Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 13
AMMSAC 2026

A Fair Ranking Mechanism for Multi-Source Evaluation Based on Random Forest and RMF-Ranking

Zitong Zhou *, Zhenyu Yang and Tianrui Zhou

Chang'an University, Xi'an, China

* Corresponding Author Email: 17868488217@163.com

Abstract. This paper investigates the fairness and stability of competition rankings in multi-source evaluation scenarios. We construct a comprehensive framework that combines audience voting reconstruction, counterfactual simulation, and analysis using random forests and SHAP explanations, and propose the RMF-Ranking fairness mechanism. This method restores unobservable voting distributions through macro- and micro-level constraints, compares the impact of different scoring mechanisms on ranking results, and further identifies structural differences between judge evaluations and audience preferences. Building on this foundation, RMF-Ranking transforms diverse evaluation results into a ranking aggregation problem and employs a Markov approximation to derive a consensus ranking, effectively mitigating the excessive dominance of any single evaluation dimension. Research indicates that this mechanism reduces extreme ranking fluctuations, enhances result robustness, and is suitable for fair evaluation and ranking decision-making scenarios involving multiple stakeholders and multiple metrics.

Keywords: RMF-Ranking Algorithm; Random Forest; SHAP Interpretability Analysis.

1. Introduction

In evaluation and ranking scenarios involving multiple stakeholders, striking a balance between expert judgment, public preferences, and result stability is a critical issue in mechanism design. Although traditional weighted or ranking aggregation methods are easy to implement, they often struggle to handle situations where evaluation sources vary significantly, voting data is incomplete, or extreme preferences are amplified, leading to final results that are overly sensitive to a single dimension. Existing research has primarily focused on rule comparisons or result interpretation, lacking a comprehensive algorithmic framework that spans from latent data recovery and mechanism simulation to the generation of fair rankings.

To address these shortcomings, this paper constructs an algorithmic model framework for multi-source evaluation systems. First, we use reverse reconstruction methods to recover the structure of audience votes that cannot be directly observed. Second, we employ counterfactual simulation to compare ranking changes under different scoring rules. Subsequently, we combine Random Forests with SHAP analysis to identify the sources of discrepancies between judge evaluations and audience preferences [1]. Finally, we propose the RMF-Ranking fairness mechanism, which transforms diverse evaluation results into a ranking aggregation problem and employs a Markov approximation to derive a more robust consensus ranking. By applying this framework to the ranking mechanisms of competition-style television programs, we validate its effectiveness in mitigating extreme popularity effects, reducing ranking volatility, and enhancing the fairness of results, thereby providing a transferable methodological reference for multi-criteria evaluation and public decision-making ranking problems.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

2. Audience Voting Reconstruction Model Based on Inverse Constraints

2.1. Model Framework

Audience voting data are not directly observable and are therefore treated as a latent variable. The votes are reconstructed indirectly using judges' scores, elimination outcomes, and audience viewership data. The model is designed as a two-layer inverse reconstruction process. The first layer determines the feasible range of total votes for each episode, while the second layer infers the audience votes received by individual contestants within that range.

At the macro level, the model is used to establish the bounds of total voting volume. Since historical viewership data are incomplete for some seasons, viewership trends from more recent seasons are transferred and combined with audience-size anchors from the premiere and finale episodes to provide a reasonable search space for vote reconstruction. At the micro level, individual vote allocation is inferred. Given judges' scores and elimination outcomes, audience votes for each contestant are reconstructed. Reverse-game constraints, a Zipf-distribution prior, temporal smoothing, and minimum-gap constraints are introduced to ensure that the reconstructed results are both rule-consistent and consistent with the long-tail characteristics commonly observed in audience voting behavior.

2.2. Data Processing and Scoring Mechanisms

Data were compiled from audience figures for premiere and finale episodes across different seasons, partial episode-level viewership records from more recent seasons, stage-specific scoring rules, and publicly reported total voting statistics. Based on this information, the total number of votes in a given episode can be estimated as:

$$V_{\text{est}} = E \times k \times r \tag{1}$$

where E denotes the estimated audience size for the episode, k is the theoretical maximum number of votes per viewer (set to 20 in this study), and r is the active voting participation rate. For regular episodes, r is assumed to range from 15% to 25%, while for finale episodes it ranges from 30% to 40%. This estimate is used primarily to define a reasonable physical boundary rather than to recover the exact number of votes cast.

Two main scoring mechanisms are considered:

1. Rank-Based System: Judges' rankings and audience rankings are converted into ordinal positions and then combined:

$$S_{\text{total}} = \text{Rank}(S_{\text{judge}}) + \text{Rank}(V_{\text{fan}}) \tag{2}$$

Contestants with poorer combined rankings face a higher risk of elimination. This mechanism tends to compress extreme differences in vote counts.

