Cognitive Computation (2024) 16:3035–3050
https://doi.org/10.1007/s12559-024-10311-2

RESEARCH

Evaluative Item-Contrastive Explanations in Rankings

Alessandro Castelnovo[1] · Riccardo Crupi[1] · Nicolò Mombelli[1,2] · Gabriele Nanino[1] · Daniele Regoli[1]

Received: 30 November 2023 / Accepted: 20 May 2024 / Published online: 10 July 2024

Abstract

The remarkable success of Artificial Intelligence in advancing automated decision-making is evident both in academia and industry. Within the plethora of applications, ranking systems hold significant importance in various domains. This paper advocates for the application of a specific form of Explainable AI—namely, contrastive explanations—as particularly well-suited for addressing ranking problems. This approach is especially potent when combined with an Evaluative AI methodology, which conscientiously evaluates both positive and negative aspects influencing a potential ranking. Therefore, the present work introduces Evaluative Item-Contrastive Explanations tailored for ranking systems and illustrates its application and characteristics through an experiment conducted on publicly available data.

Keywords Explainability · Rankings · Artificial intelligence · Machine learning · Contrastive explanation

Introduction

In today's landscape, the practice of ranking individuals has become ubiquitous and pervasive. This ranking process finds its application in a multitude of scenarios, ranging from determining creditworthiness [1], suitability for college admissions or employment, or even assessing attractiveness in the context of dating [2]. Unlike traditional scenarios where the objective is to categorically differentiate, e.g., between suitable and unsuitable items, ranking involves the arrangement of items based on their relative merits. This distinction becomes particularly relevant in contexts where a constrained number of items can be accommodated within the final selection.

Consider the scenario of a traditional credit application process. Normally, financial institutions endeavor to provide loans to all clients with the capacity to repay the borrowed funds. However, the decision-making approach shifts when the bank is constrained by a predefined limit on the number k of clients to whom loans can be granted. This constraint necessitates a ranking process—the task of prioritizing potential borrowers based on their creditworthiness. This involves not only distinguishing appropriate clients from inappropriate ones but also arranging them in an order that matches the k available loan slots. In essence, ranking emerges as a mechanism to optimize the allocation of resources, especially when the number of deserving individuals exceeds the allocation capacity.

It is worth noting that while the concept of ranking is closely related to recommendation, the two are not interchangeable. Recommendation systems refer to applications in which users exclusively access the top-k items that result from the ranking process—e.g., the top 5 suggested movies [3, 4]. On the other hand, the ranking problem concerns scenarios where the interest is placed on the top-k elements, but users have access to information about and the ranks of all the items under consideration. A similar distinction is implied by [5], who talk about non-competitive vs. competitive ranking

The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates, or its employees.

✉ Alessandro Castelnovo
alessandro.castelnovo@intesasanpaolo.com

✉ Nicolò Mombelli
n.mombelli@studenti.unibs.it

✉ Daniele Regoli
daniele.regoli@intesasanpaolo.com

Riccardo Crupi
riccardo.crupi@intesasanpaolo.com

Gabriele Nanino
naninogabriele@gmail.com

[1] Data Science & Artificial Intelligence, Intesa Sanpaolo S.p.A., Turin, Italy

[2] Dept. of Economics and Management, Univ. Brescia, Brescia, Italy

problems. This work focuses on the latter context, and as a result, our assertions cannot be directly applied to pure recommendation problems.

As it is the case for many automated decision-making, Machine Learning (ML) currently represents one of the best alternatives to generate efficient and optimized rankings, and it has become a prevalent practice to address this challenge [6]. As ML models increasingly interact with humans, who hold the ultimate decision-making responsibility, the concept of eXplainable AI (XAI) has gained significant prominence alongside the advancement of AI technologies [7]. Indeed, ML models, in particular the most advanced ones, such as Deep Neural Networks (DNN), are opaque concerning the mechanisms through which their output is related to given inputs: this is the well-known black-box effect. XAI research represents the effort to come up with techniques to make the outcomes of non-interpretable ML models more and more comprehensible, facilitating the inclusion of human involvement through *Human-in-the-loop* [8], *Human-on-the-loop* [9], and *Human-in-Control* [10] approaches. Moreover, XAI could help reduce biases resulting from the use of AI systems, allowing for an ethical analysis of the model in use [11]. Particularly in the context of ranking, a common bias, known as *position* bias [12], emerges due to an item's position: higher-ranked items significantly influence user perception, being more likely to be examined and selected by users, even in cases of unreliable system [13, 14].

Significant advancements have emerged in the development of XAI techniques that align with human cognitive processes, such as contrastive and counterfactual explanations. Significant advancements have arisen in the field of XAI that align with human cognitive processes. These include techniques like contrastive and counterfactual explanations [15, 16] and Granular Computing [17], which organizes intricate information into granules—each containing closely related internal details but loosely connected to external data.

In the field of XAI, there seems to be an overlap in the concepts of *contrastive* and *counterfactual* explanations [18]. However, they may not be epistemically equivalent. Table 1 helps to better understand the distinction between these concepts based on the problem. Specifically, contrastive explanations facilitate human comprehension by shedding light on the rationale behind choosing one outcome over another. This form of explanation is widely recognized as both effective and easily understandable. However, these explanations are well-defined and formalized primarily for classification problems, where the contrastive explanation is based on the relative position of instances with respect to the decision boundary. In the context of rankings, to the best of our knowledge, this concept still lacks proper formalization.

Contributions The main contribution of this work is to introduce and formalize contrastive explanations in the context of ranking problems. To enhance the support for human decision-making while mitigating the impact of position bias, we align our formalization with the paradigm of Evaluative AI, proposed by [19].

We call such an approach *evaluative item-contrastive explanations*. We delineate four general steps to obtain evaluative item-contrastive explanations, adhering to the principles of a good explanation outlined in [20].

Since Evaluative AI aims to provide users with the pros and cons linked to decisions proposed by automated systems, it is necessary to follow Granular Computing principles [21], which assist in organizing complex information accordingly.

Furthermore, we exemplify the practical implementation of this framework using a linear model and present experimental results across diverse domains to establish the generalizability of the approach. The first experiment pertains to its application within the domain of recruitment [22], while the second examines its efficacy in addressing credit card churn [23]. Both datasets employed in these experiments are openly accessible. Upon publication, we will provide details on code implementation to allow the reproducibility.

Given that this work represents the first attempt to derive contrastive explanations for ranking problems, we hope that it will lay the groundwork for future research in this area.

Background on Explainability

Understanding the reasons for explanations, the characteristics of a good explanation, and the distinction between contrastive and counterfactual explanations provides the necessary groundwork for formalizing our proposal to enhance interpretability in ranking systems. Furthermore, we will present the recent line of work on explainability towards

Table 1 Synthetic conceptual representation of how counterfactual and contrastive explanations articulate with different types of problems. Further details and bibliographic references will be provided in "Background on Explainability" section

Problem	Counterfactual XAI	Contrastive XAI
Classification	What to change in input to obtain a different classification	Why a specific classification occurs instead another
Recommendation	What to change in input to obtain a different recommendation	Why a specific recommendation occurs instead another
Ranking	What to change in input to Obtain a overall different rank	No proper formalization

Evaluative AI, a paradigm we believe is the most suitable to follow in shaping our proposal.

Reasons for Explanations

Explanations of outcomes in a decision-making process are useful from several perspectives, including [24]

- *Explain to justify*: to justify the decisions made using an underlying model. Explaining the reasons behind decisions enhances their justifiability and helps build trust among stakeholders.
- *Explain to discover*: to support the extraction of novel knowledge and the discovery of new relationships and patterns. By analyzing the explanations provided by AI systems, researchers can grasp hidden mechanisms and gain a deeper understanding of the data and underlying processes and phenomena.
- *Explain to control*: to enhance the transparency of an outcome, proactively confirming or identifying potential issues. Understanding system behavior provides increased visibility over potential vulnerabilities and flaws, facilitating rapid error identification and correction. This enhanced control empowers better system management
- *Explain to improve*: to aid scholars and practitioners in improving the accuracy and efficiency of their models. By analyzing the explanations, insights can be gained on how to enhance the model's performance and make it more effective in its intended task.

