XAI for Ranking

Tiantian Luo and Duanran Jing

University of Zurich, Switzerland

Abstract

Machine learning plays an important role in people's daily life, but their black-box nature makes people less convinced by their results. XAI (Explainable Artificial Intelligence) provides a solution to this problem by generating an understandable explanation for the result from them. In addition to the traditional XAI aiming for classification and regression tasks, XAI for ranking tasks has been gaining increasing attention. However, the exploration of the role visualization and interaction plays in the explanation is still limited. In our study, we visualized the explanation generated by XAI for ranking using bar charts and designed four different visualizations based on that, including static and interactive visualizations. A user study is conducted to evaluate the effects of different visualizations on understanding XAI. We found that interactions help users identify features faster, and interactive visualizations make people more confident to explain the ranking, but interactions do not always provide more content.

1. Introduction

XAI stands for Explainable Artificial Intelligence. It refers to the development and deployment of AI systems that are capable of providing understandable explanations for their decisions and actions. Traditional AI models, such as deep learning neural networks, often operate as black boxes, making complex calculations and generating predictions without providing transparent reasoning behind their outputs. This lack of interpretability poses challenges, especially in critical domains where trust, transparency, and accountability are crucial [DR20].

Researchers and practitioners in the field of XAI have developed various techniques and approaches to address these challenges. These techniques range from simple post-hoc interpretability methods, such as feature importance analysis or rule extraction, to more complex model-agnostic methods such as LIME (Local Interpretable Model-agnostic Explanations) [RSG16] and SHAP (Shapley Additive Explanations) [LL17] [LPK20].

Compared with these XAI methods mainly designed for classification and regression tasks, XAI for ranking tasks are more challenging and more unexplored. Traditional XAI methods like LIME and SHAP can not solve ranking tasks adequately because they can be viewed as aggregations of multiple predictions. [SKZA21] Given the situation, it is needed to explore the explanation for ranking tasks.

Unlike the previous studies in this field, which mainly concentrate on proposing new explanations and evaluating existing XAI methods, our study focuses more on the role of visualization and interaction plays in explanations for ranking tasks.

In this paper, our research questions are:

- Do the differences in visual design result in different levels of understanding of the XAI method?
- What type of interaction in visualization helps most to understand the XAI method?

At the same time, we also evaluated how different visualizations help in identifying features and collected people's opinions on how and in which way visualizing data can help them better understand XAI effectively.

In summary, based on the explanation generated by XAI for ranking tasks, we designed and built an interactive visualization interface to display the explanations for happiness countries ranking from World Happiness Report, and conducted a user study to evaluate different interactive visualizations.

As a result, it turns out that interactions help people identify features faster. Compared to static visualization, interactive visualizations make people more confident to explain the ranking in general. Especially, the sorting function provides the most content and the two-country comparison makes people most confident in explaining the ranking.

2. Related Work

While the traditional XAI method for classification or regression is well-developed, the explanation created for ranking tasks is still relatively unexplored, and little work has focused on the interpretability of the ranking model [ZWB*21]. Since the working principle of a ranking task is different from classification and regression, and the explanation for ranking has to provide a reason for the relationship between different items [SA19], the traditional XAI methods are not suitable to be directly applied for ranking tasks [VG19].

Recently some explanations for the ranking problem have been proposed, [ZWB*21] introduced generalized additive models into ranking tasks, and Verma et al [VG19] and Singh et al [SKZA21] focus on generating local explanations of Learning-to-Rank (LTR) models, which aim to extract the most important features that contribute to the ranking predicted by the LTR model [RB22].
Some of the previous studies provide visualization as part of explanation [SA19], however, to the best of our knowledge, there is limited work focusing on the effect of visualization on the interpretability of explanation. In our study, we focused on exploring the effect of visualizations in XAI for ranking problems.

3. XAI for Ranking

3.1. XAI Method

To process the ranking data, the main idea is to first apply an LTR model to the ranking dataset, and the XAI method is used to generate explanations for the ranking predicted by the LTR model. In our study, we applied XGBRanker from XGBBoost [CG16] to perform LTR modeling and choose the XAI method proposed by [SKZA21] which generates a subset of features and the importance scores of features as explanations, specifically, the XAI method is realized by an open-source library OmniXAI [YLL*22].