2. Percentage-Based System: Judges' score shares and audience vote shares are combined using weighted proportions:

$$S_{\text{total}} = \frac{S_{\text{judge}}}{\sum S_{\text{judge}}} + \frac{V_{\text{fan}}}{\sum V_{\text{fan}}} \tag{3}$$

Contestants with lower total scores enter the elimination zone. This mechanism preserves differences in vote magnitude and is more likely to amplify the advantage of highly popular contestants.

2.3. Construction of Macro-Level Voting Boundaries

Since episode-level viewership data are incomplete for historical seasons, the total number of votes in a single episode cannot be directly obtained from the original data. This study first extracts the

basic pattern of how audience participation changes over the course of more recent seasons, and then transfers this pattern to historical seasons that only have viewership anchors for the premiere and finale episodes. Let $t \in [0,1]$ denote the normalized progress of the season. A third-order polynomial is used to fit the changing trend in viewership:

$$S(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \text{ò} \tag{4}$$

where β_i denotes the regression coefficient, and $S(t)$ represents the nonlinear pattern of viewership changes within a season. The focus here is not the absolute popularity of different seasons, but the way audience attention changes as the season progresses.

For historical seasons, relying only on linear interpolation between premiere and finale viewership would ignore fluctuations in the middle stages. Therefore, a linear correction term $\delta^{(k)}(t) = A_k + B_k t$ is introduced, and the learned pattern $S(t)$ is mapped to the k-th historical season:

$$\hat{V}^{(k)}(t) = V_{\text{base}}^{(k)}(t) \cdot S(t) + \delta^{(k)}(t) \tag{5}$$

where A_k and B_k are determined by the boundary conditions $\hat{V}^{(k)}(0) = V_{\text{start}}^{(k)}$ and $\hat{V}^{(k)}(1) = V_{\text{end}}^{(k)}$. The resulting viewership curve passes through the known start and end anchors while also retaining possible nonlinear fluctuations during the season.

On this basis, the feasible interval for the total number of votes in each episode is constructed by combining the active voting ratio $\alpha(t)$ and the per-person voting cap C_{cap}:

$$N^{(k)}(t) \in [N_{\min}(t), N_{\max}(t)], \quad N_{\text{bound}}(t) = \hat{V}^{(k)}(t) \times \alpha(t) \times C_{\text{cap}} \tag{6}$$

where $C_{\text{cap}} = 20$. The active voting ratio $\alpha(t)$ is set to [0.15,0.25] for regular episodes and [0.30,0.40] for finale episodes. The resulting interval $[N_{\min}, N_{\max}]$ is used to restrict the search space for individual vote reconstruction.

2.4. Micro-Level Voting Reconstruction

After the macro-level boundary is determined, individual vote counts are inferred based on the elimination logic. If a safe contestant i has a lower judges' score J_i than an eliminated contestant j with judges' score J_j, then contestant i must have received more audience votes to offset the judges' score gap. This constraint is expressed as:

$$F_i \geq F_j + \kappa \cdot (J_j - J_i) + \text{ò} \tag{7}$$

where F_i and F_j denote the audience votes received by the two contestants, κ represents the conversion coefficient between the judges' score gap and audience votes, and ò is the minimum safety margin. The algorithm goes through all “safe contestant–eliminated contestant” pairs. When the votes of a safe contestant are not sufficient to explain the advancement result, the contestant's votes are raised above the lower bound implied by the constraint.

Relying only on elimination constraints can easily produce an overly even vote distribution, while reality competition shows usually have a clear head effect. A small number of highly popular contestants receive a large amount of support, while most contestants remain in the long tail. Based on this feature, a Zipf-distribution prior is introduced at the initialization stage:

$$w_r \propto \frac{1}{r^{\gamma}}, \quad \gamma = 2.5 \tag{8}$$

where w_r denotes the voting weight of the contestant ranked r, and γ controls the speed of weight decay. This prior does not directly determine the final vote counts, but provides the search process with an initial structure that is closer to real voting behavior.

Considering that fan support has some continuity, vote counts in adjacent weeks should not show large unexplained jumps. An EWMA smoothing mechanism is used to handle abnormal weekly fluctuations:

$$F_t = \lambda \cdot F_{\text{raw}} + (1-\lambda) \cdot F_{t-1} \tag{9}$$

where F_{raw} is the raw estimate obtained from the constraints in the current week, F_{t-1} is the smoothed result from the previous week, and $\lambda = 0.5$. At the same time, a minimum gap threshold δ is set to prevent key rankings from being reversed by very small disturbances.

The final reconstruction process uses an iterative rebalancing method. First, an initial voting vector $F^{(0)}$ is generated based on the Zipf distribution and the total vote count N_{total}. Then, the vote counts of contestants who do not satisfy the reverse elimination constraints are adjusted. Since the total number of votes must be conserved, the added votes are deducted proportionally from non-critical contestants. After the adjusted voting vector is processed with EWMA smoothing, the final estimated vote count $\hat{F}$ is obtained.

