

Data-Driven Analysis and Optimization of Hybrid Scoring Mechanisms in Competitive Reality Shows Using Dancing with the Stars as a Case

Zhengjia Xiao

*Haide College, Ocean University of China, Qingdao, China
sftdr0@gmail.com*

Abstract. Competitive reality shows often face a problem where absolute popularity outweighs professional merit due to opaque fan voting structures. We analyzed this imbalance by reconstructing hidden vote distributions and testing different scoring rules. We deployed an NSGA-II algorithm combined with a VIKOR model to reconstruct the underlying fan votes. The reconstructed votes matched over 90% of recorded elimination outcomes. It compared scoring methods using Spearman's rank correlation coefficient and an Entropy-weighted TOPSIS model. A percentage-based scoring system reflected audience preferences closer than the traditional ranking method and achieved a comprehensive score of 0.5796. To examine what drives judges' versus fans' votes, it trained an XGBoost-SHAP model. The results showed that judges base their scores on professional metrics, while broader social popularity factors dictate public fan voting. It proposes a Log-Dynamic Dual-Track Exchange System (LDS) that mathematically limits the marginal utility of extreme vote accumulation. Applying this system to historical data reduced ranking discrepancies by 30% to 40%. The validation of this framework currently relies on retrospective simulation rather than live audience environments.

Keywords: Hybrid scoring mechanisms, NSGA-II algorithm, entropy-weighted TOPSIS, XGBoost-SHAP model

1. Introduction

The persistent tension in modern voting systems stems from a delicate friction between information aggregation efficiency and participation equity. While Battaglini et al. provide a foundational experimental comparison showing that sequential rules can enhance information flow, their conclusions are inherently limited by a controlled laboratory setting that fails to account for the irrational, high-stakes emotional drivers dominant in reality competitions [1]. This gap is particularly glaring in "proxy" environments where, contrary to Maug's theory that public markets can resolve conflicts of interest through transparency, the lack of independent oversight in televised voting often allows sheer popularity to systematically hijack professional merit [2]. Even the secure voting frameworks proposed by Jayakumari et al. and El-Gburi et al.—which are technically solid in their use of cloud-based blockchains and El-Gamal encryption to prevent vote alteration—focus

almost entirely on data integrity [3, 4]. Missing from that picture is a hard look at the structural unfairness of the scoring logic itself, a flaw that persists even when systems are perfectly authenticated by multimodal biometrics [5].

To navigate these "black box" voting distributions, scholars have pivoted toward multi-objective optimization, yet a significant mismatch remains between algorithmic search and social reality. Hao et al. successfully improved the NSGA-II algorithm's global search capability through Levy distributions, but their focus remains strictly on engineering efficiency rather than the subjective nuances of fairness [6]. Similarly, while Bekele and Nicklow's application of NSGA-II for watershed calibration proves the algorithm's utility in high-dimensional parameter estimation, such models are rarely adapted to the stochastic task of reverse-deducing hidden public sentiment [7]. This methodological isolation means that even robust optimization tools, such as those integrated with TOPSIS by Liu et al. or evaluated through dynamic matrices by Kumari and Acherjee, remain stubbornly centered on thermodynamic or industrial outputs [8, 9]. They largely ignore the socio-psychological drivers that define an "optimal" balance between professional judgment and public desire.

It is tempting, then, to reach for interpretable machine learning as a diagnostic tool—and indeed it provides one. But it introduces a new set of contradictions that are often glossed over. Zhang and Zhao report a surprisingly high 91.0% accuracy in their corporate ESG rating setting, but this precision raises the concern that the model is merely "learning" and validating the historical biases embedded in the training data, instead of identifying true merit [10]. Although Li et al. use SHAP to successfully reveal nonlinear effects in travel behaviors, and Wang et al. rely on built-in importance methods for massive datasets, a major oversight in the current literature is the failure to realize that diagnosis does not equal regulation [11, 12]. A diagnostic tool can identify that celebrity status—rather than technical skill—is the primary driver of success, but it cannot structurally re-engineer the power struggle between an expert's score and a fan's vote, even when classic metrics like Spearman's rank correlation confirm the existence of this bias [13].

