IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 24, 2022, accepted February 20, 2022, date of publication March 14, 2022, date of current version March 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3159245

Local Explanations of Global Rankings: Insights for Competitive Rankings

HADIS ANAHIDEH [1] **AND NASRIN MOHABBATI-KALEJAHI** [2]
[1]Department of Mechanical and Industrial Engineering, University of Illinois Chicago, Chicago, IL 60607, USA
[2]Department of Information and Decision Sciences, California State University San Bernardino, San Bernardino, CA 92407, USA

Corresponding author: Nasrin Mohabbati-Kalejahi (nasrin.mohabbati@csusb.edu)

ABSTRACT Explaining complex algorithms and models has recently received growing attention in various domains to support informed decisions. Ranking functions are widely used for almost every form of human activity to enable effective decision-making processes. Hence, explaining ranking indicators and their importance are essential properties to enhance performance. Local explanation techniques have recently become a prominent way to interpret individual predictions of machine learning models. However, there has been limited investigation into explaining competitive rankings. This work proposes a hierarchical ranking explanation framework to capture local explanations for competitive rankings by defining a proper neighborhood construction approach. We explore various explanation techniques to identify the local contribution of ranking indicators based on the position of an instance in the ranking as well as the size of the neighborhood around the instance of interest. We evaluate the generated explanations for the *Times Higher Education* university ranking dataset as a benchmark of competitive ranking. The results reveal insights for a wide range of instances in the ranking list and indicate the importance of local explanations for competitive rankings.

INDEX TERMS Competitive ranking, feature importance, local explanation.

I. INTRODUCTION

Rankings provide clear-cut guidelines to illustrate comparative data for different stakeholders. University ranking, for instance, benefits prospective students to select the best option for their academic studies, encourages university-university/industry collaborations, and supports strategic plannings for resource improvement, funding opportunity enhancement, and educational capacity building. Rankings are usually based on a score function that weights specific attributes to reflect the characteristics of an entity in a list. Generally, a ranking among a set of entities is obtained based on the performance evaluation. The performance is assessed by various credentials (i.e., attributes) that mainly reflect institution success in different ways towards delivering goals and alignments with the mission. The weights in the score function reflect the influence of attributes on the ranking score. These weights are assigned by either human experts or learned by algorithms. Although in the former, the influence of these attributes on the ranking is defined a priori, they tend to be different for each entity in the ranking list. Hence, inspecting *entity-based* explanation of ranking attributes is critical for performance improvement decision-making. For example, the scoring mechanism used by U.S. News for university ranking is a linear function designed by consultation with human experts in different fields [1]. However, in the latter case, a linear regression model can learn a weight vector that transforms the multi-dimensional attribute set into a score that translates to a ranking score.

Ranking can be used to evaluate an entity in a non-competitive (e.g., loan applications) or in a competitive environment (e.g., university ranking). In the former, there is no inherent competition among entities, while in the latter, entities compete to receive a rank (or a score) in every evaluation cycle. In a non-competitive environment, the ranking shows the qualifications of each entity in the list. For example, TripAdvisor ranks hotels in a region based on the quality of their service (for example, best hotels in Los Angeles, California [2]). Hotels are not in direct competition to improve their score; instead, they employ ranking information to serve their customers. In a competitive case, on the other hand, the ranking is used to provide

The associate editor coordinating the review of this manuscript and approving it for publication was Zhipeng Cai.

administrative insights to improve the performance indicators, enhance the competitive advantage, and gain a better score in the subsequent evaluation term. As an example, *Times Higher Education* ranks universities each year considering various attributes such as graduation and retention rates, social mobility, undergraduate academic reputation, faculty resources, and financial resources [3]. University administrators take such ranking information into account for the institution's strategic planning to compete for a better place next year.

In the competitive ranking setting, although the calculated performance weighted scores and the ranking among the entities are informative in general, they do not explain the latent factor of underlying competition involved in changing the ranking of the entities. The latent competition effects are particularly important since they change the attributes' importance for different entities and impact the institutional decisions for the next evaluation period. Furthermore, competition for ranks tends to be localized, where gain or loss of ranks occurs within a few ranking positions at a time [4]. In other words, an entity's main competition is with its immediate opponents in the ranking list, where a change of position in the ranking for an entity may happen based on the effectiveness of its efforts for the evaluation cycle within a small interval of the ranking list. Consequently, the global feature importance administrated by the ranking providers, or obtained from a global model constructed on the entire list of entities, is not faithful to act based on. In this paper, we aim to investigate the local importance of attributes for competitive rankings.

Explainability of algorithmic decision-making, especially machine learning models, has recently been a field of focus and high demand [5]. The goal of an explanation is to justify the outcome of black-box or complex algorithms. Local explanation models have emerged as a popular means to understand individual predictions of classification models. A variety of model-agnostic local explanation approaches have been developed for machine learning (ML) tasks [6]–[11]. Most of them aim to provide an instance-wise explanation of the model's output as either a subset of input features or a weighted distribution of feature importance. Despite extensive local explanation research in ML, particularly for classification settings, few studies have investigated this subject for ranking tasks.

Existing work in ranking explanations is mainly in a non-competitive environment (e.g., information retrieval [12], document ranking [13], recommender systems [14], [15], etc.). Consequently, local explanation research for rankings relies heavily on random sampling for a locality construction [4], [16]. However, this approach is not valid for competitive rankings since the competition occurs in the proximity of each entity. The random sampling approaches generate out-of-proximity synthetic samples that might not even exist or be feasible to happen [10]; or they consider different entities from different positions on the ranking list for a specific neighborhood construction [17], [18].

In this paper, we propose a hierarchical ranking explainability framework to identify local explanations for competitive rankings by constructing proper neighborhoods with different sizes around each entity to identify entity-based explanations of rankings. The main intuition behind this approach is that if the importance of one attribute changes significantly for an entity from its immediate neighborhood to distant neighborhoods, it should be considered as a crucial attribute in the entity's rank. Our proposed approach includes two major steps of neighborhood construction and feature importance calculation. We construct neighborhoods in a deterministic manner considering sets of entities within the proximity of the entity of interest with different sizes. We then reflect the significance of the attributes for performance improvement within each neighborhood by exploring Explainable Artificial Intelligence (EAI) techniques [7], [19], including model-based methods as well as model-agnostic approaches. We investigate the performance of each of EAI techniques in identifying local impacts of ranking indicators and demonstrate that Shapley Value [20] is a reliable explanation technique for this matter. As a major competitive ranking setting, we consider university rankings as a benchmark in this paper to elaborate on how a global ranking fails to provide actionable insights for decisionmakers.

The rest of the paper is organized as follows: First, we discuss existing work in ranking explanations in competitive and non-competitive settings as well as the explainable university ranking literature in § II. Next, we present an overview on the EAI techniques and discuss their pros and cons in § III. In § IV, we elaborate on our proposed neighborhood construction process and present a general framework for local explanation of entities in competitive rankings. § V provides a comprehensive case study description of university ranking which is the primary problem of interest in this paper. We demonstrate the necessity of the local explanation for university ranking through a preliminary analysis using correlation technique as well as Principal Component Analysis (PCA). Next, we demonstrate the results comparing various methods based on our proposed hierarchical ranking explainability framework in § VI. Concluding remarks and future works are outlined in § VII.

II. RELATED WORK

In this section, we review the literature for explainable ranking in non-competitive and competitive settings. Considering the university ranking as a benchmark of the competitive ranking, we also present studies that mainly focus on explaining the rankings generated for universities by different rank providers.

A. NON-COMPETITIVE RANKING

The exponential growth of the web has resulted in a massive number of web pages and information overload. Thus, webpage and document ranking methods are being used to sort and recommend corresponding information for any web inquiries effectively [21]–[23]. Few of recent works employed advanced ML techniques such as Generalized Additive Models to provide global explanations for rankings [24]. Such mechanisms are considered non-competitive ranking since web pages or documents are not necessarily in direct competition to boost their rankings for cyclic evaluations. While most of the literature in the field of information retrieval, as well as the webpage and document ranking, is focused on generating effective global rankings, a few studies presented frameworks for local interpretation of such rankings [25]–[28].

Recommendation systems are widely being used to help solve personal information overload problems. Such systems leverage user performance metrics and ratings to rank the objects of interest and offer the most relevant ones as a recommendation [29], [30]. Movies/TV series ranking and recommendations are examples of such systems where user data is being utilized to identify similarities of interest, find the most relevant content, and recommend it to the target user [31], [32]. Movies/TV series ranking is categorized as non-competitive. The result of such ranking varies for each user, and the producers neither have access to the results nor can use it to gain a better rank later.

Ranking methods also can be used in financial settings such as the stock market and banking. Studies about the stock ranking problem aim to assign a rating to each stock within a portfolio based on several features involved to construct a portfolio with the highest profitability [33], [34]. Risk evaluation in the banking systems is important for strategic planning of the uncertain future. Gathering information on systemic risk factors and generating a reliable systemic risk ranking for supervision purposes can be helpful in decision-making [35]. However, the literature of local explanations of such rankings is scarce.

Rankings are also commonly used in other applications such as student admissions, organ donation-receiving, and job applications, where the ranking helps the administrators make the best decision considering their limited resources. In these applications, the competitors do not have access to the information for improving their rank in the list. As an example, Gale *et al.* [4] studied explainable and transparent ranking for high school admissions. They simplified complex ranking processes to explain the expected behaviors of ranking processes based on the design of the ranking function. To assist decisionmakers in understanding the impact of their actions on the ranking, it is critical to develop interpretable, explainable, transparent, and fair rankings.

B. COMPETITIVE RANKING

In a competitive setting, entities attempt to use past ranking information to improve their performance based on the most critical criteria for cyclic evaluations. Various sports produce ranking lists for their teams and players. Such information can help the administrators or the players try to achieve a better ranking in the future. Explaining the important factors affecting the ranking is crucial in this setting. For example, Macmillan and Smith [36] investigated the FIFA's international soccer rankings for sample selection bias using Ordinary Least Squares (OLS) models in a global ranking setting. They introduced new attributes to capture the effect of history and population in the regression model using a larger size for their sample.

Some service industries take ranking seriously and offer high-quality service to their customers. Rank providers encourage competition in this case by providing an up-to-date ranking for customers to choose the best services in the market. Montanari *et al.* [37] studied the problem of ranking nursing homes based on their capability to serve their residents. They employed a latent Markov model to define a performance index for each nursing home and proposed two ranking procedures: a) solely based on the performance indicator, b) considering the uncertainty due to its estimation in a multiple comparison perspective. Results based on a case study show the robustness of rankings obtained with respect to different model specification.

Rankings related to education have been used for a long time, where students use the information to select the best schools, and the administrators use it to attract the best students. The literature on local explanations of competitive ranking is scarce. However, a few studies in the literature of education ranking merely focus on the global explanation of the rankings. Yang *et al.* [38] developed a web-based application called ''nutritional label'' to present the ranking facts for benchmarks, including computer science departments' ranking, criminal risk assessment, and financial services. These facts are composed of a set of visual widgets that show the result of a linear model for a global ranking based on fairness, stability, and transparency.

