Explaining Learning to Rank Methods to Improve Them

Alberto Veneri[*]
alberto.veneri@unive.it
Ca' Foscari University of Venice
ISTI-CNR
Italy

ABSTRACT

State-of-the-art methods for Learning to Rank (LtR), either designed for tabular or textual data, are incredibly complex. Increasing the complexity of the models has many drawbacks, including difficulties in understanding the logic behind each prediction and a lack of trust in the system during its deployment. In this paper, which describes the author's goals during his Ph.D., there is an analysis and discussion of how we can use the ideas and tools coming from the eXplainable Artificial Intelligence (XAI) field to make the most effective methods for LtR understandable to the practitioners with the final goal of making them more efficient and/or understand better when they can be improved. The strategies adopted to achieve the aforementioned goals are different and based on the type of models analyzed, which go from more traditional LtR models based on ensembles of decision trees and using handcrafted features to fairly new neural LtR models using text data.

CCS CONCEPTS

• **Information systems** → *Language models*; • **Computing methodologies** → **Ranking**.

KEYWORDS

Explainable Artificial Intelligence, Learning to Rank, Large Language Models, Text Ranking

ACM Reference Format:
Alberto Veneri. 2023. Explaining Learning to Rank Methods to Improve Them. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583780.3616002

1 PROBLEM STATEMENT

Recently, a new research topic has emerged in the scientific literature called XAI. The research in this area aims at answering the ineludible need for AI systems to be trustworthy, fair, and understandable [7]. To be more precise, even though it gained a lot of popularity in the last few years, there has always been the need to understand the decisions made by a Machine Learning (ML) model. However, with the recent introduction of automated systems in critical application scenarios, the awareness of various stakeholders on the impact of potential undesired predictions from a model has generated a new wave of research projects and ideas on this specific subject. The need to make a model more transparent and explainable to the various stakeholders is nowadays of major importance also given the current incumbent regulations. In fact, even though it does not explicitly require an explainable model for all the possible applications, the new European Regulation called *AI Act* asks for more transparency and accountability of ML models and their development process, which can also be achieved using techniques derived from the XAI field [15]. That is particularly true in the so-called *deep-learning* area, where Artificial Neural Networks (ANN) made by billions of parameters are used to create models to solve different types of tasks, see for example the fairly new model BLOOM, with 176 billion parameters [22]. Notwithstanding ANNs are indeed models that need further investigation to make them more trustworthy, other models can be considered *black boxes*. This is the case, for example, of large ensembles of trees that are commonly used to solve efficiently a variety of real-life challenges, and they are also very popular among practitioners [19]. In the early days of its development, XAI techniques were used principally in the Computer Vision field, but now it is spreading across other research fields, including the Information Retrieval (IR) world, creating a subfield that Anand et al. called Explainable Information Retrieval (EXIR) [2].

In particular, we focus on a specific task of IR, that is the ranking process, where seminal works have been proposed to explain black-box models, e.g., [20, 21, 25], or to create new intrinsically explainable ranking algorithms, e.g., [26]. However, the research on this trend is still in its early days, and new solutions have to be found to make the current algorithm for ranking more fair, trustworthy, and accountable. In our research agenda, we are particularly interested in explaining complex LtR models to improve them in terms of efficiency or understanding in which settings they are more prone to fail and thus improve their effectiveness. In our work, we use the term LtR with the same meaning used by Bruch et al. [3], that is, without differentiating with respect to model learned using handcrafted features (usually numerical) or text. Previous works present in the literature are more focused on giving easy-to-understand explanations to end-users to justify the predictions made by a ML model (*explanation user-centric*), while we are more interested in giving an exact explanation of the system to be used by experts (*explanation system-centric*). Thus, to summarize, our previous and expected contributions focus on providing explanations of the models' predictions that are useful for practitioners who want to debug and improve the effectiveness and efficiency of their models.

[*] Supervised by Claudio Lucchese (Ca' Foscari University of Venice).

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3616002

2 STATE OF THE ART

Following the categorization proposed by [2], we can divide the EXIR field into three main lines: i) post hoc explanations, ii) explanation grounding to IR properties, iii) intrinsically interpretable methods.

