A DATASETS

We evaluate the performance of ShaRP and compare it to other local feature importance methods, using several real and synthetic datasets, with the corresponding ranking tasks. Dataset properties, along with ranker type (score-based or learned) are summarized in Table 1 and described below. We show the relationship between score and rank for score-based ranking tasks in Figure 9.

CSRankings (CSR) ranks 189 Computer Science departments in the US based on a normalized publication count of the faculty across 4 research areas: AI, Systems (Sys), Theory (Th), and Interdisciplinary (Int) [35]. We use publication data for 2013-2023, with the scoring function provided by CSRankings, a geometric mean of the adjusted counts per area, with # of sub-areas as exponent:

$$f = \sqrt[27]{(AC_{AI}^{5}+1)(AC_{Sys}^{12}+1)(AC_{Th}^{3}+1)(AC_{Int}^{7}+1)}$$

ATP Tennis (ATP) is based on publicly available 2020-2023 performance data of tennis players from the Association of Tennis Professionals (ATP) [19]. We use 2022 data that includes 5 performance-related attributes of 86 players. We select 2022 because this is the year in which data for all 5 attributes is available for the highest number of players. We use the following scoring function that we recovered from the ATP site using the scores:

$$f = 100 \times (\%\ \text{1st Serve}) + 100 \times (\%\text{1st Serve Points Won}) + 100 \times (\%\text{2nd Serve Points Won}) + 100 \times (\%\ \text{Service Points Won}) + 100 \times (\text{Avg Aces/Match}) - 100 \times (\text{Avg Double Faults/Match})$$

Times Higher Education (THE) is a dataset of worldwide university rankings [18]. It contains the university name, country, and the scores assigned to the university by Times Higher Education for teaching (TEA), research (RES), citations (CIT), income (INC), and international students (INT). We use 2020 data, for consistency with Anahideh and Mohabbati-Kalejahi [2] who also used it in their paper, with the scoring function provided by THE:

$$f = 0.3 \times TEA + 0.3 \times RES + 0.3 \times CIT + 0.025 \times INC + 0.075 \times INT$$

Moving Company. The moving company scenario [37] simulates a hiring process where job applicants are ranked based on their *qualification score*, computed as a function of their weight lifting ability, sex and race. We train two different rankers, over two scenarios:

(1) Using the original data from a previous hiring process from that company, where female applicants generally display lower weight-lifting ability than male applicants and a lower qualification score. In addition, black applicants have a lower qualification score compared to white applicants, but similar weight-lifting ability. Hence black females face greater discrimination compared to the rest of the applicants.
(2) After applying the intersectional fairness intervention proposed in the same paper over the data.

All versions of this dataset (both scenarios and train/test sets) contain 2000 tuples.

We use an eXtreme Gradient Boosting (XGB) and a Light Gradient-Boosting (LGB) Machine to model the rankings of the applicants

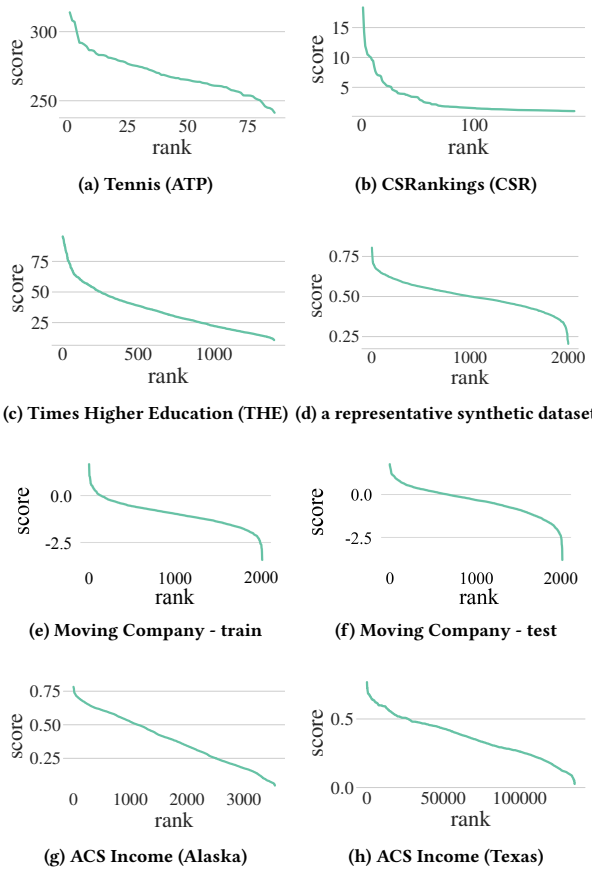


Figure 9: The relationship between an item's score (y-axis) and its rank (x-axis) for four score-based tasks.

in the training set, and infer and compute the feature contributions of the applicants in the test set, i.e., all results reported in this section correspond to the test set. The XGB ranker was defined with the pairwise ranking objective, while the LGB ranker uses the lambdarank objective.

ACSIncome. ACSIncome contains income-related data from adults in the US. It consists of 10 features: age, class of worker, educational attainment, marital status, occupation, place of birth, relationship to the reference person, work hours per week, sex, and race. The task is to predict whether the yearly income is over $50,000.

For this task we use a Random Forest Classifier (RFC) and rank the items based on the predicted probability of positive class membership.

Synthetic datasets. We also use numerous synthetic datasets to showcase specific quantitative and qualitative aspects of local feature-based explanations and metrics, and to study specific aspects of performance. These datasets contain 2,000 tuples. In five of them, items have 2 features, x_1 and x_2, distributed according to the uniform, Gaussian, or Bernoulli distributions, with varying parameters. We experiment with both independent and correlated features.

Each synthetic dataset consists of 2,000 items. We use three linear scoring functions: $f_1 = 0.8 \times x_1 + 0.2 \times x_2$, $f_2 = 0.5 \times x_1 + 0.5 \times x_2$, and $f_3 = 0.2 \times x_1 + 0.8 \times x_2$.

To explore correlations further, we create three datasets that have three Normal features x_1, x_2, and x_3, and 2,000 items. In the first dataset, all features are independent. In the second, we draw x_1 and x_2 from the 2D Gaussian and they are negatively correlated with correlation -0.8. The third feature x_3 is independent. For the third dataset, we draw the features from the 3D Gaussian. x_1 and x_2 are negatively correlated with correlation -0.8, x_1 and x_3 are positively correlated with correlation 0.6, and x_2 and x_3 are negatively correlated with correlation -0.2. For all three datasets, we use the same scoring function $f_4 = 0.33 \times x_1 + 0.33 \times x_2 + 0.34 \times x_3$.

B DISTRIBUTIONAL ANALYSIS FOR RANKING

Fixed scoring function, varying data distribution. In this experiment, we illustrate that feature importance is impacted by the data distribution of the scoring features to a much greater extent than by the feature weights in the scoring function. Further, we show that feature importance varies by rank stratum. In Figure 10, we show rank QoI for 4 synthetic datasets with the same scoring function f_2.

We observe that, while the features have equal scoring function weights, their contributions to rank QoI differ for most datasets. In D_1, the Bernoulli-distributed x_2 determines whether the item is in the top or the bottom half of the ranking, while the Gaussian-distributed x_1 is responsible for the ranking inside each half. For D_2, the uniform x_1 has higher importance because it often takes on larger values than the Gaussian x_2. In D_4, x_1 and x_2 are negatively correlated, so when one contributes positively, the other contributes negatively. Only for D_3, with two uniform identically distributed features, the median contributions of both features are approximately the same within each stratum.

Additionally, we see that feature contributions differ per rank stratum. For example, for D_3, the medians show a downward trajectory across strata. This is because they quantify the expected change (positive or negative) in the number of rank positions to which the current feature values contribute. Also for D_3, feature contributions have higher variance in the middle of the range, because a 40-60% rank corresponds to many feature value combinations.

Fixed data distribution, varying scoring function. In this experiment, we investigate the impact of the scoring function on rank and top-k QoI for two datasets. In Figure 11, we use D_3 and see that the contributions to rank QoI vary depending on the scoring function. For f_1, x_1 is the only important feature (although it carries 0.8 — and not 1.0 — of the weight). This can be explained by the compounding effect of the higher scoring function weight and higher variance of the distribution from which x_1 is drawn. Between f_2 and f_3, features x_1 and x_2 switch positions in terms of importance, and show a similar trend, despite being associated with different scoring function weights (0.5 & 0.5 vs. 0.2 & 0.8). This, again, can be explained by the higher variance of x_1, hence, x_2 needs a higher scoring function weight to compensate for lower variance and achieve similar importance.