Jajo and Harrison [39] studied the development of an index to measure universities' performance over several ranking systems. The partial least squares path modeling (PLS-PM) technique was employed to develop such an index by introducing a latent variable to measure a university's performance in a variety of ranking systems. Multiple scenarios were considered to explore the impact of variable changes for a specific university in any ranking system using the proposed performance index. McAleer *et al.* [40] evaluated the effects of the number of full-time-equivalent students (i.e., size) and the percentage of international students (i.e., internationalization) on academic rankings for private and public universities by developing linear regression models. The Times Higher Education World University Rankings dataset was used to illustrate the positive relationship between the size and internationalization for Japanese universities in 2017 and 2018.

The state-of-the-art approaches on university ranking merely focus on learning the ranking functions and evaluating different ranking systems using machine learning techniques. Frenken *et al.* [41] used a regression analysis to assess universities' research performance and the influence of structural variables (e.g., location) on the performance differences among universities. Tabassum *et al.* [42] specifically studied the correlation of university ranking indicators focusing on identifying the influential features using an outlier detection approach. Mikryukov *et al.* [43] utilized Principal Component Analysis (PCA) to identify the significant factors, latent variables, and the correlations between the latent and the basic variables. Also, score-based performance is critically evaluated for university rankings using PCA in [44]. Gale *et al.* [45] modeled the lagged rank as an independent variable to account for the stickiness of ranking using the logit technique.

C. SUMMARY OF CONTRIBUTIONS

Most of the existing related work is in non-competitive ranking settings in which the ranked entities do not change their attributes in the future to enhance their ranking. For example, in [4], the top-k entities are returned using a ranking function (i.e., a weighted function with attributes). The authors proposed a metric to assess the contribution of attributes on the final ranking outcome, which directly depends on the value of k. Particularly, the metric works to return the top-k queries based on the importance of attributes involved in ranking. Hence, the proposed framework cannot be applied to every entity listed in a fixed ranking, such as university ranking. However, in a university ranking (or any competitive ranking) setting, we need to explain the importance of attributes for different entities given their position on the ranking list. We do not necessarily want to focus on top-k entities. Even if we do, for $k = 200$ (full list of universities), the returned importance by their proposed metric is global, not local, as the approach considers all universities. For example, there is no way to find out what features are more important for a university ranked 53 using their approach. As another example in non-competitive settings, the EXTRA algorithm is designed in [14] for recommender systems and is based on user-item interactions. However, in our setting (i.e., competitive ranking of entities) there is no item to be matched with entities. Hence, the algorithm is not applicable in such settings.

In the university ranking settings (also competitive ranking) there are two most relevant papers in the literature by Gale *et al.* [45] and Johnes [44]. In [45], although they considered the local explanation of the ranking, the focus of the paper is to include the lagged rank as an independent variable to account for the stickiness of the ranking. They do not provide insights into the local importance of attributes for different entities ranked in the list. In [44], they use state-of-the-art techniques to reveal the problems of ranking scores provided by different agencies such as the *US NEWS* and *Times Higher Education* (as we also show in § V). This existing work mainly identifies groups of important attributes which are composite indices of original attributes. As a result, they provide the global impact of the attributes on the ranking and disregard the local behavior of entities.

Our proposed framework aims to show the local importance of attributes for each entity based on their location in the ranking list and their corresponding neighborhoods to provide a better entity-based insight into the competitive ranking. To the best of the authors' knowledge, there is no existing work with the focus of explaining the local impact of attributes given the position in the ranking for competitive settings. This as well as the scarcity of the relevant literature make it impossible to compare the output of our research with any baseline.

III. FEATURE IMPORTANCE METHODS

Feature selection and importance calculation have a long history in machine learning literature. Feature selection is a pre-processing step that includes eliminating non-informative and redundant information from the data. Feature selection enhances learning algorithms, increases predictive accuracy, and reduces the complexity of the results. Many studies have proposed effective and efficient methods for feature selection, as a pre-processing technique for dimensionality reduction, preserving informative attributes in high-dimensional data [46]–[48]. Feature selection techniques have also been used for explanation to justify modeling outcomes [49].

To explain ranking functions with latent variables and estimate their direct relationship with the attributes, a handful of studies exist in the literature as described in § II.

A. INTERPRETABLE MODELS

1) LINEAR REGRESSION (LR)

LR models are used to identify the dependence of a regression target (Y) on various independent features (X_i). The weighted sum of the features calculates the predicted outcome as $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$, where β_i represents the learned feature weight, β_0 is the intercept, and ϵ refers to the error in making the prediction. An advantage of linear regression models is their linearity which makes the model easy to interpret. Since the effect of features is additive, it is possible to separate them and then describe each. Assume using a linear regression model as an interpretable model. Then, the features that receive the highest weights are the explanations of the model outcome (i.e., the ones that contribute most to the outcome). It is possible to measure the importance of a feature in a linear regression model by calculating the absolute value of its t-statistic. The t-statistic is the estimated weight scaled with its standard error [50] as in equation (1):

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}. \tag{1}$$

Equation (1) shows a direct relationship between the feature weight and its importance. It indicates that the more variance the estimated weight has, the less important the associated feature is. Linear regression used as an interpretable model cannot capture the nonlinearity and often fails to identify the interactions between features by oversimplifying the relationships. As an example of using linear regression for ranking, Refenes *et al.* [33] studied the stock ranking problem, where the goal is to assign a rating to each stock within a portfolio based on several features involved. Neural network performance was compared with multiple linear regression, and it has been shown that it yielded a better in-sample and out-of-sample fitness than the linear regression model.

2) PARTIAL LEAST SQUARE (PLS)

PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ (i.e., latent variables) that are linear combinations of the original features shown in equation (2):

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j. \tag{2}$$

PLS then uses these newly created features to fit a linear model using least squares as the loss function. Unlike other dimension reduction techniques, PLS identifies these new latent variables considering the relationship with the response, Y, as well as the original features, X_is, and therefore, does a better job explaining the response. The latent variables of the PLS are the directions in a projected lower-dimensional space, where the variation of the data in the original space and with regard to the response are explained.

PLS sequentially identifies the latent variables Z_0 by fitting a linear regression on Y and X. The weights of the original features in the linear combination of equation (2) are the coefficients of the fitted linear regression. The variable with the largest weight is the most important in Z_0. From the second iteration of the PLS algorithm, we need to first decorrelate the latent variables, Z_j, and the original variables, X_j, by regressing the latter on the former and computing the residuals, which indicates the unexplained portion of the variation. The orthogonalized data will then be used to construct the next latent variable Z_1. This process continues until the predetermined number of latent variables, m, are constructed. The original variables need to be standardized before applying PLS to avoid bias in projection to lower-dimensional space. Variable Importance in Projection (VIP) scores [51], [52] are defined for each X variable, j, as the sum, over the latent variables (LV), of its PLS-weight value (ϕ_j) weighted by the percentage of explained Y variance (SSY) by each specific LV, according to equation (3):

$$w_j^2 = \sum_{m=1}^{M} \phi_{jm}^2 \frac{SSY_m \times p}{SSTot \times M}, \tag{3}$$

where M is the number of latent variables of the PLS model and p the number of X variables. When the number of observations is much smaller than the number of variables in the data set, PLS regression is the most suitable technique to analyze them. The main drawback of PLS is that it assumes a linear relationship between features and the response, which is not necessarily the case. We shall elaborate on the university ranking use case in a subsequent section and demonstrate that the relationship between features and ranking is not linear in most cases. In addition, PLS derives the latent variables, which are unexplainable, and uses the VIP equation to calculate feature importance. The SSY and $SSTot$ are calculated based on the linear regression fitted, which has the same disadvantage regarding the linearity assumption. Consequently, when this assumption is violated in the data, the PLS result is unstable [53] (i.e., this may yield a situation in which all of the features are equally important, and does not help distinguish between them).

3) DECISION TREE

The Classification and Regression Tree (CART) is a non-parametric decision tree and nonlinear machine learning model [54]. The CART algorithm utilizes a recursive binary splitting approach to partition the input variable space into smaller sub-regions. It is a classification method when the response variable is categorical and a regression method when the response is numerical. CART chooses binary splits for regression predictive modeling problems by minimizing the sum of the squared error (SSE) of the output in the resulting sub-regions. A greedy recursive binary splitting approach is performed to identify an optimal splitting variable and an optimal cut-off point that leads to the greatest possible reduction in SSE. The splitting approach continues until a stop criterion is satisfied (e.g., maximum depth of the tree). The final partitions when the CART algorithm stops are known as leaves or terminal nodes. The predicted output is the average output of the input points within each terminal node.

Feature importance in decision trees is calculated as the total decrease in the variance or Gini index of a node that uses the feature compared to the parent node. The higher the value, the more important the feature. In particular, the importance of each feature is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The method of using Gini index, assuming only two child nodes (binary tree), is as follow:

$$imp_j = \frac{n_j}{n} gini_j - \frac{n_{j(left)}}{n} gini_{j(left)} - \frac{n_{j(right)}}{n} gini_{j(right)}, \tag{4}$$

where n_j is the importance of node j, n is the total number of observations, $gini_j$ the impurity value of node j, $j(right)$ and $j(left)$ indicate the left and right child nodes on j, and imp_j is the importance of node j. The Gini impurity is calculated as $gini = p_1(1 - p1) + p2(1 - p_2)$ where $p1$ and $p2$ are class 1, 2 probabilities. Hence, the importance for each feature on a decision tree is then calculated as:

$$W_i = \frac{\sum_j imp_j}{\sum_{k \in allnodes} imp_k}, \tag{5}$$

where W_i is the importance of variable i and imp_j is the importance of node j splitting on node i.

The sum of all features' importance is scaled to 100 to achieve a normalized total reduction of criteria. This criterion indicates that the importance of each feature as a percentage of the overall model value. Using the algorithm as an interpreting model can fail since small changes in the dataset can make the tree structure unstable and result in high variance. It does not perform well as a regression method in cases with the continuous predicting values [50]. As an example of PLS application in ranking, Zhu *et al.* [34] studied the stock ranking problem in the North American stock market.

They proposed a hybrid approach for ranking stocks that combines the benefits of the decision-tree approach of CART with the linear logistic regression model, which also offers enhanced performance compared to either a stand-alone CART or logistic regression. The hybrid approach overcomes the limited sensitivity of CART to continuous variables, where it avoids the coarse-grained response produced by CART.

4) PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a widely used unsupervised method for dimension reduction and feature selection [55]. PCA method projects high dimensional data into low dimensional vectors, which have uncorrelated components. PCA has certain advantages in feature extraction, but unlike PLS, it does not consider any information regarding the response variable. PCA seeks a direction in feature space along which the data vary the most and then projects the data in this direction. The projected values are the principal components. Assuming that each of the variables has been mean normalized, the first component is the linear combination of original features as in equation (6):

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p, \tag{6}$$

where ϕ_{i1} are the parameters of the PCA, referred to as loadings of the first component. The loadings are constrained so that their sum of squares is equal to one (i.e., $\sum_{j=1}^{p} \phi_{j1}^2 = 1$). The loading vectors are the directions with the largest variance, which are, in fact, the ordered eigenvectors. Hence, the parameters of the PCA components can be obtained by an eigendecomposition. Principal components are constructed sequentially, being uncorrelated to the previous component. In particular, the second principal component Z_2 is the linear combination of features with the maximal variance of all linear combinations uncorrelated to the first component.