Post hoc explanations. One of the most common and most used methods to explain predictions of ML models in IR and other ML fields is to assign an importance score to each feature of an input instance. In this category of explanations, we can cite methods like EXS [20], LIRME [21], and DeepSHAP for Neural Information Retrieval (NeuIR) models [6]. Even though this category of explanations can be useful in different use-case scenarios, we should consider it a "shallow" explanation because it does not allow the user to understand the underlying reasoning of the model, but it only gives the end-user a score of the importance of certain features, leaving unresolved the question of "why" a model has arrived to give more importance to a document with respect to another. Another approach to the problem is to find adversarial examples, such as in [23]. The adversarial attacks on a LtR model can be really useful in understanding its pitfalls and improving its robustness. However, there is usually a goal misalignment between explanation and adversarial attacks since the latter normally try to find only a subset of samples in which the model fails, while they are usually less focused on improving the model as a whole. Finally, there are also free-text explanations that, given a black-box model, provide an explanation in natural language, e.g., [25, 24]. However, these methods can be considered more useful for end-users rather than ML developers and thus out of the scope of our research project.

Explanation by proving IR axioms. Another type of explanation specific to the IR field is the explanation by proving IR axioms. In this case, numerous studies have tried to check the adherence of NeuIR models to certain well-known IR axioms. Examples of such work are the works made by Rennings et al. [16] and the work made by Câmara and Hauff [5]. Even though this type of work wants to answer the same question that we address, i.e., what is the underlying reasoning of the model to predict that a document is more relevant for a given query, they are limited since they check only some human-created axioms. We claim, and it has also been partially proved by [5], that such axioms are good for humans to explain the relevance of a document concerning a query, but in general, what applies to humans does not directly apply to machines.

Intrinsically interpretable methods. Finally, there is a branch in EXIR that aims to create models that are easy to be explained to the end user and to the practitioners. They are usually called *intrinsically interpretable* or *white-box models*, such as [26].
Among the various categories, our research project fits better into the first and last category, i.e., in finding post-hoc explanations for complex ML models and the development of intrinsically interpretable methods. To be more precise, the main objective of our works is to make models that are easy to understand but also to "open up" the state-of-the-art black-box models to make them more interpretable.

3 APPROACH AND METHODOLOGY

Given the current state of the art, we defined two main research questions that are distinct but at the same time bounded together:

RQ1. Can we explain the function learned by a LtR model and consequentially improve its effectiveness?
RQ2. Can we explain the function learned by a LtR model and consequentially improve its efficiency?

With these research questions, we think that we tackle the explainability problem from another point of view with respect to the current literature, which can be summarized with only one question: what do we need to understand better from the current ML models used for LtR to be able to improve them?
Answering the two questions above is a hard task. Thus we planned our research agenda with several subquestions to solve both main questions. A small fraction of the most significant subquestions can be summarized as follows:

RQA. How do we measure interpretability for our specific use cases, i.e., improve the model in terms of efficiency and effectiveness?
RQB. Can we reduce the complexity of the function learned by a model for LtR and make it more interpretable, or is it better to create a model explainable by design?
RQC. What is the tradeoff between effectiveness (efficiency) and explainability in LtR models? Can we define it?

We highlight again that the research questions provided above are different from the basic goal behind most of the current explainability works, which is normally seen as a tool to explain the prediction made by the model to the end-user and is usually assumed to have no or low AI literature. In our case, we assume that the final user of the explanation is an expert and, with the right knowledge, is able to improve the model in one of the most crucial aspects in IR, which are effectiveness and efficiency. This is consistent also with one of the main goals of XAI defined by Adadi e Barreda, i.e., *explain to improve* [1].

Regarding our methodology to answer the above questions, we have decided to follow two main research threads: *i)* Create new learning algorithms for decision tree ensembles to make them more easily explainable *ii)* Zoom-in in the newly available Large Language Models (LLMs) to make them less opaque.