What is a Good Explanation

We here rely on the analysis proposed by [20], according to which humans perceive an explanation as good when it possesses four key properties: to be *contrastive*, *selected*, *social*, and *not to rely on probabilities* and statistical relationship when presenting explanations.

Contrastive explanations are designed to shed light on why a particular input yields a specific output *instead of* an alternative output [25]. They provide insights into the factors differentiating the chosen outcome from alternative possibilities.

Providing *selected* explanations means that good explanations should not include *all* the reasons that are causing an output: giving a complete account would typically require too much information, most of which would not be relevant for a given context and purpose. Therefore, a good explanation should consist of a limited but relevant subset of such information. People generally expect explanations that offer a concise and focused account of causative factors, as excessively lengthy explanations might be confusing or challenging to grasp. Existing work has already looked at selecting which features in the model were important for a decision, based on local explanations [26, 27] or on information gain [28, 29].

Explanations are more effective when they are set in the landscape of the recipient's existing beliefs and values. It is crucial for explanations to be tailored to the *social* context of the evaluator [25]. This entails not only fitting the individual's knowledge but also accommodating their self-perception and surroundings. In some cases, a mismatch between explanation and expectation can lead to under-reliance and significant loss of trust despite AI system performance [30].

To effectively communicate explanations, one should avoid incorporating probability and statistical arguments, as humans struggle with handling uncertainty [20]. Probability and statistics do not provide a clear intuition to most individuals, thus they represent a poor strategy to explain anything.

Contrastive Explanation

The main reference for the concept of contrastive explanation is [31]. The basic idea of giving a contrastive explanation of P is that of explaining why P rather Q, where P is the fact that obtained and Q is a hypothetical fact that did not occur. In order to explain P, a contrastive explanation points to the differences in the causal histories of P and Q (that plausibly made the first event happen, instead of the second). It is important to note that contrastive explanations are inherently perspectival, being relative to a defined pair made of a fact and a hypothetical fact. This perspectival nature implies that each explanation may vary, offering distinct informational content. It is debated whether the fact and the hypothetical fact of a contrastive explanation should be incompatible. Intuitively, they may not always be incompatible. However, in our context of ranking problems, we may assume that facts and hypothetical facts are indeed incompatible due to the constraints in ranking systems. For example, when dealing with the positions of items in a ranked order, practical considerations often lead to the imposition of restrictions, such as the prohibition of two items occupying the same position in a rank. Consequently, admissible contrastive explanations are typically confined to those presenting incompatible items or events.

Different kinds of contrastive explanations are recognized in the literature [20]. The more common approach is P-contrast, which consists in posing the question of why the item a has a certain property P, rather than Q. This approach aligns with the standard fact-foil structure discussed above. Additional approaches include O-contrast [32, 33], inquiring why does the item a have property P, while item b has property Q, and T-contrast [34] where the question is why does the item a have property P at time t, but property Q at time $t + \delta$?

In our formalization, detailed in "Formalizing Item-Contrastive Explanation in Ranking" section, we will extend upon the concept of O-contrast, specifically tailored to address ranking problems. This involves asking why item a has been ranked higher than item b.

Notice that the contrastive approach in ranking represents a fine-grained form of explanations: namely, it works at the level of specific couples of items. Indeed, contrastive and counterfactual explanations help clarify AI decisions by highlighting why one particular outcome was chosen over another. In this respect, Granular Computing [21, 35]—a paradigm that organizes complex information into granules, where each granule contains closely related internal information but is loosely connected to external data—naturally fits to Contrastive approaches to XAI. By merging the paradigm of Granular Computing with Contrastive and Counterfactual Explanations, XAI can be made more accessible and impactful, tailoring explanations to match human cognitive processes and diverse levels of expertise. This integration simplifies the complexities of AI decisions into more comprehensible segments, boosting user understanding, trust, and transparency in AI systems.

The Evaluative AI Paradigm

Typical implementations of XAI techniques, including those providing contrastive explanations, result in systems that return the recommended output together with its (contrastive) explanation. As argued in [36], this is likely to be a limited approach. A more effective strategy for high-stakes decisions would be to shift from recommendation-driven decision support to hypothesis-driven decision support, as proposed by [19]. Miller [19] calls his proposed paradigm "Evaluative AI": in short, instead of presenting reasons for a certain outcome or recommendation—e.g., why item a is preferable (or has a higher place in ranking) with respect to item b—evaluative implementation of contrastive explanation would present reasons *for and against* each of the two items.

Notice that Evaluative AI is still explainable AI, as [19] clarifies. This paradigm is particularly well-suited for assessing and navigating trade-offs between different factors. In ranking problems, the score assigned to each item has meaning only relative to that of all the other items. For this reason, we believe ranking problems to be quite a natural setting for a contrastive explanation within the Evaluative AI paradigm. This approach is more effective for decision support because, as argued by [19], it aligns with the cognitive decision-making process that people use when making judgments and decisions, it has the potential to effectively reduce biases that affect decisions based on rankings. In particular, it could counteract the negative influence on the cognitive decision-making process of the position bias, which arises from the very nature of the ranking framework, especially under the no-ties assumption.

Related Works on Counterfactual and Contrastive Explanation in Ranking

In the domain of contrastive and counterfactual explanations, a significant portion of the literature in supervised ML is dedicated to explaining classification and regression models [see, e.g.,18, 37]. Relatively less attention has been devoted to understanding and explaining the ranked outputs produced by these models [38]. The most notable exceptions are the works by [39] and [40].

Even if there is considerable overlap between ML models designed for classification and those designed for ranking, we want to stress the fact that contrastive and counterfactual examples designed for classification models may not be directly applicable to ranking systems. In the context of counterfactuals in ranking, the challenge extends beyond explaining what needs to be changed in order to receive a different outcome, as it is crucial not only to understand the impact of changes in a single item on its ranking but also to discern how alterations in one item reverberate and influence the rankings of other items in the list. Salimiparsa [39] contributes to adapting existing counterfactual explanations to suit ranking purposes: unlike traditional methods, the proposed approach integrates the position of an item in the list, investigating how modifications to the item can impact its ranking. The main objective is to determine the minimum change required for an item to secure a different rank compared to other items on the list.

Tan et al. [40] builds on the literature of counterfactual explanations and applies it in the context of recommendation systems. In this way, they apply a contrastive-type explainable AI technique to something very similar to a ranking problem (even if not quite the same, as we argued in the Introduction).

To the best of our knowledge, besides the works of [40] and [39], no other significant contributions have been made in the literature regarding contrastive and counterfactual explanations of ranking systems. Our work aims to address this gap by concentrating on contrastive explanations rather than solely focusing on counterfactual ones. Specifically, we strive to adapt the existing approach for contrastive explanation to the context of ranking, with a shift in focus from the foils to the items, thus enhancing the understanding of the underlying mechanisms within ranking systems.

Formalizing Item-Contrastive Explanation in Ranking

Setting the Stage: How Ranking Systems Work

We here introduce the typical setting of ranking problems. To illustrate our setup, we hereafter leverage the example of selecting candidates for a company job interview, without loss of generality.

Let $D = \{d_1, \dots, d_n\}$ denote the set of n candidates supposed to be organized in a ranking list. Each individual is identified by a set of p relevant attributes (e.g., skills) for performing a specific task that we label with $X_i = (X_i^1, \dots, X_i^p)$.

The goal is thus to assign to each candidate d_i a score Y_i that induces a (desired) permutation π of the elements in D. By $j = \pi_i$, we denote the individual i in D that is assigned to the position j in the rank, where $j = 1, 2, \dots, n$.

