Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Msc. Econometrics & Management Science

Using Embeddings for Item Similarity as Signal for Early Purchase Intention Prediction

Roos van Lookeren Campagne (471085)

Supervisor:	dr. C. Cavicchia
Second assessor:	dr. H. Akyuz
Date final version:	29th June 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Early purchase prediction of online shopping sessions has become crucial for e-commerce businesses aiming to personalise the consumer experience and boost conversion rates. This need is especially pronounced considering the smaller user bases due to privacy concerns and limited returning customers. This paper focuses on the early prediction of purchase intention using the pattern of similarity scores derived from the initial three interacted items in sessions as features. We create novel similarity scores from word2vec item recommendations based on the number of commonly recommended items and their relative relevancies based on the item's positions in recommendation lists. We construct effective classifiers and conduct extensive evaluations of the similarity scores across multiple datasets. The findings implicate that marketing strategists can leverage the similarity of the first item interaction with subsequent interactions as signal for purchase intention. This early signal can effectively be used to tailored marketing strategies to persuade customer purchases before the session ends. Notably, the Random Forest (RF) model outperforms other machine learning models in terms of F1 and Area Under Curve (AUC) scores and emphasises the importance of similarity features for accurate prediction.

Contents

Chapter 1

Introduction

Since online shopping continues to flourish and consumer data collection continues to improve, the ability to predict consumer behaviour has become increasingly important for e-commerce businesses seeking to personalise the shopping experience to drive sales and foster consumer loyalty. An inherent challenge faced by e-commerce platforms is the persistently low conversion rates, where the number of completed purchases lags significantly behind the number of initiated browsing sessions (Zhou, Mishra, Gligorijevic, Bhatia & Bhamidipati, 2019; Behera, Gunasekaran, Gupta, Kamboj & Bala, 2020). To address this challenge, even small increases in accurately identifying consumer browsing intent have been proven to be highly effective in improving the personalisation of the customer journey resulting in higher conversion rates (Blasco-Arcas, Lee, Kastanakis, Alcañiz & Reyes-Menendez, 2022; Zimmermann & Auinger, 2022). Other work further highlights the effectiveness of targeting customers during ongoing sessions, ensuring the opportunity to directly engage them with personalised marketing action (Esmeli, Bader-El-Den & Abdullahi, 2021, 2022; Alves Gomes, Meyes, Meisen & Meisen, 2022). This approach gains particular significance as an increasing number of customers prefer anonymous browsing and the majority of the started session are from non-returning customers (Tagliabue et al., 2021). Consequently, early purchase intention prediction has become a vital area of investigation for businesses striving to thrive in the rapidly evolving e-commerce sector.

On the one hand, a large number of abandoned sessions, i.e., sessions not ending in a purchase, could be attributed to informational or investigative intent, which is unlikely to convert to a purchase (Esmeli et al., 2021). On the other hand, a considerable proportion of abandoned sessions involve customers who exhibit strong purchase intent but do not complete a purchase due to factors such as a lack of offerings or misinterpretation of their behaviour. Targeted marketing strategies, such as personalised recommendations, limited-time discounts, ads, and follow-up emails, are shown to be very effective for converting customers with strong purchase intent to purchase (Chatterjee, McGinnis et al., 2010).

Existing research has primarily been focusing on predicting purchase intention based on extracted customer behaviour features from ended sessions, features such as average sessions duration, number of items clicked, and number of unique items clicked is shown to be highly correlated with the purchase intent (Mokryn, Bogina & Kuflik, 2019; Martínez, Schmuck, Pereverzyev Jr, Pirker & Haltmeier, 2020). Another example of a feature which can be extracted from the click-stream within sessions and contributes to an improved purchase prediction is the item similarity (Esmeli, Bader-El-Den & Abdullahi, 2020). Esmeli et al. (2020) incorporate the pattern of interacted item similarities within sessions as features in the prediction models, resulting in improved accuracy. The item similarity is derived from the commonly recommended items for a combination of item pairs in the first, middle, and last positions in sessions. The recommended items are provided by next item recommendation using an adapted word2vec Skip-gram model (Grbovic et al., 2015; Barkan & Koenigstein, 2016). Word2vec is originally developed to learn syntactic word associations in sentences presented in low-dimensional word embeddings (Mikolov, Sutskever, Chen, Corrado & Dean, 2013). Its adapted version learns representations of items from user interaction in sessions instead of representations of words in sentences. The underlying concept assumes that item interactions which frequently co-occur together exhibit a certain level of similarity to each other. This version has proven to be highly effective in recommendation systems (RS) for predicting relevant next items.

The research question addressed in this paper is whether the pattern of similarity of combinations of the first three item interactions within sessions, derived from word2vec item embeddings, can serve as an early prediction signal for purchase intention. These early predictive signals, which capture initial interest, enable marketing strategists to deliver targeted incentives and offers that persuade customers to purchase during ongoing sessions, thus reducing abandonment. For example, when a customer lands on the platform, a prediction of their purchase intent enables the triggering of targeted offers specifically for the items in which they have shown initial interest (Liu, Lee & Srinivasan, 2019), such as limited-time discounts or exclusive deals which can create a sense of urgency and incentivise the customer to complete the purchase. This targeted approach, at the very beginning of sessions, helps to maximise conversion rates.

By adopting and building upon the approach proposed by Esmeli et al. (2020), this research extends prior research in six ways. First, the item similarity scores between the first three items based on item embeddings have, to our knowledge, never been investigated in classification models. Unlike the research of Esmeli et al. (2020), the similarity scores derived from these interactions can serve as an early prediction signal for purchase intention. Second, we propose the use of "position-aware similarity", PAS, scores. These scores enhance the interpretation of item similarity scores by considering not only the number of common recommendations between item interaction pairs (Esmeli et al., 2020), but also incorporating the relative ranking of recommended items. For example, two item interactions with common recommendations at the top of the recommendation list can be considered more similar than recommendations that only overlap towards the end of the lists. The enriched understanding of the relationship between interactions is expected to improve the accuracy of purchase intent prediction. Third, we assess the performance effect of similarity scores when only considering the available information of the first three interactions. With this approach we aim to enhance our focus on early purchase prediction further, acknowledging the significance of the initial interactions within a session. Fourth, we extend the work of Esmeli et al. (2020) by tuning the classification model to build more effective classifiers, contributing to the overall performance improvement. Fifth, we conduct a thorough evaluation of our approach, considering multiple aspects such as performance improvement, feature importance, and feature effects. This comprehensive evaluation provides a deeper understanding of the effectiveness of our proposed method. Finally, the robustness of our approach is tested on two e-commerce datasets, ensuring the generalisability and reliability of our findings.

The structure of this paper is as follows. In Sect. 2, we provide a brief overview of related work on purchase prediction and discuss relevant models used in RS. Section 3 presents the datasets and explains the employed data processing techniques. The similarity scores, framework, and performance metrics are outlined in Sect. 4. We evaluate the item similarity scores and their performance effect on different classification models in Sect. 5. Thereafter, in Sect. 6 managerial implications and recommendations are provided. Last, we summarise our findings and suggest potential avenues for future research in Sect. 7.

Chapter 2

Related Work

In this section, we, first, provide an overview of existing research on purchase prediction models in Sect. 2.1. Thereafter, different methods to derive similar items from sessions are outlined in Sect. 2.2.

2.1 Purchase Prediction

Purchase prediction refers to the process of using binary classification models to estimate the likelihood of users making purchases (Martínez et al., 2020), which can help e-commerce strategists make appropriate marketing decisions. User representations are required as input for the purchase prediction models, which can be extracted from session data. Session data consist of a chronological sequence of users interaction in a session, e.g., click, add to cart and purchase events (Esmeli et al., 2021). Generally, for each interaction at least the event time, interacted item, and an user or a session identifier is logged. Extensive research has been devoted to identifying user behavioural features that can be extracted from session data (Sheil, Rana & Reilly, 2018; Mokryn et al., 2019; Martínez et al., 2020; Esmeli et al., 2020, 2021; Alves Gomes et al., 2022).

For example, Mokryn et al. (2019) extract the product trendiness from click-stream data. They include this feature to the temporal feature set to predict the purchase intention on different machine learning models. Temporal features capture the temporal aspect of user actions, example are the month, day and hour of the started session. Esmeli et al. (2020) investigate the effect of using the pattern of product similarity scores instead of product trendiness as feature. This product similarity score is based on the number of common recommendations for pair of items in the first, middle and last position in a session. To derive recommendations Esmeli et al. (2020) use an adapted word2vec recommendation model. In computational experiments, they test on a Random Forest (RF), Bagging and Decision Tree (DT) classification model. Results show that each model reacts differently when the item's position changes. Nevertheless, this research shows that when item similarities are used as a feature next to temporal features and click-stream features the accuracy of each model improves in terms of F1 score. Wen, Lin and Liu (2023) combine item popularity and trendiness based on various types of user feedback, also including product comments and reviews to better capture the dynamic changes in consumer preferences. They show an improvement in the prediction accuracy in terms of AUC.

These studies solely rely on historical actions from completed sessions to make purchase predictions. However, given the dynamic nature of user intentions, the prevalence of non-returning users, and the growing privacy concerns that restrict access to user data, there is an increasing demand for real-time personalization. In this context, it requires classification models to predict accurately at earliest stage of the user session and with minimal data available (Lin, Milic-Frayling, Zhou & Ch'ng, 2019; Esmeli et al., 2021; Tagliabue et al., 2021; Esmeli et al., 2022; Alves Gomes et al., 2022). Lin et al. (2019) show that Logistic Regression (LR) is robust and effective for purchase prediction based on user interaction pattern, achieving AUC score of 0.85 after considering only 50% of an ongoing session. Esmeli et al. (2021) create dynamical features and use them on DT, RF, Bagging, K-nearest neighbours (KNN), and Naive Bayes (NB) for purchase prediction in ongoing sessions. Esmeli et al. (2022) even use session information that is only available upon user arrival at e-commerce platforms. They use besides temporal features, contextual features such as location and device type.

An alternative approach to manually selecting features involves learning features directly from the sequence of user interactions. Recently, Alves Gomes et al. (2022) propose the use of word2vec embeddings to create user representations from click-stream data. The embeddings are used as input to the prediction models. The experiment shows that just-in-time purchase prediction is possible and that using word2vec representations on long short-term memory networks outperforms baseline models based on features from (Esmeli et al., 2021). According to Alves Gomes et al. (2022), one possible explanation for the effectiveness is that the embeddings are able to encode information that is challenging to be captured by manually created features. However, a notable limitation in such automated feature learning approaches is the difficulty of assessing the importance of features driving purchase intent and developing effective countermeasures based on them.

Given our objective of supporting e-commerce strategists in optimising conversion rates through the prediction of purchase intent, relying solely on item representations as predictive features is not sufficient. Nevertheless, we acknowledge the efficacy of embeddings and their potential to extract implicit consumer behaviour. Additionally, we recognise the importance of early purchase prediction. Consequently, we build upon the research conducted by Esmeli et al. (2020). One of the main contributions of this study is our focus on solely the first three items in a session when calculating item similarity. This guides e-commerce strategists to provide targeted content and recommendations at the very start of sessions to better engage users and increase the likelihood of conversion.

2.2 Session Based Recommendations Systems

Session Based Recommendation Systems (SBRS) involve providing recommendations based on a sequence of user interactions within the current session and previous anonymous sessions. Especially, in the field of e-commerce, a growing interest is observed in SBRS due to its ability to provide recommendations even to anonymous or new users. Initially, item-item similarity-based methods relying on simple heuristics between the target item and the last item were used. As soon as researchers release the analogy between words in sentences and user interactions in sessions, Natural Language Processing (NLP) techniques were leveraged and adopted for SBRS.

Among the first adopted NLP models to SBRS is Prod2Vec (Grbovic et al., 2015). It learns low-dimensional item embeddings using word2vec skip-gram model. It employs a two-layer neural network for self-supervised learning of word representations by creating word embeddings such that words sharing similar contexts within sentences appear close to each other in the vector space. This model is known to be efficient and effective in providing relevant and diverse item recommendations using the top N most similar items to the selected items (Grbovic et al., 2015; Barkan & Koenigstein, 2016).

A less widely adopted NLP model for SBRS is Global Vectors, GloVe (Pennington, Socher & Manning, 2014). It seeks to capture the local context of words like word2vec as well as global co-occurrence patterns. This is done by factorising a word-context co-occurrence matrix based on occurrences of words within a specific context window. Then word representations are learned such that the dot product of the word embeddings equals the logarithm of the probability of co-occurrence. The factorisation can be computationally expensive for a very large corpus.

A deep-learning NLP-inspired approach is the use of Recurrent Neural Networks (RNN), which can capture the sequential order of item interactions within sessions (Hidasi, Karatzoglou, Baltrunas & Tikk, 2016). However, its sequential nature restricts parallelisation, which limits scalability and efficiency, while fast training of scalable models is crucial for e-commerce businesses of which the item corpus is often very large (Ludewig, Mauro, Latifi & Jannach, 2019). Furthermore, Jannach, Ludewig and Lerche (2017) show in experiments that simpler algorithms provide more accurate recommendations and that RNN models are often not reproducible.

More recently, Transformer based architectures are shown to be an efficient alternative to the RNN-based models (Vaswani et al., 2017). These architectures provide efficient parallel training, scale well with training data and are more effective at modelling long sequences. BERT4Rec (Bianchi, Yu & Tagliabue, 2021) adopt Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee & Toutanova, 2019), for SBRS. BERT uses the entire sentence as input and creates unique embeddings for each word occurrence, based on their position and full context within the sentence. To leverage the full context the Masked Language Modelling (MLM) technique is used, which requires the model to predict the masked words during training. Bianchi et al. (2021) adopted the MLM technique also for evaluation, such that the last item in the session is masked for next item recommendation.

