An analysis of local explanation with LIME-RS*

Vito Walter Anelli[1], Alejandro Bellogín[2], Tommaso Di Noia[1], Francesco Maria Donini[3], Vincenzo Paparella[1,**] and Claudio Pomo[1,**]

[1]Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy
[2]Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain
[3]Università degli Studi della Tuscia, via Santa Maria in Gradi, 4, 01100 Viterbo, Italy

Abstract

Explainable Recommendation has attracted a lot of attention due to a renewed interest in explainable artificial intelligence. In particular, post-hoc approaches have proved to be the most easily applicable ones, since they treat as black boxes the increasingly complex recommendation models. Recent literature has shown that for post-hoc explanations based on local surrogate models, there are problems related to the robustness of the approach itself. This consideration becomes even more relevant in human-related tasks, from transparency or trustworthiness points of view – like recommendation. We show how the behavior of LIME-RS – a classical post-hoc model based on surrogates – is strongly model-dependent and does not prove to be accountable for the explanations generated.

Keywords

explainable recommendation, post-hoc explanation, local surrogate model

1. Introduction

The explanation of a recommendation list plays an increasingly important role in the interaction of a user with a Recommender System (RS) [2, 3, 4]. Given the explanation that a system can provide to a user we identify at least two characteristics that the explanation part should enforce [5, 6, 7]: (*i*) **Adherence** to reality: the explanation should mention only features that really pertain to the recommended item. (*ii*) **Constancy** in the behavior: although the explanation is generated based on some sample, and such a sample is drawn with a probability distribution, the entire process should not exhibit a random behavior to the user. Among several ways of generating explanations, we study here the application of LIME [8] to the recommendation process (LIME-RS [9]). LIME-RS is a post-hoc algorithm that can explain the predictions of any recommender in a faithful way, by approximating it locally with an interpretable model. While its black-box approach lets LIME-RS be applicable for every recommender system, the way the model is built – by drawing a huge random sample of system behaviors – makes it lose both adherence and constancy, as our experiments show.

12th Italian Information Retrieval Workshop (IIR 2022), June 29–July 1, 2022, Milan, Italy

*Extended version [1] published at the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021)

**Corresponding author.
vincenzo.paparella@poliba.it (V. Paparella); claudio.pomo@poliba.it (C. Pomo)

2. Related Work

In recent years, the theme of Explanation in Artificial Intelligence has come to the foreground, capturing the attention of the Machine Learning and related communities [5, 10, 11], among others. This trend has also touched the research field of RSs [12, 13, 14, 15, 16, 17, 18]. Explainable Recommendation is defined as the task that aims to provide suggestions to the users and make them aware of the recommendation process, explaining also why that specific object has been suggested. On the one hand, the **model-intrinsic explanation strategy** aims at creating a user-friendly recommendation model or encapsulates an explaining mechanism. On the other hand, a model-agnostic [19] approach, also known as **post-hoc** [20], does not require to intervene on the internal mechanisms of the recommendation model and therefore it does not affect its accuracy. Many post-hoc explanation methods have been proposed for recommendation models based on Matrix Factorization (MF) [20, 21, 22, 15, 23, 24, 25, 26, 27].

Our paper focuses on the operation of LIME-RS that applies the explanation model technique LIME to the recommendation domain. The goal of LIME-RS is to exploit the predictive power of the recommendation model f (treated as a black box) to generate an explanation about the suggestion of a particular item $x \in \mathcal{X}$ for a user. LIME-RS exploits a neighborhood of samples $\{x' \mid x' \in \mathcal{X}\}$ drawn from the training set according to a generic distribution, and compared to the item x to be explained, to train an interpretable model e – tipically based on a linear prediction. It seems obvious that the choice of the neighborhood is crucial within the process of explanation generation by LIME-RS. One of the disadvantages of this approach is that it sometimes fails to estimate an appropriate local replacement model; instead, it generates a model that focuses on explaining the examples and is affected by more general trends in data.

