Local Interpretable Model-Agnostic Explanations for Neural Ranking Models

Amir Hossein Akhavan Rahnama*[1], Laura Galera Alfaro*[2], , Zhendong Wang[3] and Maria Movin[4],

***Abstract*—Neural Ranking Models have shown state-of-the-art performance in Learning-To-Rank (LTR) tasks. However, they are considered black-box models. Understanding the logic behind the predictions of such black-box models is paramount for their adaptability in the real-world and high-stake decision-making domains. Local explanation techniques can help us understand the importance of features in the dataset relative to the predicted output of these black-box models. This study investigates new adaptations of Local Interpretable Model-Agnostic Explanation (LIME) explanation for explaining Neural ranking models. To evaluate our proposed explanation, we explain Neural GAM models. Since these models are intrinsically interpretable Neural Ranking Models, we can directly extract their ground truth importance scores. We show that our explanation of Neural GAM models is more faithful than explanation techniques developed for LTR applications such as LIRME and EXS and non-LTR explanation techniques for regression models such as LIME and KernelSHAP using measures such as Rank Biased Overlap (RBO) and Overlap AUC. Our analysis is performed on the Yahoo! Learning-To-Rank Challenge dataset.**

I. Introduction

Learning-to-rank (LTR) models are machine learning techniques designed to automatically learn from training data consisting of queries and corresponding ranked lists of documents (or sometimes called items) [1]. These models learn a ranking function to increase the relevance of each document to its corresponding query. LTR models are often complex since they are trained using many parameters to achieve high accuracy [2]. The complexity of ranking models can sometimes undermine their efficacy, as humans struggle to comprehend the rationale behind a particular order [3]. The absence of transparency in these so-called black-box models can cause prediction errors, biases, or even unethical behavior [4]. Hence, there is a need to understand the complex black-box models [5].

Generalized Additive Models (GAMs) are statistical models that allow for flexible, non-linear relationships between the input (predictor) and the output (response) variables. In these models, the response variable is modeled by an additive combination of smooth functions on each predictor variable. Neural Ranking GAMs [6] builds on using neural networks to model the smooth functions for each predictor variable. Because of this, Neural Ranking GAMS are intrinsically interpretable. In [6], the authors showed that Neural Rank GAMs outperformed other types of neural network-based LTR models across tabular datasets.

*Equal Contribution
[1]KTH Royal Institute of Technology, Sweden amiakh@kth.se
[2]Stockholm University, Sweden laga6199@student.su.se
[3]Stockholm University, Sweden zhendong.wang@dsv.su.se
[4]Spotify, Sweden mariamovin@spotify.se

Explanation techniques provide information about the logic behind the prediction of black-box models in a post-hoc manner, i.e., after the models are trained. Explanations come in different categories: feature attribution, counterfactual explanations, etc. Feature attributions are among the most popular explanations due to their flexibility and easy interpretation. Feature attribution presents the explanations in terms of real-valued importance scores, where each score depicts the importance of that feature to the predicted output of the black-box model [7].

Feature attribution explanations are themselves further divided into two categories: local and global explanations [8]. Global explanations provide feature importance scores to the predicted output of black-box models for the entire dataset. On the other hand, local explanations provide feature importance scores for the predicted output of the black-box model for a single data point. Global explanations summarize the dataset's important features, while local explanations excel when a user needs to understand the underlying reasons behind the (possibly wrongful) prediction of a single instance in a production machine-learning model. For example, the surprising result of a search query for a single in a music streaming app.

For LTR models, local feature attribution explanations are further categorized into point-wise [9, 10] and list-wise explanations [11]. Local point-wise explanations provide feature importance scores for the predicted output of an LTR model given a single document associated with a given query. Local list-wise explanations provide feature importance scores for the predicted output of an LTR model on the entire list of documents associated with a query. Consider the case when a user puts in the search query "The Wall album" in a music streaming app and observes that the album "Off the Wall" by Michael Jackson receives a low relevance score by the black-box LTR model. Obtaining a point-wise explanation of this document (or item) can help users understand the contribution of features such as Term frequency–Inverse document frequency (TF-IDF) to these surprisingly low relevance scores.

The main challenge in using explanation techniques lies in their evaluation [12, 13]. This is partly because the ground truth importance scores cannot be directly extracted from complex black-box models. However, since Neural Rank GAMs have Generalized Additive components that are intrinsically interpretable, we can extract the ground truth importance scores, which we refer to as the "Ground Truth"[1]. Therefore, we have a unique opportunity to evaluate local explanations

[1]The ground truth importance scores should not be confused with the definition of ground truth in supervised learning, where ground truths are discrete labels associated with data instances.

of Neural Rank models by directly comparing them to the Ground Truth.