Although the great success of BERT4Rec to improve the accuracy, personalisation, and contextual understanding in RS, it may not be suitable for our research due to specific limitations. First, BERT4Rec is designed to effectively model long sequences, which is not optimal for our e-commerce data consisting of short sessions. This can result in significant masked items during MLM such that there is barely any context left. Moreover, in the case where we want to compare recommendation lists of multiple items, these items need to be masked and almost no context will be left. Finally, unlike word2vec, BERT4Rec requires retraining for every new session, as it created unique item embeddings.

Overall, given the nature of our research and the limitations mentioned above, we find that word2vec is the most suitable choice for our purposes. It is known for its efficiency, scalability, and effectiveness in providing similar item recommendations. It is worth noting that our approach is designed with the flexibility to apply other SBRS models in future research.

Chapter 3

Data

In this research, we use two session datasets. The first one is the open-sourced RetailRocket[1] e-commerce dataset, which provides anonymous click-stream data over 4.5 months. It is widely used for testing purchase prediction models due to its real-world applicability, various types of user interactions and its established role as a benchmark dataset (Esmeli et al., 2020; Sheil et al., 2018). The second dataset is a closed data from an online furniture marketplace collected by Google Analytics over almost 3 months. We will refer to this dataset as the Closed dataset. Table 3.1 shows some details of each dataset before processing.

The RetailRocket dataset only provides user identifiers, which are inherently different from session identifiers, as an user may have multiple sessions in a given time period. To create user sessions, we apply the session creation approach of Google Analytics. This involves creating new session identifiers for user interaction sequences whenever the time between sequential interactions exceeds 30 minutes. We remove accidental and short clicks for both datasets. Sessions that last less than 2 seconds are discarded, as these sessions do not provide adequate information about user intent. Furthermore, to compute item similarities for the first three item interactions, we only keep sessions with at least three interactions.

Word2vec may struggle to generate accurate embeddings for rarely occurring items. Additionally, the method is only able to provide item recommendations for items that have been encountered during training. To mitigate this so-called cold-start problem, we remove items that occur less than twice in the dataset.

Tab. 3.1 and Tab. 3.2 summarise relevant details of each dataset before and after data pre-processing. It shows that the vast majority of sessions in both datasets do not end in a purchase. In the RetailRocket dataset, only 7.16% comprises purchase sessions, while the Closed dataset has an even smaller proportion, 0.95%. This substantial class imbalance could affect the prediction performance of the classification models (Berry & Linoff, 2004). In addition, Tab. 3.3 shows that the Closed dataset has relatively more unique item interactions among the first three items in sessions than the RetailRocket dataset. More specifically, the Closed dataset also has relatively fewer sessions where the first three item interactions are identical. This could affect the similarity scores and thus the prediction performance. Hence, the differences between the datasets provide a valuable opportunity to test the generalisation capabilities of the models.

[1] https://www.kaggle.com/datasets/RetailRocket/ecommerce-dataset

Table 3.1: Details of RetailRocket and Closed dataset before pre-processing

dataset	# unique items	# interactions	# view	# addtocart	# transactions
RetailRocket	235,061	2,756,101	2,664,312	69,332	22,457
Closed	64,365	3,842,745	3,816,970	21,313	4,462

Table 3.2: Details of RetailRocket and Closed dataset after pre-processing

dataset	# unique items	# interactions	# unique sessions	# purchase sessions	# average session length
RetailRocket	99,389	954,638	171,736	12,294	5.56
Closed	61,573	2,437,151	345,915	3,276	7.05

Table 3.3: Distribution of the number of unique items for the first three interactions in non-purchase and purchase sessions

dataset	# unique items	% in non-purchase session	% in purchase session
RetailRocket	1	16%	56%
	2	41%	28%
	3	43%	16%
Closed	1	8%	60%
	2	22%	23%
	3	70%	17%

Chapter 4

Methodology

This section provides a comprehensive overview of the steps involved in our proposed framework, as depicted in Fig. 4.1. Building upon the framework developed by Esmeli et al. (2020), our approach incorporates several key components. First, we create item embeddings from sessions by leveraging a word2vec model, which captures the similarity based on the co-occurrences of items within sessions. Detailed information regarding the model's configuration can be found in Sect. 4.1. Next, we define the similarity scores utilised as features in Sect. 4.2 and additional extracted features are outlined in Sect. 4.3. In Sect. 4.4, we introduce the classification models employed for purchase intention prediction. Last, in Sect. 4.5, we outline the evaluation metrics employed to assess the performance and effectiveness of our framework in predicting purchase intentions.

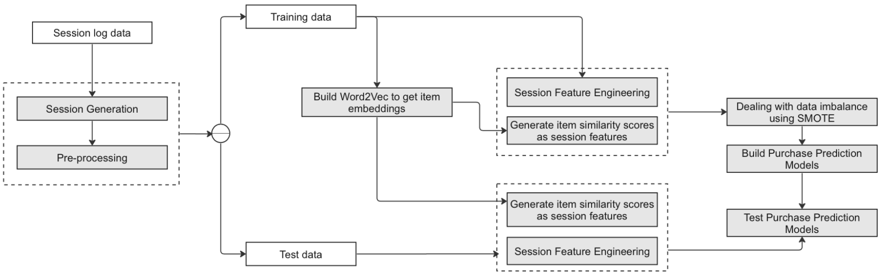
Session log data
Session Generation
Pre-processing
Training data
Test data
Build Word2Vec to get item embeddings
Session Feature Engineering
Generate item similarity scores as session features
Generate item similarity scores as session features
Session Feature Engineering
Dealing with data imbalance using SMOTE
Build Purchase Prediction Models
Test Purchase Prediction Models

Figure 4.1: Framework for session feature generation to purchase intention prediction

4.1 Hyperparameters Word2vec Model

We implement a word2vec skip-Gram model following the approach of Bader-El-Den, Teitei and Perry (2018) using the Gensim library (Řehřek & Sojka, 2010). As input, we create sequences of item ids for each session. We set the model hyperparameters according to Esmeli et al. (2020) as follows: *epochs* = 30, *vector dimension* = 100, *window*= 3, *minimal count* = 1, and other hyperparameters remain as shown by default. The model is trained only on the training set to avoid bias in the test dataset.

4.2 Similarity Score

In this section, we begin by describing how similarity scores are derived from item recommendations in the work of Esmeli et al. (2020). This serves as the foundation for our subsequent discussions on limitations, potential improvements, and the introduction of novel similarity scores. Before delving into these aspects, let us establish some notions. We denote $\mathcal{S}$ as the set of S sessions, where each session $s = (i_p, \ldots, i_P) \in \mathcal{S}$ is defined as an ordered sequence of P interacted item ids.

4.2.1 Jaccard Similarity

Esmeli et al. (2020) create item recommendations from the word2vec by selecting the N nearest neighbours in the vector space to the selected item using the cosine similarity. Then, the similarity score between item interaction i_p and i_q within session s, such that $p \neq q$ is computed by the Jaccard Similarity (JS):

$$JS(i_p, i_q) = \frac{\left|L_{i_p} \cap L_{i_q}\right|}{\left|L_{i_p} \cup L_{i_q}\right|}, \tag{4.1}$$

where the recommendation list for i_p is denoted by $L_p = R\,(i_p)$ and $L_q = R\,(i_q)$ denotes the recommendation list for i_q. We choose to set the length of the recommendation lists equal to N = 100. This ensures that items that are less relevant to the selected item are disregarded, and more precise similarity calculations are made (Esmeli et al., 2020).

JS has the limitation that it solely considers the number of common item recommendations between two given items, disregarding other important factors. For example, the ranking of the recommended items may influence the perceived similarity between two items. Overlapping items at the top of the recommendation list may be considered more similar than items that only overlap towards the end of the list. JS fails to recognise such ranking effects. Therefore, it is crucial to explore alternative similarity measures that take into account factors like item recommendation relevancy and ranking. By addressing these limitations, we can potentially enhance the accuracy and effectiveness of similarity scores used as features in purchase prediction models.

4.2.2 Position-Aware Similarity

We propose a novel similarity score called “position-aware similarity”, PAS, to address the limitations of the JS score. PAS score takes into account the relevancy of commonly recommended items based on the position within the given recommendation list relative to the item positions of the target item recommendation list. The underlying intuition is that common recommendations that are highly ranked in both recommendation lists should be accorded more weight, reflecting their presumed greater relevance and similarity. On the other hand, items that are less similar and appear lower in the lists should be assigned less weight, indicating their comparatively lower similarity. This approach captures both the number of common recommendations and the relevancy of common recommendations, ultimately aiming to provide a more comprehensive and effective approach for measuring the similarity between items.

We calculate PAS as follows. First, we calculate the position score (PS) for i_p compared to

i_q by comparing the positions of their item recommendations as follows

$$\mathrm{PS}\left(L_q|L_p\right) = \sum_{r_{i_p}, r_{i_q} \in L_p \cap L_q} \frac{1 + N - r_{i_{p(n)}}}{\log_2(r_{i_{q(n)}} + 1)}, \tag{4.2}$$

where r_{i_p} is an item in the recommendation list L_p and commonly shared with L_q and r_{i_q} is an item in the recommendation list L_q. The numerator represents the relevancy based on the position of the common recommended items in L_p, $r_{i_{p(n)}} \in L_p \cap L_q$. The n'th position of an item in r_{i_p} is presented by $r_{i_{p(n)}}$. For the target item, i_q, with a recommended item r_{i_q} in recommendation list L_q applies an analogous notion. The value of n equals 1 if an item is in the first position of the recommendation list and equals N when it is the last recommended item. Resulting in the numerator being equal to N if an item is in the first position and 1 if an item is at the last position of L_q. The denominator discounts the relevance of the given recommendations by giving higher weight to items which are at the top of the target list, L_q. It reflects the diminishing importance of the relevance of the given item recommendations as the position of these items increases in the target list.

Second, compare the positions of the commonly recommended items to an ideal list. Ideally, all items in L_p should also be in L_q. Since all recommendation lists have a length of N, we could write the ideal position similarity (IPS) as,

$$\mathrm{IPS}\left(L_q\right) = \sum_{r_{i_q} \in L_q} \frac{1 + N - r_{i_{q(n)}}}{\log_2(r_{i_{q(n)}} + 1)} = \sum_{n=1}^{N} \frac{1 + N - n}{\log_2(n + 1)}, \tag{4.3}$$

where we sum over all n positions. Now, we can obtain a normalised position score (NPS) for the recommended items in i_p by

$$\mathrm{NPS}\left(L_q|L_p\right) = \frac{\mathrm{PS}\left(L_q|L_p\right)}{\mathrm{IPS}\left(L_q\right)}. \tag{4.4}$$

NPS ranges between 0 and 1, such that a score of 1 is assigned to two items which share exact same items in their recommendation lists and a score of 0 is assigned when there are no common recommendations. The more items are in common at the top of the recommendation list the higher the value of NPS. Since we compare two recommendation lists with each other, we should consider both L_p and L_q as target lists. The final similarity score (PAS) is then computed as the average of $\mathrm{NPS}\left(L_q|L_p\right)$ and $\mathrm{NPS}\left(L_p|L_q\right)$,

$$\mathrm{PAS}\left(L_q, L_p\right) = \frac{1}{2}(\mathrm{NPS}\left(L_q|L_p\right) + \mathrm{NPS}\left(L_p|L_q\right)). \tag{4.5}$$

As an alternative to the position of item recommendations as relevancy score, we could also consider the cosine similarity between the given item and its recommendations as a measure for relevancy. The cosine similarity score indicates how close the given item and its recommended item are in the embedding space. Cosine similarity scores range between 0 and 1, where 1 represents the highest similarity and 0 indicates no similarity between items. Since we always consider the top N closest items to the given item in the embedding space as recommendations it could be that the first items in the list are relatively closer to each other in the embedding space and thus exhibit higher similarity than the last items in the list. We also experiment with whether the cosine similarity can be used as relevancy score. To this end, we adjust PS to the cosine position similarity score (CPS),

$$\mathrm{CPS}\left(L_q|i_p, L_p\right) = \sum_{r_{i_p}, r_{i_q} \in L_p \cap L_q} \frac{\cos\left(v_{i_p}, v_{r_{i_p}}\right)}{\log_2(r_{i_{q(n)}} + 1)}, \tag{4.6}$$

where $\cos(v_{i_p}, v_{r_{i_p}})$ denotes the cosine similarity between the vector representation of item i_p, denoted by v_{i_p}, and the vector representation of a common recommended item r_{i_q}, denoted by $v_{r_{i_q}}$. Ideally, we would have all items in L_p also in L_q with the exact same ranking and relevance, in terms of cosine similarities, therefore the ideal cosine position similarity (ICPS) score can be calculated as

$$\mathrm{ICPS}\left(i_q, L_q\right) = \sum_{r_{i_q} \in L_q} \frac{\cos\left(v_{i_q}, v_{r_{i_q}}\right)}{\log_2(r_{i_{q(n)}} + 1)}. \tag{4.7}$$

After normalising, we obtain the cosine position-aware similarity score (CPAS) in a similar way as PAS by taken taking the average of $\mathrm{NCPS}\left(L_q|i_p, L_p\right)$ and $\mathrm{NCPS}\left(L_p|i_q, L_q\right)$.