These observations dictate the two research questions that motivated our work. **RQ1**: *Can we trust the surrogate-based model which LIME-RS is built on, to generate always the same explanations (Constancy), or does the extraction of a different neighborhood breaks down Constancy?* **RQ2**: *Are LIME-RS explanations adherent to item content, despite the fact that the sampling function is uncritical and based only on popularity?*

3. Experiments

The datasets used for this phase of experimentation are *Movielens 1M* [28], *Movielens Small* [28], and *Yahoo!* Movies[1]. As for the models to be used in this work, we selected two well-known recommendation models that are able to exploit the information content of the items to produce a recommendation: Attribute Item kNN (Att-Item-kNN) and Vector Space Model (VSM). The implementation of both models is available in the evaluation framework ELLIOT [29, 30]. This benchmarking framework was used to select the best configuration for the two recommendation models by exploiting the corresponding configuration file[2]. After choosing the best configuration (based on the nDCG metric [31, 32]) for each of the above two models, for each user u we generated the top-10 list L_u of recommendations, and we examined the first item i_1 on L_u. Finally, each recommendation pair (u, i_1) is explained with LIME-RS. The explanation consists

[1]http://webscope.sandbox.yahoo.com/
[2]https://tny.sh/basic_limers

of a weighted vector $(g, w)_i$ where g is the genre of the movies in the dataset – *i.e.*, the features – and w is the weight associated to g by LIME-RS within the explanation. Then, this vector is sorted by descending weights to highlight, in the first positions, the genres of the movies which played a key role within the recommendation. These operations are then repeated $n = 10$ times while changing the seed each time. At this point, for each pair (u, i_1), we have a group of 10 explanations ordered by descending values of w.

RQ1. We consider only the first five features in the sorted vector representing the explanation of each recommendation. In order to verify the constancy of the behavior of LIME-RS, given a (u, i_1) pair, we exploit the n previously generated explanations for this pair. Then for $k = 1, 2, \ldots, 5$, we define G_k as the multiset of genres that appear in k-th position – for instance, if "Sci-Fi" occurs in the first position of 7 explanations, then "Sci-Fi" occurs 7 times in the multiset G_1, and similarly for other genres and multisets. Then, we compute the frequency of genres in each position as follows: given a position k, a genre g, and the number n of generated explanations for a given pair (u, i_1), the frequency f_{g_k} of g in k-th position is computed as $f_{g_k} = \frac{\|\{g \mid g \in G_k\}\|}{n}$, where $\| \cdot \|$ denotes the cardinality of a multiset. Then, all this information is collected for each user in five lists — one for each of the k positions — of pairs $\langle g, f_{g_k} \rangle$ sorted by frequency. One can observe that the computed frequency is an estimation of the probability that a given genre is put in that position within the explanation generated by LIME-RS sorted by values. Hence, the pair $\langle g, \max\left(f_{g_k}\right) \rangle$ describes the genre with the highest frequency in the k-th position of the explanation for a pair (u, i_1). Finally, it makes sense to compute the mean μ_k of the highest probability values in each position k of the explanations for each pair (u, i_1). Formally, by setting a position k, the mean μ_k is computed as $\mu_k = \frac{\sum_{j=1}^{|U|} \max\left(f_{g_k}\right)_j}{|U|}$, where U is the set of users whom it was possible to generate a recommendation for. Observing the value of μ_k, we can state to what extent LIME-RS is constant in providing the explanations until the k-th feature: the higher the value of μ_k, the higher the constancy of LIME-RS concerning the k-th feature.

RQ2. With the aim at providing an answer about the adherence to reality of LIME-RS, we make a comparison between the genres claimed to explain a recommended item and its actual genres. Indeed, the explanations about an item should fit the list of genres the item is characterized by. This means that, in an ideal case, all highly weighted features within the explanation should match the genres of the item. We intersected each explanation limited to the set E_k of its first k genres with the set of genres F_{i_1} characterizing the first recommended item, for $k = 1, 2, 3$. Upon completion of this operation for all the n explanations generated for each (u, i_1) pair, we computed the number of times we obtained an empty intersection of these sets, normalized by the total number of explanations $n \times |U|$, in order to understand to what extent an explanation is (not) adherent to the item. Formally, for a given value of k, the value $adherence_k$ is computed as $adherence_k = \frac{\sum_{j=1}^{n \times |U|} \left[\left(E_k \cap F_{i_1}\right)_j = \emptyset\right]}{n \times |U|}$, where U is the set of users of the dataset for whom it was possible to generate a recommendation, n is the number of generated explanations for each pair (u, i_1), and by $\Sigma[\cdots]$ we mean that we sum 1 if the condition inside $[\cdots]$ is true, and 0 otherwise. One can note that $adherence_k \in [0, 1]$, where a value of 1 indicates the worst case in which for none of the n explanations under consideration at least one genre of the item is in the first k features of the explanation. In contrast, the lower the value of $adherence_k$, the higher the adherence of LIME-RS.

