ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on Voting Reconstruction and Evaluation Optimization Based on XGBoost-SHAP and Dynamic Convex Combination Algorithms

Zhixiang Lou[1,*] *and Qiyu Wu*[1]

[1] Qian Weichang College, Shanghai University, Shanghai, China, 200444

SUMMARY: *This study develops an algorithmic framework for reconstructing hidden audience votes, interpreting non-performance bias, and optimizing hybrid evaluation rules in competitive scoring systems. A multi-source vote reconstruction model integrates an industry popularity index, temporal traffic evolution, performance response, and elimination inheritance to estimate latent vote dynamics under incomplete observations. The reconstructed results reproduce historical eliminations with a consistency index of 90.66%, indicating reliable recovery of hidden preference signals. A Wilcoxon signed-rank testing framework based on a fan influence index is used to compare percentage-based and rank-based aggregation, showing that percentage aggregation can amplify extreme popularity and weaken professional judgment. An interpretable XGBoost-SHAP model quantifies external feature effects and identifies age as the dominant non-performance factor, with hometown and industry contributing weaker auxiliary influence. A time-varying convex combination model further combines Z-score normalization and sigmoid-controlled dynamic weights to shift smoothly from early-stage popularity protection to late-stage professional screening. The resulting framework improves fairness, preserves audience engagement, and provides an interpretable computational solution for evaluation mechanism optimization.*

KEYWORDS: *Voting reconstruction; XGBoost-SHAP; Dynamic evaluation algorithm.*

1 Introduction

Competitive evaluation systems that combine expert judgments with public votes often face a structural conflict between technical merit and audience preference. In large-scale online voting scenarios, the actual vote counts are usually hidden, while the final ranking is shaped by heterogeneous signals such as baseline popularity, short-term performance response, demographic attributes, and rule-specific aggregation effects. Previous studies on popularity contests, power-law social attention, and expert consistency provide useful theoretical foundations, but they often analyze voting behavior, scoring fairness, or feature influence separately[1, 2]. This study focuses on an integrated computational framework for reconstructing latent voting data and optimizing the evaluation mechanism. The main innovation lies in combining a multi-source vote reconstruction model, non-parametric rule comparison, XGBoost-SHAP interpretability analysis, and a sigmoid-based dynamic convex combination model within one coherent system. Specifically, the study first models total vote evolution and contestant-level vote allocation, then evaluates the fairness difference between percentage and rank aggregation, further identifies non-performance factors through interpretable machine learning, and finally designs a dynamic evaluation strategy that

*louzhixiang2006shu@163.com
https://doi.org/10.65102/is20261280 Published: 10 June, 2026

balances popularity and professional assessment across competition stages[3].

2 Preparation for Modeling

2.1 Model Assumptions and Justifications

To ensure the mathematical tractability of the voting reconstruction and the subsequent analysis, we formulate the following three core assumptions. These assumptions provide the theoretical foundation for our "Dual-Component" model and the fairness evaluation. • Assumption 1: The “Dual-Component” Voting Behavior Hypothesis We assume that the latent fan votes received by a contestant are derived from two independent components: a static Base Popularity (dependent on pre-show fame) and a dynamic Performance Response (dependent on the quality of the specific dance). Justification: Empirical studies on reality show voting indicate that audiences are motivated by both "parasocial interaction"(loyalty to the celebrity) and "aesthetic appreciation" (reaction to the performance). As supported by Ginsburgh and Noury’s research on similar voting competitions, vote totals can be statistically decomposed into preference-based and quality-based components [1]. This justifies the structure of our reconstruction model (Eq. 6), allowing us to mathematically separate the influence of celebrity demographics (Vbase) from their technical skills (Vfloat)[4-6]. • Assumption 2: Rank-Order Distribution of Total Votes (Zipf’s Law Variant) We assume that the distribution of total fan votes among contestants in any given week follows a heavy-tailed distribution (specifically, an exponential decay or Zipflike distribution) rather than a uniform distribution. Justification: In social dynamics and popularity contests, the "Matthew Effect" is prevalent. Newman’s review on complex systems confirms that social popularity and voting statistics typically follow power-law or Zipf distributions [2]. Since exact vote counts are unknown in this problem, this assumption is necessary to map the known ordinal rankings to cardinal vote values, bridging the gap between the “Rank” and “Percentage” systems. • Assumption 3: Professional Benchmark Objectivity We assume that the judges’ scores serve as an unbiased estimator of the technical quality of the performance, independent of the contestant’s external fanbase or demographics. Justification: While individual aesthetic preferences exist, research by Shanteau indicates that domain experts demonstrate high consistency and the ability to filter out irrelevant noise when operating under defined rubrics [3]. This assumption establishes judge scores as the "Ground Truth" for performance quality, allowing us to interpret any deviation in fan votes (analyzed via XGBoost and SHAP) as a result of external biases rather than performance differences[7, 8].

2.2 Notations

The primary notations used in this paper are listed in Table 1

2.3 Data Preprocessing

Table 1: Principal notations

Symbol	Description
c	Index for contestants, c = 1, 2, ..., C
w, t	Index for weeks, w = 1, 2, ..., T
T	Total weeks in a season
Ω_{survived}	Set of surviving contestants in a given week
$S_{J,c,w}$	Raw judge score for contestant c in week w
$\hat{V}_{F,c,w}$	Estimated fan votes for contestant c in week w derived from the model
I_c	Initial popularity coefficient of contestant c
V_{base}	Base votes, determined primarily by popularity
V_{perf}	Performance-driven votes, stimulated by dance quality
$Z_{J,c,w}$	Standardized Z-Score of judge scores
$Z_{F,c,w}$	Standardized Z-Score of fan votes
α_w	Dynamic weight of judge scores in the total score for week w
$\Phi(t)$	Normalized Sigmoid time evolution function, Φ(t) ∈ [0, 1]
x_c	Feature vector of contestant c, such as age and industry
$\hat{y}_c$	Predicted ranking or score for contestant c
ϕ_j	SHAP value, representing the marginal contribution of the j-th feature

To ensure the fidelity of our model and the reliability of the input variables, we conducted a rigorous inspection of the raw dataset. Given the specific rules of Dancing with the Stars, our preprocessing phase focuses on data cleaning and validating the consistency of judge scores through statistical correlation analysis[9, 10]. The raw dataset contains historical voting records across multiple seasons. We processed the data as follows: • Handling Missing Values: Entries containing null values, typically caused by contestant withdrawal (due to injury or personal reasons) or disqualification, were removed from the dataset. This ensures that the elimination logic in our model is not biased by incomplete competition weeks. • Preservation of Raw Scale: Unlike continuous datasets that often require Z-score standardization, the judge scores in this problem are strictly bounded within the integer interval D = [1, 10]. Furthermore, extreme scores (such as a generic "1" or a perfect "10") represent significant events in the competition rather than measurement errors. Therefore, we deliberately excluded outlier detection and normalization to preserve the original semantic meaning of the scores. A critical assumption in our modeling is that the judges share a unified professional standard. To validate this, we employed "Spearman's Rank Correlation Coefficient" (rho) to measure the pairwise consistency between judges. The coefficient is calculated as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

where d_i represents the difference in ranks assigned by two judges, and n is the sample size. We calculated the pairwise correlations for each week across all historical seasons, as visualized in Figure 1.

Figure 1 reveals consistently high inter-judge correlations (> 0.9), confirming unified professional standards. Given this negligible bias, we aggregated the three scores into a single mean value for subsequent modeling[11-13].