These two threads—optimization and interpretability—have so far run in parallel, rarely intersecting in the voting fairness literature. Consequently, existing frameworks can analyze why a system is unfair but offer no mechanism to fix it. This study bridges that gap by using an integrated NSGA-II and XGBoost-SHAP architecture to deconstruct the Dancing with the Stars voting mechanism. Moving beyond mere analysis, we propose a Log-Dynamic Dual-Track Exchange System (LDS) that structurally re-anchors outcomes to actual professional talent.

2. Method

2.1. Data description and notation

Since the dataset used in this study combines official competition results (match scores, elimination paths) with publicly available web resource data containing social influence variables such as social media follower counts and search trends, it is quite appropriately structured.

The key mathematical notations used in this paper are listed in Table 1.

Table 1. Notation used in paper

Symb ol	Descriptions
$V_{i,w}$	Estimated proportion of hidden fan votes.
f_1, f_2	judges' score of contestant i in week t
ρ	Spearman's correlation metrics (encompassing percentage ρ_{Percent} , rank ρ_{Rank} , and their difference $\Delta\rho_s$).
C, V_{total}	Scaling coefficient and total public votes in the LDS core formula.

2.2. Estimation of fan voting distribution based on NSGA-II and VIKOR

To estimate the confidential proportion of fan votes, the NSGA-II algorithm was deployed. Reverse-engineering the exact distribution required balancing two conflicting demands: accurately reproducing historical elimination outcomes while restricting erratic voting spikes between weeks. Accordingly, the algorithm was configured with two primary objective functions: a historical inconsistency penalty based on the deviation between theoretical and actual eliminations, and a temporal volatility penalty minimizing the squared difference in vote proportions between consecutive weeks. Missing empirical data for some early seasons was supplemented using synthetically generated proxy records to establish historical benchmarks. Simulation constraints were implemented to enforce fundamental competition logic, briefly covering vote normalization, boundary limits, and zero-vote rules for eliminated contestants. Generating the Pareto front produced multiple viable voting matrices; to resolve these ties and minimize estimation errors across historical mismatches, the VIKOR model was applied.

2.3. Evaluation of scoring mechanisms using spearman and Entropy-weighted TOPSIS

Structural differences between the ranking method and the percentage method were evaluated by calculating the Spearman rank correlation coefficient (ρ) against the baseline fan rankings. This metric was chosen to measure the association between the strictly ordinal and heavily skewed competition rank data. Comparing the distribution characteristics and mean differences of these correlation coefficients revealed the variance between the two methods and their capacity to reflect actual fan voting behavior. An Entropy-weighted TOPSIS method was then applied, extracting proportional weights directly from the data dispersion to avoid subjective bias. Computing the relative distances of both scoring methods to the positive and negative ideal solutions determined which approach performs better under the competition rules.

2.4. XGBoost-SHAP attribution analysis of judging drivers

To understand what separates professional judgments from audience preferences, an XGBoost model was established to handle the highly non-linear interactions between a contestant's social media presence and their on-stage performance. The predictive architecture integrated feature matrices across three main dimensions: celebrity attributes (such as social media follower counts, search popularity, and state population base), professional partner qualifications (including historical rank and experience), and initial stage conditions (such as early performance scores). SHAP values

were then computed to isolate and compare the specific impacts of individual features—such as search popularity surges or performance scores—on the voting consensus.

3. Results and analysis

3.1. Estimation of fan voting results and multi-objective optimization analysis

Applying the objective functions and constraints to each season yielded a Pareto front that reflects the balance between historical reconstruction accuracy and week-to-week data smoothness. From these non-dominated solutions, the VIKOR model extracted a voting matrix that reproduced over 90% of the historical elimination trajectories, while simultaneously highlighting that contestants on the verge of elimination exhibited highly volatile voting distributions. Taking Season 2 as a representative example, the synthesized estimation of fan voting outcomes is visualized in Figure 1. The ordinate of this picture represents the votes of different players, the abscissa shows the time of the competition, and when the vote is zero, it means elimination.