4.3 Session Feature Extraction

Besides similarity scores, we extract important features from the session that are shown to be relevant for purchase intention prediction. Various works highlight the importance of temporal features, which are available at the direct start of the session, to better purchase prediction (Esmeli et al., 2020, 2021; Alves Gomes et al., 2022). Moreover, duration is shown to be one of the most important features (Esmeli et al., 2021). The vast majority of the work in this field also uses the number of interactions and the number of unique items as features, which can be extracted from the click-stream.

For comparability with Esmeli et al. (2020), we use the same feature set. That means we extract the following click-stream features: *total number of interacted items* within the session, *number of unique items* seen, total *session duration* in seconds and *average duration per item* in seconds. The temporal features we use are *month* (0 to 12), *hour of the day* (0 to 23) and *day of the week* (0 to 6) and *weekend* (0 or 1) of the session start. Last, we incrementally include the item similarity scores for different combinations of interacted item positions in the sessions to assess their impact on the purchase prediction. This encompasses considering the similarity of the first-second item, first-third, second-third, first-second and first-third, first-second and second-third, first-third and second-third, and first-second, first-third and second-third. To compare them with the similarity score of combinations of first, middle and last done by Esmeli et al. (2020), we apply the same procedure.

4.4 Purchase Prediction Models

We employ machine learning algorithms trained on a binary classification task to predict purchases, determining whether a given session results in a purchase. In line with relevant literature, we evaluate four distinct classification models, i.e., DT, Bagging, RF, and LR.

The DT algorithm (Berry & Linoff, 2004) uses decision rules to predict the value of the target variable by splitting the data into smaller subsets based on the values of input features. Bagging (Louppe & Geurts, 2012) is an ensemble method that uses multiple weak learning models, in our case DTs, trained on random subsets of the data to improve accuracy and reduce variance. The final prediction is obtained by combining the individual predictions using an aggregation function. RF (Breiman, 2001) is a specific type of bagging which trains multiple weak DTs on randomly selected subsets of features. This helps to reduce the correlation between the trees and increase the diversity of the models.

In contrast to Esmeli et al. (2020), we also implement LR in addition to ensemble methods. LR applies a logistic function to the linear combination of input features, providing interpretable estimates that represent both the importance and directionality of each feature on the probability of the target class (Hosmer Jr, Lemeshow & Sturdivant, 2013). Ensemble methods on the contrary, only provide feature importances which show the relative importance of features in the decision-making process. On LR features with larger magnitudes can dominate the learning process and potentially lead to biased coefficient estimates. Since we have features with different ranges, we apply feature scaling before building the model. DT, RF, and Bagging are invariant to feature scaling because they make decisions based on relative feature thresholds rather than absolute values. Consequently, scaling the features may not have a significant impact on the performance of these models.

To address the issue of imbalanced data, as mentioned in Sect. 3, sampling methods can be applied, such as Undersampling (Kubat, Matwin et al., 1997) and Synthetic Minority oversampling Technique (SMOTE) (Chawla, Bowyer, Hall & Kegelmeyer, 2002). Esmeli et al. (2021) test both sampling methods on RF, DT and Bagging classifiers and show that all models produce better results when SMOTE is applied. Therefore, we opt to apply SMOTE to the training datasets.

Differently to Esmeli et al. (2020), we apply hyperparameters search to build effective models. To this end, we divide the data into training and test set using a ratio of 90:10. Subsequently, hyperparameters are set by applying grid-search over 10-fold cross-validations of the training dataset. In preliminary research, we find that the first-second and, first-third item pairs contribute most to the purchase prediction. Therefore, we opt to tune the models based on this feature set to save running costs. For the first, middle and last interaction combinations, we set hyperparameters using the first-middle and middle-last similarity scores.

We use the scikit-learn library in Python to implement the classification models (Pedregosa et al., 2011). During the training of the classification models, we apply SMOTE to address the class imbalance problem. The machine used for building the framework is a 2.1 GHz Intel Core i3-10110U CPU with 16 GB RAM. After the hyperparameter search, we find equal grids for the first three item combinations and, the first, middle and last combinations. Table 4.1 summarises the hyperparameters determined by the tuning process for each model and dataset.

4.5 Evaluation Metrics

In our evaluation of classification model performance, we opt to use two common evaluation metrics, namely the F1 score and Area Under the Curve (AUC). The F1 score is computed as the harmonic mean of Recall and Precision and reflects the model's ability to accurately predict purchase sessions. It overcomes the shortcomings of stand-alone precision and recall measures, which can be misleading if the data is highly imbalanced, by balancing the trade-off between precision and recall.

Additionally, we use the AUC score to assess the model's ability to distinguish between purchase and non-purchase sessions across various thresholds. Specifically, the AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate plotted against the false positive rate at various classification thresholds. The AUC has the advantage of being independent of the classification threshold whereas the F1 score is computed using only one threshold. This allows for a more robust assessment of the model's performance on imbalanced data. Both scores range from 0 to 1, with higher values indicating better performance.

Table 4.1: Tuned hyperparameter of the prediction models on the RetailRocket and Closed dataset

Model	Parameter	RetailRocket	Closed
DT	Max depth	10	10
	Min samples split	4	2
	Min samples leaf	20	50
Bagging	N estimators	70	100
	Max depth	10	10
	Max samples	0.6	0.6
RF	N estimators	120	120
	Criterion	gini	gini
	Max Features	sqrt	sqrt
	Max depth	20	15
LR	Solver	newton-cg	newton-cg
	Max iter	50	60

Chapter 5

Evaluation

The evaluation of our methods consists of two parts. First, we analyse the PAS and CPAS scores against JS scores in Sect. 5.1. Thereafter, we evaluate the purchase prediction performance across four classification models in Sect. 5.2. We examine the prediction performance of similarity scores for combinations of the first three items in sessions as signal for early purchase intent. To this end, we benchmark against combinations of item interactions of the first, middle and last item interactions, as done by Esmeli et al. (2020) and against baseline features. For the baseline we only use temporal and click-stream features, similarity scores are excluded. To test the robustness of our approaches, we evaluate two datasets, namely the open-sourced RetailRocket dataset and a Closed dataset.

We create item pairs and test for all possible combinations of them, in total there are 7 different feature sets for each approach. We use abbreviations to present item positions and item pairs. We denote the first interacted item by *f*, the second interacted item by *s*, and the third interacted item in the session by *t*. The middle and last item interactions are denoted by *m* and *l*, respectively. The similarity between two item interactions is indicated by their positions and '-', e.g., the similarity between *f* and *s* is denoted by *f-s.* When all possible item pairs are included, we denote that by *all included.* We refer to the baseline models with *similarity not included.*

5.1 Similarity Scores Evaluation

Table 5.1 and Tab. 5.2 and, Tab. 5.3 and Tab. 5.4 present the computed similarity scores on the RetailRocket and Closed dataset, respectively. We observe that purchase sessions contain relatively a higher proportion of similar item interactions than non-purchase sessions. More than 50% of the purchase sessions have item interaction for which we get equal recommendation lists. Non-purchase, on the other hand, have a higher variation in similarity scores, thus more diverse items are clicked within these sessions. In the preceding subsections, we evaluate in more detail the scores for each dataset.

RetailRocket Dataset

In Tab. 5.1 we observe for the first three interactions that *f-t* results in the highest average similarity in non-purchase sessions while yielding relatively lower values in purchase sessions. Furthermore, Fig. 5.1 reveals that this item pair exhibits the most extreme similarity values, indicating that items are either very similar or not at all. On the other hand, *f-s* and *s-t* similarity scores emphasise more the pronounced differences between highly similar items in purchase sessions and dissimilar items in non-purchase sessions. Especially, for the bucket 0.95-1.0 for purchase sessions we observe *f-s* and *s-t* being more often highly similar compared to non-purchase sessions.

Comparing JS with PAS and CPAS scores, we see on average higher values for PAS and CPAS in Tab. 5.1 compared to JS. PAS results even in slightly higher similarity values compared to CPAS. For PAS values in non-purchase sessions, the largest increase in mean compared to JS is for the *s-t*, 32.7%, and 29.2% and 29.8% for *f-s* and *f-t*, respectively. Purchase sessions exhibit a lower increase in similarity values. The highest average increase is shown for *f-t* by 3.96% and for *f-s* and *s-t* which is 3.66% and 3.76%, respectively. These small increases in values for purchase sessions are explained by the fact that recommendation lists for items are highly similar or not identical at all. In these cases, PAS and CPAS can not make any impact compared to JS. In non-purchase sessions, there are more non-extreme similarity values for which PAS and CPAS could adjust. Also, taking into account the insights from the statistics presented in Table 3.3, we can infer that there are numerous non-identical item combinations that share relevant items within recommendation lists. Since this distinction in similarity scores between purchase and non-purchase sessions is most pronounced for *f-t*, we expect that this item pair will show higher performance when using PAS.

Table 5.1: Statistics of similarity scores between the first three item interactions for purchase and non-purchase sessions in the RetailRocket dataset

Non-purchase Sessions

	JS(f, s)	JS(f, t)	JS(s, t)	PAS(f, s)	PAS(f, t)	PAS(s, t)	CPAS(f, s)	CPAS(f, t)	CPAS(s, t)
mean	0.56	0.59	0.55	0.68	0.68	0.67	0.67	0.68	0.67
std	0.35	0.38	0.34	0.34	0.36	0.33	0.33	0.35	0.32
min	0	0	0	0	0	0	0	0	0
25%	0.27	0.24	0.27	0.45	0.40	0.45	0.48	0.43	0.48
50%	0.56	0.61	0.55	0.80	0.83	0.79	0.77	0.81	0.76
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1

Purchase Sessions

	JS(f, s)	JS(f, t)	JS(s, t)	PAS(f, s)	PAS(f, t)	PAS(s, t)	CPAS(f, s)	CPAS(f, t)	CPAS(s, t)
mean	0.76	0.72	0.77	0.81	0.76	0.81	0.81	0.76	0.81
std	0.37	0.40	0.37	0.34	0.37	0.33	0.34	0.37	0.33
min	0	0	0	0	0	0	0	0	0
25%	0.50	0.37	0.53	0.75	0.61	0.77	0.73	0.61	0.74
50%	1	1	1	1	1	1	1	1	1
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1

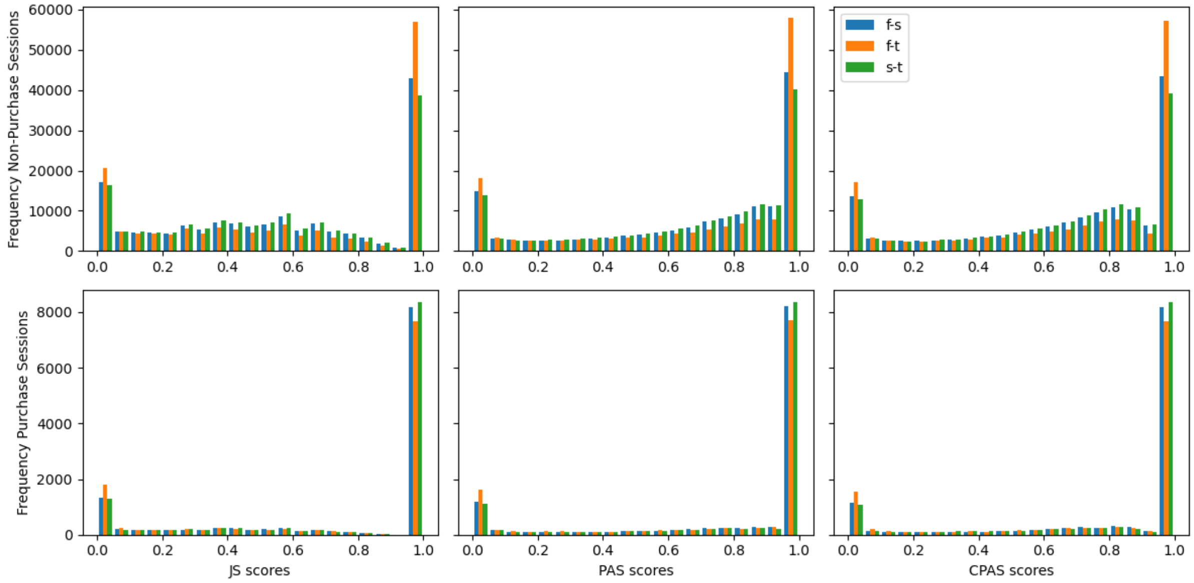

Figure 5.1: Distribution of similarity scores between the first, second and third item interactions in the RetailRocket dataset

We also evaluate the similarity scores for *f*, *m*, and *l* interactions. In Tab. 5.2 and Fig. 5.2 we see that the *m-l* are clearly more often highly similar compared to *f-m* and *f-l*. For *f-m* and *f-l* we observe a better distinction in similarity values between purchase and non-purchase sessions, being on average lower in non-purchase sessions compared to purchase sessions. Hence, we expect that the *f-m* and *f-l* will contribute the most to purchase prediction.

From these observations, we can conclude that the similarity scores extract better similarity than item-item similarity can reveal. We observe that on average the recommendations lists have more items within the top of the list in common than at the bottom resulting in higher similarity values. Furthermore, extrapolating the relevance of items with the recommendations lists could, on the one hand, help identify purchase sessions better in the classification models. On the other hand, we observe that the distinction between similarity scores in purchase and non-purchase sessions also diminishes slightly, since they are both highly similar in purchase and non-purchase sessions. This could make purchase classification also more difficult.