In addition, with regard to the methodology aspect, we highlight that we presented a unified theoretical framework for explainability [17], that even though it has been proposed to be applied in another scientific field (bioinformatics) can be easily adapted to IR. From the framework, the idea that is worth emphasizing is that there is no clear distinction in what makes a model a *white box* or a *black box*, and in principle all the models that are now considered black box can be, with the right explanation mechanism or the right distillation technique, considered white box in the future. That is why the two research threads have different starting points. On the one hand, we want to implement a model that is explainable by design, which is claimed to be the unique way to develop ML models suitable for high-stake decision scenarios [18]. On the other hand, we try to use a bland mechanistic interpretation [14] of so-called black-box models to make them interpretable and then improve them.

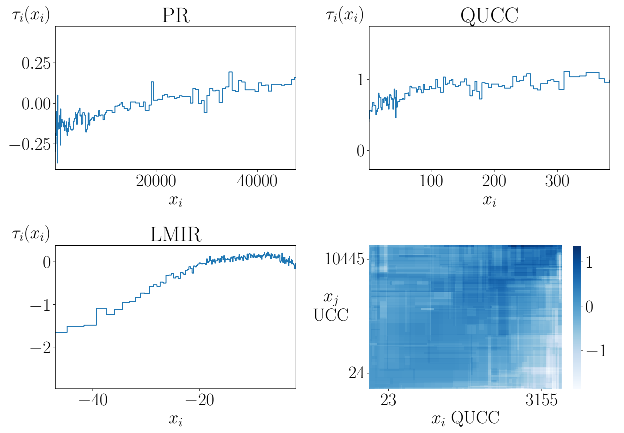

Figure 1: Example of univariate and bivariate functions presented in ILMART[11]. The acronyms of the handcrafted feature used are: PageRank (PR), Query-url Click Count (QUCC), URL Click Count (UCC), Language Model for Information Retrieval (LIMIR).

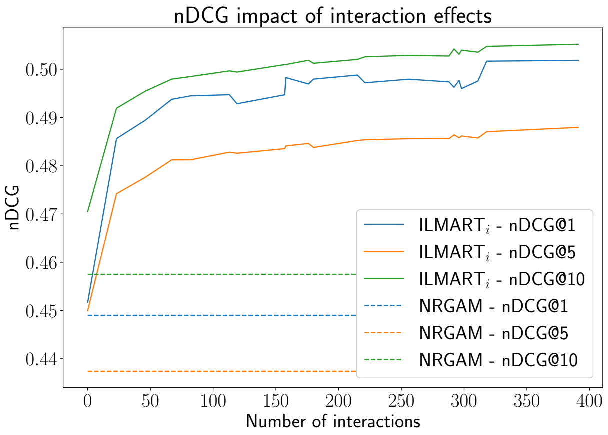

Figure 2: Initial analysis of interpretability effectiveness tradeoff presented in ILMART [11] analyzing the impact of the number of interaction effects used by the model.

4 RESULTS

Our analysis started with investigating the most common and accurate models used for the LtR task when handcrafted features are involved, which are ensembles of decision trees. Our first approach focused on explaining the prediction made by a forest and creating a surrogate Generalized Additive Model (GAM) representing the global behavior of the model without using the initial training dataset [12]. This was the first attempt at creating a *post hoc explanation* for ensembles of decision trees. Even though the results have been presented for the more classical classification and regression tasks, the same explanation strategy can be used for one of the methods based on an ensemble of decision trees currently state of the art for the LtR task, like LambdaMART [4]. This initial investigation was an initial answer to RQB, and it provided us the starting point to develop Interpretable LambdaMART (ILMART) [11], a novel learning method derived from LambdaMART that provides at the end of the learning phase a model that is a sum of univariate and bivariate functions (also called as main and interaction effects) following the equation:

$$\hat{y} = \sum_{i=1}^{N} \tau_i(x_i) + \sum_{i=1,j=1}^{i=N,j=N} \tau_{ij}(x_i, x_j)$$

Where $\hat{y}$ is the prediction made by the model, N is the number of the feature of the dataset used, τ_i is a univariate function, and τ_{ij} is a bivariate function, and the number of non-zero τ_i and τ_{ij} is controlled during the learning phase. Being such a simple function, the model can be analyzed by the practitioners in a simple way by plotting the individual functions as shown in Figure 1, where univariate functions are presented with a line plot and a bivariate function is presented with a heatmap. Considering only the functions presented in Figure 1, given the feature values, the final prediction of the model can be easily computed by summing the four functions. In this way, the model can be considered *intrinsically interpretable*. In [11], we showed that the model is similar to a GAM and outperforms by a large margin, with an nDCG gain up to 8%, the most similar competitor, NeuralRankGAM [26].