Table 1
Constancy. A value equals to 1 means that the genre(s) in the first k position(s) is always the same. **Adherence**. A value equals to 0 means one genre is always among the real genres of the movie.

	μ_1	μ_2	μ_3	μ_4	μ_5	$adherence_1$	$adherence_2$	$adherence_3$
		Att-Item-kNN						
Movielens 1M	0,9130	0,7822	0,6927	0,6288	0,5727	0,2774	0,1105	0,0488
Movielens Small	0,8830	0,7426	0,6639	0,60459	0,5616	**0,2364**	**0,0651**	**0,0180**
Yahoo! Movies	**0,9230**	**0,8016**	**0,7232**	**0,6528**	**0,5830**	0,3597	0,1202	0,0476
		VSM						
Movielens 1M	0,8929	0,7953	0,7729	0,7726	0,7801	0,5357	0,2539	0,1088
Movielens Small	0,9464	0,8636	0,8343	0,8138	0,8049	0,4384	0,1674	0,0403
Yahoo! Movies	**0,9732**	**0,9209**	**0,8887**	**0,8884**	**0,9056**	**0,1013**	**0,01348**	**0,0021**

Table 1 shows the different behaviors for Att-Item-kNN and VSM with respect to the two novel defined metrics. From the constancy point of view, Att-Item-kNN seems to guarantee a good constancy in explanations up to the third feature. This suggests that an explanation that exploits the first three features of the list produced by LIME-RS could be barely considered as reliable (i.e., reaching a constancy of 0.69 on Movielens 1M). In contrast, VSM exhibits a much more "stable" behavior, demonstrating in all cases (except for the first feature with Movielens 1M) better performance than Att-Item-kNN in terms of constancy, From the adherence point of view, the results show that Att-Item-kNN shows good performance regarding adherence and identifies 3 times out of 4 the first fundamental feature of the explanation among those present in the set of features originally associated with the item. As expected, if the number k of LIME-RS-reconstructed features increases, the number of times such a set has a nonempty intersection (with the features belonging to the item) – *i.e.,* adherence – increases. One can note that Att-Item-kNN on Yahoo! Movies shows the worst behavior in terms of adherence. VSM shows a different behavior. Despite the excellent performance regarding constancy, one can observe that on both Movielens datasets, the performance in terms of adherence is poor, and worse for Movielens 1M than for Movielens Small. Surprisingly, on Yahoo! Movies, VSM the errors are almost negligible.

4. Conclusion

In our experiments, some evidence started to emerge highlighting that the adopted explanation model is conditioned not only by the accuracy of the black-box model it tries to explain but also by the quality of the side information used to train the model. The latter result deserves to be adequately investigated to search for a link at a higher level of detail. We plan to apply our experiments also to other recommendation models, to see whether the problems with adherence and constancy that we found for the two tested models show up also in other situations. We will also investigate what impact structured knowledge has on this performance by exploiting models capable of leveraging this type of content. In addition, it would also be the case to try different reference domains with richer datasets of side information to understand what impact content quality has on this type of explainer.

References

[1] V. W. Anelli, A. Bellogín, T. D. Noia, F. M. Donini, V. Paparella, C. Pomo, Adherence and constancy in LIME-RS explanations for recommendation (long paper), in: V. W. Anelli, P. Basile, T. D. Noia, F. M. Donini, C. Musto, F. Narducci, M. Zanker, H. Abdollahpouri, T. Bogers, B. Mobasher, C. Petersen, M. S. Pera (Eds.), Joint Workshop Proceedings of the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), Virtual Event, Amsterdam, The Netherlands, September 25, 2021, volume 2960 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2960/paper11.pdf.

[2] V. Bellini, G. M. Biancofiore, T. D. Noia, E. D. Sciascio, F. Narducci, C. Pomo, Guapp: A conversational agent for job recommendation for the italian public administration, in: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2020, Bari, Italy, May 27-29, 2020, IEEE, 2020, pp. 1–7. URL: https://doi.org/10.1109/EAIS48028.2020.9122756. doi:10.1109/EAIS48028.2020.9122756.