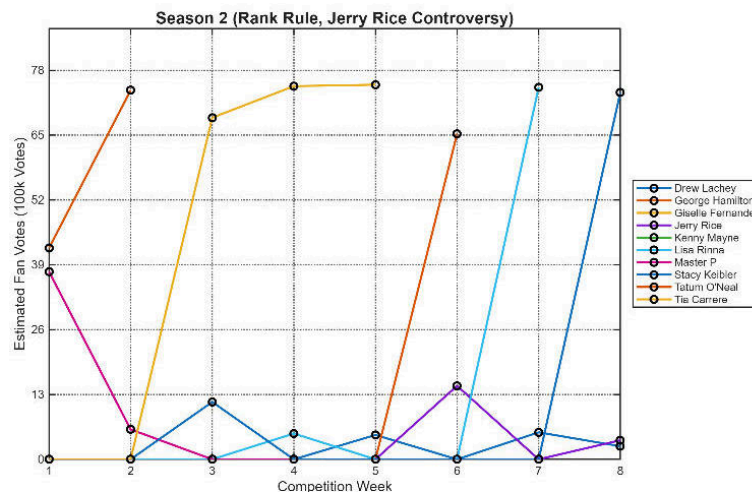


Figure 1. Simulated contestant fan voting results for season 2 (picture credit: original)

From the optimization diagnostics, it is clear that over 90% of the historical contestant elimination trajectories were perfectly reproduced by the deduced fan voting results, and accompanying Monte Carlo simulations further demonstrated that contestants on the verge of elimination had highly volatile voting distributions, thus naturally leading to the conclusion about the instability of the survival threshold under traditional counting methods.

3.2. Discrepancy and rationality evaluation of different judging methods

To establish a systematic evaluation of the two scoring mechanisms, a variance test was first conducted, providing the probability density functions illustrated in Figure 2. This graph shows that the probability density functions of the two voting methods intersect very little, which shows that the correlation between the two voting methods is weak.

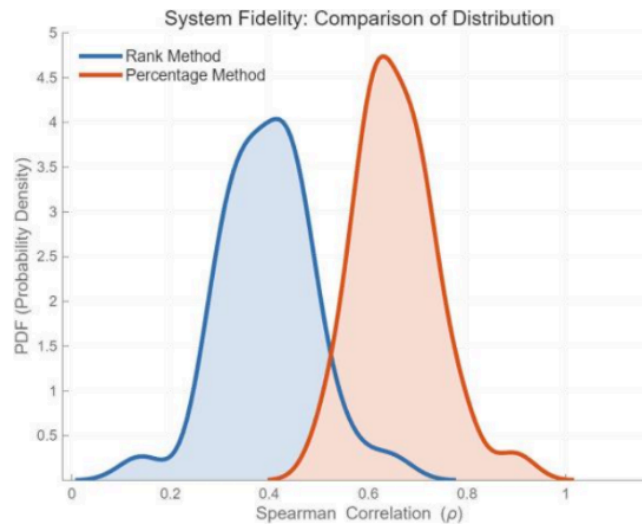


Figure 2. Probability density plots comparing the two judging methods (picture credit: original)

Calculating the Spearman correlation coefficients shows the percentage method ($\rho_{Percent} \approx 0.70$) tracks audience sentiment better than the ranking method ($\rho_{Rank} \approx 0.55$). The correlation difference ($\Delta\rho_s$) remained positive in 32 of the 34 seasons analyzed, indicating that the percentage system tends to reflect true audience preferences. The Entropy-weighted TOPSIS model evaluated both methods across four dimensions. Table 2 gives the resulting weights.

Table 2. Entropy-weighted results (table credit: original)

Evaluation Criterion	Objective Weight (EWM)	Importance
Judge Consist	0.2620 (26.20%)	No.1
Fan Consist	0.2387 (23.87%)	No.4
Controversy Rate	0.2471 (24.71%)	No.3
Suspense	0.2523 (25.23%)	No.2

Since the weights were analyzed within the TOPSIS framework, it was found that the percentage method gave a total score of 0. 5796, which is considerably higher than the ranking method's score of 0. 0273, a very clear and convincing argument can be made: the percentage method is the more appropriate baseline for ballot counting rules.