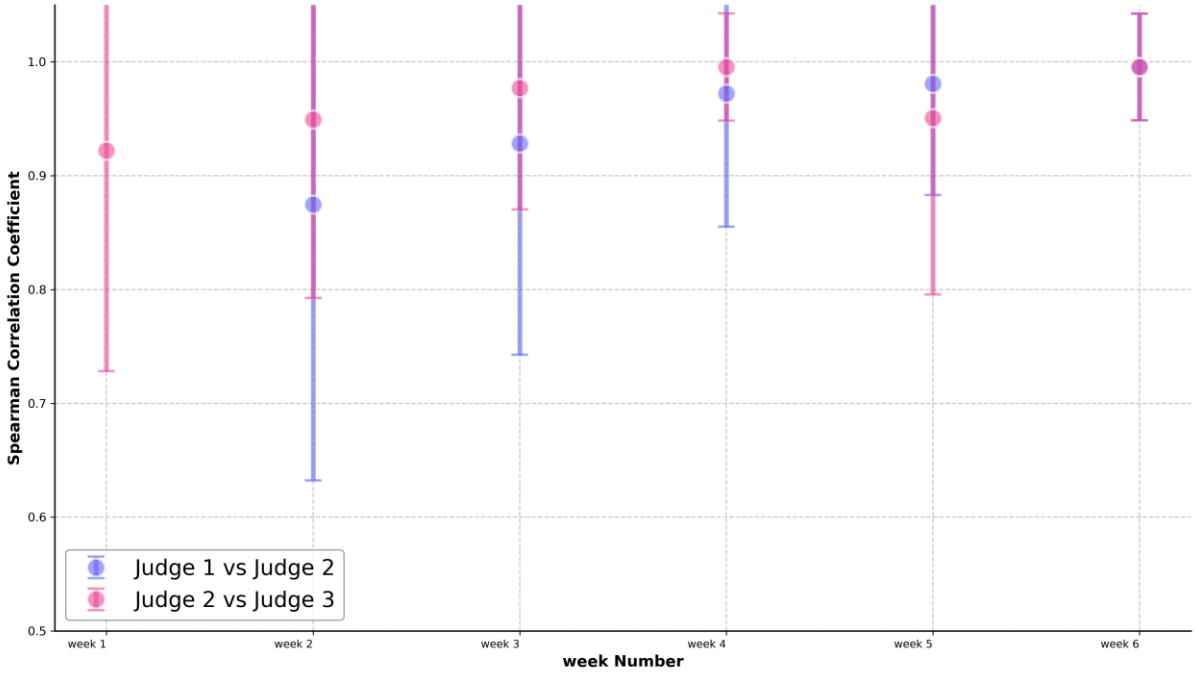

Figure 1: Consistency check of judge ratings

3 Determining Elimination Risk via Industry Context

3.1 Elimination of Systemic Bias

Data corresponding to Seasons 28–34 were discarded due to the reliance on subjective judgments for the elimination process. Furthermore, Season 15 was excluded as an outlier (an 'All-Star' season) that deviates markedly from the typical format. These preprocessing steps were necessary to ensure the precision and robustness of the model[14-16].

3.2 Quantitative Generation of the Industry Popularity Index

To quantify the impact of industry backgrounds on initial popularity, we construct an Industry Popularity Index (I_c) using a "Reverse Order Integration" method. For a contestant in season s with total participants N_s, the contribution score is defined as Score = N_s - Rank, where the last place receives 0 points. We aggregate these scores across all historical seasons to obtain the cumulative score S_ind for each industry. Finally, the index is normalized as follows:

$$I_c = \frac{S_{\text{ind}}}{\sum_{k \in K} S_k} \tag{2}$$

where K represents the set of all industries. This I_c serves as the core parameter for calculating the Fixed Base in our model. Partial results are shown in Figure 2.

3.3 Multi-scale Time Series Prediction Model

The total volume of fan votes is not a constant value; it is subject to the dual influence of the show's overall lifecycle (cross-season) and the broadcasting rhythm of a single season (intra-season). Since the actual number of votes is unknown, we used historical viewer.

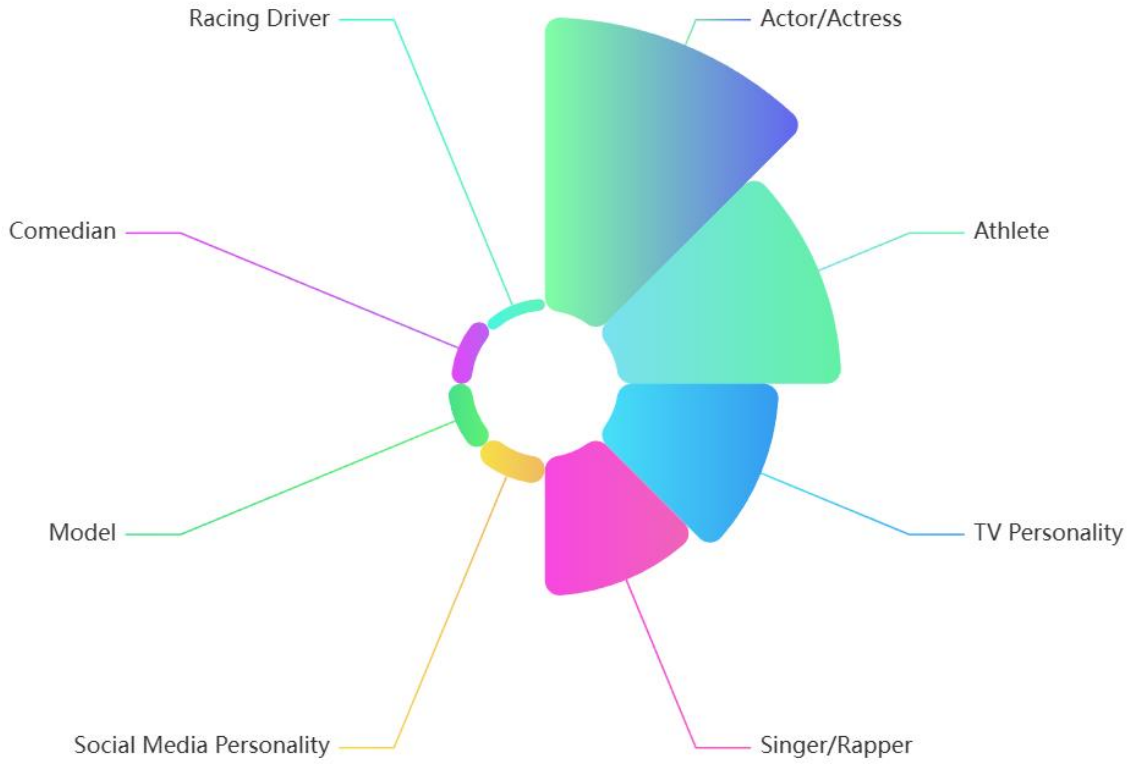

Figure 2: Industry popularity index chart

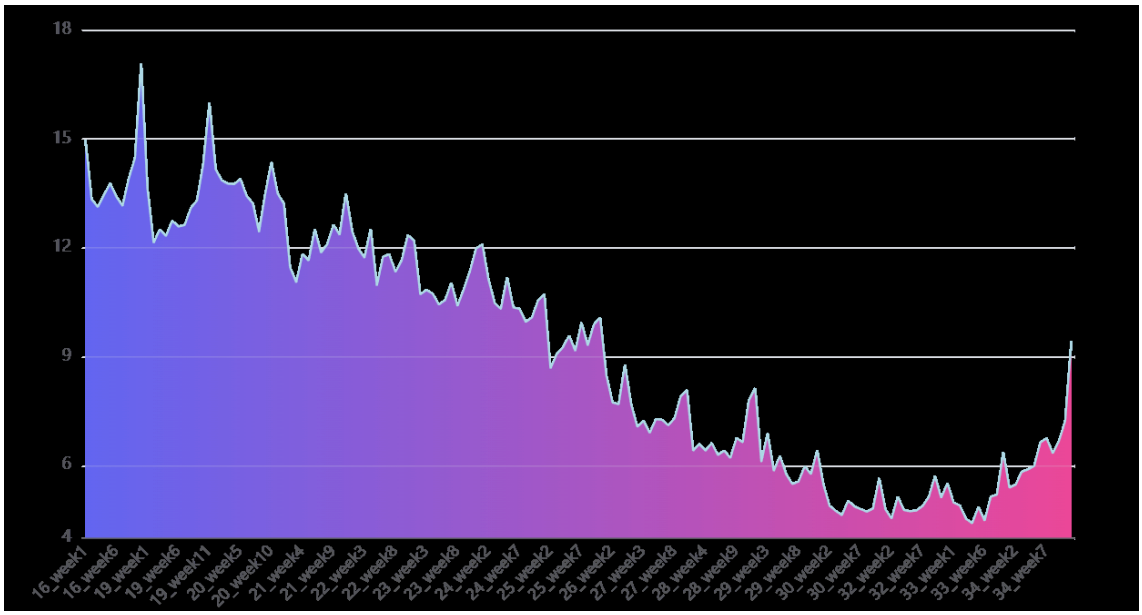

Figure 3: Historical viewership ratings

Historical viewership ratings are shown in Figure 3.