In general, a better position in the rank denotes a better *utility* of the item for the user [41, 42]. In typical ranking applications, users are interested in a sub-ranking containing the best $k < n$ items of π. In our example, the company's hiring team (the user) is interested in sorting candidates based on their suitability for the vacant position and choosing the best k to invite for an interview, k being constrained by the time and resources the company can afford to spend in the hiring process.

Item-contrastive explanations can assist the hiring manager in justifying and understanding why one candidate was positioned more favorably than another within a given pair of candidates. The goal is to enhance the freedom of action of the user by providing her information to eventually act on the members of the sub-ranking and change their position.

Evaluative Contrastive Reasoning

We want to focus on a specific type of contrastive explanation, namely the one trying to answer the following question: "Why has item a been ranked higher than item b?".

Since in ranking problems, the position of item a in the ranking list is determined not only by the score assigned to a (Y_a), but also by the scores assigned to all the other items, or, better by the contrast between Y_a and the other items' scores, the proper explanation is grounded in the concept of O-contrast explanation, which highlights the contrasts between objects (items in our context). As outlined by [32], an effective O-contrasts explanation takes the form:

"Object a has property P, while object b has property P' because a has properties $X1$, ..., Xn which object b does not have".

In our scenario, relevant are all those differences between a and b that mattered for the assignment of the respective rankings. In particular, an explanation for a given case would be of the form "a has been ranked in a better position than b because a exhibits certain properties that b does not have (or has less)".

However, answering "why has *item a* been ranked higher than *item b*?" using just the reasons in favor of the candidate ranked in the higher position could enforce the bias to the user already given by the position in the rank. Aware of these risks, we are aligned with [19] in thinking that this approach could shrink the space the user has for autonomous reflection, especially in cases of high-stakes decisions. As a consequence, her degree of control over the final outcome may be reduced. To adjust this form of explanation, we want to put the user in the position to perform further reasoning. To do so, we propose an evaluative contrastive explanation in ranking that mentions also the attributes in which item b scored better than item a but were weighted with lower importance by the ranking algorithm. The form we propose is, therefore, the following: a has been ranked in a better position than b because a exhibits certain attributes or characteristics that b does not have (or has less); however, b exhibits certain attributes or characteristics that a does not have (or have less), albeit with relatively lower importance assigned by the system.

General Approach for Implementation and Application with Linear Model

We focus on proposing guiding principles for the design and development of XAI approaches in line with our conceptualization of evaluative item-contrastive explanation, integrating the principles for a good explanation outlined in "What is a Good Explanation" section. In particular, we define the four steps for implementing a good evaluative item-contrastive explainer represented in Fig. 1, to build

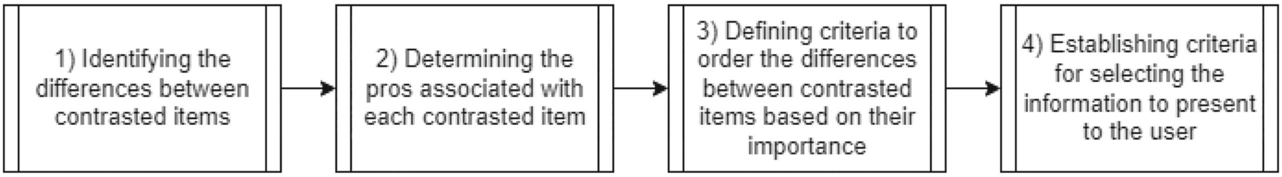
1) Identifying the differences between contrasted items
2) Determining the pros associated with each contrasted item
3) Defining criteria to order the differences between contrasted items based on their importance
4) Establishing criteria for selecting the information to present to the user

Fig. 1 Illustration of the four general steps to follow for implementing an evaluative item-contrastive explainer in line with the principles of a good explanation

an explanation that is *selected*, *contrasted*, and *evaluative*. These steps serve as general guidelines independently from the explainer adopted.

Furthermore, we present a proposal for concretely implementing these four steps starting with a rank created from a Logistic Regression (LR) model. The choice of the LR is motivated by its status of interpretable model, as acknowledged by the literature [43–45]. Therefore, we can showcase our proposed approach without the risk of complicating our exemplar analysis with technicalities specific to a given XAI or ML method. Additionally, LR is frequently utilized in conjunction with complex black-box models to provide insights and explanations regarding their underlying decision-making processes [46].

Identifying the differences between contrasted items For a linear model, the overall discrepancy between items a and b can be expressed as

$$\Delta_{a,b} = \sum_{d=1}^{p} \Delta_{a,b}^{d} = \sum_{d=1}^{p} \alpha_d (x_a^d - x_b^d), \qquad (1)$$

where p is the total number of features and α_d is the weight associated by the LR to each feature d. x_a^d and x_b^d represent the values of the regressor x^d associated to items a and b, respectively.

Determining the pros associated with each contrasted item The linear nature of LR streamlines the identification of pros and cons for each item. A positive coefficient α_d in LR, assigned to a specific feature, implies that higher values of that feature correspond to higher assigned scores. Conversely, a Negative Coefficient (NC) indicates an advantage—in terms of score—for the item with a lower value. As a consequence, a positive $\Delta_{a,b}^{d}$ implies a pro for candidate a (and a con for item b) due to feature d, while a negative value has the opposite effect.

Defining criteria to order the differences between contrasted items based on their importance A natural criterion is to compute the importance of each contribution to the overall difference as

$$|\Delta_{a,b}^{d}| = |\alpha_d (x_a^d - x_b^d)|. \qquad (2)$$

In this way, importance is contingent on both the magnitude of the weight assigned by the LR $|\alpha_d|$ and the extent of difference between the raw feature values of the two items, namely $|x_a^d - x_b^d|$.

Establishing criteria for selecting the information to present to the user This step is due to the need to reach a balance between offering a concise and selected explanation to the users while at the same time providing sufficient information for them to make a well-informed evaluation of the items under investigation. Therefore, it is clear that the final configuration embedding this trade-off can only be context-dependent. However, technically, it is possible to offer configurable methods to facilitate this decision-making process. These methods could be configured to select the top z features for each contrasted item or to pick the top feature that covers a minimum level of cumulative importance. Furthermore, a mixed method can be implemented: selecting the top features that meets a minimum threshold of cumulative importance, and additionally, including a minimum number of features as pros for the item if they were not included based on the initial criterion.

The set of presented differences should be relevant for the social context and the organizational culture of the intended user. This involves both the format used to present the explanation and the pertinence of its content with respect to the background knowledge of the recipient. This is a requirement that extends beyond the type of implementation used to generate the ranking algorithm and depends on the context of the application and its respective users. In line with the insights from [47], we suggest adopting natural language explanations and visual representation to expose the differences between contrasted items.

In particular, the textual description should be designed to empower human decision-making by promoting a comprehensive evaluation of both candidates, enhancing the understandability of the information. The explanation's sole focus should be on accentuating the disparities between the items. Consequently, it should deliberately maintain conciseness, employ straightforward syntax, and refrain from including numerical data and percentages. It is recommended to avoid using judgmental (e.g., right/wrong, good/bad) or qualifying (e.g., solid, interesting, worth noting) expressions, alongside refraining from employing expressions implying algorithmic endorsement, which may potentially influence the user's perception [48].

In the upcoming section, we present examples of the evaluative item-contrastive approach, including both textual and graphical representations, applied to real-world scenarios.

Experiment

General Setting

In order to illustrate the generalizability of the evaluative item-contrastive explanations approach, we have investigated its efficacy across two distinct domains, utilizing open-source datasets to underscore the reproducibility of the results obtained.

The first investigation focuses on the recruitment process of recent MBA graduates from an Indian college, presenting a scenario wherein a recruiter is constrained to select only a predetermined number of candidates. This analysis is conducted on the *Campus Recruitment* dataset [22][1].