Top-k access. Access to the top-k is determined by the interaction between the scoring feature weights and the distributions of these features. The top-k QoI tells us how important each feature is when we consider only access to the top-k. A positive feature contribution signifies that changing the feature's value will result in decreased

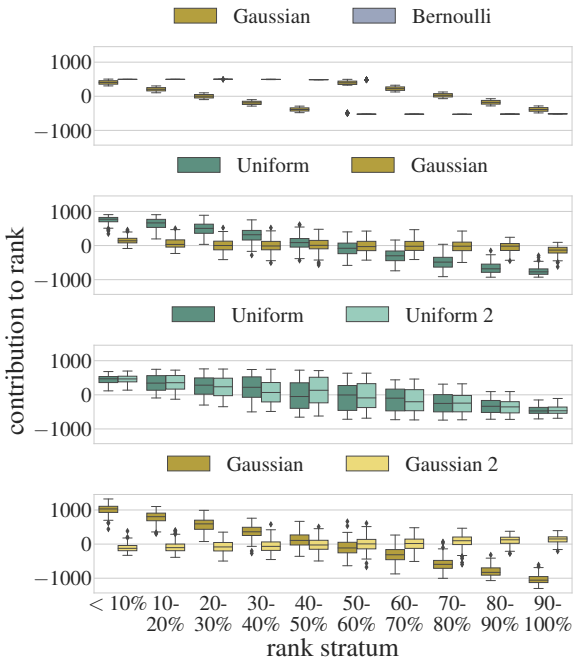


Figure 10: The rank QoI using $f_2 = 0.5 \times x_1 + 0.5 \times x_2$ for four datasets; D_1: $x_1 \sim N(0.5, 0.1)$, $x_2 \sim Bern(0.5)$; D_2: $x_1 \sim [0, 1]$, $x_2 \sim N(0.5, 0.1)$; D_3: $x_1 \sim [0, 1]$, $x_2 \sim [0, 1]$; D_4: $x_1 \sim N(0.5, 0.05)$, $x_2 \sim N(0.75, 0.016)$, with -0.8 correlation. Feature contributions are different per rank stratum and data distribution.

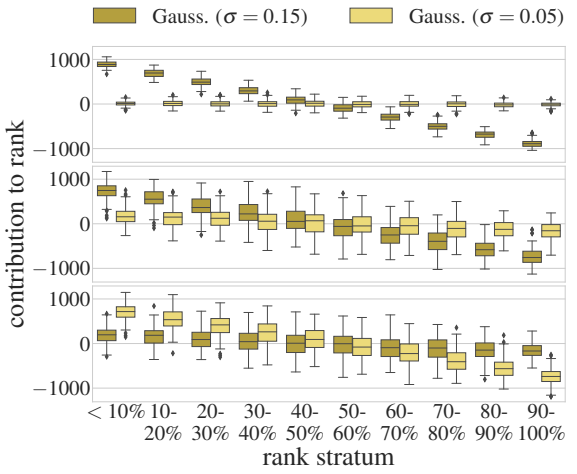


Figure 11: Rank QoI for D_5: $x_1 \sim N(0.5, 0.1)$, $x_2 \sim N(0.5, 0.05)$. Subplots correspond to different scoring functions: $f_1 = 0.8 \times x_1 + 0.2 \times x_2$ (top), $f_2 = 0.5 \times x_1 + 0.5 \times x_2$ (middle), $f_3 = 0.2 \times x_1 + 0.8 \times x_2$ (bottom).

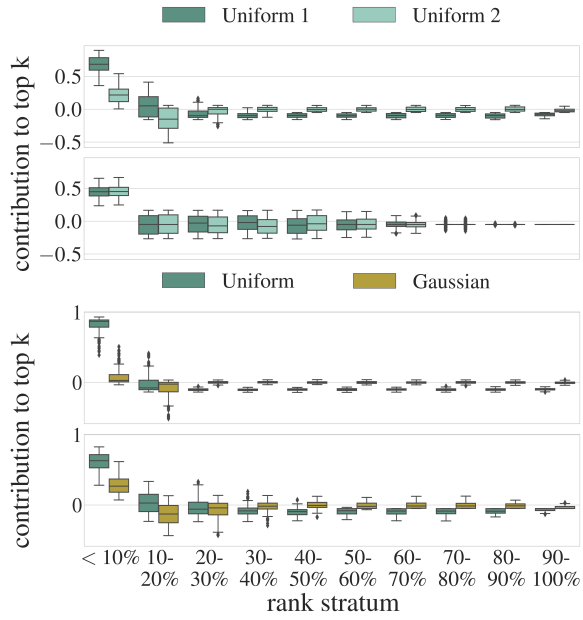


Figure 12: Top-k QoI for $k = 10\%$, D_2: $x_1 \sim [0, 1]$, $x_2 \sim N(0.5, 0.1)$, and D_3: $x_1 \sim [0, 1]$, $x_2 \sim [0, 1]$. Subplots correspond to different scoring functions: $f_1 = 0.8{\times}x_1{+}0.2{\times}x_2$ (top), $f_2 = 0.5{\times}x_1{+}0.5{\times}x_2$ (bottom).

chances of getting to the top-k. A very high (or very low) value shows that the changes are significant. Figure 12 illustrates this for datasets D_2 and D_3. When we consider two identical uniform features that have equal weights (D_2 under f_2), we first notice that their control of top-k access is identical as expected. Additionally, we see that for the top-10, changing either feature would reduce access to the top-k (the values are both very positive). However, for each stratum up to the top-70%, changing either feature can contribute either positively or negatively.

When we consider two identical uniform features (D_1) one of which has a higher weight (f_1) or dataset D_2 (under either f_1 or f_2), we see again that for the top-10, changing either feature would reduce access to the top-k. Also, we see that how high the contributions are depends on the distributions. However, we see that for the top 10%-20% changing the second, less important feature would increase the chances of getting into the top-k. For the rest of the strata, with some variations depending on the dataset and function, changing the most important feature provides non-zero probability of moving to the top-k and interestingly, this persists even for the lower strata. Evidence that items from lower strata can move to the top-k under some scoring functions and feature distributions counters the assumption of Anahideh and Mohabbati-Kalejahi [2] that changes in rank are localized.

C RANK-QOI-BASED AND SCORE-QOI-BASED EXPLANATIONS FOR CSRANKINGS

In Section 4 we discuss the differences between the rank QoI and the score QoI for the CSRankings dataset. In this section, we provide additional details for this comparison. Specifically, we demonstrate that considering different outcomes as profit functions has a profound impact on the explanations for the entire range of the ranking.

In Figure 13, we provide local Shapley value explanations for fifty universities from the CSR dataset for both the rank (Fig. 13a) and the score QoI (Fig. 13b). These universities are randomly chosen; they are approximately 25% of the dataset and span the entire ranking. Each subplot in each subfigure shows one explanation for one university, and its title shows each university's rank and score (the score is in parentheses). The universities are the same across both subfigures.

Looking at this collection of explanations, we can see how the rank and the score QoI behave significantly differently. Matching what we showed in Figure 3, the score QoI explanations become indistinguishable as we move down the ranking. Additionally, the contributions of all features become negative around rank 61 for the score QoI as opposed to 131 for the rank QoI. Finally, for the score QoI the contributions are very small for almost all universities as opposed to the rank QoI where the contributions are small for the middle of the ranks.

There are two main reasons why the behavior between the rank and the score QoI based explanations is so different. The first is that the *score-to-rank relationship is exponentially decreasing for this dataset* (see Fig.9b). This means that the top of the ranking has very high scores, and the scores quickly reach a plateau. The second is that *Shapley values explain the contribution of each feature to the distance of the outcome from the mean outcome.*

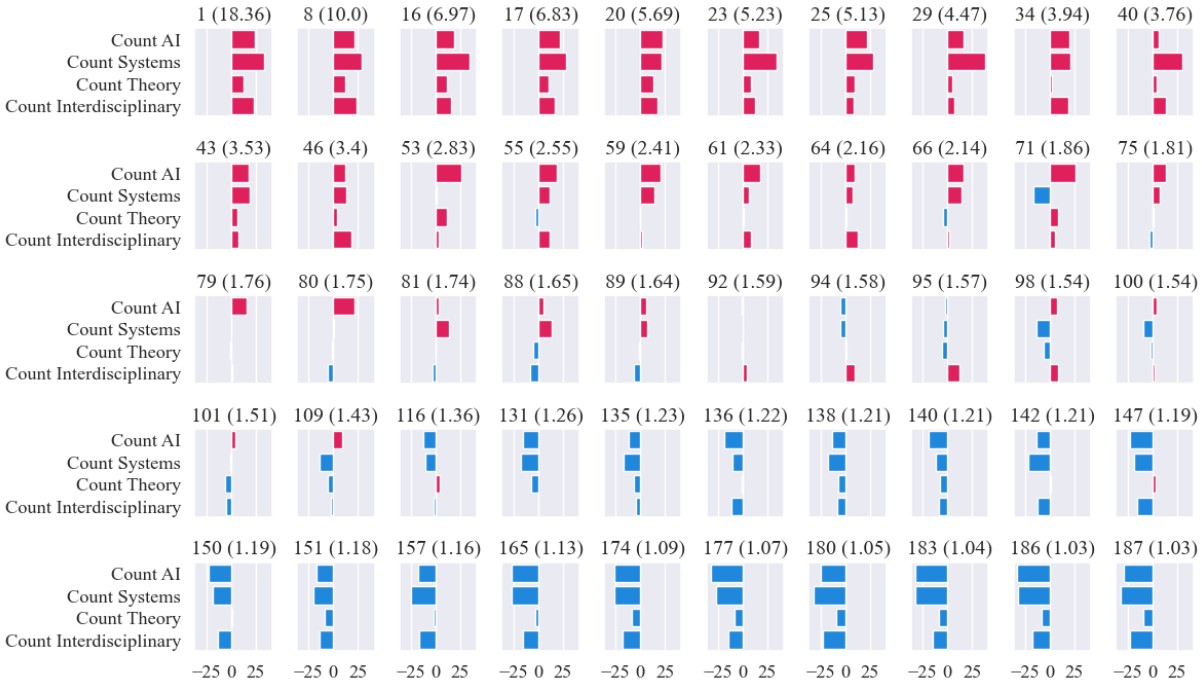
Indeed, the mean score for this dataset is 2.72, and its range is 18.36-1.03, while the mean rank is 95, and its range is 1-198. Together, these two facts mean that for the score QoI, for most items, the distance between its score and the mean score is very small. Because the score-based explanation explains the difference from the mean score, and those differences are very small for most items, the contributions are *very low* for most items. Additionally, the mean score is very influenced by the outliers at the top, so most items have *negative* contributions for all their features, even when ranked in the top 30% (e.g., the university ranked at position 61).

