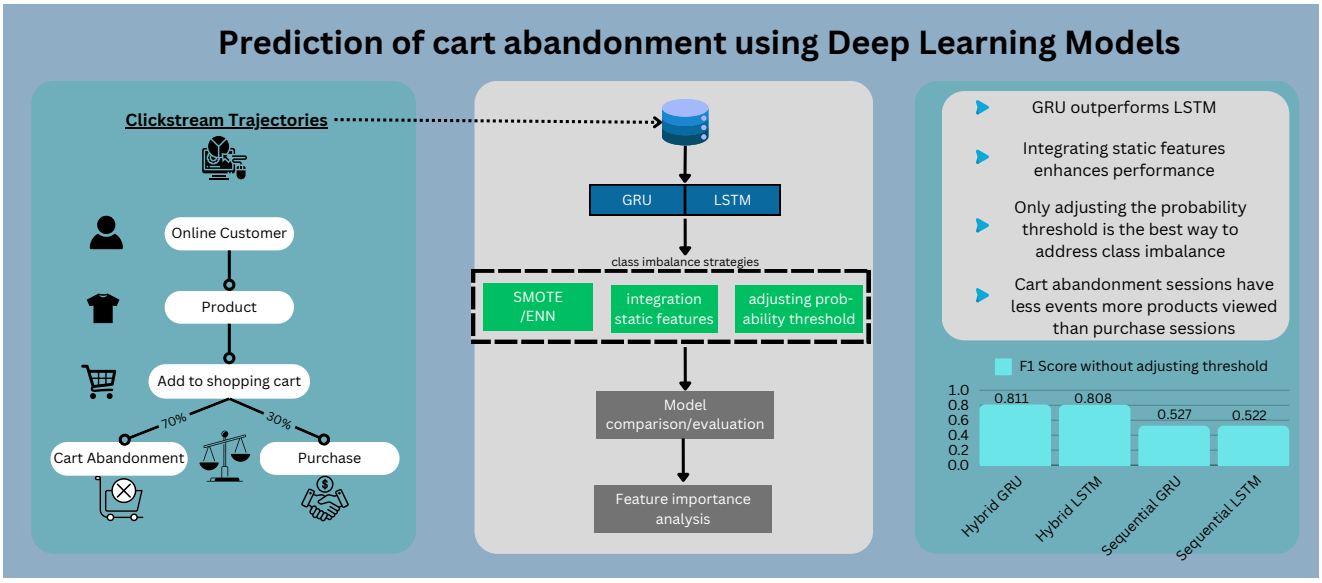
Graphical Abstract

Prediction of cart abandonment using imbalance clickstream data in online shopping

Fabian Waldmann, Gonzalo Nápoles, Yamisleydi Salgueiro



Highlights

Prediction of cart abandonment using imbalance clickstream data in online shopping

Fabian Waldmann, Gonzalo Nápoles, Yamisleydi Salgueiro

- Deep neural network models are evaluated for early prediction of cart abandonment using imbalanced data.
- We propose a probability threshold adjustment to compute the classes, which helps tackle class imbalance.
- In contrast, undersampling and oversampling methods introduce noise and bias towards the minority class.
- Integrating static features alongside sequential data significantly enhances the neural networks' performance.
- We determine the relevance of sequential and statics features for early prediction of cart abandonment.

Prediction of cart abandonment using imbalance clickstream data in online shopping

Fabian Waldmann[a], Gonzalo Nápoles[a,*] and Yamisleydi Salgueiro[b]

[a]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands
[b]Department of Industrial Engineering, Faculty of Engineering, Universidad de Talca, Campus Curicó, Chile.

ARTICLE INFO

Keywords:
Cart abandonment
Deep Learning
Class imbalance
Clickstream data
Explainable AI

ABSTRACT

Numerous machine learning techniques have been developed to predict online shopping behavior and cart abandonment. However, due to inherent class imbalance, this task remains challenging even for high-performing models like neural networks. To address class imbalance, we propose adjusting the probability threshold for deep neural networks, effectively enhancing class discrimination and performance. This method, not extensively explored in previous research, is further improved with Shapley Additive exPlanations (SHAP) for better interpretability and effectiveness and has the potential to enhance transparency and trust by revealing biases in online shopping. In this context, we found that traditional undersampling and oversampling methods introduce noise and a bias towards the minority class. Integrating static features alongside sequential data further boosts the performance of neural networks, aligning with the results of previous research.

1. Introduction

Approximately 30 % of European fashion retail sales within the fashion sector are presently conducted through digital stores (Statista, 2023). Given that online shopping is particularly popular with younger target groups, a substantial increase in digital shops is anticipated and the fundamental challenge arises to understand online customer behavior. An elementary classification represents customers who place a product in their online shopping cart and then decide whether to buy a product or abandon the products in their cart. Research has shown that around 70.19 % of items added to an online shopping cart were never purchased (Baymard-Institute, 2024). This underscores the substantial potential for scrutinizing the underlying reasons for abandonment, which can be extracted through the examination from clickstream data (Sakar, Polat, Katircioglu and Kastro, 2019). However, numerous studies in the e-commerce domain concentrate on purchase intent, whereas the associated topic of shopping cart abandonment has gained less attention.

With the emergence of deep learning and its superior predictive performance, neural networks have been used more frequently in online retail. Nevertheless, the interpretability of these black-box algorithms is challenging, potentially leading to discrimination and manipulation of consumers in an unfavorable manner (Zwitter, 2023). The European Parliament has addressed this concern in an in-depth analysis, emphasizing that an adequate level of explainability of recommendation mechanism is of great importance, especially in the E-Commerce sector (Pedreschi and Miliou, 2020). In this way, decisions made utilizing black box models can be made more trustworthy and ethical principles can be embedded (Nannini, Balayn and Smith, 2023).

This paper addresses this need and centers on an explainable approach for predicting the likelihood of online shopping cart abandonment across various customers. It compares a Markov Chain baseline with two deep learning approaches: a Gated Recurrent Unit (GRU) network (Cho, van Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk and Bengio, 2014) and a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). Given the inherent class imbalance in online shopping sessions, the performance of these sequential models will be evaluated using the F1 score.

To mitigate the negative effects of class imbalance, two strategies are implemented as imbalance data strategies for deep learning models remain relatively underexplored (Johnson and Khoshgoftaar, 2019). On the one hand, the popular oversampling Synthetic Minority Oversampling Technique (SMOTE) (Chawla, Bowyer, Hall and Kegelmeyer, 2002), and the undersampling approach, Edited Nearest Neighbors (ENN) (Wilson, 1972), are applied to the clickstream data to assess their impact on performance. On the other hand, since the outputs of the neural networks represent probabilities, the probability threshold for classifying the output is adjusted to address class imbalance. Subsequently, a feature importance analysis using SHAP (Lundberg, Erion and Lee, 2018) is conducted to identify which timesteps and actions contribute the most to the predictions and how the application of SMOTE changed the feature importance. Additionally, this research explores the impact of integrating static features alongside sequential features on the predictions.

Through the combined approach of early prediction of cart abandonment alongside the addressing of class imbalance, threshold adjustments, feature importance, and utilizing different types of features in deep learning approaches,

*Corresponding author
g.r.napoles@tilburguniversity.edu (G. Nápoles)
ORCID(s): 0009-0007-9830-8562 (F. Waldmann); 0000-0003-1936-3701 (G. Nápoles); 0000-0002-1946-0053 (Y. Salgueiro)

this study makes a significant and novel contribution to the scientific discourse.

Both GRU and LSTM deep learning models outperformed the traditional Markov Chain classifier in predicting cart abandonment. Notably, the hybrid GRU model, without addressing class imbalance and adjusting the probability threshold, achieved the highest performance with an F1 score of 0.811. The GRU and LSTM models exhibited similar behavior in terms of performance, addressing class imbalance, and feature importance analysis. Addressing the inherent class imbalance with SMOTE or ENN significantly improved training performance. However, the test performance was slightly lower compared to models that only adjusted the probability threshold without altering the underlying data.

From a societal perspective, this research demonstrated that the application of explainability techniques enhances the interpretability of the predictions, allowing the recognition of biases introduced by SMOTE and ENN. Additionally, it was concluded that online shopping customers who complete a purchase tend to have more interactions and time on the website viewing fewer products compared to those who abandon their products in the online shopping cart. With this knowledge, the user experience can be improved through advanced personalization, and the website can be better streamlined for classified buyers. From a technical perspective, it is recommended to address class imbalance by adjusting the probability threshold and incorporating static data alongside sequential data to enhance the performance of neural networks. This approach is relatively easy to implement and has received limited attention in previous research (Johnson and Khoshgoftaar, 2021).

The paper is structured as follows: Section 2 reviews related work, discussing methodologies and various approaches introduced in previous research. Building on this, Section 3 details the methodology of this study, including a description of the dataset, model selection, and evaluation processes. Section 4 presents and interprets the experimental results, offering a comparative analysis of neural networks under different class imbalance conditions. Finally, Section 5 summarizes the findings, discusses practical implications, addresses limitations, and provides suggestions for future research.

2. Related Work

This section explores the literature on various strategies for predicting cart abandonment. It covers the investigation of different machine learning models, feature selection techniques, class imbalance strategies, and interpretability analysis. By reviewing previous studies, this section highlights the strengths and weaknesses of these strategies, providing a clearer understanding.

2.1. Early Prediction of Cart Abandonment

In general, there is a substantial amount of research in the E-commerce domain, and several studies relating to cart abandonment. Most of them concentrated on the motives behind the exit of an online shop. Kukar-Kinney and Close (2010) identified various factors such as the enjoyment of shopping, the usefulness of online shopping carts as an organizational aid, and the expectation of sales as reasons behind cart abandonment. In a separate study, researchers explored the connection between customers' purchase intentions and mouse-related data. Their findings revealed that customers who are prepared to make a purchase tend to engage in more scrolling behavior (Guo and Agichtein, 2010).

Despite discovering the underlying reasons, the goal of prediction of cart abandonment is to anticipate when users are likely to leave an online shop, by analyzing sequential patterns in their navigation click data. This task presents a significant challenge since clickstream data is highly variable and influenced by various complex factors (Hatt and Feuerriegel, 2020). With sequential data, a classifier can be trained on past sessions with known class labels to learn a classification function that categorizes the current user sessions into one of the predefined classes (Koehn, Lessmann and Schaal, 2020). This predictive approach essentially involves classifying user behavior within a defined time frame, allowing us to treat the sequence of page views as time series data (Sakar et al., 2019).

One major study in this context dealt with a similar clickstream dataset predicting the user's intention, determining whether a session will involve a purchase action. Requena, Cassani, Tagliabue, Greco and Lacasa (2020) proposed a Markov Chain and an LSTM model in this investigation, with the LSTM demonstrating significantly higher precision. The authors highlight, the strong resemblances between Natural Language Processing (NLP) and clickstream prediction problems, emphasizing the sequential and discrete nature of both tasks. Given this alignment, the study advocates the application of the NLP baseline model, the Markov Chain, to address clickstream prediction challenges. In NLP, Markov Chains predict the next word based on preceding words, while in clickstream data analysis, they anticipate the next user action. Markov Chains have a fixed number of preceding events, which cannot be dynamically updated. As a result, they are limited in their ability to consider longer dependencies in the data (DelSole, 2000).