Closed Dataset

For the Closed dataset, we see in Tab. 5.3 and Tab. 5.4 that interactions within purchase sessions are significantly more similar than non-purchase sessions. In addition, Fig. 5.3 and Fig. 5.4 show that the Closed dataset reveals significantly less similarity between item interactions in non-purchase sessions compared RetailRocket dataset. This is in line with our expectations since we observed in Sect. 3 on average more unique items and especially within the first three item interactions of non-purchase sessions for the Closed dataset compared to the RetailRocket dataset.

Regarding the first three item interactions in the Closed dataset, we observe that *f-t* is the least similar in the purchase and non-purchase sessions, while this item pair reveals the highest similarity in non-purchase session on the RetailRocket dataset. This could be related to shorter

Table 5.2: Statistics of similarity scores between the first, middle, and last item interactions for purchase and non-purchase sessions in the RetailRocket dataset

Non-purchase Sessions									
	JS(f, m)	JS(f, l)	JS(m, l)	PAS(f, m)	PAS(f, l)	PAS(m, l)	CPAS(f, m)	CPAS(f, l)	CPAS(m, l)
mean	0.57	0.57	0.64	0.67	0.66	0.73	0.67	0.66	0.73
std	0.37	0.38	0.35	0.35	0.37	0.33	0.35	0.36	0.32
min	0	0	0	0	0	0	0	0	0
25%	0.24	0.21	0.35	0.40	0.34	0.56	0.43	0.38	0.57
50%	0.57	0.58	0.78	0.80	0.81	0.89	0.78	0.79	0.88
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1
Purchase Sessions									
	JS(f, m)	JS(f, l)	JS(m, l)	PAS(f, m)	PAS(f, l)	PAS(m, l)	CPAS(f, m)	CPAS(f, l)	CPAS(m, l)
mean	0.71	0.69	0.74	0.75	0.73	0.77	0.75	0.73	0.77
std	0.40	0.41	0.39	0.38	0.40	0.37	0.37	0.39	0.37
min	0	0	0	0	0	0	0	0	0
25%	0.33	0.24	0.38	0.51	0.41	0.61	0.51	0.41	0.60
50%	1	1	1	1	1	1	1	1	1
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1

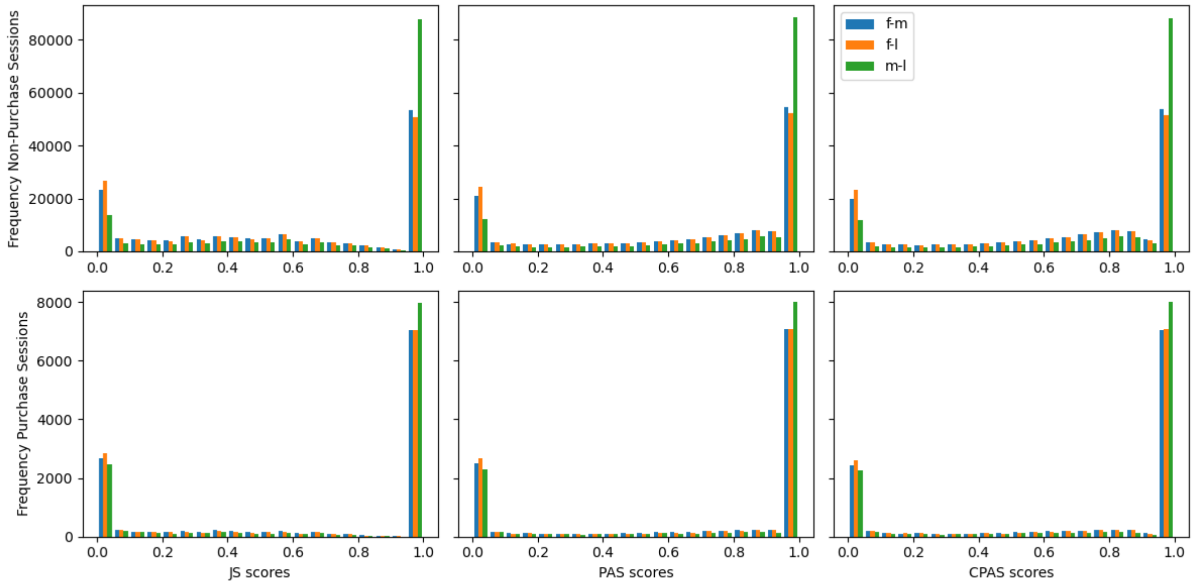

Figure 5.2: Distribution of similarity scores between the first, middle and last item interactions in the RetailRocket dataset

sessions in the RetailRocket dataset. Moreover, we see a smaller percentage change in the mean for JS and PAS. For non-purchase sessions, the percentage increase in the mean is 2.92%, 2.98% and 3.27% for *f-s*, *s-t* and *s-t*, respectively. For purchase sessions, the percentages are 1.78%, 1.52% and 1.40%, respectively. One of the main reasons for these smaller changes compared to the RetailRocket dataset is that in cases where the similarity of the recommendation list is 1 or 0, PAS and CPAS are also 1 and 0 and no correction takes place. For example, when comparing the results shown in Fig. 5.3, we observe minimal differences when using PAS and CPAS on purchase and non-purchase sessions within the 0.95-1.0 bucket. Similarly, in the 0.0-0.05 bucket, we find similar results for purchase sessions. However, for non-purchase sessions, there is a clear distinction with higher similarity scores for items that have fewer items in common. This suggests that these item pairs in purchase sessions with a few common recommendations do have relevant items in common. We hypothesise that this clear distinction in similarity scores between purchase and non-purchase sessions helps to better classify the purchase intent for all item pairs.

Regarding the first, middle, and last item interactions, we see in Fig. 5.4 that similar to the RetailRocket dataset *m-l* shows the highest similarity in both purchase and non-purchase sessions. We see that PAS and CPAS similarity scores are particularly increased from the left tail, the least similar items. Similarity scores for item pairs in the 0.95-1.0 bucket undergo small changes in both purchase and non-purchase sessions. This is likely due to the fact that these items are identical. Similar to the RetailRocket dataset, we expect *f-m* and *f-l* to contribute mostly to purchase prediction performance, where *f-l* could reveal even better purchase indication to better distinction in values between purchase and non-purchase sessions.

Table 5.3: Statistics of similarity scores between the first three item interactions for purchase and non-purchase sessions in the Closed dataset

Non-purchase Sessions									
	JS(f, s)	JS(f, t)	JS(s, t)	PAS(f, s)	PAS(f, t)	PAS(s, t)	CPAS(f, s)	CPAS(f, t)	CPAS(s, t)
mean	0.27	0.22	0.25	0.34	0.29	0.34	0.34	0.29	0.34
std	0.35	0.34	0.34	0.36	0.35	0.35	0.36	0.35	0.35
min	0	0	0	0	0	0	0	0	0
25%	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
50%	0.10	0.06	0.10	0.20	0.12	0.20	0.21	0.13	0.22
75%	0.33	0.26	0.32	0.59	0.49	0.57	0.57	0.48	0.55
max	1	1	1	1	1	1	1	1	1
Purchase Sessions									
	JS(f, s)	JS(f, t)	JS(s, t)	PAS(f, s)	PAS(f, t)	PAS(s, t)	CPAS(f, s)	CPAS(f, t)	CPAS(s, t)
mean	0.73	0.68	0.73	0.76	0.70	0.76	0.76	0.70	0.76
std	0.41	0.44	0.41	0.39	0.42	0.39	0.39	0.42	0.38
min	0	0	0	0	0	0	0	0	0
25%	0.24	0.12	0.25	0.46	0.23	0.47	0.45	0.24	0.46
50%	1	1	1	1	1	1	1	1	1
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1

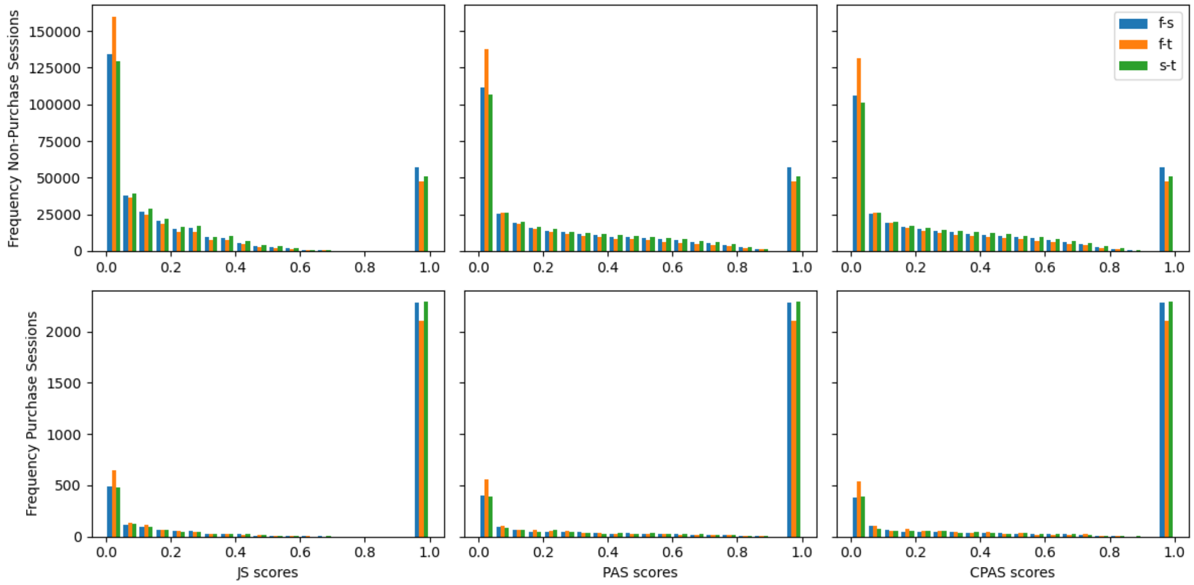

Figure 5.3: Similarity scores between the first, second and third item interactions in the Closed dataset

Table 5.4: Statistics of similarity scores between the first, middle, and last item interactions for purchase and non-purchase sessions in the Closed dataset

	Non-purchase Sessions								
	JS(f, m)	JS(f, l)	JS(m, l)	PAS(f, m)	PAS(f, l)	PAS(m, l)	CPAS(f, m)	CPAS(f, l)	CPAS(m, l)
mean	0.20	0.19	0.44	0.26	0.24	0.49	0.26	0.25	0.49
std	0.32	0.32	0.45	0.34	0.34	0.44	0.34	0.34	0.44
min	0	0	0	0	0	0	0	0	0
25%	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.03
50%	0.04	0.03	0.20	0.08	0.05	0.38	0.09	0.06	0.38
75%	0.22	0.20	1.00	0.43	0.39	1.00	0.43	0.39	1.00
max	1	1	1	1	1	1	1	1	1
	Purchase Sessions								
	JS(f, m)	JS(f, l)	JS(m, l)	PAS(f, m)	PAS(f, l)	PAS(m, l)	CPAS(f, m)	CPAS(f, l)	CPAS(m, l)
mean	0.62	0.60	0.69	0.64	0.62	0.71	0.64	0.62	0.71
std	0.46	0.47	0.44	0.45	0.45	0.43	0.44	0.45	0.42
min	0	0	0	0	0	0	0	0	0
25%	0.04	0.02	0.09	0.07	0.03	0.19	0.07	0.04	0.21
50%	1	1	1	1	1	1	1	1	1
75%	1	1	1	1	1	1	1	1	1
max	1	1	1	1	1	1	1	1	1

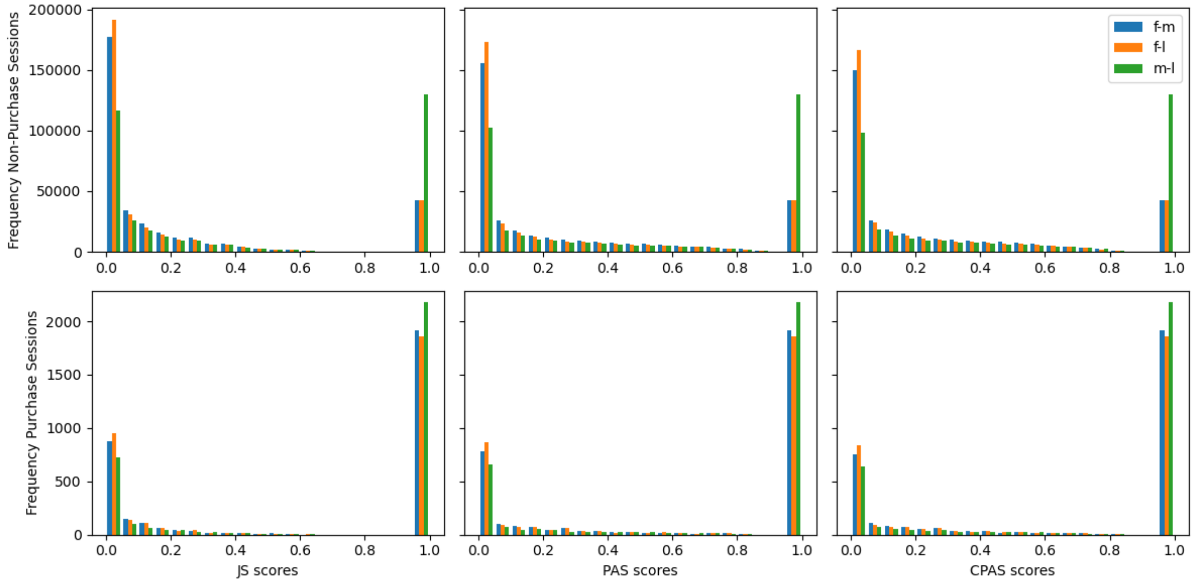

Figure 5.4: Similarity scores between the first, middle and last item interactions in the Closed dataset

5.2 Prediction Performance Evaluation

We measure prediction performance in terms of F1 and AUC. In the conducted experiments, we find that the distribution of PAS and CPAS scores is not significantly different, nor are there any significant differences in prediction performance. Therefore, in this section, we choose to compare only JS and PAS against each other and against the baseline. In Sect. 5.2.1, we evaluate the purchase prediction performance using the complete session information to create the temporal and click-stream features. To further enhance early prediction of purchase intent, it would be of great use to leverage only the information available from the first three interactions of the session. This approach is highly beneficial as it enables the purchase intent prediction even after just three interactions, which can greatly assist e-commerce strategists in executing real-time personalised marketing strategies to optimise conversion rates effectively. Therefore, we examine the early purchase intention prediction using only information available on the first three item interaction in Sect. 5.2.2.