2.5. Stability Tests of the Reconstruction Results

After the voting reconstruction process is completed, estimated audience vote counts for contestants in each episode can be obtained. Since the true audience votes are unobservable, simply presenting the reconstructed vote trajectories is not sufficient to demonstrate reliability. Therefore, the reconstructed results are evaluated from two perspectives: ranking stability and solution-space determinism.

Ranking stability is measured using Kendall's W coefficient of concordance. If the model has good robustness to noise, small random perturbations introduced within the macro-level feasible interval $[N_{\min}, N_{\max}]$ should not frequently reverse contestants' relative rankings. Monte Carlo sampling is performed within the feasible voting interval of each contestant to generate multiple potential vote distributions, and the ranking consistency coefficient is then calculated: $W \in [0,1]$.

A value of W closer to 1 indicates greater ranking stability, while a lower value suggests that the estimated ranking is more sensitive to random perturbations.

Solution-space determinism is measured by the indicator C_i:

$$C_i = 1 - \frac{v_{i,\max} - v_{i,\min}}{N_{\text{total}}} \tag{10}$$

where $v_{i,\min}$ and $v_{i,\max}$ denote the lower and upper feasible voting bounds for contestant i while keeping the elimination outcome unchanged. A value of C_i closer to 1 indicates a narrower feasible vote interval and therefore a more certain reconstruction result. Conversely, a lower value suggests that the estimated vote count still has substantial room for variation.

The test results are shown in Fig. 1. Seasons using the Rank-Based System achieve an average Kendall's W of 0.9391, which is substantially higher than the 0.5933 observed for the Percentage-Based System. Because the rank-based mechanism converts continuous vote differences into discrete ordinal positions, the impact of small perturbations on the final ranking is reduced, resulting in more stable reconstruction outcomes.

The Percentage-Based System exhibits lower ranking stability, but its overall determinism remains relatively high, with an average C_i of 0.9290. This suggests that although local rankings are more sensitive to vote fluctuations under the percentage-based mechanism, the reverse constraints still restrict feasible vote counts to a relatively narrow range. Taken together, these tests indicate that the reconstructed audience voting results are sufficiently reliable for subsequent analysis of voting mechanisms and fairness.

