Local Point-Wise Explanations of LambdaMART

Amir Hossein Akhavan Rahnama [1], Judith Bütepage [1] and Henrik Boström[1]

***Abstract*—LambdaMART has been shown to outperform neural network models on tabular Learning-to-Rank (LTR) tasks. Similar to the neural network models, LambdaMART is considered a black-box model due to the complexity of the logic behind its predictions. Explanation techniques can help us understand these models. Our study investigates the faithfulness of point-wise explanation techniques when explaining LambdaMART models. Our analysis includes LTR-specific explanation techniques, such as LIRME and EXS, as well as explanation techniques that are not adapted to LTR use cases, such as LIME, KernelSHAP, and LPI. The explanation techniques are evaluated using several measures: Consistency, Fidelity, (In)fidelity, Validity, Completeness, and Feature Frequency (FF) Similarity. Three LTR benchmark datasets are used in the investigation: LETOR 4 (MQ2008), Microsoft Bing Search (MSLR-WEB10K), and Yahoo! LTR challenge dataset. Our empirical results demonstrate the challenges of accurately explaining LambdaMART: no single explanation technique is consistently faithful across all our evaluation measures and datasets. Furthermore, our results show that LTR-based explanation techniques are not consistently better than their non-LTR-based counterparts across the evaluation measures. Specifically, the LTR-based explanation techniques consistently are most faithful with respect to (In)fidelity whereas the non-LTR-specific approaches are shown to frequently provide the most faithful explanations with respect to Validity, Completeness, and FF Similarity.**

I. Introduction

Learning-to-Rank (LTR) is an important application for machine learning. In LTR, algorithms learn to order documents (or sometimes called items) in an optimzied way based on their relevance to user queries [1]. LTR applications are omnipresent in our daily lives: online advertising, e-commerce, etc.

As the size and complexity of Learning-to-Rank (LTR) datasets increase, the LTR models are becoming more complex [2]. The LambdaMART model [1], a pairwise Gradient Boosting Tree model for Learning-To-Rank (LTR) tasks, is a powerful technique that has been shown to outperform neural ranking models for tabular data [3, 4]. While shallow decision trees can be interpretable under certain circumstances [5], ensemble boosting tree models, such as LambdaMART, often include hundreds of trees and are therefore considered black-box models [6]. In order to deploy such black-box models in real-world domains and gain the trust of users, it is vital that the logic behind the prediction of these complex models is revealed [7, 8].

Explanation techniques fill this gap by providing information about the decision-making process of complex black-box machine-learning models. Explanations can be local or global. When explanations are provided about the prediction of a single instance, they are called local explanations, and when the information is about the entire dataset, they are called global explanations. Explanation techniques represent their information in different representations. One of the most popular representations of local explanation is feature attribution, in which importance scores are allocated to features that explain their contribution to the prediction of the explained instance [9, 7]. Feature attribution-based explanation techniques can be model-agnostic, where they make no assumptions about the internal logic of the black-box model and can consequently explain the prediction of any class of machine-learning models. Due to their flexibility, these types of explanation techniques, local model-agnostic local explanations, are popular and are the focus of our study. For more details on different categories of explanation techniques, see [10].

For explaining LTR models, local model-agnostic explanations can be either point-wise or list-wise (Figure 1). Point-wise explanations provide scores that show the importance of features to the predicted output of the black-box LTR model separately for every single document in a given query [11, 12]. In contrast, list-wise explanations provide scores that explain the predicted output of black-box LTR models for a list of documents given a single query [13, 14].

Point-wise and List-wise explanations have different use cases [15, 12]. Let us consider a use-case for the point-wise explanations. An LTR model is trained to provide a list of relevant songs to a user search query in a music streaming app. The user inputs a search query, "Drake Love," and observes that the song "Love All" by Drake has received a surprisingly low predicted relevance score. We can understand what features contributed to this surprising prediction by obtaining point-wise explanations for query document pair (Drake Love, Love All). Now, we can consider a use-case for list-wise explanations. The user inputs a search query "Hotel Stockholm" and finds a list of hotels in Hornsgatan (a famous street in Stockholm) that have received surprisingly low relevance scores. We can understand the underlying contributing features for those surprising relevance scores by obtaining a list-wise explanation. Using the explanations allows model users and developers to adjust such wrongful predictions by feature scaling, de-biasing, adding interaction terms between features, or even re-training the model [16, 17]. In this study, we focus on point-wise explanations of LTR models[1].

Local explanations have a lot of potential, but there is a caveat associated with them: their evaluation. The challenge is that the ground truth importance scores cannot be directly extracted from the complex black-box models [18, 19, 20, 21]. However, several measures for evaluating local explanations have been proposed in the literature [13, 14, 22], which we use

[1] KTH Royal Institute of Technology. Corresponding Author: amiakh@kth.se

[1] For brevity, we may refer to local model-agnostic point-wise explanations as simply explanations in our study.

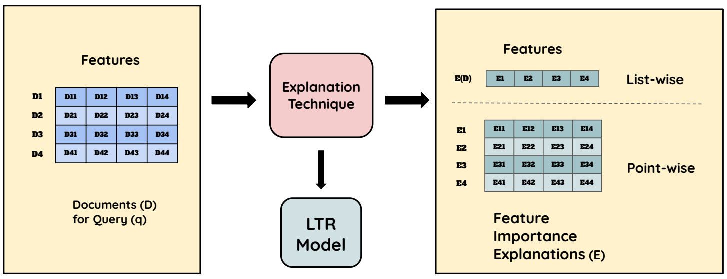


Fig. 1: List-wise $E(D)$ and Point-wise Explanations E_i for D_i ($i = 1, ..., 4$) have different interpretations and utility. Point-wise explanations show us the importance of features for the change in the predicted rank of each document D_i **separately**. In contrast, list-wise explanations show us the importance of features for the list of documents D. It is possible to aggregate the point-wise explanations for all documents D_i to obtain list-wise explanations.

in our study. Explanation techniques cannot excel in providing faithfulness without extensive and rigorous evaluation studies since it has been shown that they can fail in providing faithful explanations [23, 24, 25].