In our study, we investigate the faithfulness of different variants of Local Interpretable Model-Agnostic Explanation (LIME) techniques for explaining Neural Rank GAM models. We propose our variation of LIME with different sampling techniques such as Gaussian, SMOTE, Latin Hypercube Sampling (LHS), and Deterministic LIME (DLIME)[2]. We then evaluate our proposed techniques against the point-wise explanations of Locally Interpretable Ranking Model Explanation (LIRME) [10] and Explainable Search (EXS) [9], and non-LTR explanations of LIME in its official implementation [14] and SHapley Additive exPlanations (SHAP) [15] on the Yahoo! Learning-To-Rank Challenge dataset.

Our study is the first study to evaluate the explanations of Neural Rank GAM models. Moreover, we are the first to evaluate the local explanations of LTR models using ground truth importance scores. We evaluate the explanations using the Rank Biased Overlap (RBO) measure. Moreover, in our study, we propose a measure called Overlap AUC for evaluating local explanations using ground truth[3]. The code of our experiments is available at https://github.com/amir-rahnama/neural_ranking_exp.

More specifically, our main findings are as follows:

1) The faithfulness of Neural Ranking GAM explanations depends on two main factors: the predicted rank of the explained documents and the explanation sample size.
2) No single LIME-based explanations can be faithful with respect to the two aforementioned factors using RBO and Overlap measures in all cases.
3) Our proposed LIME explanations based on Gaussian, DLIME, and LHS sampling provide the most faithful explanations based on Overlap and RBO for the majority of cases, outperforming point-wise explanations techniques of LIRME, EXS, LIME (official implementation), and KernelSHAP.
4) For specific choices of the explanation sample size parameters and when explaining documents ranked second in the test set queries, the LTR-based explanations of LIRME and EXS Score (S) can provide the highest faithfulness based on the RBO measure.
5) We show that our proposed LIME explanation with SMOTE sampling excels at reflecting the explained documents' predicted rank in providing its local explanations.
6) We highlight that generated samples of explanation techniques can be largely imbalanced depending on the predicted rank of explained documents. We postulate that this challenges developing faithful explanation techniques for LTR models.

II. Related Work

To the best of our knowledge, there have not been studies on local point-wise explanations of Neural Rank GAM models. Moreover, no evaluation study has focused on comparing LIRME and EXS explanations for tabular datasets.

LIME-based explanations of LIRME [10] and EXS [9] were originally developed and evaluated on text datasets. In their original study of LIRME, the authors showed that LIRME explanations are both faithful based on Consistency and Correctness. In the study, Consistency was calculated as the similarity between the top-K important features of LIRME explanations as its sample size increased. Correctness was defined as the similarity between the tokens in top-K important features and relevant terms in the text datasets. No systematic evaluation exists in the original study of EXS.

In [16], LIME-based list-wise explanations of RankLIME were shown to be more faithful than the explanations of LIRME and EXS. However, the list-wise explanations are outside the scope of our study.

III. Background

A. Local Point-wise Explanations

Let $X = (q, D)$ be the dataset comprising of query q with a list of m documents $D \in \mathbb{R}^m$, where document d_i is represented by a feature vector $d \in \mathbb{R}^M$, with M as the size of the feature vector.

Learning-to-Rank (LTR) models learn a ranking function $f : D \rightarrow \Pi^M$ from the data, which outputs a predicted score π_i for each document d_i, indicating its relevance to the query q.

LTR models can be optimized using point-wise, pairwise, or list-wise loss functions. The point-wise explanation technique $g : d_i \rightarrow \mathbb{R}^M$ provides $\Phi \in \mathbb{R}^M$, where ϕ_j (for $j = 1, \ldots, M$) is the importance score of feature j with respect to $f(D)$.

B. Neural Ranking GAMs

The neural Generalized Additive Ranking Model is an additive ranking model. For each document d with m features $d = [d_1, d_2, ..., d_m]$, the ranking score is:

$$f(d) = f_1(d_1) + f_2(d_2) + \ldots + f_m(d_m) \tag{1}$$

where each feature is scored by a corresponding sub-model, and the overall ranking score $f(d)$is the sum of all the component f_j outputs where $j = 1, ..., M$. Each component is a standalone feed-forward network. The model is overall *implicitly* interpretable, given that the contribution of each feature d_i to the final ranking score $f(d)$ can be easily allocated to the output of $f_i(d_i)$ where $i = 1, ..., m$. Note that no interaction terms exist between features.

C. LIME and KernelSHAP

The goal of LIME explanations is to allocate an importance score to each feature in the explained instance d with respect to the predicted output of a black-box model f.