With PCA, a biplot can be set up to represent the variables and observations as a graph, thereby allowing the correlations and associations to be visualized to facilitate interpretation of the associations and variation, and consequently, aid in the selection of attributes. The PCA biplot gives an overview of the similarities and differences between attributes and the interrelationships between them. Consequently, biplot analysis is a provisioning tool to identify less important attributes in the ranking. In this paper, we utilize biplot and will not use PCA for feature importance calculation. Due to the low interpretability of principal components, PCA is not considered as an explanation technique. As an example of using PCA in ranking, Fang *et al.* [35] combined five popular systemic risk measurement rankings (i.e., SRISK, Leverage, CoVaR, VaR and CAPM-$\beta\times$ MV) for 16 Chinese banks by applying the PCA model. Their proposed framework gathers the main information on systemic risk and generates a reliable systemic risk ranking for supervision purposes.

B. MODEL-AGNOSTIC METHODS

A model-agnostic approach is an alternative approach for the explanation and interpretability of machine learning models. These methodologies treat the original model as a black-box and extract post-hoc explanations, which involves learning an interpretable model on the predictions of the black-box model [56], [57], and evaluating the sensitivity of model through perturbing input sample points [10], [58], [59].

1) LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

LIME aims to explain individual predictions of black-box machine learning models [10]. LIME approximates complex model functions by locally fitting linear models to random permutations of the original training set to provide point-wise explanations. It helps to identify and explain which features are contributing the most to the prediction. LIME starts with perturbing each feature in the training dataset individually and creating a new subset of sample points. The newly generated samples are taken from a distribution centered at the training data and not necessarily around the instance of interest, which can be problematic [50]. LIME calculates a similarity score as the distance between new sample and the original data. Considering p features, it trains an interpretable model on the new subset of samples, which is weighted by the similarity score. The output of this model (i.e., features weight) explains the machine learning model's local behavior.

Mathematically, LIME aims to minimize a loss function $\mathcal{L}(f, g, \pi_x)$, where f, g, and π_x refer to the original model, an explanation model, and similarity kernel measuring the proximity of a new sample z (i.e. perturbed sample point) to x (i.e, original data point), respectively. Defining $\Omega(g)$ as a measure of complexity (as opposed to interpretability) of the explanation g, the explanation by LIME is obtained by equation (7):

$$\xi(x) = \underset{g}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_n) + \Omega(n). \tag{7}$$

Equation (7) demonstrates that samples are generated around x and weighted by π_x to approximate $\mathcal{L}(f, g, \pi_x)$. Such approximation aims to learn the local behavior of the original model f and measures how close the explanation $\xi(x)$ is to the prediction of the original model. One way to generate new sample points is to perturb the instance of interest by sampling from a $Normal(0, 1)$ distribution and applying the inverse operation of mean-centering and scaling, based on the means and standard deviation in the data. This results in the total loss of the covariance structure where the samples lie outside the space of a real dataset [60]. Another way for sampling is to use an exponential smoothing kernel where it takes two data instances and returns a similarity measure to define the neighborhood. The size of such neighborhood is defined by the size of the kernel width, which crucially affects the performance. Smaller the size, lower the effect of farther instances to the model output. There is not a unified way to find the best kernel width [50].

There are a few studies that use LIME for text-based ranking models explaining the relevance between a single query document pair [27], [61], and binary sentimental analysis as a classification task [62]. The literature of using LIME for competitive ranking explanation is scarce.

2) SHAPLEY ADDITIVE EXPLANATIONS

Shapley Value (SV) originally were introduced in the 1950s to measure contributions of individual players to a cooperative game [63]. A coalitional game consists of a set of N players and a characteristic function v which maps subsets $S \subseteq \{1, 2, \ldots, N\}$ to a real value $v(S)$, satisfying $v(\emptyset) = 0$. Suppose we could replay the game with all possible combinations of (a subset of) players and observe the resulting team score $v(S)$ (aka payout). We could then divide the overall payout among the players based on their average contributed value across all possible subteams to which they were added. This individual payout is the player's Shapley Value. SV payout scheme is proven to be:

- **Efficient**: the Shapley Values of all players should sum up to the total payout,
- **Symmetric**: the payout of two players are the same if they add the same value in all team combinations,
- **Dummy-sensitive**: a player that never improves a subteam's performance when it is added to the subteam should have a Shapley Value of zero,
- **Additive**: in case of a combined payout (say we add two game bonuses), the combined Shapley Values of a player across these two games is the sum of the individual game's Shapley Values.

In machine learning applications, considering the success of a team as an output, each player's contribution can be considered as the feature importance. SV has been recently used as an interpretable model-agnostic explanations to calculate feature importance [58]. SV is an intuitive tool to understand both global feature importance across all instances in a dataset and instance-wise feature importance in black-box machine learning models.

The general idea of SV is based on the inclusion and exclusion of variables to calculate their individual impacts on the prediction. This is calculated by fitting a linear model on the full dataset but considering only a subset of features for prediction of an instance. The subset of features are all possible coalition subsets of features, which makes the SV impractical predictive modeling applications with a large number of features. The estimation methods are proposed to reduce this computational complexity. Instead of calculating the Shapley Values using all possible coalitions, only a subset of these coalitions is selected in a random manner for the calculation.

Let S be a subset of features that does not include the feature for which we calculate the importance. Let M be the full set of features. Given a model $g(x)$ trained to predict $f(x)$, the marginal contribution of feature i to the model's prediction and accordingly to the score $f(X)$ is:

$$\phi_i = \sum_{S \subseteq M \setminus i} \frac{|S|!(|M| - |S| - 1)!}{|M|!}[g(S \cup i) - g(S)], \quad (8)$$

where $S \cup i$ is the subset that includes features in S plus feature i and $S \subseteq M \setminus i$ indicates all sets S that are subsets of the full set of features M, excluding feature i.

In the ranking context, SV identifies the distribution of a model's prediction resulting from an input feature vector over the individual features. To the best of our knowledge no work in the literature exists that uses SV for ranking explanations.

IV. LOCAL EXPLANATION FOR GLOBAL RANKINGS

While the explanation space itself and methods to generate explanations are widely known in practice for classification tasks, little has been explored on leveraging explainability techniques for ranking tasks, and more specifically, for *competitive ranking*. In this section, we first describe our hierarchical neighborhood construction idea and then describe the utilization of EAI techniques (both model-based and model-agnostic approaches) within each neighborhood to identify the local impact of features.

Let $\mathcal{N}$ be a list of $|\mathcal{N}| = n$ ranked entities with a set of attributes $x = < x_1, \ldots, x_p >$ indicating p ranking parameters and total score of f. The score is the ranking function that is normally defined as $f(x) = \sum_{i=1}^{p} w_i * x_i$, where $w_1, \ldots, w_p$ are the global weights of each attributes given such that $\sum_{i=1}^{p} w_i = 1$. Let (x^s, f^s) indicates the attribute and the score of an entity ranked s. Hence, $s \succ s'$ implies $f^s \geq f^{s'}$. We assume f is a monotonic ranking function, i.e., having two entities with equal attributes except one, ranking function, f, of the two entities follows the dominance direction of the single unequal attribute.

In the competitive ranking, if two entities compare equally, they obtain the same ranking. This will accordingly leave a gap in the numbers that represent the ranking of the universities. If r universities receive the same ranking, then there will be $r-1$ units of the gap in the ranking numbers succeeding the last university that compared equally. This ranking strategy is commonly used in competitions since it ensures that if two or more participants tie for a ranking position, the positions of all those listed worse remain unchanged [64], [65].

The global weights, w_i, are defined considering all the entities in the ranking list, $\mathcal{N}$, and are also independent of the rank of the entity. However, the competition between entities and the performance improvement strategies occurs within certain proximity for each entity. To investigate the local influence of attributes on ranking, we propose a hierarchical neighborhood-based feature importance calculation, which is more aligned with the nature of ranking where entities compete with their adjacent peers. Our proposed hierarchical feature selection approach partitions entities into subsets (i.e., **neighborhood**s), based on their adjacency in the ranking list. Consequently, instead of similarity-based partitioning, we consider a subset of entities located in the same ranking range to capture the local impact of features.

For each entity ranked s let $\mathcal{H}^s = \{\mathcal{H}_1^s, \ldots, \mathcal{H}_{k_s}^s\}$ be the set of k_s neighborhoods with the one side size of $z_j, \forall j = 1, \ldots, k_s$. More specifically, each neighborhood is constructed considering z_{k_s} entities below s and z_{k_s} entities above it; $\mathcal{H}_j^s = \{(x^{s-z_j}, f^{s-z_j}), \ldots, (x^{s+z_j}, f^{s+z_j})\}$. We refer to the first neighborhood $\mathcal{H}_1^s$ as the *immediate* neighborhood, where entity s is surrounded by $2 * z_{k_s}$ number of neighbors. The second neighborhood, $\mathcal{H}_2^s$, is called *near*, and the third and largest one, $\mathcal{H}_3^s$, is named *far* neighborhood. Note that if the one side size of the neighborhood z_j is larger than rank s, the neighborhood includes only the ones that do not exceed the range of the ranking, i.e., $s - z_j > 0$ and $s + z_j < |\mathcal{N}|$. The neighborhood structure is hierarchical, meaning that each neighborhood encompasses its predecessor neighborhood members as well as the new ones.

To clarify the neighborhood concept for each entity in the ranking list, let us provide an example. Consider $s = 5$ and $z_j = 5$. The immediate neighborhood, then, will be $\mathcal{H}_1^s = [1, 2, 3, 4, 6, 7, 8, 9, 10]$, which includes 10 universities surrounding the university number 5 in the list. Other neighborhoods, $\mathcal{H}_j^s$, can be constructed similarly.

The underlying distribution of the attributes impacts how much each factor contributes to the final ranking. Even though each feature's global contribution in ranking is known, different distributions of attributes in different neighborhoods around entities of varying rank result in substantial differences in their importance. This indicates the impact of unknown latent attributes not explicitly considered in the weighted sum score function for ranking.

Consequently, learning the distribution of attributes locally better explains the ranking mechanism for decision processes. Considering the hierarchical neighborhood construction for each entity reveals the explicit contribution of each attribute on the final ranking for each entity based on the local behavior of its neighbors, rather than being misled by the whole set of entities in the ranking. The local explanation of the attributes results in having proportional expectations on the outcome corresponding to a change in various attributes, i.e., derived by a performance improvement strategy.