As discussed in multiple works, for instance [4, 24] explanations should differ when the outcomes and the items are different. In these figures, we can see that this is not the case for the score QoI-based explanations. Items ranked in the middle of the ranking (e.g., item ranked in position 92) have similar explanations to items ranked in the bottom of the ranking (e.g., item ranked in position 183).

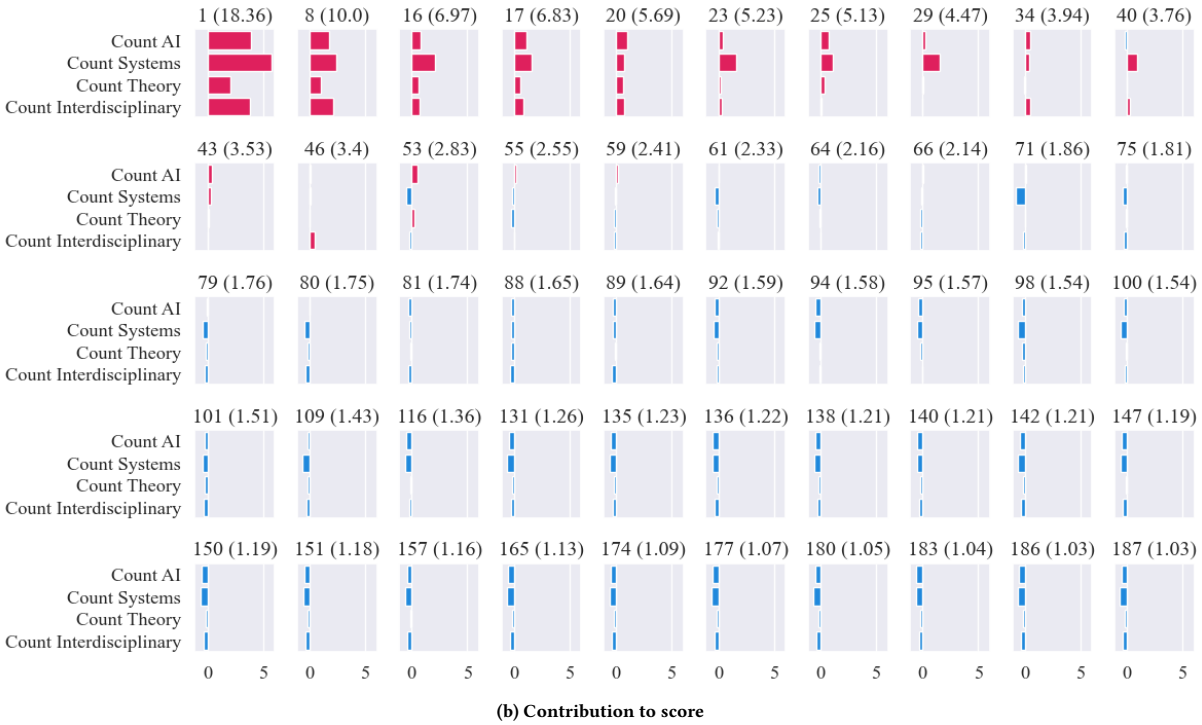
This behavior of the local explanations, coupled with the fact that the score QoI is not able to know when the rank changes (see Section 1), argues for using the rank QoI when explaining rankings.

D IMPLEMENTATION OF HIL

HIL [39] is the only other method that recommends the usage of ranks as a profit function for individual explanations in ranking.



(a) Contribution to rank



(b) Contribution to score

Figure 13: Shapley value explanations for fifty random universities for the rank QoI and the score QoI. The title of each subplot indicates the rank of each university and also contains its score in parentheses. The exponentially decreasing score-to-rank relationship and the dependence of Shapley values on the mean make score explanations indistinguishable and negative for most of the items.

While this method is not general, we are interested in comparing it with our rank QoI. This was not straightforward because the method is available as a web app that works only for linear weight scoring functions and datasets of two Gaussian features. To compare the rank-relevance contributions introduced in that paper to the rank QoI, we adapted their method using their definitions and code. This implementation is available alongside our own. Further, we extended their method to work with the specific non-linear scoring function used by CS Ranking, by changing the way that Std rank and Std score (discussed below) are computed.

More specifically, because HIL [39] works only with linear weight scoring functions, they do not provide a full Shapley values implementation but use the linear weights to approximate Shapley values assuming feature independence, see Corollary 1 in [21] and also [34]. This is a well-established method to compute Shapley values for linear weights also implemented by SHAP, so we do not compare with this part of the method. In addition, HIL defines two methods to acquire feature contributions: "standardized Shapley values" and "rank relevance Shapley values," which we will call Std score and Std rank, respectively. Those are not calculated using the linear weight method described above, but rather directly from the weights, and without using the mean score or rank. For an item $\mathbf{v}$, each feature i contribution for Std score is $\phi_i = \frac{\beta_i \mathbf{v}_i}{\sum_{\mathbf{u}\in\mathcal{D}} f(u)}$, where β_i is the weight for feature i. In other words, the contribution of each feature for each item is the score contribution of this feature over the sum of all scores for all items. Similarly, for Std rank, the contribution of feature i for an item $\mathbf{v}$ is $\phi_i = \beta_i \mathbf{v}_i \alpha_{\mathbf{v}}$, where $\alpha_{\mathbf{v}}$ is a scaling factor used to transform the score of the specific item to the rank of the specific item calculated as $\alpha_{\mathbf{v}} = \frac{(\max_{r\in r_{\mathcal{D}}}(r) - r^{-1}_{\mathcal{D}(\mathbf{v})}) \sum_{\mathbf{u}\in\mathcal{D}} f(u)}{\max_{r\in r_{\mathcal{D}}}(r) f(\mathbf{v})}$. Note that neither of the two formulas is computing Shapley values; rather, they assign a contribution to the features based on the linear weights, and the score and rank. This implies that our rank QoI is the only rank QoI for Shapley values.

E ADDITIONAL DETAILS ON METHOD COMPARISONS

E.1 Fidelity

We provide more details on the Fidelity results discussed in Section 8.2.2. We compute the Fidelity of all the methods that have that property across all datasets. We use SHAP and LIME out-of-the-box so their performance is not perfect (although extremely good). We make this choice to highlight the importance of using exact Shapley values when computing local explanations, where the error in each separate explanation is important as each explanation impacts a separate person.

E.2 Agreement between Explanations

Figure 14 presents agreement between ShaRP and all other methods averaged across all datasets. We use rank and score QoIs for this comparison, as they match those used by the methods we evaluate. Kendall's tau distance is computed to enable cross-method comparisons. We observe that explanations vary significantly by method, regardless of the QoI. ShaRP aligns most closely with LIME and SHAP across both rank and score QoIs. HRE, which relies on localized information, naturally differs. However, even among HRE

Table 4: Fidelity across all methods across all datasets.

dataset	LIME score	SHAP score	ShaRP score	ShaRP rank	HIL score
ATP	0.98	1.00	1.00	1.00	0.14
CSR	0.95	0.99	1.00	1.00	0.85
THE	0.94	0.97	1.00	1.00	0.64
Syn 0	0.95	0.99	1.00	1.00	0.37
Syn 1	0.95	0.99	1.00	1.00	0.29
Syn 2	0.95	0.99	1.00	1.00	0.35

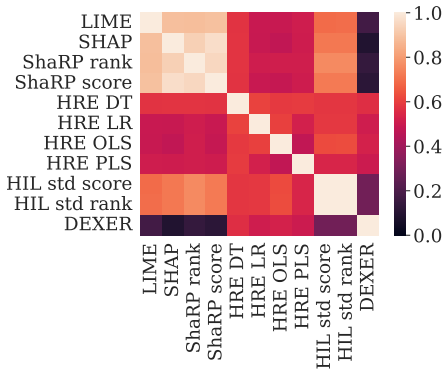


Figure 14: Method agreement averaged across all datasets

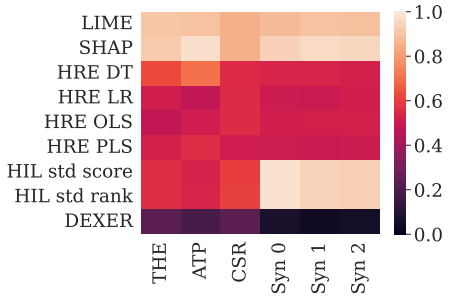
variants, explanations remain inconsistent. The two HIL methods and the two ShaRP methods produce similar explanations despite using different QoIs, suggesting that explanation consistency depends more on the method than the QoI. In contrast, DEXER, which fits a linear regression to the ranking output and applies SHAP, differs greatly from all methods, indicating that rank cannot be effectively explained without a rank QoI.

Figure 15 provides a per-dataset visualization of the agreement between the explanations of the methods in Section 8.2.