The second experiment explores customer churn within the credit card industry, whereby churn denotes the cessation of card usage and subsequent closure of the associated credit card account. Here, we simulate the role of an employee tasked with retention efforts under budget constraints. The study employs the *Credit Card Customer* dataset [23][2].

In each experiment, we introduce the input dataset, the designated target variable and a brief description of the data preparation phase.[3] As elucidated in "General Approach for Implementation and Application with Linear Model" section, in both scenarios, we exploited an LR model to forecast the placement of an item on the basis of the values of several features. More precisely, to attain the dual goals of dropping non-significant features and reaching satisfactory performances, a pre-processing phase of backward step-wise feature selection has been used. In particular, we employed the p-value as a metric to choose the candidate feature to be removed at each step and the significance level (p-value less than 5%) as a criterion whether to retain or drop the selected feature. Furthermore, to enhance the reliability and generalizability of our experimental findings, we employed a 5-fold stratified Cross Validation. This approach ensures that the findings are not dependent on a particular partition of the dataset.

The LR model is then applied out-of-sample on the remaining set of candidates to extract the corresponding ranking scores. As a concrete illustration of the item-contrastive approach, we conduct a comparative analysis of two elements extracted from this sample, elucidating the guiding rationale behind their respective positioning through both graphical and textual means. This textual description is generated through an automated function.

Experiment 1: Recruitment

The *Campus Recruitment* dataset on academic and employability factors influencing placement consists of records on job placement of 215 students from an Indian University campus. In particular, it contains information about students' education, from secondary school to post-graduate specialization. Other information about the education system and the working experience is also present. The schema of the dataset is presented in Table 2. We refer to [22] for additional info on the data.

In our experiment, we use `STATUS` as binary target variable (1 placed, 0 not placed). The dataset counts 148 hired and 67 unemployed students.

During the data preparation phase, categorical features have been one-hot encoded, while numeric features were pre-processed via standard scaling in order to make their coefficients comparable for the evaluative phase. The LR learned coefficients for the selected features are shown in Fig. 2. Notably, the analysis reveals the significance of attending commercial or scientific programs during the high secondary school, along with the grades attained in such studies. Moreover, students with working experience seem to be strongly advantaged with the perspective of job placement. Table 3 displays the rank and the significant features of the 10 candidates with the highest score. In the following, we shall employ this set of ten candidates as working examples to showcase our approach.

Constructing Evaluative Item-Contrastive Explanations for Recruitment

We assume that the resulting rank obtained from the LR model represents the ordered list of candidates presented to hiring managers for selecting a limited number (e.g., $k = 5$) of candidates for interviews. Following our methodology, we consider an exemplar scenario in which managers start by engaging in pairwise comparisons. For instance, a recruiter may be interested into evaluate the rationale behind the positioning of candidates ranked at positions 5 (candidate `00079`) and 6 (candidate `00188`). This particular choice addresses a larger gap among both the numeric features and the final score of the items. However, it is paramount to emphasize that our approach is not reliant on these individual cases.

Aligning to what delineated in "General Setting" section, in this showcase, the explanation returned by the system comprises both graphical comparisons and textual support. As mentioned in "General Approach for Implementation and Application with Linear Model" section, the displayed amount of information depends on the context. In our example, given that the number of features considered by the final model has already been filtered by a feature-selection procedure, a scenario in which all available information is provided to the user is outlined.

Figure 3 provides a comprehensive explanation by incorporating both computed model weights and feature importance. The graphical depiction showcases feature contributions, with the length of each bar indicating the magnitude of the contribution and direction indicating the respective item to which it is provided. Null bars represent no relevant contribution for any candidate. In particular, the contribu-

[1] The dataset is publicly available at Campus Recruitment dataset.

[2] The dataset is publicly available at Credit Card Customer dataset.

[3] A more detailed discussion of data preparation can be found in the code used to perform experiments that will be made openly available upon acceptance.

Table 2 Schema of the *Campus Recruitment* dataset. For each variable, we report the name along with the description, the data type, and the domain

Variable	Description	Type	Domain
SL_NO	Student id	Numeric	[1; 215]
GENDER	Student gender	Categorical	Female/Male
SSC_P	Secondary education percentage grade, 10th grade	Numeric	[0; 100]
SSC_B	Secondary education board of education	Categorical	Central/others
HSC_P	High secondary education percentage grade, 12th grade	Numeric	[0; 100]
HSC_B	High secondary education board of education	Categorical	Central/others
HSC_S	Specialization of high secondary education	Categorical	Science/Art/Commerce
DEGREE_P	Undergraduate degree percentage grade	Numeric	[0; 100]
DEGREE_T	Field of undergraduate studies	Categorical	Comm&Mgmt/Sci&Tech/others
WORKEX	Previous work experience	Categorical	Yes/no
ETEST_P	Employability test percentage	Numeric	[0; 100]
SPECIALIZATION	Type of MBA specialization	Categorical	Mkt&HR/Mkt&Fin
MBA_P	MBA percentage grade	Numeric	[0; 100]
STATUS	Placement status	Categorical	Placed/not placed
SALARY	Job salary, if any	Numeric	[0; +∞[

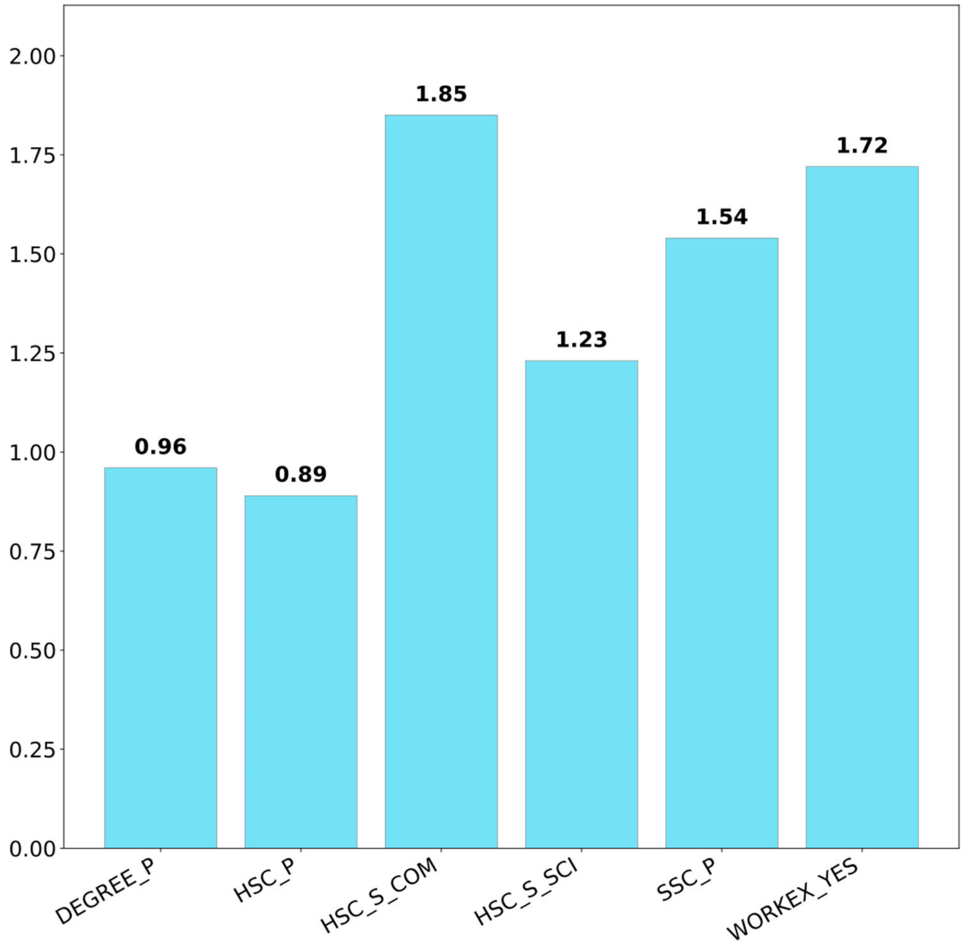

Fig. 2 Coefficients of the LR model for the recruitment case. The analysis underscores the importance of participation in commercial/scientific programs during high school, as well as the related grades achieved. Students with work experience appear to have an advantage

Table 3 Top 10 candidates sorted by model's output scores. Each entity is provided with the identification code, the ranking position and the LR model's forecast (SCORE). Additionally, the most influential features as determined by the pre-processing algorithm have been included. Candidates with shaded background are those chosen to elucidate the functionality of the proposed solution as discussed in "Constructing Evaluative Item-Contrastive Explanations for Recruitment".