Starting from this work, we moved towards RQC to better analyze the interpretability and effectiveness tradeoff of a model like ILMART. The initial results were also published in [11], and are proposed again in Figure 2 where we showed that the number of τ_{ij} used in the final model has a huge impact on the final prediction accuracy. In a still unpublished extension of the work, we analyzed different strategies to select the minimum number of τ_i and τ_{ij}, and we showed that it is possible to achieve good results using a very limited number of main and interaction effects. Indeed, the lower the number of τ_i and τ_{ij} functions used, the greater the explainability of the model (RQA). In the same analysis, we also showed that it is possible to make a very efficient representation of the model and being able to achieve performance comparable to state-of-art methods for efficient scoring of ensembles of decision trees, such as [10] (RQ2). The study presented in [11] and its extension concludes our first planned analysis of LtR methods based on ensembles of decision trees and using handcrafted features.

Our current ongoing work is related to the explainability of LtR models using text data. In our preliminary and unpublished study, we saw that the best way to create an explainable model that is also accurate is to distill an already trained large model, such as MonoBERT [13], using in the distillation process the knowledge about the model behavior gained from the explanation of its predictions. One way that seems promising is to analyze the embedding space created and check possible recurrent and interesting patterns [9][1]. A lot of work has to be done in this direction, but the preliminary results show that there are interesting correlations between the geometrical properties of the space created and the

[1] Post-acceptance note: short paper just accepted for publication at CIKM 2023, the description of the method was at a high level to preserve anonymity.

predictions made that can be exploited. Therefore, we plan to continue in this direction and try to distill a simpler function from the complex one learned by a LLM used for ranking.

5 CONCLUSION AND FUTURE WORKS

In this research project, our main goals are to understand and approximate the function learned by the current state-of-the-art ML models used for LtR and to design new explainable by-design models that are competitive with the state-of-the-art. In our works, the explanations of a model are intended for practitioners and developers who want to improve the accuracy and/or the efficiency of their models, creating a feedback loop that can be iterated continuously to enhance the developed systems.

On the one hand, the results obtained in the scenario where numerical features are used are promising and show that it is possible to improve the efficiency of a LtR model by simplifying the function learned, i.e., making them also more interpretable, giving a partial positive answer to RQ2. On the other hand, the tradeoff between model complexity and effectiveness seems difficult to manage, so improving the model accuracy (RQ1) only from the explanation provided by the model seems difficult, and the only way to improve the effectiveness is usually just by increasing the complexity.

Regarding the scenario where text data is used, there is still a lot of work to do to understand the inner workings of current ML models, especially if we consider the now widely used LLMs. Try to understand and simplify the functions learned by LLMs is hard due to the enormous amount of parameters used, but it is essential to make the ground of a new area of LtR models, where we move from the "alchemy" [8] of tweaking part of the models in order to achieve better performance towards a more in-depth understanding of their inner working. Our work in this direction started by analyzing the geometry behind the hidden space defined by the model, finding patterns of evolution of the embeddings passing through the transformer blocks. If a good explanation of the functions learned is found, the following steps regarding the improvement in terms of efficiency and effectiveness become easier.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). en. *IEEE Access*, 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052.

[2] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable Information Retrieval: A Survey. (Nov. 2022). doi: 10.48550/arXiv.2211.02405.

[3] Sebastian Bruch, Claudio Lucchese, and Franco Maria Nardini. 2023. Efficient and Effective Tree-based and Neural Learning to Rank. English. *Foundations and Trends® in Information Retrieval*, 17, 1, (May 2023), 1–123. Publisher: Now Publishers, Inc. doi: 10.1561/1500000071.

