Feature Attribution Explanations of Session-based Recommendations
Appendix

Simone Borg Bruun[1,2][0000−0003−1619−4076], Maria Maistro[2][0000−0002−7001−4817], and Christina Lioma[2][0000−0003−2600−2701]

[1] Alka Forsikring, Denmark
[2] University of Copenhagen, Denmark
simone.bruun@alka.dk
{mm,c.lioma}@di.ku.dk

A Implementation Details

A.1 Optimal Hyperparameters for Recommendation Models

The optimal hyperparameters for the recommendation models are in Tab. A.1.

Table A.1: Optimal hyperparameters.

	GRU4REC		SASRec	
	Diginetica	30Music	Diginetica	30Music
Batch Size	128	128	256	256
Units	256	256	64	128
Dropout	0.3	0.4	0.4	0.3

A.2 Training Procedure for Explainability Models

In LIME, we use 5000 samples as the size of the neighborhood to learn the linear model. In Deep Shap, we integrate over the 100 most recent sessions from the training set, as it is only necessary (and most efficient) to integrate over a smaller sample of the training set. In IG, we use as baseline a session with all the interactions replaced by the padding value ignored by the model (e.g., 0), as this should act as an input with no information. We use Deep Shap and IG directly on the logit values (predicted from the model). In LIME, we convert the logit values to probabilities and train local classification models.

B Additional Analysis

We find examples where the faithfulness of our method is better than LIME, Deep Shap and IG to illustrate how the difference affects the resulting explanations. Fig. B.1 presents an example with attribution scores (normalized) generated for a session with a sequential dependency between interaction i_1 and i_2

	Sequential Dependency i_1 i_2	i_3	i_4	Rank Correlation	Attribute Correlation
LIME	-1.60 (-0.50, -1.10)	1.17	0.43	0.33	-0.96
Deep Shap	-1.28 (-0.71, -0.57)	-0.18	1.46	-1.00	-0.40
IG	-1.16 (-0.60, -0.56)	-0.33	1.49	-1.00	-0.30
Session-based Occlusion	1.00	-1.01	0.01	1.00	1.00

Fig. B.1: Example of attribution scores (normalized) generated for a session with a sequential dependency between two interactions. For LIME, Deep Shap and IG, the attribution score for the sequential dependency is computed as the sum of their individual scores indicated in parentheses.

	i_1	i_2	Repeated Interaction i_3 i_4	i_5	i_6	Rank Correlation	Attribute Correlation
LIME	-0.31	-0.44	-0.97 (-0.54, -0.43)	-0.32	2.04	0.40	0.97
Deep Shap	-0.52	-0.49	-0.78 (-0.41, -0.37)	-0.19	1.98	0.00	0.98
IG	-0.34	-0.43	-0.90 (-0.45, -0.45)	-0.37	2.04	0.40	0.98
Session-based Occlusion	-0.43	-0.45	-0.44	-0.46	1.78	1.00	1.00

Fig. B.2: Example of attribution scores (normalized) generated for a session with a pair of repeated interactions (i.e., interaction i_3 and i_4 is with the same item). For LIME, Deep Shap and IG, the attribution score for the repeated interactions is computed as the sum of their individual scores indicated in parentheses.

(detected by our method). The session is an example from the Diginetica dataset and the GRU4REC model. For LIME, Deep Shap and IG, the attribution score for the interactions with a sequential dependency is computed as the sum of their individual scores. We see that all three methods find that the individual scores for these two interactions are negative, but the joint attribution score computed by our method is positive. This is likely because the isolated contribution to the prediction of these two interactions is negative, but the model has learned a sequential dependency, so when they appear in this particular order, they have a positive contribution to the prediction. We see that this also affects the attribution score of the other interactions in the session, as the attribution scores for i_3 and i_4 found by LIME, Deep Shap and IG are very different from the ones found by our method. This is because the additive feature attribution methods try to distribute the contributions between the interactions. In this example, both the rank correlation and the attribute correlation are worse with LIME, Deep Shap and IG than with our method.

Fig. B.2 presents an example of attribution scores (normalized) generated for a session where i_3 and i_4 are interactions with the same item (i.e., repeated interactions). This session is an example from the Diginetica dataset and the SASRec model. We see that LIME, Deep Shap and IG assign to each of the repeated interactions approximately the same contribution as our method assigns to both of them in total. This is likely because this interaction has the same contribution to the prediction no matter how many times it appears in the session. This is captured by our method, but not by the standard additive feature attribution. In this example, the attribute correlation is almost the same for all the methods, but the rank correlation is lower with LIME, Deep Shap and IG than with our method, because they assign too much importance to the repeated interactions.