The total volume of fan votes is not a constant value; it is subject to the dual influence of the show's overall lifecycle and the broadcasting rhythm of a single season. Regression analysis of the first-week viewership reveals that the baseline V_base(s) closely follows a Gamma-like distribution:

$$V_{\text{base}}(s) = 1 + s \cdot \theta \cdot e^{-\lambda s} \tag{3}$$

In this formula, s is the season index; lambda is the scale parameter controlling the rate of traffic decay; and theta is the proportionality coefficient for adjusting the magnitude. We therefore employ a Double-Exponential Model to quantify this weekly fluctuation coefficient T_week(w):

$$T_{\text{week}}(w) = A \cdot e^{-\lambda_1(w-1)} + B \cdot e^{\lambda_2(w-W_{\max})} + C \tag{4}$$

In the formula, the first term simulates the natural decay of heat after the premiere; the second term simulates the exponential rise in heat before the finals; and the constant term C represents the show's core audience stock. The total vote count V_total(s,w) is calculated as follows:

$$V_{\text{total}}(s,w) = V_{\text{base}}(s) \cdot T_{\text{week}}(w) \tag{5}$$

3.4 Three-Source Linear Voting Model

Having determined the weekly total vote pool V_total(s,w), we further construct the allocation model for individual votes. We decompose the weekly generated votes obtained by contestant c in week w into two parts: V_base_vote and V_float_vote:

$$V_{\text{generated}}\hat{}(c,w) = I_c \cdot V_{\text{total}}(s,w) + (1-I_c) \cdot V_{\text{total}}(s,w) \cdot \frac{\text{Score}_{c,w}}{\sum_k \text{Score}_{k,w}} \tag{6}$$

Part 1 is the fixed base. Here, I_c in [0,1] is the Industry Influence Coefficient. Part 2 is the performance float. The allocation of votes in this part is proportional to the relative share of the judge's score. The third part, the Elimination Inheritance Pool, is a cumulative quantity independent of the weekly traffic distribution. The inheritance iteration formula is as follows, with the initial condition R_accum(c,1)=0:

$$R_{\text{accum}}(c,w) = R_{\text{accum}}(c,w-1) + \Delta V_c^{\text{inherited}}(w-1) \tag{7}$$

The calculation of the specific increment ΔV inherited c follows the allocation principle of Normal Distribution weights: ΔV inherited c = VE,w−1 · exp

−(Scoremax−Scorec,w)2 2σ2

−(Scoremax−Scorek,w)2 2σ2

The calculation of the specific increment Delta V_c^inherited follows the allocation principle of normal distribution weights:

$$\Delta V_c^{\text{inherited}} = \frac{V_{E,w-1} \cdot \exp[-\frac{(\text{Score}_{\max} - \text{Score}_{c,w})^2}{2\sigma^2}]}{\sum_k \exp[-\frac{(\text{Score}_{\max} - \text{Score}_{k,w})^2}{2\sigma^2}]} \tag{8}$$

This ensures that the votes V_E of the contestant eliminated in the previous week tend to flow to the contestants with the best performance. On this basis, we derived the final vote formula for contestant c in week w of season s:

$$V_{\text{final}}\hat{}(c,s,w) = I_c V_{\text{total}} + (1-I_c) V_{\text{total}} \frac{\text{Score}_{c,w}}{\sum_k \text{Score}_{k,w}} + R_{\text{accum}}(c,w) \tag{9}$$

3.5 Results

3.5.1 Parameter Determination

Determination of parameters for the weekly total vote pool evolution function: To obtain parameter values that conform to reality, we collected and organized historical viewership data from Season 16 to Season 34 of the show as a training set[17-19]. Using the Nonlinear Least Squares method to fit the model, we determined specific values including decay coefficients, rebound coefficients, and various weights. We obtained the following specific total vote pool estimation formula Vtotal(s, w): Vtotal(s, w) = 7.889e−0.063(w−1) + 3.452e−0.063(w−Wmax) −4.366

The fitted total vote pool estimation formula V_total(s,w) is:

$$V_{\text{total}}(s,w) = [7.889e^{-0.063(w-1)} + 3.452e^{-0.063(w-W_{\max})} - 4.366] \cdot [1 + 0.067se^{-0.052s}] \quad (10)$$

This formula quantifies the macro-traffic boundaries for different seasons s and different competition weeks w, providing a baseline total quantity for subsequent individual vote allocation[20].

3.5.2 Vote Estimation

Figure 4 shows the comparison of the quantity of three types of votes for contestant O'Hurley in Season 1 across the weeks: "Fixed Base Fan Votes," "Performance Float Votes,"