[4] Christopher J C Burges. [n. d.] From RankNet to LambdaRank to LambdaMART: An Overview. en, 19.

[5] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. en. In *Advances in Information Retrieval* (Lecture Notes in Computer Science). Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, (Eds.) Springer International Publishing, Cham, 605–618. isbn: 978-3-030-45439-5. doi: 10.1007/978-3-030-45439-5_40.

[6] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the Interpretability of Neural Retrieval Models using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'19). Association for Computing Machinery, New York, NY, USA, (July 2019), 1005–1008. isbn: 978-1-4503-6172-9. doi: 10.1145/3331184.3331312.

[7] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. en. *Science Robotics*, 4, 37, (Dec. 2019). doi: 10.1126/scirobotics.aay7120.

[8] Matthew Hutson. 2018. Has artificial intelligence become alchemy? *Science*, 360, 6388, (May 2018), 478–478. Publisher: American Association for the Advancement of Science. doi: 10.1126/science.360.6388.478.

[9] Claudio Lucchese, Giorgia Minello, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2023. Can embeddings analysis explain large language model ranking? en. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM, (Oct. 2023). doi: 10.1145/3583780.3615225.

[10] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. 2015. Quickscorer: a fast algorithm to rank documents with additive ensembles of regression trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 73–82.

[11] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2022. ILMART: Interpretable Ranking with Constrained LambdaMART. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '22). Association for Computing Machinery, New York, NY, USA, (July 2022), 2255–2259. isbn: 978-1-4503-8732-3. doi: 10.1145/3477495.3531840.

[12] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2023. GAM Forest Explanation. en. (2023). doi: 10.48786/EDBT.2023.14.

[13] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. arXiv:1910.14424 [cs]. (Oct. 2019). doi: 10.48550/arXiv.1910.14424.

[14] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: an introduction to circuits. *Distill*. doi: 10.23915/distill.00024.001.

[15] Cecilia Panigutti et al. 2023. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, New York, NY, USA, (June 2023), 1139–1150. doi: 10.1145/3593013.3594069.

[16] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. en. In *Advances in Information Retrieval* (Lecture Notes in Computer Science). Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, (Eds.) Springer International Publishing, Cham, 489–503. isbn: 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_32.

[17] Matteo Rizzo, Alberto Veneri, Andrea Albarelli, Claudio Lucchese, Marco Nobile, and Cristina Conati. 2023. A theoretical framework for ai models explainability with application in biomedicine. (2023). arXiv: 2212.14447 [cs.AI].

[18] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. en. *Nature Machine Intelligence*, 1, 5, (May 2019), 206–215. doi: 10.1038/s42256-019-0048-x.

[19] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. en. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, 4, (July 2018). doi: 10.1002/widm.1249.

[20] Jaspreet Singh and Avishek Anand. 2019. EXS. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, (Jan. 2019). doi: 10.1145/3289600.3290620.

[21] Manisha Verma and Debasis Ganguly. 2019. Lirme: locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'19). Association for Computing Machinery, Paris, France, 1281–1284. isbn: 9781450361729. doi: 10.1145/3331184.3331377.

[22] BigScience Workshop et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 [cs]. (June 2023). Retrieved July 5, 2023 from http://arxiv.org/abs/2211.05100.

[23] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023. PRADA: Practical Black-box Adversarial Attacks against Neural Ranking Models. en. *ACM Transactions on Information Systems*, 41, 4, (Oct. 2023), 1–27. doi: 10.1145/3576923.

[24] Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards Explainable Search Results: A Listwise Explanation Generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '22). Association for Computing Machinery, New York, NY, USA, (July 2022), 669–680. isbn: 978-1-4503-8732-3. doi: 10.1145/3477495.3532067.

[25] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query Understanding via Intent Description Generation. en. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, Virtual Event Ireland, (Oct. 2020), 1823–1832. isbn: 978-1-4503-6859-9. doi: 10.1145/3340531.3411999.

[26] Honglei Zhuang et al. 2021. Interpretable ranking with generalized additive models. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (WSDM '21). Association for Computing Machinery, Virtual Event, Israel, 499–507. isbn: 9781450382977. doi: 10.1145/3437963.3441796.