Additionally, a separate study comparing shopping behavior prediction methods reiterates the comparison between Markov Chain and LSTMs. Apart from their overarching finding that LSTMs outperform the Markov Chain, the study further underscores that abandonment sequences exhibit more similarities to purchase sessions than to sessions focused solely on browsing (Toth, Tan, Fabbrizio and Datta, 2017). Comparative evaluations of GRU and LSTM networks within the realm of deep learning reveal similar overall performance, with task-specific nuances influencing their effectiveness (Zhou, Zhang, Li, Li, Zhao, Wang and Wang, 2023). The performance is contingent upon the data's characteristics. LSTM tends to yield superior results for long-term sequences due to its adeptness at recognizing long-term dependencies. On the other hand, GRU networks demonstrate superiority in terms of convergence speed, CPU

time efficiency, parameter updating, and generalization capabilities. (Chung, Gulcehre, Cho and Bengio, 2014).

Another foundational study on shopping cart abandonment prediction was conducted by Hatt and Feuerriegel (2020). They analyzed clickstream data using Markov Chains as a baseline and contrasted it with LSTM and extended Markov Chain models. In addition to predicting shopping cart abandonment, this study places emphasis on identifying users who are at risk of abandonment early in their session, considering factors such as time spent on pages. For early prediction, Bogina, Kuflik and Mokryn (2016) also found that incorporating the temporal dynamics between actions can improve predictive accuracy whether sessions end with a "purchase" or not.

2.2. Dealing with Class Imbalance

Class imbalance is a major problem in machine learning methodologies and can significantly affect performance. Research has shown that the sensitivity to imbalances increases with increasing problem complexity and that non-complex, linearly separable problems are not affected by all levels of class imbalances (Johnson and Khoshgoftaar, 2019). In the E-Commerce domain, an intrinsic imbalance is inherited since the majority of the sessions end without any transaction. If the models are trained with imbalanced data, this will lead to a bias in the prediction of conversion. For example, all predictions will be classified as non-purchases since they represent the majority class (Esmeli, Bader-El-Den and Abdullahi, 2021). Identifying users who are most likely to convert is therefore both potentially very valuable and very difficult due to class imbalance and general noise in browsing data (Bigon, Cassani, Greco, Lacasa, Pavoni, Polonioli and Tagliabue, 2019).

To address this problem, numerous studies have proposed a minority oversampling technique such as SMOTE for clickstream data. SMOTE operates by identifying the k-nearest neighbors within the feature space of the minority class, subsequently generating synthetic samples through interpolation between the selected instance and its nearest neighbors (Chawla et al., 2002). Since deep learning approaches are sensitive to class imbalance, they were only outperforming other machine learning approaches after balancing the class distribution using SMOTE (Sakar et al., 2019). The study from Alex, Jhanjhi, Humayun, Ibrahim and Abulfaraj (2022) applied the method to time series forecasting in conjunction with an LSTM network. The implementation showed enhanced accuracy and outperformed the LSTM model without SMOTE.

However, it cannot be concluded that the use of SMOTE improves the performance of neural networks under all circumstances. The study by Hooshyar, Azevedo and Yang (2023) concludes that, depending on the characteristics of the data and the inconsistency in the distribution between the training and test data, there may also be a slight drop in performance. By generating synthetic samples based solely on neighboring instances, SMOTE fails to capture more intricate patterns. This can result in the creation of noisy or unrealistic instances, potentially leading to overfitting or diminished generalization performance of the model (Teslenko, Sorokina, Khovrat, Huliiev and Kyriy, 2023).

Next to oversampling techniques, also undersampling techniques proved to be beneficially, in some cases even outperforming methodologies like SMOTE (Esmeli et al., 2021). Among the most effective undersampling methods is the ENN Rule. It operates by removing instances from the majority class that are misclassified by their nearest neighbors from the minority class. This approach ensures the preservation of valuable instances that contribute to defining decision boundaries (Wang, Chukova and Nguyen, 2023). In a study across ten datasets, Alejo, Sotoca, Valdovinos and Toribio (2010) conducted experiments to assess the accuracy of ENNs combined with neural networks. Their findings suggested that while classification accuracy generally improved, aggressive instance removal by ENN had an adverse effect in some cases, leading to the loss of crucial boundary information (Wang et al., 2023).

Another method discussed in the literature involves adjusting the probability threshold for binary classification. Since certain machine learning algorithms, such as neural networks, output a probability between 0 and 1, the classification threshold can be adjusted to enhance performance (Johnson and Khoshgoftaar, 2021). This approach can be particularly useful for addressing class imbalances, as the probability distribution may be skewed towards the majority class rather than centered around 0.5. Consequently, adjusting the threshold can be beneficial in improving the F1 score (Lipton, Elkan and Naryanaswamy, 2014).

2.3. Explainable AI for sequential Data

In recent literature, explainable artificial intelligence (XAI) (Flores, Flores and Winograd, 1986) has gained increased attention due to the advancement of complex machine learning models. Its objective is to enhance the transparency, reliability, and accountability of AI systems by explaining the inner workings of black-box models (Thalpage, 2023). Various techniques have been proposed for explaining sequential data and deep learning approaches.

The study by Theissler, Spinnato, Schlegel and Guidotti (2022) stands out as a flagship study in the realm of XAI methods for sequential and time series data. Their comparative analysis of various methods led them to the finding that SHAP demonstrates superior capability in capturing the time points pertinent to the classification behavior of the model. SHAP is a post-hoc method that operates independently from the machine learning model itself, offering insights into what the model has learned post-training without altering its underlying structure (Lundberg et al., 2018). This capability enables SHAP to generate both local and global explanations, making it a reliable tool across various types of data (Saranya and Subhashini, 2023). Therefore, SHAP stands out as the only method capable of processing different types of data due to its additive property. This allows all

sequence-related features to be combined and reported as a single feature (Molnar, 2022).

Roshan and Zafar (2023) explored the potential of using SHAP to enhance the performance of deep learning models by identifying the most relevant features. They employed the KernelSHAP method, which employs a unique weighted linear regression approach to assess the importance of each feature. Furthermore, Villani, Lockhart and Magazzeni (2022) demonstrated the successful application of KernelSHAP to time series tasks. However, both studies acknowledge that despite yielding consistent results, the high computational costs remains a significant drawback. As a solution, the SHAP documentation suggests sampling from the data and employing k-means clustering to work with representative data (Lundberg et al., 2018).

Due to the non-intelligible nature of time series, it is much more difficult to use only attributions and their relevance values for explainability (Theissler et al., 2022). Despite these challenges, Requena et al. (2020) adopted a strategy of explaining clickstream data by identifying the most influential time steps in a subset of the dataset using SHAP. These critical features or time periods can subsequently be integrated into recommendation systems to mitigate potential shopping cart abandonment.

2.4. Combining sequential and static Data

While most recent studies primarily focus on user-item interactions using sequential data, few machine learning approaches incorporate additional static data available in e-commerce (Bauer and Jannach, 2023). Since sequential clickstreams can typically be grouped and aggregated into static attributes or features, they offer stable high-level information. Meanwhile, sequential data aids in understanding dynamic behavior and transitions (Melnykov, 2016).

Chandramohan and Ravindran (2018) conducted a study comparing the performance of an artificial neural network using sequential data versus static data. They found that the model built on hand-crafted static data outperformed the one trained on sequential data in terms of F1 score. However, a hybrid approach combining both types of data was not explored in this study.

One of the earliest studies, utilizing a hybrid approach was conducted by Wang, Konolige, Wilson, Wang, Zheng and Zhao (2013), which initially built a support vector machine model using only click events as sequential data. Subsequently, the study incorporated static click characteristics, such as average session length and average time between clicks, with the user's corresponding clickstream data. Given that the task involved predicting sybil detection in clickstream analysis, the dataset was highly imbalanced. The study concluded that the hybrid models significantly outperformed the sequential models, particularly in reducing false negatives, and were more resistant to the inherent bias of the majority class.

Another study by Sheil, Rana and Reilly (2020) using GRU and LSTM networks demonstrated that combining sequential and static data can enhance the differentiation between purchase and cancellation sequences, which are often very similar. A further experiment with deep neural networks reinforced the overarching finding that combining sequential and static data for predicting customer churn significantly improves the performance of neural networks (Mena, Coussement, Bock, Caigny and Lessmann, 2023). However, it was noted as a limitation that there is a significant need to explain the different contributions of sequential and static data and to use XAI tools that can support both types of data.

2.5. Literature Gap

Numerous studies in the field of e-commerce have aimed to predict online customer behavior. Due to the variable nature of clickstream data and the inherent class imbalance, this task has proven challenging and highly dependent on the data. In our literature review, we investigated various models, class imbalance strategies, explanatory methods, and feature selection procedures for clickstream data. Notably, few studies have specifically focused on cart abandonment, and explanations for this type of data were not sufficiently considered (Theissler et al., 2022). Since previous research has typically addressed individual components in isolation, our approach of combining these methodologies offers a novel contribution to the scientific discourse.

3. Research Methodology

In this section, we provide a comprehensive overview of the dataset and outline the various stages of the experimental setup. We begin by describing the raw dataset, followed by detailed methods for data exploration and processing. After splitting the data, we train the selected models, introduced in the literature review, using different class imbalance strategies and fine-tune their hyperparameters. The models' performances are then measured and compared using the F1 score on the test set. Finally, we conduct an error analysis and evaluate feature importance using the SHAP method. Figure 1 provides a visual representation of the complete research process.

3.1. Dataset Description

The data set of the "SIGIR 2021 E-Commerce Workshop Data Challenge" contains 36 million events and is one of the first data sets with such a large scope (Tagliabue, Greco, Roy, Yu, Chia, Bianchi and Cassani, 2021). It is a publicly accessible clickstream file containing nearly 5 million anonymized shopping sessions, with each line representing a separate interaction within a session. The click events were recorded over a period of three months, spanning from January 15, 2019 to April 15, 2019. A session encompasses a sequence of browsing events occurring within a designated time frame. The time threshold for defining a session is set at 30 minutes, in accordance with industry standards. Therefore, if two actions are executed by the same user with a time gap exceeding 30 minutes, they are assigned to separate sessions. The raw variables included in the dataset are listed in Table 1.

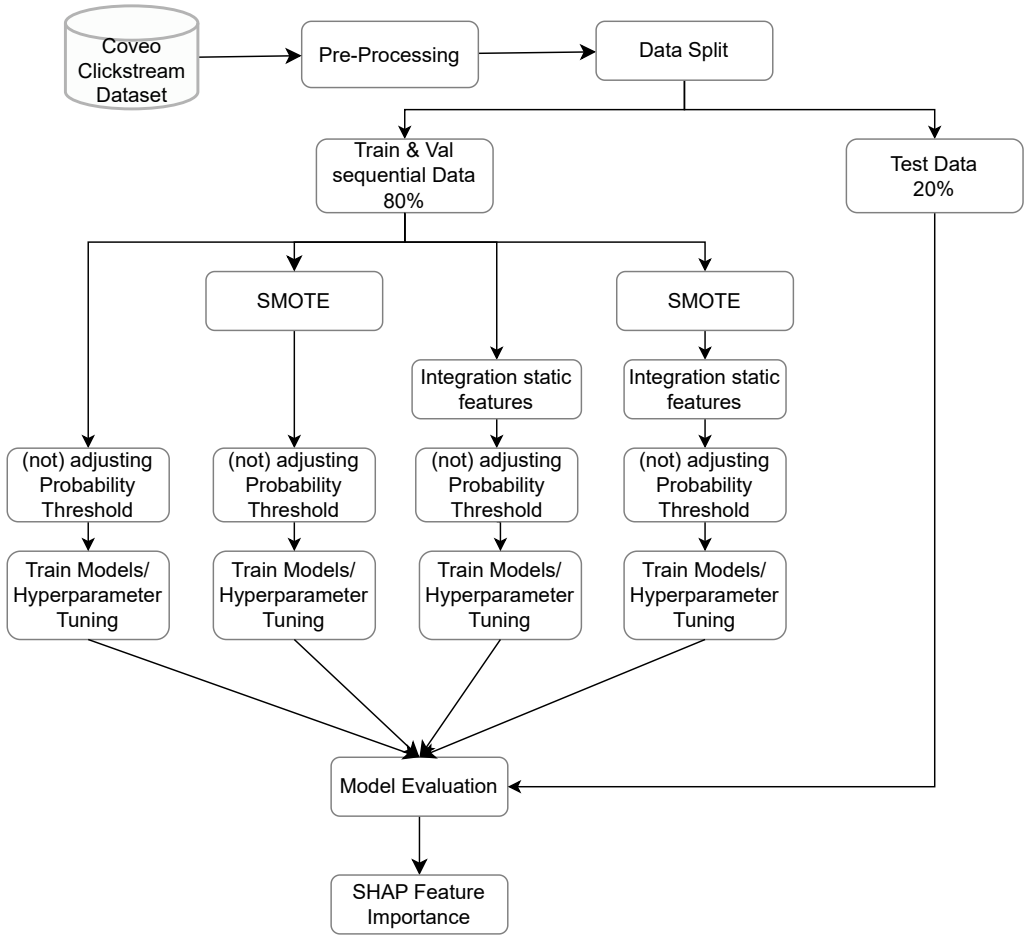


Figure 1: Workflow of research methodology. Following data preprocessing and exploratory data analysis, various neural network architectures are trained and evaluated using the F1 score to determine the best-performing model.