ID	RANK	SCORE	DEGREE_P	HSC_P	HSC_S_COM	HSC_S_SCI	SSC_P	WORKEX_YES
00034	1	0.99933	81.0	65.0	0	1	87.0	1
00029	2	0.99648	67.5	76.5	1	0	76.76	1
00139	3	0.9959	73.0	64.0	0	1	82.0	1
00097	4	0.99578	76.0	70.0	0	1	76.0	1
00079	**5**	**0.99418**	**64.5**	**90.9**	**0**	**1**	**84.0**	**0**
00188	**6**	**0.9872**	**67.0**	**65.5**	**0**	**1**	**78.5**	**1**
00140	7	0.98367	59.0	70.0	1	0	77.0	1
00070	8	0.98218	66.0	73.0	0	1	73.0	1
00063	9	0.9769	67.4	64.2	0	1	86.5	0
00072	10	0.97364	71.0	70.29	1	0	75.0	0

tion of each feature towards the final score is computed as percentage on the overall discrepancy. Candidate 00079 predominantly benefits from having recorded higher marks during secondary education, with high-secondary education (HSC_P) contributing the most and (low-)secondary grades (SSC_P) approximately half of it. Conversely, candidate 00188 derives primary support from prior work experience, with a smaller contribution from higher marks in the bachelor's degree. Finally, since they both attended the same high-secondary studies (namely, scientific studies), this feature is not a discriminator among the two of them.

Alongside the visual representation comes the textual explanations that, in our approach, is structured as the following example:

The available information regarding Candidate 00079 and Candidate 00188 suggests that both individuals are qualified for the job. Candidate 00079 is ranked higher than Candidate 00188 according to the current algorithm reasoning. However, the ultimate decision remains within your control, offering the option to alter this ranking if desired. Characteristics in favor of Candidate 00079 include a higher score in HSC_P and a higher score in SSC_P. Characteristics in favor of Candidate 00188 include a higher score in DEGREE_P and having previous working experience.

This comparative analysis serves the dual purpose of either confirming the validity of the existing rank or potentially prompting adjustments to the final candidate selection for interviews.

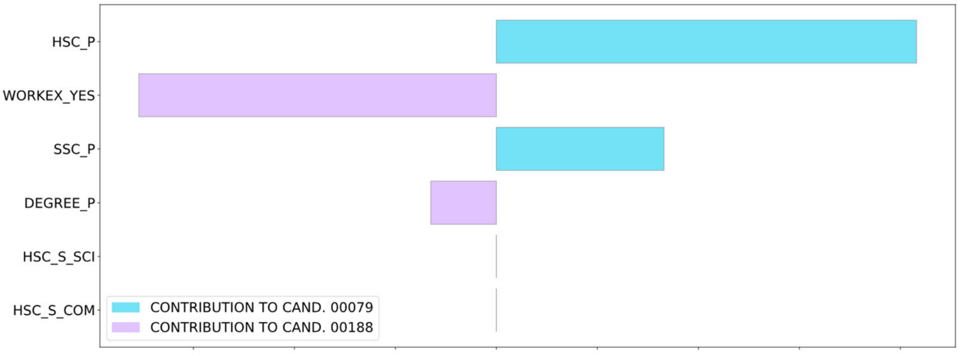

Fig. 3 Feature contributions to support the disparity in ranking among candidates 00079 and 00188. While candidate 00079 is favored by having higher marks during the secondary school, candidate 00188 benefits from having previous work experience and higher marks during the bachelor degree. Since they both attended the same high-secondary studies, no contribution is provided by this feature

Experiment 2: Churn

The *Credit Card Customer* dataset serves as a comprehensive repository of churn activity data pertaining to credit card holders within a specific financial institution. Comprising approximately 10,000 records and 21 columns, the dataset encompasses a wide array of demographic and customer relationship information pertinent to the institution's clientele. The schema and the domain of the source are showcased in Table 4. We refer to [23] for additional info on the data.

ATTRITION_FLAG has been selected as target variable for this scenario. In particular, since only 16% of the customer considered ceased to use a credit card, oversampling has been exploited to re-balance the target variable to around 30%. During the data preparation phase, ordinal categorical variables such as INCOME_CATEGORY have been cast to numeric, while non ordinal ones have been one-hot encoded. During this phase, we also took care of removing variables with high correlation level that could undermine the performances of the LR model. Moreover, numeric features were pre-processed via standard scaling in order to make their coefficients comparable for the evaluative phase.

The LR model is utilized as outlined in "General Setting" section. The acquired coefficients of the model pertaining to the features selected through the aforementioned procedure are depicted in Fig. 4. The propensity to churn is more evident for customers who have been inactive for several months in the last year, for those who have had a high number of contacts, and for those who have dependents. Customers who have increased the number of transactions in the fourth quarter compared to the first, who are married/single, or who have a high number of other relationships with the bank, on the other hand, show a lower propensity to churn. Table 5 collects the customers most likely to start churn actions along with the features found significant by the algorithm. This sample will be utilized in the prosecution to offer a tangible illustration of item-contrastive explainability.

Constructing Evaluative Item-Contrastive Explanations for Churn

The outlined scenario pertains to an employee positioned within a financial institution ascribed with the task of executing customer retention strategies. However, constrained by budgetary and time limitations inherent to their role, the employee endeavors to exclusively engage with clients deemed susceptible to churn. The proposed ranking system constitutes the initial phase in the employee's decision-

Table 4 Schema of the *Credit Card Churn* dataset. For each variable, we report the name along with the description and the data type. A sample of the possible values is also reported to provide an intuitive comprehension of the domain

Variable	Description	Type	Domain
CLIENTNUM	Client identifier for the customer holding the account	Numeric	$[0; +\infty[$
ATTRITION_FLAG	Customer activity	Binary	1 if closed account, else 0
CUSTOMER_AGE	Customer's age in years	Numeric	$[0; +\infty[$
GENDER	Customer's gender	Categorical	M=Male, F=Female
DEPENDENT_COUNT	Number of dependents	Numeric	$[0; +\infty[$
EDUCATION_LEVEL	Educational qualification of the account holder	Categorical	e.g. high school, graduate
MARITAL_STATUS	Customer's marital status	Categorical	e.g., married, single, divorced
INCOME_CATEGORY	Annual Income category of the account holder	Categorical	e.g. $< \$40K$, $\$40K$ - $60K$
CARD_CATEGORY	Type of card	Categorical	Blue, silver, gold, platinum
MONTHS_ON_BOOK	Period of relationship with bank	Numeric	$[0; +\infty[$
TOTAL_RELATIONSHIP_COUNT	Total no. of products held by the customer	Numeric	$[0; +\infty[$
MONTHS_INACTIVE_12_MON	No. of months inactive in the last 12 months	Numeric	$[0; 12]$
CONTACTS_COUNT_12_MON	No. of contacts in the last 12 months	Numeric	$[0; +\infty[$
CREDIT_LIMIT	Credit limit on the credit card	Numeric	$[0; +\infty[$
TOTAL_REVOLVING_BAL	Total revolving balance on the credit card	Numeric	$[0; +\infty[$
AVG_OPEN_TO_BUY	Open to buy credit line (average of last 12 months)	Numeric	$[0; +\infty[$
TOTAL_AMT_CHNG_Q4_Q1	Change in transaction amount (Q4 over Q1)	Numeric	$[0; +\infty[$
TOTAL_TRANS_AMT	Total transaction amount (last 12 months)	Numeric	$[0; +\infty[$
TOTAL_TRANS_CT	Total transaction count (last 12 months)	Numeric	$[0; +\infty[$
TOTAL_CT_CHNG_Q4_Q1	Change in transaction count (Q4 over Q1)	Numeric	$[0; +\infty[$
AVG_UTILIZATION_RATIO	Average card utilization ratio	Numeric	$[0; +\infty[$