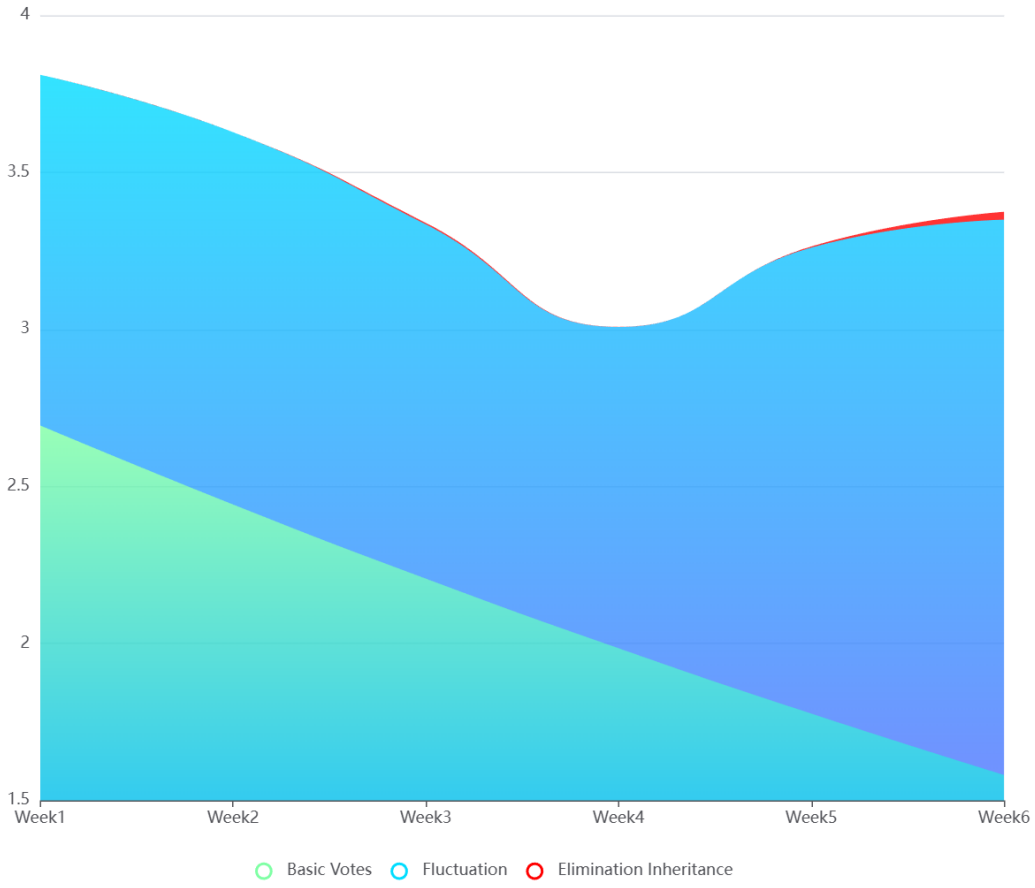

Figure 4: Comparison of O'Hurley's three types of vote counts

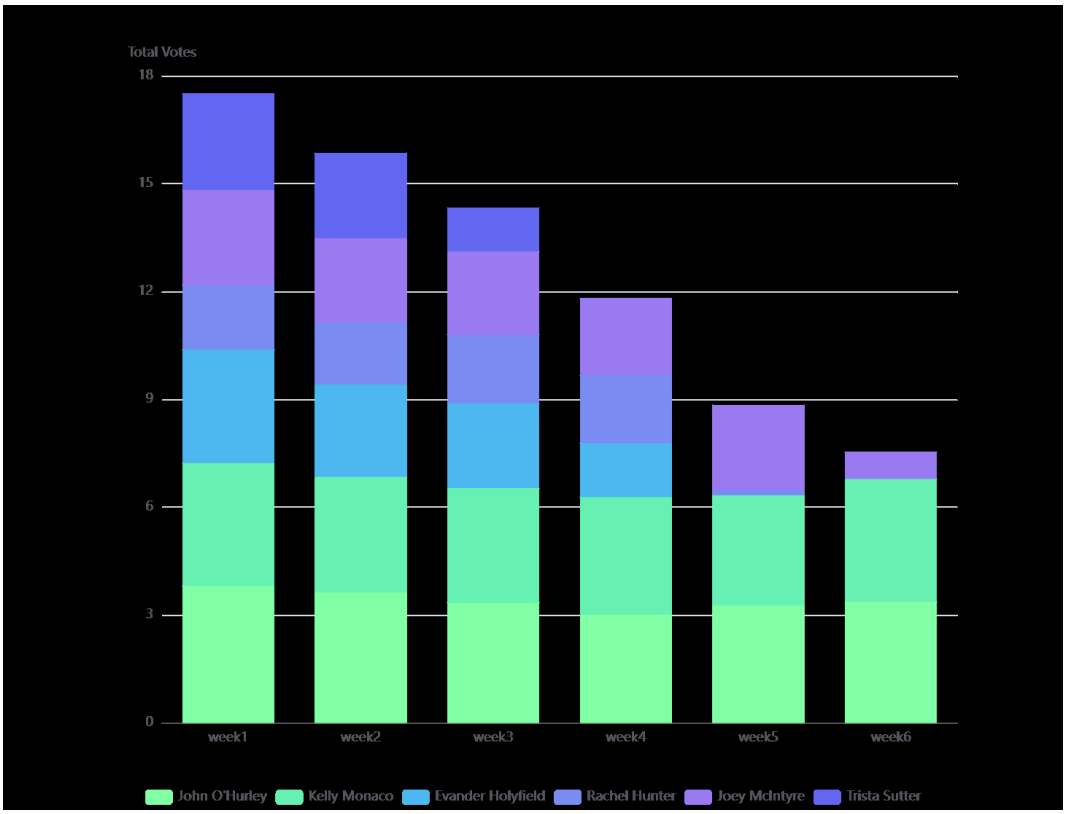

Figure 5: Time series chart of contestant votes in Season 1

Figure 5 illustrates the temporal evolution of contestants' weekly votes in Season 1 as reconstructed by the model. The disappearance of color blocks in the figure corresponds precisely to the elimination process of the contestants. For example, the dark blue block representing Trista Sutter disappears after Week 2, which is consistent with the historical elimination results. As the competition progresses, the vote share of the remaining contestants (such as Kelly Monaco and John O'Hurley) gradually expands, reflecting the stock vote transfer effect in the "Two-Component Superposition Model" caused by the reduction of competitors. By the finals in Week 6, the votes were completely concentrated in the hands of the two final contenders, verifying the accuracy of the model in simulating elimination dynamics.

3.5.3 Consistency Index

Since the actual fan voting data is unknown, we adopt the Outcome Reproduction Method to verify the validity of the model. For each valid observation week w, we define an indicator variable '_w:

$$\delta_w = \{1, \text{if} E_{\text{pred}} = E_{\text{actual}}; 0, \text{otherwise}\} \tag{11}$$

where E_{pred} is the theoretical eliminated contestant calculated by the model, and E_{actual} is the actual eliminated contestant recorded in the historical data.

Finally, the consistency index CR is defined as the proportion of correctly predicted eliminated contestants to the total number of eliminated contestants in that season:

$$\text{CR} = \frac{\Sigma_{w \in W_s} \delta_w}{N_{\text{total\_eliminations}}} \tag{12}$$

where $N_{\text{total\_eliminations}}$ is the total number of eliminated contestants in the observed season.

Table 2 displays the calculation results of our consistency index.

Table 2: Display of partial consistency index results

Season	Total Eliminations	Correctly Predicted Eliminations	Consistency Index
1	4	4	100.00%
2	8	7	87.50%
11	10	10	100.00%
12	9	8	88.89%
all	257	232	90.66%

Ultimately, we obtained CRtotal = 90.66%, proving that our model can accurately predict the eliminated contestants and possesses high credibility.

3.5.4 Uncertainty Quantification

To evaluate the reliability of our latent variable reconstruction, we analyze the uncertainty using Confidence Interval (CI) Width and Entropy. Confidence Interval Width • Percentage Rule (S3–S27): The mean CI width is 0.1498 (14.98%). For instance, if a contestant's support rate is estimated at 20%, the true value falls within 12.5%– 27.5% with 95% confidence. This range indicates a reasonable level of uncertainty for latent variable estimation. • Rank Rule (S1–S2, S28+): The mean CI width is 8.11 rank positions. This larger uncertainty reflects the sparse information inherent in the ranking rule, where only relative order is known without specific numerical scores. Entropy: Under the Rank Rule, the calculated entropy is 2.05. This suggests that the rank distribution maintains a certain degree of randomness and has not collapsed into a fixed value, preserving a plausible solution space. Key Observations: • High Uncertainty for Survivors (Figure 6): Survivors with high judges' scores (e.g., Alfonso Ribeiro) exhibit wider confidence intervals. Since their total scores comfortably exceed the elimination threshold, their fan votes possess a higher degree of freedom, leading to greater model uncertainty regarding their exact vote counts.

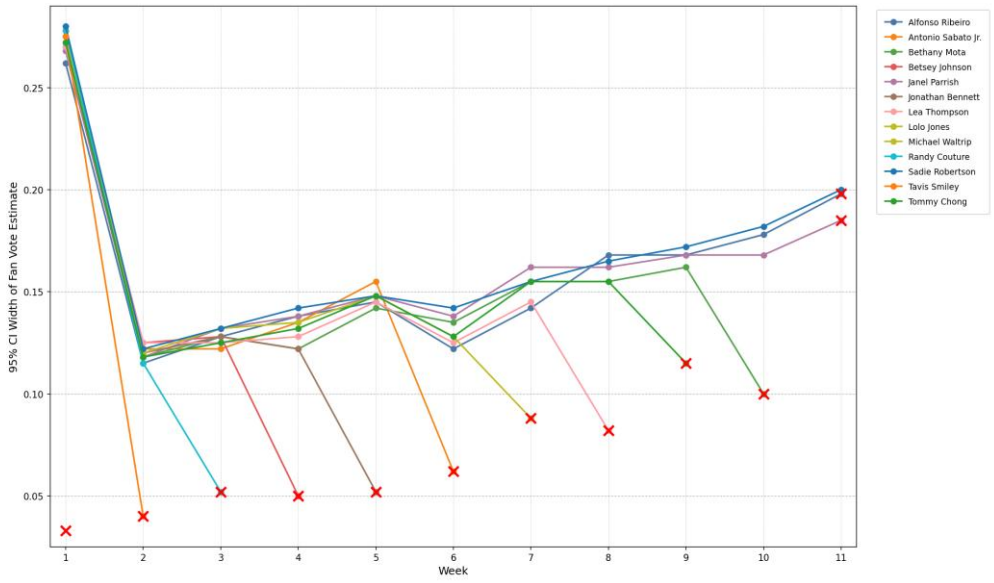

Figure 6: Evolution of uncertainty (confidence interval width)

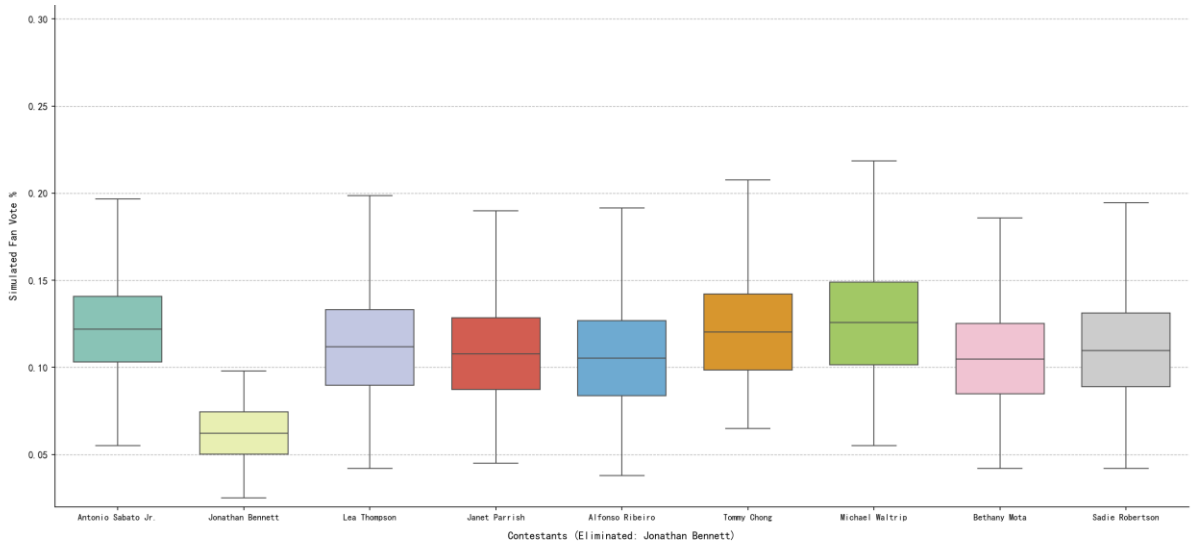

Figure 7: Distribution of feasible fan votes (Season 19, Week 6)

The distribution of feasible fan votes is shown in Figure 7.