Previous research on clickstream data has been criticized for its low external validity. However, the dataset used in this study offers greater representativeness as it captures the web trajectories of typical users in the online industry (Requena et al., 2020). The selection of this dataset is justified by its depiction of an e-commerce shop with realistic patterns in terms of traffic, conversion rate, and, most importantly, abandonment rate.

3.2. Pre-Processing and exploratory Data Analysis

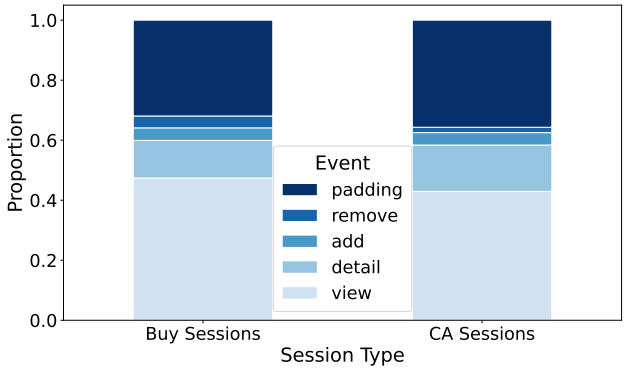
In preparation for pre-processing, we conduct an exploratory data analysis to gain a better understanding of the data. The initial phase of data preprocessing involves the exclusion of sessions that lack an “add-to-cart” event. Given the dataset’s highly imbalanced nature, characterized by a predominance of mere browsing sessions, a significant portion of the data is discarded at this stage. For labeling purposes, sessions that include a "purchase" event are marked with a "1", while those that feature an "add" event but no subsequent "purchase" are assigned a label of "0" for the GRU and LSTM networks. The Markov Chain uses a next-event prediction approach, where a separate label is not required. As demonstrated in Figure 2b, the frequency of cart abandonment sessions far exceeds that of purchase sessions. This disparity indicates a significant class imbalance which, as previously discussed, can lead to considerable challenges in our analysis.

While cart abandonment and purchase sequences share numerous similarities, noticeable differences only exist in the frequency of "detail" actions as depicted in Figure 2a. In the next step, actions are converted from character strings to numerical codes based on their frequency. The most frequently occurring action, "view," is assigned the lowest numerical code, "1", followed by "detail" as "2", "add" as "3", "remove" as "4", and "purchase" as "5". This approach significantly reduces the memory required for data storage and processing.

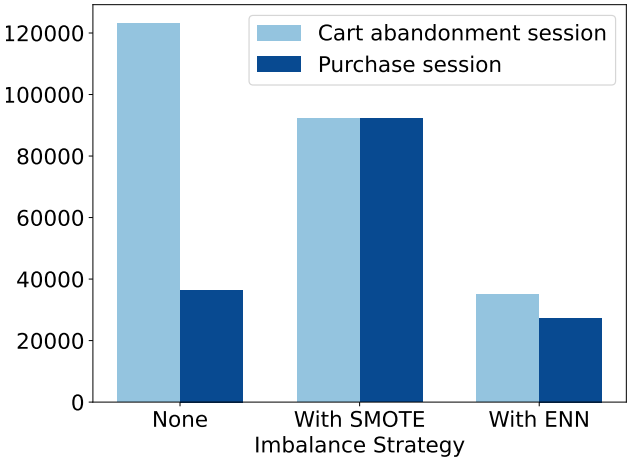
Furthermore, following the exclusion of purchase actions, sessions that are either too short or long are also filtered out. Adhering to the methodology proposed by Requena et al. (2020), sessions comprising fewer than 5 events are discarded. In an effort to promptly identify cart abandonment while minimizing computational demands, sessions exceeding 30 events are likewise eliminated. For sessions that last fewer than 30 events and longer than 5 events,

Table 1
Features of the raw data.

Name	Type	Data Description
session_id_hash	string	The hashed identifier signifies a shopping session, grouping events within a 30-minute timeframe. If a user returns after 31 minutes, a new session identifier is assigned.
event_type	enum	Event types according to the Google Protocol include “pageview" and "event."
product_action	enum	The product action can be one of detail, add, purchase, remove. If the field is empty, it describes a simple page view, such as the FAQ page, without associated products.
product_sku_hash	string	If the event is a product event, hashed identifier of the product in the event.
server_timestamp_epoch_ms	int	Epoch time, in milliseconds. As a further anonymization technique, the timestamp has been shifted by an unspecified amount of weeks, keeping intact the intra-week patterns.
hashed_url	string	Hashed url of the current web page.



(a) Event distribution in sessions



(b) Class distribution by imbalance strategy

Figure 2: Comparative analysis of event and class distributions.

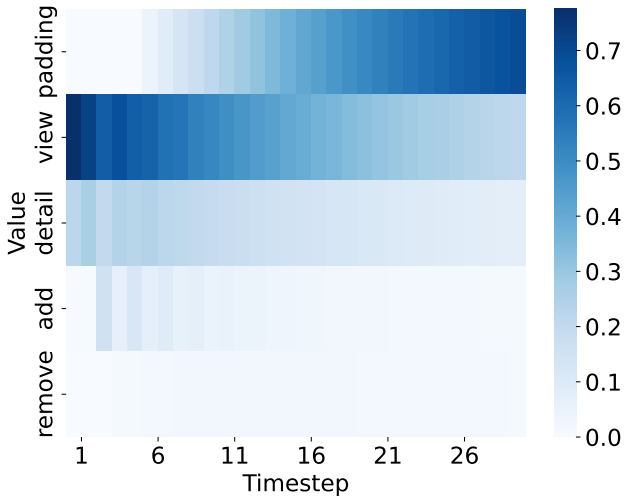
padding encoded as "0" is employed to standardize session lengths, which is necessary for the GRU and LSTM models.

The diagrams in Figure 3 illustrate profile matrices, which depict sequences of categorical events over time. These matrices reveal the frequency and probability of each action occurring at each time step.

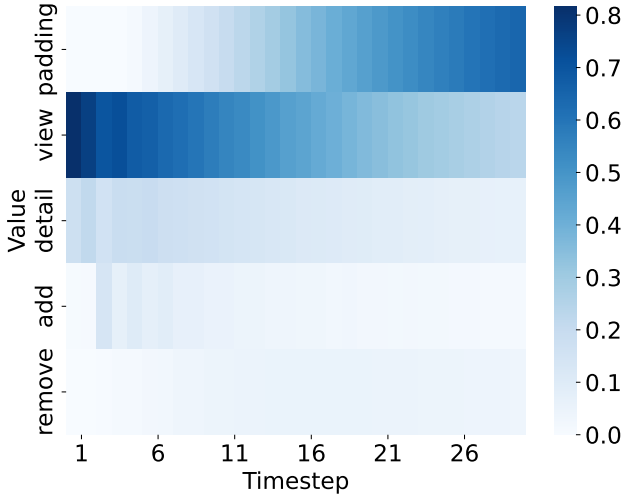
As shown in the profile matrices, the probabilities of various actions at each time step closely resemble each other. This similarity complements the results in Figure 2a and emphasizes the difficulty for machine learning models to distinguish between the two classes, which is consistent with the results of Toth et al. (2017) regarding the high similarity between shopping cart abandonment and purchase sequences. In the final phase of data processing, the coded events are consolidated into a list format, where each row represents a session characterized by the sequence of events. This method produces a sequential data representation as depicted in Figure 4.

The static features were selected based on characteristics that exhibit the greatest differences between cart abandonment and purchase sessions in the exploratory data analysis, enabling the model to more effectively distinguish between the two classes. After evaluating various hand-crafted features, three independent features were selected: “nevents," indicating the length of a session; "unique_products_viewed," showing how many products were viewed in one session; and "avg_dwell_time," measuring the average time between different clicks. The features are also in line with the selected static features of Wang et al. (2013) and Sheil et al. (2020). This approach allows for the combination of sequential and static features as input for the neural network, thereby creating a hybrid model.

The data was conclusively divided into three distinct subsets inspired by Requena et al. (2020) approach: 60% for the training set, 20% for the validation set, and 20% for the test set.



(a) Cart abandonment sessions



(b) Purchase sessions

Figure 3: Profile matrices illustrating the likelihood of each action occurring at each timestep.

3.3. Model Selection

As outlined in the related work, various machine learning approaches have been employed for analyzing customer behavior. The baseline model uses a 1st-order Markov Chain, where the probability depends solely on the previous clicked page. Despite the consideration of more history in higher-order Markov Chains, a first-order Markov Chain is preferred for its simplicity and straightforward implementation (Lakshminarayan, Kosuru and Hsu, 2016). To implement a first-order Markov Chain, it is essential to construct a transition matrix, as depicted in Figure 5, which represents the probabilities of transitioning from one state to another. Utilizing this approach, "purchase" can be effectively predicted as the subsequent action, serving as our target label.

The performance of this baseline model is compared with two deep learning approaches, namely LSTM and GRU. Both networks are types of recurrent neural network (RNN) architectures used primarily for sequence prediction tasks. While both share the common goal of addressing the vanishing gradient problem, they do so with different internal structures and mechanisms. On one hand, LSTMs feature a more complex architecture with a generally higher parameter count compared to GRUs, leading to slower training times. However, LSTMs tend to excel with larger datasets or longer sequences (Cahuantzi, Chen and Güttel, 2021). On the other hand, GRUs are regarded as more efficient, often outperforming LSTMs in scenarios with limited training data or when faster model training is necessary.

The architectures of the GRU and LSTM networks align with the configurations proposed by Sakar et al. (2019) and Requena et al. (2020), featuring one hidden layer consisting of approximately 30 neurons. Binary cross-entropy serves as the loss function and the networks are optimized using the Adam optimizer over ten epochs. The output, a predicted binary class probability, is produced by a dense layer equipped with a single node and sigmoid activation.

To implement the hybrid model, the sequential data is initially processed by an input layer. The output from the input layer is then concatenated with the static features and passed through a hidden dense layer, allowing the model to learn the interactions between the two data types. The remaining configurations of the hybrid models are kept identical to those of the sequential models to ensure performance comparability.