Let $w_{i_{(j)}}^s$ be the weight of attribute x_i in neighborhood $\mathcal{H}_j^s$. We define $\mathcal{W}_j^s$ as the set of importance weights for all attributes in neighborhoods j, $\forall j = 1, \ldots, k_s$, for entity ranked s. To determine the weights of each attribute, x_i, $\forall i = 1, \ldots, p$ in each neighborhood, k_s, we utilize a variable importance calculation technique (i.e. EAI techniques) within each neighborhood.

The local importance of features, $w_{i_{(j)}}^s$, is less biased towards a group of entities, in which one or a subset of features are more representative than the ones in other groups. Therefore, it provides better explainability regarding the performance of entity s. Algorithm 1, represents the proposed neighborhood-based explanation framework. Steps 2-5 indicate the upper side neighborhood construction, and steps 8-12 indicate the lower-side neighborhood construction. Step 11 shows the calculation of variable importance using the techniques described in § III.

The hierarchical neighborhood construction allows us to evaluate the local impact of features calculated by EAI precisely. The variation in feature importance from an immediate neighborhood to a distant neighborhood confirms the necessity of the local importance of a feature for an entity. The sensitivity of an EAI technique is measured by the total variation of the feature importance across k_s neighborhoods of entity s as in equation (9):

Algorithm 1 Neighborhood-Based Explanation

```
Input: $s$
 1: $\mathcal{H}_j^s = \emptyset$;
 2: for $j = 1$ to $k_s$ do
 3:     for $l = 1$ to $z_j$ do
 4:         if $s - l > 0$ then
 5:             $\mathcal{H}_j^s = (x^{s-l}, f^{s-l}) \cup \mathcal{H}_j^s$
 6:         end if
 7:         $\mathcal{H}_j^s = (x^s, f^s) \cup \mathcal{H}_j^s$
 8:     end for
 9:     for $l = 2$ to $z_j + 1$ do
10:         if $s + l < \mathcal{N}$ then
11:             $\mathcal{H}_j^s = (x^{s+l}, f^{s+l}) \cup \mathcal{H}_j^s$
12:         end if
13:     end for
14: end for
15: for $j = 1$ to $k_s$ do
16:     $\mathcal{W}_j^s \leftarrow EAI(\mathcal{H}_j^s, x)$
17: end for
18: return $\mathcal{W}^s$
```

$$var(w_i^s) = E[(w_i^s - \frac{1}{k_s}\sum_{j=1}^{k_s} w_{i_{(j)}}^s)^2], \quad \forall i = 1, \ldots, p. \tag{9}$$

This is a necessary step since we do not have access to the actual local important features to evaluate the performance of different EAI techniques. In classification settings, a predefined set of features for classification is known in advance and are considered important features (i.e., "golden features"). This, however, is not a possible approach for competitive rankings, where we do not have any control over the ranking list.

Considering a distinguished competition in each neighborhood for an entity, a method that reflects a larger variance for feature importance across neighborhoods is a more reliable model for local explanations.

EAI techniques are mainly designed to explain any classifier. However, in this paper, we have a different setting, i.e. ranking. In order to adopt EAI for competitive ranking, in this paper, we consider ranking as a numerical response and seek to identify the contribution of each feature on the response.

- **Model-based Methods**: Considering ranking as Y and ranking indicators as X, we fit Linear Regression, Partial Least Squares, and Decision Tree with each neighborhood $\mathcal{H}^s$ to calculate the local importance of features.
- **Model-agnostic Methods**: To apply LIME to ranking, we first need a model on which the ranking is to be explained. The reason is that LIME is designed

to explain a predictive model and cannot be applied independently. In this paper, we use an ML model on the ranking to map Y and X in each neighborhood first, then apply LIME to explain the contribution and importance of each feature. SV, however, can be adopted directly on the ranking without requiring a model since it uses a linear model in its algorithm for contribution calculation. SV considers the rankings as the predicted values of a black-box model (i.e. ranking function) and calculates the feature importance through its exclusion/inclusion process by fitting a local interpretable model. However, the inbuilt locality of SV that is performed on the entire dataset will not consider the localized competition. That is why, in this paper, we use SV within each neighborhood to limit the sample generation of SV to a proper locality for competitive rankings.

V. UNIVERSITY RANKING

Various organizations provide university rankings yearly with the sole goal of comparing each university's resources. University administrations take such rankings seriously to advertise their programs, prioritize their resource allocations, and compete for a better place in the following evaluation cycle. Making random decisions for investing in a factor such as increasing the number of international students or allocating more teaching resources will not necessarily improve the entity's rank for the next year since an adjacent university might do the same. Also, any global ranking provided by university ranking agencies fails to capture entities' underlying competition. In addition, the nature of the competition is different for each entity based on location in the ranking list. For example, strategies of ranking improvement for a university ranked five do not necessarily benefit a university ranked 50 on the list. Therefore, the competition to improve the rank among universities is different based on the location and neighborhood they are located in. It is critical to construct a model to assess the locally faithful importance of the attributes and provide insights into the nature of the underlying competition to assist university administrators in developing proper strategies to improve their rankings. This helps decisionmakers identify the best subset of actions to take at the entity level.

Most academic rankings attempt to measure the quality of university education and research based on the research productivity, teaching quality, and the number of students and faculty/staff, etc. Utilizing the *Times Higher Education* ranking dataset [66], we identify the contribution of each ranking attribute to the overall ranking. This will allow us to get an insight into the actual importance and differentiation across different ranking groups.

First, we performed a PCA analysis and produced a *Biplot*, shown in Figure 1, to compare and identify the contribution of different attributes to the ranking. Although biplots are mainly a two-dimensional exploratory graph, they provide a summarized analysis of the directions of variation in the data, mainly the importance (i.e., loadings) of attributes and components in one plot. To better visualize the impact of the attributes across different ranking groups, we considered four groups of rankings and labeled them as [1, 20], [20, 50], [50, 100], and [100, 200]. The loadings are plotted as vectors in the space.

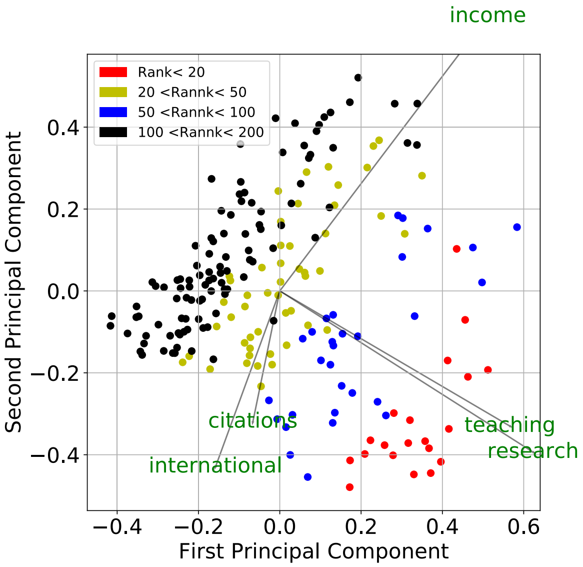


FIGURE 1. Biplot of ranking indicators using PCA analysis.

As illustrated in Figure 1, the PCA result indicates that *teaching* and *research* are the main factors for the first principal component, which explains the largest variation in the ranking since the two attributes have large positive loadings. *Income*, *international*, and *citations* are less important to the first component, as they are far from the horizontal axes and are highly loaded features for the second component. Our findings show that the top 20 universities (red dots in the plot) tend to have higher teaching and research indicators. Furthermore, we can observe that middle-rank universities (blue dots) have citations and international as the critical indicators for ranking. Lastly, for lower ranked universities, income plays the main role in the ranking variation, and all other factors have a lesser overall impact. In particular, these universities are on the negative portion of the scale (i.e., lower scores than average) related to teaching and research indicators. The loadings of attributes in each component are provided in Table 1. Relationships between features shown in PCA-Biplot graphics can be evaluated according to the angles between vectors. The smaller the angle between any two traits, the more closely related they are.

Although the preliminary results based on Biplot demonstrate that the global importance of attributes differs across different ranking groups, it is not as precise as it should be for more granular level decision-making.

Next, we compared the PLS result to discuss the local and global importance of attributes more percisely under different scenarios shown in Figure 2. Since the *Times Higher Education* ranking calculation equation is a linear weighted summation of attributes with "30%:teaching", "7.5%:internationa",

TABLE 1. PCA loadings.

	Original Attributes				
Components	Teaching	International	Research	Citations	Income
PC 1	0.68	-0.09	0.68	0.01	0.25
PC 2	-0.10	-0.46	-0.15	-0.67	0.55
PC 3	0.12	-0.85	-0.07	0.22	-0.45
PC 4	0.16	0.17	0.12	-0.70	-0.66
PC 5	0.70	0.13	-0.70	0.02	0.07

"30%:research", "30%:citations", and "2.5%:income", the coefficients of a linear regression approximation mapping ranking and attributes are precisely equal. This indicates the global impact of the attributes, which is obviously aligned with the weights provided by the *Times Higher Education* report [3].

In Figure 2, we can observe the pairs that the plot builds using two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable, while the scatter plots on the lower triangle show the correlation between two variables. The pair plot is used to figure out which attributes are best for explaining the impact of each variable on the response variable of interest, in our case of university ranking. The pair plot demonstrates between variable correlations present in the dataset. For example, the left-most plot in the fifth row shows the scatter plot of "world rank" versus "citations". The color-coding of the points indicates universities with different groups of rankings categorized as "top-20", "20-40", "40-80", "80-100", "100-150", and "150-200", for visualization purpose. We can observe that, for some variables, there exist some global trends with the ranking (e.g., "research"). However, for others, there is no global pattern (e.g., "international"). Looking more precisely at the groups of ranking, we can identify some local correlations. Focusing on a subgroup of universities in group "40-80" ranked between 50 to 60, Figure 2 (a) and (b) demonstrate the local trend exists for "income" and "international" variables versus world rank, which is in contrast with the global pattern.

Finally, we evaluated the *Times Higher Education* ranking in 2020 to show the nonlinear nature of the ranks, shown in Figure 3. Such insights can help understand the relationship between the rank and the score for each entity. It also emphasizes that in a competitive ranking, several entities might get the same rank, which can change the system's dynamic and the inability of some explanatory models to capture the feature importance correctly.