We have noticed several gaps in the literature on explainability for learning to rank models. Firstly, the two LTR-specific point-wise explanation techniques, Locally Interpretable Ranking Model Explanation (LIRME) [11] and EXplainable Search (EXS) [12] are not evaluated for explaining LTR models trained on tabular benchmark. As mentioned earlier, LambdaMART is considered the state-of-the-art model on these datasets [3, 4]. Secondly, the current studies have employed a subset of the proposed evaluation measures in their studies, even though in recent years, more evaluation measures have been proposed in the literature [14, 22]: In [11], the authors LIRME is only evaluated based on Explanation Consistency consistent[2]). There are no evaluation measures available in the work of [12] for EXS explanations, and to our knowledge, no study has compared the point-wise explanations of LIRME and EXS to this date. Thirdly, the current studies have not evaluated the LTR-based explanation techniques against their non-LTR-based counterparts. Lastly, the implementations of the local explanation techniques for LTR models are not publicly available and open.

In this work, we aim to fill the above gaps. We evaluate local point-wise explanations of the state-of-the-art ranking model LambdaMART trained on tabular LTR datasets. We have adapted the two aforementioned local LTR-based point-wise explanation techniques, i.e., LIRME and EXS, to work on tabular data[3] and will compare them against non-LTR-based local explanations generated by LIME [8], KernelSHAP [26], and Local Permutation Importance (LPI) [27]. The evaluation is performed with an extensive set of evaluation measures: Completeness and validity [14], Explanation Consistency [11], (In)fidelity [22], Fidelity [13], and Feature Frequency[4]). Moreover, the evaluation will include the LTR tabular benchmark datasets of LETOR 4 (MQ2008), Microsoft Bing Search (MSLR-WEB10K), and Yahoo. Finally, to enable reproducibility, we have released the code for implementing these techniques and their evaluation in https://github.com/amir-rahnama/p_exps_lambdamart.

The main research question for the study is whether a single explanation technique can provide faithful explanations of LambdaMART based on all evaluation measures on our studied datasets. Moreover, we would like to investigate whether there is clear evidence that LTR-based explanation techniques consistently provide more faithful explanations compared to the non-LTR-specific techniques based on our evaluation measures.

The key findings from our study are: 1) No single explanation technique can provide faithful explanations of LambdaMART on all our studied dataset considering all evaluation measures. 2) LTR based explanations such as LIRME and EXS outperform the non-LTR-specific techniques with respect to the (In)fidelity metric for all datasets. 3) The non-LTR-specific techniques LIME, SHAP, and LPI outperform LIRME and EXS with respect to Validity, Completeness, and Decision Path Feature Frequency in the majority of datasets. 4) To our surprise, random explanations are most faithful based on the Fidelity metric for MQ2008 and Yahoo datasets. 5) Overall, there are large disagreement among explanations across all datasets. 6) LIME explanations tend to favor features that are used for splitting closer to the root note of trees of LamdbaMART in the Yahoo dataset.

II. Background

In this section, we first briefly introduce the point-wise local explanation techniques that we will investigate in this

[2] Explanation consistency is defined in Section II-D1

[3] The original studies have only implemented these techniques for models trained on text data. See II for more details

[4] The evaluation measures are defined in Section II-D

work. After that, we will overview the non-LTR explanation techniques of LIME, SHAP and LPI. Lastly, we provide an overview of the explanation evaluation measure.

A. Local Point-Wise Explanations

Let $X = (q, D)$ where $D \in \mathbb{R}^N$ is the list of m documents for a query q and $d_i \in \mathbb{R}^n$ the i-th document in that list. Each document is assumed to be represented by a feature vector of discrete and/or real values $d \in \mathbb{R}^M$ where M is the size of the feature vector.

Learning-to-Rank (LTR) models learn the ranking function f rank function $f : D : \mathbb{R}^{M\times N} \rightarrow \Pi^M$ from the data. The function f outputs the predicted score (rank) π_i for the i-th document. This predicted score (rank) represents its relevance to the query q. In parts of our study, we denote the predicted score of f for documents D by S or predicted ranks of f for documents D (in descending order) by R.

LTR models are optimized using point-wise, pairwise, or list-wise loss functions. Point-wise loss evaluates the relevance of individual documents to a query by comparing predicted relevance scores against true relevance scores. Pairwise Loss Function compares pairs of documents for a given query to ensure that a more relevant document is ranked higher than a less relevant one. The list-wise Loss function considers the entire list of documents for a query, optimizing the ranking of the whole list according to the relevance scores. LambdaMART is a pairwise LTR model shown to approximate list-wise objective functions [28].

A point-wise explanation technique $g : d_i \in R^M$ provides $\Phi \in \mathbb{R}^M$ where ϕ_j $(i = 1, ..., N)$ is the score of feature j that explains its importance with respect to $S(d_i)$ or $R(d_i)$ where i can take a single value between $i = 1, ..., M$.

1) LIRME: LIRME [11] is an extension of LIME explanations [8] that is adjusted for explaining learning to rank models. The current version of LIRME does not work with tabular data. Therefore, we made adjustments to suit our tabular use case. The main part of this change was the adaptation of LIRME's sampling to the interpretable quantile sampling for tabular datasets as described in [29]. This is because LIRME's original study uses interpretable sampling and representation for text datasets. We briefly overview this sampling process, but see the aforementioned study for more details.