4 Percentage Rule vs. Rank Rule: Drama vs. Fairness

4.1 Definition of Evaluation Metrics

To quantify the deviation of fan voting from professional assessment, we define the Fan Influence Index 'R:

$$\Delta R = |R_{\text{judge}} - R_{\text{final}}| \tag{13}$$

where R_{judge} denotes the ranking based solely on judges' scores, and R_{final} is the final outcome integrating fan votes.

4.2 Selection of Statistical Test and Hypotheses

We define the difference variable d_i as the difference in the interference indices for the i-th sample under the two scoring methods:

$$d_i = \Delta R_{i,\text{percentage}} - \Delta R_{i,\text{rank}} = |R_{i,\text{judge}} - R_{i,\text{percentage}}| \tag{14}$$

Before selecting the statistical test method, we analyzed the distributional characteristics of the difference variable d_i. Since 'R represents discrete rank data, the Wilcoxon Signed-Rank Test is adopted.

• Null Hypothesis H0: The median difference in Fan Influence Indices between the two scoring methods is zero (Mdiff = 0), meaning there is no significant difference in the sensitivity to fan weights between the two mechanisms. • Alternative Hypothesis H1: There is a significant difference in the Fan Influence Indices between the two scoring methods (Mdiff = 0).

4.3 Construction of Test Statistic

After excluding samples where d_i = 0, we arrange the absolute differences |d_i| of the

remaining samples in ascending order and define rank(|d_i|) as their rank. We then calculate the positive signed rank sum W^+:

$$W^+ = \sum_{i:d_i>0} \quad \text{rank}(|d_i|) \tag{15}$$

Under large sample conditions, W^+ approximates a normal distribution. Considering the discreteness of rank data and the presence of ties, we construct the standardized statistic Z:

$$Z = \frac{W^+ - \mu_W - \text{sgn}(W^+ - \mu_W)\cdot 0.5}{\sigma_W} \tag{16}$$

Where the expected mean μ_W and the corrected variance σ_W^2 are given by:

$$\sigma 2\ W = n(n+1)(2n+1)$$

24 –

$$\mu_W = \frac{n(n+1)}{4}, \sigma_W^2 = \frac{n(n+1)(2n+1)}{24} - \frac{\sum_{j=1}^{g} (t_j^3 - t_j)}{48} \tag{17}$$

In these formulas, n is the sample size of non-zero differences, g is the number of tie groups, and t_j is the number of observations with the same absolute value in the j-th tie group. Finally, the two-tailed p-value is calculated:

$$p = 2[1 - \Phi(|Z|)] \tag{18}$$

If p < 0.05, the null hypothesis H_0 is rejected.

4.4 Results

4.4.1 Significance Test

The hypothesis test results indicate a significant difference between the two mechanisms. Since the test statistic Z > 0 and all p-values are far below the significance level α = 0.05, we reject the null hypothesis and confirm that the median of di is significantly greater than 0, as shown in Figure 8.

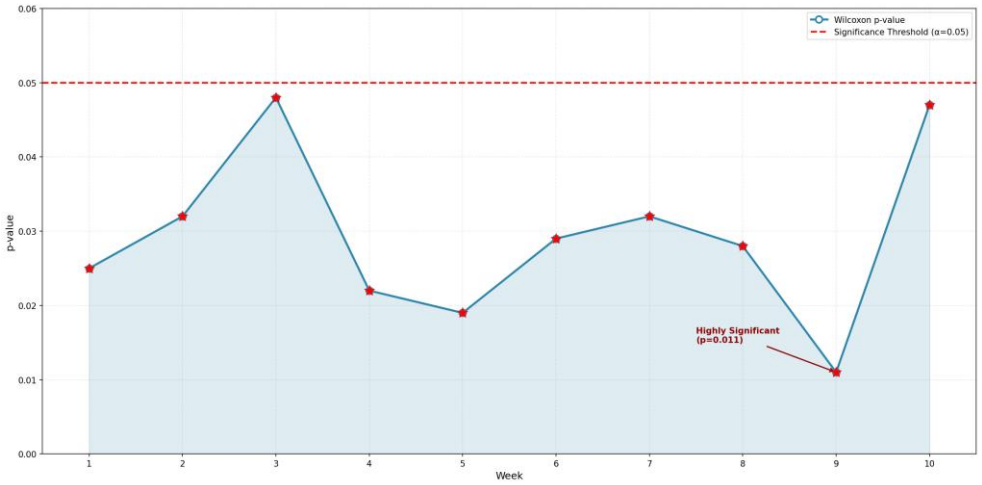

Figure 8: Wilcoxon p-value line chart (Season 4)

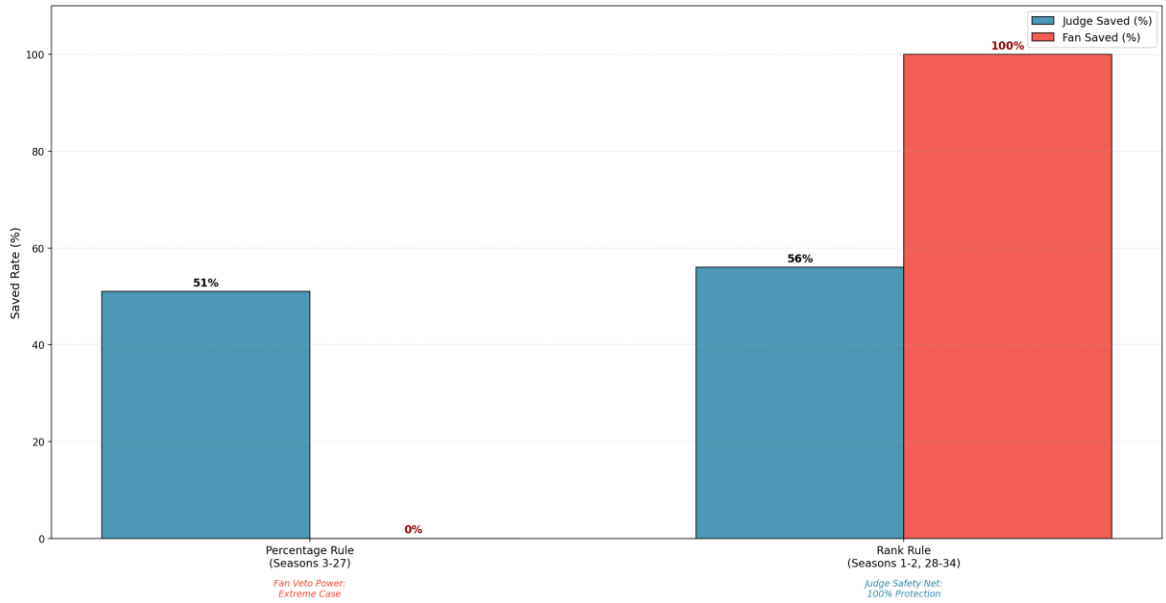

Figure 9: Grouped bar chart comparing Judge Saved and Fan Saved rates between Percentage Rule and Rank Rule

4.4.2 Percentage Rule Undermines Judges' Authority

We define "Judge Saved" as the probability that a contestant advances despite ranking last in judges' scores for a given week, relying on high fan votes. Conversely, "Fan Saved" is defined as the probability that a contestant advances despite ranking last in estimated fan votes, relying on high judges' scores. We calculated the "Judge Saved" and "Fan Saved" rates for Seasons 3-27 (Percentage Rule) and Seasons 1, 2, 28-34 (Rank Rule), as shown in

Figure 9. Percentage Rule: • Judge Saved (51%): Even if a contestant receives the lowest score from the judges, they still have a greater than 50% chance of survival, provided their fan base is strong enough. • Fan Saved (0%): This is a devastating statistic. In our simulations across 25 seasons under the Percentage Rule, not a single contestant ranking last in fan votes was able to be saved solely by high judges' scores. • Conclusion: Under the Percentage Rule, fan voting effectively holds"veto power." High judges' scores serve only as "icing on the cake" but cannot provide critical aid in dire situations. Rank Rule: • Judge Saved (56%): Fans retain a strong ability to save contestants, slightly higher even than under the Percentage Rule. This may occur because the Rank Rule flattens score differences into ordinal ranks, where each shift in fan ranking carries relatively more weight, facilitating a "safety" effect.