[3] V. W. Anelli, Y. Deldjoo, T. D. Noia, A. Ferrara, F. Narducci, How to put users in control of their data in federated top-n recommendation with learning to rank, in: C. Hung, J. Hong, A. Bechini, E. Song (Eds.), SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021, ACM, 2021, pp. 1359–1362. URL: https://doi.org/10.1145/3412841.3442010. doi:10.1145/3412841.3442010.

[4] C. Ardito, T. D. Noia, E. D. Sciascio, D. Lofú, G. Mallardi, C. Pomo, F. Vitulano, Towards a trustworthy patient home-care thanks to an edge-node infrastructure, in: R. Bernhaupt, C. Ardito, S. Sauer (Eds.), Human-Centered Software Engineering - 8th IFIP WG 13.2 International Working Conference, HCSE 2020, Eindhoven, The Netherlands, November 30 - December 2, 2020, Proceedings, volume 12481 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 181–189. URL: https://doi.org/10.1007/978-3-030-64266-2_11. doi:10.1007/978-3-030-64266-2_11.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38. URL: https://doi.org/10.1016/j.artint.2018.07.007. doi:10.1016/j.artint.2018.07.007.

[6] N. Tintarev, J. Masthoff, Explaining recommendations: Design and evaluation, in: Recommender Systems Handbook, Springer, 2015, pp. 353–382.

[7] F. Gedikli, D. Jannach, M. Ge, How should I explain? A comparison of different explanation types for recommender systems, Int. J. Hum. Comput. Stud. 72 (2014) 367–382. URL: https://doi.org/10.1016/j.ijhcs.2013.12.007. doi:10.1016/j.ijhcs.2013.12.007.

[8] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, R. Rastogi (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[9] C. Nóbrega, L. B. Marinho, Towards explaining recommendations through local surrogate models, in: C. Hung, G. A. Papadopoulos (Eds.), Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, Limassol, Cyprus, April 8-12, 2019, ACM,

2019, pp. 1671–1678. URL: https://doi.org/10.1145/3297280.3297443. doi:10.1145/3297280.3297443.

[10] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[11] J. Chakraborty, K. Peng, T. Menzies, Making fair ML software using trustworthy explanation, in: 35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020, IEEE, 2020, pp. 1229–1233. URL: https://doi.org/10.1145/3324884.3418932. doi:10.1145/3324884.3418932.

[12] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Found. Trends Inf. Retr. 14 (2020) 1–101. URL: https://doi.org/10.1561/1500000066. doi:10.1561/1500000066.

[13] V. W. Anelli, T. D. Noia, E. D. Sciascio, A. Ragone, J. Trotta, How to make latent factors interpretable by feeding factorization machines with knowledge graphs, in: C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. F. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon (Eds.), The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I, volume 11778 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 38–56. URL: https://doi.org/10.1007/978-3-030-30793-6_3. doi:10.1007/978-3-030-30793-6_3.

[14] G. P. Polleti, H. N. Munhoz, F. G. Cozman, Explanations within conversational recommendation systems: improving coverage through knowledge graph embedding, in: 2020 AAAI Workshop on Interactive and Conversational Recommendation System. AAAI Press, New York City, New York, USA, 2020.

[15] D. Pan, X. Li, X. Li, D. Zhu, Explainable recommendation via interpretable feature mapping and evaluation of explainability, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 2690–2696. URL: https://doi.org/10.24963/ijcai.2020/373. doi:10.24963/ijcai.2020/373.

[16] K. Tsukuda, M. Goto, Dualdiv: diversifying items and explanation styles in explainable hybrid recommendation, in: T. Bogers, A. Said, P. Brusilovsky, D. Tikk (Eds.), Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019, ACM, 2019, pp. 398–402. URL: https://doi.org/10.1145/3298689.3347063. doi:10.1145/3298689.3347063.

[17] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 765–774. URL: https://doi.org/10.1145/3331184.3331254. doi:10.1145/3331184.3331254.

[18] G. Cornacchia, F. M. Donini, F. Narducci, C. Pomo, A. Ragone, Explanation in multi-stakeholder recommendation for enterprise decision support systems, in: A. Polyvyanyy, S. Rinderle-Ma (Eds.), Advanced Information Systems Engineering Workshops - CAiSE 2021 International Workshops, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings, volume 423 of *Lecture Notes in Business Information Processing*, Springer, 2021, pp. 39–47. URL: https://doi.org/10.1007/978-3-030-79022-6_4. doi:10.1007/978-3-030-79022-6_4.