3.3. XGBoost-SHAP attribution analysis of judging drivers

Since the question is about what factors besides the scoring mechanics contribute to a contestant's success, the global feature importance distribution from the XGBoost-SHAP model makes it very clear: " Initial Gap " and " Early Stage Score " have absolute SHAP magnitudes far exceeding those of all other variables (see Figure 3). The ordinate in this honeycomb chart indicates the SHAP value of the variable (greater than 0 indicates a positive effect), while the color bar indicates the value of the feature itself, with redder indicating a higher value.

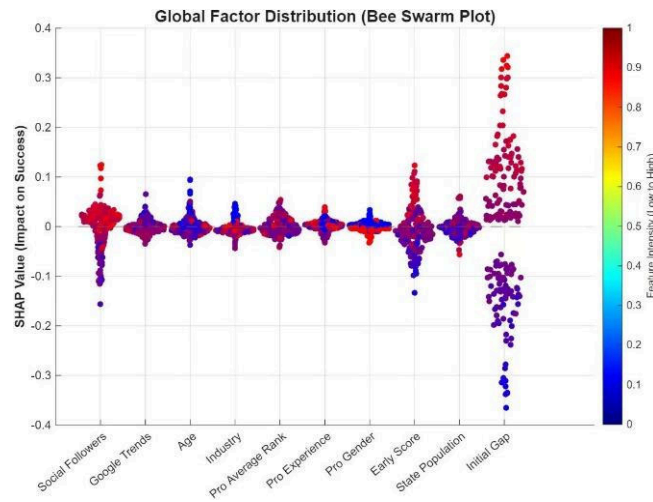


Figure 3. Global factor distribution (picture credit: original)

These early-stage metrics act as the core driving factors determining a contestant's ultimate performance trajectory. When decomposing the heterogeneous preferences between professional judges and audiences, a clear dichotomy emerged: judges' scores are significantly anchored by technical and professional metrics (where early scores exhibited an average SHAP value of ≈ 0.115), whereas fan voting behavior relies heavily on broader social popularity signals, such as Google search trends (average SHAP value ≈ 0.075).

3.4. Corrective efficacy and sensitivity testing of the Dynamic Dual-Track Mechanism (LDS)

Because there have been recurring public controversies when extreme popularity outranks professional competence, this paper naturally and logically proposes a Log - Dynamic Dual-Track Exchange System (LDS), whose operational flowchart is given in Figure 4.

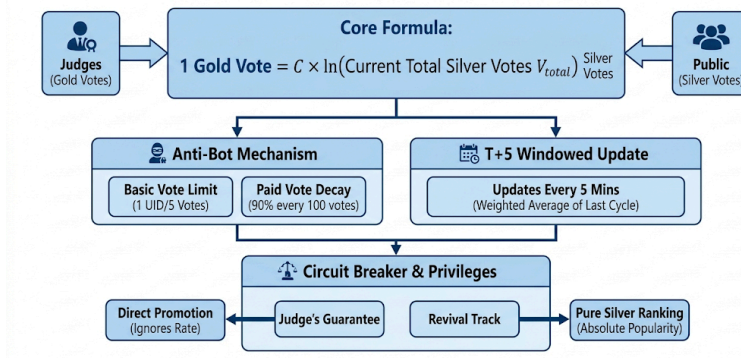


Figure 4. Operational flowchart of the LDS model (picture credit: original)

Simulation tests gave very clear and convincing evidence of the excellent corrective properties of the new architecture: by using a logarithmic function to deliberately suppress the marginal utility of extreme vote accumulation, the LDS system was tested on highly controversial historical cases, and the result was that it re-anchored final rankings to the contestants' actual professional performances, thus reducing ranking deviation by 30% to 40%. More importantly, a rigorous comparative analysis, presented in multi-dimensional radar charts (Figure 5), showed that the LDS system outperforms

traditional weighted aggregation systems in all relevant dimensions: fairness, stability, and internal consistency.