5.2.1 Performance Using Complete Session Information

In this section, we first evaluate the purchase prediction on the RetailRocket dataset and thereafter on the Closed dataset.

RetailRocket Dataset

In Tab. 5.5 we see that the similarity score approaches outperform the baseline on four out of four prediction models.Regarding DT and Bagging, we see that *f-t* and *s-t* mainly contributes to better performance. However, relatively more feature importance is assigned *f-t* (Fig. A.1). Also, considering item pair combinations, *f-t* contributes most to the prediction accuracy.

RF and LR reveal slightly different patterns. First, we find that RF allocates overall more feature importance to the similarity scores compared to DT and Bagging (Fig. A.1). Second, relatively more importance is placed on *f-s* and *s-t* than *f-t*, while *f-t* is the dominant predictor on DT and Bagging. Especially, *s-t* contributes to better performance on LR for both similarity scores. In combination with other item pairs, *s-t* is the most dominant. Interestingly, for LR we see that *f-t* even has minimal effect and even a negative effect on the purchase intention when all similarity scores are included on LR. This could be due to the relatively high similarity of *f-t* in both purchase and non-purchase sessions. Since the data is highly imbalanced, this could result in a bias toward indicating non-purchase intention, also indicated by relatively lower AUC scores on LR. As presented in Sect. 5.1 the similarity values of *f-s* and *s-t* reveal a better distinction in similarity values of purchase and non-purchase session. Similarity features *f-s* and *s-t* show on LR a positive effect on the purchase intention, indicating that the interactions of subsequently three very similar items could be a signal of early purchase intention.

Table 5.5: Purchase prediction performance on different classifications models using JS and PAS scores for the first three item interactions in complete sessions of the RetailRocket dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
				Jaccard Similarity (JS) Score				
similarity not included	0.363	0.856	0.359	0.872	0.373	0.864	0.356	0.839
f-s	0.359	0.856	0.360	0.869	0.372	0.865	0.356	0.839
f-t	0.363	0.861	0.366	0.871	0.372	0.866	0.350	0.835
s-t	0.363	0.860	0.366	0.872	0.373	0.867	0.359	0.843
f-s, f-t	0.364	0.860	**0.372**	0.872	0.376	0.867	0.354	0.839
f-s, s-t	0.362	0.858	0.362	0.869	0.373	0.866	0.355	0.843
f-t, s-t	0.362	0.860	0.369	0.872	**0.377**	0.868	**0.359**	0.843
all included	0.364	0.861	0.372	**0.873**	0.374	**0.869**	0.357	**0.845**
				Position Aware Similarity (PAS) Score				
f-s	0.361	0.857	0.361	0.869	0.373	0.864	0.349	0.839
f-t	0.365	**0.862**	0.367	0.871	0.369	0.866	0.350	0.836
s-t	0.363	0.859	0.361	0.872	0.376	0.866	0.354	0.840
f-s, f-t	0.366	0.862	0.371	0.872	0.374	0.868	0.349	0.839
f-s, s-t	0.358	0.854	0.362	0.869	0.376	0.867	0.354	0.840
f-t, s-t	0.365	0.861	0.368	0.872	0.372	0.868	0.355	0.841
all included	**0.368**	0.859	0.370	0.872	0.375	0.869	0.352	0.842

The prediction results for similarity scores between the *f, m*, and *l* item interactions are shown in Tab. 5.6. We clearly see that *f-m* outperforms the baseline in DT, Bagging and RF in case of using JS and PAS. We show in Sect. 5.1 that this item pair also has the most extreme value distinction between purchase and non-purchase sessions. Our finding that *f-m* contributes most to the purchase intention accuracy is similar to what Esmeli et al. (2020) find. Considering the feature importance scores (Fig. A.2), we note only little differences between JS and PAS for DT, Bagging and RF. In these models, we see relatively lower importance of *m-l* compared to *f-m*, *f-l*. This indicates that *f-m* is a real contributor to better purchase performance.

Again LR shows a different pattern. On this model *m-l* shows the highest performance among similarity features. However, this similarity feature exhibits lower importance compared to *f-m*, *f-l* . More specifically for PAS, we see that *m-l* has little and even a negative effect on LR purchase intention when used with other item pairs. However, in neither combination the baseline is outperformed, which suggests that LR is not able to find information patterns in the similarity scores.

To draw some general conclusions on the RetailRocket dataset, we achieve for both item combinations, i.e., the first three items and the first, middle, and last item comparable levels of performance. Although the differences between the similarity features are relatively small, *f-s*, *s-t*, and *f-m* contribute most to signalling early purchase prediction. The best prediction performance is achieved for Bagging in terms of AUC and for RF in terms of F1. These results suggest that on the RetailRocket dataset, RF is best able to predict purchase intention correctly while Bagging is better able to distinguish the purchase and non-purchase sessions. The performance of Bagging can be explained by the fact that it trains multiple DTs on bootstrapped samples and averages their predictions. Consequently, the ensemble can collectively capture a wider range of patterns enhancing its ability to discriminate between purchase and non-purchase sessions and leading to a higher AUC. On the other hand, RFs superior performance in terms of F1 score can be attributed to its capacity to effectively capture complex decision boundaries. RF trains multiple decision trees on a random selection of features, which can lead to lower correlation among the trees and improved generalisation. The individual decision trees within the RF ensemble collectively learn different aspects of the data, contributing to improved performance in identifying purchase intentions.

Table 5.6: Purchase prediction performance on different classifications models using JS and PAS for the first, middle and last item interactions in complete sessions of the RetailRocket dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
			Jaccard Similarity (JS) Score					
similarity not included	0.363	0.856	0.359	0.872	0.373	0.864	**0.356**	**0.839**
f-m	0.365	**0.862**	0.369	**0.873**	0.375	0.869	0.348	0.835
f-l	0.361	0.857	0.364	0.869	0.370	0.864	0.348	0.836
m-l	0.359	0.855	0.358	0.871	0.375	0.865	0.354	0.837
f-m, f-l	0.364	0.857	**0.372**	0.872	0.374	0.868	0.348	0.836
f-m, m-l	0.363	0.858	0.367	0.872	0.378	0.869	0.351	0.835
f-l, m-l	0.357	0.852	0.363	0.869	0.373	0.866	0.352	0.836
all included	0.365	0.857	0.371	0.873	0.373	0.868	0.349	0.835
			Position Aware Similarity (PAS) Score					
f-m	0.365	0.861	0.369	0.873	0.377	**0.870**	0.350	0.836
f-l	0.362	0.858	0.364	0.870	0.375	0.865	0.350	0.837
m-l	0.363	0.856	0.359	0.871	0.371	0.865	0.352	0.838
f-m, f-l	0.366	0.857	0.370	0.872	0.377	0.867	0.350	0.836
f-m, m-l	0.365	0.860	0.367	0.872	0.377	0.868	0.349	0.836
f-l, m-l	0.358	0.852	0.363	0.869	0.370	0.866	0.349	0.837
all included	**0.367**	0.858	0.370	0.873	**0.378**	0.868	0.348	0.836

Closed Dataset

Looking at the performances on the Closed dataset in Tab. 5.7, we generally observe lower F1 and higher AUC scores compared to the RetailRocket dataset. Higher AUC scores could be due to a clearer distinction in feature values between purchase and non-purchase features. Lower F1 scores can be explained by the fact that the Closed dataset is more imbalanced than the RetailRocket dataset shown in Sect. 3, making it harder to accurately predict purchase sessions. Despite the clearer distinction in similarity scores between purchase and non-purchase sessions, we see that in all models it is harder to outperform the baseline using similarity scores. This could also be explained by the strong imbalance of the data, where the high similarity scores of purchase sessions make up a very small proportion of the data. Additionally, for SMOTE it is more difficult to create accurate synthetic samples of purchase sessions.

The *f-s* interaction pair is the most influential in purchase prediction performance while including additional features does not significantly enhance performance except for LR. Interestingly, JS scores show significantly better performance than PAS on the Closed dataset also more importance is placed on JS's similarity features, while the difference between JS and PAS on the RetailRocket dataset is minimal. Moreover, the relative importance of *f-s*, *f-t*, and *s-t* differ between using JS and PAS in DT and Bagging, indicating their sensitivity to minor value changes. RF shows the highest prediction performance in terms of F1 and outperforms the baseline, with *f-s* playing a key role in predicting purchase intention.

Table 5.7: Purchase prediction performance on different classifications models using JS and PAS for the first three items interaction in sessions of the Closed dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
				Jaccard Similarity (JS) Score				
similarity not included	**0.123**	0.890	**0.128**	**0.917**	0.150	0.904	0.119	0.917
f-s	0.123	0.886	0.127	0.909	0.150	**0.916**	0.111	0.925
f-t	0.111	0.878	0.123	0.908	0.144	0.908	0.113	0.922
s-t	0.113	0.860	0.124	0.906	**0.151**	0.913	0.115	0.924
f-s, f-t	0.110	0.869	0.116	0.899	0.148	0.912	0.124	0.926
f-s, s-t	0.105	0.859	0.115	0.898	0.140	0.912	0.125	0.929
f-t, s-t	0.104	0.874	0.116	0.903	0.143	0.908	0.126	0.925
all included	0.109	0.868	0.116	0.899	0.143	0.911	**0.128**	**0.929**
				Position Aware Similarity (PAS) Score				
f-s	0.102	0.874	0.120	0.908	0.150	0.916	0.112	0.923
f-t	0.104	**0.893**	0.114	0.906	0.144	0.909	0.113	0.920
s-t	0.108	0.881	0.119	0.908	0.147	0.911	0.118	0.921
f-s, f-t	0.107	0.872	0.116	0.903	0.146	0.912	0.122	0.923
f-s, s-t	0.101	0.877	0.109	0.905	0.142	0.910	0.122	0.926
f-t, s-t	0.106	0.887	0.113	0.906	0.143	0.908	0.124	0.922
all included	0.107	0.862	0.116	0.901	0.139	0.909	0.122	0.926

Regarding the performance of the similarity scores between the *f*, *m*, and *l* item interactions, we generally see in Tab. 5.8 no major differences in the level of performance compared to the *f*, *s*, and *t* item interactions in Tab. 5.7. For the Closed dataset, we see that the classification models significantly place more importance on *f-l* and *f-m* than *m-l* (Fig. A.5). This could be explained by the fact that this similarity feature shows the clear distinction in values between purchase and non-purchase session making providing more clear information for the decision process. We observe that in most cases also *f-l* contributes to better prediction of purchase intention, though only outperforming the baseline on RF and LR. For this similarity feature, we find that RF is the best performing model in terms of F1 and outperforming the baseline.

Table 5.8: Purchase prediction performance on different classifications models using JS and PAS for the first, middle and, last item interactions in complete sessions of the Closed dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
				Jaccard Similarity (JS) Score				
similarity not included	**0.123**	**0.890**	**0.128**	**0.917**	0.150	0.904	0.119	0.917
f-m	0.108	0.880	0.115	0.906	0.145	0.911	0.114	0.921
f-l	0.123	0.888	0.127	0.909	**0.156**	**0.916**	**0.126**	0.920
m-l	0.105	0.878	0.117	0.911	0.152	0.911	0.114	0.919
f-m, f-l	0.113	0.876	0.122	0.902	0.150	0.910	0.113	0.921
f-m, m-l	0.110	0.877	0.115	0.905	0.148	0.910	0.118	0.922
f-l, m-l	0.118	0.880	0.127	0.907	0.154	0.914	0.117	0.921
all included	0.113	0.870	0.122	0.899	0.150	0.909	0.116	**0.922**
				Position Aware Similarity (PAS) Score				
f-m	0.108	0.879	0.113	0.907	0.143	0.913	0.116	0.918
f-l	0.123	0.886	0.127	0.908	0.151	0.913	0.125	0.918
m-l	0.105	0.883	0.117	0.911	0.150	0.914	0.115	0.918
f-m, f-l	0.115	0.873	0.122	0.903	0.146	0.910	0.115	0.919
f-m, m-l	0.111	0.879	0.115	0.906	0.144	0.911	0.117	0.920
f-l, m-l	0.119	0.880	0.126	0.907	0.150	0.913	0.117	0.919
all included	0.114	0.871	0.121	0.901	0.143	0.907	0.116	0.920

5.2.2 Performance Using Information from First Three Item Interactions

So far we have used temporal and click-stream features from complete session information. Now, we examine the performance effect of using only the information on the first three items to the early purchase intention prediction.