3.2. Data

We started with the top hit songs ranking provided by Spotify. Spotify dataset has some basic features like danceability, speechness, length of the song, and so on. But later on, we realized that a very important feature, the social media effect, that will influence the ranking of a song magnificently, is not quantified in the Spotify dataset unfortunately. In this situation, the result generated by the XAI method is not that convincing.
The dataset we use in our study is the country ranking of happiness from The World Happiness Report 2015-2019 [kag]. Each year's ranking consists of 50 countries and each instance has corresponding 6 features to describe its characteristics, which are Healthy life expectancy, GDP per capita, Freedom, Generosity, Corruption, and Social Support.

4. Visualizations and Interface Design

The basic component of the visualization is the bar chart. A single bar represents a feature, and since each country is described by 6 features, correspondingly a country is represented by a bar chart consisting of 6 bars. The XAI method generated a subset of features and importance scores of features as the original explanation, and in our design, the different features are encoded by colors, and the value of each feature is represented by length. In addition, the importance level is encoded by the height of the bars, the taller the bar, the more important role the feature plays in the explanation.
Four different visualizations are provided in our interface as shown in Figure 2. Visualization 1 2a is static visualization. Based on Visualization 1, we provided three different interactions: in Visualization 2 2b, when users hover over a single bar, the corresponding feature's bars will be highlighted in all the countries' charts and the exact value of this feature in each country will be displayed. In Visualization 3 2c, when the users click on a single bar, all countries will be ordered based on the value of the feature users clicked. In Visualization 4 2d, users are able to select 2 countries to compare. We present the ranking of 20 countries on one single page, and users are able to navigate to different interaction pages by the selection button on the interface.

5. User Study

In order to test the effect of different interactions, we conducted a user study. Participants need to play with visualizations and finish the tasks we designed.

5.1. Participants

5 Participants are recruited for our study. All of the participants have a computer science background.

5.2. User Study Procedure

During the user study, we have two major focuses: How the different interactions help users finish certain tasks related to identifying features and the effect of different interactive visualizations on helping them understand the XAI.

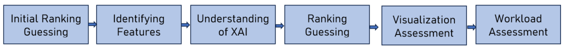
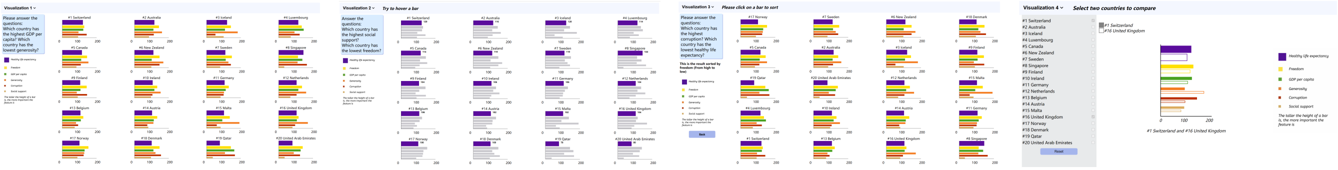


Figure 1: User study procedure

The procedure of our user study is described in Figure 1. After we introduced the basic information of our study, we asked the participants to have an initial ranking guessing about which country has the higher ranking in three pairs of anonymous countries based on the static visualization only. When ready for further steps, participants were asked to find out two countries that meet the requirements about certain features in the first three visualizations, and we measured the time they used to finish the tasks and the accuracy of the answers. In this part, we wanted to evaluate the effects of different visualizations on helping users identify features.
Then, we let the participants interact with each visualization respectively, and we conducted a semi-structured interview after they finished interacting with each interaction. The interview focused on the assessment of their understanding of XAI, for example, their confidence level of understanding the explanation for ranking, how much information is provided in visualization, if they observed any pattern from the visualizations, and they were asked to briefly explain how the ranking is generated to us. We collected their textual answers and used a 5-point Likert scale for their confidence level and the information provided. In this part, we focused on the effects of helping users understand the explanation generated by XAI.
After completing the exploration of all four visualizations and interactions, the participant had the same rank-guessing quiz again consisting of the same three pairs of countries as in the beginning to see if they wanted to modify their answers after exploration. Finally, we conducted a semi-structured interview to ask for their opinions and suggestion about our interactive visualizations. The participant then completed a 5-point scale NASA TLX questionnaire.



Visualization 1 ⌄

Please answer the questions: Which country has the highest GDP per capita? Which country has the lowest generosity?

Visualization 2 ⌄ *Try to hover a bar*

Answer the questions: Which country has the highest social support? Which country has the lowest freedom?