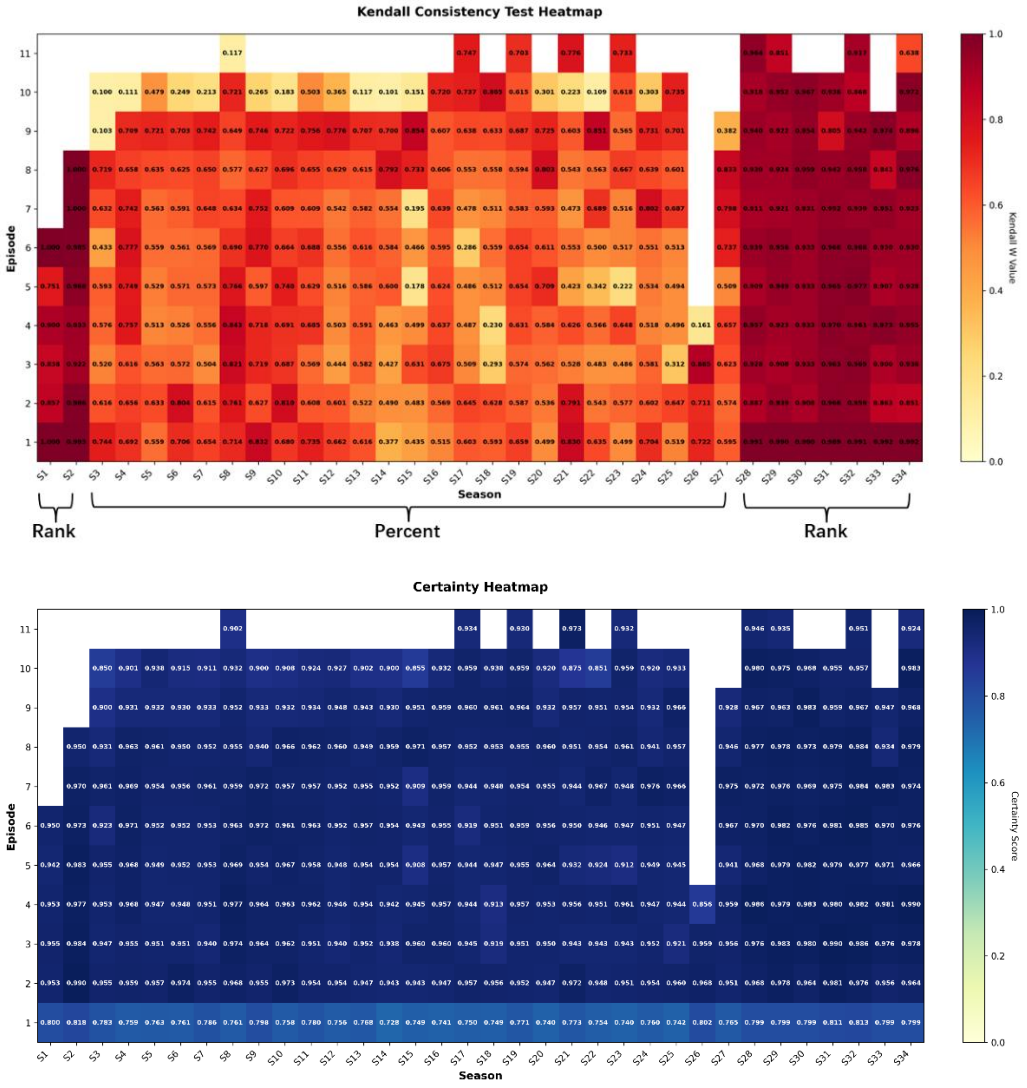

Fig. 1 Consistency and determinism analysis

3. Fairness Analysis of Scoring Mechanisms Based on Counterfactual Simulation

3.1. Counterfactual Simulation Framework

Based on the reconstructed audience voting results, this study further analyzes how different scoring mechanisms affect final rankings [2]. Specifically, with judges' scores and audience votes held unchanged, the Rank-Based System and the Percentage-Based System are separately applied to recalculate season rankings, and the resulting ranking changes are then compared.

Because voting scales differ greatly across seasons, directly using raw vote counts may introduce bias. Therefore, a dimensionless voting indicator is constructed:

$$V_{\text{index}} = \frac{x_i - \bar{x}}{\bar{x}} \tag{11}$$

where x_i denotes the audience votes received by contestant i, and $\bar{x}$ denotes the average vote count in the same episode. This indicator is used to measure a contestant's audience support relative to the average level. Judges' evaluations are represented by the average score:

$$S_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{12}$$

where s_i is the original judges' score, and N is the number of judges in that round. For contestants whose rankings improve under the alternative mechanism but who lack records in later rounds, their historical average performance is used for completion, so that the two mechanisms remain comparable.

3.2. Structural Differences Between Scoring Mechanisms

To identify the main sources of differences between scoring mechanisms, contestants whose ranking changes exceed 3 places are selected for K-Means clustering analysis [3]. The results are shown in Fig. 2.

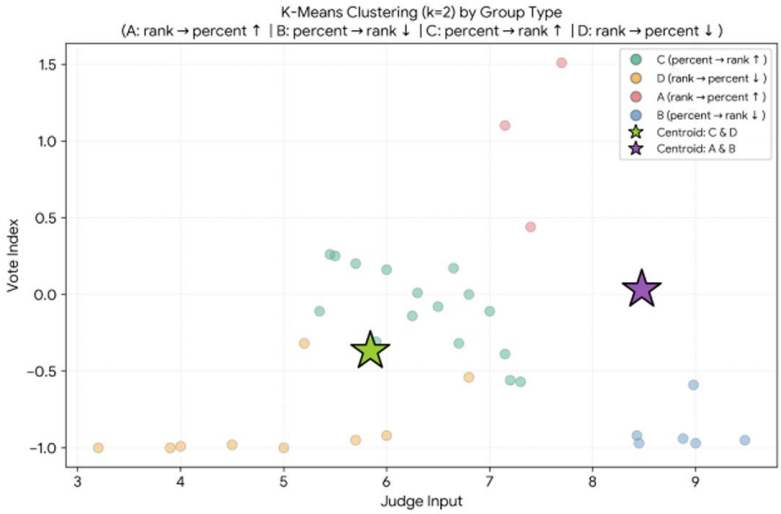

Fig. 2 K-means result diagram

The clustering results show that contestants can be roughly divided into three groups. The first group consists of "high-score, high-vote" contestants. They receive both strong judges' evaluations and high audience support, and therefore remain competitive under both mechanisms. Since the Percentage-Based System preserves differences in vote magnitude, the advantage of these contestants is further amplified.

The second group consists of "high-score, low-vote" contestants. These contestants perform well technically but have a relatively weaker audience base. Under the Percentage-Based System, they are more likely to be squeezed out by highly popular contestants. After votes are converted into discrete rankings under the Rank-Based System, the effect of extreme vote gaps is weakened, which offers some protection for technically strong contestants.

The third group consists of "high-vote, low-score" contestants. They mainly rely on audience support to maintain their ranking and are more sensitive to the scoring mechanism. The Percentage-Based System tends to amplify their popularity advantage, while the Rank-Based System limits the further expansion of vote-based advantages.

This shows that the two mechanisms reflect different evaluation orientations. The rank-based mechanism places greater emphasis on technical performance and suppresses popularity inflation through discretization, while the percentage-based mechanism gives more weight to audience preference and makes it easier for highly popular contestants to gain a competitive advantage.

3.3. Analysis of Historical Controversial Cases

To test whether the counterfactual simulation can explain controversies observed in actual competitions, several representative historical cases are selected for comparison. The results are shown in Table 1.