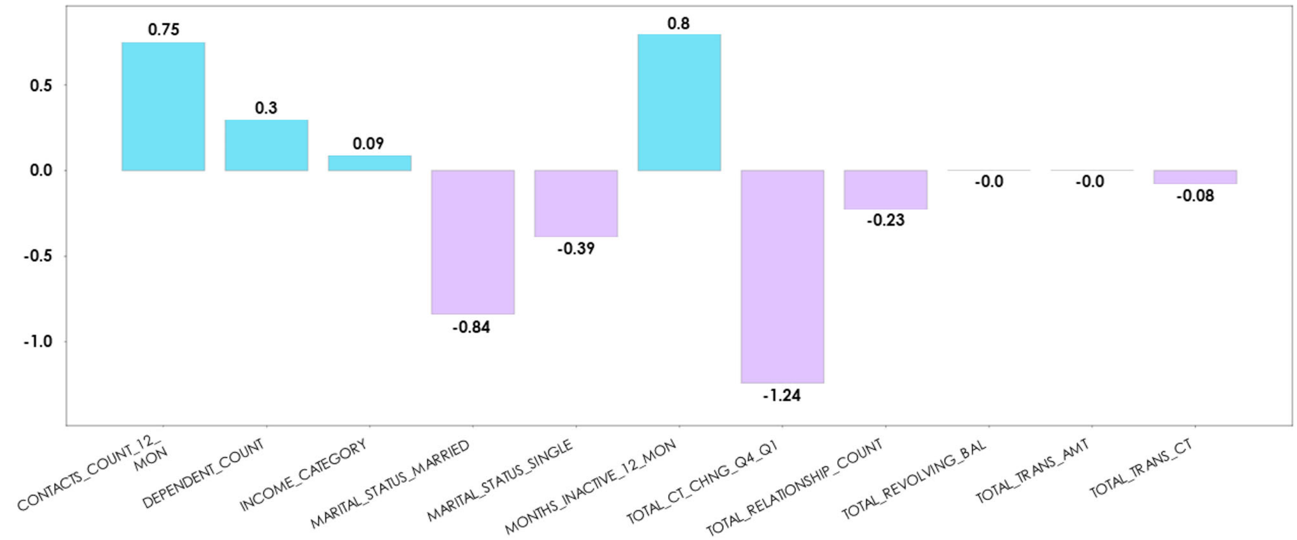

Fig. 4 Coefficients of the LR model for the churn scenario. The variables driving customer churn encompass the number of interactions with the institution, the period of inactivity, and the count of dependents. Conversely, features indicative of credit card retention include the activity in Q4 compared to Q1, marital status, and the number of overall relationships with the institution

making process, subsequently enabling an evaluation of client positioning to discern the factors underlying their ranking relative to others.

For instance, the employee may wish to compare clients ranked 6 (identifier 794560833) and 19 (identifier 719808558). This selection is entirely non-binding within the approach proposed in this study and merely serves narrative purposes. Hence, Fig. 5 is provided to the user to aid in understanding the relative positioning of the two clients, indicating for each the features that highlight a greater inclination for churn activity with respect to the other. In light of the coefficients' signs illustrated in Fig. 4, as expounded upon in "General Approach for Implementation and Application with Linear Model" section, it is important to remark that a NC implies that the likelihood of churn is heightened for the client exhibiting a lower value for the associated feature. Conversely, a positive coefficient indicates a greater predisposition towards churn for the client demonstrating a higher value. The approach suggests that client 794560833 is primarily favored by having a greater number of contacts in the last 12 months, a higher number of dependents, and a greater decrease in activity in the last quarter (feature with NC). On the other hand, client 719808558 receives a greater contribution from having a lower number of relationship with the institution (feature with NC), a lower total number of transactions (feature with NC), and from being inactive for a greater number of months in the last year.

The structure of the textual explanation associated to this use case should comply with the guidelines depicted in the following example:

The available information regarding Customer 794560833 and Customer 719808558 suggests that both clients may engage in churn activities. Customer 794560833 is ranked higher than Customer 719808558 according to the current algorithm reasoning. However, the ultimate decision remains within your control, offering the option to alter this ranking if desired. Characteristics in favor of Customer 794560833 include a higher level of CONTACTS_COUNT_12_MON and DEPENDENT_COUNT, along with a smaller value for TOTAL_CT_CHNG_Q4_Q1. Characteristics in favor of Customer 719808558 include a lower number of TOTAL_RELATIONSHIP_COUNT and TOTAL_TRANS_CT, along with a higher value of MONTHS_INACTIVE_12_MON.

Due to this explanatory capability, the employee gains enhanced insight into the algorithm's preference for one client over another, while also obtaining valuable contrasting information regarding the client with a lower final rank.

Discussion

As discussed throughout the manuscript, the primary objective behind our proposition of Evaluative Item-Contrastive Explanations is to support users in evaluating pairs of items, stimulating their critical judgment toward expressing a preference.

Table 5 Top 25 customers for the credit card churn use case. For each customer, the features that the LR model found significant are represented. It is also reported the probability of churn and the ranking assigned by the algorithm. Customers with shaded background are those chosen to elucidate the functionality of the proposed solution as discussed in "Constructing Evaluative Item-Contrastive Explanations for Recruitment" section

CLIENTNUM	RANK	SCORE	CONTACTS_ COUNT_ 12_MON	DEPENDENT_ COUNT	INCOME_ CATE- GORY	MARITAL_ STATUS_ MARRIED	MARITAL_ STATUS_ SINGLE	MONTHS_ INACTIVE_ 12_MON	TOTAL_ CT_CHNG_ Q4_Q1	TOTAL_ RELATION- SHIP_ COUNT	TOTAL_ REVOLV- ING_ BAL	TOTAL_ TRANS_ AMT	TOTAL_ TRANS_CT
815539233	1	0.99159	4	4	4.0	0	1	3	0.0	1	214	1201	22
809599158	2	0.97966	4	4	4.0	0	1	3	0.4	2	0	1626	28
827451333	3	0.97257	3	2	1.0	1	0	6	0.423	3	0	2615	37
713104458	4	0.96892	5	2	2.0	0	1	3	0.562	3	0	998	25
772695033	5	0.96789	4	5	4.0	0	1	4	0.211	3	2517	1575	23
794560833	**6**	**0.96325**	**6**	**3**	**2.0**	**1**	**0**	**3**	**0.308**	**6**	**0**	**1913**	**34**
711267633	7	0.96205	4	3	1.0	0	0	4	0.333	5	316	2271	36
709537683	8	0.95048	3	3	4.0	1	0	4	0.364	3	9226	563	15
720122808	9	0.9499	3	2	3.0	1	0	4	0.273	4	511	797	14
712509108	10	0.94869	5	3	1.0	0	1	3	0.478	1	532	8356	68
712578933	11	0.94694	6	5	2.0	1	0	2	0.517	3	0	2500	44
713865333	12	0.94436	3	4	1.0	1	0	4	0.75	3	0	827	21
714192933	13	0.94424	5	3	1.0	0	1	3	0.257	3	0	1976	44
807379833	14	0.94026	3	3	2.0	0	0	3	0.417	2	145	2563	34
720181758	15	0.93428	4	0	5.0	1	0	3	0.385	3	0	850	18
714581358	16	0.93008	4	3	1.0	0	1	3	0.44	3	0	2204	36
714915633	17	0.9218	6	0	3.0	0	1	3	0.6	4	0	1793	40
780613758	18	0.9196	5	5	1.0	0	1	3	0.82	4	0	17093	111
719808558	**19**	**0.91717**	**3**	**2**	**1.0**	**1**	**0**	**4**	**0.421**	**2**	**0**	**886**	**27**
716448783	20	0.90642	3	3	2.0	1	0	4	0.478	2	0	1225	34
804668658	21	0.90345	4	0	1.0	0	1	2	0.053	3	0	947	20
788794758	22	0.89704	4	4	2.0	0	1	2	0.462	2	0	2061	38
711205758	23	0.89257	3	3	4.0	0	1	3	0.433	1	0	2534	43
708193008	24	0.8914	4	4	4.0	0	1	3	0.654	3	0	1616	43
721074183	25	0.89031	3	1	3.0	1	0	6	1.0	4	1196	1111	24