The LIME explanations are obtained as follows. LIME generates new samples based on adding Gaussian distribution taken from the training data's mass center. After repeating this process T times, the sample D' is created. LIME then

[2]These sampling techniques are described in Section IV-A.

[3]These measures are defined in Section VI-C.

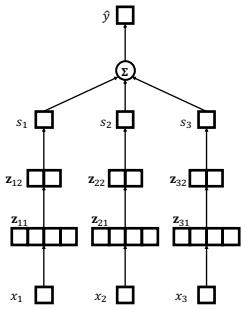

Fig. 1: A graphical illustration of different components in Neural ranking GAMs [17]. We can extract ground truth importance scores from the additive components of the model.

weights these samples using an exponential kernel function $k(d, D')$. After that, the black-box model f is used as an oracle to generate labels for these samples, i.e., $f(D')$. After performing Larspath feature selection to eliminate features with co-linearity, an interpretable surrogate g is trained on new samples with that subset of features selected by Larspath and their sample weights and labels to minimize the loss: $\xi(f, g, \pi_x)$. The explanations E are the weight W_g of the surrogate model.

In [14], the authors show a geometrical interpretation of this process. The surrogate model aims to fit a linear model to the vicinity around the explained instance.

KernelSHAP [15] is a variation of SHAP that uses a combinatorial kernel function that is shown to guarantee certain theoretical properties, such as fairness in LIME explanations. LIME and SHAP were originally proposed to explain supervised learning models. They can provide pointwise explanations of LTR models when they are used for regression models.

D. LTR-based Explanations

LIRME [10] and EXS [9] are examples of explanation techniques that have adapted LIME for LTR models. In this section, we provide an overview of these techniques.

In their original study, LIRME and EXS used an interpretable sampling of text data. In this study, based on the proposal of [18], we have adopted LIRME and EXS's sampling to quantile interpretable sampling process that is the equivalent sampling but for tabular datasets.

This sampling process transforms the explained instance into a binary interpretable representation based on quantiles of features. Feature values of explained instances are allocated the number of bins they fall into. A sampling process generates new samples d' from the explained instance d. The samples are generated by randomly selecting a subset of features in d, and then, for each selected feature, one of four bins is randomly selected. If the selected bin from the generated sample equals the bin of the feature value in an explained instance, the sample receives a value of one and zero otherwise. The sampling process is performed T times to create $D' = \{d'_1, ...d'_T\}$ where T is a hyper-parameter. For more details on this process, see [18].

LIRME trains a Ridge surrogate model on pairs of $(D', f(D'))$ with the following loss function:

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

where Θ is the weight of the surrogate model and hence are LIRME explanations.

EXS, on the other hand, uses a Linear SVM surrogate and has three labeling processes built for labels y, which leads to three variants in the experiment comparison: Score-based (S), Top-K binary (B), and Rank-based (R). In *Score-based (S)*, 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 that query. *Top-K binary (B)* 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 (R)*, the label of d' is zero if its rank is less than the top-K document in the query. Otherwise, the label equals $1 - \frac{R(d')}{k}$. EXS uses a hinge square loss or epsilon-insensitive loss function to train its surrogate, depending on the type of labeling used:

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

where T is the sample size of perturbed documents and Θ is the parameter of the surrogate linear SVM model g and hence are the EXS explanations.

IV. Methodology

In our study, we propose different adaptations of the LIME explanation for explaining Learning-To-Rank models. Our adaptation has some differences with LIRME, EXS, LIME, and SHAP. The most important difference is our sampling process. Secondly, we skip the Lars path feature selection process after training our surrogate model.

A. Sampling

As we mentioned, the first difference is that we do not sample based on quantile and binary representations like LIRME and EXS. As other studies have shown [8, 19], transforming the data into binary representations comes with a limitation: we are operating in a data space that is different than the original data space, and moreover, there is an information loss. We propose four sampling techniques for LIME explanations of LTR models on tabular datasets: Gaussian, SMOTE, Latin Hypercube Sampling (LHS), and Deterministic LIME (DLIME).

Gaussian sampling introduces perturbations to each feature of the original instance by adding random noise drawn from a normal distribution.

SMOTE [20] is a variation of the Synthetic Minority Oversampling Technique that randomly selects one of the k-nearest neighbors to the instance explained and then creates new samples by interpolating between the feature values of pairs of instances.

LHS [21] applies a structured approach to sample across feature distributions. Formally, for each dimension j, $i = 1, 2, ..., M$, LHS divides the range of possible values into T intervals and samples uniformly within each interval. Additionally, LHS ensures that only one sample is taken from each interval along each dimension, which avoids the clustering of samples. The process can be summarized as follows: 1) Divide each dimension into T equal intervals; 2) Randomly select one sample from each interval along each dimension; 3) Permute the samples randomly within each dimension to eliminate any remaining order dependencies. The resulting set of samples provides a more evenly distributed coverage of the multidimensional space.