LIRME generates its explanations by generating samples from the explained instance d. The sampling technique divides each feature into quantiles. A binary representation is created by binning the feature values of the explained document into quartiles. Each feature from the explained document receives its corresponding bin numbers to which the feature value belongs. The sampling technique then generates new samples d' based on the explained document d by randomly sampling a set of features in d. After that, a bin number is generated for each randomly selected feature. If the newly generated bin number is equal to the bin number of that feature in the explained document, then $d'_j = 1$, and otherwise, $d'_j = 0$. This process is repeated T times and the set $D' = \{d'_1, ..., d'_T\}$ is created where T is a hyper-parameter. A kernel function k weights these new samples with the explained documents. After obtaining the predictions of the black-box model f on these samples, $f(D')$, LIRME trains a Ridge surrogate model g on pairs of $(D', S(D'))$ with the following loss function:

$$\mathcal{L}(D', f(D'), k) = \sum_{j=1}^{T} k(d'_j, d)(g(d'_j) - f_{d'_j})^2 + \alpha|\Theta| \quad (1)$$

where α is the coefficient of L1 regularization. The explanations of LIRME are the weights of surrogate model g, i.e. Θ.

2) EXS: EXS [12] is a local explanation technique tailored for LTR models largely based on LIRME. Similarly to LIRME, EXS does not work with tabular data, and we made the same changes in the sampling process for LIRME to adapt EXS to tabular datasets. However, EXS differs from LIRME in two major ways. Firstly, the surrogate model is a linear SVM model. Secondly, three labeling processes are built for EXS to generate y: Score-based (S), top-K binary (B), and rank-based (R). In score-based, label equals $1 - \frac{R(d')-R(d_1)}{R(d_1)}$ where $R(d_1)$ is the rank of the top-1 document in the query we aim to explain. Top-K binary generates a label one for sample d' if its predicted rank is larger than the rank of the Top-K document for the query. In Rank-based, the label of d' is zero if its rank is less than the top-K document in the query. Otherwise, the label equals $1 - \frac{R(d')}{k}$. In the study, the top-K document, i.e., the anchor, is usually set to be among the top predicted documents [12]. EXS uses a hinge square loss or epsilon-insensitive loss function to train its surrogate, depending on the type of labeling used.

$$\begin{aligned}\mathcal{L}(D', y, k) = \sum_{j=1}^{T} k(d'_j, d) y (\max(0, 1 - \Theta^T D')\\ + (1 - y)\max(0, 1 + \Theta^T D')\end{aligned}$$

where y is the label selected depending on one of the approaches described above, and T is the sample size. The parameter of the surrogate linear SVM model g, i.e., Θ, is the EXS explanation.

B. LIME and SHAP

Even though LIME [8] and KernelSHAP [26] are not developed for explaining LTR models, they can provide point-wise explanations of LTR models by casting the problem as a regression problem.

There are some key differences between LIME and SHAP. The most significant difference is the choice of kernel function that weights the generated samples. LIME uses an exponential kernel, while SHAP uses a discrete combinatorics kernel. Moreover, unlike other techniques, LIME and SHAP use Larspath feature selection after training their surrogate model. Moreover, LIME and SHAP use Gaussian sampling instead of the quantile sampling of LIRME and EXS. In this approach, new instances are added by adding Gaussian noise with the

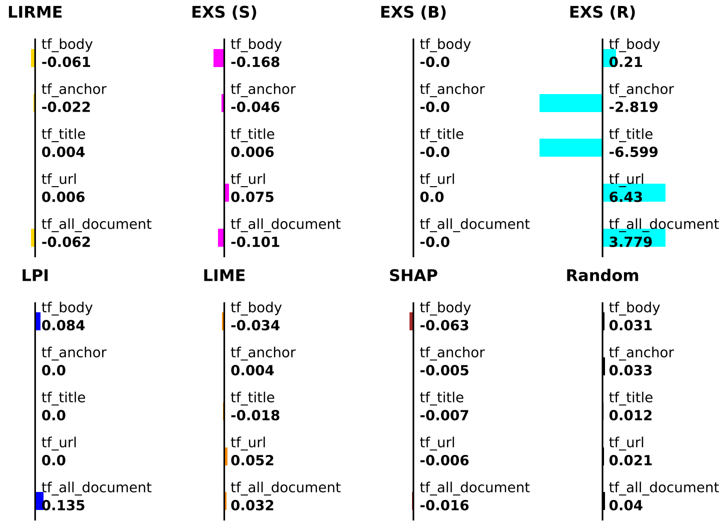


Fig. 2: Local explanation of LIRME, EXS, LIME, SHAP, and LPI of LambdaMART for a single document in the MQ2008 test dataset. The predicted relevance of the document is -0.82, ranked third among sixteen other documents for test query 18401. The explanation shows the importance scores of the first five features in MQ2008 to the predicted relevance score of -0.82.

mean adjusted to the average of each feature in the training dataset. See [9, 8] for more details.

1) LPI: Local Permutation Importance [27] is an extension of Permutation Importance [30] for obtaining local explanations. LPI does not have a surrogate model but obtains its explanations with a simple yet effective algorithm. The importance score for feature $j = 1, ..., M$ in explained document d is computed by replacing the value of that feature with other unique values of the same feature in the dataset X_j and creating d'_j. We record the change in the predicted score of the black-box model f before and after this replacement, $|f(d) - f(d'_j)|$ for T unique values of feature j in the dataset. The process is iterated for all features independently. The importance score is then calculated as $\sum_T |f(d) - f(d'_j)|/T$, i.e., the average absolute change of the predicted relevance scores after replacing each feature with all unique T feature values.