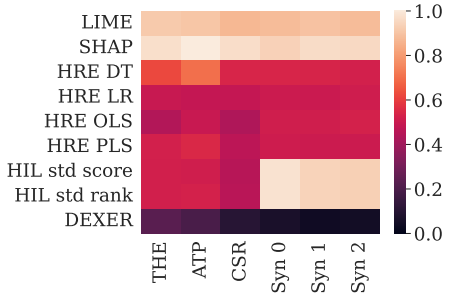
In Fig. 15a We visualize Kendall's tau explanation distance correlation of ShaRP using the rank QoI with all other methods across every dataset. In Fig. 15b, we plot the same result for ShaRP using the score QoI. As in the aggregated method agreement plot (Fig. 14), ShaRP is very similar to SHAP and LIME for both QoIs. As expected it is more similar for the score QoI but not identical that is perhaps because we used SHAP out-of-the-box which applies some approximation parameters for running time optimization. Similarly, ShaRP behaves similarly to what we discussed in Fig. 14 to all methods across the datasets except the HIL methods for the Synthetic datasets. We hypothesize that this is because the HIL methods are able to perform better for those datasets due to their Synthetic nature.

E.3 Sensitivity

We provide the sensitivity analysis for all methods for CSRankings in Fig. 16. In addition, to HRE LR, DEXER, LIME, SHAP, ShaRP rank, and HIL std rank that we presented in Fig. 7, we also plot HRE DT in Fig. 16a, HRE OLS in Fig. 16c, HRE PLS in Fig. 16d,



(a) Method Agreement between Sharp using the rank QoI and all methods across all datasets



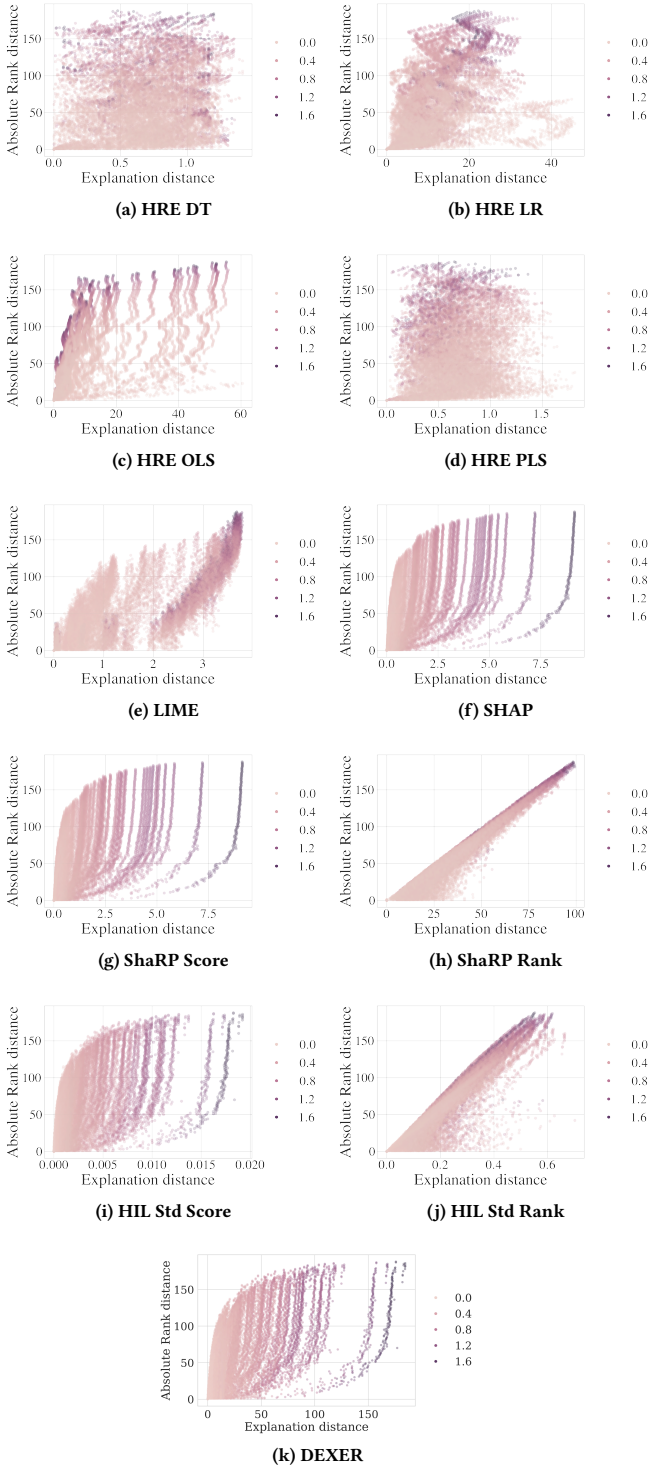
(b) Method Agreement between Sharp using the score QoI and all methods across all datasets

Figure 15: Method Agreement

ShaRP score in Fig. 16g, HIL std score in Fig. 16i. We see that all HRE methods perform similar or worse than HRE-LR, this is unsurprising as all these methods are used locally. We also see that both HIL std score and ShaRP score perform similar to SHAP, this is also expected. HIL std score and DEXER are very similar which is revealing our inability to predict the rank using the ranked output of the model. Specifically, HIL std score is assuming knowledge of the weights used by the model and uses them directly to compute the feature importance. DEXER is assuming black-box access to the ranked output only and fits a linear regression model on the ranking. Nevertheless, judging from these results, it appears that DEXER is explaining the score (and not the rank) and is learning the model weights to do so. Inadvertently, we also show that the choice of the explanation method makes a big difference to the final explanation.

To further compare ShaRP with rank QoI and HIL with Std rank, we present Figure 17. Even though both methods are appropriate for the ranking task we are examining, in this figure, we see that ShaRP with rank QoI (Figures 17a, 17c, and 17e) can capture the full range of different ranks and features, and that groups the items more successfully. HIL with Std rank cannot capture the difference of feature values for ATP (Figure 17b), or the similarly ranked items that have different feature values for the Synthetic experiment (the middle area close to the x-axis of Figure 17f. Both methods perform similarly for THE (Figures 17c and 17d).

Finally, we present an analysis of ShaRP using the score QoI and the rank QoI for the CSR dataset but for a score task (instead of



(a) HRE DT **(b) HRE LR**

(c) HRE OLS **(d) HRE PLS**

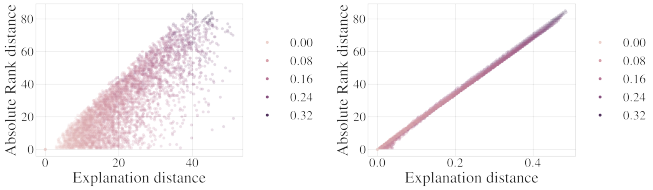
(e) LIME **(f) SHAP**

(g) ShaRP Score **(h) ShaRP Rank**

(i) HIL Std Score **(j) HIL Std Rank**

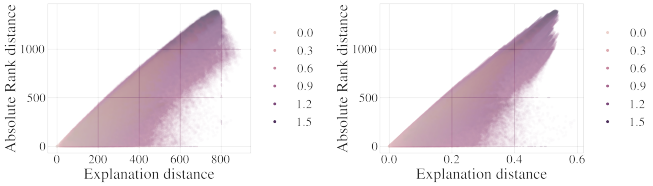
(k) DEXER

Figure 16: Sensitivity analysis for the CSRankings dataset for all methods.



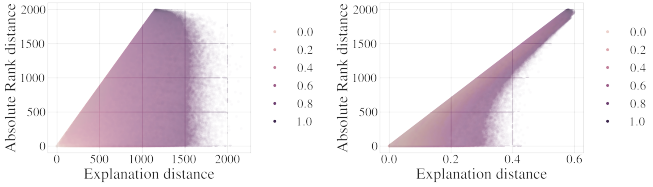
(a) ShaRP Rank ATP

(b) HIL Std Rank ATP



(c) ShaRP Rank THE

(d) HIL Std Rank THE



(e) ShaRP Rank Synthetic-0

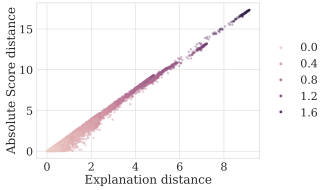
(f) HIL Std Rank Synthetic-0

Figure 17: Sensitivity analysis for the ATP, THE, and Synthetic dataset 0 for the methods using the rank QoI.

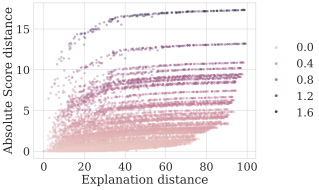
rank). The goal of this analysis is to show that the sensitivity of the methods that use a score QoI is very high when we are explaining a score task. In other words, if we are trying to explain the score, then the methods that use a score-based profit function perform the best as it is fully expected.

The task we are going to explain is the score of the CSRankings scoring function. We choose this task for two reasons, first we already provided the results of the CSRankings ranking task and we can draw a direct comparison. Secondly, we have a ranking for that dataset and we can plot the methods that use the rank QoI for juxtaposition. Note that it is entirely redundant to use a rank-based QoI method in this case. In fact, it is redundant to even produce a ranking as we are asking an explainability question about the score. But we are choosing to provide this information to showcase that each explainability task needs each own profit function and the choice of the profit function makes a big difference to the final explanation.