Finally, *DLIME* first generates samples using LHS and then selects a subset of them by applying Agglomerative Clustering and choosing the cluster that contains the nearest neighbors of the explained document (target instances). Agglomerative Clustering [22] is a hierarchical clustering technique used to group similar data points into clusters. It starts with each data point considered as a single cluster and iteratively merges the closest pairs of clusters until a predefined stopping criterion is met. Let n be the number of data points and d be the dimensionality of the data. The process can be summarized as follows: 1) Start with n clusters, each containing a single data point; 2) Compute the pairwise distance or similarity between all clusters; 3) Merge the two closest clusters based on a linkage criterion (e.g., single linkage, complete linkage, average linkage); 4) Update the distance matrix to reflect the distances between the new cluster and the remaining clusters; 5) Repeat steps 2-4 until a stopping criterion is met, such as reaching a desired number of clusters or a specified threshold distance. Specifically, agglomerative clustering is computationally intensive, particularly for large datasets, as it requires computing the pairwise distances between all data points at each iteration. However, it often produces interpretable hierarchical structures that can be visualized using dendrograms.

B. Training the surrogate

The second difference between our approach and LIME and KernelSHAP is that we skip the Larspath feature selection step in LIME and SHAP. Moreover, our labeling process is similar to the Top-K binary labeling of EXS. To re-iterate, the generated sample based on d is labeled one if $R(q, d')$ is greater than $R(q, d_k)$, being d_k the k-th ranked document for that given query.

V. Evaluation

As mentioned earlier in Section I, evaluating local explanations is challenging and an open research problem [12]. However, in the case of Neural Rank GAMs, we have access to the ground truth importance scores from the Generalised Additive Model components.

Because of this, we can evaluate local explanations by directly comparing them to the ground truth importance scores obtained from Neural Rank GAM models. We make use of two measures: *RBO* and our proposed *Overlap AUC*.

The *RBO* [23] measure compares two ranked lists, and allocates a numeric value between zero and one to represent their similarity. The measure is calculated as follows:

$$RBO = (1-p) \times \sum_{k=1}^{n} \frac{p^k \times \min(k,m)}{k * m} + p^n \times \frac{n}{m} \qquad (3)$$

where p is a parameter between 0 and 1, indicating the weight assigned to ranks, n is the depth at which the overlap is calculated, and m is the length of the reference list. The measure includes $min(k, m)$ to ensure the calculation does not go beyond the length of the lists. Faithful explanations have a large value of RBO similarity to the ground truth.

Overlap AUC is our proposed measure for evaluating the faithfulness of explanations with respect to the ground truth. For calculating Overlap AUC, we first select the top-K important features from an explanation ϕ and ground truth vector λ where $K = 1, ..., M$ and M is the total number of features in the dataset.

$$\text{Overlap}(k, \phi, \lambda) = \frac{|\text{Top}(k, \phi) \cap \text{Top}(k, \lambda)|}{k}, \qquad (4)$$

For each value of K, Overlap(k, ϕ, λ) allocates a value between zero and one to represent the similarity. See Figure 6 for an example of how Overlap is calculated for explanations with explanations of sample sizes 500 and 2000.

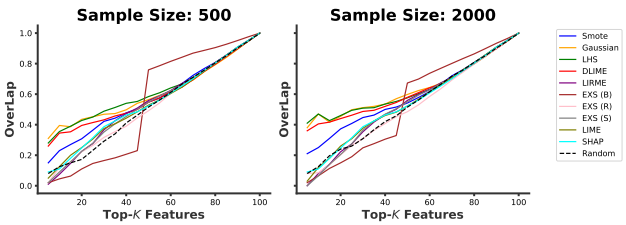

Fig. 2: The Overlap of explanations with Ground truth for documents with predicted rank of two in Yahoo dataset.

In order to reduce the dependence on the value of K, we calculate Overlap AUC by marginalizing over the values of K and calculating the area under the curve:

$$\text{Overlap\_AUC}(\phi, \lambda) = \int_0^N \text{Overlap}(k, \phi, \lambda)\, dk, \qquad (5)$$

where $K = 1, ..., M$. Based on this, larger values of Overlap AUC indicate that the generated explanations are more faithful.

VI. Experiments

In this section, we provide the result of our empirical investigation. Firstly, we describe the setup of our experiments in Section VI-A. After that, we peek into the ground truth importance scores obtained from Neural Ranking GAM models in Section VI-B. Our main evaluation analysis is presented in Section VI-C.

