data. This supports the inclusion of tree-based models (XGBoost, LightGBM) in future comparisons with the recurrent neural networks evaluated in this study. Finally, an important direction for future work is the integration of sequential models with explainable modules, enabling both high predictive performance and transparent decision-making for marketing and analytics teams.

References

- [1] J. A. Bastos, M. I. Bernardes, Understanding Online Purchases with Explainable Machine Learning, *Information*, vol. 15, no. 10, art. 587, 2024.
- [2] J. Bilski, B. Kowalczyk, L. Dymova, M. Xiao, Accelerating Neural Network Training with FS-GQR: A Scalable and High-Performance Alternative to Adam. *Journal of Artificial Intelligence and Soft Computing Research*, 15(2), 2025. doi:10.2478/jaiscr-2025-0006.
- [3] N. Chaudhuri, G. Gupta, V. Vamsi, I. Bose, On the platform but will they buy? Predicting customers' purchase behavior using deep learning, *Decision Support Systems*, vol. 149, 113622, 2021.
- [4] S. Chen, X. Wang, H. Zhang, J. Wang, Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine, *Expert Systems with Applications*, vol. 173, 114756, 2021.
- [5] S. Chen, Z. Xu, D. Xu, X. Gou, Customer purchase prediction in B2C e-business: A systematic review and future research agenda, *Expert Systems with Applications*, vol. 252, 124261, 2024.
- [6] F. Ehsani, M. Hosseini, Customer purchase prediction in electronic markets from clickstream data using the Oracle meta-classifier, *Operational Research*, vol. 24, no. 1, art. 11, 2024.
- [7] R. Esmeli, M. Bader-El-Den, H. Abdullahi, Towards early purchase intention prediction in online session based retailing systems, *Electronic Markets*, vol. 31, no. 3, pp. 697–715, 2021.
- [8] M. Gabryel, M. Kocić, A. Gabryel Prediction of Future Customer Purchases in an Online Store Based on Web Session Behavioral Data, In: *Proceedings in The 25th International Conference on Artificial Intelligence and Soft Computing 2026*, accepted for publication.
- [9] M. Gabryel, K. Grzanek, Y. Hayashi. Browser fingerprint coding methods increasing the effectiveness of user identification in the web traffic. *Journal of Artificial Intelligence and Soft Computing Research* 10.4, pp. 243-253, 2020.
- [10] M. Gabryel, E. Kocić, M. Kocić, Z. Patora-Wysocka, M. Xiao, M. Pawlak, Accelerating User Profiling in E-Commerce Using Conditional GAN Networks for Synthetic Data Generation, *Journal of Artificial Intelligence and Soft Computing Research*, vol. 14, no. 4, 2024.

- [11] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in *Advances in Neural Information Processing Systems* 35, 2022.
- [12] W.-C. Kang, J. McAuley, Self-Attentive Sequential Recommendation, in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.
- [13] S. Kim, W. Shin, H.-W. Kim, Predicting online customer purchase: The integration of customer characteristics and browsing patterns, *Decision Support Systems*, vol. 177, 114105, 2024.
- [14] D. Koehn, S. Lessmann, M. Schaal, Predicting online shopping behaviour from clickstream data using deep learning, *Expert Systems with Applications*, vol. 150, 113342, 2020.
- [15] Martínez, C. Schmuck, S. Pereverzyev Jr., C. Pirker, M. Haltmeier, A machine learning framework for customer purchase prediction in the non-contractual setting, *European Journal of Operational Research*, vol. 281, no. 3, pp. 588–596, 2020.
- [16] Requena, G. Cassani, J. Tagliabue, C. Greco, L. Lacasa, Shopper intent prediction from clickstream e-commerce data with minimal browsing information, *Scientific Reports*, vol. 10, 16983, 2020.
- [17] M. Romaszewski, P. Sekuła, P. Głomb, M. Cholewa, K. Kołodziej, Through the Thicket: A Study of Number-Oriented LLMs Derived from Random Forest Models. *Journal of Artificial Intelligence and Soft Computing Research*, 15(3), 2025, doi:10.2478/jaiscr-2025-0014.
- [18] C. O. Sakar, S. O. Polat, M. Katircioglu, Y. Kastro, Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks, *Neural Computing and Applications*, vol. 31, pp. 6893–6908, 2019.
- [19] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, *Information Fusion*, vol. 81, pp. 84–90, 2022.
- [20] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1441–1450, 2019.
- [21] H. Vaghari, M. Hosseinzadeh Aghdam, H. Emami, Diarec: Dynamic intention-aware recommendation with attention-based context-aware item attributes modeling, *Journal of Artificial Intelligence and Soft Computing Research*, vol. 14, no. 2, 2024.
- [22] C. Zhang, J. Liu, S. Zhang, Online Purchase Behavior Prediction Model Based on Recurrent Neural Network and Naive Bayes, *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 4, pp. 3461–3476, 2024.



Marcin Gabryel earned his PhD degree in computer science at Czestochowa University of Technology, Poland, in 2007. He is an assistant professor in the Department of Artificial Intelligence at Czestochowa University of Technology. His research focuses on developing new methods in computational intelligence and data mining. He

has published over 60 research papers. His present research interests include deep learning architectures and their applications in databases and security.

<https://orcid.org/0000-0002-6701-0460>



Milan Kocić received his BEng Computer Science degree from Gdańsk University of Technology. He has been working as software engineer at Spark Digitup since 2016, mainly focusing on maintenance and development of marketing fraud protection systems and ecommerce chatbots.

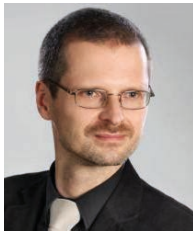
His areas of work include device fingerprinting, anomaly detection and integration with AI solutions.

<https://orcid.org/0000-0002-6257-0936>

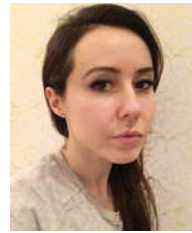


Aleksandra Gabryel graduated with a BEng in Computational Engineering from AGH University of Krakow. She is currently continuing her studies at the same institution, pursuing an MSc in Technical Computer Science and further developing her competences in modern information technologies, data analysis, and artificial intelligence. She

is interested in the application of artificial intelligence in IT systems, process automation, and the development of solutions that support efficient application and data management.



Marek Kisiel-Dorohinicki conducts research in the field of agent systems and metaheuristic computations. He is also interested in software engineering problems. For many years, he has been running prestigious projects financed by the National Centre for Research and Development in the area of Security and Defense, implementing many implementations of the developed software. He performs a number of organizational and management functions in the structure of the University, being, among others, the Director of the Implementation Management Center and the AGH Center for Security Technologies. He serves as Dean of the Faculty of Computer Science at AGH University of Krakow.
<https://orcid.org/0000-0002-8459-1877>



Zofia Patora-Wysocka is a professor at the University of Social Science in Łódź, Poland. She received the Ph.D. degree from the Częstochowa University of Technology, Częstochowa, Poland in 2008, and the D.Sc. degree in economic sciences from the WSB University in Dąbrowa Górnicza in 2020. Her research interest includes change management, routine dynamics and strategy, practice theory, science and technology studies, and applications of data mining and artificial intelligence methods in management.
<https://orcid.org/0000-0002-0429-0207>