C. Key Similarities and Differences

In Table I, we summarize and clarify the difference between the explanation techniques. The table helps us analyze the reasons behind the empirical success and failures of these techniques on our studied datasets later in Section III and Section IV. The explanation techniques generally differ in the way they generate samples, their kernel function, the labeling technique they use, the surrogate models, and their objective functions.

In Figure 2, we show an example of feature importance scores from all our studied explanations for the first five features in the MQ2008 dataset. In this example, the predicted relevance score of the document is -0.82, and the importance scores show the contribution of the first five features to this predicted relevance.

D. Evaluation Measures for Local Explanations

As mentioned in Section I, evaluating local explanations is challenging as the ground truth importance scores cannot be directly extracted from black-box models. However, in the literature on explainability, several evaluation measures are proposed.

Explanation Consistency (Section II-D1) measures the sensitivity of explanation techniques with respect to their hyperparameters, e.g., sample size. Validity and Completeness (Section II-D2) measure the change in the predicted score of the explained document after nullifying important and unimportant features from its explanation. Fidelity (Section II-D3 and Infidelity (Section II-D4) are based on the product between the explained document and its explanation. Lastly, Feature Frequency (Section II-D5 is based on the similarity of explanations with a baseline: the frequency of features used for splits along the decision paths of tree-based models.

1) Explanation Consistency: Explanation consistency [11] is one of the desired properties of local explanation techniques that employ surrogate models, e.g., LIME, SHAP, EXS, and LIRME. Explanation consistency measures the change in the top-K ($k \ll M$) important features as the explanation sample size increases. The logic behind this is that as the sample size grows, these explanations must become consistent since the surrogate model has more information about the vicinity of the document it explains [9]. Consistent explanations show minimal changes in their set of top important features as their sample size increases and reaches a plateau.

2) Validity and Completeness: Validity (Completeness) measures the change in the predicted score of explained documents after the top-K important (unimportant) features from their explanations are nullified [14, 31] in the explained document d. The change in the predicted scores is calculated across cutoff points of K, and after averaging the values

Name	Sampling	Kernel	Labeling	Surrogate	Objective
LIRME	Quantile	exp	Scores	Ridge	Weighted MSE
EXS	Quantile	exp	Anchor	SVM	Squared Hinge
LIME	Gaussian	exp	Scores	Ridge	Weighted MSE & Larspath
SHAP	Gaussian	discrete	Scores	Ridge	Weighted MSE & Larspath
LPI	Replacement	None	Scores	None	Change in Prediction & None

TABLE I: Key Differences between the explanation techniques in our study. LIRME and EXS point-wise and LIME and SHAP are non-LTR explanation techniques.

across all documents, the AUC of the chart is calculated as proposed by [18]. Faithful explanations based on these measures have small (large) values of Validity (Completeness). Nullification is performed by replacing the feature values with their average values in the datase. We provide separate analyses of Validity and Completeness based on changes in predicted scores and ranks, and our cutoff values for K include $[0.1, 0.2, 0.3, 0.4, 0.5]$ percent of features in datasets as proposed in [18]. See Figure 3 for an example of these two measures. In Section III-E, we report the AUC values for these measures.

3) Fidelity: In [13], the authors proposed Fidelity for evaluating explanations of LTR models. Given a local explanation ϕ and a document d and a black-box model f, the fidelity is calculated as mean squared error between $d \cdot \phi$ and $f(d)$). Faithful explanations have large values of Fidelity.

4) (In)fidelity: In [22], the authors proposed (In)fidelity for evaluating local explanations. In this measure, we first calculate the product between the explanation ϕ and the explained document after significant perturbations d', i.e., $\phi{\cdot}d'$. Then, the mean squared error is calculated between $\phi \cdot d'$ and $f(d) - f(d')$. In our study, we replace the top-20% of features in the explained document with their corresponding average values for significant perturbations. Faithful explanations have small values of (In)fidelity.

5) Feature Frequency Similarity: In tree-based models, features that appear on the decision path of a single document play a significant role in the prediction of that document. The feature frequency is proposed and used in most tree-based models for obtaining global explanations [32, 33]. In our study, we calculate the feature frequency on the decision path of each single document. Note that one feature can be used multiple times to split along the decision path. For LambdaMART models, we average the frequencies over all trees. We use the Kendal Tau correlation between local explanations and the feature frequency vector as the similarity measure. The local explanations that provide the largest similarity to this vector are considered more faithful.

6) Pairwise Similarity: The pairwise similarity shows the agreement between pairs of two explanations from two different explanation techniques [34]. We use the Kendal Tau correlation between the absolute importance scores from two explanations of a single document to measure pairwise similarity.

III. Experiments

In this section, we present the empirical result of evaluating the explanation techniques for the LambdaMART model trained on Web10K, Yahoo, and MQ2008 datasets. After describing the experimental setup, we present the global feature importance scores obtained from LambdaMART in Section III-B. In Section III-D, we show the agreement between explanations using pairwise similarity. In Section III-C, we discuss the evaluation of explanation based on Explanation Consistency. In Section III-E, the evaluation of explanation using the Validity, Completeness (In)fidelity, Fidelity, and Feature Frequency similarity are presented. Lastly, we investigate the relation between the median depth of features across all trees in LambdaMART and their feature importance scores obtained from different explanations.