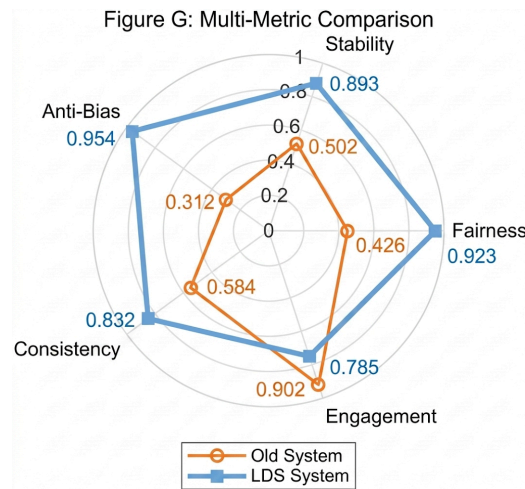


Figure 5. Multi-metric comparison of voting systems (photo/picture credit: original)

4. Conclusion

It reconstructed hidden fan voting distributions using an NSGA-II model, matching over 90% of historical elimination records. TOPSIS and XGBoost-SHAP analyses confirmed that percentage-based scoring tracks audience intent closer than ranking methods, while exposing how initial social popularity dictates outcomes over professional evaluations. We developed the Log-Dynamic Dual-Track Exchange System (LDS) to address this imbalance. This mechanism applies logarithmic scaling to limit the marginal utility of extreme fan voting. Retrospective simulations indicated the LDS reduces historical ranking discrepancies by 30% to 40% compared to weighted aggregation models. The current framework relies strictly on historical proxy records. Its validation is therefore restricted to retrospective simulation. Future implementation must test the LDS in live-voting environments to observe actual audience behavioral shifts under vote decay rules. Researchers could also integrate unstructured sentiment analysis to capture unquantified aesthetic preferences.

References

- [1] Battaglini, M., Morton, R., & Palfrey, T. (2007). Efficiency, equity, and timing of voting mechanisms. *American Political Science Review*, 101(3), 409–424.
- [2] Maug, E. G. (1999). How effective is proxy voting? Information aggregation and conflict resolution in corporate voting contests. *Information Aggregation and Conflict Resolution in Corporate Voting Contests* (March 25, 1999).
- [3] El-Gburi, J., Srivastava, G., & Mohan, S. (2022). Secure voting system for elections. *International Journal of Computer Aided Engineering and Technology*, 16(4), 497–511.
- [4] Jayakumari, B., Sheeba, S. L., Eapen, M., Anbarasi, J., Ravi, V., Suganya, A., & Jawahar, M. (2024). E-voting system using cloud-based hybrid blockchain technology. *Journal of Safety Science and Resilience*, 5(1), 102–109.
- [5] Omoze, S., Omaji, S., & Edegbe, G. N. (2025). Machine learning-based multimodal biometric authentication system (facial and fingerprint recognition) for online voting systems. *ABUAD Journal of Engineering Research and Development (AJERD)*, 8(1), 122–128.
- [6] Hao, J., Yang, X., Wang, C., Tu, R., & Zhang, T. (2022). An improved NSGA-II algorithm based on adaptive weighting and searching strategy. *Applied Sciences*, 12(22), Article 11573.
- [7] Bekele, E. G., & Nicklow, J. W. (2007). Multi-objective automatic calibration of SWAT using NSGA-II. *Journal of Hydrology*, 341(3–4), 165–176.

- [8] Kumari, A., & Acherjee, B. (2025). Comparative effectiveness of data normalization methods in ARAS for multi-criteria decision-making across industrial applications. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 32(4).
- [9] Liu, Q., Zeng, Z., Yang, D., Liu, W., Lu, X., Zhou, L., & Jing, D. (2025). Multi-objective optimization and performance analysis of a CPV/T system based on NSGA-II and entropy weight TOPSIS method. *International Journal of Heat and Mass Transfer*, 255, Article 127810.
- [10] Zhang, J., & Zhao, Z. (2025). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems with Applications*, 295, Article 128809.
- [11] Li, X., Shi, L., Shi, Y., Tang, J., Zhao, P., Wang, Y., & Chen, J. (2024). Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost. *Applied Geography*, 166, Article 103264.
- [12] Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, 11, Article 44.
- [13] Wisniewski, S. J., & Brannan, G. D. (2026). Correlation (coefficient, partial, and Spearman rank) and regression analysis. StatPearls Publishing. <http://europepmc.org/books/NBK606101>