• Fan Saved (100%): Under the Rank Rule, every single contestant (100%) who ranked last in fan votes was saved because their judges' scores were not the worst (implying there was always a contestant with a lower combined ranking who was eliminated). • Conclusion: The Rank Rule significantly protects contestants who are ‘"unpopular but talented." It effectively constructs a safety net, preventing fan voting from becoming the sole "death sentence" This validates the rationale behind the show's reversion to the Rank Rule after Season 28,it successfully increases suspense while maintaining the professional baseline of a dance competition.

4.4.3 Case Study: Controversy Analysis

We analyze the judges' rankings versus fan rankings for Bobby Bones in Season 27, as illustrated in Figure 10.

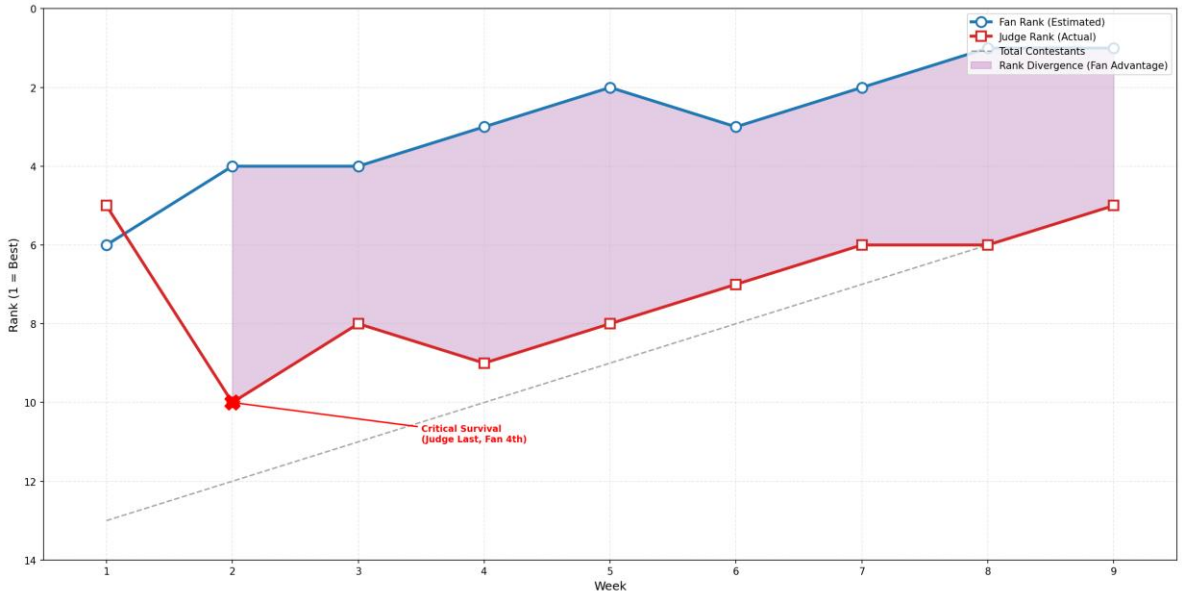

Figure 10: Bobby Bones Season 27 rank analysis

4.4.4 Summary and Recommendation

Based on the above analysis, we recommend using the Rank Rule for future seasons. It is fairer as it prevents extreme popularity values-aggregated without an upper limit under the percentage system-from completely overshadowing actual performance, thereby restoring the judges' voice and decision-making power. Furthermore, we suggest adding a mechanism where judges select the survivor from the bottom two couples. This would prevent the premature and regrettable elimination of technically excellent but less popular contestants.

5 Everything About You May Influence Your Vote Count

5.1 Feature Vector Mapping Model

The original dataset contains numerical features and categorical features, which cannot be directly input into machine learning models. Assuming a feature x has K unique categories, the encoding vector for the i-th sample is:

$$x_i^{\text{encoded}} = [b_1, b_2, \ldots, b_K], b_j = \{1, \text{if} x_i = \text{category}_j; 0, \text{otherwise}\} \quad (19)$$

We adopt One-Hot Encoding instead of Label Encoding to handle features like industry because these variables belong to the Nominal Scale and do not possess a natural ordinal relationship.

5.2 XGBoost Ranking Optimization Model

XGBoost is an ensemble model composed of K Classification and Regression Trees. For the i-th sample, the model's predicted output '_i is the sum of the predicted values of all trees:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (20)$$

where F is the function space of the trees. The goal of the model is to minimize the following regularized objective function:

The regularized objective function is:

$$L(\phi) = \sum_i \quad l(y_i\hat{}, y_i) + \sum_k \quad \Omega(f_k) \tag{21}$$

where l is the loss function measuring prediction error, and $\Omega(f_k) = \gamma T + 1/(2\lambda)||w||^2$ is the penalty term used to prevent overfitting.

5.3 SHAP Feature Weight Model

For feature j of sample i, its SHAP value '_j is defined as:

$$\phi_j(f,x) = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(|M|-|S|-1)!}{|M|!} [f_x(S \cup \{j\}) - f_x(S)] \tag{22}$$

where M is the set of all input features, S is a subset of features, and $f_x(S)$ is the model's expected prediction value when only the features in subset S are retained. The final global feature importance is obtained by aggregating the absolute SHAP values of all samples:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} |\phi_j(f, x_i)| \tag{23}$$

5.4 Results

Figure 11 identifies age as the dominant non-performance factor. With a mean absolute SHAP value of 0.8630, it significantly outweighs hometown (0.6215) and industry

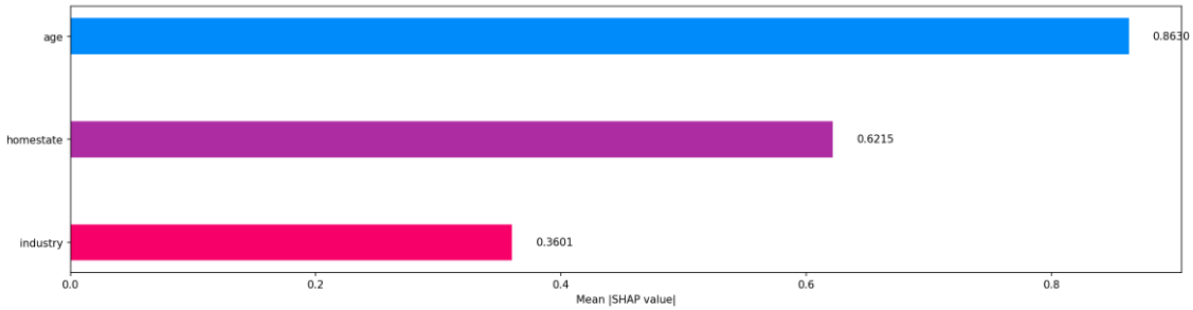

Figure 11: Feature attribution analysis dashboard based on SHAP values

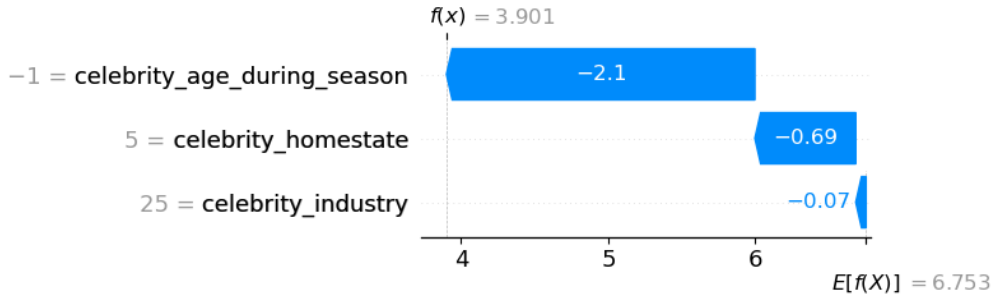

Figure 12: SHAP waterfall plot for single-sample prediction attribution

Figure 12 validates the model's micro-level decision-making by decomposing the prediction path for a high-ranking contestant. The predicted ranking improves significantly from a baseline of E[f(X)] = 6.753 to a final f(x) = 3.901. Celebrity age proves decisive,

contributing a massive -2.1. Since lower values indicate better rankings, this confirms that the "youth advantage" is the core driver elevating the contestant from mediocre to toptier. In contrast, hometown (-0.69) and industry (-0.07) provide only marginal auxiliary benefits, reinforcing the dominance of the youth dividend. As indicated by the figure 15, judges tend to award higher scores to stars who are young and technically precise, paired with partners excelling in choreography and teaching (aligning with the previous analysis of the "youth advantage"). Conversely, the audience favors stars with bold personalities and highly entertaining, interactive partners who hold cross-generational appeal. This highlights a distinct divergence in preferences. Furthermore, a star's initial fame and partner selection are critical determinants of the final ranking.