3.4. Hyperparameter Tuning

The Markov Chain model, serving as the baseline, is implemented in its simplest form and thus does not undergo hyperparameter tuning. Conversely, hyperparameter optimization for the LSTM and GRU networks, both with and without the application of SMOTE, is conducted using the GridSearch Tuner from the Keras neural network library (Chollet et al., 2015). The parameter grid includes learning rates of 0.01 and 0.001, batch sizes of 32, 64, and 128, and dropout rates of 0, 0.25, and 0.5, as similarly specified by Bigon et al. (2019).

Upon completing one iteration over each parameter combination, the validation loss is computed. The grid search results indicate that sequential models, particularly those not utilizing SMOTE, perform comparably across various hyperparameters, suggesting limited optimization potential with the existing architecture.

As shown in the contour plot in Figure 6 the inclusion of dropout layers has been observed to enhance performance exclusively in sequential models that incorporate SMOTE. Sequential models without addressing class imbalance and all hybrid models performed best without introducing dropout. Variations in batch sizes or learning rates do not significantly impact the validation loss of the models.

3.5. Model Comparison and Evaluation

As previously described in the context of hyperparameter tuning, the validation loss is calculated to assess model

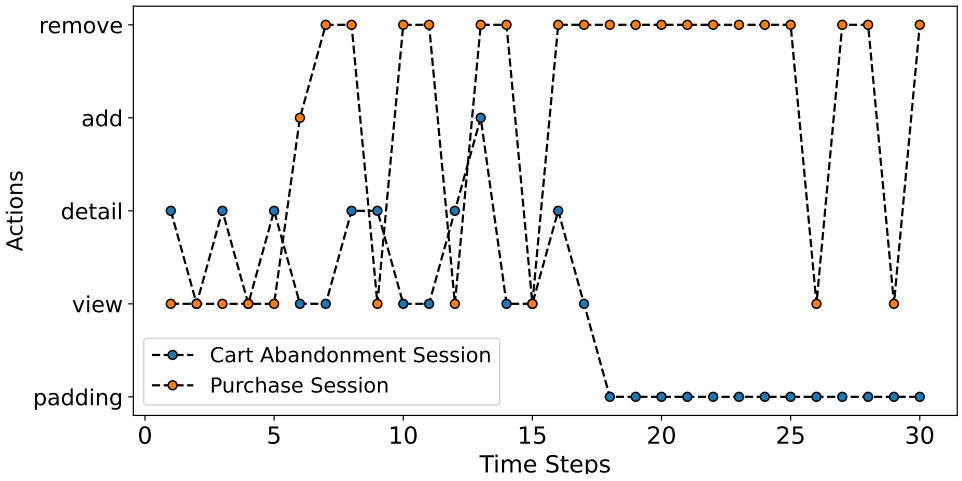


Figure 4: Visualization of clickstream trajectories inspired by Requena et al. (2020).

Figure 5: Transition matrix of actions within sessions.

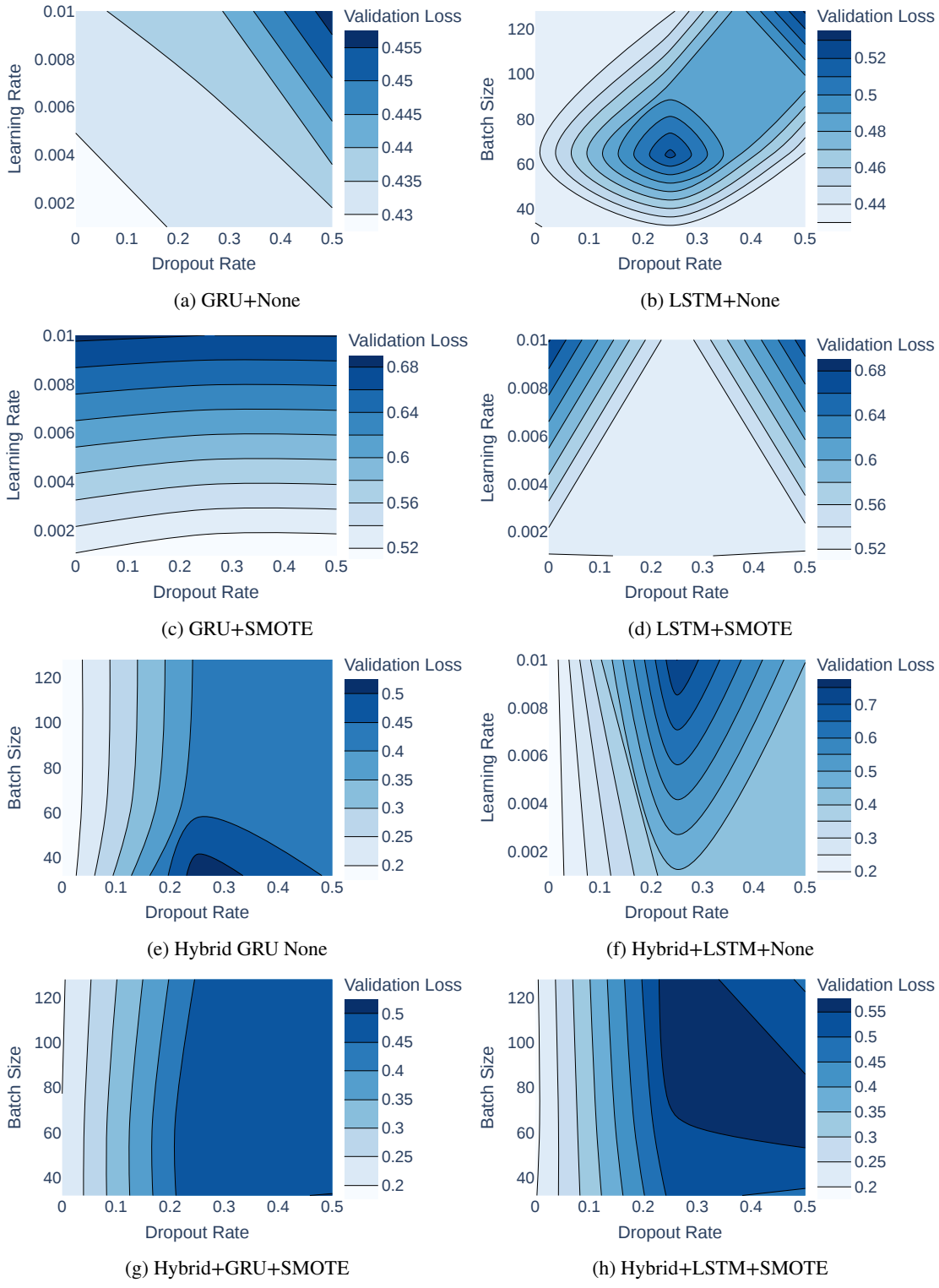
performance. However, the final evaluation of the optimal hyperparameters and the best model across various imbalanced and balanced datasets is determined using the F1 score. As the harmonic mean of these two metrics, the F1 score is particularly valuable for handling imbalanced datasets where accuracy alone might provide a skewed perspective (Wegier and Ksieniewicz, 2020). To determine the optimal F1 score, we perform a threshold analysis using binary search on the training dataset for the neural networks. This involves adjusting the probability threshold at which the network classifies an output as either a cart abandonment or a purchase. The results obtained from the optimal threshold are then compared to those where the threshold was not adjusted and kept at 0.5. As illustrated in Figure 7, the score varies significantly across different thresholds, especially for neural networks without addressing class imbalance, underscoring its dependency on precise threshold settings.

The final versions of all models are assessed using the test set. The Markov Chain model, not subject to hyperparameter tuning, does not require a split into training and validation sets. In contrast, the GRU and LSTM models are trained using both the training and validation sets. Additionally, SMOTE is applied to the training dataset to correct and compare for class imbalances.

An Oracle model, as proposed by Requena et al. (2020), is furthermore introduced as an upper benchmark for evaluating the sequential models. The Oracle model is designed with full knowledge of the true class distribution within the test set. It assigns to each session an empirical probability based on its likelihood of belonging to the purchase class. For instance, if a session appears five times in the test set and includes a purchase action in three of those instances, the model would assign a probability of 0.6 (3/5). Such an approach is particularly insightful for early prediction strategies, helping to address the significant challenge of managing duplicates within the dataset (Requena et al., 2020).

3.6. Error Analysis and Feature Importance

Concurrent with the model comparison, an error analysis is undertaken, incorporating the use of a confusion matrix to evaluate model performance across two classes. This visualization clearly indicates whether the model confuses one class for another (Ulitzsch, Ulitzsch, He and Lüdtke, 2022). Additionally, a calibration curve is constructed to examine how accurately the model's predicted probabilities match the actual outcomes. This involves plotting the mean predicted probability for each bin against the observed probability for that bin. Alongside, the Brier score is employed to assess both the calibration and the sharpness of the predictions, providing a comprehensive single-value measure. As both these analytical tools are invaluable for quantifying the impact of modifications to the model on its reliability, they are utilized



(a) GRU+None

(b) LSTM+None

(c) GRU+SMOTE

(d) LSTM+SMOTE

(e) Hybrid GRU None

(f) Hybrid+LSTM+None

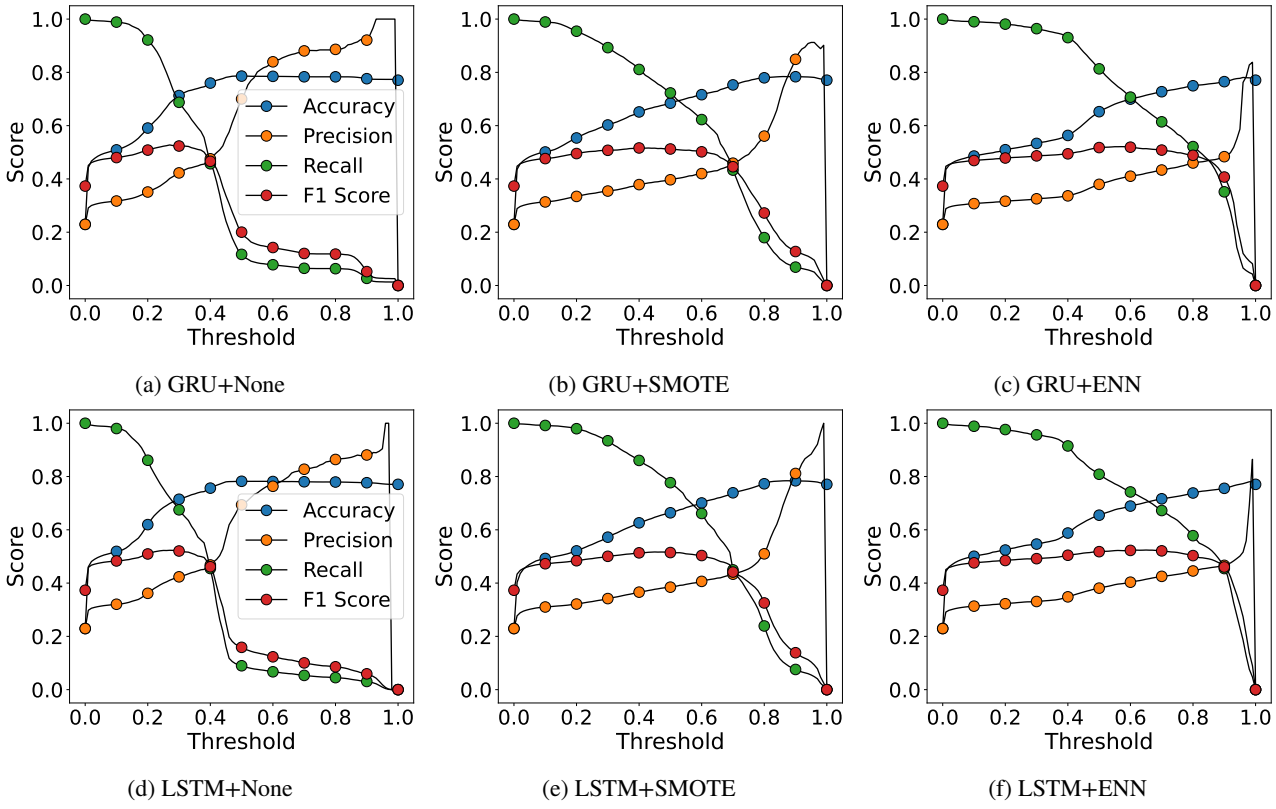
(g) Hybrid+GRU+SMOTE

(h) Hybrid+LSTM+SMOTE

Figure 6: Contour plots of key hyperparameters for each model.

in harmony with previously mentioned metrics (Park, Park, Kim and Park, 2021).