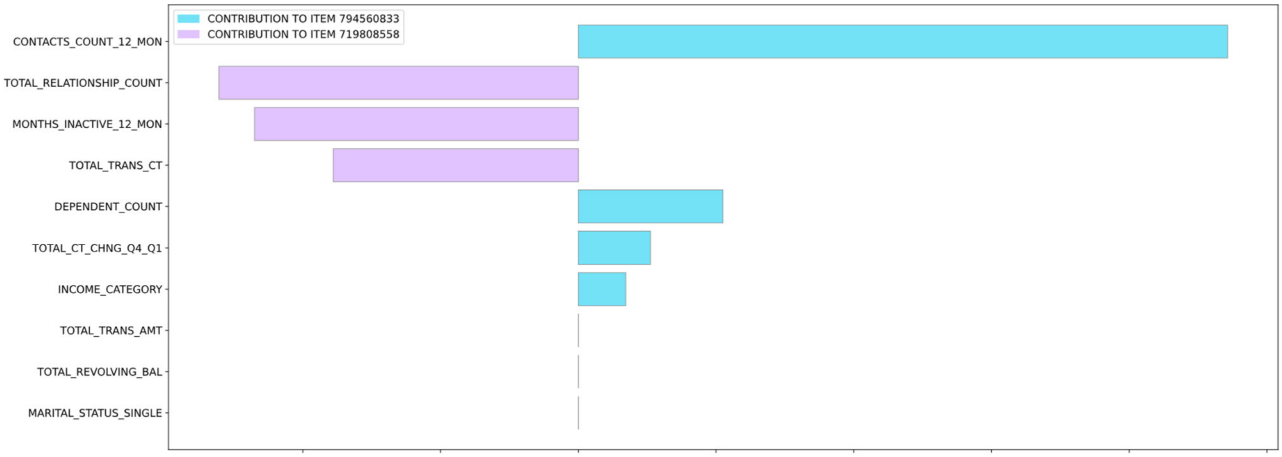

Fig. 5 Feature contributions to support the disparity in ranking among customers `794560833` and `719808558`. Client `794560833` is favored by a greater number of contacts in the last year, a higher number of dependents, and a greater decrease in activity in the last quarter. Client `719808558` receives a greater contribution from having a lower number of relationship with the institution, a lower total number of transactions, and from being inactive for more time

Additionally, in line with [19], it is essential to remember that evaluative AI, despite its paradigm shift, is still a form of explainable AI and, as such, must adhere to the standard reasons for generating explanations, as outlined in "Reasons for Explanations" section.

In this discussion, we argue that our approach remains valid also for addressing the goals of *justify*, *discover*, and *improve*. Furthermore, within the context of ranking systems, its application becomes particularly suitable in the domains of *control*.

The pros highlighted for each candidate serve as the *justification* of the logic endorsed by the algorithm. Furthermore, they may contribute to the *discovery* of new insights. However, attributes favoring specific items, not acknowledged by the user, could be instrumental in guiding system *improvements*. When users disagree with the provided justifications, their dissent prompts a request for system enhancements.

Finally, the entire formalization of the evaluative item-contrastive explanation is crafted to empower human *control* over the final decision-making process. To illustrate this, consider the following possible scenarios:

1. The user agrees with the outcome justification, and it is satisfied with the item's relative positions.
2. The user agrees with the outcome justification but is unsatisfied with the item's relative positions.
3. The user disagrees with the outcome justification but is satisfied with the item's relative positions.
4. The user disagrees with the outcome justification, and it is unsatisfied with the item's relative positions.

The first case represents a scenario in which the user is in agreement with both the proposed ranking and the provided justification for why it resulted in that way. In this situation, the user will confirm the existing ordering without suggesting any improvements to the system.

In the second scenario, the user encounters a situation where she concurs with the system's provided justification but opts to modify the order of items. This circumstance, although it may initially appear counter-intuitive, underscores the importance of the evaluative paradigm, which presents the advantages and disadvantages of contrasting items. It is a clear demonstration of the user's empowerment in control: agreement with the justification indicates contentment with the system's reasoning and decision-making process. However, this does not imply blind adherence; the user has access to contextual elements, allowing decisions that may differ from the system's recommendations. It may be the case that contextual elements influence the position of more than one item in the output of the ranking. Items that are near in position may be similar so it should not surprise. To ensure that the ranking is correctly evaluated by the user, we suggest contrasting couples of items recursively so that potential adjustments are operated.

The third and fourth scenarios, although resulting in different decisions from the user regarding the confirmation of the provided rank, are both marked by dissatisfaction with the justification. The user's reaction underscores their profound understanding of the organizational context and professional expertise. Regardless of confirmation, they recognize the need to improve the ranking algorithm, aligning its logic more closely with their well-informed judgment. Case three can easily be the product of chance, as two different sets of reasons may well generate the same ranking. In this case, an alignment is needed. In fact, a justification to other stakeholders is more easily provided by the users if they understand and share the AI system's reasons for a certain outcome. In case four, the system could appear to be faulted or useless, but it serves important epistemic functions. In our understanding, explanation should serve to improve the AI system. However, improvement cannot happen without the reflection of the user on the result of the system. To be empowered, the user exploits the output of the system as a source of knowledge and an object of reflection. So, also in case four, the systems shape the decision-making process helping the growth of the user. However, it is essential to discuss potential directions for algorithm improvement to ensure agreement among different users, as unexpected reasoning could lead to the discovery of novel knowledge.

Summarizing, considering the four potential scenarios users may encounter in ranking settings, we contend that an evaluative contrastive explanation is suitable for enhancing human oversight and control over the decision-making process.

Conclusion

In this work, we introduced and formalized the application of contrastive explanations as an effective methodology for explaining Machine Learning models for ranking. In particular, we want to stress that such an approach has the merit of highlighting to the decision-maker the key elements both supporting and contrasting a proposed rank, with the ultimate goal of putting her in the most appropriate position to make an informed decision. This, in turn, helps mitigate the impact of position bias in ranking problems.

In this respect, our approach is aligned with [19], calling for a paradigm shift from a *passive* decision-maker that can only take or reject the model's outcome to an ever more *active* decision-maker that can truly use the model as a support to extract information on the problem at hand.

By contrasting a pair of candidates in the proposed rank, the decision-maker can leverage granular information on what characteristics are pushing the score of one candidate above that of the other, but also what positive characteristics the lower-ranked candidate has, albeit not sufficient to be ranked higher—given the model.

Due to their granular character, contrastive explanations in rankings readily fit into the broad Granular Computing paradigm, increasing the impact and approachability of XAI and enabling explanations to be tailored to a wide variety of expertise and human cognitive processes.