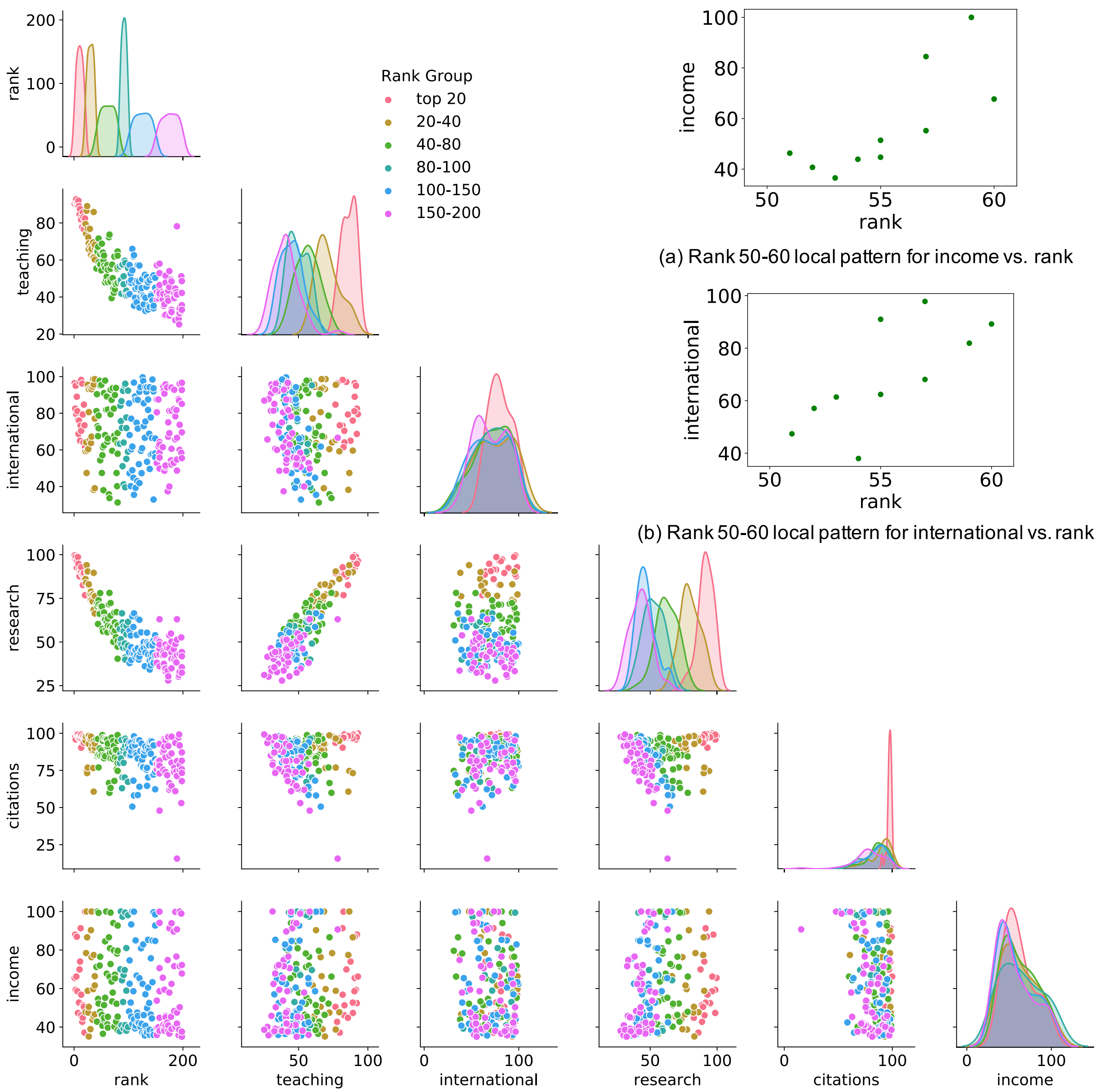
Figure 3 (a) shows the nonlinear relationship between the total score and the world rank in specific ranges such as [40-80]. In these ranges, the universities are less sticky in ranking and do not show a linear relationship between their score and ranking (see the sudden changes of the slope and the gaps). The distribution of the world ranks and the total scores along with their central tendency are depicted in Figure 3 (b) and (c), respectively. For these plots, the y-axis shows the density or the normalized counts of observations for each subset. Figure 3 (b) indicates that there are universities with the same ranks in several subsets, especially above 100 in the ranking list. This also highlights the reason for seeing gaps in the ranking, as in Figure 3 (a), for the ranges such as [140-160]. Figure 3 (c) shows that the total score distribution has a positive skew meaning that the scores for lower-ranked universities, especially for ranges [40-80], are very close. As we move further in the ranking list, the scores differ significantly. This emphasizes the nonlinear relationship between score and rank as in Figure 3 (a) for the corresponding range. It also shows that the competition is tight and different for the lower ranked universities compared to the higher ranked ones.

VI. RESULTS

A. DATASET DESCRIPTION

We analyzed the *Times Higher Education* [66] university ranking dataset for year 2020 to evaluate the local explanation frameworks using different models as in § III including Linear Regression (**LR**), Partial Least Squares (**PLS**), Decision Trees (**CART**), Shapely Value (**SV**), and Local Interpretable Model-Agnostic Explanations (**LIME**). We use X_1, X_2, X_3, X_4, and X_5 `teaching`, `international`, `research`, `citations`, and `income`, respectively, as the ranking determinants (also called attributes). We filtered out the universities with rankings above 200 since there were labeled as ranges instead of precise rankings. To evaluate the performance of the proposed framework using the above methods for all the ranges of the ranking dataset, we designed different scenarios to consider universities with a wide range of rankings. We categorized the scenarios into two groups of below 101 in the ranking, namely [5, 15, 25, 45, 65, 85, 101], and above 101, namely [110, 120, 130, 141, 149, 157, 175]. Note that in the 2020 world ranking list by the *Times Higher Education* [66], no university has been ranked 100. Therefore, we used 101st university since it is in the targeted scenario border. Also, there is a gap in ranking for some universities above 101. As mentioned before, in a competitive ranking problem, some ranking spots are not necessarily filled by any entity, while various entities have the same rank in the list. Therefore, we had to select existing rankings with respect to their scatter for the second scenario for our university benchmark problem. We defined three neighborhoods with sizes of 5, 10, and 15 as *immediate*, $\mathcal{H}_1$, *near*, $\mathcal{H}_2$, and *far*, $\mathcal{H}_3$, neighborhoods. It should be noted that since **LIME** works with random sampling, it generates different outputs in each run, where setting a random seed does not necessarily yield to the same output [67]. Therefore, we report the average of 30 runs for **LIME** in the results subsection.

For **LR**, **CART**, and **PLS** we utilized Scikit-learn packages. We adapted the basic **SV** algorithm implemented in [68]. For **LIME** we used python implementation provided in [69].



(a) Rank 50-60 local pattern for income vs. rank

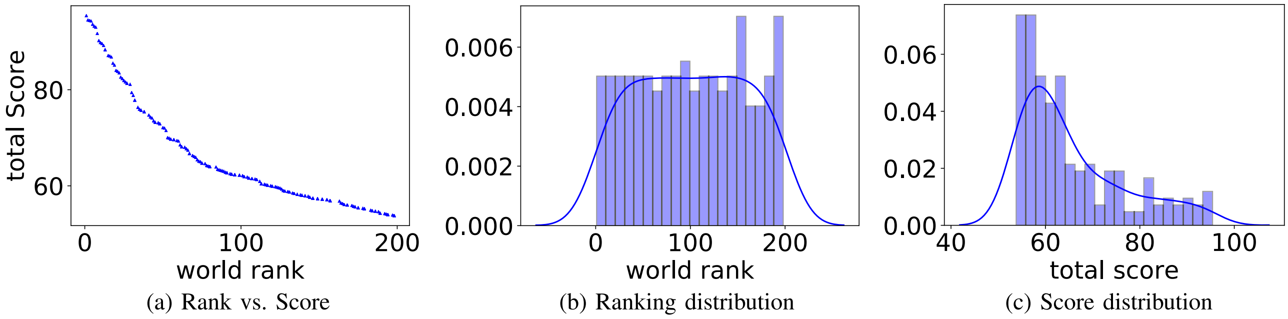
(b) Rank 50-60 local pattern for international vs. rank

FIGURE 2. Pairs plot to show local and global importance of attributes for different rank groups.

B. RESULTS INTERPRETATION

Figure 4 provides the results of our proposed framework using the interpretable techniques in identifying the local explanations of the ranking for universities below 101 in the ranking list. The heatmaps help to identify which features are most related to the ranking within each of the $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ neighborhoods. Each heatmap has two vertical axes, one representing the three neighborhoods and the other representing the feature importance values. Since each technique results in a different range for the importance values, we scaled them to be between 0 and 1 using a maxmin scaling approach. There are also two horizontal axes: one shows the specific instance in the ranking list (e.g., a university ranked 5 in the list of 2020 ranking is shown as Uni.5), and the other represents the attributes X_1, X_2, X_3, X_4, and X_5.

As observed in Figure 4 (a), **LR** is not able to distinctly explain the local impact of the features across neighborhoods in the majority of cases. Moreover, **LR** acknowledges all features as equally important within each neighborhood. As in Figure 4 (c), **CART**, on the other hand, emphasizes



(a) Rank vs. Score (b) Ranking distribution (c) Score distribution

FIGURE 3. Characteristics of the university ranks and scores using Times 2020.

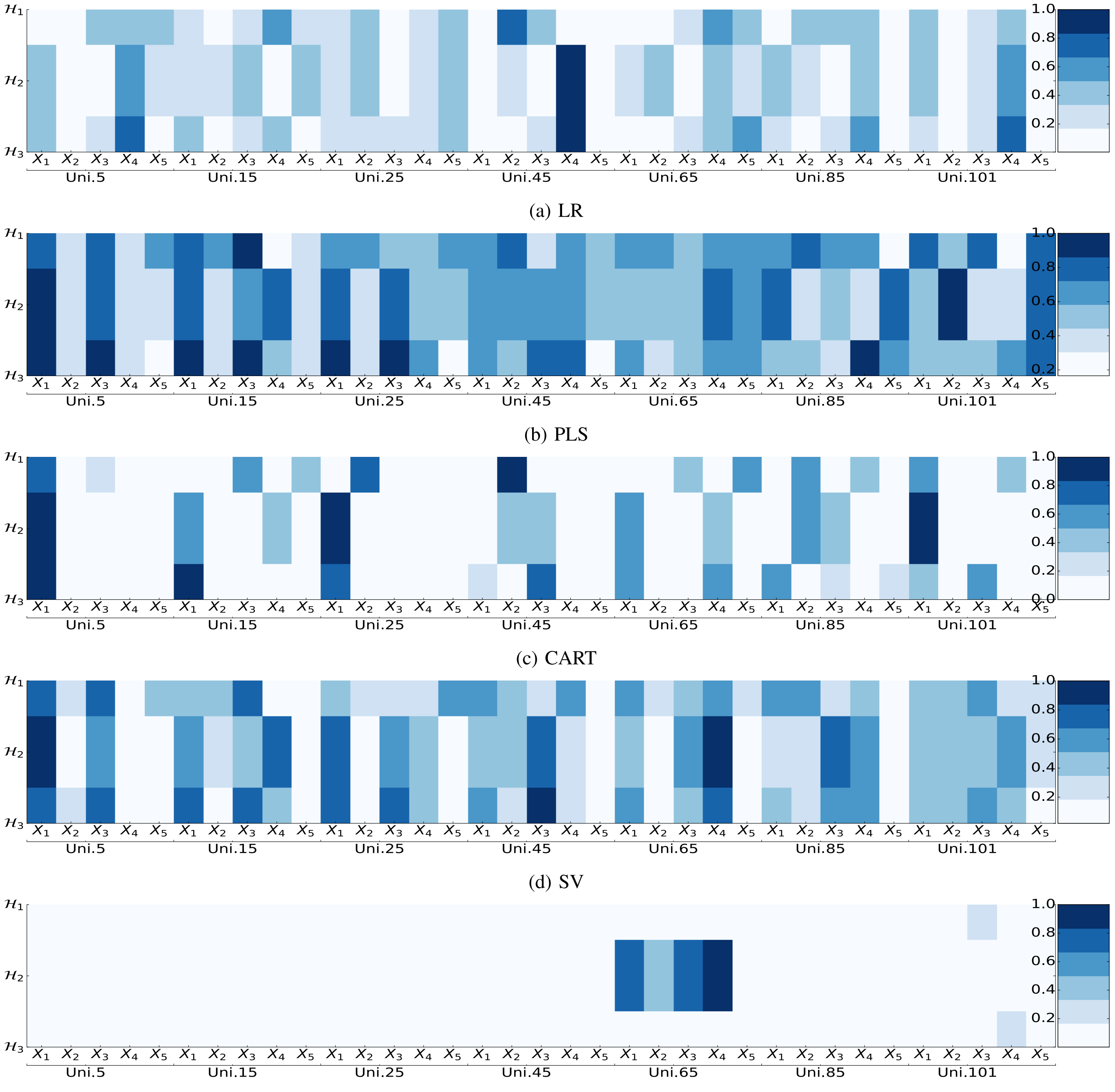
a single attribute and creates a sparse decision space within each neighborhood, which provides a limited point of view for the decisionmakers. As expected, CART disperses the output, which refers to its limited sensitivity to regression settings (ranking in our case). Consequently, the local impact of the attributes (especially the ones that have been dropped) is not properly reflected across neighborhoods. One can similarly confirm the underestimation of **LIME**, in Figure 4 (e), in identifying the importance of the attributes in many cases either within or across neighborhoods. This indicates the disadvantage of using **LIME** for competitive rankings, where the local perturbation approach of **LIME** generates non-existing instances and utilizes them to calculate the feature importance. As we can observe, both **PLS**, Figure 4 (b), and **SV**, Figure 4 (d), are capable of capturing the local impact of features across different neighborhoods and in most of the considered university instances.