Table 1. Analysis of historical controversial cases

Contestant	Feature	R	P	Change
Rice	Low scores	2	1	↑1
Palin	Lowest judge	3	1	↑2
Cyrus	Long-term last place	5	11	↓6
Bones	Low scores	1	4	↓4
Kardashian	Low score, high votes	11	7	↑4
Castroneves	High Score Low Votes	6	11	↓5

Most controversial cases are related to excessive weight being placed on audience voting. Some contestants whose technical performance was not especially strong achieved higher rankings, or even became possible title contenders, under the Percentage-Based System because of strong audience support. When the Rank-Based System is applied, their vote advantage is discretized, and their rankings decline noticeably. This suggests that the rank-based mechanism can, to some extent, reduce the advantage gained purely from popularity.

Some contestants with lower popularity but higher judges' scores also benefit from the Rank-Based System. Although their audience support is limited, their higher technical scores allow them to remain competitive. By contrast, under the Percentage-Based System, this type of contestant is more likely to be eliminated quickly.

Combining Fig. 2 and Table 1, the two scoring mechanisms do not have an absolute advantage over each other. Instead, they represent different trade-offs between technical evaluation and public preference. From the perspective of fairness, the Rank-Based System is better at limiting extreme popularity effects and is more helpful for preserving the role of technical evaluation in final rankings.

3.4. Boundary-Elimination Correction Mechanism

The Rank-Based System can weaken extreme vote gaps, but it is still a linear aggregation rule. When judges' scores and audience votes are in clear conflict, simply adding rankings may not properly handle abnormal cases near the elimination boundary. Therefore, a boundary-elimination correction mechanism is introduced on top of the rank-based rule as a supplement to the linear mechanism.

This mechanism does not directly eliminate the contestant with the lowest combined ranking. Instead, it first places contestants in the bottom range into a pending status and then introduces expert adjudication. On the one hand, this can protect “high-score, low-vote” contestants and prevent technically stronger contestants from being eliminated too early because of a weaker audience base. On the other hand, it also sets a minimum technical threshold for advancement, so that the final outcome is not determined entirely by audience voting.

To compare the effects of different mechanisms, a representative controversial season is selected for counterfactual simulation. The results are shown in Table 2.

Table 2. Comparison of different ranking methods for Season 27

Percent	Rank	LDS
Bobby Bones	Evanna Lynch	Evanna Lynch
Milo Manheim	Milo Manheim	Milo Manheim
Evanna Lynch	Alexis Ren	Alexis Ren
Alexis Ren	Bobby Bones	Juan Pablo
Juan Pablo	Juan Pablo	Bobby Bones
Joe Amabile	John Schneider	John Schneider
DeMarcus Ware	DeMarcus Ware	DeMarcus Ware
John Schneider	Mary Lou Retton	Mary Lou Retton

Table 2 shows that the Percentage-Based System is more sensitive to skewed vote distributions. When a contestant receives substantially more audience support than others, the continuous proportional calculation fully preserves this voting advantage, which may cover up the contestant's weaker judges' score. The Rank-Based System produces a different result. It converts continuous votes into discrete ordinal rankings, compressing the marginal advantage of contestants with extremely high votes. Even if a contestant leads by a large margin in audience voting, this advantage only translates into the corresponding ranking score, and the vote gap is not amplified without limit.

The boundary-elimination correction mechanism further plays a secondary correction role. It does not change the overall ranking of all contestants, but focuses on abnormal cases near the elimination boundary. When a low-vote, high-score contestant falls into the danger zone, expert adjudication can offer a corrective channel. When a high-vote, low-score contestant enters the safe zone mainly because of popularity, the mechanism can also impose some restriction.

Overall, combining the Rank-Based System with the boundary-elimination correction mechanism is more effective than the Percentage-Based System alone in limiting the influence of extreme audience preferences, and it is also more suitable than the pure rank-based rule for handling elimination-boundary cases. This suggests that a fair mechanism should not rely only on a single linear weighted rule. Instead, a correction step for boundary cases should be added beyond ranking aggregation.

4. Preference Difference Analysis Based on Random Forest and SHAP

The counterfactual simulation results show clear differences between judges' evaluations and audience voting. To further explain the sources of these differences, this study uses a random forest regression model together with SHAP analysis to examine how different features affect judges' scores and audience votes.

4.1. Random Forest and SHAP Analysis Framework

Audience voting and judges' scores are affected by many factors, including contestants' performances, age, and professional background. The relationships among these variables may be nonlinear and may also involve interaction effects. Compared with traditional linear models, random forest can capture more complex mapping relationships through ensemble learning with multiple decision trees, and it also has relatively good robustness [4].

To improve model interpretability, a SHAP analysis framework is introduced to quantify the contribution of each variable and identify the key factors affecting audience voting and judges' scores. This makes it possible to compare differences in preference structure between the two evaluation systems.