The SHAP method, is utilized to assess feature importance. Due to the high computational demands of SHAP, a subset of 500 sessions is analyzed to manage resources effectively. To further optimize the computational efficiency, a k-means clustering approach is applied directly to the SHAP values, allowing similar values to be grouped together, thereby simplifying the complexity of the dataset (Arslan, Lebichot, Allix, Veiber, Lefebvre, Boytsov, Goujon, Bissyande and Klein, 2022). This analysis includes a global ranking of the most significant time steps, and a bee



(a) GRU+None (b) GRU+SMOTE (c) GRU+ENN

(d) LSTM+None (e) LSTM+SMOTE (f) LSTM+ENN

Figure 7: Threshold analysis for sequential models.

swarm plot for visual representation but also incorporates local explanations to facilitate detailed comparisons between specific sessions.

4. Empirical Studies

In the following, the numerical results of the presented models under different class imbalance strategies are described and interpreted.

4.1. Comparison of sequential Models in predicting Cart Abandonment

Following the training phase and the determination of the optimal hyperparameters, the overall F1 score for every model was computed. Table 2 presents the performance metrics of all models across the training, validation, and test datasets after adjusting the probability threshold.

The performance metrics indicate that the Markov Chain, serving as the baseline model, is significantly outperformed by the two deep learning approaches with adjustment of the probability threshold. This is partly due to the effectiveness of the GRU and LSTM networks, but also due to the inability of the Markov Chain to adapt to modifications in the implementation and the corresponding sequential data, hindering its ability to make accurate predictions. Since the Markov Chain tries to predict the next action and the probabilities of transitioning to a buy action are particularly low, the Markov Chain is strongly influenced by the inherent class imbalance.

Table 2
Model F1 scores without addressing class imbalance

Adjustment	Dataset	Markov	GRU	LSTM	Oracle
Yes	Train	-	0.530	0.521	-
	Val	-	0.522	0.514	-
	Test	-	0.527	0.522	-
No	Train	0.173	0.205	0.158	-
	Val		0.205	0.158	-
	Test	0.176	0.200	0.158	0.850

As a result, only about 6% of sessions are predicted as buy sessions. As this greatly underestimates purchase sessions, it can be concluded that first-order Markov Chains that predict purchase as the next action are not suitable for this type of clickstream data. The results that the Markov Chains are outperformed by GRU and LSTM are thus in line with Requena et al. (2020) and Toth et al. (2017).

However, GRU and LSTM networks perform similarly to the Markov Chain when the probability threshold is set to the default value of 0.5. This illustrates how the inherent bias towards the majority class significantly affects the neural network. As highlighted in the exploratory data analysis, the

machine learning models struggle to differentiate between cart abandonment and purchase sessions because of their similar patterns. Consequently, the observed performance discrepancies can largely be attributed to noise from these ambiguous sessions within the clickstream data. The inability of the Oracle model to perfectly classify cart abandonment and purchase sessions underscores the complexity of the dataset. The presence of duplicate sessions that lead to inconsistent class labels, explains the relatively modest performance of the GRU and LSTM networks, even with adjusting the probability threshold.

In line with the findings of Zhou et al. (2023), both GRU and LSTM models exhibit comparable performance, with the GRU model slightly outperforming the LSTM. This superior performance of the GRU can be attributed to its use of shortened sequences within the dataset, optimized for early prediction. Given that LSTM's strengths lie in capturing long-term dependencies, its advantages are less pronounced in these shortened sessions. This observation aligns with the findings by Chung et al. (2014), who noted that GRU models adapt and effectively identify crucial features sooner due to their fewer parameters.

Given that sequences are truncated after 30 events, and the average length of buy sessions exceeds this limit, LSTM models can still be used effectively for general prediction tasks. However, GRUs are recommended for early detection tasks such as predicting cart abandonment, where timely intervention is crucial.

4.2. SMOTE's Effect on Prediction Performance

The second experiment assessed the impact of applying SMOTE to the training data on model performance. Table 3 presents a comprehensive comparison of the outcomes across all models, with and without the adjustment of the probability threshold, under three scenarios: without addressing class imbalance, with the application of SMOTE, and with the use of the ENN method.

Addressing class imbalance with SMOTE leads to a slight improvement in the Markov Chain baseline, yet its performance remains suboptimal. This highlights the inadequate suitability of the current implementation of the Markov Chain.

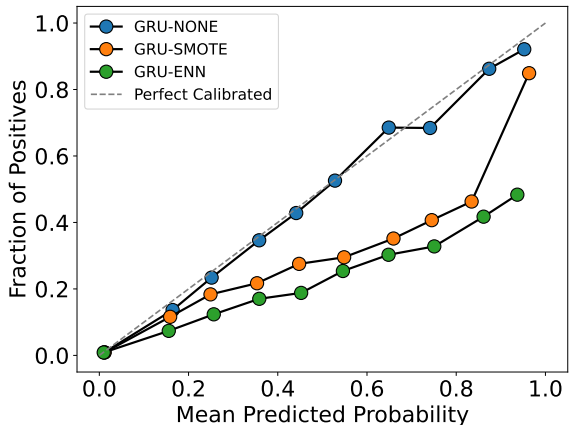
For the GRU and LSTM networks, both models exhibit in general similar performance characteristics. When using SMOTE or ENN, the performance on the test set of these neural networks can be significantly increased if the probability threshold is not adjusted. This improvement can be attributed to a more balanced distribution of the two target labels in the training set. Without addressing class imbalance and without adjusting the threshold value, neural networks significantly underestimate the number of purchase sessions, resulting in a very low F1 score. By using SMOTE or ENN, the number of predicted purchases increases substantially, thereby improving the performance of the neural networks. This demonstrates that the poor performance of the networks due to class imbalance can be effectively addressed by the two presented strategies for sequential data.

Table 3
F1 scores for sequential models under various imbalance strategies and probability threshold adjustments.

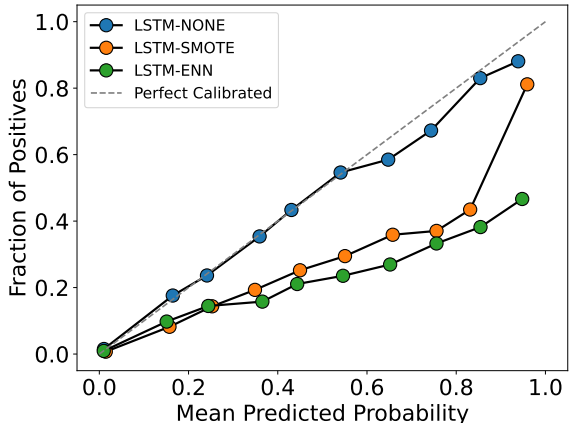
Strategy	Adjustment	Dataset	Markov	GRU	LSTM
None	Yes	Train	-	0.530	0.521
		Val	-	0.522	0.514
		Test	-	0.527	0.522
	No	Train	0.173	0.205	0.158
		Val		0.205	0.158
		Test	0.176	0.200	0.158
SMOTE	Yes	Train	-	0.789	0.789
		Val	-	0.515	0.516
		Test	-	0.514	0.516
	No	Train	0.185	0.782	0.789
		Val		0.510	0.515
		Test	0.187	0.512	0.514
ENN	Yes	Train	-	0.825	0.831
		Val	-	0.492	0.497
		Test	-	0.494	0.501
	No	Train	-	0.810	0.814
		Val	-	0.511	0.514
		Test	-	0.517	0.517

Additionally, a significant improvement in training performance is observed when using SMOTE or ENN compared to models trained without these techniques. However, considering the models with an adjusted probability threshold, this improvement is offset by a minor reduction in performance on the test set. This suggests that models using SMOTE or ENN cannot generalize effectively to unseen instances. The failed capability of generalization is also evident in Figure 8. The figure features calibration curves that illustrate the performance of probabilistic classifiers by plotting the mean predicted probability against the proportion of positive outcomes (actual results). Ideally, a predicted probability of 0.5 should align with about 50 % of the observations being positive. In this study, a positive observation is defined as a cart abandonment session.

The calibration curves indicate that both the GRU and LSTM models, when not addressing class imbalance, are reasonably well-calibrated. However, the curves for models using SMOTE and ENN on the training data show an underestimation of the actual likelihood of cart abandonment as they deviate more from the perfectly calibrated line. This suggests that the class imbalance strategies have

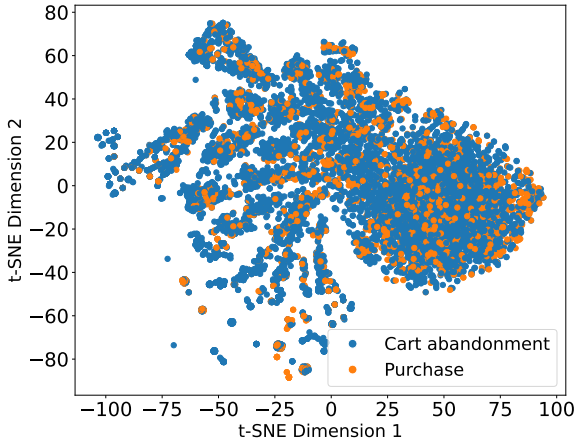


(a) GRU

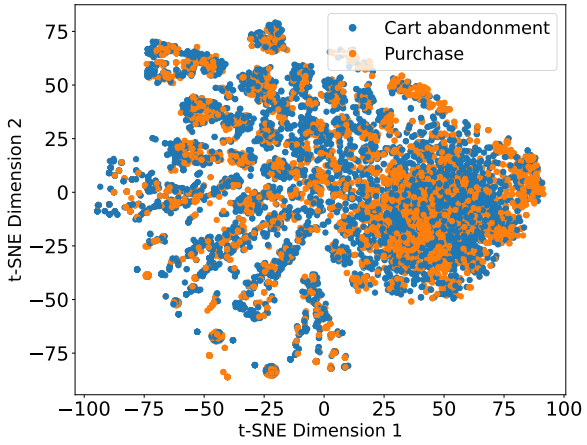


(b) LSTM

Figure 8: Calibration curves showing the reliability of predicted probabilities from classification models.



(a) without SMOTE



(b) with SMOTE

Figure 9: t-SNE visualizations for dimensionality reduction, highlighting the variance within the dataset.

introduced a stronger shift towards the probability of buying sessions, which is effective when the threshold is not adjusted. Nonetheless, if the threshold is adjusted, this results in a slight decrease in performance.

This reduced ability to distinguish between the two classes can also be explained using Figure 9.

The t-SNE visualization simplifies the complex sequential data into two dimensions, facilitating a deeper understanding of the relationships within the data. It preserves the pairwise distances between individual sessions. For instance, in the dataset at hand, the algorithm interprets a "View" action as number 1 and measures the distance to an "Add" action, which is encoded as number 3. From the visualization in the figure, it is apparent that the orange dots, representing purchase sessions, occupy the same space as the blue dots, which signify cart abandonment sessions.