A closer look at the range of [40-80] in the ranking list shows the ineffectiveness of **PLS** in distinguishing attribute importance across the three neighborhoods. Therefore, **PLS** does not reflect the local impact of attributes precisely. Moreover, **PLS** feature importance identification works similar to the global impact since the importance of attributes becomes less distinct in *immediate* and *near* neighborhoods. For example, **PLS** identifies all attributes to be important for Uni.45 especially in the *near* neighborhood. As we discussed in section V, the [40-80] range in the 2020 university ranking list is a critical interval due to the fact that these universities are less sticky in ranking and do not follow the linear relationship between score and ranking. This indicates that the local importance of features for the universities in this range significantly differs from their global importance due to the presence of latent factors (e.g., competition). In contrast, in the critical interval, **SV** in Figure 4 (d), strictly captures the dissimilarity in the importance value of each feature and explains the local impact of all features for different ranking entities across the list.

For example, see Uni.45 in Figure 4 (d), where X_1 and X_4 (i.e., teaching and citation) are identified by **SV** to be the most influential ranking factors in the *immediate* neighborhood, $\mathcal{H}_1$. However, moving from the *immediate* $\mathcal{H}_1$ to the *near* neighborhood $\mathcal{H}_2$, X_3 (i.e., research) becomes the most important attribute. For the *far* neighborhood, $\mathcal{H}_3$, X_3 contributes significantly to the ranking of the university. Such information implies that if university 45 wants to improve its ranking for the next evaluation cycle in competition with its close competitors, spending the budget on the teaching and citation agenda is better. Otherwise, to compete more globally ($\mathcal{H}_2$ or $\mathcal{H}_3$), its strategic plan should be more focused on research. Such interpretation of the outcome of our proposed framework provides an informative guideline for performance improvement for decisionmakers and higher-education policymakers.

Figure 5 provides a deeper insight into the influence of important features across defined localized neighborhoods for universities ranked below 101. The vertical axis represents the feature importance values. The figure has two horizontal axes: one shows the specific instance in the ranking list, and the other represents the attributes X_1, X_2, X_3, X_4, and X_5. The bars for each attribute for each instance indicate the average importance (i.e. $1/3 \times [(\mathcal{W}_1^s + \mathcal{W}_2^s + \mathcal{W}_3^s)]$) across neighborhoods. The larger the bar, the higher the importance of a feature in all neighborhoods combined. The size of the errorbars also demonstrates the variance of feature importance values across neighborhoods on each bar in Figure 5. In other words, each error bar shows the range of the change of a feature importance value moving from the *immediate* neighborhood, $\mathcal{H}_1$, to the *far* neighborhood, $\mathcal{H}_3$. Higher variance means that the value of feature importance for a specific attribute changes dramatically in the three neighborhoods. Note that the error bars indicate the model's sensitivity in feature importance calculation to the neighborhood size. As a result, a model with a larger errorbar better depicts variations in the importance of a feature from one neighborhood to the next. As we can observe in Figure 5, **SV** has larger error bars compared to other methods in the majority of cases.

As an example, for the university ranked 45 (i.e., Uni.45) Figure 5 (d) shows a large variance for attribute X_3 across three neighborhoods. Referring to Figure 4, we observe that the feature importance of X_3 dramatically increases moving from the *immediate* neighborhood, $\mathcal{H}_1$, to the *far* neighborhood, $\mathcal{H}_3$. This is important information for the



(a) LR

(b) PLS

(c) CART

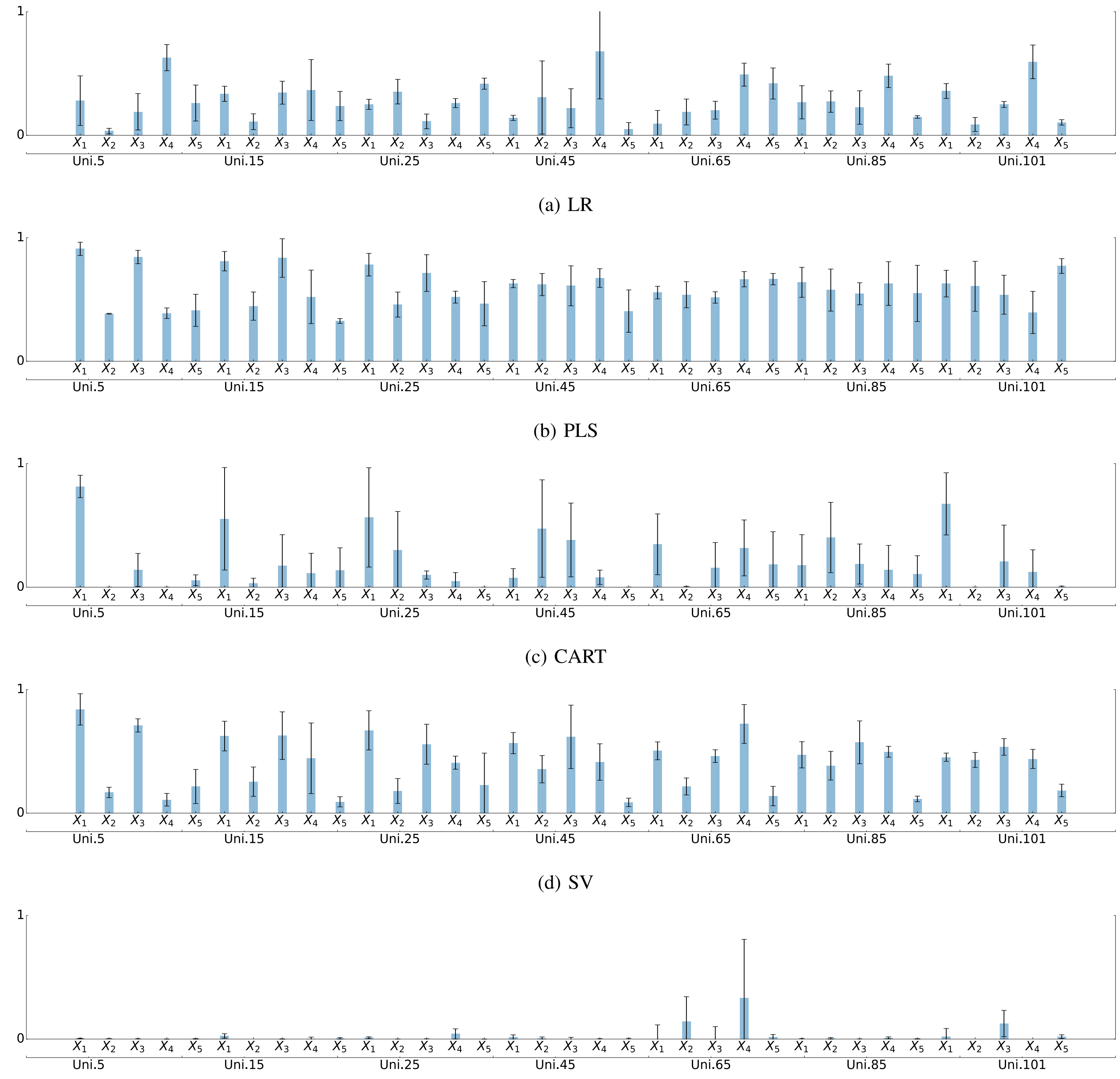
(d) SV

(e) LIME

FIGURE 4. **Heatmap of feature importance for different neighborhoods of university instances below 101.**

decisionmakers of university 45, first, to determine in which neighborhood they intend to compete for the next evaluation cycle, and then prioritize their resource allocation to the most important attribute based on their budget. More importantly, the attribute with a large variation across neighborhoods plays the key role in the competition within the neighborhood in which the attribute has the largest importance. In cases where the university wants to secure a rank in competition within a larger neighborhood ($\mathcal{H}_3$) of universities (e.g., making a significant jump to a lower rank), it needs to invest in X_3 more than other attributes.