4.2. Feature Construction

During data preprocessing, invalid records caused by elimination or missing values are removed, and only data from the periods when contestants actually participated in the competition are retained. The model mainly uses three types of features.

The first type is the average performance score, which is used to measure contestants' technical level:

$$S = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{13}$$

where S_i denotes the valid judges' score received in week i.

The second type is age-related features. Considering the relatively wide age range, Z-score standardization is applied:

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

where x is the contestant's age, and μ and σ denote the sample mean and standard deviation, respectively.

The third type is professional background and identity features. Occupational categories are encoded using One-Hot Encoding, including major categories such as actors, athletes, and singers. Identity variables are used to examine the possible influence of regional factors.

4.3. Results Analysis

Two random forest models are built separately: the audience model uses the reconstructed audience votes as the prediction target, while the judges' model uses the average judges' score as the prediction target. The model evaluation results are shown in Table 3.

Table 3. Model evaluation

Model Type	R^2	MAE	MAPE
Judge Model	0.9369	0.2484 (points)	3.71%
Audience Model	0.9724	0.3691 (million votes)	19.54%

Table 3 shows that both models achieve high fitting accuracy. The R^2 of the judges' model is 0.9369, while the R^2 of the audience model is 0.9724, indicating that the selected features can explain scoring and voting outcomes well.

The variable importance results are shown in Fig. 3.

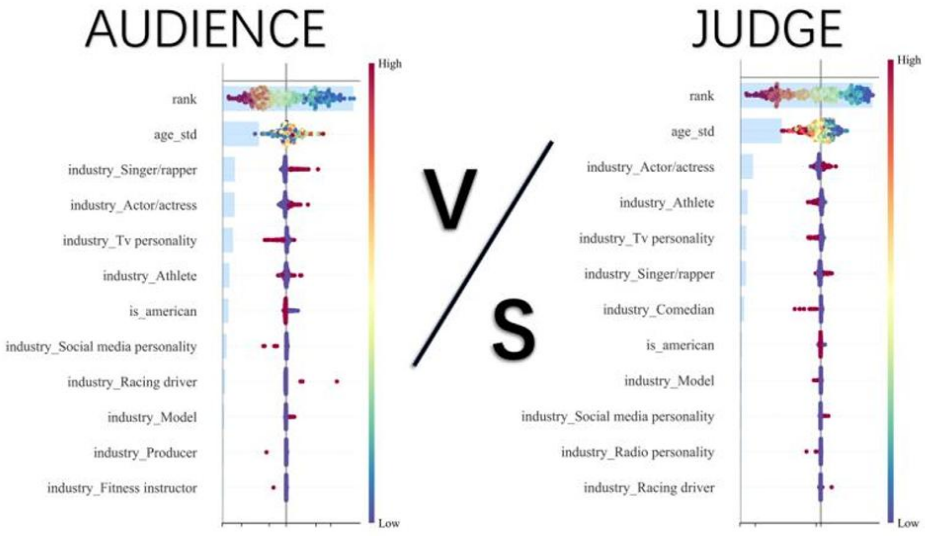

Fig. 3 Importance analysis chart

Fig. 3 shows that ranking and age are important factors in both models, suggesting that contestants' overall performance remains an important foundation for competition outcomes. However, the two evaluation systems do not show fully consistent preferences regarding professional background.

The judges' model gives more weight to groups such as actors and athletes, who tend to have advantages in body control. These contestants are usually more stable in movement completion, stage performance, and technical details, so they are more likely to receive higher scores. The audience model assigns higher importance to singers and other performance-oriented professions, suggesting that audience voting is influenced not only by technical level, but also by public recognition and fan base.

The importance of identity variables remains low, indicating that regional factors have limited influence on the final results. Audience voting and judges' scores are driven more by individual performance and professional characteristics than by external identity attributes.

This result further shows that judges and audiences do not evaluate contestants using the same set of standards. Judges focus more on technical ability, while audiences are more easily influenced by public image and fan effects. The structural difference between these two types of preferences also explains why ranking disagreements arise under different scoring mechanisms.

5. RMF-Ranking Fair Ranking Mechanism

5.1. Design Motivation

The counterfactual simulation and feature analysis show a clear tension between professional evaluation and public preference in existing scoring mechanisms. The Percentage-Based System can preserve continuous differences in audience votes, but it can easily amplify the influence of extreme popularity. The Rank-Based System weakens vote inflation through discretization, but it also loses part of the information contained in score differences. Relying on only one scoring logic makes it difficult to balance technical performance, audience participation, and result stability at the same time.

Based on this issue, this study proposes the RMF-Ranking mechanism. This mechanism does not directly apply linear weighting to judges' scores and audience votes. Instead, it converts the results generated by different evaluation systems into a ranking aggregation problem and searches for a final ranking that is as close as possible to the consensus across multiple evaluation sources, thereby reducing the excessive dominance of any single evaluation dimension [5].