In Fig. 18, we evaluate the similarity of explanations for pairs of similar items *when we attempt to explain the score*. For each pair of items, we compute three distances: (1) Euclidean distance between the explanations (x-axis); (2) distance between the *scores* (instead of rank) of the two items (y-axis); and (3) Euclidean distance between the items in terms of their feature values (hue, where lighter means closer). To make the plot, we place one item (the reference item) at position (0,0) and use a scatter point for each other item (neighbor), indicating the distance in ranks and the distance of the explanations. The color of the scatter point indicates the distance between the



(a) ShaRP Score



(b) ShaRP Rank

Figure 18: Sensitivity analysis for the CSRankings dataset for all methods when the task we are trying to explain is a score task. Compared to Figure 16 we see that the methods that are using a score QoI are performing better.

features of the reference point and the neighbor. We then overlay the plots for all items in the dataset, so that all items are used as reference points.

Unlike Fig. 16, we now expect to see items that are both similar in terms of their features and *scored* near each other to have similar explanations. We would still expect all points to be on or near the diagonal line $y = x$, with the hue getting darker as we move away from the origin, *if their explanations successfully explain the score.*

In Fig. 18, we see that indeed the score-based method has the desired shape we discussed in Section 8.2.1. ShaRP score is extremely similar and almost entirely fits the $y = x$ line. ShaRP rank, appears to be providing explanations that do not depend on the score distance between the items' outcomes (y-axis) or the feature distance between the items (hue) as expected.

This analysis shows how QoI selection is important when providing an explanation. The score is unable to perform well for a ranking task since it estimates the impact of each feature on the score outcome and similarly, it is completely unreasonable to use a rank QoI when explaining the score.

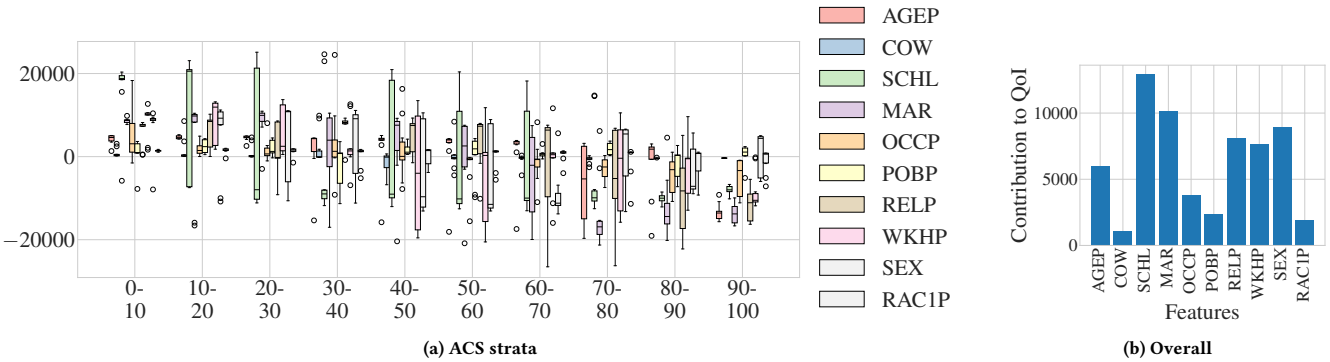
F ADDITIONAL RESULTS FOR ACSINCOME

In Figure 19, we present the overall and strata results for the second ACSIncome dataset we used, Texas, that was previewed in Section 8.1. As discussed in that section, the feature importance shifts notably compared to Alaska, shown in Figure 6. The biggest changes are in age (AGEP), education (SCHL), work hours per week (WKHP), and race (RAC1P). These differences highlight the usefulness of explanations, the necessity of working with multiple subsets of similar data, and the ability of our method to capture distributional shifts.

G ADDITIONAL RESULTS ON EFFICIENCY AND APPROXIMATION

In this section, we present the extended results previewed in Section 8.3.2.

In Table 5 we include the running times of ShaRP for ACSIncome, AK, when varying the maximum coalition size or the sample size. As discussed in 8.3.2, we include both cold and warm start results, and the fidelity for each setting. Fidelity is high for any sample size for this dataset, and while it declines more when varying the coalition size, it remains over 0.8 for both the score and rank QoIs



(a) ACS strata

(b) Overall

Figure 19: Feature contribution on ACS Income (Texas) to the rank QoI (a) across strata and (b) overall.

Table 5: Time experiment results over the ACS Income (Alaska) dataset. Running times (reported in seconds) for varying coalition sizes are measured using a fixed sample size of 100, while running times for varying sample sizes are measured using a fixed coalition size of 9. All results are reported by averaging results over 10 tuples, 3 runs each.

		Rank			Score		
max coal. size	sample size	Time (cold)	Time (warm)	Fidelity	Time (cold)	Time (warm)	Fidelity
1	100	1.98	1.87	0.810	0.37	0.17	0.850
3	100	6.46	3.14	0.857	5.410	1.41	0.887
5	100	18.57	5.82	0.904	17.70	4.13	0.924
7	100	24.51	7.06	0.951	23.34	5.41	0.961
9	100	24.86	7.18	0.991	23.66	5.52	0.993
9	20	45.10	12.60	0.994	43.35	11.02	0.996
9	50	95.27	28.79	0.995	92.87	27.58	0.996
9	100	160.19	55.80	0.997	154.56	54.54	0.997
9	250	292.22	137.50	0.998	282.04	136.82	0.999
9	500	445.00	271.93	0.999	428.96	270.44	0.999
9	1,000	708.37	542.77	0.999	689.55	536.26	0.999
9	3348	1,956.79	1,830.24	1.000	1,960.67	1,816.78	1.000

for any coalition size and is over 0.9 for both QoIs for coalition size 5 and above.

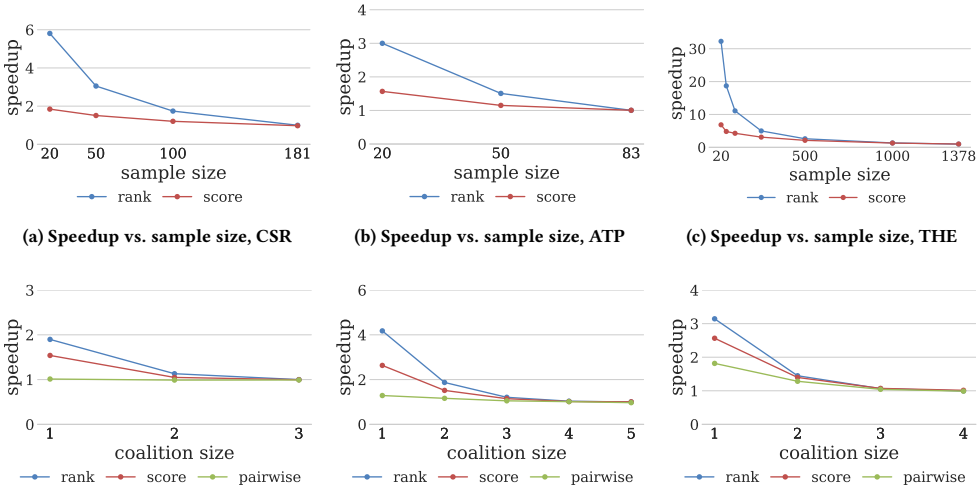
In Figure 20 we present the speedup vs. sample size, and speed-up vs max coalition size for THE, CSR, and ATP. We already presented the results for ACSIncome, AK in Figure 8. We observe similar results but scaled down due to the dataset sizes. In Figure 21, we present the corresponding fidelity for both sample size and max coalition size. We observe that fidelity is very high for all sample sizes, and almost identical or better to the fidelity of ACSIncome, AK for all max coalition sizes.

In Figures 22 and 23, we present the method agreement between the approximation and the exact computation for CSRankings (CSR). We omit method agreement results for the other datasets, where ShaRP performs similarly. In 22a and 22b, we present agreement of the approximation when we vary sample size for the rank and the score QoI. We evaluate the agreement using the Jaccard Index (considering the top-2 features), Kendall's tau distance, and the Euclidean distance of the feature vectors (converted to unit vectors). Here, we see that performance is similar for both QoIs. The Jaccard index is over 0.9 for any sample size, and is the distance metric with the worst performance for both QoIs. This is worth noting as shorter explanations are often considered more interpretable [24]. Agreement is similar or higher for all QoIs when we vary maximum coalition size, see Figure 23a- 23c.

H FOCUS GROUP PROTOCOL AND RESULTS

In this section we provide more details on the focus group study described in Section 9.