For universities with rankings above 101, Figure 6 shows the heatmaps of different interpretation techniques. The results depicted in this figure are consistent with the observations for universities below 101 in the ranking list. **LR**, **CART**, and **LIME** fail to differentiate the feature importance values across different neighborhoods of several entities. **SV** performs better compared to **PLS**. Consider, for



(a) LR

(b) PLS

(c) CART

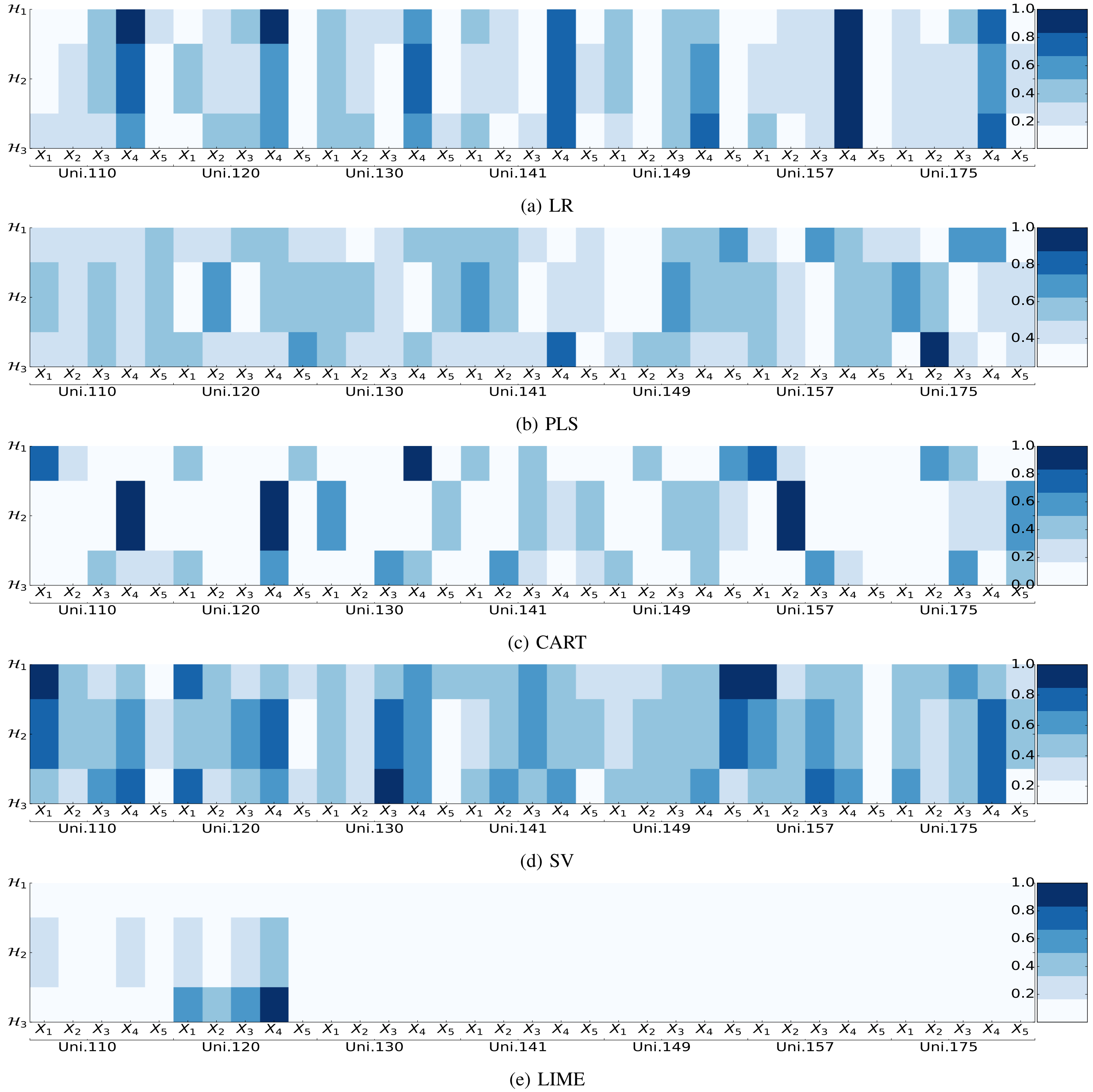
(d) SV

(e) LIME

FIGURE 5. **Barplot of mean and variance of feature importance for different neighborhoods of university instances below 101.**

example, the university ranked 157 (i.e., Uni.157). **SV** indicates how moving from the *immediate* neighborhood to the *far* neighborhood changes the priority of resource allocation for more effective competition. In contrast, **PLS** assigns relatively equal importance to the attributes across different neighborhoods. In addition, **SV** clearly distinguishes the important attributes within each neighborhood. For example, for Uni.157 in Figure 6 (d), X_1 (i.e., teaching) is identified by **SV** to be the most influential ranking factor in the *immediate* neighborhood, $\mathcal{H}_1$. However, moving from the *immediate* $\mathcal{H}_1$ to the *near* neighborhood $\mathcal{H}_2$, X_3 (i.e., research) becomes important as well. For the *far* neighborhood, $\mathcal{H}_3$, X_3 contributes significantly to the ranking of the university. This information helps the university administrators to compete with targeted competitors.

Figure 7 presents the average and variance of features' importance across neighborhoods for universities above 101. Recall that the higher variance means that the value of



(a) LR

(b) PLS

(c) CART

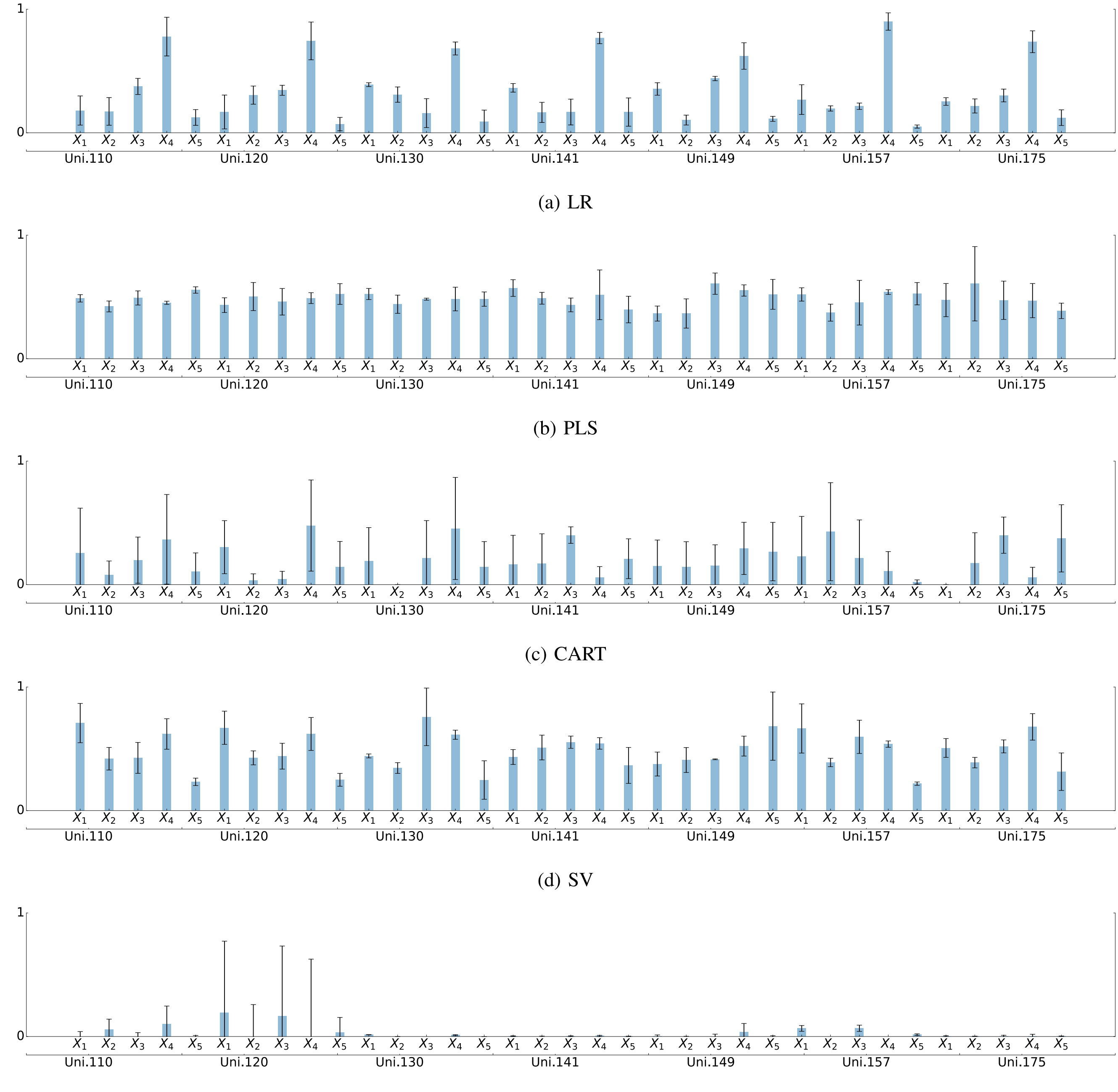
(d) SV

(e) LIME

FIGURE 6. Heatmap of feature importance for different neighborhoods of university instances above 101.

feature importance for a specific attribute changes dramatically across neighborhoods. As we can observe in Figure 7, **SV** still has larger error bars compared to **PLS** method. **PLS** demonstrates similar importance of attributes across neighborhoods, unlike **SV** that precisely distinguishes between variables.

We initially investigated the ranking dataset provided by *Times Higher Education* over two consecutive years to identify a university that exactly follows the weights obtained through our local explanation strategy. However, there exists no entity with an exact match. Since the ground truth of the local importance of attributes is not available, it is not possible to evaluate the correctness of any strategy. To determine whether the important attributes recommended by the proposed models are the truly important ones, we would need to implement a real-world label collection procedure (i.e., the



(a) LR

(b) PLS

(c) CART

(d) SV

(e) LIME

FIGURE 7. Barplot of mean and variance of feature importance for different neighborhoods of university instances above 101.

effectiveness of recommended strategy based on our provided weights) from universities that followed our proposed outcome to improve their ranking for the next evaluation period.

VII. CONCLUSION

In this study, we first investigated the global impact of universities in different rank groups. We then demonstrated that the important attributes differ for different rank groups. Moreover, we showed that the importance of attributes does not follow the global impact reported by the *Times Higher Education*. More specifically, the local influence of attributes depends on the rank position as well as the neighborhood consideration. We used correlation analysis and Principal Component Analysis for this purpose. We proposed a hierarchical ranking explanation framework to identify the local impact of attributes across the ranking list. Our proposed technique considers a hierarchical neighborhood construction to reveal the impact of attributes for a given instance. The discovery of the local impact of involved attributes in ranking potentially reflects the latent competition influence on the ranking. We used the most common feature selection techniques to identify the importance of attributes. More importantly, we introduced leveraging explainable AI techniques, including Shapely Value, as the most successful local and pointwise explanation methodologies for a competitive ranking explanation.

We evaluated the performance of our proposed framework using the *Times Higher Education* dataset for year 2020 as the baseline. We considered two groups of model-based and model-agnostic feature selection techniques along with our proposed framework to identify the importance of attributes in different neighborhoods. The results indicate that model-based techniques fail to distinguish between attributes' importance within each neighborhood. Partial Least Square and Shapely Value outperform other techniques in reflecting the local importance of attributes across the neighborhood hierarchy. However, Shapely Value competitively demonstrates the local importance of attributes for different university instances, especially for those in the critical ranking interval which are less sticky to the ranking and do not follow the linear relationship between score and ranking. Our findings in the experiment section advocate Shapely Value for the local explanation of competitive rankings.

In this study, we considered the ranking as static observations and we assumed the time effect is not significant. The framework can be applied for different rankings over different years. However, we aim to consider the time effect on the local importance of attributes for competitive ranking as a future direction. Moreover, we only considered super-categories of attributes for university ranking (i.e., research, teaching, etc.) and did not include the pre-weighted indicators (e.g., reputation survey, Institutional income per staff, etc. for teaching [3]). Due to the data limitation and accessibility we only focused on the publicly available attributes. Note that our proposed framework can be generalized to incorporate any size of attributes.