5.2. Ranking Aggregation Framework

Suppose there are n contestants in a given round: $C = c_1, c_2, \ldots, c_n$.

There are also k input ranking lists: $R = \tau_1, \tau_2, \ldots, \tau_k$.

In this study, the input rankings mainly come from the results calculated under the Rank-Based System and the Percentage-Based System. To measure differences between rankings, Kemeny Distance is introduced. For any two rankings τ_a and τ_b, it is defined as:

$$K(\tau_a, \tau_b) = |(c_u, c_v) : \tau_a(c_u) < \tau_a(c_v), \tau_b(c_u) > \tau_b(c_v)| \tag{15}$$

This distance represents the number of inverted pairs between two rankings. The optimization objective for the final ranking σ is:

$$\sigma^* = \arg\min_{\sigma \in P_n} \sum_{i=1}^{k} K(\sigma, \tau_i) \tag{16}$$

where P_n denotes the set of all possible rankings. This objective does not directly select the result of a single mechanism. Instead, it looks for a consensus ranking with the smallest overall distance across multiple evaluation sources.

5.3. Markov Approximation Solution

Since directly solving the ranking aggregation problem is computationally costly, this study uses the MC4 method for approximate solution. First, the state transition matrix P is constructed. For any two contestants c_i and c_j, the number of input rankings in which c_j is ranked above c_i is counted:

$$W_{ji} = \sum_{m=1}^{k} I\left(\text{rank}(c_j) < \text{rank}(c_i) \text{ in } \tau_m\right) \tag{17}$$

If most rankings consider c_j to be better than c_i, then:

$$P_{ij} = \frac{1}{n}, \quad \text{if } W_{ji} > \frac{k}{2} \tag{18}$$

Otherwise, $P_{ij} = 0$. To make sure that the probabilities in each row sum to 1, the diagonal element is defined as:

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij} \tag{19}$$

The stationary distribution of the Markov chain is then solved. Let the stationary vector be: $\pi = [\pi_1, \pi_2, \ldots, \pi_n]$. It satisfies $\pi P = \pi$, $\sum_{i=1}^{n} \pi_i = 1, \quad \pi_i \geq 0$.

Power Iteration is used for the iterative solution: $\pi^{(t+1)} = \pi^{(t)} P$.

The iteration stops when $|\pi^{(t+1)} - \pi^{(t)}| < 10^{-6}$. Finally, π_i represents the relative advantage of contestant c_i in the aggregated ranking. The final ranking is obtained by sorting π_i from large to small:

$$\sigma(i) < \sigma(j) \Leftrightarrow \pi_i > \pi_j \tag{20}$$

In this way, the complex ranking aggregation problem is converted into a Markov chain stationary distribution problem, reducing computational cost while retaining useful information from different evaluation systems.

5.4. Performance Evaluation of RMF-Ranking

To test the ability of RMF-Ranking to coordinate disagreement among multiple evaluation sources, this study compares the distribution of ranking changes under different mechanisms, with particular attention to significant ranking fluctuations: $|\Delta \text{Rank}| \geq 5$.

This indicator is used to measure the extent to which changes in scoring rules affect final rankings. If many contestants show significant ranking changes, the results are sensitive to the mechanism design, and the system is not stable enough.

The statistical results are shown in Fig. 4.

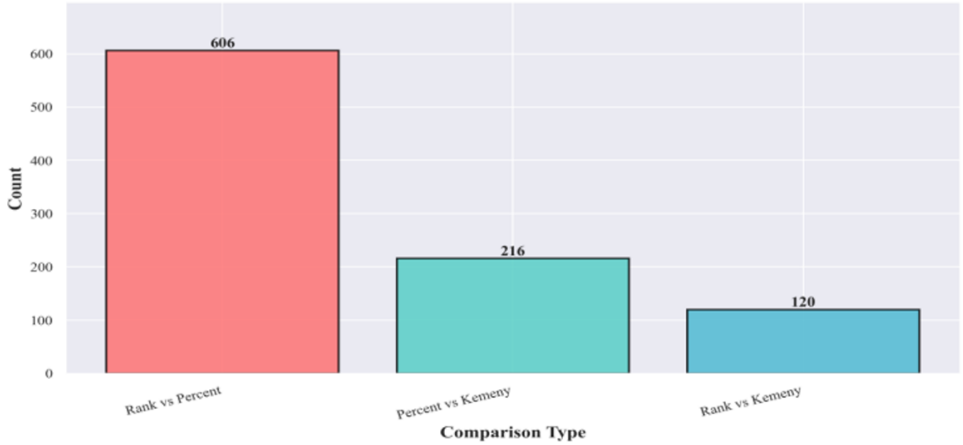

Fig. 4 Distribution of comparison types

Fig. 4 shows clear differences between the original Rank-Based System and the Percentage-Based System, with some contestants experiencing large ranking changes under the two mechanisms. After RMF-Ranking is introduced, the number of significant ranking differences decreases, and large ranking fluctuations are reduced by about 44.6%. This suggests that RMF-Ranking can retain

information from multiple evaluation sources while reducing conflict between mechanisms, making the final ranking closer to an overall consensus [6].