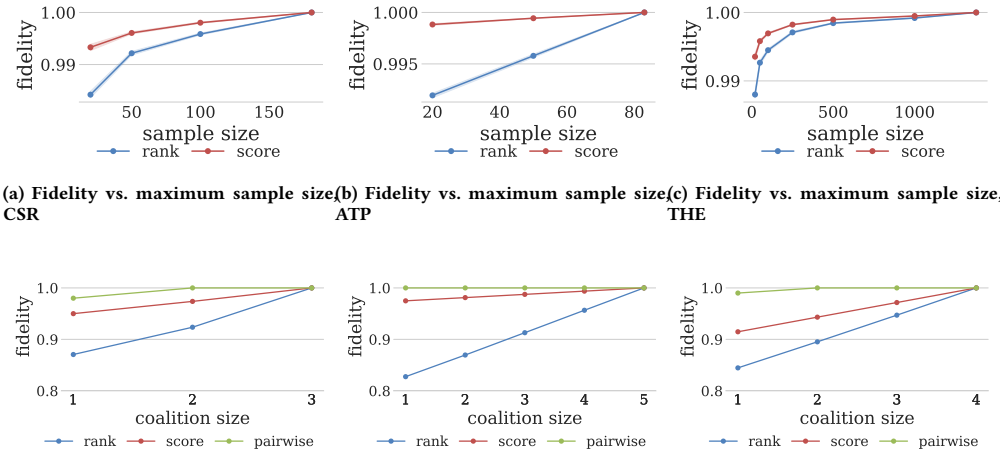
The goal of the focus group was to evaluate the usability of rank-based and score-based explanations. We conducted the focus group between members of our institution. For this reason, we chose CSRankings as the dataset since we assumed that it will be of interest to the participants. To understand the user understanding of group-based and rank-based explanations, we randomly selected a subset of the CSRankings schools, we produced explanations for each school using either the score or the rank QoI, we divided



(a) Speedup vs. sample size, CSR (b) Speedup vs. sample size, ATP (c) Speedup vs. sample size, THE

(d) Speedup vs. maximum coalition size, CSR (e) Speedup vs. maximum coalition size, ATP (f) Speedup vs. maximum coalition size, THE

Figure 20: Computational time performance of approximation for CSRankings (CSR), ATP Tennis (ATP), and Times Higher Education (THE). Speedup is computed in comparison to exact computation times, reported in Table 2.



(a) Fidelity vs. maximum sample size, CSR (b) Fidelity vs. maximum sample size, ATP (c) Fidelity vs. maximum sample size, THE

(d) Fidelity vs. maximum coalition size, CSR (e) Fidelity vs. maximum coalition size, ATP (f) Fidelity vs. maximum coalition size, THE

Figure 21: Fidelity of approximation for CSRankings (CSR), ATP Tennis (ATP), and Times Higher Education (THE) using different sample sizes and maximum coalition sizes.

the participants to two groups Score-Group and Rank-Group, and presented each group with a series of identical questions about the score or the rank explanations correspondingly.

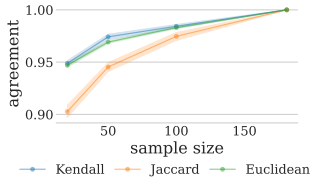
In this section, we detail the study protocol in Subsection H.1 and then we present the extended results in Subsection H.2.

H.1 Study Protocol

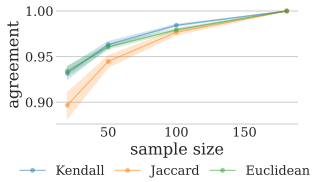
The study consisted of the four parts listed below. In this section we provide details for each part.

(1) Enrollment form
(2) Introductory document
(3) Score-based or rank-based tasks
(4) Exit discussion

Enrollment form. The enrollment form collected the educational background of the participants (optional text box), their highest academic degree (BS/BA, MS/MA, PhD, Other), their field of study (required text box), their relevant background (text box optional), their familiarity with AI explainability (scale 1-5, where 1 means

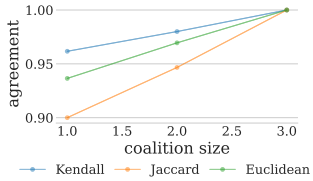


(a) Agreement of ShaRP across different sample sizes using maximum coalition size for rank QoI

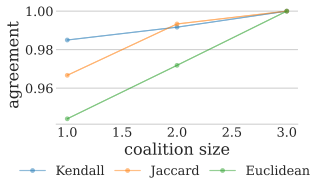


(b) Agreement of ShaRP across different sample sizes using maximum coalition size for score QoI

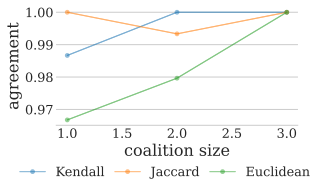
Figure 22: Agreement of ShaRP for CSRankings when varying the sample size.



(a) Agreement of ShaRP across different coalitions sizes using maximum sample size for rank QoI

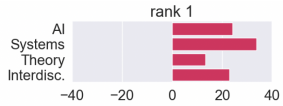


(b) Agreement of ShaRP across different coalitions sizes using maximum sample size for score QoI



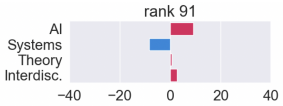
(c) Agreement of ShaRP across different coalitions sizes using maximum sample size for pairwise QoI

Figure 23: Agreement of ShaRP for CSRankings when varying the coalition size.

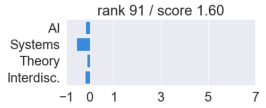


(a) Rank-Group category (i) visualization 1

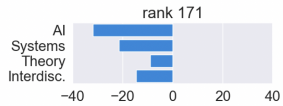
(b) Score-Group category (i) visualization 1



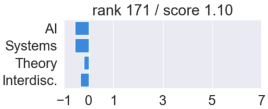
(c) Rank-Group category (i) visualization 2



(d) Score-Group category (i) visualization 2



(e) Rank-Group category (i) visualization 3



(f) Score-Group category (i) visualization 3

Figure 24: Focus group example figures for questions of type (i): understanding the rank of a specific department

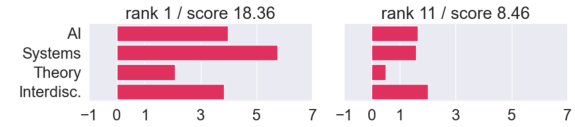
unfamiliar and 5 expert), their familiarity with Shapley value-based methods (scale 1-5, where 1 means unfamiliar and 5 expert), and their familiarity with the CSRankings dataset (scale 1-5, where 1 means unfamiliar and 5 expert).

Introductory document. We provide the introductory document in Section I and briefly summarize it here. The introductory document provided a short description of algorithmic rankers, the CSRankings dataset, and ShaRP , and then proceeded to explain the task. To explain the task, we included three the figures, similar to the figures used in the tasks. As we explain in the next paragraph, the task involves interpretation of individual or sets of Shapley value explanations. So, using the figures we provided information on how to read Shapley value explanations such as, distinguishing the features that negatively or positively impact the outcome, understanding the feature importance's magnitude, understanding the metric-unit of the explanation (which depends on the QoI), and finally the Shapley value efficiency property.

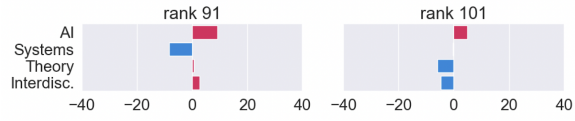
Tasks. The tasks consisted of three categories. Each category had a different objective and different questions. The categories were (i) understanding the rank of a specific department (3 departments, 4 questions for each), (ii) understanding why one department is ranked higher than another (3 department pairs, 2 questions for each), and (iii) understanding feature importance trends across the ranking (2 sets of 6 departments, 2 questions for each). To select the items presented in the study, we sampled 9 universities from CSRankings, 3 from the top, 3 from the middle, and 3 from the bottom of the ranking at random. We generated explanations for all of them using our method and plotted them on the same axes so they are comparable.

Figure 24 contains one example of the images used in the study for the questions of type (i) for both Rank-Group (left column) and

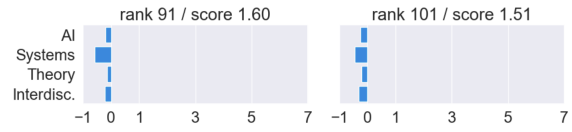
(a) Rank-Group category (ii) visualization 1



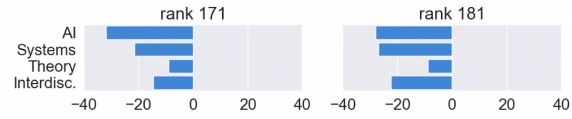
(b) Score-Group category (ii) visualization 1



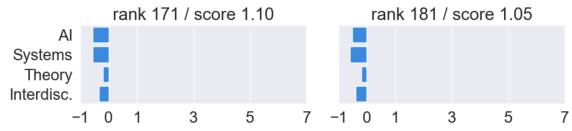
(c) Rank-Group category (ii) visualization 2



(d) Score-Group category (ii) visualization 2



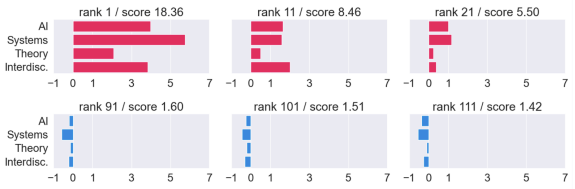
(e) Rank-Group category (ii) visualization 3



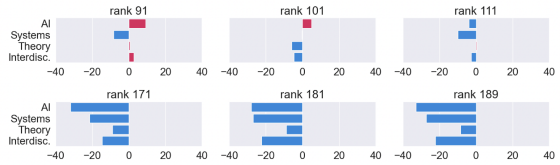
(f) Score-Group category (ii) visualization 3

Figure 25: Focus group example figures for questions of type (ii): understanding why one department is ranked higher than another

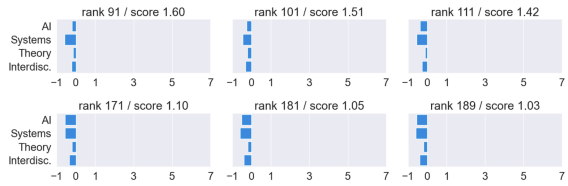
(a) Rank-Group category (iii) visualization 1



(b) Score-Group category (iii) visualization 1



(c) Rank-Group category (iii) visualization 2



(d) Score-Group category (iii) visualization 2

Figure 26: Focus group example figures for questions of type (iii): understanding feature importance trends across the ranking

Score-Group (right column). Each image was presented separately accompanied by four questions. Each question was followed by a 5-Likert-scale confidence question. The first question for this category asked the participants to select the feature that contributed to the department being at its respective rank the most *overall*. The second asked for the feature that contributed the least *overall*. The third, for the feature that contributed the most *positively*. And, finally, the fourth one for the feature that contributed the most *negatively*. All questions asked the participants to select the correct answer among the options. The options listed all features (AI, Systems, Theory, Interdisciplinary) and also included "Don't know" as an option. The last two questions also had "No features contributed positively/negatively" as an option.