REFERENCES

[1] R. Morse, "How us news calculated the 2017 best graduate schools rankings," US NEWS, Washington, DC, USA, Tech. Rep., 2016. [Online]. Available: https://www.usnews.com/education/best-graduate-schools/articles/how-us-news-calculated-the-rankings

[2] TripAdviser. (2021). *Los Angeles Hotels and Places to Stay*. Aug. 28, 2021. [Online]. Available: https://www.tripadvisor.com/Hotels-g32655-Los_Angeles_California-Hotels.html

[3] Times Higher Education. (2021). *Methodology for Overall and Subject Ranking for the Times Higher Education World University Rankings*. Oct. 1, 2021. [Online]. Available: https://www.timeshighereducation.com/sites/default/files/breaking_news_files/the_2021_world_university_rankings_methodology_24082020final.pdf

[4] A. Gale and A. Marian, "Explaining monotonic ranking functions," *Proc. VLDB Endowment*, vol. 14, no. 4, pp. 640–652, Dec. 2020.

[5] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[6] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 883–892.

[7] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[9] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–9.

[12] P. Sen, D. Ganguly, M. Verma, and G. J. F. Jones, "The curious case of IR explainability: Explaining document scores within and across ranking models," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2069–2072.

[13] J. Zhan, J. Mao, Y. Liu, M. Zhang, and S. Ma, "An analysis of BERT in document ranking," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1941–1944.

[14] L. Li, Y. Zhang, and L. Chen, "EXTRA: Explanation ranking datasets for explainable recommendation," 2021, *arXiv:2102.10315*.

[15] M. T. Hoeve, M. Heruer, D. Odijk, A. Schuth, and M. D. Rijke, "Do news consumers want explanations for personalized news rankings," in *Proc. FATREC Workshop Responsible Recommendation*, 2017, pp. 1–6.

[16] M. ter Hoeve, A. Schuth, D. Odijk, and M. de Rijke, "Faithfully explaining rankings in a news recommender system," 2018, *arXiv:1805.05447*.

[17] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 180–186.

[18] S. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the Shapley value," *Other Words*, vol. 1, p. 98Eqr, 2005.

[19] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, "Building explainable artificial intelligence systems," in *Proc. AAAI*, 2006, pp. 1766–1773.

[20] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5491–5500.

[21] E. Momeni, C. Cardie, and N. Diakopoulos, "A survey on assessment and ranking methodologies for user-generated content on the web," *ACM Comput. Surveys*, vol. 48, no. 3, pp. 1–49, Feb. 2016.

[22] F. Ali and S. Khusro, "Content and link-structure perspective of ranking webpages: A review," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100397.

[23] J. Leonhardt, K. Rudra, M. Khosla, A. Anand, and A. Anand, "Fast forward indexes for efficient document ranking," 2021, *arXiv:2110.06051*.

[24] 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 *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 499–507.

[25] J. Kang, S. Freitas, H. Yu, Y. Xia, N. Cao, and H. Tong, "X-rank: Explainable ranking in complex multi-layered networks," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 1959–1962.

[26] M. Wang, J. Kang, N. Cao, Y. Xia, W. Fan, and H. Tong, "Graph ranking auditing: Problem definition and fast solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3366–3380, Oct. 2021.

[27] M. Verma and D. Ganguly, "LIRME: Locally interpretable ranking model explanation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 1281–1284.

[28] J. Singh and A. Anand, "Model agnostic interpretability of rankers via intent modelling," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 618–628.

[29] H. Chen, X. Chen, S. Shi, and Y. Zhang, "Generate natural language explanations for recommendation," 2021, *arXiv:2101.03392*.

[30] C.-E. Dias, V. Guigue, and P. Gallinari, "Personalized attention for textual profiling and recommendation," in *Proc. EARS@ SIGIR*, 2019.

[31] D. A. Reddy and G. Narasimha, "A multi-criteria decision approach for movie recommendation using machine learning techniques," in *Intelligent System Design*. Singapore: Springer, 2021, pp. 425–433.

[32] A. L. Zanon, L. Souza, D. Pressato, and M. G. Manzato, "WordRecommender: An explainable content-based algorithm based on sentiment analysis and semantic similarity," in *Proc. Brazilian Symp. Multimedia Web*, Nov. 2020, pp. 181–184.

[33] A. N. Refenes, M. Azema-Barac, and A. D. Zapranis, "Stock ranking: Neural networks vs multiple linear regression," in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, pp. 1419–1426.

[34] M. Zhu, D. Philpotts, R. Sparks, and M. J. Stevenson, "A hybrid approach to combining CART and logistic regression for stock ranking," *J. Portfolio Manage.*, vol. 38, no. 1, pp. 100–109, Oct. 2011.

[35] L. Fang, B. Xiao, H. Yu, and Q. You, "A stable systemic risk ranking in China's banking sector: Based on principal component analysis," *Phys. A, Stat. Mech. Appl.*, vol. 492, pp. 1997–2009, Feb. 2018.

[36] P. Macmillan and I. Smith, "Explaining international soccer rankings," *J. Sports Econ.*, vol. 8, no. 2, pp. 202–213, May 2007.

[37] G. E. Montanari and M. Doretti, "Ranking nursing homes' performances through a latent Markov model with fixed and random effects," *Soc. Indicators Res.*, vol. 146, nos. 1–2, pp. 307–326, Nov. 2019.

[38] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. Jagadish, and G. Miklau, "A nutritional label for rankings," in *Proc. Int. Conf. Manage. Data*, May 2018, pp. 1773–1776.

[39] N. K. Jajo and J. Harrison, "World university ranking systems: An alternative approach using partial least squares path modelling," *J. Higher Educ. Policy Manage.*, vol. 36, no. 5, pp. 471–482, Sep. 2014.

[40] M. McAleer, T. Nakamura, and C. Watkins, "Size, internationalization, and university rankings: Evaluating and predicting times higher education (THE) data for Japan," *Sustainability*, vol. 11, no. 5, p. 1366, Mar. 2019.

[41] K. Frenken, G. J. Heimeriks, and J. Hoekman, "What drives university research performance? An analysis using the CWTS Leiden ranking data," *J. Informetrics*, vol. 11, no. 3, pp. 859–872, Aug. 2017.

[42] A. Tabassum, M. Hasan, S. Ahmed, R. Tasmin, D. M. Abdullah, and T. Musharrat, "University ranking prediction system by analyzing influential global performance indicators," in *Proc. 9th Int. Conf. Knowl. Smart Technol. (KST)*, Feb. 2017, pp. 126–131.

[43] A. Mikryukov and M. Mazurov, "The task of improving the university ranking based on the statistical analysis methods," in *Proc. Int. Conf. Artif. Intell., Med. Eng., Educ.* Cham, Switzerland: Springer, 2020, pp. 65–75.

[44] J. Johnes, "University rankings: What do they really show?" *Scientometrics*, vol. 115, no. 1, pp. 585–606, Apr. 2018.

[45] R. Grewal, J. A. Dearden, and G. L. Llilien, "The university rankings game: Modeling the competition among universities for ranking," *Amer. Statistician*, vol. 62, no. 3, pp. 232–237, Aug. 2008.

[46] H. Almuallim and T. G. Dietterich, "Learning Boolean concepts in the presence of many irrelevant features," *Artif. Intell.*, vol. 69, nos. 1–2, pp. 279–305, 1994.

[47] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 284–292.

[48] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," 2009, *arXiv:0912.3924*.

[49] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. IJCAI Workshop Explainable AI (XAI)*, vol. 8, 2017, pp. 8–13.

[50] C. Molnar. *Interpretable Machine Learning*. Morrisville, NC, USA: Lulu.com, 2020.

[51] S. Wold, W. Johansson, and M. Cocchi, "PLS—Partial least- squares projections to latent structures," *3D QSAR Drug Des.*, vol. 1, pp. 523–550, 1993.

[52] S. Favilla, C. Durante, M. L. Vigni, and M. Cocchi, "Assessing feature relevance in NPLS models by VIP," *Chemometric Intell. Lab. Syst.*, vol. 129, pp. 76–86, Nov. 2013.

[53] H. Anysz, Ł. Brzozowski, W. Kretowicz, and P. Narloch, "Feature importance of stabilised rammed Earth components affecting the compressive strength calculated with explainable artificial intelligence tools," *Materials*, vol. 13, no. 10, p. 2317, May 2020.

[54] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. Wadsworth int," *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[55] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.

[56] M. Craven and J. Shavlik, "Extracting tree-structured representations of trained networks," *Advances in Neural Information Processing Systems*, vol. 8, 1995, pp. 24–30.

[57] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1803–1831, 2010.

[58] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, pp. 1–18, Jan. 2010.

[59] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 5686–5697.

[60] Ribeiro. (2019). *LIME Python Package*. Aug. 28, 2021. [Online]. Available: https://github.com/marcotcr/lime

[61] J. Singh and A. Anand, "EXS: Explainable search using local model agnostic interpretability," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 770–773.

[62] I. van der Linden, H. Haned, and E. Kanoulas, "Global aggregations of local explanations for black box models," 2019, *arXiv:1907.03039*.

[63] L. S. Shapley, *A Value for N-Person Games*. Princeton, NJ, USA: Princeton Univ. Press, 2016.

[64] J. Liang, J. Bisnett, A. Hylton, J. Sang, and C. Yu, "Parallel computation of standard competition rankings over a sorted array," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2020, pp. 1243–1249.

[65] A. Schram, J. Brandts, and K. Gërxhani, "Social-status ranking: A hidden channel to gender inequality under competition," *Exp. Econ.*, vol. 22, no. 2, pp. 396–418, Jun. 2019.

[66] Times Higher Education. (2020). *Worlds University Ranking*. Oct. 1, 2021. [Online]. Available: https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats

[67] M. T. C. Ribeiro. (2018). *Unstable Explanations When, No, Random Seed is Assigned in LIME Explain Instance*. Sep. 20, 2021. [Online]. Available: https://github.com/marcotcr/lime/issues/119

[68] S. Pal. (2019). *Shapley Value Regression*. Nov. 18, 2021. [Online]. Available: https://github.com/sarbadal/Shapley-Value-Regression

[69] M. T. C. Ribeiro. (2020). *LIME*. Nov. 18, 2021. [Online]. Available: https://pypi.org/project/lime/

HADIS ANAHIDEH received the Ph.D. degree in industrial engineering from the University of Texas at Arlington. She is currently a Research Assistant Professor with the Mechanical and Industrial Engineering Department, University of Illinois Chicago. Her research interests include sequential optimization, active learning, statistical learning, explainable AI, and algorithmic fairness.

NASRIN MOHABBATI-KALEJAHI received the Ph.D. degree in industrial and systems engineering from Auburn University. She is currently an Assistant Professor with the Information and Decision Sciences Department, California State University San Bernardino. Her research interests include decision-making under uncertainty, large scale optimization, and operations research and machine learning interface.