A. Setup

This study uses the Yahoo! Learning-To-Rank Challenge dataset [24]. This publicly accessible dataset includes two sets, namely Set1 and Set2. Set1 is commonly used for learning to rank evaluation and consists of three partitions for training, validation, and testing. Each document in this dataset is represented by 700 numerical features, normalized to a range of $[0, 1]$ using inverse cumulative distribution. The specific meaning of each feature is not disclosed. The documents are labeled with relevance labels ranging from 0 to 4.

For data preprocessing, we reduced the dimensionality of the data from 700 features to 100 by performing feature selection. This is because the large majority of features do not have discriminative or predicted power. In our feature selection process, we excluded the features that appeared in less than fifty percent of the documents in the training dataset. Then, we conducted a correlation study between features and their relevance scores to find the top 100 features with the highest discriminative power.

We have used the official Tensorflow implementation of Neural GAM models. For training the Neural ranking GAM model, we chose the hyper-parameter configuration in the original study [6]. With the defined partitions of training and testing, we obtained a Normalized Discounted Cumulative Gain (NDCG) score [25] of 77.89% for the trained model. The NDCG score measures a ranking algorithm's quality by assessing the retrieved items' relevance and considering their positions in the result list.

in the Neural Rank GAM model, the ground truth importance scores for feature j are extractable from the weight of the component called "feature j subscore" where $i = 1, ..., M$. See our implementation code for more details[4].

To evaluate explanations, we randomly selected 20 queries from the test set, each with 23 associated documents. For each query, we explain the 2nd and 10th-ranked documents by the Neural Ranking GAM model since we are interested in investigating the effect of the predicted rank of documents on the faithfulness of explanations.

For SMOTE sampling, the number of neighbors is set to 10. This value is the optimum minimum for the surrogate loss among the values between 3 and 20.

For LIME and SHAP, we use all the test data as background datasets. This choice has been shown to provide the most optimal performance in [26]. For the LHS sampling [21], the number of clusters is set to 3. The choice is made after observing this value provides the max silhouette scores among the clusters in the range 2 to 11. For EXS, the anchor ranked document is set to the predicted rank of the explained document, i.e., rank 10 and 2. Additionally, our random baseline (referred to as "Random" in the comparison) generates importance scores uniformly at random for all features.

[4]More specifically, see line 279 of the file: generating_exp.py.

B. Ground truth importance scores

In Fig. 3, we present the ground truth importance scores obtained from Neural Ranking GAMs for documents ranked second (left) and 10th (right). Note that since we have performed feature selection and have reduced the set of features to the one with predictive scores, most features have absolute importance scores greater than zero in the figure.

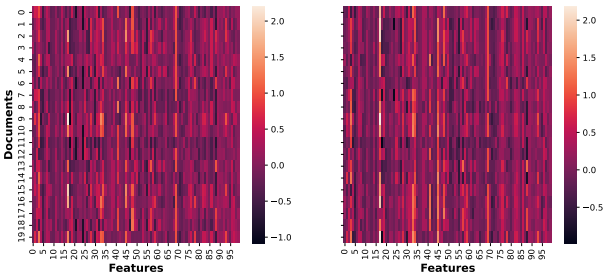

Fig. 3: The ground truth importance scores of Neural Ranking GAMs for documents ranked 2nd (Left) and 10th (Right).

In Fig. 4 and 5, we provide the frequency of top-10 important features obtained from the explanations and the ground truth importance scores (see Fig. 3) for documents with predicted rank of two and ten, respectively. For the documents with predicted rank 2, our proposed explanation techniques based on SMOTE and LHS sampling can detect the top-1 important feature from the ground truth. On the other hand, no explanation technique has detected the top-1 important features for documents with predicted rank of 10. We provide an intuition for this in Section VI-E.

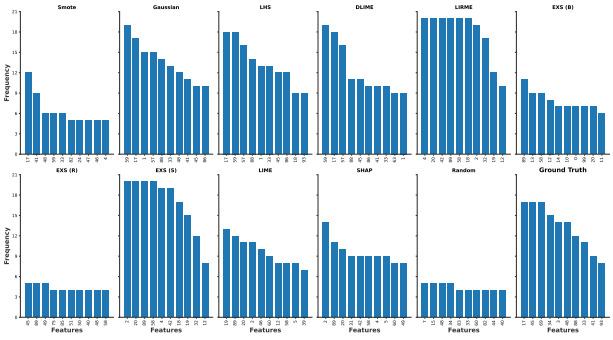

Fig. 4: The frequency of Top-10 Important Features from Explanations and Ground Truth for the explanations of all test documents with the predicted rank of 2.