The reason for the overlapping clusters in the t-SNE visualization can be traced back to the complexity of the sequential data, where similar or even identical sessions are classified into different categories. Despite the increase in buying sessions from using SMOTE, the two classes still fail to separate into distinct groups. This indicates that the synthetic samples introduced by SMOTE also contribute to noise within the dataset. Therefore, a significant challenge arises as the networks memorize the noise from oversampling, limiting their ability to generalize to broader patterns in the non-oversampled validation and test datasets. This issue is compounded by the introduction of dropout techniques, which, while improving performance on the test data with SMOTE, do not overcome the fundamental problem of noise retention.

Similarly, networks show the same behavior of overfitting with the undersampling approach since valuable information from cart abandonment sessions is discarded with this approach. The resulting discrepancy between the class

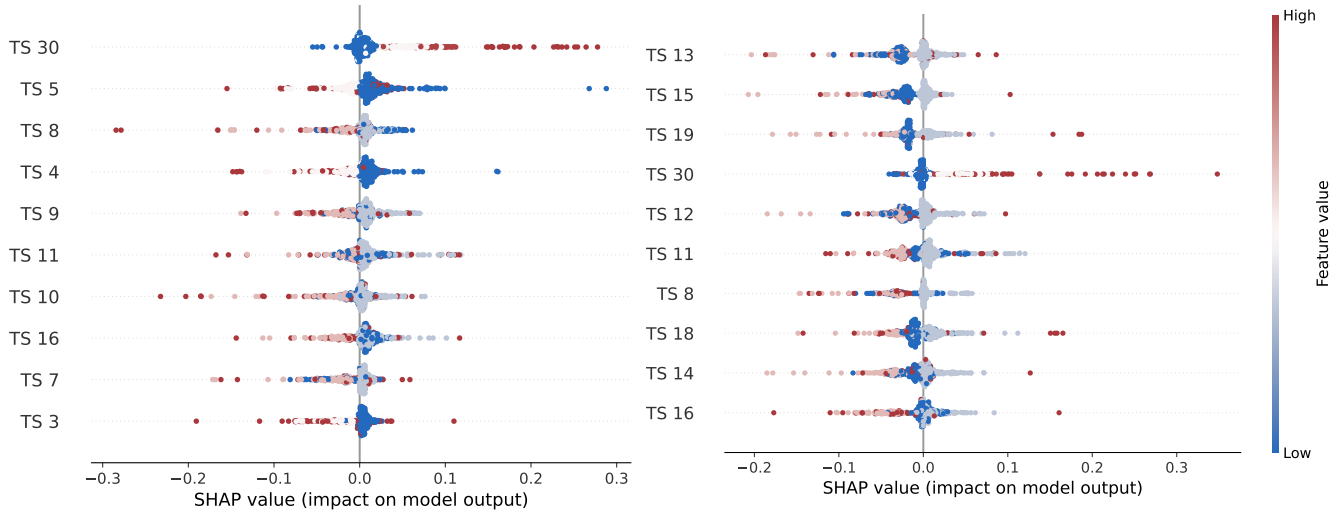


Figure 10: SHAP beeswarm plot for GRU (left) and LSTM (right) models without SMOTE

distributions in the training and test data introduces a significant bias in the probability estimates and results in overconfidence in the minority class of purchase sessions. This suggests that both k-nearest neighbor methods struggle to enhance the clarity of clustering if there is no clear discrimination before addressing class imbalance.

Therefore, adjusting the probability threshold without changing the underlying structure yields the best results, aligning with Johnson and Khoshgoftaar (2021). Additionally, the introduction of noise and reduced generalization through SMOTE is consistent with the findings of Teslenko et al. (2023) and Hooshyar et al. (2023). However, it cannot be concluded that the introduction of SMOTE generally leads to lower performance, as it improves performance when the probability threshold is not adjusted. This suggests that while SMOTE can mitigate the bias towards the majority class, its effectiveness is highly dependent on the specific data in the online commerce domain.

4.3. SMOTE's Effect on Feature Importance

After evaluating the performance of the machine learning models, we further explore how SMOTE affects feature importance determined by SHAP. To maintain consistency and manage computational demands, we used the same stratified random sample of 500 sessions from the test set across each model.

Figure 10 illustrates the feature importance as measured by SHAP for both the GRU and LSTM models. The features represented are different time steps within a session, abbreviated as "TS" and numbered from 1 to 30. In this figure, the features are ranked by their importance. Notably, time step 30 for instance had the most significant impact on the GRU model's predictions across the 500 analyzed sessions.

The feature values, encoded as numbers, represent different actions, as described in section 3.2. In the analysis, blue values, which correspond to frequent actions like "padding" and "view," exert less influence on the predictions compared to less frequent actions such as "add" and "remove." The features contribute both negatively and positively to the predictions, where negative SHAP values indicate a likelihood of cart abandonment sessions, and positive values suggest a model tendency towards purchase sessions.

Figure 11 presents the SHAP bee swarm plot for the GRU and LSTM models with SMOTE applied. As with the models not utilizing SMOTE, the rare actions continue to have the most significant impact on distinguishing between the two classes. However, in both models, the SHAP values that contribute to the prediction of cart abandonment now tend to be more extreme and cover a wider range. The contributions to the purchase sessions, on the other hand, tend to decrease. An explanation for this could be that after applying SMOTE, both neural networks continue to detect significant patterns indicative of cart abandonment. With a more balanced class distribution, the patterns in cart abandonment sessions become even more distinctive indicators during training for the neural networks. This results in higher weights for cart abandonment and correspondingly more pronounced negative SHAP values. Nevertheless, this effect cannot be observed in the contributions for the purchase sessions. Since the purchase sessions share the same feature space as the cart abandonment session, the synthetic instances of the purchase sessions introduce noisy purchase sessions through the application of SMOTE. As a result, the neural networks recognize the noisy and duplicate instances and do not adjust their perception of feature importance for predicting purchases.

The introduced bias towards buying sessions introduced during model training is also reflected in Figure 12.

The two SHAP force plots describe the same session predicted once with the GRU model without SMOTE (Figure 12a) and once with the GRU model with SMOTE (Figure 12b). The true label is cart abandonment, correctly predicted by the model without SMOTE and incorrectly predicted by the GRU model with SMOTE. The most critical insight from the charts is the base value, which signifies the average

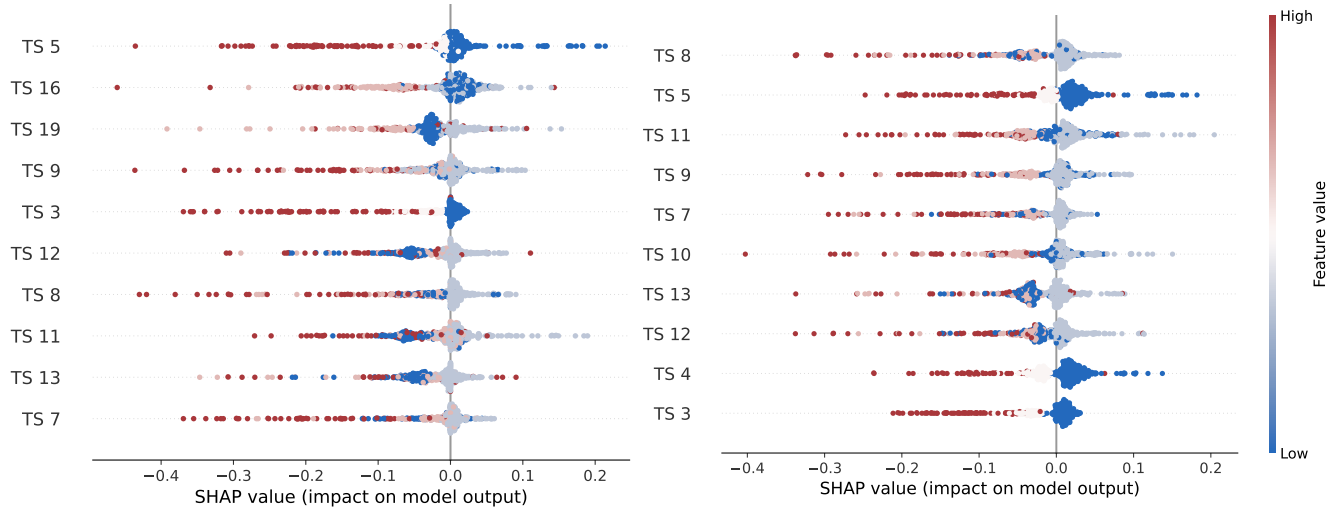
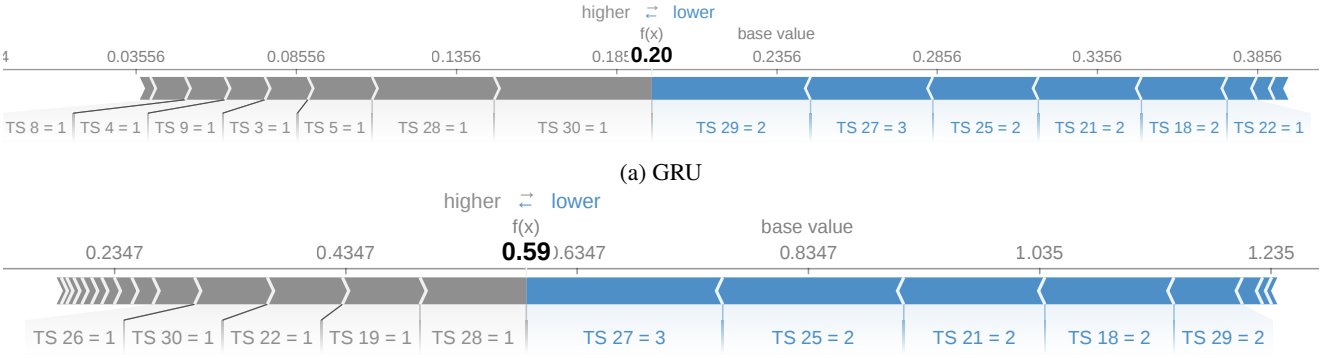


Figure 11: SHAP beeswarm plot for GRU (left) and LSTM (right) models with SMOTE.



(a) GRU

(b) GRU SMOTE

Figure 12: SHAP force plots for one individual session.

output of the model across the sample dataset utilized. In this instance, the base value indicates the average probability of a buying session, noticeably higher in the GRU model with SMOTE, demonstrating the introduced bias.

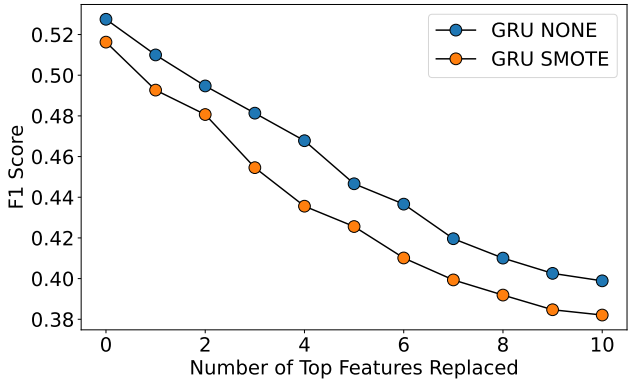
A further finding from the analysis of feature importance can be seen in the ranking of the timesteps. For the GRU model, the timestep remains roughly the same as seven out of ten of the most important timesteps are also included in the top ten features of their counterpart with SMOTE. For LSTM the ranked timesteps change more as there are only four common. This has also an impact on the performance of the neural networks as can be seen in the degradation curves in Figure 13. Performance degradation curves serve as indicators of the effectiveness of the feature importance measures as the most important features are substituted with their majority values for this timestep.