A. Experimental Setup

The datasets included in this study, MQ2008, Web10k, and Yahoo LTR datasets, have 800, 10000, and 29921 queries with document pairs with 46, 137, and 699 features, respectively. We have used the LightGBM implementation of LambdaMART [32]. We have used LightGBM implementation of LambdaMART and have kept the default parameters as they achieve the state-of-the-art performance in all datasets as shown in [3], i.e., nDCG@5 score of 0.75, 0.72, and 0.46, and nDCG@10 score of 0.79, 0.76 and 0.48, respectively in each dataset.

The evaluation considers point-wise explanations of over 100 randomly selected queries from the test set of each dataset. The sum of all associated documents for these queries is 607, 3479, and 462 as the MQ2008, Web10k, and Yahoo LTR datasets, respectively.

For EXS explanations, we set the anchor document to the document that achieves the top 10 percent of the ranks among the other documents. This is because choosing an anchor ranked higher or lower in the lists induces a large imbalance between the generated labels of documents and, as a result, causes the surrogate model not to converge. In all the evaluations, we evaluate the explanations after ranking the features based on their absolute importance scores, as is common practice in tabular dataset [8, 26, 7]. This way, the important features are positioned at the top of the ranked list, regardless of the sign of their importance scores.

For all LIME-based explanations, the background dataset is the entire training set. The random explanation baseline allocates uniformly random importance scores between -1 and

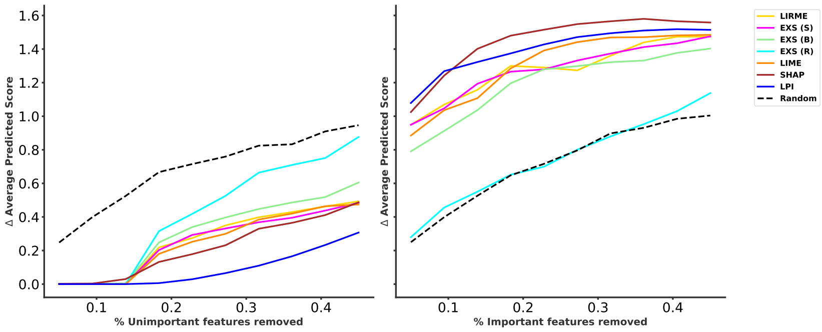


Fig. 3: MQ2008: Validity (Left) and Completeness (Right) of explanations with a varying number of top-K important and unimportant features in the dataset. Faithful explanations provide low (large) values for Validity (Completeness). Note that all explanations except EXS (R) for Completeness are more faithful than our random baseline based on both measures.

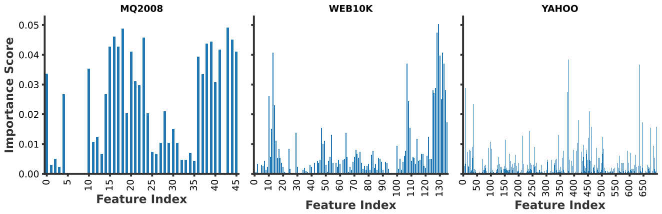


Fig. 4: Global Feature Importance Scores of LambdaMART

1 for all features. For details on the implementation, we refer readers to our code.

B. Global Feature Importance

In this section, we present the global feature importance scores of LambdaMART to bring an intuition about our studied dataset (Figure 4). The global importance of features is based on the number of times features are used for splits in the nodes of LambdaMART for all documents in our training datasets. Notice that in MQ2008, unlike Web10k and Yahoo, the feature importance scores are more evenly distributed.

C. Explanation Consistency

In this section, we measure the Explanation Consistency of our LIME-based explanations, e.g., LIRME, EXS, LIME, and SHAP, based on their sample size. Given sample sizes $T = [500, 1000, 2000, 30000, 40000, 5000]$, the explanation consistency at sample size T is the similarity of the top 50 percent of important features between explanation at time T and $T-1$. The similarity metric is Jaccard Similarity. The consistency of the top 50 percent important features in faithful explanations is expected to increase and reach a plateau. The plateau happens when all different perturbations of explained documents are nearly created, and generating more samples does not necessarily lead to significant changes in the information captured by the surrogate model.

In Figure 5, we can see that the consistency for several explanation techniques converges to a fixed value as the sample size grows as expected. There are a few exceptions. For example, EXS (Top-K Rank) in MQ2008, LIME, SHAP, and EXS (Top-K Binary) in the Yahoo dataset. There are similar trends between the consistency of explanations in the MQ2008 and Web10K datasets. SHAP provides the largest faithfulness relative to other explanation techniques in MQ2008 and Web10K datasets, while LIME is the most consistent explanation on the Yahoo dataset. EXS (Top-K Rank) and EXS (TOp-k) Binary in MQ2008, along with EXS (Score) in the Yahoo dataset, show a relatively low change in the values for consistency as their sample size grows.

Based on our result, we set 3000, 4000, and 5000 as the selected sample size for all explanations when explaining LambdaMART on the MQ2008, Web10k, and Yahoo, respectively. We chose the plateau threshold, the sample size value, since beyond that value, increasing the sample size does not make large changes to the consistency among the top 50% of important features. Moreover, we chose a similar sample size for all explanations for a fair comparison, as we need to allocate an equal computational budget to all explanations with sample size hyper-parameters.

D. Pairwise Explanation Similarity

In this section, we measure the agreement among explanations by measuring their Pairwise Similarity. Pairwise Similarity is calculated by measuring the Kendal Tau correlation between a pair of explanations of all documents in test queries. In Figure 6), we see the average similarity values among the top-50% of important features among explanations across all datasets. Overall, we can see that the average pairwise similarity, or agreement, between techniques is not large, except for a few cases: EXS (S) and LIRME for MQ2008, LPI, and SHAP for Web10k and Yahoo datasets. The disagreement confirms that the design choices behind each explanation technique (Table I) do lead to substantially different explanations in terms of feature importance scores.