RetailRocket Dataset

Table 5.9 shows that the overall performance descreases by on average 0.1, even for the baseline models when only the information of the first three interactions is used on the RetailRocket dataset. Furthermore, when including all similarity features, the four classification models consistently outperform the baseline, with PAS scores achieving higher F1 and AUC values than to JS. We observe similar patterns of the similarity features on the prediction performance as in Sect. 5.2.1 on the RetailRocket, though higher feature importances (Fig. 5.9). The significance of *f-s* becomes even more apparent when using PAS, while the performance of *f-t* and *s-t* worsens compared to JS, suggesting the importance of PAS in extracting underlying similarity information, particularly with limited information.

Table 5.9 shows that *f-s* is the main contributor to higher F1 scores and *f-t* to higher AUC. The high AUC scores of *f-t* could also question again the bias, as indicated in Sect. 5.2.1 on the RetailRocket dataset. For the combination of *f-s* and *f-t*, *f-t* is relatively more important on DT and Bagging and *f-s* on RF (Fig. A.3). Furthermore, we see in Tab. 5.9 that the good complementarity of these item pairs for purchase prediction is again confirmed by the set *f-s*, *f-t*. Interestingly, we find that all similarity features exhibit negative effects on the likelihood of purchase, while we find for *f-s* and *s-t* positive effects when the complete session information is used (Fig. A.3). In Sect. 5.2.1, we find that the similarity values are relatively higher in purchase sessions than in non-purchase sessions, suggesting that higher similarity increases the likelihood of purchase intention. Since only the baseline features are changed and the similarity scores remain the same for the first three items, we may question the stability of LR in this case.

Closed Dataset

Tab. 5.10 shows an average decrease of 0.07 in prediction scores compared to using the complete session information (Tab. 5.7 Moreover, JS scores attain higher prediction scores than PAS when only the first three session interactions are used, which was also found with the complete session information. This is contradictory to the RetailRocket dataset in Tab. 5.9, PAS results in higher scores on DT Bagging and RF.

The main difference in feature importance across similarity scores on the Closed dataset is observed for DT and Bagging, where similarity features become less important in the case of PAS, especially *s-t* (Fig. A.6). Furthermore, we observe that the similarity features become less important when only the first three items are used compared to the complete session information (Fig. A.4), suggesting the relatively greater importance of baseline features and difficulty to extract informative decision boundaries from similarity features when information is more limited. This observation might explain why the baseline generally performs better. For LR, none of the similarity features outperforms the baseline, although *s-t* shows almost equal performance. Similar to the estimated coefficients of the RetailRocket dataset on LR (Fig. A.3), we see that in most cases the similarity features have a negative effect on the likelihood of purchase (Fig. A.6), which we do not observe when using the complete session information (Fig. A.4), again calling into question the stability of these models.

Interestingly, we find that *f-t* and *s-t* are the most contributing predictors to better performance across classification models, whereas in Tab. 5.7 we observed that *f-s* was the most contributing to better prediction performance, though much importance is placed on *f-s* (Fig.A.6). This could imply that *f-s* reveals interesting patterns, but when considering the first three items it is harder to capture complex decision boundaries.

Table 5.9: Purchase prediction performance on different classifications models using JS and PAS on only the first three item interactions in sessions of the RetailRocket dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
			Jaccard Similarity (JS) Score					
similarity not included	0.281	0.733	0.281	0.739	0.287	0.740	0.271	0.723
f-s	0.274	0.722	0.279	0.732	0.281	0.736	0.250	0.725
f-t	0.267	0.742	0.270	0.753	0.269	0.757	0.272	0.744
s-t	0.264	0.726	0.264	0.731	0.273	0.735	0.246	0.721
f-s, f-t	0.284	0.755	0.289	0.760	0.293	0.765	0.281	0.747
f-s, s-t	0.263	0.721	0.269	0.728	0.284	0.738	0.242	0.721
f-t, s-t	0.254	0.740	0.259	0.749	0.265	0.756	0.271	0.741
all included	0.299	0.759	0.302	0.766	0.305	0.770	0.275	0.746
			Position Aware Similarity (PAS) Score					
f-s	0.266	0.720	0.278	0.731	0.282	0.735	0.253	0.726
f-t	0.259	0.745	0.267	0.752	0.270	0.756	0.281	0.745
s-t	0.263	0.724	0.265	0.729	0.274	0.734	0.248	0.723
f-s, f-t	0.275	0.751	0.285	0.759	0.291	0.765	**0.283**	**0.748**
f-s, s-t	0.272	0.722	0.285	0.727	0.283	0.735	0.248	0.723
f-t, s-t	0.253	0.744	0.256	0.750	0.264	0.757	0.280	0.744
all included	**0.305**	**0.759**	**0.304**	**0.766**	**0.307**	**0.770**	0.280	0.748

Table 5.10: Purchase prediction performance on different classifications models using JS and PAS on only the first three item interactions in sessions of the Closed dataset

	DT		Bagging		RF		LR	
Similarity Feature(s)	F1	AUC	F1	AUC	F1	AUC	F1	AUC
			Jaccard Similarity (JS) Score					
similarity not included	0.083	0.806	0.081	**0.819**	0.094	**0.817**	**0.087**	**0.850**
f-s	0.071	0.782	0.075	0.795	0.097	0.785	0.078	0.849
f-t	0.093	0.781	0.093	0.797	0.099	0.799	0.083	0.849
s-t	0.088	0.779	0.090	0.798	**0.107**	0.801	0.087	0.849
f-s, f-t	0.071	0.779	0.076	0.789	0.093	0.788	0.079	0.849
f-s, s-t	**0.096**	**0.784**	0.093	0.793	0.101	0.793	0.078	0.849
f-t, s-t	0.095	0.779	0.094	0.800	0.104	0.799	0.084	0.848
all included	0.100	0.781	**0.099**	0.792	0.103	0.794	0.079	0.850
			Position Aware Similarity (PAS) Score					
f-s	0.078	0.778	0.080	0.795	0.090	0.792	0.080	0.849
f-t	0.089	0.780	0.080	0.798	0.091	0.796	0.083	0.848
s-t	0.086	0.776	0.087	0.796	0.102	0.799	0.087	0.849
f-s, f-t	0.071	0.776	0.074	0.787	0.090	0.792	0.080	0.849
f-s, s-t	0.089	0.776	0.090	0.788	0.097	0.794	0.080	0.849
f-t, s-t	0.086	0.771	0.088	0.792	0.096	0.794	0.082	0.847
all included	0.089	0.780	0.089	0.788	0.097	0.788	0.080	0.849

Chapter 6

Managerial Implications

Our evaluation of the results provides valuable insights for the practical implications, recommendations, and limitations of using item similarity as a feature for early purchase prediction in the e-commerce sector.

First, regarding the feature similarity sets we observe variate results across models in terms of feature importance and best similarity features. When comparing the first-middle-last (fml) and first-second-third (fst) combinations on the complete session information, we found a minimal difference in prediction performance, despite the fact that the fml similarity scores show generally better distinction in similarity values between purchase and non-purchase sessions. Moreover, if we compare the prediction performance of features extracted from the complete session versus the first three session interactions, we see that in the case of more limited information, overall performance decreases. Since there is a clear need for early prediction of purchase intention, we recommend further research on improving similarity representations. Additionally, the use of other features in combination with the similarity features which can be extracted at the beginning could be investigated. For example, (Esmeli et al., 2022) show that contextual features which are already available at the very start of the session such as device and IP location are effective for early purchase intention prediction.

Second, regarding the choice of prediction models, we found that RF classifiers consistently outperformed other models, showing the highest performance and feature importance across various similarity scores. Therefore, we recommend the RF model as the primary choice when building purchase prediction models using our selected feature sets.

Third, when comparing the JS and PAS similarity scores, we find that PAS results in overall higher values than JS indicating that items share even more similarity. On the one hand, this increase diminishes the distinction between values for purchase and non-purchase sessions. On the other, this results in different relative importances and performance scores among similarity features, though no significant differences in the performances are found. Consequently, the results do not directly indicate which similarity score is better to use, and further investigation is needed to determine the most suitable choice.

Fourth, despite the fact that our findings of best-performing similarity features are different across the two datasets, we could say that similarities between the first and subsequent items contribute most to purchase intention detection. On the RetailRocket dataset, *f-t*, *s-t* and *f-m*,*m-l* achieve the best performance on the RF model, in which *f-t* and *f-m* are relatively more important and *f-t* within the combination even has a negative effect on the purchase. Only using the available information from the first three item interactions on this dataset highlights the importance of the similarity scores. On the Closed dataset, *f-s* and *f-l* performed the best as a standalone feature in terms of AUC and F1 on the Closed dataset when the complete session information is used, including more similarity features does not necessarily result in better performance. Contrary, if we only consider the first three item interactions *f-t* and *s-t* show higher prediction performance, although *f-s* gets also significant importance assigned. These findings strongly suggest that the initial interacted items within sessions reflect the user's initial purchase interest. Furthermore, the observed similarity between this first item and subsequent items within the session serves as a valuable indicator of the user's underlying purchase intention.

It is important to acknowledge the limitations of our study. Despite adopting the approach to Esmeli et al. (2020), we did not achieve comparable F1 scores. This suggests that there may be additional factors or techniques employed in their study that were not reported or considered in their work. Further investigation and clarification of their results are necessary to draw comparative conclusions. Furthermore, the short duration of e-commerce sessions introduces overlap between the first-middle-last and first-second-third item interactions, which may affect interpretability of the models. Additionally, it is important to examine the feature similarity values frequently to capture possible bias during prediction. Last, an one-unit increase in similarity features reveals on LR opposite directional effects on the likelihood of purchase intention across item combinations and when limited session information is used. This questions the stability of LR when using similarity features.

To conclude, the results suggest that the similarity of the first item with subsequent item interactions plays a significant role in predicting users' purchase intention. Considering that e-commerce sessions, especially purchase sessions, are often short, targeting users as soon as possible is critical. Therefore, we recommend using *f-s* and *f-t* on RF for purchase intention prediction. The directional effect on the purchase intention of the similarity features should be further investigated. Additionally, we recommend an integration with RS to provide recommendations or discounts for relevant items. Last, we suggest further investigation into selecting features next to the similarity features, which can contribute to an improved early purchase prediction.

Chapter 7

Conclusion

In this research, we investigate whether the patterns of the initial three interacted item similarities can be used as signal for early purchase intention prediction. We hypothesise that the similarity of the first session interactions reveals initial customer interest and could already be used to detect purchase intention. By leveraging the similarity between the first three items in the session, we aim to assist e-commerce businesses by identifying purchase intention not only during an ongoing session but also in the very early stages. With this approach, conversion rates can be optimised by using the purchase signal for targeted marketing strategies. We use item recommendations from word2vec embeddings to create similarity scores from two different e-commerce datasets and use them as features in various classification models. By implementing this approach, we demonstrate the effectiveness of item similarity as signal for early purchase intention prediction.

Compared to Esmeli et al. (2020), we create "position-aware similarity" scores, named PAS, which are not only based on the number of common recommendations for selected item pairs, but we also consider the relevancy of the item in the recommendations lists. Although this approach results in more comprehensive scores, the distinction between similarity scores in purchase and non-purchases diminished and it leads to varying levels of feature importances. In term of prediction performance, PAS does not yield significant improvements when compared to solely considering the number of common recommendations in the classification models.

However, our results show that the similarity score between the first three items in the session can be effectively used as signal for early purchase prediction. Specially, we find that the similarity between the first and preceding items contributes most to identifying purchase intention. We recommend using RF for inferring purchase intention based on these similarities. This model attains the best performance among the investigated models in terms of its F1 and AUC scores and in terms of its robustness. Additionally, it also allocates relatively more importance to the similarity features compared to other models, indicating that these truly contribute to purchase detection.

For future research, we are aiming to use the predictions to create automatised personalised marketing actions to drive users with uncertain purchase intentions towards a purchase or to engage users toward the company's goal. In this respect, we could integrate our approach into recommendation systems and, if there is a purchase intent, examine which items the consumer is most interested in and thus should be recommended or discounted, for example. Furthermore, this research highlights the effective performance of retrieving similarity scores using a simple word embedding-based model. However, since item similarity scores are computed independently from the purchase prediction our approach is also applicable to other recommendation models. We could further improve the embedding representation such that even better recommendations are provided using side information. In the field of purchase prediction, we see that temporal features have a positive effect on the prediction. Further research could address how such features could be included in the item representation so that these features do not have to be manually extracted. For example, Vasile, Smirnova and Conneau (2016) show that using the category when creating the embeddings improves item representations. Finally, it could also be interesting to evaluate whether the diversity instead of similarity between item interactions can be considered as a new feature for purchase prediction.

References

Alves Gomes, M., Meyes, R., Meisen, P. & Meisen, T. (2022). Will this online shopping session succeed? predicting customer's purchase intention using embeddings. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)* (pp. 2873–2882).

Bader-El-Den, M., Teitei, E. & Perry, T. (2018). Biased random forest for dealing with the class imbalance problem. *IEEE transactions on neural networks and learning systems*, *30*(7), 2163–2172.

Barkan, O. & Koenigstein, N. (2016). Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)* (pp. 1–6).

Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S. & Bala, P. K. (2020). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services*, *53*, 101799.

Berry, M. J. & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bianchi, F., Yu, B. & Tagliabue, J. (2021). Bert goes shopping: Comparing distributional models for product representations. In *Proceedings of the 4th Workshop on e-Commerce and NLP* (pp. 1–12).