Both models initially achieved higher F1 scores compared to their counterparts trained with SMOTE. The performance degradation for the GRU models is similarly steep, indicating that the strategy for handling class imbalance does not significantly affect the assessment of feature importance. This implies that the GRU model is relatively robust to the noise introduced by SMOTE, as its performance remains consistent. However, the LSTM model without SMOTE exhibits a steeper slope, suggesting that the top ten ranked time steps are more accurate compared to the LSTM model with SMOTE. The reason for this may be due to the more complex architecture of the LSTM model, which makes it more susceptible to overfitting to noise. This offers an additional reason, beyond the slightly superior performance, for implementing GRU models for the early prediction of cart abandonment.

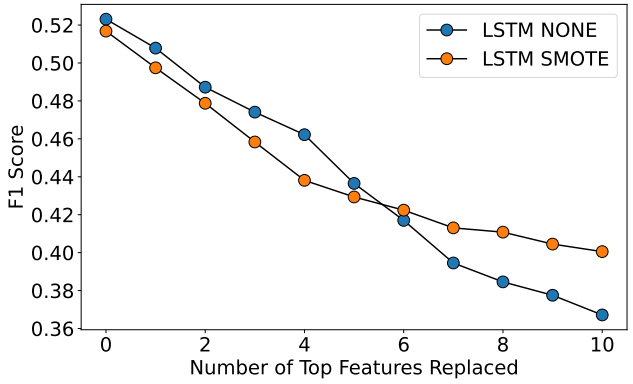
4.4. Impact of static Data Integration

The fourth experiment investigated the impact of including static features in the sequential models creating hybrid models. Table 4 presents the performance of the neural networks in terms of F1 score of the hybrid models under different conditions of adjusting the threshold and class imbalance strategies.

The results clearly indicate that the hybrid models significantly outperform all types of sequential models. The best-performing model on the test set is the GRU model, which, without considering class imbalance and adjusting



(a) GRU



(b) LSTM

Figure 13: Performance degradation curves.

the probability threshold, achieved an F1 score of 0.811. The LSTM model performs comparably, with an F1 score of 0.808. Both models exhibit similar performance across various data sets, indicating strong generalization capabilities. Therefore, incorporating the three hand-crafted static features alongside the sequential clickstream data significantly enhances performance and is in line with Sheil et al. (2020) and Mena et al. (2023). In contrast to the sequential models, the hybrid models without a class imbalance strategy and without adjusting the threshold perform as well as the models using class imbalance strategies. This indicates that the included static features are informative enough to mitigate the negative impact of inherent bias on performance. Hence, the hybrid model's performance without addressing class imbalance and without adjusting the threshold remains competitive with hybrid models using class imbalance strategies, as shown in Figure 14.

This suggests that the introduction of static features leads to better differentiation, thereby mitigating the inherent bias towards cart abandonment. Therefore, the hybrid neural networks exhibit greater robustness to varying threshold values compared to the sequential models, as their probabilities are more distinctly shifted towards 0 or 1. However, the trade-off for including static features is the more challenging implementation of early prediction and live recommendations for online shoppers. Adjusting the threshold value without

Table 4
F1 scores for hybrid GRU and LSTM models using sequential and static data under various imbalance strategies and probability threshold adjustments.

Strategy	Adjustment	Dataset	GRU	LSTM
None	Yes	Train	0.814	0.810
		Val	0.810	0.807
		Test	0.811	0.808
	No	Train	0.799	0.785
		Val	0.797	0.784
		Test	0.794	0.778
SMOTE	Yes	Train	0.908	0.918
		Val	0.783	0.791
		Test	0.781	0.791
	No	Train	0.906	0.916
		Val	0.793	0.800
		Test	0.789	0.798
ENN	Yes	Train	0.917	0.910
		Val	0.743	0.720
		Test	0.745	0.723
	No	Train	0.917	0.909
		Val	0.744	0.744
		Test	0.746	0.745

changing the underlying structure leads to the best results for both hybrid and sequential models and thus proves to be the best method for dealing with class imbalances in sequential data.

The SHAP visualization in Figure 15 shows how strongly the combined characteristics contribute to the predictions.

The figure shows that the length of the sessions with the specific actions at each timestep contributed the most to the prediction. Utilizing the additive property of SHAP, the contributions of all actions at each timestep within a session were aggregated into a single feature, "sequence_features," representing a single point on the scale. Lower feature values for the sequential features indicate buying sessions, while high feature values for "nevents" suggest that buying sessions involve more actions overall. This implies that purchase sessions are characterized by sequences with more frequent actions, such as viewing products, resulting in lower aggregated SHAP values. Additionally, customers who make a purchase tend to spend more time on the site, engaging in exploratory interactions while only viewing a few unique products.

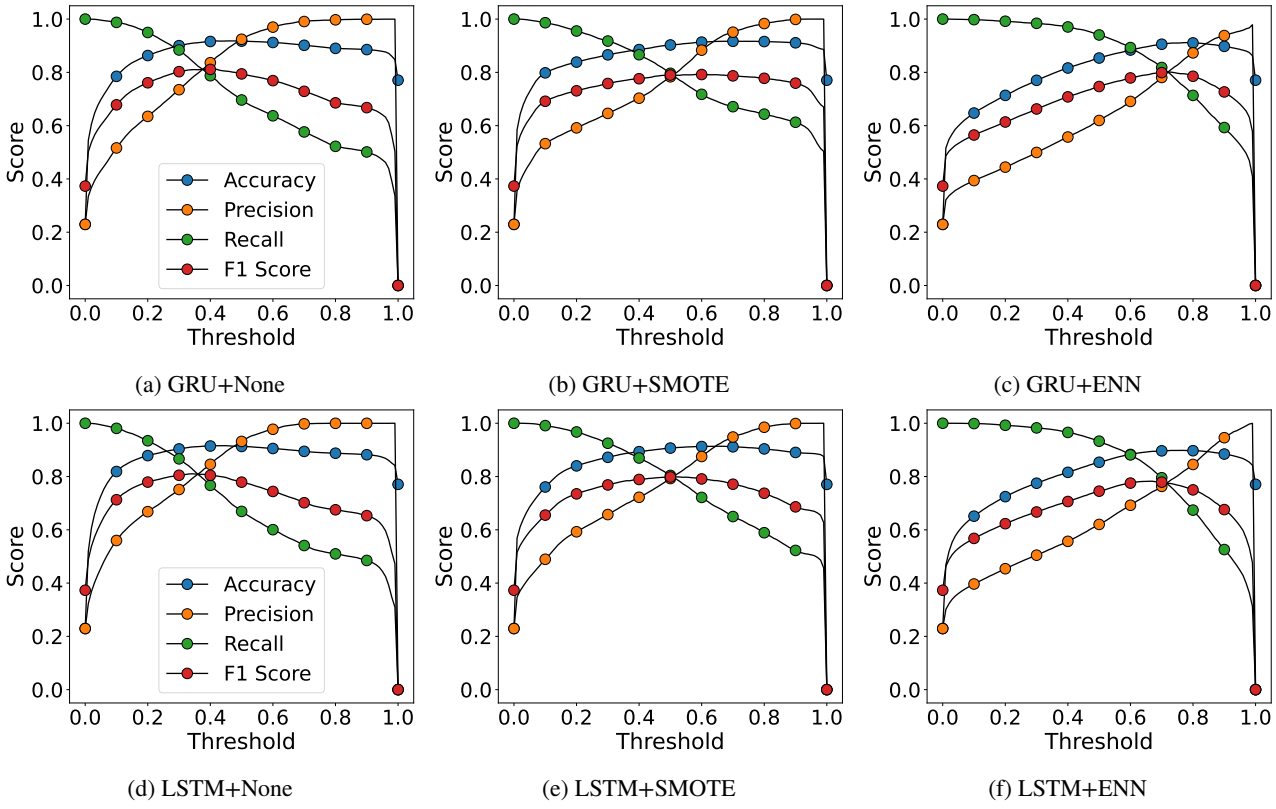


Figure 14: Threshold analysis for all hybrid models.

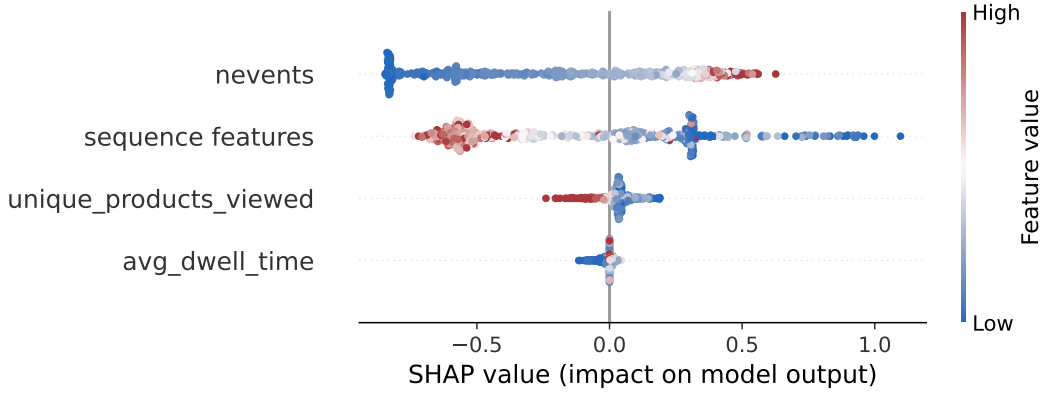


Figure 15: SHAP summary plot of the hybrid GRU model.

On the other hand, cart abandonment sessions involve fewer events and display higher aggregated SHAP values per session. This suggests that customers who abandon their cart tend to have shorter but more decisive interactions, such as quickly adding items to the cart and then removing them again. The higher number of products viewed, which pushes the SHAP values to the left, indicates that customers who browse a wide variety of products without making a decision are more likely to abandon their carts. These findings, combined with the observation that a lower average dwell time suggests cart abandonment, align with the results of the exploratory data analysis and (Guo and Agichtein, 2010).

5. Conclusion

This study offered a comprehensive overview of addressing the inherent class imbalance in clickstream data. Due to the dataset's similar characteristics for cart abandonment and purchase sessions, applying SMOTE introduced noise and a strong bias towards the minority class. This led to better performance without adjusting the threshold for sequential data but also reduced generalization for the validation and

test sets. By not altering the underlying data and adjusting the probability threshold, the best generalization and performance were achieved. Since this method is easy to implement and works effectively with more powerful hybrid neural network models, it is recommended for use in the online commerce domain. Through a feature importance analysis conducted by SHAP, it was also shown that purchase sessions tend to have more events and fewer products viewed than cart abandonment sessions.

Given that the dataset is one of the few representative data sources ensuring a certain degree of external validity, these findings contribute significantly to the scientific discourse and provide valuable insights for enhancing the user experience. As such, this study constitutes a step towards a clearer understanding of clickstream data, the behavior of neural networks, and the coherent class imbalance strategies.

5.1. Limitations

The study encountered several challenges due to the complexity of the data and computational constraints. The code was executed in the Google Colab environment, which offers a GPU backend and 52 GB of RAM. However, a limitation due to computational constraints was that only a single iteration was performed for each hyperparameter combination. Additionally, incorporating more hyperparameters could have potentially enhanced the performance and robustness of the models.

Associated with this, the representativeness of using 500 samples to derive feature importance is open to scrutiny. Given that SHAP is a computationally intensive method and resources were constrained, future research should consider increasing the sample size to improve the robustness and reliability of the findings.

Another limitation arises when considering the features of the hybrid models. The integration of "nevents" requires the session length to be determined only after completion, which precludes the possibility of making real-time recommendations and adjustments.

Overall, these limitations in computational resources and feature selection highlight the need for further academic research in this area.