The distribution of ranking change magnitudes is shown in Fig. 5.

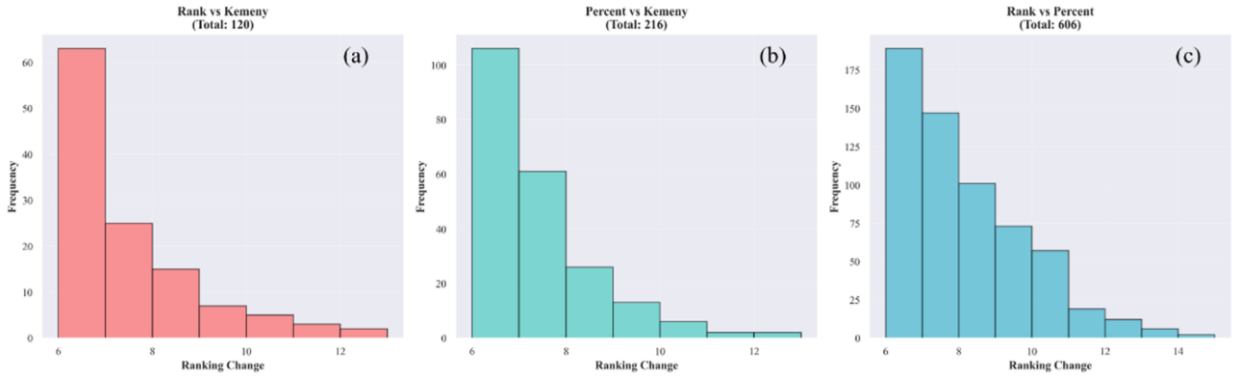

Fig. 5 Distribution histogram of ranking changes

Under the original mechanisms, ranking changes show a clear long-tail pattern, with some contestants experiencing ranking shifts of more than 12 places. Such extreme changes usually indicate that the final result is overly influenced by one evaluation dimension. In contrast, ranking changes under RMF-Ranking are more concentrated, extreme fluctuations are clearly reduced, and most adjustments are kept within a smaller range.

Taken together, Fig. 4 and Fig. 5 show that RMF-Ranking can reduce structural disagreement between different evaluation systems and improve the stability of the final ranking. For multi-source evaluation settings, this mechanism provides a relatively robust fair ranking solution.

6. Conclusion

This paper proposes a fair ranking framework for multi-source evaluation scenarios, encompassing vote reconstruction, counterfactual simulation, preference interpretation, and ranking aggregation. The study indicates that proportional mechanisms tend to amplify popularity advantages, whereas ranking mechanisms can mitigate the influence of extreme preferences. In stability tests, the ranking mechanism achieved an average Kendall's W of 0.9391; RMF-Ranking further integrates multi-source evaluation information, reducing significant ranking fluctuations by approximately 44.6% and enhancing the robustness and fairness of the results. This framework can be extended to scenarios such as social science evaluation, expert review, and public participation in decision-making, providing an algorithmic reference for balancing professional judgment and social preferences. Future research could further incorporate dynamic behavioral data to optimize adaptive fair ranking models.

References

[1] Guo Kewei, Ren Guojun, Jia Jiyuan. Assessment of the Resilience of the Arable Land System in Yunxi County and Analysis of Its Driving Factors Based on Random Forests and SHAP [J]. Central South Agricultural Science and Technology, 2025, 46(S1): 145–150.

[2] Yin Zhichao, Liu Jiayi, Li Feng. The Impact of the New Individual Income Tax Policy on Household Consumption: A Counterfactual Analysis Based on a Nonlinear Consumption Function [J]. Economic Research Journal, 2026, 61(2): 213-236.

[3] Ren Haoyuan, Jin Jing, Wang Yanyan, et al. Data Mining of Temperature and Humidity Data in Coastal Areas Based on the K-Means Clustering Algorithm and Bayesian Network Model [J]. Aerospace Environmental Engineering, 2025, 42(5): 494-503.

[4] Tan Zhicheng. A Study on the Determinants of Commercial Housing Prices in Qingyuan City Based on the Random Forest Model [D]. Guangdong University of Technology, 2025. DOI: 10.27029/d.cnki.ggdgu.2025.002458.

[5] Dong Hao, Wang Hefei, Lei Jiaqi. A Markov Chain-Based Passenger Flow Allocation Model for Tourism Rail Transit [J]. Research on Urban Rail Transit, 2022, 25(9): 38-44. DOI:10.16037/j.1007-869x.2022.09.008.

[6] Xiao Fuxian. Construction and Stability Analysis of the Chinese Stock Market Correlation Network from an Evolutionary Network Perspective [D]. Wuhan Textile University, 2023. DOI:10.27698/d.cnki.gwhxj.2023.000758.