Visualization 3 ⌄ *Please click on a bar to sort*

Please answer the questions: Which country has the highest corruption? Which country has the lowest healthy life expectancy?

This is the result sorted by freedom (From high to low)

Visualization 4 ⌄ *Select two countries to compare*

(a) Visualization 1

(b) Visualization 2

(c) Visualization 3

(d) Visualization 4

Figure 2: Interface

6. Result

6.1. Identifying Features

To evaluate how different interactive visualizations help users identify features, we asked them to find out the highest and lowest value of 2 features in each visualization. In this part of the study, only visualization 1 (static visualization) 2a, visualization 2 (highlighting with hover) 2b, and visualization 3 (sorting by one feature) 2c are involved. In each visualization, different features are asked to prevent users from memorizing the answers.

The result is shown as follows. Figure 3 shows the average time (in seconds) for finishing the tasks and Figure 4 shows the correctness rate for the tasks. From the results, we can observe that in general, with the help of interaction, users spent less time finding the answer and sorting makes it easy to answer the question correctly.

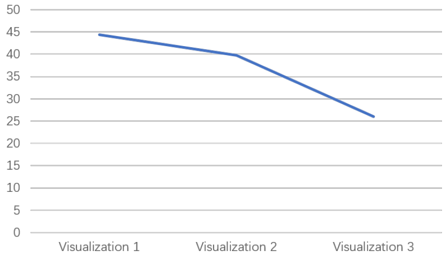


Figure 3: The time spent for finishing tasks in each visualization

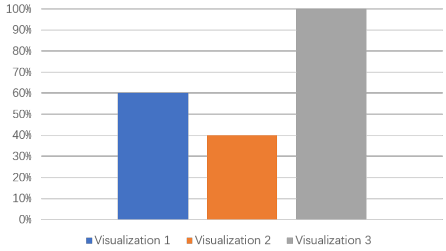


Figure 4: The correctness for finishing tasks in each visualization

Interestingly, we observed there are two types of participants. Some participants spent less time doing the task in Visualization 2 and then spent even less time doing the task in Visualization 3. However, some participants spent more time in Visualization 2 but significantly less time in Visualization 3. In our observation, this difference appeared maybe because, in Visualization 2, some users need more time on reading instructions and understanding how the interaction works.

6.2. Understanding XAI

In general, the visualization of the XAI for the ranking method is able to explain the ranking well under the condition that there is no outlier. Firstly, all visualizations were rated as easy to understand, and most participants stated they understand healthy life expectancy is the most important feature, which is the explanation supposed to represent. Moreover, given basic information about static visualization, the user achieved a 90% accuracy rate in both rank-guessing quizzes when there is no outlier. However, according to the feedback from our interviews and the decreasing accuracy on quiz question including outlier case, participants found difficulties to decide which country has a higher ranking when an outlier is involved, even after fully interacting with the visualizations.

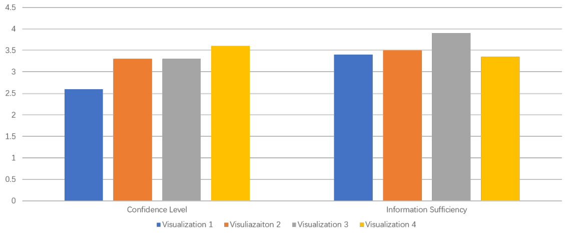


Figure 5: Average scores on participants' confidence level and information sufficient

The preference for visualization and interaction is subjective, however, in general, users gained more confidence through interactions based on the increasing average score of the answers, as shown in Figure 5. Visualization 4 provides the function to compare two items separately, which makes users feel most confident to explain the ranking if they were asked to do so. Interaction does not always provide more content, but we observed the scores about information sufficiency of the first three visualizations are increasing, which shows that interaction can provide more content when the fundamental layout stays the same. Visualization 3 with the sorting function provides the most content.

Moreover, we observed that the outliers are sometimes easier to detect with visualization, even though outliers can confuse users, make them feel less confident, and doubt the pattern they observed. Besides, unclear labeling also caused confusion in understanding XAI. For example, participants were not sure about the definition of Generosity and Corruption, and they tended to assume Corruption has a negative effect on the ranking of happiness, which introduced an additional workload for them to process the information.