E. Evaluation

In this section, we present results for the remaining evaluation measures, i.e., (In)fidelity, Validity, Completeness, and Feature Frequency similarity. As mentioned earlier, faithful explanations should exhibit small values of Infidelity and

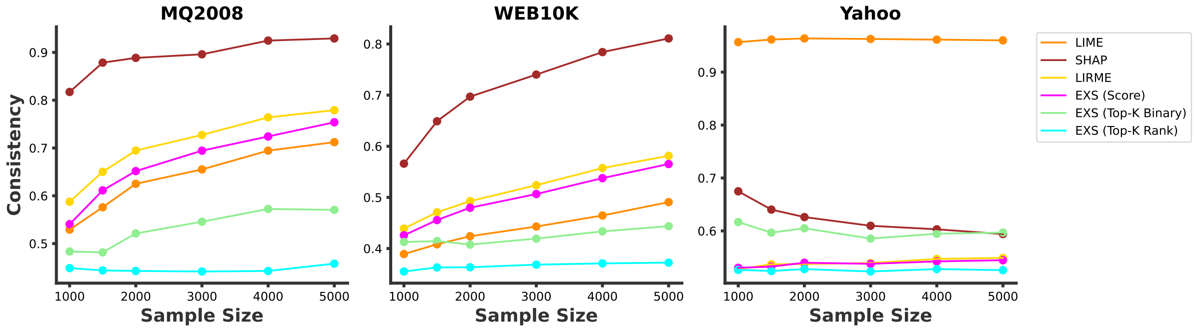
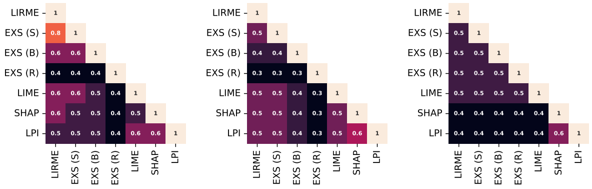


Fig. 5: Explanation consistency of top-50 percent of important features using Jaccard Similarity. The first point of the chart is the comparison between the sample size of 500 to 400 for MQ2008 and 700 to 1000 for Yahoo and Web10k datasets.



(a) MQ2008 (b) Web10K (c) Yahoo

Fig. 6: The average pairwise similarity between explanations of test documents based on Kendal Tau in each dataset.

Validity and large values of Fidelity, Completeness, and FF similarity.

In Tables II, III and IV, we see the average value of each evaluation measure for explanations of all documents associated with test queries of the MQ2008, Web10k, and Yahoo datasets. For MQ2008, non-LTR-based explanations provide faithful explanations for the majority of measures: SHAP for Feature Frequency similarity and Completeness and LPI for validity. In only one measure, i.e., Infidelity, EXS (R) provides the most faithful explanations.

In the Web10k Dataset, LPI is the most faithful explanation for Feature frequency similarity, Completeness, and Validity. On the other hand, LIRME is the most faithful explanation based on Fidelity and Infidelity.

In the Yahoo dataset, LPI is the most faithful explanation based on Feature Frequency similarity and Validity. LIME is the most faithful explanation based on Completeness, while EXS (R) is the most faithful explanation based on Infidelity.

Surprisingly, our random baseline is the most optimal explanation based on Fidelity for MQ2008 and Web10K.

To summarize the results in the previous tables and for a clearer overview of the faithfulness of each explanation technique, we analyze the rank of all explanation techniques based on every evaluation measure across all datasets In Figure 7. The results are the ranked values of Tables II, III and IV. We have adjusted the ranks so that lower ranks indicate more faithfulness for all measures. Overall, we can see that SHAP and LPI consistently rank lower across numerous measures and datasets except for the Fidelity measure. Among the LTR-based explanations, LIRME provides relatively low ranks for the Web10k dataset, yet the ranks for other measures and datasets are larger than those of non-LTR-based explanations.

	FF ↑	Fidelity ↑	Completeness ↑	Validity ↓	Infidelity ↓
LIRME	0.46	5.58	227.27	42.77	3.28
EXS (S)	0.45	5.33	227.08	40.72	3.64
EXS (B)	0.38	6.52	212.85	49.13	4.61
EXS (R)	0.23	3.5	127.02	67.69	**2.96**
LIME	0.37	4.03	233.63	39.78	3.57
Shap	**0.57**	4.07	**261.62**	33.21	4.53
LPI	0.51	4.42	251.25	**11.07**	3.97
Random	-0.01	**8.66**	124.45	119.29	7.76

TABLE II: MQ2008: Average values of evaluation measure across test documents. The bold values indicate the most optimal explanation for each measure.

	FF ↑	Fidelity ↑	Completeness ↑	Validity ↓	Infidelity ↓
LIRME	0.4	**1.3**	326.03	14.11	**0.11**
EXS (S)	0.39	1.11	325.38	14.94	1.12
EXS (B)	0.22	1.27	240.94	59.3	0.45
EXS (R)	0.01	0.92	81.91	135.04	0.75
LIME	0.26	1.22	330.7	16.77	1.63
Shap	0.5	1.27	296.43	3.77	1.02
LPI	**0.53**	0.84	**333.84**	**0.07**	0.16
Random	0	1.25	106.78	109.22	0.83

TABLE III: Web10k: Average values of evaluation measure across test documents. The bold values indicate the most optimal explanation for each measure.

F. *Effect of Depth*

In the structure of decision trees in LambdaMART, features utilized for splitting in nodes with shallower depths, closer to the root node, are regarded as more important [30]. This is because a larger number of documents are likely to traverse through these nodes along the decision paths of the tree.