We showcased our proposal with two experiments, with an emphasis on the characteristics that explanations should possess in order to be truly effective and informative for the user. From this viewpoint, the Evaluative Item-Contrastive approach has been assessed within the realms of recruitment and credit card churn. Both scenarios have utilized open-source datasets to emphasize the replicability of the attained outcomes. Our experiment exploits a simple linear model, thus directly using model weights as a means for extracting positive and negative contributions to the model outcomes. In our future research, we aim to extend the methodology to develop a post-hoc explanation mechanism that offers item-evaluative contrastive reasoning, independent of the black-box model that generates the rank.

Author Contributions Alessandro Castelnovo conceived the main ideas and supervised the entire work. Gabriele Nanino helped in structuring the paper and in shaping the main concepts. Nicoló Mombelli performed all the experiments and produced plots and tables. Riccardo Crupi helped with insights and comments. Daniele Regoli helped in shaping the structure and in the writing of the manuscript. All authors reviewed the manuscript.

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement. No funding was obtained for this study.

Code and Data Availability The data used in this work are available openly at https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement and https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn/. The python code used to perform experiments is openly available at https://github.com/nmombelli/EICER.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Sadok H, Sakka F, Maknouzi MEHE. Artificial intelligence and bank credit analysis: a review. Cogent Econ Financ. 2022;10(1):2023262. https://doi.org/10.1080/23322039.2021.2023262.

2. Asudeh A, Jagadish H, Stoyanovich J, Das G. Designing fair ranking schemes. In: Proceedings of the 2019 international conference on management of data; 2019. p. 1259–1276.
3. Viganò E. The right to be an exception to predictions: a moral defense of diversity in recommendation systems. Philos Technol. 2023;36(3):1–25. https://doi.org/10.1007/s13347-023-00659-y.
4. Zhang Q, Lu J, Jin Y. Artificial intelligence in recommender systems. Complex Intell Syst. 2021;7:439–57.
5. Anahideh H, Mohabbati-Kalejahi N. Local explanations of global rankings: insights for competitive rankings. IEEE Access. 2022;10:30676–93. https://doi.org/10.1109/ACCESS.2022.3159245.
6. Rahangdale A, Raut S. Machine learning methods for ranking. Int J Softw Eng Knowl Eng. 2019;29(06):729–61. https://doi.org/10.1142/S021819401930001X.
7. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
8. Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Gener Comput Syst. 2022;135:364–81.
9. Li N, Adepu S, Kang E, Garlan D. Explanations for human-on-the-loop: a probabilistic model checking approach. In: Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. SEAMS '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 181-187. Available from: https://doi.org/10.1145/3387939.3391592.
10. Nothwang WD, McCourt MJ, Robinson RM, Burden SA, Curtis JW. The human should be part of the control loop? In: 2016 Resilience Week (RWS); 2016. p. 214–220.
11. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.
12. Joachims T, Granka L, Pan B, Hembrooke H, Gay G. Accurately interpreting clickthrough data as implicit feedback. In: Acm Sigir Forum. vol. 51. Acm New York, NY, USA; 2017. p. 4–11.
13. Gupta A, Johnson E, Payan J, Roy AK, Kobren A, Panda S, et al. Online post-processing in rankings for fair utility maximization. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining; 2021. p. 454–462.
14. Zehlike M, Yang K, Stoyanovich J. Fairness in ranking, part i: score-based ranking. ACM Comput Surv. 2022;55(6):1–36.
15. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL Tech. 2017;31:841.
16. Dhurandhar A, Chen PY, Luss R, Tu CC, Ting P, Shanmugam K, et al. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf.
17. Alfeo AL, Cimino MG, Gagliardi G. Concept-wise granular computing for explainable artificial intelligence. Granul Comput. 2023;8(4):827–38.
18. Stepin I, Alonso JM, Catala A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access. 2021;9:11974–2001.
19. Miller T. Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; 2023. p. 333–342.
20. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif intell. 2019;267:1–38.
21. Bargiela A, Pedrycz W. Granular computing. In: Handbook on Computer Learning and Intelligence: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation. World Scientific; 2022. p. 97–132.
22. Roshan B.: Campus recruitment. Accessed: November 2023. https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement.
23. Credit card churn. Accessed: March 2024. https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn/.
24. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
25. Hilton DJ. Conversational processes and causal explanation. Psychol Bull. 1990;107(1):65.
26. Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller KR. How to explain individual classification decisions. J Mach Learn Res. 2010;11:1803–31.
27. Robnik-Šikonja M, Kononenko I. Explaining classifications for individual instances. IEEE Trans Knowl Data Eng. 2008;20(5):589–600.
28. Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong WK, Too much, too little, or just right? Ways explanations impact end users' mental models. In,. IEEE Symposium on visual languages and human centric computing. IEEE. 2013;2013:3–10.
29. Kulesza T, Burnett M, Wong WK, Stumpf S. Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th international conference on intelligent user interfaces; 2015. p. 126–137.
30. Papenmeier A, Englebienne G, Seifert C. How model accuracy and explanation fidelity influence user trust in AI. In: IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019; 2019.
31. Lipton P. Contrastive explanation. Royal Institute of Philosophy Supplement. 1990;27:247–66. https://doi.org/10.1017/s1358246100005130.
32. Van Bouwel J, Weber E. Remote causes, bad explanations? J Theory Soc Behav. 2002;32(4):437–49.
33. Weber E, van Bouwel J. The living apart together relationship of causation and explanation: a comment on Jean Lachapelle. Philos Soc Sci. 2002;32(4):560–9. https://doi.org/10.1177/004839302237837.
34. Malandri L, Mercorio F, Mezzanzanica M, Nobani N, Seveso A. ContrXT: generating contrastive explanations from any text classifier. Inf Fusion. 2022;81:103–15. https://doi.org/10.1016/j.inffus.2021.11.016.
35. Yao JT, Vasilakos AV, Pedrycz W. Granular computing: perspectives and challenges. IEEE Trans Cybern. 2013;43(6):1977–89. https://doi.org/10.1109/TSMCC.2012.2236648.
36. Hoffman RR, Miller T, Clancey WJ. Psychology and AI at a crossroads: how might complex systems explain themselves? Am J Psychol. 2022;135(4):365–78. https://doi.org/10.5406/19398298.135.4.01.
37. Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery. 2022; p. 1–55. https://doi.org/10.1007/s10618-022-00831-6.
38. Anahideh H, Mohabbati-Kalejahi N. Local explanations of global rankings: insights for competitive rankings. IEEE Access. 2022;10:30676–93.
39. Salimiparsa M. Counterfactual explanations for rankings. Proceedings of the Canadian Conference on Artificial Intelligence. 2023. https://caiac.pubpub.org/pub/9aov4tmt.

40. Tan J, Xu S, Ge Y, Li Y, Chen X, Zhang Y. Counterfactual explainable recommendation. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21. New York, NY, USA: Association for Computing Machinery; 2021; p. 1784-1793. Available from: https://doi.org/10.1145/3459637.3482420.
41. Singh A, Joachims T. Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD ICKDDM; 2018; p. 2219–2228.
42. Alimonda N, Castelnovo A, Crupi R, Mercorio F, Mezzanzanica M. Preserving utility in fair top-k ranking with intersectional bias. In: International Workshop on Algorithmic Bias in Search and Recommendation. Springer; 2023; p. 59–73.
43. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics. 2019;8(8). https://doi.org/10.3390/electronics8080832.
44. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2020;10(5):e1379.
45. Castelnovo A, Cosentini A, Malandri L, Mercorio F, Mezzanzanica M. FFTree: a flexible tree to handle multiple fairness criteria. Inform Process Manag. 2022;59(6):103099. https://doi.org/10.1016/j.ipm.2022.103099.
46. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv (CSUR). 2018;51(5):1–42.
47. Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. Inform Process Manag. 2023;60(1):103111. https://doi.org/10.1016/j.ipm.2022.103111.
48. Passi S, Vorvoreanu M. Overreliance on AI literature review. Microsoft Research. 2022.