C. Evaluation

In this section, we present our evaluation of the faithfulness of our studied explanation techniques beyond visual inspec-

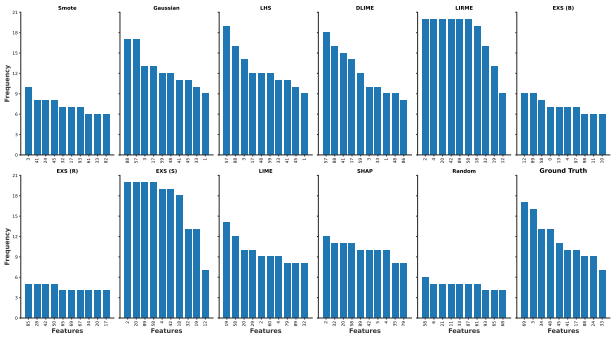

Fig. 5: The frequency of Top-10 Important Features from Explanations and Ground Truth for the explanations of all test documents with the predicted rank of 10.

tions of previous sections. Table I and II show the faithfulness of explanations for documents with the predicted rank of second and tenth, respectively. Note that we have included the results with varying sample sizes for a conclusive comparison.

Overall, we can see that our proposed explanations provide the most faithful explanations across numerous measures and sample sizes. What is most important is that the faithfulness of these explanations is consistent with varying values of sample size. However, there are a few exceptions to this. In Table I, EXS (S) provides the most faithful explanations based on RBO for sample sizes 2000 and 5000 and is on par with LIRME for sample size 3000. In Table II, LIME, SHAP, and Random explanations provide the most faithful explanations based on the RBO measure for sample size values of 500 and 1000.

There is a clear explanation behind the faithfulness of random baseline explanations with smaller sample size values. In smaller sample sizes, the surrogate model is trained on a small subset of data that includes only a few angry changes in the explained documents and their predicted output by the black-box model. Because of this, our explanations are as faithful as a random baseline.

By comparing the results from Table I and II, we can see that our proposed explanations, along with the majority of explanations, are more faithful for documents in predicted ranks of 10 compared to those of predicted rank second. This can indicate that the faithfulness of LIME-based explanations depends on the predicted rank of explained documents. We analyze this phenomenon later in Section VI-E.

D. Overlap based on Predicted Rank

One natural question is to what extent two explanations from a single explanation technique overlap for two documents at two predicted ranks associated with the same query.

For our investigation, we can measure the overlap of top-K important features between two explanations from each explanation technique for two documents, one with the predicted ranks of two and another one with the predicted rank of ten for the same associated query.

We expect that if explanation techniques show a high level of overlaps between the explanation of documents with different predicted ranks, they may not have leveraged the importance of the predicted rank of explained documents efficiently in their explanations.

In Fig. 6, we see the result for explanations of documents averaged over all test queries. In the figure, we can see that our proposed explanation based on SMOTE sampling shows the least increase of overlap as values of K increase. This can partially explain the success of SMOTE sampling in the results from Table I and 5. On the other hand, LIME, LIRME, EXS Binary, and EXS Score (S) show the largest overlap between their explanations for documents between the two ranks. The result is surprising, particularly for EXS (S), as its labeling process is also defined based on the predicted rank of explained documents.

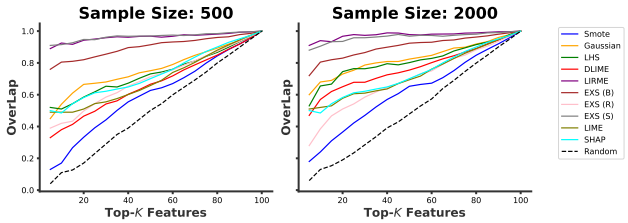

Fig. 6: The average Overlap of Explanations of documents with rank two and ten with varying Top-K important features for sample sizes 500 and 2000.

E. Sampling imbalance

In the previous section, we showed that our proposed explanation techniques provide more faithful explanations for documents with predicted ranks of ten instead of two. We have identified that the cause of this phenomenon is a sample imbalance problem.

In our labeling process, i.e., EXS's Top-K binary (B) labeling, depending on the predicted rank of the explained document by black-box model f, the labels generated by the black box can be largely imbalanced in a given sample. This is because, as we have realized, achieving the predicted relevance scores of documents in the top (or bottom) ranks is increasingly harder than those with moderate ranks in the list of documents associated with queries. In Fig. 7, we can see an example of this phenomenon. In SMOTE sampling, the number of generated samples with label one can incrementally increase. The results are averaged over all test documents and query pairs. This phenomenon affects the explanations at both tails, namely, the documents ranked at the top and bottom of the list.

To address this issue, we added an extra step to our sampling process: oversampling using the SMOTE technique for the minority class. In Fig. 8, we show the difference in Overlap measure between our original method in comparison to when using the samples generated by each sampling technique have gone through an extra step of oversampling. We can see that oversampling does improve the median Overlap faithfulness scores of some sampling techniques, e.g., LHS and DLIME

TABLE I: Predicted Rank 2: Faithfulness of explanations of Neural GAM model with different sample sizes. Bold values indicate the most faithful explanations for each measure.