Figure 25 contains one example of the images used in the study for the questions of type (ii) for both Rank-Group (left column) and Score-Group (right column). Each pair of images was presented separately, accompanied by two multiple-choice questions. Each question was followed by a 5-Likert-scale confidence question. The first question for this category asked the participants to select all features that were helping the department on the left outrank the department on the right. The second asked for the participants to select all features that were hurting the department on the left relative to the department on the right. The answers listed all features (AI, Systems, Theory, Interdisciplinary), "None," and "Don't know."

Finally, Figure 26 contains one example of the images used in the study for the questions of type (iii) for both Rank-Group (left column) and Score-Group (right column). Each group of images was presented separately, accompanied by two multiple-choice questions. Each question was followed by a 5-Likert-scale confidence

question. The first question for this category asked the participants to select up to 2 features that were helping the departments in the top row the most in comparison to the departments in the bottom row. The second asked for the participants to select the features that were hurting the rank/score of the departments on the bottom row the least in comparison to the departments in the top row. The answers listed all features (AI, Systems, Theory, Interdisciplinary), and "Don't know."

Discussion. The last part of the focus group study was an open discussion that lasted approximately 30 minutes. During the discussion, prompting questions were asked and the participants were encouraged to expand on their thoughts. The prompting questions were the following:

(1) What are your impressions of the explanations you just reviewed?
(2) Do you feel the explanations provided sufficient information to answer the questions accurately?
(3) Is there any additional or alternative information you would have preferred to receive?
(4) Do you have any other comments or feedback regarding the explanations or your overall experience?

H.2 Results

We found that Score-Group performed worse than Rank-Group. The results are presented in detail in Table 6. Rank-Group participants managed to answer correctly 73% of the time, in contrast to 67% for Score-Group participants. Additionally, Score-Group participants were less confident in their answers, 4.15/5 and 3.90/5 measured in a 5-Likert scale, for Rank-Group and Score-Group respectively.

Looking at the results per question category, we see that Rank-Group performed better for the questions of category (i) and (ii), scoring 90% and 57% correctly versus 86% and 36%. however, they performed worse for the questions of category (iii). The reason for Score-Group performing better in the last category appears to be the second question of each group codified as "Which feature hurt the bottom row the least?" in the table. Our hypothesis for why this happened is that it is easier to answer this question correctly when looking at the Score-Group plots in 26, unlike most of the other questions. (The right answer here is "Theory".)

The confidence of the participants in Rank-Group is overall higher for all questions. It is worth noting, however, that as the

Table 6: Performance in total and for each type of question. Confidence is reported on a 5 Likert scale.

			Rank-only		Score-only	
Type	Visualization	Which feature(s)	% correct	Avg. Conf.	% correct	Avg. Conf.
(i)	Figure 24a/24b	Contributed the most	100.00%	4.43	100.00%	4.83
(i)	Figure 24a/24b	Contributed the least	100.00%	4.57	100.00%	4.83
(i)	Figure 24a/24b	Contributed most positively	100.00%	4.71	100.00%	4.83
(i)	Figure 24a/24b	Contributed most negatively	85.71%	4.57	100.00%	4.67
(i)	Figure 24c/24d	Contributed the most	85.71%	4.14	66.67%	4.00
(i)	Figure 24c/24d	Contributed the least	71.43%	4.29	83.33%	4.00
(i)	Figure 24c/24d	Contributed most positively	85.71%	4.57	83.33%	4.33
(i)	Figure 24c/24d	Contributed most negatively	85.71%	4.57	83.33%	4.67
(i)	Figure 24e/24f	Contributed the most	100.00%	4.29	66.67%	2.50
(i)	Figure 24e/24f	Contributed the least	100.00%	3.71	100.00%	4.17
(i)	Figure 24e/24f	Contributed most positively	85.71%	4.43	83.33%	4.33
(i)	Figure 24e/24f	Contributed most negatively	85.71%	4.29	66.67%	3.00
(i)		Total	90.48%	4.38	86.11%	4.18
(ii)	Figure 25a/25b	Helped the 1st of the pair	14.29%	4.29	0.00%	4.67
(ii)	Figure 25a/25b	Hurt the 1st of the pair	71.43%	4.14	0.00%	4.17
(ii)	Figure 25c/25d	Helped the 1st of the pair	100.00%	3.86	66.67%	4.17
(ii)	Figure 25c/25d	Hurt the 1st of the pair	100.00%	3.86	83.33%	4.17
(ii)	Figure 25e/25f	Helped the 1st of the pair	57.14%	4.00	16.67%	2.17
(ii)	Figure 25e/25f	Hurt the 1st of the pair	0.00%	4.00	50.00%	2.67
(ii)		Total	57.14%	4.02	36.11%	3.67
(iii)	Figure 26a/26b	Helped the top row the most	42.86%	3.71	33.33%	3.67
(iii)	Figure 26a/26b	Hurt the bottom row the least	14.29%	3.50	83.33%	3.67
(iii)	Figure 26c/26d	Helped the top row the most	42.86%	3.57	16.67%	3.33
(iii)	Figure 26c/26d	Hurt the bottom row the least	71.43%	3.71	83.33%	3.00
(iii)		Total	42.86%	3.63	54.17%	3.42
All		Total	72.73%	4.15	66.67%	3.90

questions get harder, the confidence for either group does not accurately reflect the accuracy of their answers. For example, participants of both groups were overall more confident when answering incorrectly for the questions in category (ii).

The discussion portion of the focus group study also yielded different results for Rank-Group and Score-Group. For Rank-Group, participants discussed the questions and the visualization choices. While the participants of Score-Group also mentioned these points, they additionally expressed distrust in both the ranking process and the dataset during the discussion. This is consistent with [1], who used a school admissions dataset and showed that (score-based) SHAP exhibited greater and unexplained variability in the trust of the system by users compared to other methods.

Finally Score-Group participants noted that a single score-based explanation provides no insight into the overall ranking process. This is expected, as score-based explanations focus solely on the score of an item, without relating it to its position in the ranking. The x-axis represents the score, and the contributions are derived from it, making it difficult to infer how ranks change. Participants emphasized that understanding the ranking process requires viewing multiple explanations. They appreciated that the study allowed them to examine several explanations at once, which helped them form a clearer understanding of how the ranking works.

In summary, our results provide preliminary evidence that rank-based explanations are a better fit for ranking tasks as compared to score-based explanations. We are working to refine the user study protocol based on participants' feedback and to scale up the sample size to observe clearer trends.

I FOCUS GROUP MATERIALS

In this section, we present the materials used for the focus group described in Section 9.

Rank-Group

Introduction

AI tools are used to make important decisions, including in lending, school admissions, and hiring. These systems are often complex, and their decisions are difficult to interpret. In our project, we are interested in explaining the decisions of *algorithmic rankers*.

We will illustrate this with the help of *CSRankings* (https://csrankings.org), which ranks **189 computer science departments** at US-based universities based on the publication record of their faculty. Publications fall within four areas: **AI**, **Systems**, **Theory**, and Interdisciplinary (which we'll abbreviate as "**Interdisc.**"). Normalized publication counts in these areas are the **features** used by *CSRankings* to rank departments relative to each other. (The scoring formula is unimportant and we omit it here.) In *CSRankings*, **1** is the highest (best) rank, **189** is the lowest (worst) rank, and **95** is the median rank.

The goal of our project, called *ShaRP* (Shapley Values for Rankings and Preferences), is to explain the contribution of each feature to the score, rank, or some other outcome for each item. As the name of the project suggests, we use *Shapley values* to generate these explanations. **You will be helping us assess the effectiveness of explanations of a department's *rank*.**

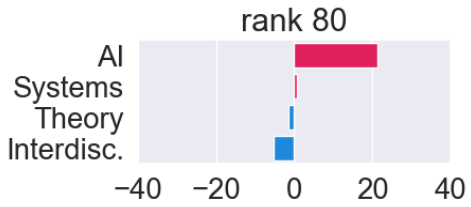


Figure 1: Explanation of the *rank* of a department that appears at position 80.

Figure 1 shows how each feature influences the rank of a department at position 80. The **blue** bars show that **Theory** and **Interdisc.** **negatively** impact the department's rank, moving it below the median rank, with **Interdisc.** having a larger effect (–5 rank positions, as shown on the x-axis). In contrast, the **red** bars show that **AI** and **Systems** contribute **positively**, moving the department above the median rank, with **AI** having the strongest positive impact (+20 rank positions). Overall, **AI contributes the most** to this department being at rank 80 because the contribution of this feature (represented by the length of the bar in Figure 1) has the highest magnitude.