6.3. Visualization Assessment

Overall, users are satisfied with the interactive visualizations we presented. The main problem they had is that they had difficulties understanding Corruption, which caused a lot of confusion. Also, users think there is too much information on one page because each country has 6 bars and there are 20 countries in total, which leads to too many colors and bars on a page.

Based on the interactive visualizations we presented, it will be more helpful if we also present value in sorting, or if multiple countries comparison is supported. Users also wanted more variety in visualizations like radar charts, parallel coordinate charts, and bubble charts. To understand XAI better, it will be nice if the ranking score is provided or if we could give more information about the partial importance of each feature, like if GDP per capita increases by one unit, how much the ranking will increase.

6.4. NASA TLX Results

Since there is no physical work involved in our study, we removed the questions about physical workload. From the results, we concluded that users did not experience too much mental, and temporal workload with average scores around 3. They are fine with their performance with an average score of 3 and spent a suitable amount of effort with an average score of 3 and did not suffer from frustration with an average score of 1.6.

7. Threats to Validity

In the paper, the main concerns are acquiescence bias and randomized problems. First of all, all the participants were recruited from our social circle so that they may give us the "good answer" instead of the "true answer". In this case, it can lead to bias in the result. Secondly, in identifying features part, the tasks for every participant are in the same order. There is still a tiny chance that participants may already know and notice the answer to the next question. If we randomized the order of the tasks for each participant, then we could fully rule out the problem of knowing the answer in advance. Additionally, we had a relatively small dataset. The World Happiness Report changed the features they measured before 2015 and after 2019. To keep the dataset consistent, we had to choose to use the data from 2015-2019. Also, the number of participants is relatively small, which may also cause bias in the result of the study.

8. Conclusion

This study mainly aims at evaluating how interactive visualizations help people understand XAI. From the user study, it turns out that in static visualization with important information properly encoded, people can already determine ranking with a high accuracy rate, if there is no outlier. Interactions like highlighting or sorting can help people identify features more efficiently. Regarding understanding XAI, the sorting function can provide the most content to people, but interactions do not always provide more content itself, and interactions perform well to help users increase their confidence in understanding XAI, especially the two-country comparison makes people more confident to explain how the ranking is generated themselves.

By conducting the study, we realized visualization can help people detect outliers easier sometimes. The existence of outliers can cause confusion among people. Also, when visualizing data, unclear labeling or people's stereotypes can cause confusion and make it more difficult for people to understand the ranking. More attention should be paid to future work to avoid this problem.

There is still much potential in understanding XAI by interactive visualization. Potential work in the future can be focused on providing more interactions, more information, or more different types of charts and investigating if they can help understand XAI better.

References

[CG16] Chen T., Guestrin C.: Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794. 2

[DR20] Das A., Rad P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020. arXiv:2006.11371. 1

[kag] World Happiness Report — kaggle.com. https://www.kaggle.com/datasets/unsdsn/world-happiness. [Accessed 25-Jun-2023]. 2

[LL17] Lundberg S. M., Lee S.-I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017). 1

[LPK20] Linardatos P., Papastefanopoulos V., Kotsiantis S.: Explainable ai: A review of machine learning interpretability methods. *Entropy 23*, 1 (2020), 18. 1

[RB22] Rahnama A. H. A., Butepage J.: Evaluating local model-agnostic explanations of learning to rank models with decision paths. *arXiv preprint arXiv:2203.02295* (2022). 2

[RSG16] Ribeiro M. T., Singh S., Guestrin C.: "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (2016), pp. 1135–1144. 1

[SA19] Singh J., Anand A.: Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (2019), pp. 770–773. 1, 2

[SKZA21] Singh J., Khosla M., Zhenye W., Anand A.: Extracting per query valid explanations for blackbox learning-to-rank models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (2021), pp. 203–210. doi:10.1145/3471158.3472241. 1, 2

[VG19] Verma M., Ganguly D.: Lirme: locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019), pp. 1281–1284. 1, 2

[YLL*22] Yang W., Le H., Laud T., Savarese S., Hoi S. C. H.: Omnixai: A library for explainable ai, 2022. arXiv:2206.01612. 2

[ZWB*21] Zhuang H., Wang X., Bendersky M., Grushetsky A., Wu Y., Mitrichev P., Sterling E., Bell N., Ravina W., Qian H.: Interpretable ranking with generalized additive models. In

Proceedings of the 14th ACM International Conference on Web Search and Data Mining (2021), pp. 499–507. 1, 2