	FF ↑	Fidelity ↑	Completeness ↑	Validity ↓	Infidelity ↓
LIRME	0.45	6.63	158.48	5.04	3.38
EXS (S)	0.44	7.45	159.39	5.46	4.42
EXS (B)	0.39	7.69	132.67	13.52	5.23
EXS (R)	0.33	2.94	82.65	15.85	**2.11**
LIME	0.52	8.29	**186.65**	3.83	5.29
Shap	0.45	7.1	170.94	7.1	5.69
LPI	**0.58**	7	167.68	**0**	3.7
Random	-0	**8.54**	61.87	62.95	7.37

TABLE IV: Yahoo: Average values of evaluation measure across test documents. The bold values indicate the most optimal explanation for each measure.

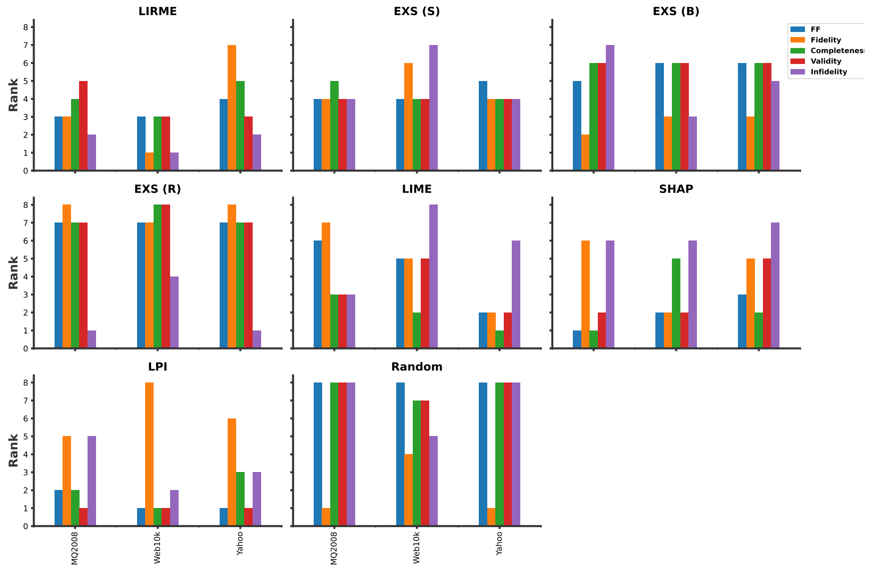


Fig. 7: The average rank of all explanations across all measures in our benchmark datasets. Lower ranks indicate more faithful explanations for all measures.

We investigate the median depth of features among the top-K important features in each explanation. As we increase the value of K, it can be expected that the median rank of the feature set should also rise for most explanations since it can increase the inclusion of less significant features used in nodes with greater depth.

Figure 8 shows the result of our analysis, averaged over all documents for all test queries in each dataset and across all trees in LambdaMART. We can see that most explanations follow the expected trend with few exceptions. The average median depth of features in EXS (R) for Web10K does not change as we increase the vales of K, while it decreases in the Yahoo dataset.

This is an expected behavior of EXS (R) as it allocates importance to features that can change the relevance scores of the explained document only if they are larger than the rank or predicted scores of the anchor documents. As we mentioned in Section II-A2, the anchor documents are set to be the top-rank documents. Since the features with large depths are considered less important, EXS (R) allocates very small values of importance to them. However, the trends for LIME explanations in Yahoo datasets are surprising as LIME is expected to set importance on any feature for which its change in value can improve the predicted relevance score of LambdaMART, even in smaller values.

IV. Discussion

Our experiments show that LTR-based explanation techniques of LIRME and EXS do not strongly outperform the non-LTR-based explanations of LPI, LIME, and SHAP. We would like to present some reasons as to why they lack faithfulness.

By comparing the difference between LIME and LIRME in Table I, we can argue that the sampling technique of LIRME can be a potential limitation of this technique. This is because the main difference between LIRME and LIME is their sampling techniques. LIME is based on Gaussian sampling and LIRME is based on interpretable quantile sampling. One possible improvement to LIRME is by abandoning the idea of an interpretable sampling process and replacing it with Gaussian sampling of LIME.

By comparing the difference between EXS and LIRME in the same table, we can argue that the low faithfulness of EXS can be traced back to its labeling process. This is particularly evident for EXS (S) and EXS (B) approaches. During our experiments, we noticed that samples generated by EXS (B) are largely imbalanced. One possible solution for this is to use oversampling techniques on top of the EXS (B) sampling process.

V. Concluding Remarks

We evaluated the local pointwise explanation of LambdaMART models trained on the Yahoo, Microsoft Bing Search (MSLR-WEB10K), and LETOR 4 (MQ2008) datasets. In the investigation, the LTR pointwise explanation techniques LIRME and EXS were compared to the non-LTR explanation techniques LIME, SHAP, and LPI. We used an extensive set of evaluation measures; Explanation Consistency, Pairwise Similarity, Validity, Completeness, Feature Frequency Similarity, and (In)Fidelity.

We showed that explanations are mostly optimal based on specific evaluation measures and no single explanation technique is faithful for all studied evaluation measures and across all our datasets. As a result, we can conclude that providing faithful explanations of LambdaMART is no silver bullet.