Shapley values have the following property: their sum indicates how far the item's outcome is from the expected outcome. In our example, the outcome is the item's *rank*, and the expected outcome is the median rank (95 in this dataset of 189 items). Consequently, departments ranked lower in the list will have more **negative** feature contributions (shown in **blue**), while departments higher up in the list will have more **positive** contributions (shown in **red**).

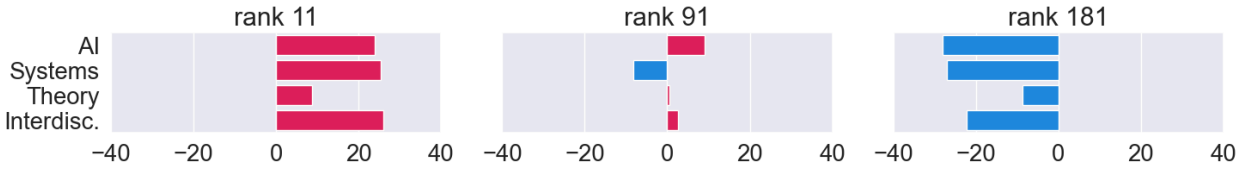


Figure 2: Explanation of the *ranks* of depts. at positions 11 (left), 91 (middle), and 181 (right).

Figure 2 shows feature importance for three among 189 departments in *CSRankings*.

- The first department (Figure 2, left) is ranked high, at position **11**; all of its features are contributing **positively** to its rank (all are **red**).
- The second department (Figure 2, middle) is ranked at position **91**, slightly above the median rank of 95; some of its features are contributing **positively** and others **negatively**. Overall, the positive contributions outweigh the negative contributions.
- The third department (Figure 2, right) is ranked low in the list, at position **181**; the contributions of all of its features are **negative**.

In this study, we will ask you to identify the features that are most informative for explaining the *rank* of an individual department, as illustrated in **Figures 1** and **2**. We will also ask you to identify features that are most informative for comparing the ranks of several departments. This is illustrated in **Figure 3**, which shows 3 top-ranked departments in the first row, 3 middle-ranked departments in the second row, and 3 low-ranked departments in the third row.

Figure 3: Explanation of the *ranks* of 9 departments: 3 top-ranked departments are in the 1st row, 3 middle-ranked are in the 2nd row, 3 low-ranked are in the 3rd row.

Score Group

Introduction

AI tools are used to make important decisions, including in lending, school admissions, and hiring. These systems are often complex, and their decisions are difficult to interpret. In our project, we are interested in explaining the decisions of *algorithmic rankers*.

We will illustrate this with the help of *CSRankings* (https://csrankings.org), which ranks **189 computer science departments** at US-based universities based on the publication record of their faculty. Publications fall within four areas: **AI**, **Systems**, **Theory**, and Interdisciplinary (which we'll abbreviate as "**Interdisc.**"). Normalized publication counts in these areas are the *features* used by *CSRankings* to rank departments relative to each other. (The scoring formula is unimportant and we omit it here.) In *CSRankings*, **18.36** is the highest (best) score, **1.03** is the lowest (worst) score, and **2.72** is the mean score.

The goal of our project, called *ShaRP* (Shapley Values for Rankings and Preferences), is to explain the contribution of each feature to the score, rank, or some other outcome for each item. As the name of the project suggests, we use *Shapley values* to generate these explanations. **You will be helping us assess the effectiveness of explanations of a department's *score*.**

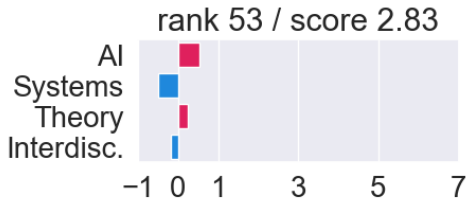


Figure 1: Explanation of the *score* of a department that appears at position 53.

Figure 1 shows how each feature influences the score of a department at position 53, with score 2.83. The **blue** bars show that **Systems** and **Interdisc.** **negatively** impact the department's score, moving it below the mean score, with **Systems** having a larger effect (–0.5 score points, as shown on the x-axis). In contrast, the **red** bars show that **AI** and **Theory** contributed **positively**, moving the score of the department above the mean score, with **AI** having the strongest positive impact (+0.6 score points). Overall, **AI contributes the most** to this department having a score of 2.83 because the contribution of this feature (represented by the length of the bar in Figure 1) has the highest magnitude.

Shapley values have the following property: their sum indicates how far the item's outcome is from the expected outcome. In our example, the outcome is the item's *score*, and the expected outcome is the mean score (2.72 in this dataset). Consequently, departments with scores below the mean score will have more **negative** feature contributions (shown in **blue**), while departments with scores above the mean will have more **positive** contributions (shown in **red**).

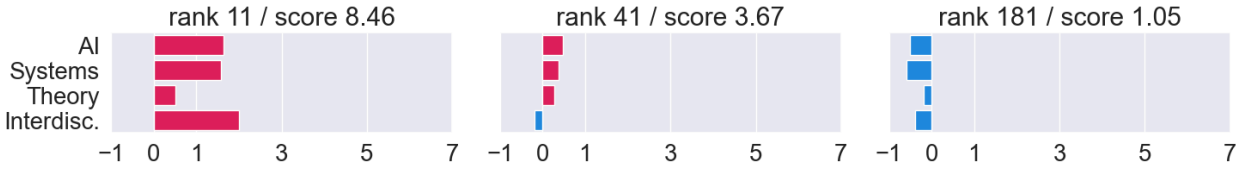


Figure 2: Explanation of the *scores* of depts. at positions 11 (left), 41 (middle), and 181 (right).

Figure 2 shows feature importance for three among about 189 CS departments.

- The first department (Figure 2, left) is ranked high, at position 11. Its score is **8.46**; all of its features are contributing **positively** to its score (all are **red**).
- The second department (Figure 2, middle) is ranked at position 41. Its score is **3.67**, which is above the mean score of 2.72; some of its features are contributing **positively** to the score and others are contributing **negatively**. Overall, the positive contributions outweigh the negative contributions.
- The third department (Figure 2, right) is ranked low in the list, at position 181. Its score is **1.05**; the contributions of all of its features are **negative**.

In this study, we will ask you to identify the features that are most informative for explaining the *score* of an individual department, as illustrated in **Figures 1** and **2**. We will also ask you to identify features that are most informative for comparing the scores of several departments. This is illustrated in **Figure 3**, which shows 3 top-ranked departments in the first row, 3 middle-ranked departments in the second row, and 3 low-ranked departments in the third row.

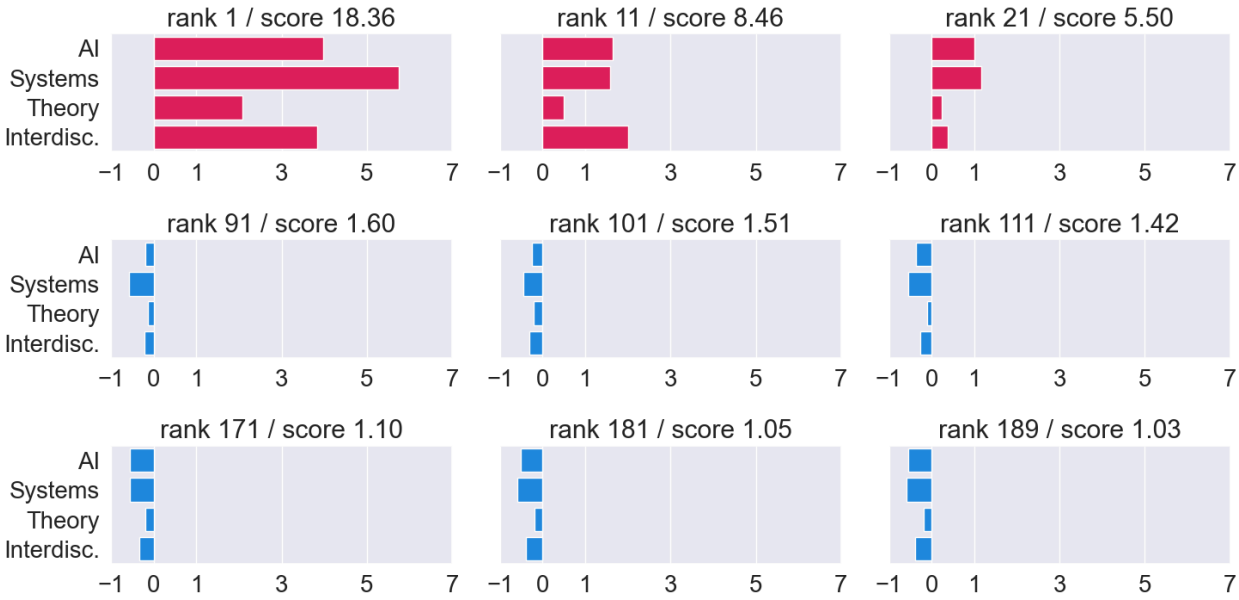


Figure 3: Explanation of the *scores* of 9 departments: 3 top-ranked departments are in the 1st row, 3 middle-ranked are in the 2nd row, 3 low-ranked are in the 3rd row.