Sample Size	500		1000		2000		3000		5000	
Measure	RBO	Overlap	RBO	Overlap	RBO	Overlap	RBO	Overlap	RBO	Overlap
SMOTE	0.18	2.05	0.2	2.13	0.19	2.12	0.2	2.13	0.19	2.12
Gaussian	0.19	**2.36**	0.2	2.41	0.19	**2.44**	0.2	2.43	0.2	2.44
LHS	0.2	2.34	0.21	**2.43**	0.19	**2.44**	0.2	**2.44**	0.21	**2.46**
DLIME	**0.22**	2.28	**0.23**	2.36	0.22	2.39	0.19	2.42	0.19	2.42
LIRME	0.2	1.99	**0.23**	1.98	0.2	1.98	**0.22**	1.98	0.2	1.98
EXS (B)	0.12	1.98	0.12	1.99	0.14	1.96	0.11	1.98	0.15	1.96
EXS (R)	0.2	1.93	0.21	1.96	0.21	1.96	0.2	1.89	0.2	1.9
EXS (S)	0.18	1.99	0.19	1.98	**0.24**	1.97	**0.22**	1.98	**0.22**	1.99
LIME	0.2	1.92	0.19	1.95	0.22	1.95	0.19	1.95	0.19	1.94
SHAP	0.2	1.96	0.2	1.96	0.2	1.96	0.2	1.96	0.2	1.96
Random	0.2	1.9	0.2	1.85	0.19	1.9	0.21	1.9	0.19	1.93

TABLE II: Predicted Rank 10: Faithfulness of explanations of Neural GAM model with different sample sizes. Bold values indicate the most faithful explanations for each measure.

Sample Size	500		1000		2000		3000		5000	
Measure	RBO	Overlap	RBO	Overlap	RBO	Overlap	RBO	Overlap	RBO	Overlap
Smote	0.19	2.07	**0.22**	2.09	0.2	2.12	0.18	2.13	0.19	2.13
Gaussian	**0.21**	2.21	0.21	**2.27**	0.21	**2.3**	0.21	**2.3**	0.19	**2.31**
LHS	0.2	**2.23**	0.19	**2.27**	**0.23**	**2.3**	**0.23**	**2.3**	**0.23**	**2.31**
DLIME	0.2	2.14	**0.22**	2.26	0.21	2.26	0.22	2.27	0.21	2.29
LIRME	0.18	1.97	0.21	1.97	0.21	1.96	0.2	1.96	0.2	1.97
EXS (B)	0.11	1.96	0.12	1.97	0.15	1.96	0.11	1.98	0.16	1.94
EXS (R)	0.19	1.88	0.21	1.88	0.2	1.87	0.19	1.89	0.19	1.98
EXS (S)	0.18	1.96	0.19	1.96	0.19	1.96	0.2	1.96	0.22	1.96
LIME	**0.21**	1.97	0.19	1.98	0.18	1.97	0.2	1.96	0.2	1.97
SHAP	**0.21**	1.98	0.21	1.98	0.21	1.98	0.21	1.98	0.21	1.98
Random	**0.21**	1.91	**0.22**	1.89	0.19	1.95	0.19	1.88	0.2	1.93

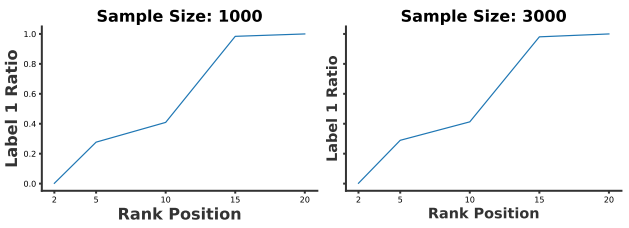

Fig. 7: SMOTE sampling: The average ratio of our generated samples obtaining label 1 when our labeling process is EXS (Top-K binary) for explaining test documents with predicted rank of 2.

for the sample size of 500 and SMOTE for the sample size of 2000, but only to a small degree. We consider this problem to arise in other explanation techniques and believe that future studies can investigate this problem further and propose alternative solutions to fix this problem.

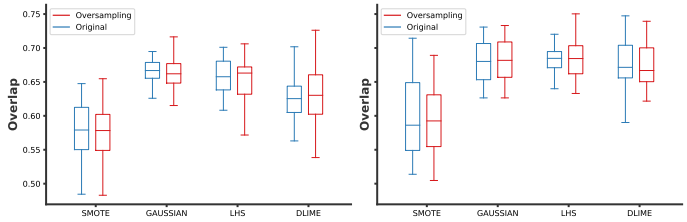

Fig. 8: The box plot showing the Overlap measure when using oversampling after our sampling process for sample sizes 500 (Left) and 2000 (Right). The results are for the explanation of documents ranked second in our Yahoo test dataset.