5.2. Future Research

Future research could develop a more targeted approach for identifying purchase sessions which would enhance practical applications in online shopping. Given that misclassified shopping cart abandonments (false positives) are generally less detrimental than misclassified purchases (false negatives), refining model development and incorporating evaluation metrics such as the weighted F1 score could provide significant benefits for practical implementations.

This study focused on k-nearest-neighbor strategies for addressing class imbalance in sequential data. Exploring alternative strategies, such as employing generative adversarial networks or T-SMOTE, presents promising alternatives. Both approaches have been shown to enhance performance further, as demonstrated in Doan, Veira, Ray and Keng (2019) and Zhao, Luo, Qiao, Wang, Rajmohan, Lin and Zhang (2022).

From a societal perspective, analyzing a clickstream dataset that includes sociodemographic characteristics of customers can be crucial. Understanding the typical characteristics of a buyer, how class imbalance strategies introduce or remove instances, and how neural networks adopt this bias is valuable for considering the ethical implications in the online commerce domain. By enhancing the transparency and interpretability of these black-box models, we not only deepen our understanding but also contribute to making the digital world more fair and reliable for users.

References

Alejo, R., Sotoca, J.M., Valdovinos, R.M., Toribio, P., 2010. Edited Nearest Neighbor Rule for Improving Neural Networks Classifications. pp. 303–310. doi:10.1007/978-3-642-13278-0_39.

Alex, S.A., Jhanjhi, N., Humayun, M., Ibrahim, A.O., Abulfaraj, A.W., 2022. Deep lstm model for diabetes prediction with class balancing by smote. Electronics 11, 2737. doi:10.3390/electronics11172737.

Arslan, Y., Lebichot, B., Allix, K., Veiber, L., Lefebvre, C., Boytsov, A., Goujon, A., Bissyande, T., Klein, J., 2022. On the suitability of shap explanations for refining classifications, in: In Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022), SCITEPRESS - Science and Technology Publications. pp. 395–402. doi:10.5220/0010827700003116.

Bauer, J., Jannach, D., 2023. Hybrid session-aware recommendation with feature-based models. User Modeling and User-Adapted Interaction doi:10.1007/s11257-023-09379-6.

Baymard-Institute, 2024. Cart abandonment rate statistics 2024. URL: https://baymard.com/lists/cart-abandonment-rate.

Bigon, L., Cassani, G., Greco, C., Lacasa, L., Pavoni, M., Polonioli, A., Tagliabue, J., 2019. Prediction is very hard, especially about conversion. predicting user purchases from clickstream data in fashion e-commerce. ArXiv .

Bogina, V., Kuflik, T., Mokryn, O., 2016. Learning item temporal dynamics for predicting buying sessions, in: Proceedings of the 21st International Conference on Intelligent User Interfaces, ACM. pp. 251–255. doi:10.1145/2856767.2856781.

Cahuantzi, R., Chen, X., Güttel, S., 2021. A comparison of lstm and gru networks for learning symbolic sequences. ArXiv doi:10.1007/978-3-031-37963-5_53.

Chandramohan, T.N., Ravindran, B., 2018. A neural attention based approach for clickstream mining, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, Association for Computing Machinery, New York, NY, USA. p. 118–127. URL: https://doi.org/10.1145/3152494.3152505, doi:10.1145/3152494.3152505.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357. doi:10.1613/jair.953.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation .

Chollet, F., et al., 2015. Keras. https://keras.io.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555.

DelSole, T., 2000. A fundamental limitation of markov models. Journal of the Atmospheric Sciences 57, 2158–2168. doi:10.1175/1520-0469(2000)057<2158:AFLOMM>2.0.CO;2.

Doan, T., Veira, N., Ray, S., Keng, B., 2019. Generating realistic sequences of customer-level transactions for retail datasets. ArXiv .

Esmeli, R., Bader-El-Den, M., Abdullahi, H., 2021. Towards early purchase intention prediction in online session based retailing systems. Electronic Markets 31, 697–715. doi:10.1007/s12525-020-00448-x.

Flores, T., Flores, T.T., Winograd, T., 1986. Understanding Computers and Cognition: A New Foundation for Design. Intellect Books.

Guo, Q., Agichtein, E., 2010. Ready to buy or just browsing? detecting web searcher goals from interaction data, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA. p. 130–137. URL: https://doi.org/10.1145/1835449.1835473, doi:10.1145/1835449.1835473.

Hatt, T., Feuerriegel, S., 2020. Early detection of user exits from clickstream data: A markov modulated marked point process model, in: Proceedings of The Web Conference 2020, Association for Computing Machinery, New York, NY, USA. p. 1671–1681. URL: https://doi.org/10.1145/3366423.3380238, doi:10.1145/3366423.3380238.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hooshyar, D., Azevedo, R., Yang, Y., 2023. Augmenting deep neural networks with symbolic knowledge: Towards trustworthy and interpretable ai for education. arXiv:2311.00393.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. Journal of Big Data 6, 27. doi:10.1186/s40537-019-0192-5.

Johnson, J.M., Khoshgoftaar, T.M., 2021. Thresholding Strategies for Deep Learning with Highly Imbalanced Big Data. Springer Singapore, Singapore. pp. 199–227. URL: https://doi.org/10.1007/978-981-15-6759-9_9, doi:10.1007/978-981-15-6759-9_9.

Koehn, D., Lessmann, S., Schaal, M., 2020. Predicting online shopping behaviour from clickstream data using deep learning. Expert Systems with Applications 150, 113342. doi:10.1016/j.eswa.2020.113342.

Kukar-Kinney, M., Close, A.G., 2010. The determinants of consumers' online shopping cart abandonment. Journal of the Academy of Marketing Science 38, 240–250. doi:10.1007/s11747-009-0141-5.

Lakshminarayan, C., Kosuru, R., Hsu, M., 2016. Modeling complex clickstream data by stochastic models: Theory and methods, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. p. 879–884. URL: https://doi.org/10.1145/2872518.2891070, doi:10.1145/2872518.2891070.

Lipton, Z.C., Elkan, C., Naryanaswamy, B., 2014. Optimal thresholding of classifiers to maximize f1 measure, in: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 225–239.

Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles. ArXiv .

Melnykov, V., 2016. Model-based biclustering of clickstream data. Computational Statistics & Data Analysis 93, 31–45. doi:10.1016/j.csda.2014.09.016.

Mena, G., Coussement, K., Bock, K.W.D., Caigny, A.D., Lessmann, S., 2023. Exploiting time-varying rfm measures for customer churn prediction with deep neural networks. Annals of Operations Research doi:10.1007/s10479-023-05259-9.

Molnar, C., 2022. Interpretable Machine Learning. 2 ed. URL: https://christophm.github.io/interpretable-ml-book.

Nannini, L., Balayn, A., Smith, A.L., 2023. Explainability in ai policies: A critical review of communications, reports, regulations, and standards in the eu, us, and uk, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 1198–1212. URL: https://doi.org/10.1145/3593013.3594074, doi:10.1145/3593013.3594074.

Park, S.Y., Park, J.E., Kim, H., Park, S.H., 2021. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean Journal of Radiology 22, 1697. doi:10.3348/kjr.2021.0223.

Pedreschi, D., Miliou, I., 2020. Artificial intelligence (ai): New developments and innovations applied to e-commerce, EPRS: European Parliamentary Research Service.

Requena, B., Cassani, G., Tagliabue, J., Greco, C., Lacasa, L., 2020. Shopper intent prediction from clickstream e-commerce data with minimal browsing information. Scientific Reports 10, 16983. doi:10.1038/s41598-020-73622-y.

Roshan, K., Zafar, A., 2023. Using kernel shap xai method to optimize the network anomaly detection model. arXiv:2308.00074.

Sakar, C.O., Polat, S.O., Katircioglu, M., Kastro, Y., 2019. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks. Neural Computing and Applications 31, 6893–6908. doi:10.1007/s00521-018-3523-0.

Saranya, A., Subhashini, R., 2023. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. Decision Analytics Journal 7, 100230. doi:10.1016/j.dajour.2023.100230.

Sheil, H., Rana, O., Reilly, R., 2020. Understanding ecommerce clickstreams: A tale of two states, in: KDD Deep Learning Day. ACM(2018).

Statista, 2023. Fashion e-commerce worldwide - statistics & facts. URL: https://www.statista.com/topics/9288/fashion-e-commerce-worldwide/#topicOverview.

Tagliabue, J., Greco, C., Roy, J.F., Yu, B., Chia, P.J., Bianchi, F., Cassani, G., 2021. Sigir 2021 e-commerce workshop data challenge. arXiv:2104.09423.

Teslenko, D., Sorokina, A., Khovrat, A., Huliiev, N., Kyriy, V., 2023. Comparison of dataset oversampling algorithms and their applicability to the categorization problem. Innovative Technologies and Scientific Solutions for Industries , 161–171doi:10.30837/ITSSI.2023.24.161.

Thalpage, N., 2023. Unlocking the black box: Explainable artificial intelligence (xai) for trust and transparency in ai systems. Journal of Digital Art & Humanities 4, 31–36. doi:10.33847/2712-8148.4.1_4.

Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R., 2022. Explainable ai for time series classification: A review, taxonomy and research directions. IEEE Access 10, 100700–100724. doi:10.1109/ACCESS.2022.3207765. helllo.

Toth, A.R., Tan, L.H.S., Fabbrizio, G.D., Datta, A., 2017. Predicting shopping behavior with mixture of rnns, in: eCOM@SIGIR. URL: https://api.semanticscholar.org/CorpusID:59528391.

Ulitzsch, E., Ulitzsch, V., He, Q., Lüdtke, O., 2022. A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. Behavior Research Methods 55, 1392–1412. doi:10.3758/s13428-022-01844-1.

Villani, M., Lockhart, J., Magazzeni, D., 2022. Feature importance for time series data: Improving kernelshap. ArXiv abs/2210.02176. URL: https://api.semanticscholar.org/CorpusID:252715840.

Wang, A.X., Chukova, S.S., Nguyen, B.P., 2023. Synthetic minority oversampling using edited displacement-based k-nearest neighbors. Applied Soft Computing 148, 110895. doi:10.1016/j.asoc.2023.110895.

Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B.Y., 2013. You are how you click: clickstream analysis for sybil detection, in: Proceedings of the 22nd USENIX Conference on Security, USENIX Association, USA. p. 241–256.

Wegier, W., Ksieniewicz, P., 2020. Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms. Entropy 22, 849. doi:10.3390/e22080849.

Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics SMC-2, 408–421. doi:10.1109/TSMC.1972.4309137.

Zhao, P., Luo, C., Qiao, B., Wang, L., Rajmohan, S., Lin, Q., Zhang, D., 2022. T-smote: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification, in: Raedt, L.D. (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization. pp. 2406–2412. URL: https://doi.org/10.24963/ijcai.2022/334, doi:10.24963/ijcai.2022/334. main Track.

Zhou, D., Zhang, Y., Li, Y., Li, K., Zhao, B., Wang, M., Wang, N., 2023. Research on prediction method of uav heat seeking navigation control based on gru networks, in: Yan, L., Duan, H., Deng, Y. (Eds.), Advances in Guidance, Navigation and Control, Springer Nature Singapore, Singapore. pp. 3874–3881.

Zwitter, A., 2023. Handbook on the Politics and Governance of Big Data and Artificial Intelligence. Edward Elgar Publishing. doi:10.4337/

9781800887374.

CRediT authorship contribution statement

Fabian Waldmann: Software, Validation, Formal analysis, Investigation, Visualization, Writing - Original Draft. **Gonzalo Nápoles:** Conceptualization, Methodology, Visualization, Supervision, Writing - Review & Editing. **Yamisleydi Salgueiro:** Resources, Writing - Review & Editing.