Blasco-Arcas, L., Lee, H.-H. M., Kastanakis, M. N., Alcañiz, M. & Reyes-Menendez, A. (2022). The role of consumer data in marketing: A research agenda. *Journal of Business Research*, *146*, 436–452.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chatterjee, P., McGinnis, J. et al. (2010). Customized online promotions: Moderating effect of promotion type on deal value, perceived fairness, and purchase intent. *Journal of Applied Business Research (JABR)*, *26*(4).

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.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Asso- ciation for Computational Linguistics: Human Language Technologies* (Vol. 1, p. 4171–4186).

Esmeli, R., Bader-El-Den, M. & Abdullahi, H. (2020). Using word2vec recommendation for improved purchase prediction. In *2020 international joint conference on neural networks (ijcnn)* (pp. 1–8).

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.

Esmeli, R., Bader-El-Den, M. & Abdullahi, H. (2022). An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain. *Journal of Business Research*, *147*, 420–434.

Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N., Savla, J., Bhagwan, V. & Sharp, D. (2015). E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1809–1818).

Hidasi, B., Karatzoglou, A., Baltrunas, L. & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. *Proceedings of the International Conference on Learning Representations*.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jannach, D., Ludewig, M. & Lerche, L. (2017). Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts. *User Modeling and User-Adapted Interaction*, *27*, 351–392.

Kubat, M., Matwin, S. et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, p. 179).

Lin, W., Milic-Frayling, N., Zhou, K. & Ch'ng, E. (2019). Predicting outcomes of active sessions using multi-action motifs. In *Ieee/wic/ACM international conference on web intelligence* (pp. 9–17).

Liu, X., Lee, D. & Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*, *56*(6), 918–943.

Louppe, G. & Geurts, P. (2012). Ensembles on random patches. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I 23* (pp. 346–361).

Ludewig, M., Mauro, N., Latifi, S. & Jannach, D. (2019). Performance comparison of neural and non-neural approaches to session-based recommendation. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 462–466).

Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C. & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3), 588–596.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mokryn, O., Bogina, V. & Kuflik, T. (2019). Will this session end with a purchase? inferring current purchase intent of anonymous visitors. *Electronic Commerce Research and Applications*, *34*, 100836.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning*

Research, *12*, 2825–2830.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Řehřek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora.

Sheil, H., Rana, O. & Reilly, R. (2018). Predicting purchasing intent: Automatic feature learning using recurrent neural networks. *arXiv preprint arXiv:1807.08207*.

Tagliabue, J., Greco, C., Roy, J.-F., Yu, B., Chia, P. J., Bianchi, F. & Cassani, G. (2021). Sigir 2021 e-commerce workshop data challenge. *SIGIR eCom 2021*.

Vasile, F., Smirnova, E. & Conneau, A. (2016). Meta-prod2vec: Product embeddings using side-information for recommendation. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 225–232).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 5998–6008.

Wen, Z., Lin, W. & Liu, H. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. *Systems*, *11*(5), 255.

Zhou, Y., Mishra, S., Gligorijevic, J., Bhatia, T. & Bhamidipati, N. (2019). Understanding consumer journey using attention based recurrent neural networks. In *Proceedings of the 25th ACM sigkdd international conference on knowledge discovery & data mining* (pp. 3102–3111).

Zimmermann, R. & Auinger, A. (2022). Developing a conversion rate optimization framework for digital retailers—case study. *Journal of Marketing Analytics*, 1–11.

Appendix A

Similarity Feature Importances and Coefficients

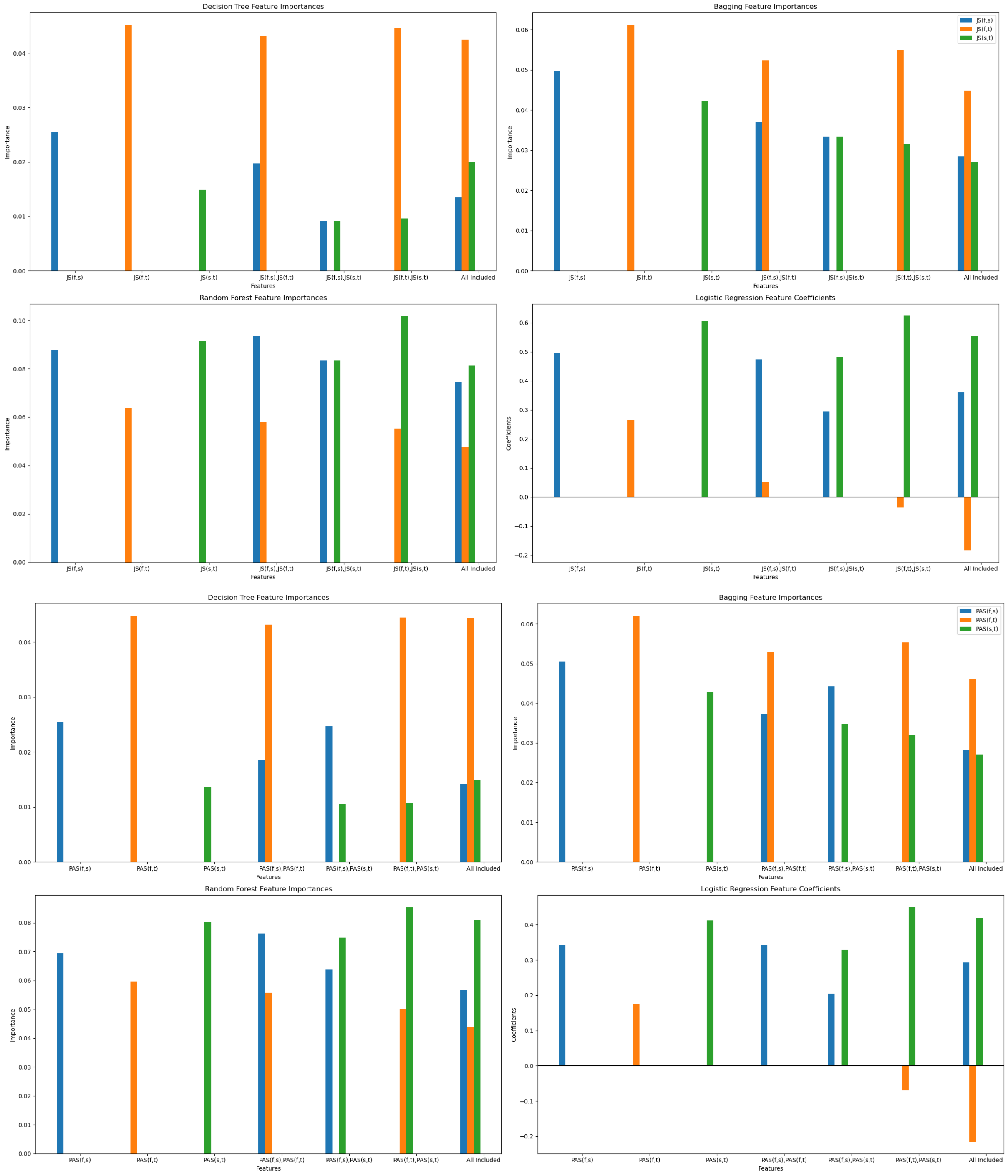

Figure A.1: Similarity feature importances from first three item interactions on DT, Bagging, RF, and estimated coefficients on LR using the RetailRocket dataset

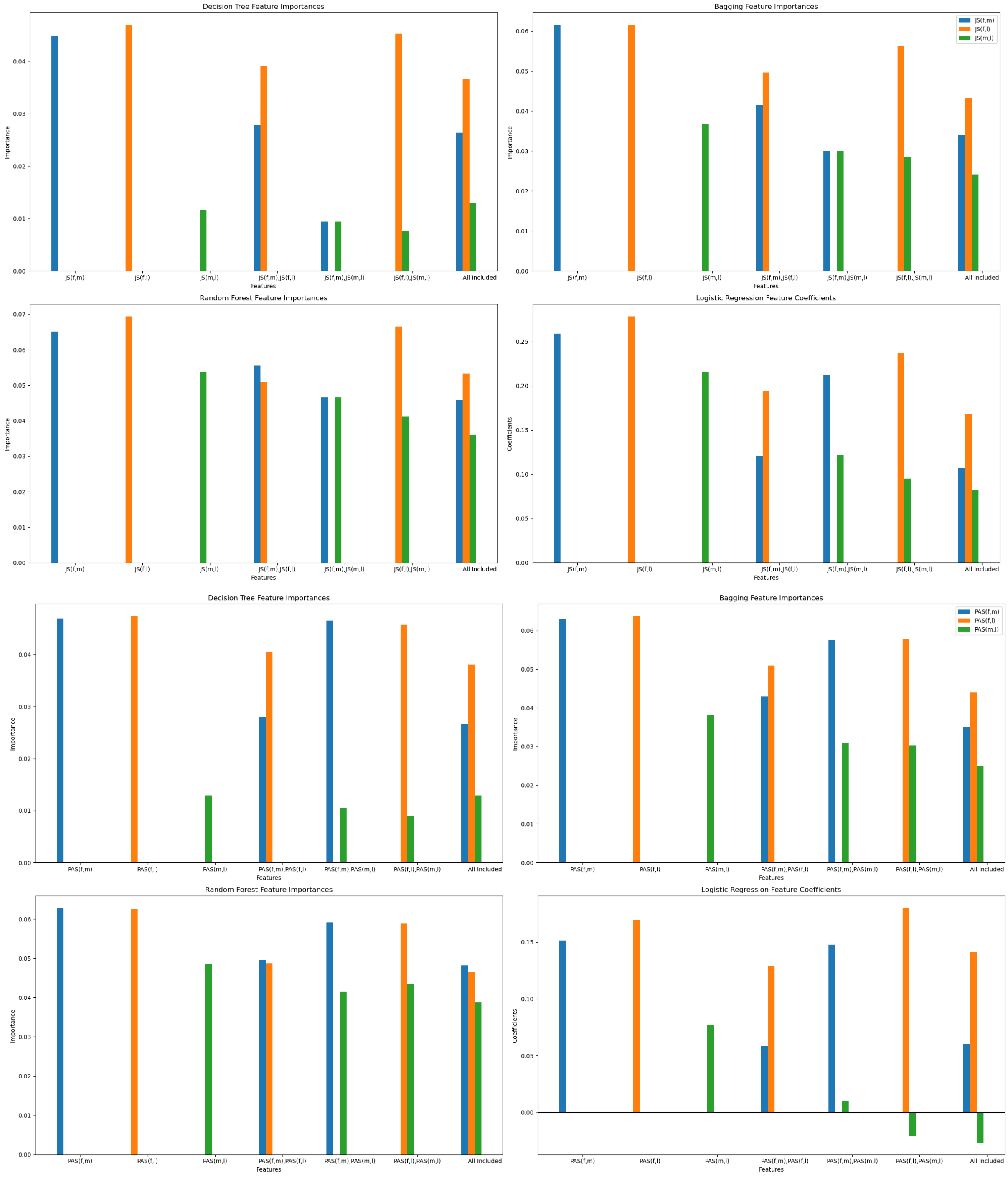

Figure A.2: Similarity feature importances from the first, middle, and last item interactions on DT, Bagging and RF and estimated coefficients on LR using the RetailRocket dataset

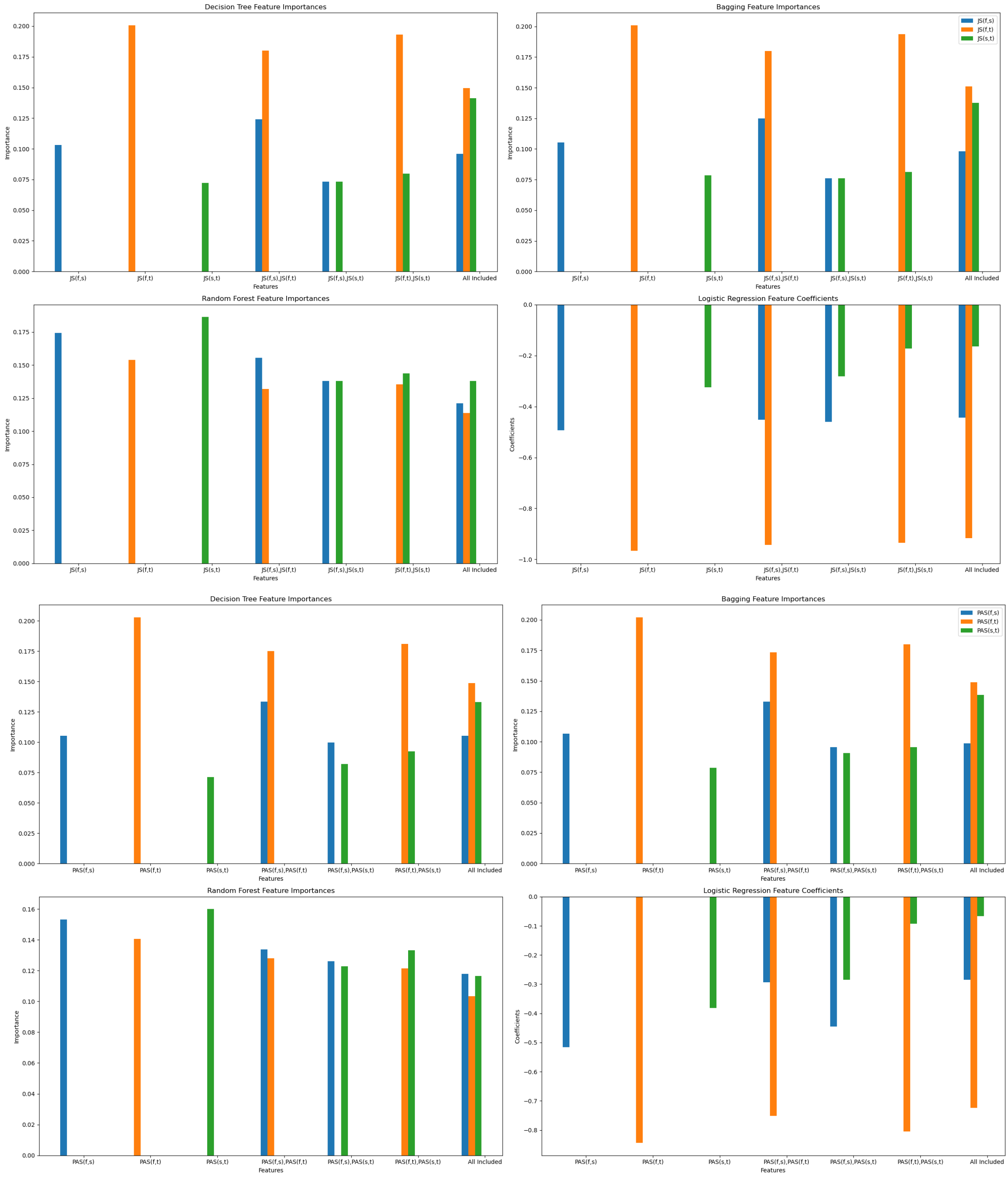

Figure A.3: Similarity feature importances on DT, Bagging and RF and estimated coefficients on LR using only available information on the first three item interactions on the RetailRocket dataset