[19] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, X. Xie, A reinforcement learning framework for explainable recommendation, in: IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018, IEEE Computer Society, 2018, pp. 587–596. URL: https://doi.org/10.1109/ICDM.2018.00074. doi:10.1109/ICDM.2018.00074.

[20] G. Peake, J. Wang, Explanation mining: Post hoc interpretability of latent factor models for recommendation systems, in: Y. Guo, F. Farooq (Eds.), Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, ACM, 2018, pp. 2060–2069. URL: https://doi.org/10.1145/3219819.3220072. doi:10.1145/3219819.3220072.

[21] Y. Tao, Y. Jia, N. Wang, H. Wang, The fact: Taming latent factor models for explainability with factorization trees, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 295–304. URL: https://doi.org/10.1145/3331184.3331244. doi:10.1145/3331184.3331244.

[22] J. Gao, X. Wang, Y. Wang, X. Xie, Explainable recommendation through attentive multi-view learning, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 3622–3629. URL: https://doi.org/10.1609/aaai.v33i01.33013622. doi:10.1609/aaai.v33i01.33013622.

[23] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, K. Järvelin (Eds.), The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, ACM, 2014, pp. 83–92. URL: https://doi.org/10.1145/2600428.2609579. doi:10.1145/2600428.2609579.

[24] F. Fusco, M. Vlachos, V. Vasileiadis, K. Wardatzky, J. Schneider, Reconet: An interpretable neural architecture for recommender systems, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 2343–2349. URL: https://doi.org/10.24963/ijcai.2019/325. doi:10.24963/ijcai.2019/325.

[25] M. Tsang, D. Cheng, H. Liu, X. Feng, E. Zhou, Y. Liu, Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=BkgnhTEtDS.

[26] V. W. Anelli, T. D. Noia, E. D. Sciascio, A. Ferrara, A. C. M. Mancino, Sparse feature factorization for recommender systems with knowledge graphs, in: H. J. C. Pampín, M. A. Larson, M. C. Willemsen, J. A. Konstan, J. J. McAuley, J. Garcia-Gathright, B. Huurnink, E. Oldridge (Eds.), RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021, ACM, 2021, pp. 154–165. URL: https://doi.org/10.1145/3460231.3474243. doi:10.1145/3460231.3474243.

[27] V. W. Anelli, T. D. Noia, P. Lops, E. D. Sciascio, Feature factorization for top-n recom-

mendation: From item rating to features relevance, in: Y. Zheng, W. Pan, S. S. Sahebi, I. Fernández (Eds.), Proceedings of the 1st Workshop on Intelligent Recommender Systems by Knowledge Transfer & Learning co-located with ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 27, 2017, volume 1887 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 16–21. URL: http://ceur-ws.org/Vol-1887/paper3.pdf.

[28] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 19:1–19:19. URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[29] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2405–2414. URL: https://doi.org/10.1145/3404835.3463245. doi:10.1145/3404835.3463245.

[30] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, V-elliot: Design, evaluate and tune visual recommender systems, in: H. J. C. Pampín, M. A. Larson, M. C. Willemsen, J. A. Konstan, J. J. McAuley, J. Garcia-Gathright, B. Huurnink, E. Oldridge (Eds.), RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021, ACM, 2021, pp. 768–771. URL: https://doi.org/10.1145/3460231.3478881. doi:10.1145/3460231.3478881.

[31] V. W. Anelli, T. D. Noia, E. D. Sciascio, C. Pomo, A. Ragone, On the discriminative power of hyper-parameters in cross-validation and how to choose them, in: T. Bogers, A. Said, P. Brusilovsky, D. Tikk (Eds.), Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019, ACM, 2019, pp. 447–451. URL: https://doi.org/10.1145/3298689.3347010. doi:10.1145/3298689.3347010.

[32] W. Krichene, S. Rendle, On sampled metrics for item recommendation, in: R. Gupta, Y. Liu, J. Tang, B. A. Prakash (Eds.), KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, ACM, 2020, pp. 1748–1757. URL: https://doi.org/10.1145/3394486.3403226. doi:10.1145/3394486.3403226.