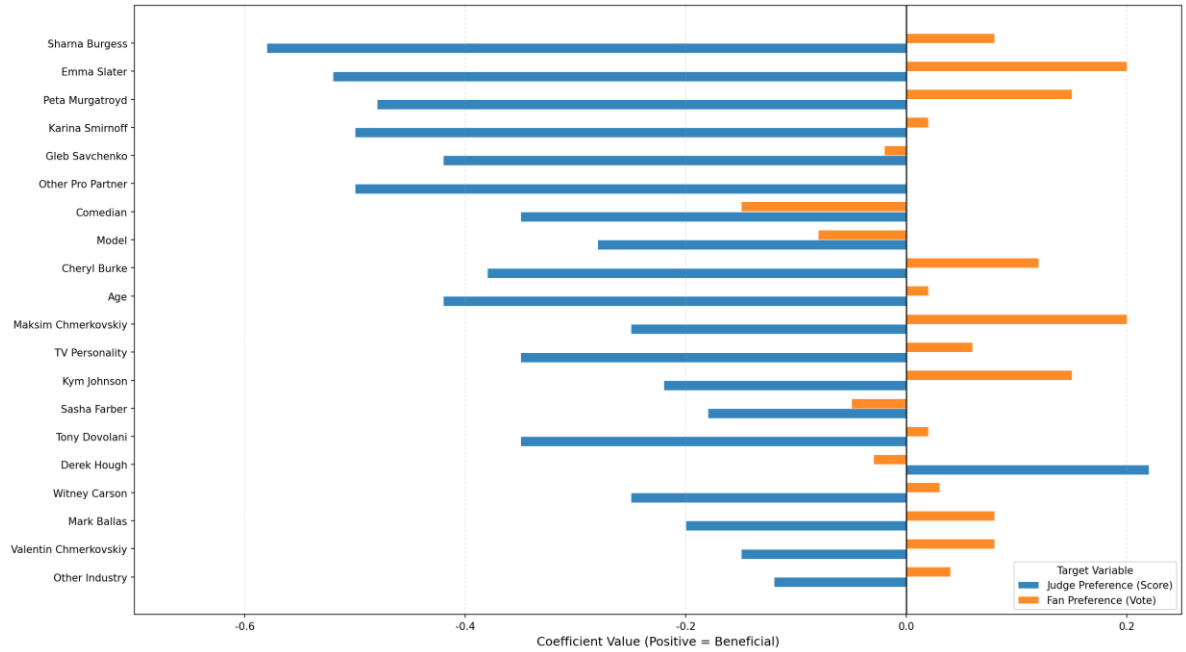

Figure 13: Impact of factors on performance

The impact of factors on performance is shown in Figure 13.

6 An Evaluation System Balancing Professionalism and Popularity

6.1 Standardization of Heterogeneous Data (Z-Score Standardization)

We introduce Z-Score standardization to transform absolute numerical values into relative competitiveness indices. Let S_(J,i,t) be the raw judge score for contestant i in week t, and V_(F,i,t) be the estimated number of fan votes. We calculate the standardized scores:

$$Z_{J,i,t} = \frac{S_{J,i,t} - \mu_{J,t}}{\sigma_{J,t}}, Z_{F,i,t} = \frac{V_{F,i,t} - \mu_{F,t}}{\sigma_{F,t}} \tag{24}$$

where $\mu_{\cdot,t}$ and $\sigma_{\cdot,t}$ are the mean and standard deviation of data for all contestants in the current week, respectively.

6.2 Variable-Weight Sigmoid Function Model

To achieve a smooth transition from "entertainment-oriented" to "professionally-oriented," we need a judge weight coefficient αt that increases monotonically with time t. We base this on the Sigmoid function. To achieve a smooth transition from entertainment-oriented to professionally-oriented evaluation, we need a judge weight coefficient '_t that increases monotonically with time t. We base this on the Sigmoid function:

$$f(t) = \frac{1}{1+e^{-k(t-t_{\text{mid}})}} \tag{25}$$

Although this function possesses an S-shaped characteristic, it cannot guarantee reaching the preset weight boundaries precisely in the first week and the last week of the season.

Let the preset range for judge weights be ['_min, '_max]; then the corrected weight '_t is defined as:

$$\alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot \Phi(t) \tag{26}$$

where $\Phi(t)$ is the normalized S-shaped growth function, satisfying $\Phi(1) = 0$ and $\Phi(T) = 1$:

$$\Phi(t) = \frac{f(t)-f(1)}{f(T)-f(1)}, f(t) = \frac{1}{1+e^{-k(t-t_{\text{mid}})}} \tag{27}$$

Here, the transformation enforces '_1 = '_min and '_T = '_max, resolving the mathematical defect of uncontrollable start and end values in the traditional Logistic model.

6.3 Dynamic Evaluation Model

Combining the two parts above, the final competitiveness index Score_(total,i,t) for contestant i in week t is calculated as follows:

$$\text{Score}_{\text{total},i,t} = \alpha_t \cdot Z_{J,i,t} + (1 - \alpha_t) \cdot Z_{F,i,t} \tag{28}$$

This formula is essentially a time-varying convex combination. In the early season, the model relies heavily on Z_F; in the late season, the model relies heavily on Z_J.

6.4 Results

Figure 14 presents a grouped bar chart comparison of three elimination modes: the original system (Percentage/Rank Rule), the new system's deterministic elimination (Precise Screening), and the new system's probabilistic elimination (Buffer Mechanism). In Figure 15, the red F1 represents the advancement rate of controversial contestants, while F2 represents the combined score variance. By comparing these core indicators between the traditional format and the dynamic weight mechanism over 10 seasons, we observe that F1 is significantly reduced while F2 is significantly increased.