VII. Discussion

Based on our empirical investigation of Yahoo datasets, we can see the LIME-based explanations of LIRME and EXS fail to consistently provide faithful explanations based on the ground truth extracted from the Neural Rank GAM model.

We can identify a set of limitations of these techniques by means of comparison. There are two main differences between our proposed approaches and LIRME and EXS.

First is the sampling process. The sampling techniques of LIRME and EXS are quantile-based sampling. Quantile-based sampling replies on interpretable binary representations of tabular data. In our proposed LIME explanations, we have abandoned this step, and we can see a clear indication of improvement in the faithfulness of our local point-wise explanations.

The second difference between EXS and our proposed LIME explanations is that, unlike EXS with its linear SVM surrogate, we use LIME's original Ridge classifier.

We showed that the overlap between explanations of a single technique for two documents at predicted rank 2nd and 10th is a reliable indicator of the failure of LIRME and EXS explanation techniques. This can directly show that these techniques do not leverage the predicted scores information in their explanations.

VIII. Conclusion

In our study, we evaluated local point-wise explanations of a state-of-the-art LTR model, Neural Ranking GAM models. Given that this model has intrinsically interpretable components based on the Generalized Additive Model, we extracted the ground truth importance scores and evaluated local ex-

planations using two evaluation measures, namely RBO and Overlap AUC.

Overall, our proposed explanations provide the most faithful explanations across numerous measures, sample sizes, and predicted ranks of explained documents, except in a few cases. For documents with the predicted rank of two, EXS (S) provides the most faithful explanations based on RBO for sample sizes 2000 and 5000 and is on par with LIRME for sample size 3000. For documents with a predicted rank of ten, LIME, SHAP, and Random explanations are the most faithful based on the RBO measure.

Among all explanations, we showed that our proposed explanation based on SMOTE sampling excels at using the predicted rank information for obtaining its explanations. We showed this by calculating the Overlap of important features between the explanations of documents at the second and tenth rank in each query.

We showed that our proposed explanation technique suffers from a class imbalance problem. This phenomenon happens for the labels of generated samples when explaining documents with top or low ranks in a list of documents. We consider the sample imbalance problem to be an important challenge in providing faithful local explanations for LTR models. Even though our extra oversampling step showed small improvements in faithfulness, we consider this problem to be an open research problem in this domain.

References

[1] H. Li, "Learning to rank," in *Learning to Rank for Information Retrieval and Natural Language Processing*. Springer, 2009, pp. 1–9.

[2] H. Yang and T. Gonçalves, "Field features: The impact in learning to rank approaches," *Applied Soft Computing*, vol. 138, p. 110183, 2023.

[3] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux, "Model interpretability through the lens of computational complexity," *Advances in neural information processing systems*, vol. 33, pp. 15 487–15 498, 2020.

[4] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Systems*, vol. 28, no. 6, pp. 2335–2355, 2022.

[5] A. Zytek, I. Arnaldo, D. Liu, L. Berti-Equille, and K. Veeramachaneni, "The need for interpretable features: Motivation and taxonomy," *ACM SIGKDD Explorations Newsletter*, vol. 24, no. 1, pp. 1–13, 2022.

[6] H. Zhuang, X. Wang, M. Bendersky, A. Grushetsky, Y. Wu, P. Mitrichev, E. Sterling, N. Bell, W. Ravina, and H. Qian, "Interpretable learning-to-rank with generalized additive models," *arXiv preprint arXiv:2005.02553*, 2020.

[7] 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.

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

[9] J. Singh and A. Anand, "Exs: Explainable search using local model agnostic interpretability," 2018.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

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

[17] H. Zhuang, X. Wang, M. Bendersky, A. Grushetsky, Y. Wu, P. Mitrichev, E. Sterling, N. Bell, W. Ravina, and H. Qian, "Interpretable ranking with generalized additive models," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 499–507.

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

[19] 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.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 42, no. 1, pp. 55–61, 2000.

[22] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[23] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 4, pp. 1–38, 2010.

[24] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview," in *Proceedings of the learning to rank challenge*. PMLR, 2011, pp. 1–24.

[25] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[26] H. Yuan, M. Liu, L. Kang, C. Miao, and Y. Wu, "An empirical study of the effect of background data size on the stability of shapley additive explanations (shap) for deep learning models," *arXiv preprint arXiv:2204.11351*, 2022.