Our other research question was whether the LTR-specific explanation techniques outperform the non-LTR-specific tech-

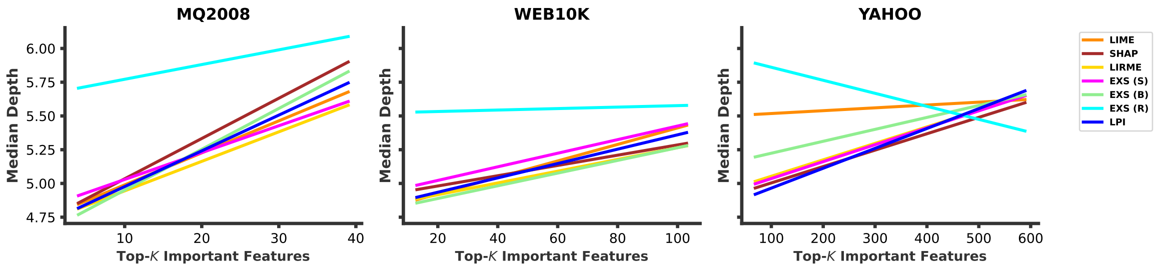


Fig. 8: The relationship between the median depth of features in the top-K important features in each explanation. The results are averaged across all test query document pairs. We expect the median rank of features to increase as the values of K increase for most explanations.

niques. The presented results give some support for a positive answer, when evaluating performance using (In)fidelity. On the contrary, for the measures of Validity and Feature Frequency Similarity, LPI and SHAP were observed to outperform all competing techniques.

Even though LPI does not include a surrogate model, it was shown to outperform LIRME and EXS across numerous measures. Based on this, we propose developing and evaluating surrogate-free explanations as a future direction for our study.

In our experiments, we showed that random baseline explanations showed faithfulness to the Fidelity measure for MQ2008 and Yahoo datasets. We argue that further studies need to further investigate the Fidelity measure proposed by [13].

Another possible future direction is to study the link between model accuracy, the number of features, and the performance of local explanations for LTR models similar to the investigations made for local explanations of classification and regression models in [35, 9].

Our study has several limitations. Firstly, the conclusions made in our study about which explanations are most optimal apply only to LambdaMART and the studied datasets. Secondly, even though certain explanations are shown to be faithful based on a specific evaluation measure in our study, local explanations need to be evaluated using human subjects before they are deployed in high-stake decision-making domains.

References

[1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 89–96.

[2] O. Chapelle, Y. Chang, and T.-Y. Liu, “Future directions in learning to rank,” in *Proceedings of the Learning to Rank Challenge*. PMLR, 2011, pp. 91–100.

[3] Z. Qin, L. Yan, H. Zhuang, Y. Tay, R. K. Pasumarthi, X. Wang, M. Bendersky, and M. Najork, “Are neural rankers still outperformed by gradient boosted decision trees?” *Proceedings of International Conference on Learning Representations*, 2021.

[4] Z. Hu, Y. Wang, Q. Peng, and H. Li, “Unbiased lambdamart: an unbiased pairwise learning-to-rank algorithm,” in *The World Wide Web Conference*, 2019, pp. 2830–2836.

[5] Y. Izza, A. Ignatiev, and J. Marques-Silva, “On explaining decision trees,” *arXiv preprint arXiv:2010.11034*, 2020.

[6] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[7] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “" 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*, 2016, pp. 1135–1144.

[9] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, “General pitfalls of model-agnostic interpretation methods for machine learning models,” in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2020, pp. 39–68.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[11] M. Verma and D. Ganguly, “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.

[12] J. Singh and A. Anand, “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.

[13] T. Chowdhury, R. Rahimi, and J. Allan, “Rank-lime: Local model-agnostic feature attribution for learning to rank,” *arXiv preprint arXiv:2212.12722*, 2022.

[14] J. Singh, M. Khosla, W. Zhenye, and A. Anand, “Extract-

ing 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.

[15] T. Chowdhury, R. Rahimi, and J. Allan, "Rank-lime: local model-agnostic feature attribution for learning to rank," in *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 2023, pp. 33–37.

[16] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Gimenez-Abalos, "Focus! rating xai methods and finding biases," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–8.

[17] A. Jain, M. Ravula, and J. Ghosh, "Biased models have biased explanations," *arXiv preprint arXiv:2012.10986*, 2020.

[18] C.-Y. Hsieh, C.-K. Yeh, X. Liu, P. Ravikumar, S. Kim, S. Kumar, and C.-J. Hsieh, "Evaluations and methods for explanation through robustness analysis," *Proceedings of International Conference on Learning Representations*, 2021.

[19] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

[20] A. H. A. Rahnama, J. Bütepage, P. Geurts, and H. Boström, "Can local explanation techniques explain linear additive models?" *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 237–280, 2024.

[21] A. H. Akhavan Rahnama, "The blame problem in evaluating local explanations and how to tackle it," in *European Conference on Artificial Intelligence*. Springer, 2023, pp. 66–86.

[22] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[24] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*. PMLR, 2020, pp. 9269–9278.

[25] A. H. A. Rahnama and H. Boström, "A study of data and label shift in the lime framework," *arXiv preprint arXiv:1910.14421*, 2019.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[27] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 655–670.

[28] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview," *Learning*, vol. 11, no. 23-581, p. 81, 2010.

[29] D. Garreau and U. von Luxburg, "Looking deeper into tabular lime," *arXiv preprint arXiv:2008.11092*, 2020.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] J. Singh, M. Khosla, and A. Anand, "Valid explanations for learning to rank models," *arXiv preprint arXiv:2004.13972*, 2020.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[33] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[34] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *arXiv preprint arXiv:2202.01602*, 2022.

[35] A. H. A. Rahnama, J. Bütepage, P. Geurts, and H. Boström, "Can local explanation techniques explain linear additive models?" *Data Mining and Knowledge Discovery*, pp. 1–44, 2023.