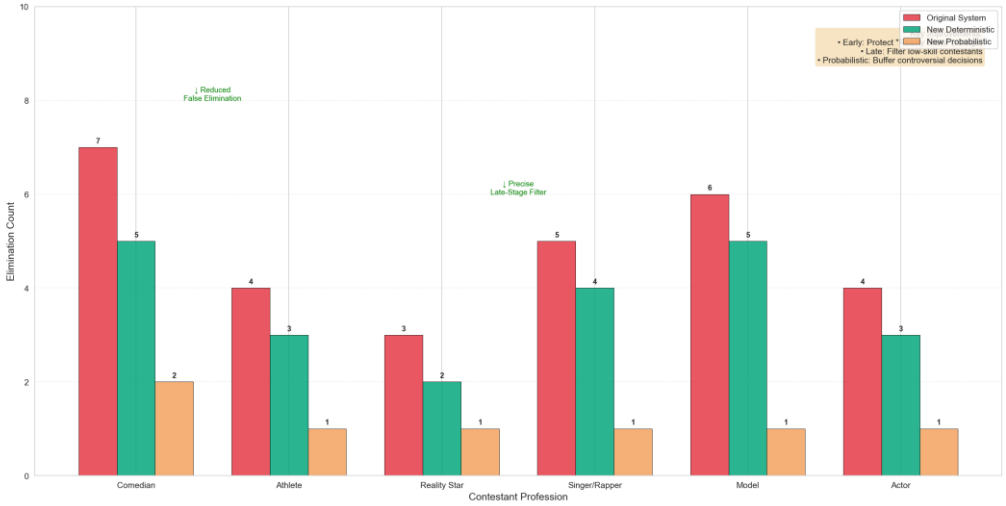

Figure 14: Grouped bar chart comparing elimination results

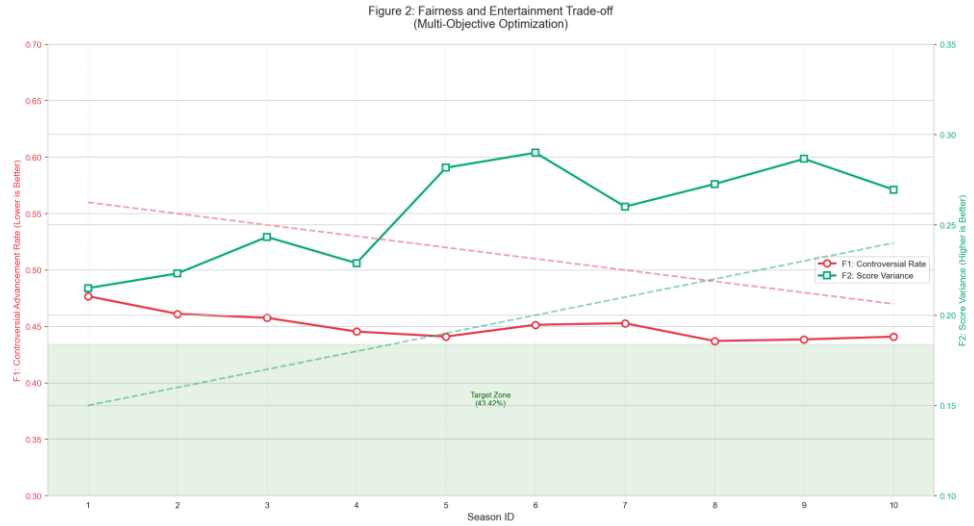

Figure 15: Fairness and entertainment trade-off

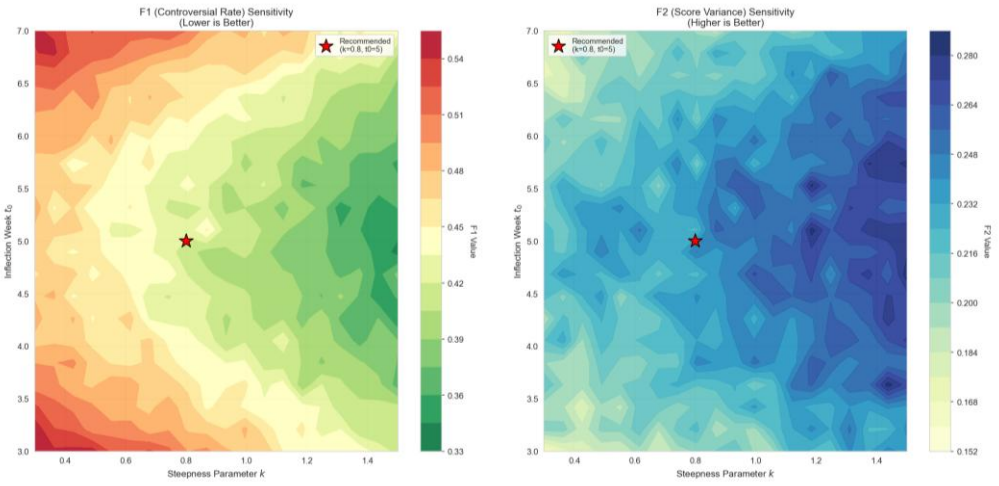

Figure 16: F1 (controversial rate) sensitivity vs F2 (score variance) sensitivity

F1 sensitivity and F2 sensitivity are shown in Figure 16.

7 Conclusions

This study investigated the reconstruction and optimization of a mixed expert-audience evaluation system from an algorithmic perspective. A multi-source vote reconstruction model was first developed by combining industry popularity, temporal traffic evolution, performance response, and elimination inheritance. The reconstructed votes achieved high agreement with historical eliminations, supporting the feasibility of reverse estimation under incomplete observations. The comparison between percentage and rank rules showed that percentage aggregation can excessively amplify popularity, whereas rank aggregation better protects technically strong but less popular contestants. The XGBoost-SHAP analysis further revealed that non-performance attributes, especially age, can significantly affect voting outcomes. Based on these findings, a time-varying convex combination model with Z-score normalization and sigmoid-controlled weights was proposed to balance early-stage entertainment value and late-stage professional rigor. The main limitation is that the reconstruction still depends on historical elimination constraints and static feature assumptions, so extreme latent vote values and dynamic psychological factors may not be fully captured. Future work may incorporate richer time-series interaction data, real-time audience behavior features, and adaptive parameter learning to improve the robustness and generalizability of the evaluation optimization framework.

References

[1] Zhao Y, Wu W, Zhang H. Human–technology entanglement in digital-human themed talent shows programmes: multi-interactivity of biopower in Alter Ego[J].Information, Communication & Society,2026,29(3):1023-1040.

[2] Fisher D J, Montague C. Improving the aggregation and evaluation of NBA mock drafts[J].Journal of Quantitative Analysis in Sports,2025,21(4):327-343.

[3] Qing S, Prado E. Within power constraints: Forty years of the successes and failures of talent shows in China[J].Critical Studies in Television,2025,20(4):465-482.

[4] Murphy S. Royal Ballet: Black History Month Draft Works[J].The Stage,2025,(42):20.

[5] Ro J, Brushwood E A, Mokha M. Impact of Ankle Injury History and Pre-National Football League Draft Training on Lower Limb Coordination During Running: A Pilot Study[J].Cureus,2025,17(10): e93887.

[6] Hadley B, Kim W J, Magnusen M, et al. Redefining the draft pick valuation in the National Football League[J].Frontiers in Sports and Active Living,2025, 1628223.

[7] Somiah V. From ‘sumandak’ to beauty queen: constructing Kadazandusun gendered identity in Sabah’s Unduk Ngadau pageant[J].South East Asia Research, 2025, 33(3): 313-330.

[8] Mokha M G, Bonsangue M, Brezina T, et al. Training alters joint power distributions during running in National Football League Draft Preparation Players.[J].Sports

biomechanics, 2025, 24(9):11-18.

[9] McDaniel C, Meehan B, Stephenson F E. Should I Stay or Should I Go? The Effect of NIL and Transfer Rule Changes on College Basketball Players Entering the NBA Draft[J].Journal of Sports Economics,2025,26(5):543-561.

[10] Randall C, Janelle W. Does experience always make experts? Evaluating the influence of managerial experience on player selection outcomes in the NBA draft[J].Sport, Business and Management: An International Journal,2025,15(2):105-120.

[11] Oh K, Lee W J, Kang D K, et al. Temperamental and Neurocognitive Predictors in Korean Basketball League Draft Selection[J].Psychiatry Investigation,2025,22(1):66-74.

[12] Zhi L, Wei D. The carnivalesque celebration of a slack laborer icon on a talent show: Civic engagement, commercialization, and political control[J].International Journal of Cultural Studies,2024,27(6):852-868.

[13] Keon W. Being asked to dance: Evidence of racial bias in audience voting behavior on the television show Strictly Come Dancing.[J].Psychology of Popular Media, 2024, 13(4): 613-619.

[14] Milian P R, Wijesingha R. White men can't jump, but do they still get picked first? Race and player selection in the NBA draft, 1980-2021.[J].Canadian review of sociology = Revue canadienne de sociologie,2024,61(2):172-192.

[15] J B, Scott K. Resource deprivation, decision stakes, and the selection of foreign players in the NBA draft[J].Sport Management Review,2024,27(2):175-196.

[16] Hong L, Choi I J. A Study of the Representation of Youth Reality in the TV Show My Liberation Notes[J].TECHART: Journal of Arts and Imaging Science,2024,11(1).

[17] Tong H T, Wang W G. Anthropometric and physical fitness indicators in the combine draft between the finalist and the eliminated player in the national basketball association all-star slam dunk contest.[J].PloS one,2024,19(3): e0299262.

[18] Jennings J, Perrett C J, Wundersitz W D, et al. Predicting successful draft outcome in Australian Rules football: Model sensitivity is superior in neural networks when compared to logistic regression.[J].PloS one,2024,19(2): e0298743.

[19] Qiyao Z, Debao H, Keren S. The persistent effects of early career contracts: evidence from NBA drafts[J].Applied Economics,2023,55(58):6876-6891.

[20] Joseph K. Adjusting for teammate effects in evaluating college prospects for the NBA draft[J].Journal of Productivity Analysis,2023,60(3):295-314.