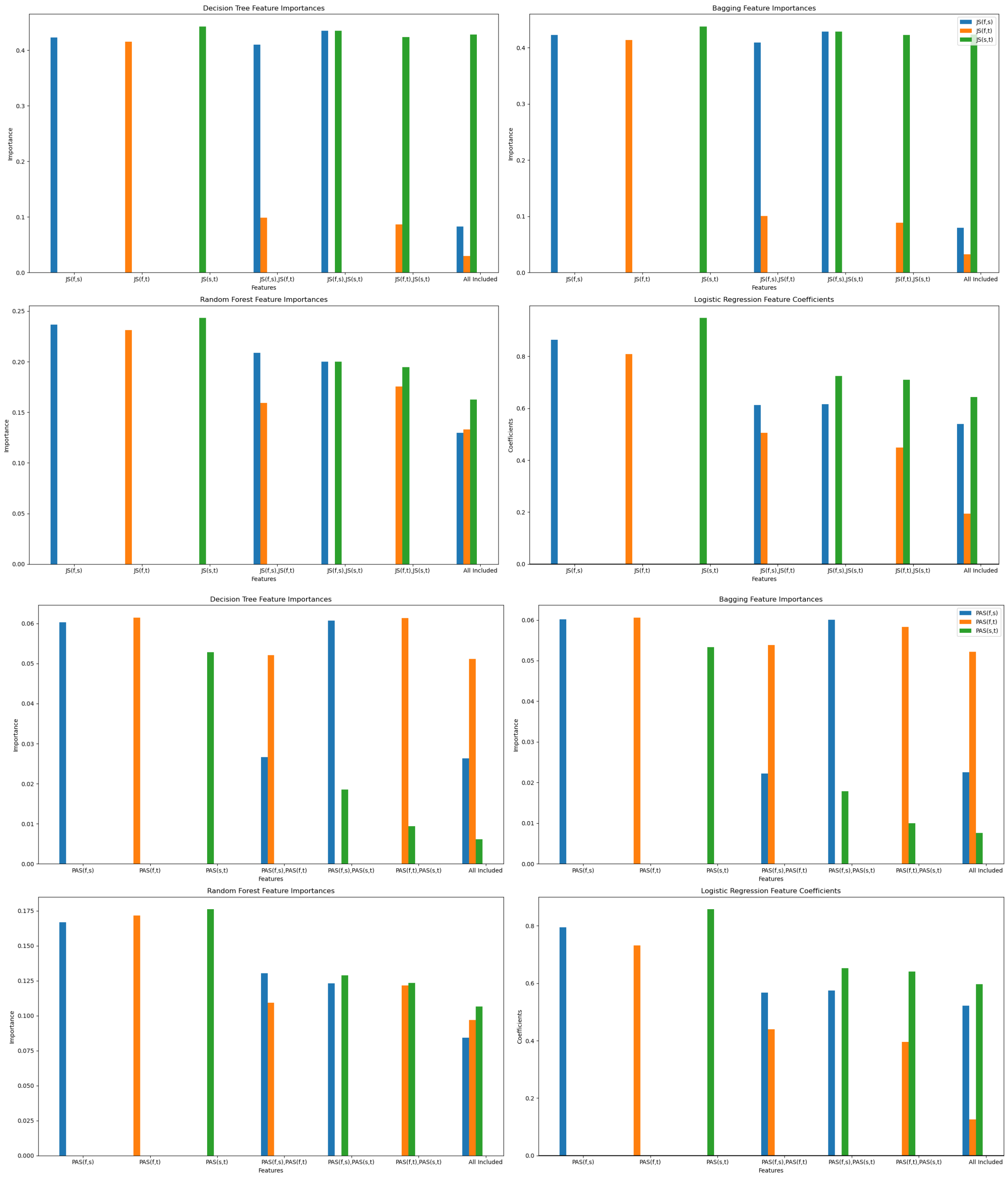

Figure A.4: Similarity feature importances from first three item interactions on DT, Bagging, RF, and estimated coefficients on LR using the Closed dataset

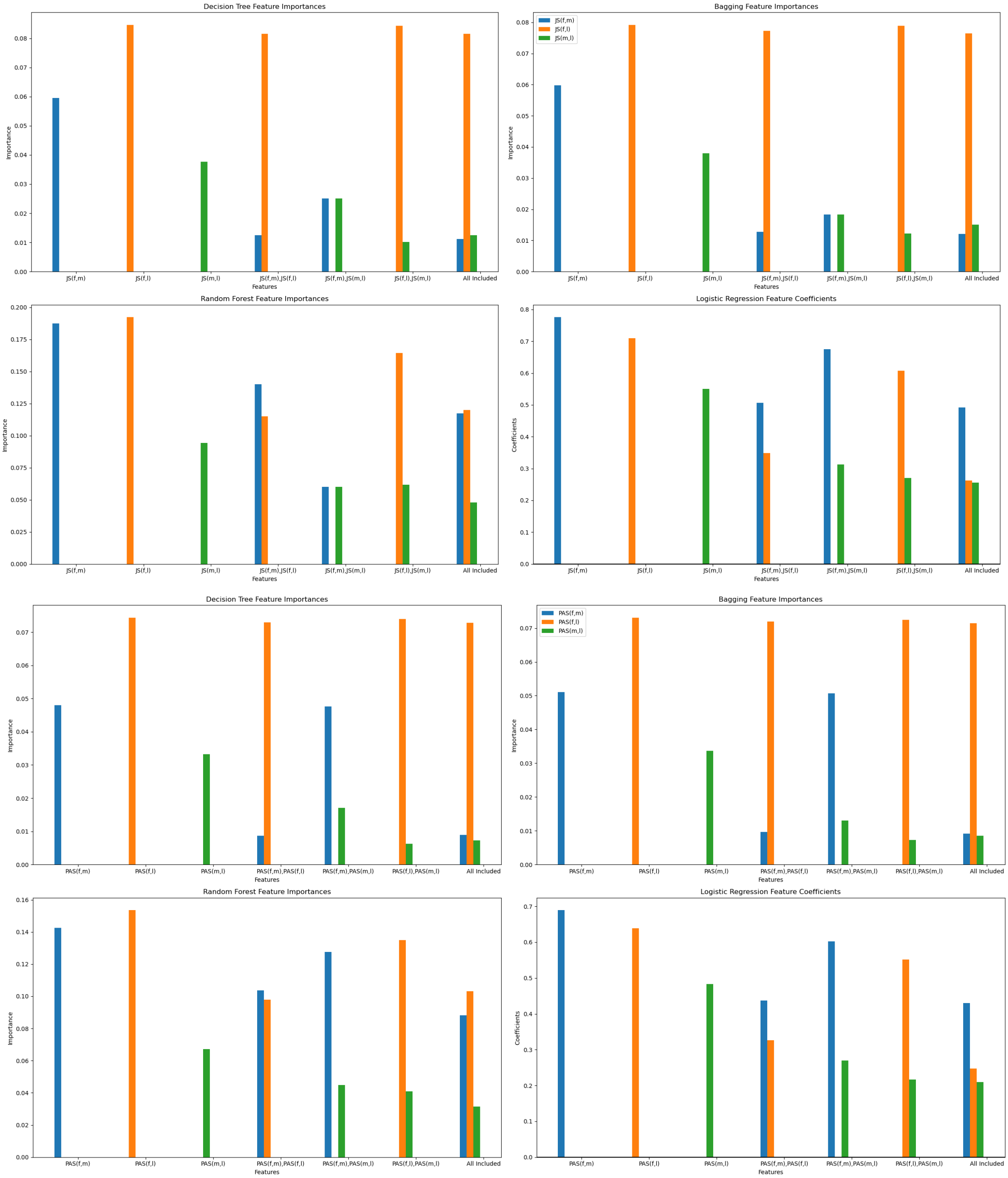

Figure A.5: Similarity feature importances from the first, middle, and last item interactions on DT, Bagging, RF, and estimated coefficients on LR using the Closed dataset

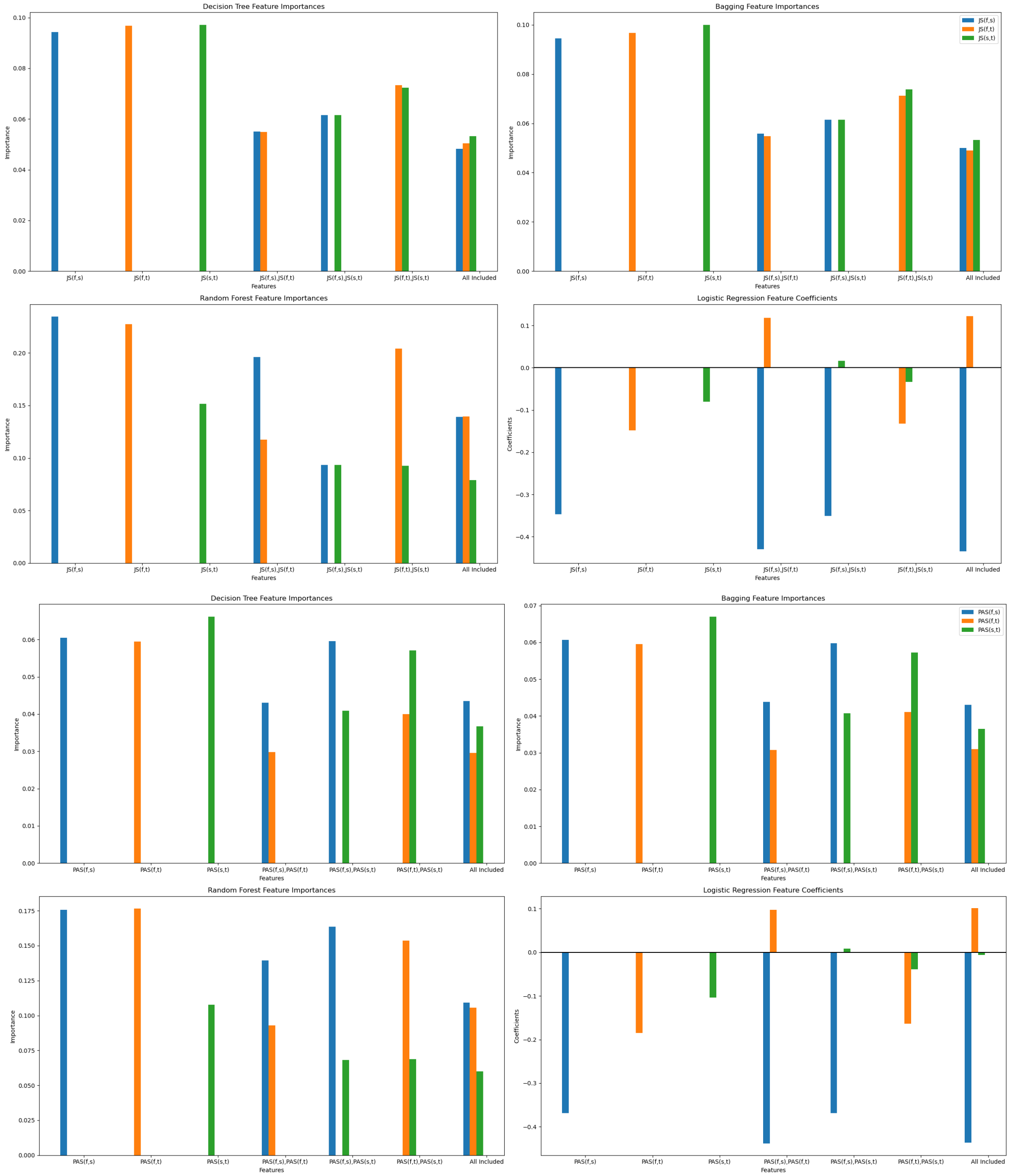

Figure A.6: Similarity feature importances on DT, Bagging and RF and estimated coefficients on LR using only available information on the first three item interactions on the Closed dataset